Electronic Research Archive

ERA, 34(6): 3804–3842.
DOI: 10.3934/era.2026172
Received: 23 January 2026
Revised: 11 March 2026
Accepted: 09 April 2026
Published: 11 May 2026

http://www.aimspress.com/journal/era

Research article

Trustworthy and interpretable stacking via TabNet-driven feature generation for breast cancer diagnosis

Lin Xia[1], Yoona Chung[1], Liqiu Suo[1], Jeongmin Hong[1] and Eunchan Kim[1,2,*]

[1] Department of Information Systems, Hanyang University, Seoul 04763, South Korea
[2] Department of Artificial Intelligence, Hanyang University, Seoul 04763, South Korea

*** Correspondence:** Email: eckim@hanyang.ac.kr.

Abstract: Trustworthy analytics for healthcare require models that are not only accurate but also interpretable and robust under distributional perturbations. In this paper, we propose an interpretable stacked ensemble framework that repurposes TabNet from an end-to-end classifier into an attention-guided feature generator for downstream learners. We constructed a dual-channel stacking architecture in which TabNet-derived embeddings and original tabular features were fed into heterogeneous gradient-boosted base learners (XGBoost and LightGBM) to enhance representation diversity, and were integrated by an interpretable logistic-regression meta-learner. For transparent and unbiased evaluation, we employed nested stratified cross-validation with fixed-budget hyperparameter tuning with systematic ablation studies. Experiments on the public Wisconsin Diagnostic Breast Cancer dataset showed that the proposed model achieves strong and stable performance (average accuracy 97.8% ± 1.0% under nested cross-validation) compared to a single TabNet baseline and conventional ensemble variants. Moreover, we assessed robustness under out-of-distribution-style covariate perturbations by injecting Gaussian noise at varying intensities, demonstrating that the stacking design mitigates the noise sensitivity of TabNet-derived representations and maintains a balanced sensitivity–specificity trade-off via adaptive thresholding. Overall, the proposed framework provides a reproducible template for combining deep tabular representation learning with explainable ensemble decision-making toward reliable data science applications in high-stakes domains.

Keywords: computational intelligence; trustworthy analytics; robustness; explainable artificial intelligence; stacking ensemble; TabNet feature generation; healthcare tabular data

Abbreviations: AI: Artificial intelligence; BC: Breast cancer; CAD: Computer-aided diagnosis; CI: Confidence interval; CKD: Chronic kidney disease; FNA: Fine needle aspiration; GBDT: Gradient-boosted decision tree; IARC: International Agency for Research on Cancer; IQR: Interquartile range; LR: Logistic regression; OOD: Out-of-distribution; OOF: Out-of-fold; PCA: Principal component analysis; RF: Random forest; SHAP: Shapley additive explanations; SVM: Support vector machine; WDBC: Wisconsin Diagnostic Breast Cancer; XAI: Explainable artificial intelligence.

1. Introduction

Breast cancer (BC) is one of the most common malignant tumors worldwide and a leading cause of cancer-related mortality among women [1,2]. According to the 2022 Global Cancer Statistics Report released by the International Agency for Research on Cancer (IARC), approximately 2.3 million new breast cancer cases were diagnosed globally, making it the most commonly diagnosed cancer among women [3]. This trend underscores the urgent need for more accurate and efficient diagnostic technologies.

Conventional diagnostic approaches, including mammographic screening and histopathological examination, remain the clinical gold standard; however, these procedures are labor-intensive, time-consuming, and often subject to inter-observer variability [4]. Their accuracy may vary depending on clinician experience, and early-stage or subtle lesions frequently pose diagnostic challenges [5]. To alleviate these limitations, artificial intelligence (AI) driven computer-aided diagnosis (CAD) systems have emerged as promising tools that enhance workflow efficiency and diagnostic consistency [6]. Recent deep-learning-based CAD systems have further automated feature generation and decision support [7], but many rely on handcrafted features or exhibit limited interpretability, restricting their applicability in clinical practice [8].

Although many CAD systems have focused on imaging-based analysis, breast cancer diagnosis often relies heavily on structured clinical and biochemical measurements [9]. Tabular medical datasets, especially those used for breast cancer risk assessment are typically small, heterogeneous, and governed by complex nonlinear relationships. Researchers have explored heterogeneous feature selection strategies for complex biomedical data to better capture informative feature interactions in such heterogeneous environments [10,11]. As a result, deep learning methods often underperform strong classical models such as gradient-boosted decision trees (GBDTs) and may suffer from reduced stability and weak calibration when applied to clinical tabular data [12–14]. These limitations highlight the need for models that combine the representational flexibility of deep learning with the robustness and interpretability expected in medical decision-support systems [15].

To address these challenges, researchers have explored hybrid pipelines that integrate deep feature learning with GBDT-based classifiers [16]. However, a key methodological gap persists. Most deep learning models operate as end-to-end black-box classifiers [17], offering little transparency into intermediate representations [18], while conventional stacked ensembles typically provide identical handcrafted feature sets to all base learners. Moreover, researchers employing weighted ensemble strategies have noted difficulties in selecting diverse and complementary base learners [19], which can potentially constrain representational richness and weaken generalization.

To address these challenges, we introduce a new interpretable stacked ensemble framework tailored for breast cancer diagnosis using structured clinical data. First, we redefine the role of TabNet by employing it not as a standalone classifier but as a feature generator capable of generating sparse,

attention-guided representations that enrich downstream learning. Building on these representations, we design a dual-channel stacking architecture in which TabNet-derived embeddings and original tabular features are delivered to heterogeneous base learners (XGBoost and LightGBM), thereby increasing feature diversity and enhancing the ensemble's generalization capacity.

Additionally, to ensure clinical transparency, we further implement a multi-level interpretability pipeline based on Shapley additive explanations (SHAP), enabling clear inspection of TabNet's internal feature selection, the behavior of each base learner, and the meta-learner's final decision logic. Finally, using rigorous nested cross-validation and fixed-budget hyperparameter tuning, we empirically demonstrate that each component of our architecture contributes meaningfully to performance improvements over baseline TabNet and conventional ensemble models.

The major contributions of this study can be summarized as follows:

1) We repurpose TabNet from a standalone classifier into an attention-guided feature generator that produces informative representations for downstream learners in tabular medical prediction tasks.
2) We propose a dual-channel stacked ensemble architecture that integrates TabNet-derived embeddings with original tabular features and leverages heterogeneous base learners (XGBoost and LightGBM) to improve predictive robustness and feature diversity.
3) We provide an integrated interpretability analysis using SHAP-based explanations for the base learners and the meta-learner to offer transparent insights into feature importance and model decision pathways in the proposed ensemble framework.

2. Literature review

2.1. Binary machine learning approaches for breast cancer diagnosis

Machine learning has become a central technology for developing CAD systems in breast cancer, enabling automated prediction and diagnostic support across clinical settings [20]. Within the domain of tabular medical data, methodological families have been widely adopted, each offering different strengths depending on the complexity, scale, and structure of clinical datasets.

Traditional statistical and machine learning models such as logistic regression (LR), k-nearest neighbors, decision trees, random forests (RF), and support vector machines (SVM) have long been used in healthcare. They remain standard baselines in many clinical applications, as demonstrated by studies evaluating their performance across disease prediction tasks [21–25]. Their continued adoption is largely driven by their relative interpretability and stable behavior on small-to-medium-sized tabular datasets [26]. In addition, there is a growing number of studies utilizing traditional machine learning models, such as a study that achieved 95.1% accuracy using an ensemble model combining LR, SVM, and RF [27].

These models remain widely used across structured clinical prediction tasks, including major chronic disease risk prediction, heart disease prognosis, radiation-induced toxicity prediction, pressure ulcer risk assessment, and early infectious disease diagnosis from routine clinical data, underscoring their broad practical value in healthcare applications.

Among these, GBDT models such as XGBoost and LightGBM have emerged as the dominant classifiers for tabular biomedical data because of their strong predictive performance, resistance to overfitting, and ability to capture complex nonlinear feature interactions [13,28].

In particular, machine learning studies have extensively entailed the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is also adopted in this study, to evaluate classification performance. Ghiasi and Zendehboudi [29] demonstrated that decision tree–based ensemble methods, including RF and Extremely Randomized Trees, achieved classification accuracies exceeding 97%, highlighting their robustness in handling tabular breast cancer data. Similarly, Arravalli et al. [30] conducted a comprehensive comparative analysis of multiple machine learning classifiers and reported that tree-based and boosting ensemble models consistently outperformed other approaches in terms of accuracy and area under the receiver operating characteristic curve, with a stacked ensemble model achieving an F1-score of 83%.

Collectively, these findings indicate that ensemble learning, particularly tree-based ensemble methods, has been widely validated as an effective and reliable approach for breast cancer diagnosis on tabular datasets such as WDBC, reinforcing their position as state-of-the-art machine learning models in this domain.

Deep learning-based approaches constitute the second major methodological direction. Artificial neural networks and multilayer perceptrons have demonstrated improved capability in modeling nonlinear patterns that traditional machine learning models may fail to capture. Studies on the WDBC dataset have shown that artificial neural network-based models can outperform strong classical baselines such as SVM and RF in certain diagnostic settings [31]. Moreover, specialized deep architectures for tabular learning such as TabNet have been proposed to incorporate sequential attention and sparse feature selection, enabling deep representation learning tailored to structured clinical data [32].

Despite these advances, challenges remain. Deep learning models frequently require large datasets to avoid overfitting, while many medical datasets, including those used for breast cancer diagnosis, are limited in size, heterogeneous, and characterized by subtle nonlinear interactions. These limitations often result in deep models performing inconsistently compared with GBDT methods, motivating the need for hybrid or ensemble frameworks that combine the representational strengths of deep learning with the robustness of tree-based approaches.

2.2. Opportunities and challenges in stacking integration strategies

Stacked generalization (stacking), originally introduced by Wolpert [33], provides a principled framework for integrating multiple heterogeneous learning algorithms. Rather than selecting a single best model, stacking leverages the complementary strengths of diverse base learners by training a meta-learner to combine their outputs. This strategy has demonstrated measurable performance improvements across medical prediction tasks. For example, a stacking framework using GBDT models as base learners and a random forest meta-learner achieved an accuracy of 93.44% for heart disease diagnosis, outperforming individual constituent models [34]. These results highlight the potential of stacking to enhance diagnostic accuracy in biomedical applications, particularly when datasets contain complex nonlinear relationships or noise.

A critical design element in stacking is the choice of meta-learner. LR is among the most widely adopted meta-learners in machine learning and clinical predictive modeling due to its robustness, low variance, and inherent interpretability [35]. LR mitigates overfitting by imposing a simple linear decision boundary over the predictions of high-capacity base learners, producing well-calibrated outputs that are especially desirable in medical settings where transparency and stability are essential [36]. Although

more expressive meta-learners such as neural networks or GBDT variants can be used, research consistently shows that simple linear meta-learners frequently outperform them by preventing the compounding of errors from overfit base models, particularly in small tabular medical datasets [37].

Despite its advantages, stacking research suffers from an overlooked architectural limitation: Nearly all base learners are trained on the exact same set of raw input features. Whether the models consist of LR, SVM, GBDT, or artificial neural network architectures, they typically receive an identical representation of the data. This homogeneous input constraint restricts the diversity of the learned feature transformations and decision boundaries. Since stacking derives its strength from the diversity of perspectives among base learners, the lack of differentiated feature inputs limits the expressive capacity of the ensemble and reduces the meta-learner's ability to integrate complementary information effectively [38].

In medical predictive modeling, this limitation is particularly problematic. Tabular clinical datasets often contain subtle, high-order interactions that may be captured by deep architectures but are overlooked by tree-based methods, or vice versa [39]. Without differentiated representations, base learners tend to converge toward similar decision patterns, diminishing the theoretical benefits of stacked generalization and leading to only marginal improvements over single-model baselines [33,40].

These challenges reveal the need for more advanced stacking architectures that incorporate distinct and complementary feature spaces for each base learner, enabling genuinely diverse decision pathways. This motivates the central idea of our study: Leveraging TabNet as a feature generator to produce high-level, attention-guided representations that differ fundamentally from the original tabular features, thereby enabling heterogeneous base learners to operate on enriched and distinct inputs.

2.3. Explainability challenges in complex models

As machine learning systems become more deeply integrated into clinical workflows, explainability for black box models has emerged as a critical requirement for ensuring clinician trust, accountability, and regulatory acceptance [41]. A wide range of explainable artificial intelligence (XAI) approaches, such as SHAP, local interpretable model-agnostic explanations, and gradient-based attribution methods, have been widely applied to interpret single-model architectures in medical domains, including liver diabetes classification [42], disease prediction [43], and cancer survival analysis [44]. These techniques provide valuable insight into how individual models utilize clinical features, yet they primarily address monolithic models, offering explanations at a single level of abstraction.

In contrast, stacking-based ensemble frameworks pose more complex interpretability challenges. A stacked ensemble consists of multiple base learners whose outputs are integrated through a meta-learner, forming a multi-layer decision pipeline. Explaining such systems requires clarifying not only how each base model processes the input, but also how the meta-learner synthesizes their predictions, producing a final decision. Researchers have attempted to interpret ensemble models by applying SHAP to tree-based ensembles [45,46]. However, these approaches are generally limited to shallow ensembles or homogeneous models, and they do not provide end-to-end interpretability across all layers of a stacked architecture [47].

This limitation is particularly consequential in medical diagnosis, where the opacity of multi-model systems amplifies the black-box problem that hinders the adoption of deep learning models. Even if individual components of an ensemble are interpretable, the absence of a unified framework

that connects feature selection, base-learner behavior, and meta-learner integration prevents clinicians from understanding the rationale behind predictions [48]. As noted in clinical AI literature, lack of transparency remains one of the most significant barriers to real-world deployment of machine learning systems, even when performance is high [49].

Despite the growing interest in XAI, no work provides a comprehensive interpretability paradigm for stacked ensembles incorporating deep feature representation and tree-based learners. Researchers interpret components in isolation, failing to bridge the full decision chain from feature generation to final classification. This leaves an important methodological gap [50]. Accordingly, a systematic multi-level explanation framework is needed, which can capture the interactions between deep models, boosted trees, and meta-learning layers in a clinically meaningful way.

We address this gap by introducing an end-to-end interpretability framework designed for TabNet-driven stacked ensembles. By applying SHAP at multiple stages, including TabNet's sparse attention mechanism, quantifying base learner contributions, and elucidating the meta-learner's decision logic, we provide a coherent and clinically aligned explanation pipeline that has not been explored in ensemble-based diagnostic modeling.

2.4. TabNet in medical and tabular clinical applications

TabNet has gained significant attention as one of the first deep learning architectures explicitly designed for tabular data, leveraging sequential attention and sparse feature selection to learn structured representations [32]. A clinical application demonstrated that TabNet can effectively model heterogeneous variables and non-linear interactions [51], suggesting strong potential for medical prediction tasks.

Building on this foundation, TabNet has been applied to clinical problems, including disease severity classification, metabolic risk assessment, and chronic disease modeling where its sparse attention mechanism helps capture complex, high-order relationships [52,53]. In the context of chronic kidney disease (CKD) staging among patients with diabetes, TabNet has been utilized for multi-stage classification based on clinical and laboratory indicators that exhibit complex nonlinear and high-order interactions. Researchers reported that this approach achieved strong predictive performance (approximately 94% accuracy) while improving model interpretability through XAI, thereby offering clinically relevant insights into CKD progression [54]. These studies demonstrate TabNet's capacity to model nonlinear relationships in clinical data while maintaining a degree of interpretability through sparse attention mechanisms.

Researchers have also explored advanced deep learning architectures and ensemble strategies for medical diagnosis. For example, adaptive convolutional neural network ensembles have been applied to genetic biomarker discovery in neurological disorders, while dual-enhanced deep learning frameworks have been proposed for early disease classification in medical time-series data [55,56]. These approaches demonstrate the growing interest in combining learning mechanisms to improve predictive performance and robustness in clinical decision-support systems.

In the context of cancer, several applications have examined deep learning approaches, including TabNet as end-to-end classifiers for diagnosing liver cancer, breast cancer, and other malignancies using structured clinical or imaging-derived features [57]. While these works highlight the potential of deep tabular models, TabNet is consistently used as a standalone classifier rather than a component in a broader ensemble architecture. Furthermore, these researchers do not investigate TabNet's potential

role as a feature generator, despite its architectural suitability for producing rich, attention-weighted embeddings that may benefit downstream learners. Moreover, these researchers do not explore how TabNet's internal representations could be integrated into broader ensemble pipelines to enhance feature diversity or improve downstream learning.

Efforts, such as the hybrid TabNet-XGBoost ensemble by Yasmeen et al. [58], have attempted to bridge these gaps using TabNet for feature selection. However, these models suffer from input homogeneity, where all base learners operate on identical raw features limiting the ensemble's ability to leverage truly diverse representations. More importantly, this increased architectural complexity often exacerbates the 'black-box' nature of the system. As Hildt [59] argues, technical performance alone cannot foster clinical trust; without an interpretability framework that aligns with medical reasoning, even high-performing ensembles remain opaque and impractical for real-world diagnostic decision-making.

These gaps motivate the need for an architectural redesign that repositions TabNet not merely as a classifier but as a feature generator that supplies diverse and meaningful representations for downstream learners. By integrating TabNet-generated embeddings into a dual-channel stacking framework and complementing them with an end-to-end interpretability pipeline, our study provides one of the first systematic explorations of using TabNet as a feature generator within a clinically interpretable stacked ensemble for breast cancer diagnosis.

2.5. Positioning of this research

The limitations identified across learning paradigms range from the reliance on handcrafted features in classical machine learning models, to the overfitting tendencies of deep learning methods, to the input homogeneity constraints of conventional stacking frameworks, highlighting the need for architectures that balance predictive performance with clinical transparency. Similarly, although ensemble models offer advantages in medical diagnosis, their lack of differentiated feature spaces restricts model diversity, and the absence of end-to-end interpretability limits their suitability for clinical decision-support systems.

TabNet presents an appealing direction for structured medical data due to its sparse attention mechanism and ability to learn high-level feature representations. However, prior medical applications of TabNet have exclusively used it as a standalone classifier or as a simple component within shallow ensembles, without exploring its potential as a feature generator capable of enriching other learners. Moreover, TabNet-based studies do not address the interpretability challenges of multi-layer ensemble architectures, nor do they consider the methodological implications of integrating TabNet representations into decision-making pathways.

These observations collectively reveal a methodological gap: No research systematically leverages TabNet-derived representations to construct differentiated feature channels within a clinically interpretable stacking ensemble, nor provides unified explanations across feature generation, base learners, and meta-level decision-making.

To address this gap, we propose an integrative framework that repurposes TabNet as a feature generator, enabling the creation of enriched and complementary feature spaces for heterogeneous base learners (XGBoost and LightGBM). By combining these dual feature channels within a logistic regression meta-learner that was chosen for its robustness, calibration, and interpretability, we construct an ensemble that maintains high predictive performance while remaining suitable for clinical

interpretation. Further, by incorporating a multi-stage SHAP-based interpretability pipeline, the framework provides transparent, end-to-end explanations of how features are selected, transformed, and combined across the diagnostic model.

Through this combination of methodological advances, our study contributes to next-generation data science applications by (i) enabling trustworthy analytics via end-to-end interpretability, (ii) improving robustness under OOD-style perturbations that mimic realistic measurement uncertainty, and (iii) providing a modular computational-intelligence pipeline that can be deployed or extended across other structured-data domains beyond breast cancer.

3. Methods

3.1. Dataset and preprocessing

We utilized the WDBC dataset. Features in the data were computed from digitized images of fine needle aspiration (FNA) biopsies of breast masses to describe the characteristics of cell nuclei within the images [31,60]. The dataset contained no missing values and comprised 569 samples, each labeled as benign or malignant: 357 were benign and 212 were malignant. It included 30 feature values, including radius, texture, perimeter, and area. All feature variables were utilized in this study.

The diagnostic labels were converted into a binary variable using LabelEncoder. The numerical scales and ranges of different features varied significantly (e.g., radius_mean, area_mean, and smoothness_mean). Therefore, the data were scaled using RobustScaler to place all features on a comparable scale [61], facilitating their subsequent use in the TabNet model.

The transformation implemented by RobustScaler is given in Eqs (1) and (2):

$$x_i' = \frac{\mathrm{x}_i - \mathrm{median(X)}}{\mathrm{I}QR(X)}, \tag{1}$$

$$IQR = Q_3(x) - Q_1(x). \tag{2}$$

where median(x) is the median of all sample values for this feature (i.e., the second quartile Q_2), $Q_1(x)$ is the first quartile of this feature (i.e., the 25th percentile), $Q_3(x)$ is the third quartile of this feature (i.e., the 75th percentile), and the denominator $Q_3(x) - Q_1(x)$ represents the interquartile range (IQR).

On a comparable scale, it also facilitates subsequent principal component analysis (PCA) to fairly evaluate each feature's contribution to the overall variance.

We employed nested cross-validation, utilizing 5-fold stratified cross-validation to evaluate the performance of the outer layer model. This approach ensures balanced category proportions and guarantees the stability and fairness of the evaluation results. The inner loop was used to generate stacked features and construct the model. Within each outer loop iteration, the training set was fed into a 4-fold stratified cross-validation process to generate out-of-fold (OOF) prediction features for the sub-learners. These sub-learner features were derived from samples not directly used in training, effectively mitigating data leakage risks.

3.2. Model

TabNet is a deep learning model designed for structured data. Its core sequential attention

mechanism selects a semantically meaningful subset of features for processing at each decision step [32].

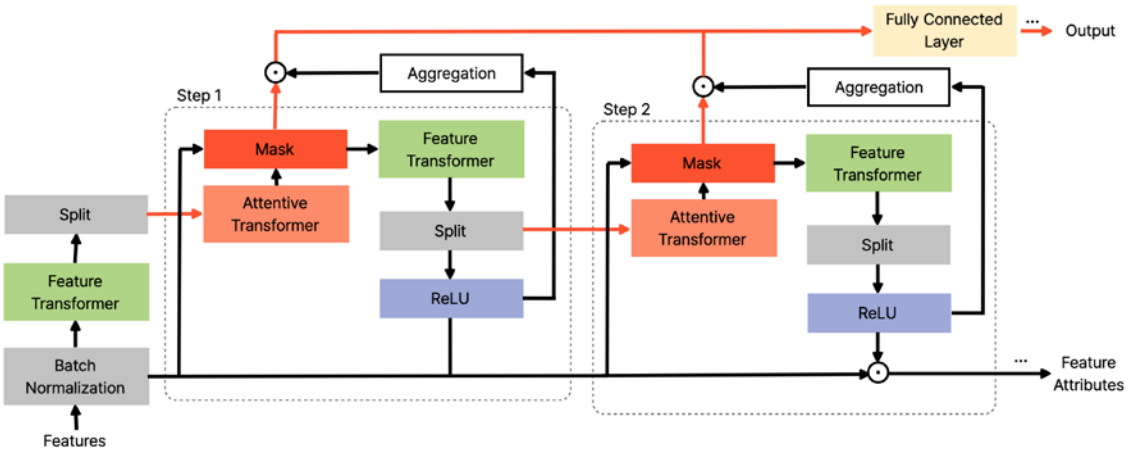

Figure 1. Overview of the TabNet encoder architecture.

As shown in Figure 1, the TabNet encoder comprises a Feature Transformer, Attentive Transformer, and feature masking component. A split block divides the processed representations, allocating one part to the Attentive Transformer in the subsequent step and the other part to the overall model output. At each decision step, the feature selection mask provides interpretable information about the operation of the model, and these masks can be aggregated to determine the global feature importance.

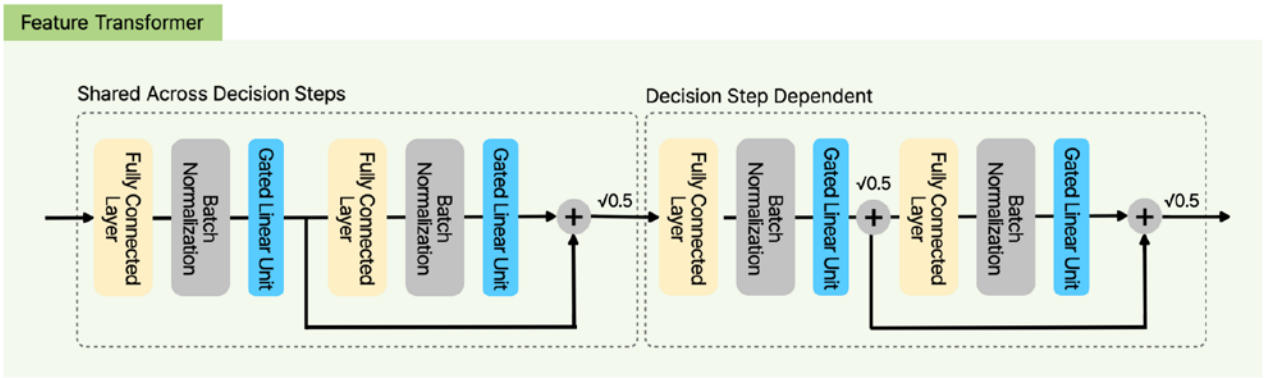

Figure 2. Schematic of the TabNet Feature Transformer block architecture.

Figure 2 illustrates an example of a Feature Transformer block, a 4-layer network in which two layers are shared across all decision steps and the other two layers are specific to the decision steps. Each layer consists of a fully connected layer, batch normalization layer, and gated linear unit nonlinearity.

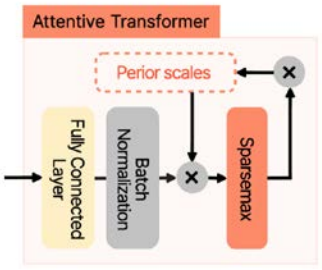

Figure 3. Schematic of the TabNet Attentive Transformer block architecture.

Figure 3 illustrates an example of an Attentive Transformer block. This single-layer mapping is modulated by prior scale information that aggregates the frequency of utilization of each feature prior to the current decision step. Sparsemax normalization of the coefficients yields a sparse selection of salient features. The core of this mechanism is mathematically defined by the masking function, as shown in Eq (3):

$$M[i] = sparemax\big(P[i-1] \cdot h_i(a[i-1])\big), \tag{3}$$

where $M[i]$ is the feature selection mask at step i, $P[i-1]$ is the prior scale term, and the Sparsemax function ensures that only a sparse subset of features is selected.

To ensure that the learned features are discriminative and interpretable, the model is trained by minimizing a total loss function that combines classification performance with a sparsity penalty, as shown in Eq (4):

$$L_{total} = L_{classification} + \lambda_{sparse} L_{spares}, \tag{4}$$

where $L_{classification}$ is the standard classification loss, and L_{spares} encourages sparsity of the feature selection masks across all decision steps.

3.3. Proposed stacking model

As shown in Figure 4, the hybrid model used in this study consists of three layers: Feature generation based on TabNet, a heterogeneous base learner, and a meta-learner employing LR. The proposed architecture is explicitly designed to balance rich deep feature representation with deployment feasibility under resource-constrained clinical environments.

TabNet excels in automatically learning highly abstract, information-rich feature combinations and representations from raw data, a capability validated in complex medical diagnostic tasks [54]. Its multi-step attention mechanism optimizes the decision-making process, progressively focusing on the most critical features while constructing deep, nonlinear feature transformations. This enables the automated generation of high-quality features. As a feature generator, in Figure 5, TabNet provides three types of features: final embedded features (32-dimensional), multi-step features (120-dimensional), and attention weights (30-dimensional).

The final embedded features are deep abstract features that capture complex nonlinear relationships among raw features. Multi-step features are aggregates generated at different decision

steps. Attention weights directly reflect the model's focus on different features during decision-making.

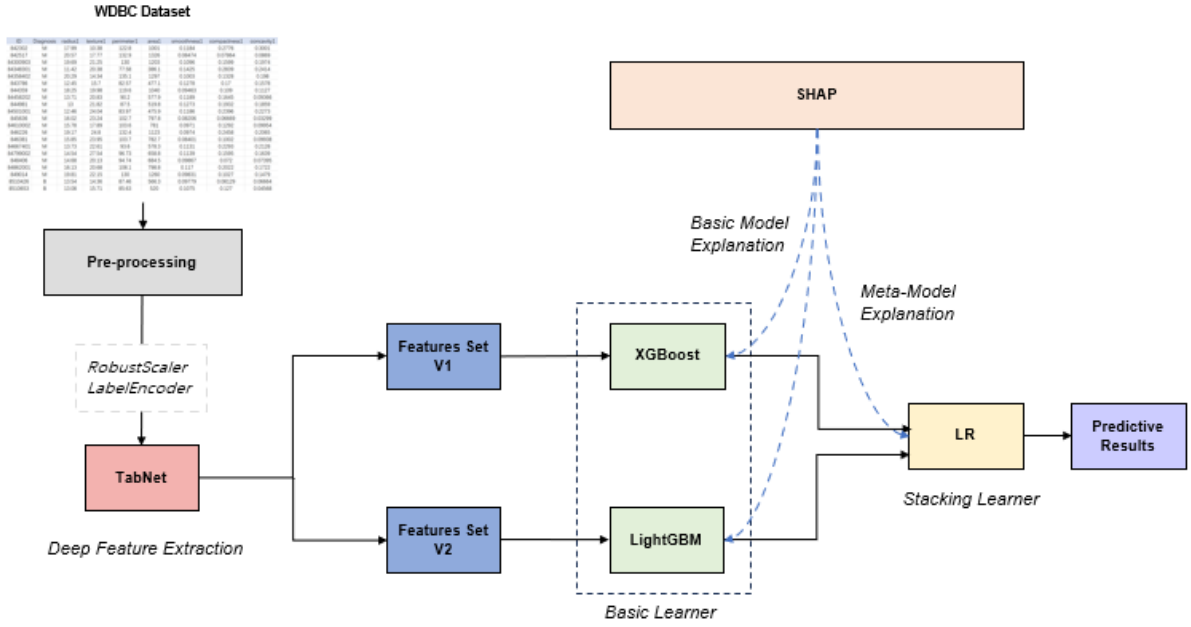

Figure 4. Overall framework of the proposed stacking model.

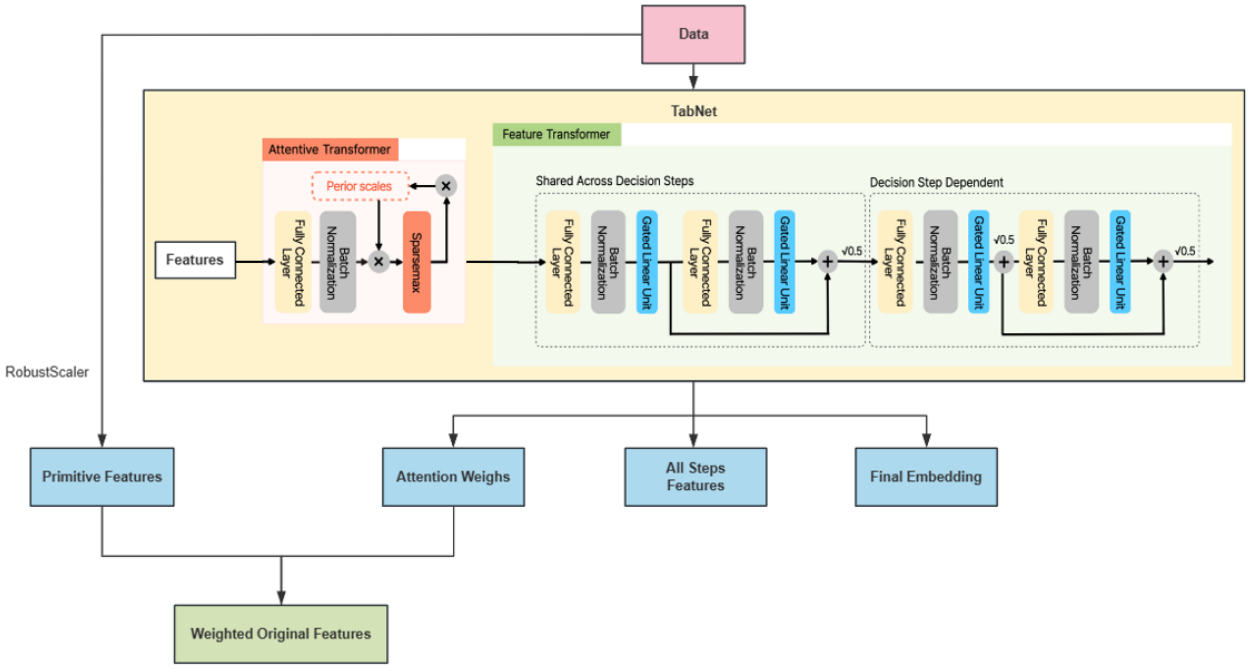

Figure 5. Architecture of the proposed TabNet-based feature generator.

3.3.1. Base learners

XGBoost and LightGBM serve as base learners in the stacking ensemble model. These two

models were selected as heterogeneous base learners owing to their distinct yet complementary strengths within the gradient boosting framework, consistent with findings that tree-based ensembles outperform deep learning on tabular data [28] and perform robustly across disease datasets [62], with the aim of maximizing ensemble diversity.

XGBoost is renowned for its robustness and regularization capabilities, making it effective across datasets, whereas LightGBM offers high efficiency and a unique leaf-wise tree growth strategy, potentially capturing different data patterns and feature interactions than the level-wise approach of XGBoost. XGBoost and LightGBM learn from the data from different perspectives and output prediction probabilities. The core idea is to add new decision trees iteratively to optimize a common objective function [63]. This objective function typically consists of a loss term and regularization term to balance the prediction accuracy and complexity of the model, as shown in Eq (5):

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)\,. \tag{5}$$

In XGBoost, the regularization term $\Omega(f_t)$, which penalizes the complexity of the tree, is defined as Eq (6):

$$\Omega(f) = \gamma T + \frac{1}{2}\,\lambda \sum_{j=1}^{T} w_j^2\,, \tag{6}$$

where T is the number of leaves, w_j is the score of the j-th leaf, and γ and λ are hyperparameters that control the complexity [64].

As a member of the GBDT family, LightGBM adopts an objective function formulation like that defined in Eqs (5) and (6). Its primary contributions, however, lie in substantially improved training efficiency achieved through two key algorithmic innovations.

First, gradient-based one-side sampling (GOSS) accelerates training by prioritizing instances with large gradient magnitudes, which carry more informative signals for model updates, while randomly sampling from instances with smaller gradients. This strategy reduces computational cost without significantly compromising model accuracy.

Second, exclusive feature bundling (EFB) reduces effective feature dimensionality by grouping mutually exclusive features that rarely take nonzero values simultaneously into compact feature bundles. This enables more efficient tree construction, particularly in high-dimensional sparse settings.

Owing to these distinct algorithmic optimizations and their use of different input feature representations, XGBoost and LightGBM act as heterogeneous base learners. This heterogeneity provides the meta-learner with diverse and complementary predictive signals, enhancing the generalization capability of the proposed stacking framework.

3.3.2. Feature Set V1 for XGBoost

XGBoost is characterized by its robustness and strong regularization capabilities. Therefore, a comprehensive global-view Feature Set, V1, was constructed for XGBoost. As shown in Figure 6, it comprises three components: First, preprocessed raw features (Primitive Features); second, final embedding features output from the last layer of TabNet (Final Embedding); and third, multi-step

features derived from PCA dimensionality reduction of intermediate features across all decision steps in TabNet (All Steps Features).

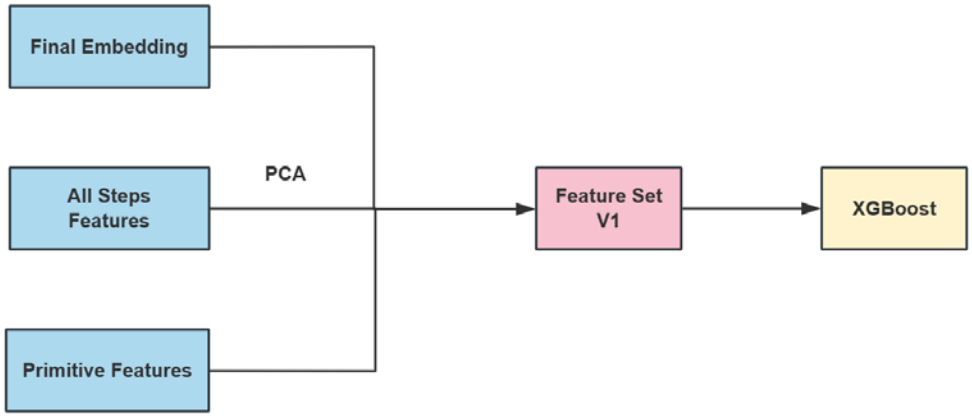

Figure 6. Architecture of the Feature Set V1 for the XGBoost-based stacking model in the proposed framework.

Given the high-dimensional nature of the stepwise deep features extracted by TabNet, dimensionality reduction was applied prior to stacking. Under our configuration (4 decision steps × 30 features), TabNet produces a 120-dimensional feature space that exhibits strong multicollinearity and redundancy. Feeding such raw high-dimensional features directly into downstream learners may increase the risk of overfitting and computational inefficiency. To address this issue, PCA was employed to orthogonalize the feature space and suppress redundant variance.

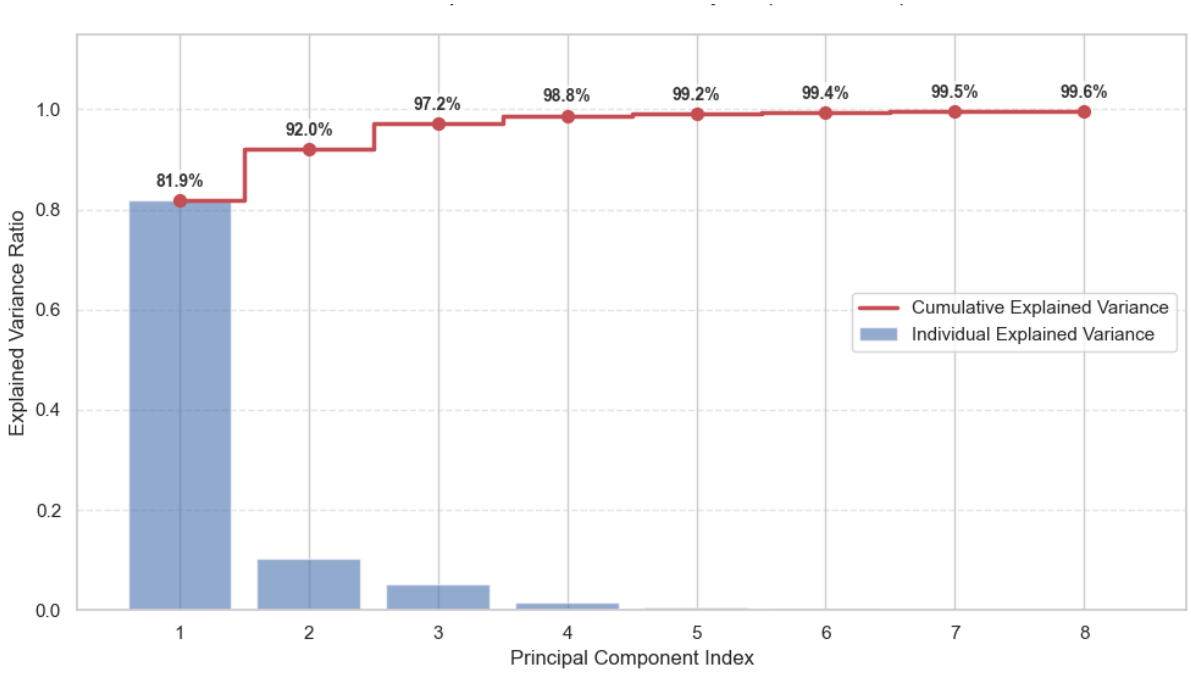

Figure 7. Individual and cumulative explained variance of the principal components.

The number of retained PCA components was determined through cumulative variance analysis. As illustrated in Figure 7, the first eight principal components explain 99.62% of the total variance in the original 120-dimensional feature space. This result demonstrates that dimensionality reduction from 120 to 8 components preserves virtually all discriminative information, with less than 0.4% variance loss, providing a quantitative justification for selecting eight components as compact yet informative inputs for the downstream XGBoost learner.

In parallel, the LightGBM branch utilizes attention-weighted features derived from TabNet's Attentive Transformer rather than heuristic feature scaling. TabNet learns a decision mask $M^{(i)}$ for each feature at every decision step, normalized via Sparsemax and regulated by the sparse entropy loss in Eq (7):

$$L_{sparse} = \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{d=1}^{D} \frac{-M_{b,d}[i] \log (M_{b,d}[i] + \epsilon)}{N_{steps} \cdot B} \tag{7}$$

which statistically enforces sparsity by penalizing dispersed attention. Consequently, the learned attention weights reflect feature relevance. We leverage this property by rescaling input features according to their attention-induced variance, enabling LightGBM to focus on robust clinical biomarkers while reducing noise.

Overall, the combination of PCA-based orthogonalization (retaining 99.62% of total variance) and statistically regularized attention weighting provides a principled balance between information preservation and noise reduction, forming a robust feature foundation for the proposed stacking framework. This combination provides XGBoost with rich, multi-level information spanning raw to abstract representations.

3.3.3. Feature Set V2 for LightGBM

LightGBM employs an efficient leaf-wise generation strategy, enabling rapid model convergence but making it more sensitive to noise in the data [57]. Feature Set V2, constructed for LightGBM, represents an attention-guided focused view. Using TabNet's attention mechanism preprocessing and intelligent filtering, the signals of important features in the dataset were amplified and those of secondary or noisy features were attenuated. Moreover, LightGBM can leverage the feature importance information learned by TabNet more directly. As shown in Figure 8, it consists of two parts: First, Weighted Original Features, which are the original features weighted by TabNet's attention mask; and second, the Final Embedding.

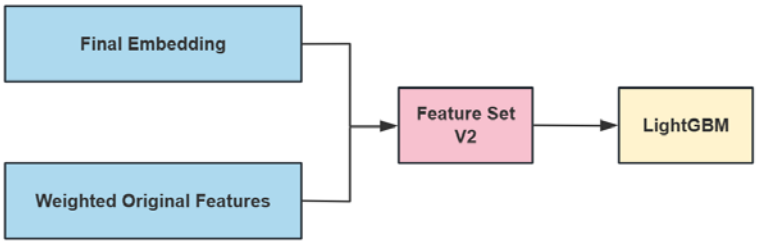

Figure 8. Architecture of the Feature Set V2 for the LightGBM-based stacking model in the proposed framework.

To capture the non-linear feature importance learned by the deep model, we proposed a variance-based feature weighting strategy derived directly from TabNet's latent representations. Unlike traditional feature selection methods that rely on static metrics, our approach utilized the dynamic output of the Feature Transformer, which processes features filtered by the Attentive Transformer.

Let $X \in R^{B\times F}$ denote the original input batch, where B is the batch size and F is the number of features. Let $E \in R^{B\times F}$ represent the output embeddings from the Feature Transformer layer. We defined the global attention weight vector $w \in R^{F}$ based on the activation variance of these embeddings in Eq (8):

$$w = Normalize(Var(E)) \tag{8}$$

where $Var(\cdot)$ computes the variance across the batch dimension, and $Normalize(\cdot)$ ensures the weights sum to 1. The final attention-weighted feature set $X_{weighted}$ is obtained via element-wise multiplication in Eq (9):

$$X_{weighted} = X \cdot w \tag{9}$$

The rationale for using activation variance as a proxy for importance is grounded in TabNet's sparsity regularization (L_{sparse}). During training, the model minimized a loss function that included a sparsity penalty for the feature selection masks. Consequently, features deemed irrelevant by the model were suppressed, resulting in near-zero constant activations (low variance) in the Feature Transformer. Conversely, discriminative features exhibited high activation variance across samples. By modulating the original features with w, we effectively suppressed noise and amplified the clinically significant signals identified by the deep network before feeding them into the downstream GBDT classifiers.

3.3.4. Meta-learner

The core function of the meta-learner is to analyze and intelligently combine the prediction results from the base learners, aligning with the hierarchical learning paradigms reviewed in meta-learning literature [65]. As shown in Figure 9, the input to the meta-learner is a brand-new, two-dimensional feature matrix. Each row of the feature matrix represents a sample, and each column represents the prediction probability from a base learner. The first column contains the XGBoost model's prediction of malignant for the sample, and the second column contains the LightGBM model's probability of predicting malignant for the sample.

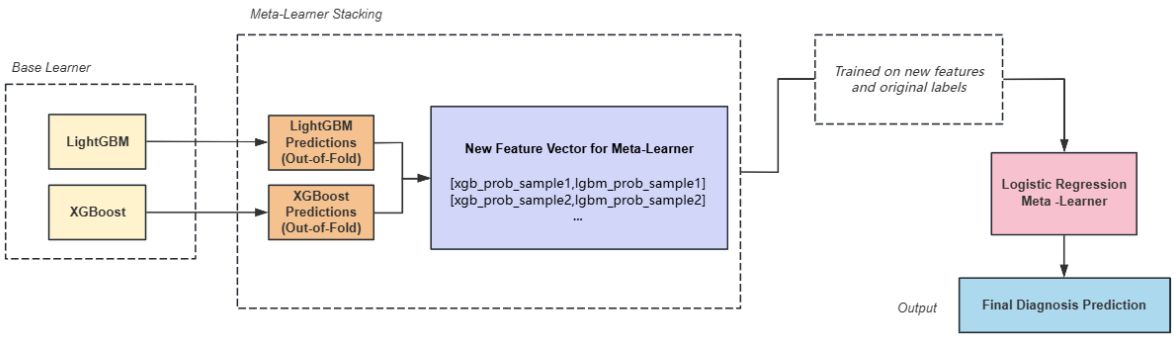

Figure 9. Architecture of the meta-learner in the proposed stacking framework.

The meta-learner employs an LR model to analyze the predictions from XGBoost and LightGBM, optimally combining the outputs of both base models linearly to generate the final probability prediction. Given the complexity of the XGBoost and LightGBM base learners, we selected a simple linear model (logistic regression, LR) as the meta-learner to maintain interpretability and reduce the risk of overfitting. LR maps the output of a linear model to the (0,1) interval using the Sigmoid function to derive a probability value, as shown in Eq (10):

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}, where\ z = W^T X + b. \tag{10}$$

The model learns the optimal weights W and bias b by minimizing the Binary Cross-Entropy Loss function defined in Eq (11):

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})]. \tag{11}$$

The OOF prediction mechanism ensures that the combination strategy learned by the model possesses strong generalization capabilities.

3.3.5. Computational efficiency and deployment optimization

To ensure the proposed framework is practical for real-world clinical deployment, particularly under resource-constrained environments, we implemented specific architectural optimizations. We reduced the embedding dimension ($N_d = N_a = 32$), the number of decision steps ($N_{steps} = 4$), and the hidden layer width of the TabNet component, resulting in a lightweight configuration with approximately 103k parameters. This configuration was designed to uphold the model's deep representational integrity while significantly reducing its computational footprint, thereby enabling low-latency inference on standard CPUs. The baseline TabNet model was retained as a reference deep tabular architecture, whereas the proposed lightweight configuration reflected a deployment-oriented design choice that prioritized computational efficiency.

Furthermore, we introduced a "cached embeddings" inference mode to address scenarios requiring ultra-low latency. In this mode, deep embeddings were precomputed offline once, and real-time prediction relied solely on the optimized basic learners (XGBoost and LightGBM) and the lightweight logistic regression meta-learner. The reported inference latency for this mode excluded the one-time offline cost of embedding precomputation, thereby decoupling feature generation from the inference phase and enabling sub-millisecond prediction speeds without retraining.

To improve clarity and reproducibility of the proposed framework, the complete training and evaluation pipeline is summarized in Algorithm 1.

To prevent data leakage, all preprocessing operations, including RobustScaler and PCA, were fitted exclusively on the training partition within each fold and subsequently applied to the corresponding validation or test partitions. Additionally, TabNet was trained independently within each inner fold to extract out-of-fold latent representations used for stacking.

Algorithm 1. TabNet attention-guided stacking framework

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, number of outer folds $K_{out} = 5$, and inner folds $K_{in} = 4$.

Output: Predicted probabilities $\hat{P}_{test}$ for the outer-test set and optimized classification metrics.

1) Data Preprocessing: Fit RobustScaler on the training partition only and apply the fitted scaler to the validation/test partition.
2) Outer Cross-Validation (Model Evaluation):
   - Split $\mathcal{D}$ into K_{out} folds. For each fold $k \in \{1, \dots, K_{out}\}$:
   - Define Outer-Train X_{tr_full} and Outer-Test X_{test}.
3) Inner Cross-Validation (OOF Feature Generation):
   - Split X_{tr_full} into K_{in} folds. For each inner fold $j \in \{1, \dots, K_{in}\}$:
   - Divide into Inner-Train X_{tr} and Inner-Validation X_{va}.
   - Step A: TabNet on Optimization & Feature Extraction
     - Train TabNet on X_{tr} with early stopping based on X_{va}.
     - Extract Deep Features: f_{emb} (Embeddings), f_{steps} (Multi-step features), and w_{attn} (Attention weights).
   - Step B: Data-Leakage-Free PCA (Crucial for Reviewers)
     - Fit PCA(f_{steps_tr}) only on Inner-Train to get projection matrix P_j.
     - $f_{pca\_va} \leftarrow$ Transform f_{steps_va} using P_j.
   - Step C: Heterogeneous Feature Set Construction
     - $\mathcal{V}_{1\_tr} \leftarrow [X_{tr}, f_{emb\_tr}, f_{pca\_tr}]$
     - $\mathcal{V}_{1\_va} \leftarrow [X_{va}, f_{emb\_va}, f_{pca\_va}]$ (Used for XGBoost).
     - $\mathcal{V}_{2\_tr} \leftarrow [(X_{tr} \otimes w_{attn}), f_{emb\_tr}]$
     - $\mathcal{V}_{2\_va} \leftarrow [(X_{va} \otimes w_{attn}), f_{emb\_va}]$ (Used for LightGBM).
   - Step D: Base Learner Training
     - Train XGBoost on $\mathcal{V}_{1\_tr}$ and LightGBM on $\mathcal{V}_{2\_tr}$.
     - Generate validation predictions: P_{xgb_va} and P_{lgb_va}
     - Aggregate OOF predictions: P_{xgb_oof} and P_{lgb_oof}
4) Meta-learning (stacking):
   - Aggregate OOF predictions to form Meta-Feature Matrix $M = [P_{xgb\_oof}, P_{lgb\_oof}]$.
   - Train Logistic Regression (Meta-classifier) on M and y_{tr_full}.
5) Final Prediction:
   - Generate average predictions from base learners trained on the full outer-training set and apply them to X_{test}(using all K_{in} models).
   - Input average base predictions into the Meta-classifier to obtain predicted probabilities $\hat{P}_{test}$.
   - Optimize the classification threshold T using Youden's J statistic computed only from the OOF p redictions generated on the outer-training set.
   - Apply the optimized threshold T to $\hat{P}_{test}$ without further tuning.
   - Calculate $\hat{y} = 1(\hat{P}_{test} \geq T)$

3.4. SHAP

To address the black box issue in deep learning and complex mixed models and enhance model credibility in clinical applications, we employed the SHAP framework to construct a comprehensive, multi-level interpretability analysis system like Figure 10. Originating from cooperative game theory,

SHAP provides a theoretically grounded, consistent, and reliable model interpretation method by assigning an importance value (Shapley value) to each feature in a specific prediction. Its core concept is to explain the prediction of a complex model, $f(x)$, as a linear function of its simplified binary features, $g(x')$, as shown in the formula for its additive feature attribution model Eq (12):

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i' , \tag{12}$$

where $g(x')$ is the explanation model, M is the number of input features, ϕ_0 is the base value (average prediction over the training data), and ϕ_i is the SHAP value for feature i. Each SHAP value, ϕ_i, is calculated by determining a feature's average marginal contribution across all possible feature coalitions, as defined by the classic Shapley value formula in Eq (13):

$$\phi_{i(f,x)} = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S|_1)!}{|F|\ !} [f_x(S \cup \{i\}) - f_x(S)] . \tag{13}$$

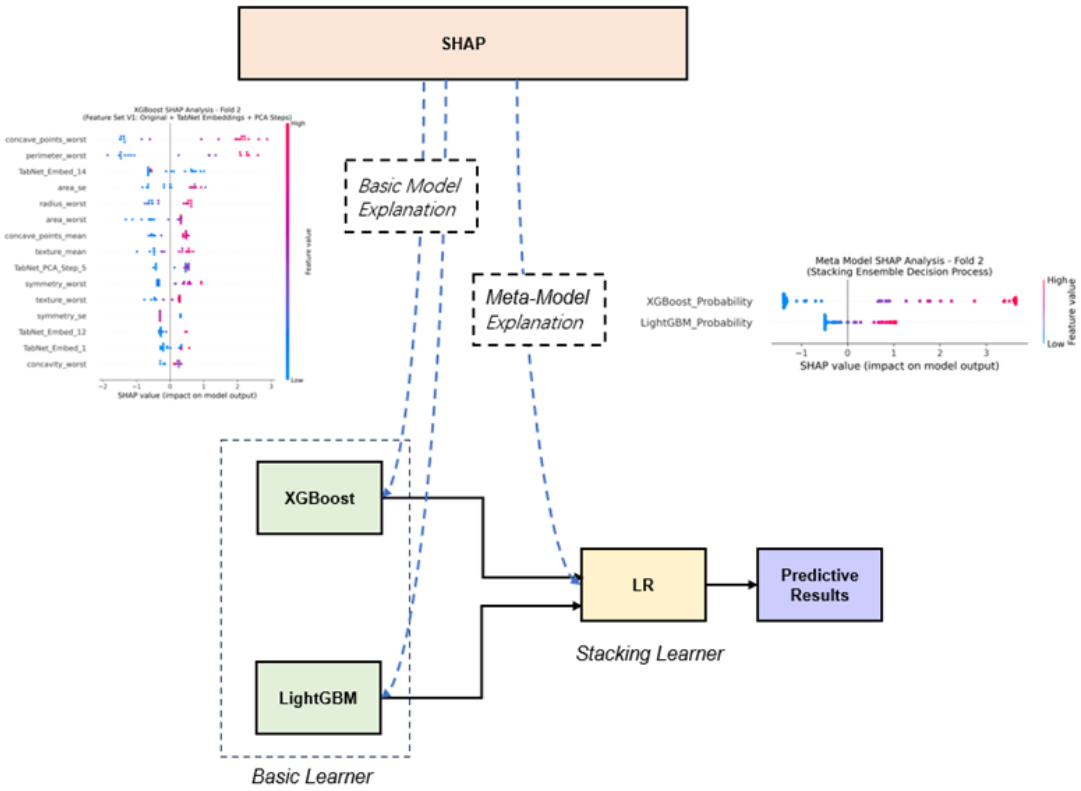

Figure 10. SHAP-based multi-level interpretability of proposed framework.

Based on this foundation, we constructed a comprehensive top-down, end-to-end interpretability framework that systematically dissects the internal decision logic of the model. This framework provides an interpretability analysis for the decision process of the meta-learner and feature contributions within the XGBoost and LightGBM base learners.

3.4.1. Meta-learner decision process analysis

To investigate the integration mechanism of the stacked ensemble model, we analyzed the logistic regression meta-learner using SHAP's Linear Explainer. Since logistic regression acts as a generalized linear model where the output is a weighted sum of inputs, it inherently aligns with the additive feature attribution principle. Unlike model-agnostic approaches that rely on sampling approximations, Linear Explainer leverages this linearity to compute exact Shapley values directly from the model's coefficients. This calculation follows the Linear SHAP formulation derived in Corollary 1 of Lundberg and Lee [66], as defined in Eq (14), for a linear model $f(x)$ approximated by a weighted sum of M input features.

$$f(x) = \sum_{j=1}^{M} w_j x_j + b\,, \tag{14}$$

The exact Shapley value ϕ_j for the j feature can be computed directly from the model's weight coefficients:

$$\phi_j(f, x) = w_j\big(x_j - E\big[x_j\big]\big) \tag{15}$$

By utilizing Eq (15), we mathematically quantified the decision weight assigned to each base model. This enables us to explicitly trace how the meta-learner balances the probabilistic outputs from the heterogeneous base learners to derive the final diagnostic outcome.

3.4.2. Base learner feature importance analysis

To gain a deeper insight into the underlying rationale behind the model decisions, we employed SHAP's Tree Explainer to separately analyze the two base learners.

Given that XGBoost and LightGBM are GBDT models, Tree Explainer is the optimal analytical tool. Unlike model-agnostic methods, Tree Explainer utilizes the internal tree structure to perform exact computations. As demonstrated by Lundberg et al. [67], this algorithm reduces the computational complexity from exponential time $O(TL2^M)$ to polynomial time $O(TLD^2)$, where T is the number of trees, L is the max number of leaves, D is the max depth, and M is the number of features. This efficiency enables consistent and granular interpretation of high-dimensional feature sets [67].

We assigned distinct feature sets to independent learning paths: The XGBoost model was trained on Feature Set V1, which integrates primitive features with TabNet-derived final embeddings and PCA-reduced step features, while the LightGBM model utilized Feature Set V2, combining deep embeddings with attention-weighted original features. This step revealed which raw features (such as concave_points_worst) and depth features extracted from TabNet play the most critical roles in the model's judgments across two independent learning paths.

By integrating analyses at both levels, we constructed an end-to-end, full-chain interpretability framework. This framework provides insights at the macro and micro levels: At the macro level, it reveals the fusion strategy of the meta-learners; and at the micro level, it quantifies the impact of individual features on the base learners. This framework not only clarifies the origin of final predictions but also establishes clear traceability between decision logic and original input variables, thereby enhancing model transparency and its potential for clinical applications.

3.5. Experimental setup

3.5.1. Software environment and libraries

All experiments in this study were conducted in a Python 3.8.6 (64-bit) environment and relied on a suite of open-source scientific computing libraries. The hardware consisted of an Intel(R) Core (TM) i5-8250U CPU and Intel(R) UHD Graphics 620 GPU, running on a 64-bit operating system.

To ensure reproducibility, all experiments were conducted with a fixed random seed (42). The core implementation was based on key libraries, including Pandas (2.0.3), NumPy (1.24.3), Scikit-learn (1.3.2), PyTorch (2.4.1+cpu), PyTorch-TabNet (4.1.0), XGBoost (2.1.4), and LightGBM (4.6.0). The complete source code for reproducing the results presented in this paper is publicly available on GitHub at https://github.com/harimm-ll/An-Interpretable-Stacking-Ensemble-with-TabNet-Enhanced-Features-for-Breast-Cancer-Diagnosis.

- Data processing and feature engineering: We utilized the Pandas library for dataset reading and management; NumPy library for efficient numerical computations and array operations; and Scikit-learn library for data preprocessing, including feature scaling via RobustScaler, label encoding using LabelEncoder, and dimensionality reduction using PCA.
- Model construction: The core model architecture implementation was as follows:
  - The feature generator utilized the PyTorch-Tabnet library, built upon the PyTorch framework.
  - Two heterogeneous base learners were implemented via the XGBoost and LightGBM libraries.
  - The meta-learner employed a logistic regression model from scikit-learn.
- Hyperparameter optimization: Hyperparameter tuning was performed using the Optuna framework for tree-based models and random search for deep learning models within a nested cross-validation setting.
- Model interpretation: To achieve end-to-end interpretability analysis, we employed the SHAP library to compute and visualize feature importance across all model components.

3.5.2. Hyperparameter optimization

To systematically optimize model performance, hyperparameter tuning was conducted within a nested cross-validation framework. The external evaluation protocol employed 5-fold stratified cross-validation to ensure robust and reliable performance estimates.

Each optimization task was limited to a search budget of 40 trials to balance computational efficiency and exploration of the parameter space. To ensure a fair comparison across models, all algorithms were optimized under this identical search budget. Tree-based models (LR, XGBoost, and LightGBM) were optimized using the Optuna optimization framework, while deep learning models (FT-Transformer, Tab-Transformer, and TabNet) were tuned using random search due to their higher-dimensional parameter spaces.

Within each fold of the nested cross-validation framework, hyperparameter optimization was performed independently on the training portion of the data, and model selection was based on validation performance. This design prevents information leakage and ensures that the final performance estimates reflect consistent trends across folds rather than dataset-specific configurations.

Early stopping with a patience of 10 epochs was applied during training to reduce the risk of overfitting on the relatively small WDBC dataset. At the epoch level, validation loss was monitored to

stabilize training convergence, while model selection was guided by the F1-score computed from OOF predictions. The detailed hyperparameter search spaces and the optimal configurations obtained during the optimization process are provided in Appendix A.

4. Results

We expanded our comparative analysis to include five strong baselines, covering gradient boosting trees and advanced deep learning architectures. Specifically, we evaluated meticulously tuned XGBoost and LightGBM models, which are widely regarded as the gold standard for tabular data, as well as two transformer-based tabular deep learning baselines, FT-Transformer and Tab-Transformer. In addition, TabNet was adopted as our primary deep learning baseline.

Although gradient boosting models such as XGBoost and LightGBM are conventional techniques, they remain among the most effective approaches for structured tabular data, particularly in medical datasets with limited sample sizes. In the proposed framework, these models are not used as standalone predictors but rather as complementary learners that exploit the latent feature representations generated by TabNet.

The results summarized in Table 1 demonstrate that the proposed model consistently outperforms all baseline methods across accuracy, sensitivity, and specificity. In addition to the mean performance values, 95% confidence intervals are reported to provide an indication of performance variability across cross-validation folds.

Table 1. Comparison of accuracy, sensitivity, and specificity across machine learning models.

Model	Acc (%)			Sens (%)			Spec (%)		
	Mean	CI	SD	Mean	CI	SD	Mean	CI	SD
LR	93.1*	3.7	3.7	87.5*	6.3	6.3	96.6 ns	3.4	3.4
XGBoost	94.9*	2.0	1.4	91.5*	6.3	5.1	96.8 ns	2.7	2.8
LightGBM	93.8*	3.2	2.4	88.9*	9.4	2.1	96.9 ns	3.1	3.3
FT-Transformer	95.2*	2.5	1.5	91.5*	7.2	7.1	97.2 ns	1.5	2.4
Tab-Transformer	92.9*	2.3	2.7	88.2*	3.3	3.7	95.8 ns	2.3	2.6
TabNet (baseline)	95.6*	1.7	1.4	96.6ns	8.4	4.1	94.9*	3.8	2.7
Proposed Model	97.8	2.0	1.0	98.1	1.3	1.9	97.2	2.2	1.2

Note: Statistical significance compared with the proposed model: *$p < 0.05$, **$p < 0.01$, ns = not significant, evaluated using a two-tailed corrected resampled t-test.

As shown in Table 1, tree-based ensemble models such as XGBoost and LightGBM achieve strong overall performance, confirming their effectiveness for tabular medical data. Among deep learning approaches, FT-Transformer demonstrates competitive results, indicating that Transformer-based architectures can be viable alternatives for tabular learning. Notably, the TabNet baseline outperforms most conventional baselines in terms of sensitivity, highlighting its ability to capture clinically relevant patterns associated with positive cases.

In addition to achieving the highest mean performance across all evaluation metrics, the proposed model demonstrates stable predictions across cross-validation folds. The proposed model attains an accuracy of 97.8%, sensitivity of 98.1%, and specificity of 97.2%. Compared with the strongest baseline (TabNet, 95.6% accuracy), the proposed model improves accuracy by approximately 2.2%

while maintaining balanced gains in sensitivity and specificity. These results suggest that the proposed framework effectively leverages complementary feature representations and modeling strategies, resulting in robust and well-balanced performance across positive and negative classes.

To provide additional statistical evidence for the observed performance differences, we conducted pairwise comparisons between the proposed model and each baseline using the two-tailed corrected resampled t-test proposed by Nadeau and Bengio. This test adjusts the variance estimate to account for dependencies between cross-validation folds caused by overlapping training samples.

Among these baselines, TabNet serves as the primary deep learning reference model and is therefore discussed in greater detail. When compared with the TabNet baseline, the proposed model demonstrates consistent improvements across evaluation metrics under the corrected resampled t-test framework. In particular, the proposed model achieves higher overall accuracy and sensitivity, which is particularly important in clinical cancer screening as it reflects a reduced false-negative rate. Moreover, the improved specificity indicates that the enhanced sensitivity does not come at the expense of incorrectly classifying benign cases. These results further support the robustness and clinical relevance of the proposed ensemble framework.

It is worth noting that the number of outer folds in the nested cross-validation scheme is limited (five folds), which inherently constrains the statistical power of hypothesis testing. Therefore, the reported p-values and confidence intervals should be interpreted cautiously and primarily serve as supporting statistical evidence.

While the proposed model achieves strong performance across all evaluation metrics, it is important to emphasize that the primary contribution of this study is not to compete for marginal state-of-the-art gains. Instead, we focused on the repurposing of TabNet as an interpretable feature generator and its integration within an ensemble framework designed to balance predictive performance and model transparency.

By leveraging TabNet-derived representations in a complementary manner rather than relying on a single end-to-end deep learning model, the proposed architecture prioritizes interpretability, modularity, and clinical usability. This design choice is particularly relevant in medical decision-making contexts, where transparent reasoning and stable performance are often more critical than incremental improvements in benchmark metrics.

4.1. Ablation study

Building upon these comparative results, we further investigated the factors underlying the observed performance improvements by systematically analyzing the contribution of each core component of the proposed model. To this end, we conducted a comprehensive ablation study, the results of which are presented in Table 2. By sequentially removing or combining key components, we were able to isolate and quantify the contribution of each component to the overall performance.

The ablation study provides further insight into how each component contributes to the final performance of the proposed model. As shown in Table 2, the TabNet model trained directly on the original tabular dataset provides a strong baseline performance, particularly in sensitivity, indicating its effectiveness in capturing discriminative patterns from tabular clinical features. Stage 1 evaluates several baseline models trained directly on the original dataset, including TabNet, XGBoost, LightGBM, and simple ensemble combinations.

Table 2. Ablation study on model composition and feature-source design of the proposed dual-channel stacking architecture.

Ablation stage	Feature source	Model	Nested cross-validation		
			Acc (%)	Sens (%)	Spec (%)
Stage 1	Raw-only (original dataset)	TabNet	95.6	96.6	94.9
		XGBoost	94.9	91.5	96.8
		LightGBM	93.8	88.9	96.9
		XGBoost+LightGBM	94.2	93.4	94.7
		XGBoost+LightGBM+LR	94.5	91.4	96.3
Stage 2	TabNet-only (TabNet extracted)	TabNet+XGBoost	94.4	93.9	94.7
		TabNet+LightGBM	93.8	93.4	94.1
		TabNet+XGBoost+LightGBM	96.1	96.2	96.1
		TabNet+XGBoost+LightGBM+LR	95.7	93.4	97.2
Stage 3	Dual-Set (TabNet extracted + XGB/LGBM)	Proposed Model	97.8	98.1	97.2

Comparing the raw-feature baselines in Stage 1 shows that individual tree-based models such as XGBoost and LightGBM achieve comparable performance, while simple stacking combinations provide only limited improvement when trained on the original feature space. This staged comparison enables a clear assessment of how performance evolves from raw-feature baselines (Stage 1), to homogeneous stacking based on TabNet-derived features (Stage 2), and finally to the dual-channel stacking architecture proposed in this study (Stage 3).

In contrast, combining multiple base learners (XGBoost and LightGBM) trained on identical TabNet-derived features leads to moderate improvements across accuracy and specificity, demonstrating the benefit of ensemble diversity even under homogeneous feature inputs. When logistic regression is further added to the stacking ensemble, the overall accuracy slightly decreases compared with the XGBoost + LightGBM configuration. This result suggests that the inclusion of a linear meta-learner does not always provide additional benefits when the base learners already capture most of the nonlinear decision structure.

More importantly, the proposed model differs from these intermediate configurations by introducing heterogeneous feature inputs. In this setting, XGBoost and LightGBM receive different feature representations (original tabular features and TabNet-derived embeddings), enabling the ensemble to exploit complementary information from raw clinical features and deep representations.

As shown in Table 2, the dual-channel stacking architecture achieves the best performance with an accuracy of 97.8%, improving upon the best homogeneous stacking configuration (TabNet + XGBoost + LightGBM, 96.1%) by approximately 1.7%.

4.1.1. Performance of TabNet and single base learners

When used as a standalone classifier in Stage 1, the TabNet model achieves an accuracy of approximately 95.6%, confirming its effectiveness in capturing complex patterns in tabular clinical datasets.

However, when TabNet-derived features are directly combined with a single downstream learner,

the overall performance slightly decreases. The accuracy of the TabNet + XGBoost configuration is approximately 94.4%, while that of the TabNet + LightGBM configuration is 93.8%. One possible explanation for this is that the feature representations generated by TabNet contain multiple levels of abstraction, which may introduce redundancy or increased feature complexity when used directly by a single downstream learner. In such cases, individual tree-based models may not fully exploit the complementary information contained in these representations, potentially affecting model efficiency and generalization.

Under these configurations, the TabNet + LightGBM combination performs slightly worse than TabNet + XGBoost. This difference may be related to the distinct tree-growth strategies employed by the two algorithms. XGBoost adopts a level-wise growth strategy, which may provide more stable splits when handling complex engineered features. In contrast, LightGBM utilizes a leaf-wise growth strategy, which, although computationally efficient, may be more sensitive to feature redundancy or noise in certain settings.

4.1.2. Impact of heterogeneous ensemble and meta-learning

In stage 2, when TabNet + XGBoost + LightGBM are deployed concurrently and their predictions are integrated via simple averaging, the model accuracy improves significantly to 96.1%, surpassing that of the original TabNet baseline. This result suggests that combining heterogeneous base learners can improve predictive performance by leveraging complementary modeling strategies.

Although simple averaging of base-learner outputs provides a reasonable baseline, the ablation results indicate that introducing a meta-learner may provide additional improvements in overall accuracy and clinically relevant metrics. Unlike static averaging, which assigns equal weights to all base learners, the meta-learner learns a data-driven fusion rule by using base-learner probabilities as input features.

In our experiments, the full dual-channel stacking architecture achieves 97.8%, outperforming the best homogeneous TabNet-only stacking configuration (96.1%). This heterogeneous input design enables the ensemble to utilize complementary information from raw clinical features and deep representations, which likely contributes to the performance improvements observed in Table 2.

4.2. SHAP analysis results

To better understand the decision-making mechanisms of the base learners and to examine whether the learned patterns align with clinically plausible factors, we conducted SHAP analyses for the XGBoost and LightGBM components in the dual-channel stacking architecture. The results highlight the key drivers of model predictions and provide interpretable evidence that the model's reasoning is broadly consistent with commonly reported pathology-related criteria in the literature.

4.2.1. Analysis of a decision-making basis in XGBoost (Feature Set V1)

Figure 11 demonstrates that XGBoost's decision logic is deeply rooted in breast cancer pathology, primarily driven by features quantifying nuclear pleomorphism, a key criterion in the Elston and Ellis histological grading system [68].

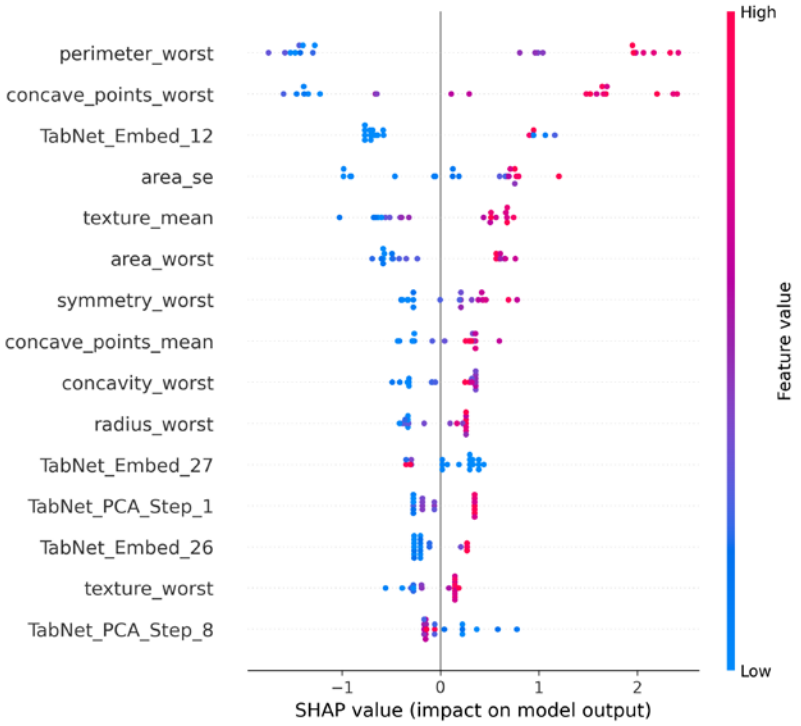

Figure 11. SHAP feature importance for the XGBoost component (representative fold 2).

Specifically, "perimeter_worst" and "concave_points_worst" emerge as the top two most influential predictors. This primacy of nuclear contour irregularity and size aligns with Wolberg et al. [69], who identified nuclear morphology as a strong independent prognostic indicator of tumor grade and invasive potential. Furthermore, the high rankings of "area_se" and "area_worst" confirm that the model correctly associates enlarged, irregular nuclear dimensions with malignancy, which is consistent with the prognostic findings of Aaltomaa et al. [70].

The abstract deep feature "TabNet_Embed_12", part of the Final Embedding vector derived from the Feature Transformer layer in TabNet, ranked third, notably surpassing classical morphological features such as "area_se". As shown in Figure 11, high values of this embedding (indicated by red dots) consistently yield substantial positive SHAP values. This indicates that XGBoost successfully fuses interpretable clinical indicators with complex abstract patterns extracted by TabNet. This 'deep intelligence' captures latent multi-feature interactions that complement traditional morphological analysis, thereby significantly enhancing the model's predictive precision without sacrificing the interpretability provided by the top-ranking clinical features.

4.2.2. LightGBM (Feature Set V2) with attention mechanism validation

As shown in Figure 12, the SHAP analysis of the LightGBM model not only reaffirms the clinical findings but, more crucially, provides compelling empirical evidence for the effectiveness of the TabNet attention mechanism in simulating the diagnostic logic of pathologists. Figure 12 visually demonstrates the dominance of attention-weighted features. In LightGBM's importance ranking, most top-ranked features are original clinical indicators adjusted by TabNet attention weights. The high consistency between the feature importance learned by this model and established clinical pathology indicators is summarized in Table 3.

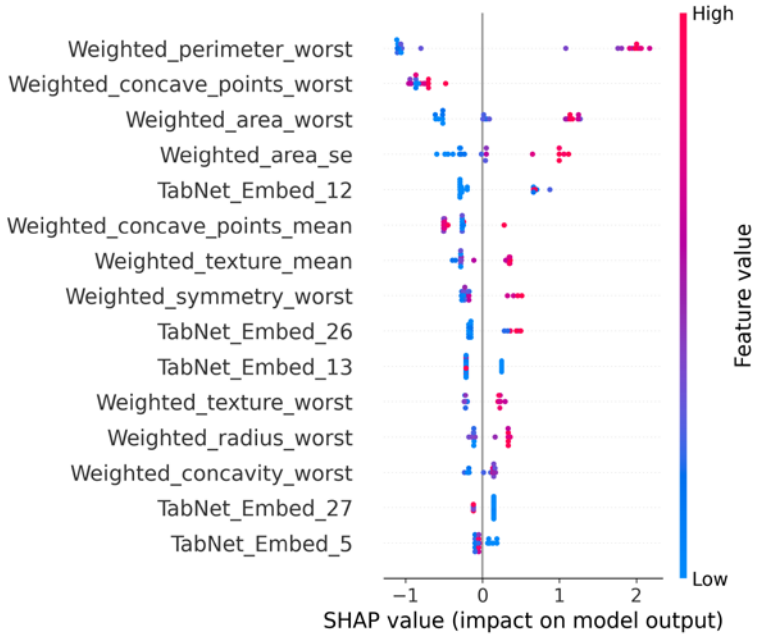

Figure 12. SHAP feature importance for the LightGBM component (representative fold 2).

Table 3. Alignment with clinicopathological indicators.

Features	Pathological indicators
Weighted_concave_points_worst	Severe nuclear contour irregularity
Weighted_perimeter_worst / Weighted_area_worst / Weighted_radius_worst	Nuclear size/Morphometry
Weighted_texture_worst	Nuclear chromatin texture
Weighted_symmetry_worst	Nuclear symmetry/Shape regularity
Weighted_concavity_worst	Degree of nuclear concavity

Based on this correspondence, we can conduct a more in-depth analysis. The most influential feature, "Weighted_concave_points_worst," directly quantifies severe irregularities in nuclear contours. This holds biological significance, as distortions and concavities in the nuclear envelope contour represent not only key morphological markers of cellular malignancy but also potentially reflect the loss of nuclear laminin stability and cytoskeletal reorganization, which are biological foundations for tumor cells acquiring invasive and metastatic capabilities [69,71]. The subsequent Weighted_perimeter_worst and Weighted_area_worst jointly reflect nuclear size, confirming the model's sensitivity to nuclear pleomorphism, a core criterion in the internationally recognized histological grading system established by Elston and Ellis [68], typically associated with cellular aneuploidy and uncontrolled proliferative activity.

Furthermore, the model responds to Weighted_texture_worst (reflecting nuclear chromatin patterns, often linked to fundamental dysregulation of gene expression [72]), Weighted_symmetry_worst (a key morphological indicator used to differentiate malignant from benign cytology [73]), and Weighted_concavity_worst (quantifying irregularity from another

dimension). Collectively, these highly valued features establish a decision foundation that is highly consistent with pathologists' diagnostic logic. The collective high ranking of this series of weighted features strongly indicates that TabNet's attention mechanism successfully guides the downstream LightGBM model, enabling it to automatically focus on the most diagnostically valuable multidimensional pathological information constituting the assessment of nuclear atypia.

Equally crucial is that even after the attention mechanism filters and weighs the original features, the abstract deep features generated by TabNet, "TabNet_Embed_14" still exerts significant influence, as demonstrated in Figure 12. Using the transparency provided by SHAP analysis, we observe a deeper synergistic effect, proving that LightGBM's decision-making process embodies a dual intelligence. SHAP enables us to observe that the model not only relies on human-interpretable key clinical indicators amplified via the attention mechanism but also integrates complex abstract patterns unearthed by deep learning, which these indicators alone cannot fully capture.

The granular insights from the SHAP analysis fully elucidate the intricate decision-making steps of the LightGBM base learner, showing precisely the mechanism of balancing weighted clinical features against abstract deep features. The resulting transparency enhances its credibility as a reliable component within the Stacking framework.

Collectively, these features represent nuclear morphological irregularities (concavity, size, symmetry, and chromatin texture), all of which are consistent with the cytopathological indicators of malignancy. This alignment underscores the clinical validity of the model's interpretive logic.

4.2.3. Meta-learner decision process analysis

We employed SHAP's Linear Explainer to analyze the meta-learned LR model and investigate the generation of diagnostic predictions by our Stacking ensemble model.

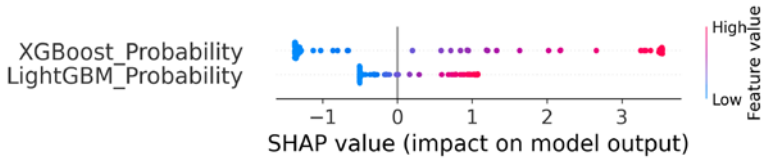

Figure 13. Global interpretability of the stacking meta-learner (representative fold 2).

As shown in Figure 13, the SHAP analysis of the meta-learner reveals a clear, robust, and intelligent decision fusion process. The chart demonstrates that the prediction probabilities from both base learners (XGBoost_Probability and LightGBM_Probability) are positively correlated with the final prediction outcome. Specifically, when a base learner predicts a higher probability of malignancy (red dots in the figure), its contribution to the final model output (SHAP value) also increases, strongly driving the result toward malignancy. The reverse also holds true. This demonstrates that the meta-learner learned fusion rules consistent with medical logic.

The meta-learner also relies on inputs from both base learners simultaneously without disregarding either's opinion. This validates the effectiveness of our dual-channel stacking architecture and confirms that the two models work synergistically. Moreover, a closer inspection reveals an intriguing detail: Although both base learners contribute equally when predicting benign (blue points, negative SHAP values), the meta-learner assigns higher weight to XGBoost_Probability when

predicting malignant. XGBoost_Probability's high probability values could generate SHAP values exceeding +3.0, while LightGBM_Probability's positive contributions remain relatively modest. This indicates the meta-learner learns a nuanced arbitration rule: It treats XGBoost as a more decisive expert in identifying highly suspicious malignant cases while balancing input from both learners in other scenarios.

4.3. Out-of-distribution robustness under covariate perturbations

To assess OOD-style robustness under covariate perturbations, we performed a controlled stress test by injecting additive Gaussian noise with varying standard deviations (σ) into the test inputs. This setting emulated measurement noise and minor distributional shifts that could occur in real-world clinical data pipelines. We examined (i) the stability of TabNet-derived representations, (ii) the robustness benefit of heterogeneous tree-based learners, and (iii) the behavior of adaptive thresholding via Youden's J statistic under perturbed probability distributions. The results are shown in Figure 14.

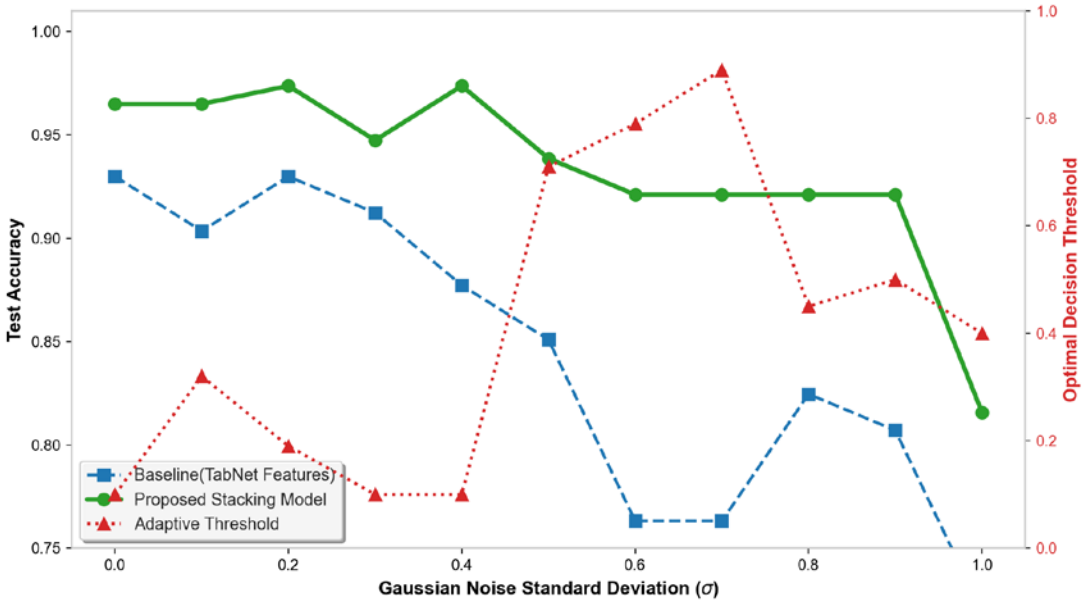

Figure 14. Model performance under different Gaussian noise levels (σ).

As the noise level increases, the model relying solely on TabNet-generated features exhibits a noticeable performance degradation. This behavior reflects the sensitivity of attention-based deep representations to input perturbations, where noise can distort learned attention patterns and propagate through the feature embeddings.

In contrast, the proposed stacking model maintains substantially more stable performance across noise intensities. This robustness arises from the complementary integration of deep representations and tree-based learners. While TabNet provides semantic feature embeddings, XGBoost and LightGBM inherently suppress noise through discrete splitting and ensemble averaging. The logistic regression meta-learner further enhances stability by adaptively reweighting base learner contributions as feature reliability changes.

Figure 14 also illustrates the evolution of the optimal decision threshold selected by maximizing Youden's J statistic. Rather than relying on a fixed threshold, the adaptive threshold adjusts to shifts in

the predicted probability distribution induced by noise, helping preserve a balanced trade-off between sensitivity and specificity.

Overall, these results demonstrate that although TabNet-derived features alone are sensitive to noise, their integration within the proposed stacking framework effectively mitigates this limitation, yielding a robust prediction pipeline suitable for noisy clinical data.

5. Discussion

5.1. Clinical significance and trustworthiness

A key strength of the proposed model is its end-to-end interpretability, which is essential for clinical decision-support systems. By integrating SHAP-based explanations at the base-learner and meta-learner levels, the decision-making process can be transparently decomposed from raw clinical features to the final diagnostic outcome.

For instance, a high-risk prediction can be explained by the meta-learner assigning greater importance to the XGBoost branch, which highlights features such as concave_points_worst. This finding is consistent with established pathological knowledge that irregular and distorted nuclear morphology is a hallmark of malignancy. Such alignment between model reasoning and known clinical biomarkers enhances trustworthiness [49,59,74] and supports the model's role as an assistive diagnostic tool rather than a black-box predictor.

From a practical perspective, the proposed framework is well suited for integration as a decision-support component within existing diagnostic workflows. By providing transparent risk estimates and feature-level explanations, the model can assist clinicians in prioritizing ambiguous cases and support improved diagnostic consistency.

5.2. Methodological novelty compared to existing research

The proposed approach differs from prior ensemble and TabNet-based studies in two fundamental aspects. First, with respect to stacking strategies, most frameworks, including the stacked integration proposed by Daza et al. [34], train all base learners on identical raw input features. Although learner heterogeneity is introduced at the algorithmic level, the homogeneity of feature inputs limits representational diversity and constrains the potential benefits of stacked generalization.

In contrast, our dual-channel architecture explicitly introduces heterogeneous feature spaces: XGBoost receives a comprehensive feature set combining original features, TabNet-derived embeddings, and PCA-compressed stepwise features, whereas LightGBM operates on attention-weighted features filtered by TabNet's Attentive Transformer. This differentiation at the input level provides the meta-learner with richer and more complementary predictive signals.

Second, regarding the role of deep learning models, FH-TabNet-based studies [75] rely exclusively on TabNet at each stage of the multi-stage pipeline. Although effective, these designs use TabNet as the sole predictive module rather than leveraging its learned representations in combination with other model families. Similar paradigms can also be observed in earlier deep belief network-based breast cancer classification approaches [76], where deep models are used exclusively for end-to-end prediction.

In contrast, we adopt a different perspective by repurposing TabNet as a feature generator rather than a classifier. This paradigm is conceptually aligned with trends in medical image analysis, where

deep networks are frequently used to generate informative representations that are subsequently leveraged by classical models or ensembles [77]. By adapting this idea to the tabular clinical domain and combining it with attention-guided feature selection, our framework enables deep representation learning while maintaining modularity, interpretability, and ensemble diversity.

5.3. Computational efficiency and deployability

Table 4. Computational footprint of candidate models (mean ± std over outer folds). Train time is reported per outer fold. Inference latency is measured with batch size = 1. "Proposed (cached embeddings)" reports the prediction-time cost, excluding the offline cost of precomputing TabNet embeddings.

Model	Train time/fold (s)	Inference latency (ms/sample)	Peak memory (MB)	Params/Trees
LR	0.007 ± 0.001	0.004 ± 0.002	0.067 ± 0.005	31 params
XGBoost	0.224 ± 0.014	0.014 ± 0.006	0.159 ± 0.033	100 trees (depth = 5)
LightGBM	0.122 ± 0.055	0.003 ± 0.001	1.23 ± 0.180	126 trees (leaves = 31)
FT-Transformer	11.729 ± 1.570	0.150 ± 0.029	0.150 ± 0.067	~16 k params
TabNet	32.960 ± 5.846	0.464 ± 0.096	1.97 ± 0.100	~526 k params
Proposed (full)	44.051 ± 18.054	0.300 ± 0.053	5.56 ± 0.430	Lightweight TabNet(103k) + 167 trees + LR(3)
Proposed (cached embeddings)	44.051 ± 18.054	0.018 ± 0.001	5.56 ± 0.430	167 trees + LR(3)

Note: All experiments were run on Intel(R) Core(TM) i5-8250U CPU (PyTorch-CPU). Peak memory denotes the maximum incremental system RAM usage during model fitting or inference, excluding the dataset loaded in memory. Absolute values are implementation- and hardware-dependent; the table is intended for relative comparison under a consistent environment.

Computational Efficiency and Deployability. Table 4 summarizes the relative computational footprint of the proposed framework and representative baselines under a consistent CPU environment. Tree-based models (XGBoost/LightGBM) exhibit sub-millisecond inference due to efficient tabular traversal, whereas deep tabular models (TabNet, FT-Transformer) incur higher latency from neural forward passes. Our "Proposed (full)" mode includes lightweight TabNet-based feature generation at prediction time but remains competitive in inference latency (0.300 ms/sample) while offering improved robustness and interpretability through modular stacking. For deployment settings where repeated scoring is required (e.g., batch evaluation, monitoring, or repeated what-if analyses), we further provide a "cached embeddings" mode: TabNet embeddings are precomputed offline once, and prediction relies only on the tree learners and a lightweight logistic-regression meta-learner, reducing latency to 0.018 ms/sample. Importantly, the reported inference latency for the cached embeddings mode excludes the one-time offline cost of precomputing TabNet embeddings. This two-mode design aligns with the special issue's emphasis on resource-constrained computational intelligence by enabling a practical trade-off between end-to-end online feature generation and ultra-fast inference via caching. Note that caching is most suitable for batch or repeated scoring, while real-time streaming scenarios may require online embedding generation.

5.4. Dataset limitations and generalizability

Several limitations of this study should be acknowledged. First, the model was evaluated exclusively on the WDBC dataset. Although WDBC is one of the most widely adopted benchmark datasets in breast cancer CAD research, it is relatively small and curated, and therefore may not fully reflect the variability, noise, and heterogeneity encountered in real-world clinical settings. Accordingly, our primary objective of this study is methodological validation and architectural exploration rather than direct clinical deployment.

Second, the limited dataset size constrains claims regarding generalizability. While nested cross-validation and fixed-budget hyperparameter tuning were employed to mitigate overfitting, model performance may vary when applied to datasets with different distributions, levels of class imbalance, or feature characteristics. Real-world clinical cohorts often exhibit distribution shifts that can affect calibration, sensitivity, and the stability of feature attributions.

Furthermore, real-world clinical measurements are often affected by various sources of noise and acquisition uncertainty, which can impact the stability of deep feature representations. To examine this aspect, we conducted an additional robustness analysis by introducing Gaussian noise with varying magnitudes to the input features.

Although evaluation on additional large-scale clinical datasets would further strengthen the assessment of generalizability, the current experiments provide an initial methodological validation of the proposed framework under standard benchmark conditions and controlled noise perturbations.

The results indicate that TabNet-derived features, when used in isolation, exhibit increased sensitivity to noise perturbations. However, this sensitivity is substantially mitigated within the proposed stacking architecture, where the complementary robustness of tree-based learners and the adaptive weighting mechanism of the meta-learner help stabilize predictions. This observation suggests that the proposed framework is better suited to realistic clinical conditions, where data noise is unavoidable, despite being evaluated on a controlled benchmark dataset.

5.5. Implications for clinical integration and future work

The proposed framework is best viewed as a transparent decision-support system rather than a standalone diagnostic tool. Its modular architecture enables flexible integration into clinical information systems, where it could assist clinicians by highlighting influential biomarkers and providing interpretable risk scores alongside routine assessments. Furthermore, its reliance on structured tabular data makes it broadly applicable to various clinical contexts beyond breast cancer diagnosis.

In future work, we will focus on external validation using larger, multi-center datasets and, where possible, prospective clinical studies. Such efforts are necessary to assess robustness under real-world variability and to evaluate performance across patient populations. Additionally, further experiments on datasets with different distributions and imbalance ratios will help clarify how the proposed architecture generalizes beyond benchmark settings.

Overall, despite being evaluated on a single benchmark dataset, we establish a principled and interpretable framework for combining deep tabular representation learning with ensemble modeling. By reframing TabNet as a feature generator within a stacked architecture, the proposed approach provides a scalable methodological foundation that can be extended and validated in future clinical research.

6. Conclusions

In this paper, we proposed a novel interpretable stacking ensemble model that successfully integrates the feature generation power of TabNet with the predictive accuracy of heterogeneous tree-based models for breast cancer diagnosis.

Our major contributions are the novel use of TabNet as a feature generator, design of a dual-channel stacking architecture with differentiated feature inputs to enhance model diversity, and construction of an end-to-end SHAP-based framework for multi-level interpretability.

Despite promising results, this study had several limitations: 1) It was conducted on a single, relatively small public dataset (i.e., WDBC), which may limit the external validity of our findings. 2) The features generated by TabNet, although powerful, could be sensitive to noise in the input data. 3) The model has not been validated in a real-world clinical setting.

In future studies, researchers should focus on validating the model on larger, multi-center datasets to ensure its robustness and generalizability. Further investigation into the noise sensitivity of TabNet-generated features is also warranted. In future work, we will explore multimodal integration combining tabular and imaging data, as well as validation on prospective clinical datasets. Ultimately, prospective clinical trials are necessary to assess the real-world utility of this model and its potential for integration into clinical diagnostic workflows. With further validation on multi-center datasets, this framework could serve as a trustworthy decision-support tool to assist pathologists in breast cancer diagnosis. This study demonstrated that interpretability and predictive power are not mutually exclusive.

Beyond the directions outlined above, several additional research avenues merit further investigation. First, systematic studies on improving the robustness of TabNet-generated representations, including noise-aware training strategies and stability-regularized feature generation, could further enhance reliability under real-world clinical conditions. Second, researchers may explore uncertainty estimation and calibration techniques within the stacking framework to support safer decision-making in high-risk diagnostic scenarios. Third, although we focused on breast cancer diagnosis, the proposed architecture is not disease-specific and may be extended to other clinical risk prediction tasks involving structured tabular data.

These future efforts will help position the proposed framework not only as a high-performing benchmark model, but also as a flexible and generalizable methodological template for interpretable clinical AI.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by the research fund of Hanyang University (HY-202600000001873). Lin Xia and Yoona Chung contributed equally to this work.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. J. Zhang, Y. Lu, N. Zhang, Z. Yu, H. Li, R. He, et al., Global burden of female breast cancer and its association with socioeconomic development status, 1990–2044, *Cancer Rep.*, **6** (2023), e1827. https://doi.org/10.1002/cnr2.1827
2. M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, et al., Current and future burden of breast cancer: Global statistics for 2020 and 2040, *Breast*, **66** (2022), 15–23. https://doi.org/10.1016/j.breast.2022.08.010
3. F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, et al., Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, **74** (2024), 229–263. https://doi.org/10.3322/caac.21838
4. L. E. M. Duijm, M. W. J. Louwman, J. H. Groenewoud, L. V. van de Poll-Franse, J. Fracheboud, J. W. Coebergh, Interobserver variability in mammography screening and effect of type and number of readers on screening outcome, *Br. J. Cancer*, **100** (2009), 901–907. https://doi.org/10.1038/sj.bjc.6604954
5. W. A. Berg, C. Campassi, P. Langenberg, M. J. Sexton, Breast imaging reporting and data system: Inter- and intraobserver variability in feature analysis and final assessment, *AJR Am. J. Roentgenol.*, **174** (2000), 1769–1777. https://doi.org/10.2214/ajr.174.6.1741769
6. A. Labrada, B. D. Barkana, A comprehensive review of computer-aided models for breast cancer diagnosis using histopathology images, *Bioengineering*, **10** (2023), 1289. https://doi.org/10.3390/bioengineering10111289
7. T. Alam, W. C. Shia, F. R. Hsu, T. Hassan, Improving breast cancer detection and diagnosis through semantic segmentation using the Unet3+ deep learning framework, *Biomedicines*, **11** (2023), 1536. https://doi.org/10.3390/biomedicines11061536
8. R. Masud, M. Al-Rei, C. Lokker, Computer-aided detection for breast cancer screening in clinical settings: scoping review, *JMIR Med. Inf.*, **7** (2019), e12660. https://doi.org/10.2196/12660
9. L. Guo, Y. Xie, J. He, X. Li, W. Zhou, Q. Chen, Breast cancer prediction model based on clinical and biochemical characteristics: Clinical data from patients with benign and malignant breast tumors from a single center in South China, *J. Cancer Res. Clin. Oncol.*, **149** (2023), 13257–13269. https://doi.org/10.1007/s00432-023-05181-4
10. L. Zhao, Y. Zhang, X. Luo, Y. Zhang, Y. M. Cheung, K. Li, Selecting heterogeneous features based on unified density-guided neighborhood relation for complex biomedical data analysis, in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (2023), 771–778. https://doi.org/10.1109/BIBM58861.2023.10386030
11. Y. Zhang, X. Chen, L. Zhao, Y. Ji, P. Liu, Y. M. Cheung, Online heterogeneous feature selection, *IEEE Trans. Cybern.*, **56** (2025), 2224–2237. https://doi.org/10.1109/TCYB.2025.3635888
12. K. A. Ahmed, I. Humaira, A. R. Khan, M. S. Hasan, M. Islam, A. Roy, et al., Advancing breast cancer prediction: Comparative analysis of ML models and deep learning-based multi-model ensembles on original and synthetic datasets, *PLoS One*, **20** (2025), e0326221. https://doi.org/10.1371/journal.pone.0326221
13. A. Y. Yıldız, A. Kalayci, Gradient boosting decision trees on medical diagnosis over tabular data, in *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, IEEE, (2025), 1–8. https://doi.org/10.1109/ICAD65464.2025.11114069
14. H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, G. Cai, Classification prediction of breast cancer based on machine learning, *Comput. Intell. Neurosci.*, **2023** (2023), 6530719. https://doi.org/10.1155/2023/6530719

15. C. Meaney, X. Wang, J. Guan, T. A. Stukel, Comparison of methods for tuning machine learning model hyper-parameters: With application to predicting high-need high-cost health care users, *BMC Med. Res. Methodol.*, **25** (2025), 134. https://doi.org/10.1186/s12874-025-02561-x
16. A. Maleki, M. Raahemi, H. Nasiri, Breast cancer diagnosis from histopathology images using deep neural network and XGBoost, *Biomed. Signal Process. Control*, **86** (2023), 105152. https://doi.org/10.1016/j.bspc.2023.105152
17. L. Yan, Z. Liang, H. Zhang, G. Zhang, W. Zheng, C. Han, et al., A domain knowledge-based interpretable deep learning system for improving clinical breast ultrasound diagnosis, *Commun. Med.*, **4** (2024), 90. https://doi.org/10.1038/s43856-024-00518-7
18. D. Muhammad, M. Bendechache, Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis, *Comput. Struct. Biotechnol. J.*, **24** (2024), 542–560. https://doi.org/10.1016/j.csbj.2024.02.012
19. W. N. Ismail, H. A. Alsalamah, E. Mohamed, GA-Stacking: A new stacking-based ensemble learning method to forecast the COVID-19 outbreak, *Comput. Mater. Contin.*, **74** (2023), 3945–3976. https://doi.org/10.32604/cmc.2023.031194
20. J. A. M. Sidey-Gibbons, C. J. Sidey-Gibbons, Machine learning in medicine: A practical introduction, *BMC Med. Res. Methodol.*, **19** (2019), 1–64. https://doi.org/10.1186/s12874-019-0681-4
21. S. Nusinovici, Y. C. Tham, M. Y. C. Yan, D. S. W. Ting, J. Li, C. Sabanayagam, et al., Logistic regression was as good as machine learning for predicting major chronic diseases, *J. Clin. Epidemiol.*, **122** (2020), 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002
22. M. A. Khan, T. Mazhar, M. M. Yaqoob, M. B. Khan, A. K. Jilani Saudagar, Y. Y. Ghadi, et al., Optimal feature selection for heart disease prediction using modified Artificial Bee Colony and K-nearest neighbors, *Sci. Rep.*, **14** (2024), 26241. https://doi.org/10.1038/s41598-024-78021-1
23. M. Alizade-Harakiyan, A. Khodaei, A. Yousefi, H. Zamani, A. Mesbahi, Decision tree-based machine learning algorithm for prediction of acute radiation esophagitis, *Biochem. Biophys. Rep.*, **42** (2025), 101991. https://doi.org/10.1016/j.bbrep.2025.101991
24. J. Song, Y. Gao, P. Yin, Y. Li, Y. Li, J. Zhang, et al., The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms, *Risk Manage. Healthcare Policy*, **14** (2021), 1175–1187. https://doi.org/10.2147/RMHP.S297838
25. A. Qaiser, S. Manzoor, A. H. Hashmi, H. Javed, A. Zafar, J. Ashraf, Support vector machine outperforms other machine learning models in early diagnosis of dengue using routine clinical data, *Adv. Virol.*, **2024** (2024), 5588127. https://doi.org/10.1155/2024/5588127
26. H. Yuan, K. Yu, F. Xie, L. Liu, S. Sun, Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare, *Med. Adv.*, **2** (2024), 205–237. https://doi.org/10.1002/med4.75
27. N. Hariprasad, S. Duraimurugan, M. Sushmitha, S. Vedha Shri, A comparative study of machine learning classifiers and ensemble method for breast cancer detection using XAI technique, in *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, IEEE, (2024), 1–5. https://doi.org/10.1109/DELCON64804.2024.10866753
28. R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, *Inf. Fusion*, **81** (2022), 84–90. https://doi.org/10.1016/j.inffus.2021.11.011
29. M. M. Ghiasi, S. Zendehboudi, Application of decision tree-based ensemble learning in the classification of breast cancer, *Comput. Biol. Med.*, **128** (2021), 104089. https://doi.org/10.1016/j.compbiomed.2020.104089

30. T. Arravalli, K. Chadaga, H. Muralikrishna, N. Sampathila, D. Cenitta, R. Chadaga, et al., Detection of breast cancer using machine learning and explainable artificial intelligence, *Sci. Rep.*, **15** (2025), 26931. https://doi.org/10.1038/s41598-025-12644-w
31. A. Rovshenov, S. Peker, Performance comparison of different machine learning techniques for early prediction of breast cancer using Wisconsin breast cancer dataset, in *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, IEEE, (2022), 1–6. https://doi.org/10.1109/IISEC56263.2022.9998248
32. S. Ö. Arık, T. Pfister, TabNet: Attentive interpretable tabular learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826
33. D. H. Wolpert, Stacked generalization, *Neural Networks*, **5** (1992), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
34. A. Daza, J. Bobadilla, J. C. Herrera, A. Medina, N. Saboya, K. Zavaleta, et al., Stacking ensemble-based hyperparameters for diagnosing heart disease: Future works, *Results Eng.*, **21** (2024), 101894. https://doi.org/10.1016/j.rineng.2024.101894
35. H. W. Zhang, Y. R. Wang, B. Hu, B. Song, Z. J. Wen, L. Su, et al., Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases, *Sci. Rep.*, **14** (2024), 28575. https://doi.org/10.1038/s41598-024-80210-x
36. W. Van Loon, M. Fokkema, B. Szabo, M. de Rooij, Stacked penalized logistic regression for selecting views in multi-view learning, *Inf. Fusion*, **61** (2020), 113–123. https://doi.org/10.1016/j.inffus.2020.03.006
37. W. Van Loon, M. Fokkema, B. Szabo, M. de Rooij, View selection in multi-view stacking: Choosing the meta-learner, *Adv. Data Anal. Classif.*, **19** (2024), 579. https://doi.org/10.1007/s11634-024-00587-5
38. M. Garouani, A. Barhrhouj, O. Teste, XStacking: An effective and inherently explainable framework for stacked ensemble learning, *Inf. Fusion*, **124** (2025), 103358. https://doi.org/10.1016/j.inffus.2025.103358
39. L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data, in *Advances in Neural Information Processing Systems*, (2022), 507–520. https://doi.org/10.52202/068431-0037
40. K. Tumer, J. Ghosh, Error correlation and error reduction in ensemble classifiers, *Connect. Sci.*, **8** (1996), 385–404. https://doi.org/10.1080/095400996116839
41. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, *BMC Med.*, **17** (2019), 195. https://doi.org/10.1186/s12916-019-1426-2
42. L. P. Joseph, E. A. Joseph, R. Prasad, Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture, *Comput. Biol. Med.*, **151** (2022), 106178. https://doi.org/10.1016/j.compbiomed.2022.106178
43. D. Kumar, B. Bakariya, C. Verma, Z. Illes, LivXAI-Net: An explainable AI framework for liver disease diagnosis with IoT-based real-time monitoring support, *Comput. Methods Programs Biomed.*, **270** (2025), 108950. https://doi.org/10.1016/j.cmpb.2025.108950
44. R. O. Alabi, A. A. Mäkitie, M. Elmusrati, A. Almangush, Y. Tiblom Ehrsson, G. Laurell, Machine learning explainability for survival outcome in head and neck squamous cell carcinoma, *Int. J. Med. Inf.*, **199** (2025), 105873. https://doi.org/10.1016/j.ijmedinf.2025.105873

45. Y. Nohara, K. Matsumoto, H. Soejima, N. Nakashima, Explanation of machine learning models using Shapley additive explanation and application for real data in hospital, *Comput. Methods Programs Biomed.*, **214** (2022), 106584. https://doi.org/10.1016/j.cmpb.2021.106584
46. S. H. B. Fard, A. H. Baharvand, A. H. Poursaeed, M. Doostizadeh, Explainable stacked ensemble model for short-term net load forecasting, *IET Renewable Power Gener.*, **19** (2025), e70145. https://doi.org/10.1049/rpg2.70145
47. A. Rauschenberger, E. Glaab, M. A. van de Wiel, Predictive and interpretable models via the stacked elastic net, *Bioinformatics*, **37** (2021), 2012–2016. https://doi.org/10.1093/bioinformatics/btab046
48. Q. Teng, Z. Liu, Y. Song, K. Han, Y. Lu, A survey on the interpretability of deep learning in medical diagnosis, *Multimed. Syst.*, **28** (2022), 2335–2355. https://doi.org/10.1007/s00530-022-00960-4
49. J. Fehr, B. Citro, R. Malpani, C. Lippert, V. I. Madai, A trustworthy AI reality-check: The lack of transparency of artificial intelligence products in healthcare, *Front. Digital Health*, **6** (2024), 1267290. https://doi.org/10.3389/fdgth.2024.1267290
50. I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel, et al., A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges, *Inf. Med. Unlocked*, **51** (2024), 101587. https://doi.org/10.1016/j.imu.2024.101587
51. M. Zahedi, S. Das, P. K. Dabla, S. Kandar, MICE-based enhanced TabNet modeling for CVD risk prediction in T2DM patients: a graph-based approach with KS-validated SMOTE and class weight optimization, *Biomed. Signal Process. Control*, **112** (2026), 108672. https://doi.org/10.1016/j.bspc.2025.108672
52. H. Wang, J. Ding, W. Wang, S. Li, L. Song, J. Bai, Enhancing predictive accuracy for urinary tract infections post-pediatric pyeloplasty with explainable AI: An ensemble TabNet approach, *Sci. Rep.*, **15** (2025), 2455. https://doi.org/10.1038/s41598-024-82282-1
53. C. Li, Y. Wang, M. Li, Y. Zheng, Y. Luo, W. Zhong, GAT-enhanced TabNet model with heterogeneous tabular and dependency graph information feature fusion for multi-disease coexistence risk prediction, *Comput. Methods Programs Biomed.*, **241** (2025), 109080. https://doi.org/10.1016/j.cmpb.2025.109080
54. M. N. H. Chowdhury, M. B. I. Reaz, S. H. M. Ali, M. L. Crespo, S. Ahmad, G. M. Salim, et al., Deep learning for early detection of chronic kidney disease stages in diabetes patients: A TabNet approach, *Artif. Intell. Med.*, **166** (2025), 103153. https://doi.org/10.1016/j.artmed.2025.103153
55. A. Zeng, H. Rong, D. Pan, L. Jia, Y. Zhang, F. Zhao, et al., Discovery of genetic biomarkers for Alzheimer's disease using adaptive convolutional neural networks ensemble and genome-wide association studies, *Interdiscip. Sci. Comput. Life Sci.*, **13** (2021), 787–800. https://doi.org/10.1007/s12539-021-00470-3
56. T. Xie, Z. Tan, H. Xiao, B. Sun, Y. Zhang, DE3S: Dual-enhanced soft-sparse shape learning for medical early time series classification, in *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (2025), 1891–1896. https://doi.org/10.1109/BIBM66473.2025.11357232
57. H. Qi, Y. Hu, R. Fan, L. Deng, Tab-Cox: An interpretable deep survival analysis model for patients with nasopharyngeal carcinoma based on TabNet, *IEEE J. Biomed. Health Inf.*, **28** (2024), 4937–4950. https://doi.org/10.1109/JBHI.2024.3397955
58. R. Yasmeen, L. Khan, A. Choi, Heart disease prediction using hybrid TabNet architecture with stacked ensemble learning, *Front. Physiol.*, **16** (2025), 1665128. https://doi.org/10.3389/fphys.2025.1665128

59. E. Hildt, What is the role of explainability in medical artificial intelligence? A case-based approach, *Bioengineering*, **12** (2025), 375. https://doi.org/10.3390/bioengineering12040375
60. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., A survey on deep learning in medical image analysis, *Med. Image Anal.*, **42** (2017), 60–88. https://doi.org/10.1016/j.media.2017.07.005
61. P. J. Rousseeuw, A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, 1987. https://doi.org/10.1002/0471725382
62. S. Uddin, A. Khan, M. E. Hossain, M. A. Moni, Comparing different supervised machine learning algorithms for disease prediction, *BMC Med. Inf. Decis. Making*, **19** (2019), 281. https://doi.org/10.1186/s12911-019-1004-8
63. J. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, **29** (2001), 1189–1232. https://doi.org/10.1214/aos/1013203451
64. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, (2016), 785–794. https://doi.org/10.1145/2939672.2939785
65. T. Hospedales, A. Antoniou, P. Micaelli, A. Storkey, Meta-learning in neural networks: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 5149–5169. https://doi.org/10.1109/TPAMI.2021.3079209
66. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems*, (2017), 4765–4774.
67. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, et al., From local explanations to global understanding with explainable AI for trees, *Nat. Mach. Intell.*, **2** (2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9
68. C. W. Elston, I. O. Ellis, Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up, *Histopathology*, **19** (1991), 403–410. https://doi.org/10.1111/j.1365-2559.1991.tb00229.x
69. W. H. Wolberg, W. N. Street, O. L. Mangasarian, Importance of nuclear morphology in breast cancer prognosis, *Clin. Cancer Res.*, **5** (1999), 3542–3548.
70. S. Aaltomaa, P. Lipponen, M. Eskelinen, E. Alhava, K. Syrjänen, Nuclear morphometry and mitotic indexes as prognostic factors in breast cancer, *Eur. J. Surg.*, **157** (1991), 319–324.
71. D. J. Niranjan Pandian, A. Ramdas, M. M. Ambroise, Image analysis-assisted nuclear morphometric study of benign and malignant breast aspirates, *J. Microsc. Ultrastruct.*, **9** (2021), 114–118. https://doi.org/10.4103/Jmau.Jmau_17_20
72. A. Kashyap, M. Jain, S. Shukla, M. Andley, Role of nuclear morphometry in breast cancer and its correlation with cytomorphological grading of breast cancer: A study of 64 cases, *J. Cytol.*, **35** (2018), 41–45. https://doi.org/10.4103/Joc.Joc_237_16
73. W. H. Wolberg, O. L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *Proc. Natl. Acad. Sci. U. S. A.*, **87** (1990), 9193–9196. https://doi.org/10.1073/pnas.87.23.9193
74. J. Keyl, P. Keyl, G. Montavon, R. Hosch, A. Brehmer, L. Mochmann, et al., Decoding pan-cancer treatment outcomes using multimodal real-world data and explainable artificial intelligence, *Nat. Cancer*, **6** (2025), 307–322. https://doi.org/10.1038/s43018-024-00891-1
75. S. Khademi, Z. Hajiakhondi-Meybodi, G. Vaseghi, N. Sarrafzadegan, A. Mohammadi, FH-TabNet: Multi-class familial hypercholesterolemia detection via a multi-stage tabular deep learning network, in *2024 32nd European Signal Processing Conference (EUSIPCO)*, EURASIP, (2024), 1416–1420. https://doi.org/10.23919/EUSIPCO63174.2024.10715254

76. A. A. M. A. E. Zaher, A. Eldeib, Breast cancer classification using deep belief networks, *Expert Syst. Appl.*, **46** (2015), 139–144. https://doi.org/10.1016/j.eswa.2015.10.015
77. A. Mahbod, N. Saeidi, S. Hatamikia, R. Woitek, Evaluating pre-trained convolutional neural networks and foundation models as feature extractors for content-based medical image retrieval, *Eng. Appl. Artif. Intell.*, **150** (2025), 110571. https://doi.org/10.1016/j.engappai.2025.110571

Appendix

A. Hyperparameter search space and optimal configurations

A.1. Hyperparameter tuning protocol

Hyperparameter optimization was performed within the nested cross-validation framework described in Section 3.5. Tree-based models (logistic regression, XGBoost, and LightGBM) were optimized using the Optuna optimization framework, while deep learning models (FT-Transformer, Tab-Transformer, and TabNet) were tuned using random search due to their higher-dimensional parameter spaces.

All models were optimized under an identical search budget of 40 trials to ensure fairness in the comparison. Early stopping with a patience of 10 epochs was applied during training to mitigate overfitting on the relatively small WDBC dataset. At the epoch level, validation loss was monitored to stabilize convergence, while the final hyperparameter configuration was selected based on the F1-score obtained from OOF predictions in the nested cross-validation procedure.

Table A1. LR hyperparameters.

Hyperparameter	Search space	Optimal value
C	(1e-4, 100)	0.1
penalty	$\{L_1, L_2\}$	L_2
class_weight	{None, balanced}	balanced

Table A2. XGBoost hyperparameters.

Hyperparameter	Search space	Optimal value
n_estimators	(100, 1000)	666
max_depth	(3, 10)	6
learning_rate	(0.01, 0.3, log = True)	0.14987
subsample	(0.6, 1.0)	0.76446
colsample_bytree	(0.6, 1.0)	0.73699
reg_alpha	(0.0, 1.0)	0.46945
reg_lambda	(0.0, 1.0)	0.47206

Table A3. LightGBM hyperparameters.

Hyperparameter	Search space	Optimal value
num_leaves	(10, 300)	146
learning_rate	(0.01, 0.3, log = True)	0.07731
subsample	(0.6, 1.0)	0.82083
colsample_bytree	(0.6, 1.0)	0.66995
reg_alpha	(0.0, 1.0)	0.76540
reg_lambda	(0.0, 1.0)	0.44640
min_child_weight	(1, 10)	5

Table A4. FT-Transformer hyperparameters.

Hyperparameter	Search space	Optimal value
n_blocks	{2, 3, 4}	3
d_token	{8, 16, 32}	32
learning_rate	{1e-5, 5e-5, 1e-4, 3e-4}	1e-5
weight_decay	{1e-6, 1e-5, 1e-4, 1e-3}	1e-6
batch_size	{32, 64, 128}	64
dropout	{0.1, 0.2, 0.3}	0.2

Table A5. Tab-Transformer hyperparameters.

Hyperparameter	Search space	Optimal value
dim	{16, 32, 48, 64}	64
depth	{1, 2, 3, 4}	1
heads	{2, 4, 8}	8
attn_dropout	{0.2, 0.3, 0.4, 0.5}	0.5
ff_dropout	{0.1, 0.2, 0.3}	0.1
weight_decay	{1e-5, 1e-4, 1e-3}	1e-3
learning_rate	{1e-5, 5e-5, 1e-4, 3e-4}	1e-5

Table A6. TabNet hyperparameters.

Hyperparameter	Search space	Optimal value
n_a, n_d	{8, 16, 24, 32}	32
n_steps	{3, 4, 5, 6}	4
gamma	{1.0, 1.2, 1.4, 1.6, 1.8}	1.8
lambda_sparse	{1e-4, 1e-3, 1e-2}	1e-4
learning_rate	{1e-3, 1e-2, 2e-3, 5e-3}	1e-3
weight_decay	{1e-5, 1e-4, 1e-3, 1e-2}	1e-4