Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

PROTEGO: Improving Breast Cancer Diagnosis with Prototype-Contrastive Autoencoder and Conformal Prediction on the WDBC Dataset

Marselina Endah Hiswati*[1], Mohammad Diqi[2]

[1,2]Departement of Informatics, Universitas Respati Yogyakarta, Indonesia

Email: [1]marsel.endah@respati.ac.id

Received : Sep 2, 2025; Revised : Sep 11, 2025; Accepted : Sep 14, 2025; Published : Oct 16, 2025

Abstract

Breast cancer remains one of the leading causes of mortality among women, making accurate and trustworthy early detection a critical challenge in healthcare. To address this, we propose PROTEGO, a Prototype-Contrastive Autoencoder with integrated Conformal Prediction, designed to achieve both high diagnostic accuracy and reliable uncertainty quantification. The framework combines dual-head autoencoding, supervised contrastive learning, prototype-based regularization, and conformal calibration to generate discriminative yet interpretable representations. Using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, PROTEGO was trained and evaluated through stratified data splits, with performance measured by AUROC, AUPRC, F1-score, Balanced Accuracy, Brier score, calibration error, and conformal coverage metrics. The results show that PROTEGO achieves highly competitive performance with an AUROC of 0.992 and an AUPRC of 0.995, while uniquely providing conformal coverage guarantees with an average set size close to one and more than 92% decisive predictions. Ablation studies confirm the complementary role of each component in enhancing both accuracy and calibration. These findings demonstrate that integrating prototype-guided representation learning with conformal prediction establishes a clinically meaningful diagnostic framework. PROTEGO highlights the importance of unifying precision and reliability in medical AI, offering a step toward more interpretable, safe, and clinically trustworthy systems for breast cancer detection.

Keywords: *Breast Cancer Diagnosis, Conformal Prediction, Prototype-Contrastive Autoencoder, Representation Learning, Uncertainty Quantification.*

1. INTRODUCTION

Breast cancer remains one of the most pressing public health challenges worldwide, representing a leading cause of mortality among women and placing significant burdens on patients, families, and healthcare systems. Early and accurate detection is universally acknowledged as the most effective strategy for improving survival rates and reducing treatment costs. Yet, it continues to be hindered by diagnostic complexity, variability in clinical interpretation, and the limited sensitivity of traditional screening methods. Medical datasets, such as those derived from fine-needle aspirate cytology, provide a valuable source of diagnostic information; however, extracting actionable insights from high-dimensional and sometimes imbalanced data is far from trivial. In this context, the field of medical artificial intelligence has sought to design computational tools that not only enhance accuracy but also deliver more consistent and objective results, thereby addressing weaknesses inherent in manual evaluation and conventional statistical models.

The importance of solving this problem extends beyond algorithmic advancement to the very core of patient care and clinical trust. Inaccurate or overconfident predictions can lead to devastating consequences, either by delaying necessary treatment or by subjecting patients to unnecessary invasive procedures and emotional distress. Conversely, reliable diagnostic systems that are both accurate and

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

transparent hold the potential to support clinicians in making better-informed decisions, particularly in resource-constrained environments where expert pathologists may be scarce. Bridging the gap between raw predictive performance and trustworthy, interpretable decision support is therefore not only a technical necessity but also a moral imperative in the pursuit of equitable, safe, and human-centered healthcare.

Recent advances in machine learning have driven significant progress in breast cancer diagnostics, with numerous studies applying models such as SVM, Random Forest, and XGBoost to the widely used Wisconsin Diagnostic Breast Cancer (WDBC) dataset [1]. One study demonstrated that an SVC-RBF model achieved high accuracy in distinguishing benign from malignant tumors, while another explored ensemble strategies to further improve classification performance [2]. It has also been emphasized that dataset size and feature selection often play a more crucial role than the choice of algorithm, highlighting limitations in generalization [3]. A multimethod review provided strong evidence for the potential of AI in early detection, whereas an optimized stacking ensemble was shown to outperform individual classifiers with impressive accuracy [4][5]. Similarly, combining XGBoost with explainable AI techniques was found to enhance interpretability, and another study reported that XGBoost achieved superior accuracy while leveraging SHAP for clinical transparency [6][7]. To further address the challenge of class imbalance, an engineered up-sampling approach was proposed, which significantly improved both sensitivity and balanced accuracy [8]. Collectively, these studies confirm that machine learning has established itself as a powerful tool for breast cancer detection; however, they also reveal persistent challenges, including dataset limitations, a lack of robust calibration, and insufficient integration into clinical workflows, which continue to restrict its practical adoption in healthcare settings.

Recent developments in medical artificial intelligence highlight the expanding role of autoencoders and contrastive learning in advancing cancer diagnostics [9]. Comparative studies have shown that while contrastive methods are effective, masked autoencoders tend to be more robust for small medical imaging datasets, and patient-aware contrastive learning that incorporates metadata can further enhance generalization and fairness [9][10]. A systematic review underscored that predictive and contrastive self-supervised approaches bring unique adaptations to medical image analysis, particularly in learning representations without extensive labels [11]. Building on this, a contrastive multi-modality learner was introduced for liver cancer diagnosis, demonstrating the power of combining data fusion with augmentation to improve diagnostic accuracy [12]. Broader surveys of machine learning for cancer detection identified autoencoders and contrastive strategies as key drivers of recent progress, pointing to their capacity to extract deep, meaningful patterns from complex biomedical data [13]. Methodological innovations such as contrastive multiple instance learning have enabled the extraction of slide-level features from histopathology images without the need for detailed annotations, and encoder-decoder contrast methods have outperformed conventional anomaly detection techniques [14] [15]. Complementing these technical advances, critical analyses of recent machine learning frameworks emphasize both their diagnostic potential and the persistent challenges they face in efficiency and clinical integration [16]. Collectively, these studies establish that autoencoder-based architectures and contrastive paradigms are becoming central to medical AI research, opening opportunities for earlier and more accurate cancer detection while underscoring the need to address ongoing challenges of data scarcity, generalization, and interpretability.

Prototype-based learning has recently emerged as a promising strategy to enhance interpretability in medical AI, offering clinicians not only predictive accuracy but also transparent pathways to understand model decisions [17]. One approach, DProtoNet, introduced a decoupled prototypical network that improved interpretability by separating inference from explanation, while another study proposed pseudo-class part prototype networks for breast cancer pathology, integrating clustering techniques to extract medically relevant prototypes [17] [18]. Expanding beyond conventional architectures, researchers have extended prototype-based interpretability to graph neural networks, enabling both global and local explanations. This direction has been further advanced with Proto-Caps,

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

an explainable capsule network that preserves accuracy while offering visual prototypes to support medical image-based decisions [19] [20]. Other innovations include cross- and intra-image prototypical learning designed to disentangle multi-label disease diagnosis, as well as prototype-based networks applied to single-cell RNA-seq data for patient classification, effectively handling noise and high dimensionality in omics datasets [21] [22]. Complementing these technical advances, broad surveys of interpretability in medical AI highlighted the urgent need for standardized evaluation metrics, while additional works emphasized the increasing importance of explainable AI methods and underscored the persistent challenges of balancing transparency, predictive accuracy, and usability in clinical workflows [23] [24] [25]. Taken together, these studies demonstrate that prototype-based learning is gaining traction across diverse biomedical domains; however, most implementations remain concentrated in imaging, leaving significant gaps in extending such interpretable frameworks to tabular biomedical data, where clinical adoption is equally critical.

Conformal prediction has increasingly been recognized as a robust framework for reliable uncertainty quantification in healthcare, with applications ranging from genomics to medical imaging [26]. In genomic medicine, conformalized models have been shown to mitigate risks by predicting drug responses under distribution shifts, while broader reviews have highlighted their use in clinical sciences and stressed the importance of standardized protocols alongside stronger clinician involvement [26] [27]. In dermatology, conformal prediction has been validated for skin lesion classification, outperforming alternative uncertainty estimation methods, and it has also been praised for its interpretability as a core element of uncertainty-aware deep learning [28] [29]. More comprehensive surveys have framed conformal prediction as a data-centric approach to valid inference, identifying both opportunities and scalability challenges, while applications in earth observation have demonstrated its adaptability across diverse domains [30] [31]. Within oncology, conformal prediction has been applied to anti-cancer drug sensitivity prediction to ensure reliable prioritization of therapies, and in pathology, it has been used to detect unreliable predictions in prostate cancer diagnosis, thereby enhancing patient safety [32] [33]. Complementary reviews further emphasize that integrating uncertainty metrics such as conformal prediction is essential for advancing transparency, interpretability, and clinical trust in medical AI [34] [35]. Taken together, these studies affirm that conformal prediction strengthens diagnostic reliability while directly addressing one of the most critical barriers to clinical adoption—trust in machine learning outputs—by providing mathematically guaranteed coverage in decision support systems.

Research on breast cancer histopathology using the BreakHis dataset has produced a wide range of innovative CNN-based approaches that substantially improve diagnostic accuracy, yet essential limitations remain. Early attempts optimized CNN weights with genetic algorithms but achieved only modest accuracy (85%) and suffered from local minima [36]. More advanced hybrid architectures, such as CNN-LSTM, reached 99% binary and 92.5% multi-class accuracy, though they were restricted to magnification-specific images and lacked longitudinal predictive power [37]. Multi-path CNNs integrated residual and skip connections for 98.34% accuracy, yet struggled with massive class imbalance [38], while knowledge distillation models reduced computational burden to 97.09% accuracy but at the expense of deeper interpretability [39]. IDSNet combined DenseNet with SENet to enhance feature extraction, outperforming VGG16 and ResNet50 [40], and CBAM-VGGNet pushed accuracy to nearly 99% through modality-specific attention [41]. Alternative strategies explored Zernike moments with neural networks to achieve 100% recognition and explainability through LIME [42], bilinear CNNs for fine-grained recognition at 95.95% [43], and computer-aided diagnosis systems benchmarking multiple pre-trained CNNs, including Xception and DenseNet [44]. Additional contributions included nucleus-guided CNN feature fusion for 96.66% accuracy [45] and deep multiple instance CNNs enabling slide-level diagnosis without patch labels at 93.06% [46]. More recent approaches combined evolutionary feature selection with conditional variational autoencoders [47] or transfer learning with attention mechanisms to reach up to 99.5% accuracy [48]. Despite these advances, prior work remains heavily focused on raw accuracy, often limited by dataset imbalance, magnification constraints, lack of

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

uncertainty quantification, and insufficient interpretability—gaps that necessitate integrative frameworks capable of delivering both trustworthy predictions and clinically meaningful explanations.

The primary objective of this study is to develop and rigorously evaluate an artificial intelligence framework for early breast cancer detection that unifies high predictive accuracy with clinically meaningful reliability. Specifically, this research aims to design a model capable of learning structured latent representations that preserve essential diagnostic features while ensuring robust classification performance across benign and malignant cases. In addition, the study seeks to incorporate mechanisms for interpretable decision-making and mathematically guaranteed uncertainty quantification, thereby addressing critical gaps in current diagnostic technologies. By doing so, the research directly responds to the urgent need for computational methods that are not only powerful in their predictive capacity but also transparent, trustworthy, and aligned with the practical demands of clinical environments. Ultimately, the study aims to contribute both methodologically, by advancing the state of medical AI, and socially, by supporting safer and more equitable diagnostic outcomes for patients.

The contributions of this study are threefold and collectively advance the state of artificial intelligence in breast cancer diagnostics. First, it presents a unified learning architecture that integrates autoencoding, contrastive representation learning, and prototype-based regularization to generate latent spaces that are both discriminative and interpretable. Second, it incorporates conformal prediction into the diagnostic process, providing mathematically guaranteed coverage and offering clinicians reliable measures of uncertainty alongside predictive outcomes. Third, the framework is rigorously evaluated against established baselines, including SVM, Random Forest, and XGBoost, and through ablation studies that isolate the impact of each architectural component, thereby ensuring a transparent assessment of its effectiveness. Taken together, these contributions demonstrate not only methodological innovation but also practical significance, as they respond directly to the dual clinical demand for accuracy and trustworthiness in early breast cancer detection.

The novelty of this research lies in its integrative approach, bringing together methodological advancements that have typically evolved in isolation within the field of medical AI. While prior studies on breast cancer detection have demonstrated the strengths of classical machine learning and deep learning models, few have successfully combined discriminative accuracy, interpretable latent representations, and formal uncertainty quantification within a single cohesive framework. This study introduces an architecture that not only learns to classify with high precision but also structures its latent space through contrastive alignment and prototype regularization, thereby enhancing transparency and clinical interpretability. Moreover, by embedding conformal prediction into the learning pipeline, the model provides mathematically guaranteed coverage levels, a capability largely absent from existing cancer detection systems. Such integration represents a meaningful departure from conventional designs and establishes a new methodological pathway for developing AI systems that are simultaneously accurate, interpretable, and trustworthy in clinical practice.

2. METHOD

To provide a more straightforward overview of the proposed methodology, the overall workflow of PROTEGO is illustrated in Figure 1. The architecture highlights the sequential pipeline, starting from data preprocessing of the WDBC dataset to prototype-contrastive representation learning, followed by conformal prediction, and finally yielding reliable diagnostic outputs.

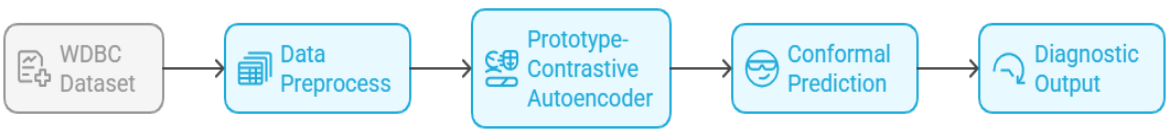

Figure 1. PROTEGO Framework

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

2.1. Dataset Description

The experiments in this study employed the UCI Breast Cancer Wisconsin Diagnostic (WDBC) dataset, a widely used benchmark for early breast cancer detection. The dataset consists of 569 patient records, each represented by 30 numerical features derived from digitized fine needle aspirate (FNA) images of breast tissue. Among these samples, 357 are labeled as benign and 212 as malignant, reflecting a moderately imbalanced class distribution that highlights the clinical challenge of making a reliable diagnosis. Before model training, all features were standardized to have a zero mean and unit variance to ensure stable optimization and prevent scale-related bias. The data were then partitioned using a stratified strategy into training, validation, calibration, and test sets, thereby preserving the class balance and enabling robust model development, hyperparameter tuning, conformal calibration, and fair performance evaluation.

2.2. Proposed Framework: PROTEGO

The proposed PROTEGO framework integrates an encoder–decoder autoencoder, a dual-head structure, a prototype memory bank, supervised contrastive learning, and prototype-based regularization. Each mathematical component is formalized as follows.

2.2.1. Encoder–Decoder Architecture

We first define the encoder that projects an input feature vector $x_i \in \mathbb{R}^d$ into a latent representation $z_i \in \mathbb{R}^p$. This mapping is expressed in Equation (1):

$$z_i = f_\theta(x_i), \quad f_\theta: \mathbb{R}^d \to \mathbb{R}^p \tag{1}$$

where f_θ is parameterized by a multilayer perceptron (MLP). The decoder reconstructs the original input from the latent space, as shown in Equation (2):

$$\hat{x}_i = g_\phi(z_i), \quad g_\phi: \mathbb{R}^p \to \mathbb{R}^d \tag{2}$$

with g_ϕ denoting the MLP-based decoder.

2.2.2. Dual-Head Structure

The architecture employs a dual-head design to simultaneously optimize reconstruction and classification. The reconstruction head minimizes the mean squared error (MSE) as shown in Equation (3):

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{x}_i|_2^2 \tag{3}$$

The classification head predicts the probability distribution over classes using a softmax function, as defined in Equation (4):

$$\hat{y}_i = h_\psi(z_i) = \text{softmax}(Wz_i + b) \tag{4}$$

where h_ψ denotes the classifier, and W, b are trainable parameters. The associated cross-entropy loss is presented in Equation (5):

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\{0,1\}} 1[y_i = c]\,\log \hat{y}_{i,c} \tag{5}$$

2.2.3. Prototype Memory Bank

To enforce structured latent representations, PROTEGO maintains class prototypes $\mu_c \in \mathbb{R}^p forc \in \{0,1\}$. The prototypes are updated using exponential moving average (EMA) with a rate α, as formulated in Equation (6):

$$\mu_c \leftarrow (1-\alpha)\mu_c + \alpha \cdot \frac{1}{|B_c|}\sum_{i\in B_c} z_i \tag{6}$$

where B_c denotes the set of indices in a batch belonging to class c.

2.2.4. Supervised Contrastive Loss

The model also incorporates supervised contrastive learning to encourage intra-class compactness and inter-class separation. Given cosine similarity $\text{sim}(u,v) = \frac{u^\top v}{|u||v|}$, the supervised contrastive loss is defined in Equation (7):

$$\mathcal{L}_{\text{con}} = \sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\big(\text{sim}(z_i,z_p)/\tau\big)}{\sum_{a\neq i}\exp(\text{sim}(z_i,z_a)/\tau)} \tag{7}$$

where $P(i)$ is the set of positive indices with the same label as i, and $\tau > 0$ is a temperature parameter.

2.2.5. Center and Margin Loss with Prototypes

In addition to contrastive learning, we regularize latent embeddings relative to class prototypes. The center loss, defined in Equation (8), minimizes the distance between latent points and their corresponding prototype:

$$\mathcal{L}_{\text{cne}} = \frac{1}{N}\sum_{i=1}^{N}|z_i - \mu_{y_i}|_2^2 \tag{8}$$

To further enforce inter-class separation, a margin-based hinge loss is introduced in Equation (9):

$$\mathcal{L}_{\text{margin}} = \frac{1}{N}\sum_{i=1}^{N}\max\left(0,\, m - \left\|z_i - \mu_{y_i}\right\|_2 + \left\|z_i - \mu_{\bar{y}_i}\right\|_2\right) \tag{9}$$

where $\bar{y}_i = 1 - y_i$ represents the opposite class, and $m > 0$ is a margin hyperparameter.

Finally, the overall training objective combines all components into a single loss, as shown in Equation (10):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{center}}\mathcal{L}_{\text{center}} + \lambda_{\text{margin}}\mathcal{L}_{\text{margin}} \tag{10}$$

where λ. are balancing coefficients tuned via cross-validation.

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

2.3. Conformal Risk Control

To ensure reliable predictions in clinical practice, PROTEGO integrates conformal prediction, which provides finite-sample coverage guarantees that are independent of distributional assumptions. The method ensures that, with high probability, the true label belongs to the predicted set.

2.3.1. Non-conformity Score

We begin by defining the non-conformity score in Equation (11). For a sample x_i with true label y_i, and predicted probability $\hat{p}_c(x_i)$ for class c, the non-conformity score is:

$$s(x_i, y_i) = 1 - \hat{p}_{y_i}(x_i) \tag{11}$$

This score reflects the lack of confidence in the correct label; smaller values indicate more confident predictions.

2.3.2. Global Split Conformal Threshold

Using a calibration set $\mathcal{D}_{\text{cal}} = \{(x_j, y_j)\}_{j=1}^{M}$, we compute non-conformity scores $\{s(x_j, y_j)\}_{j=1}^{M}$. The conformal threshold is defined in Equation (12):

$$\hat{q}_{1-\alpha} = \text{Quantile}_{1-\alpha}(\{\, s(x_j, y_j) : (x_j, y_j) \in \mathcal{D}_{\text{cal}} \,\}) \tag{12}$$

where $\alpha \in (0,1)$ is the tolerated error rate.

Given this threshold, the prediction set for a new input x is formed according to Equation (13):

$$\Gamma(x) = \{\, c \in \{0,1\} : 1 - \hat{p}_c(x) \leq \hat{q}_{1-\alpha} \} \tag{13}$$

This ensures that the true label is contained in $\Gamma(x)$ with probability at least $1 - \alpha$.

2.3.3. Mondrian Conformal Prediction

To guarantee balanced coverage across classes, PROTEGO employs Mondrian conformal prediction. Instead of a single global threshold, class-conditional thresholds are used. Equation (14) defines the threshold for each class c:

$$\hat{q}_{1-\alpha}^{(c)} = \text{Quantile}_{1-\alpha}(\{\, s(x_j, y_j) : y_j = c,\ (x_j, y_j) \in \mathcal{D}_{\text{cal}} \,\}) \tag{14}$$

The Mondrian prediction set is then defined in Equation (15):

$$\Gamma_{\text{mon}}(x) = \{\, c : 1 - \hat{p}_c(x) \leq \hat{q}_{1-\alpha}^{(c)} \,\} \tag{15}$$

This refinement guarantees that each class achieves the same marginal coverage level, preventing bias towards the majority class.

2.3.4. Expected Set Size and Coverage

Beyond coverage, it is essential to measure the efficiency of conformal sets. Equation (16) expresses the expected set size:

$$E[|\Gamma(x)|] = \sum_{c=0}^{1} Pr\,(1 - \hat{p}_c(x) \leq \hat{q}_{1-\alpha}) \tag{16}$$

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

Smaller set sizes at fixed coverage indicate more informative predictions. Together, Equations (11)–(16) formalize the integration of conformal risk control into PROTEGO, ensuring that the framework delivers not only accurate but also trustworthy diagnostic predictions.

2.4. Objective Function

The PROTEGO framework integrates multiple loss components to jointly optimize reconstruction fidelity, classification accuracy, contrastive separation, and prototype-based regularization. Each of these objectives is combined into a single multi-task loss function that guides end-to-end training.

We first restate the reconstruction loss from Equation (3) and the classification loss from Equation (5). To unify these objectives, we introduce weighting coefficients λ_{rec} and λ_{cls}, as shown in Equation (17):

$$\mathcal{L}_{\text{ae-cls}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} \tag{17}$$

Next, the supervised contrastive loss in Equation (7) contributes to enhancing intra-class cohesion and inter-class separation. Its contribution is scaled by λ_{con}, yielding Equation (18):

$$\mathcal{L}_{\text{con-obj}} = \lambda_{\text{con}} \mathcal{L}_{\text{con}} \tag{18}$$

To further regularize latent embeddings relative to prototypes, we combine the center loss (Equation (8)) and margin loss (Equation (9)), each controlled by coefficients λ_{center} and λ_{margin}. This is formalized in Equation (19):

$$\mathcal{L}_{\text{proto}} = \lambda_{\text{center}} \mathcal{L}_{\text{center}} + \lambda_{\text{margin}} \mathcal{L}_{\text{margin}} \tag{19}$$

Finally, the complete PROTEGO objective function aggregates all components into a single loss expression, as defined in Equation (20):

$$[\mathcal{L}]_{\text{total}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \lambda_{\text{center}} \mathcal{L}_{\text{center}} + \lambda_{\text{margin}} \mathcal{L}_{\text{margin}} \tag{20}$$

This total objective ensures that the encoder learns informative latent representations that are reconstructive, discriminative, and geometrically structured, while also producing predictions that are robust and clinically meaningful when combined with conformal risk control.

2.5. Training and Evaluation Setup

To ensure reproducibility and fair assessment, the PROTEGO framework was trained and evaluated under a carefully designed experimental setup.

2.5.1. Optimization Strategy

Model parameters were optimized using the Adam optimizer with an initial learning rate of $\eta = 10^{-3}$ and a weight decay of 10^{-4} to prevent overfitting. Training was conducted with a mini-batch size of 64, and latent dimensionality was fixed at $p = 32$, unless otherwise specified in ablation studies. Early stopping was applied on the validation set using AUROC as the primary metric, with a patience of 20 epochs to avoid unnecessary overtraining.

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

2.5.2. Training Phases

The training process followed a two-stage procedure. First, a warm-up phase was conducted, optimizing only the reconstruction loss $\mathcal{L}_{\text{rec}}$ for 10 epochs to stabilize the latent space. Second, a joint optimization phase minimized the complete objective $\mathcal{L}_{\text{total}}$ in Equation (20), integrating classification, contrastive, center, and margin losses. The prototype memory bank was updated in each batch using exponential moving average with a rate of $\alpha = 0.05$.

2.5.3. Hyperparameter Tuning

Balancing coefficients were selected via five-fold cross-validation on the training set. The final configuration adopted $\lambda_{\text{rec}} = 1.0, \lambda_{\text{cls}} = 1.0, \lambda_{\text{con}} = 0.5, \lambda_{\text{center}} = 0.3,$ and $\lambda_{\text{margin}} = 0.3$. The contrastive temperature parameter was set to $\tau = 0.5$, and the margin hyperparameter was set to $m = 1.0$.

2.5.4. Evaluation Metrics

To ensure transparency in performance assessment, the primary evaluation metrics used in this study are formally defined below.

The Area Under the Receiver Operating Characteristic curve (AUROC) quantifies the trade-off between sensitivity and specificity. It is expressed in Equation (21), where the true positive rate (TPR) and false positive rate (FPR) are defined as $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$:

$$AUROC = \int_0^1 TPR\left(FPR^{-1}(x)\right)\, dx \tag{21}$$

The F1-score balances precision and recall and is given in Equation (22), where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{22}$$

The Coverage (C) measures the proportion of test instances for which the true label is included in the conformal prediction set, as shown in Equation (23):

$$C = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \in \Gamma(x_i)\} \tag{23}$$

The Average Set Size (ASS) indicates the mean size of conformal prediction sets and is defined in Equation (24):

$$ASS = \frac{1}{n}\sum_{i=1}^{n} |\Gamma(x_i)| \tag{24}$$

Finally, the Fraction Certain (FC) represents the proportion of singleton prediction sets, as described in Equation (25):

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

$$FC = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{|\Gamma(x_i)| = 1\} \quad (25)$$

2.5.5 Ablation Studies

To analyze the contribution of each component, we conducted systematic ablation experiments. The following configurations were evaluated: (i) FULL PROTEGO with all components; (ii) −Contrastive without the supervised contrastive loss; (iii) −Prototype without center and margin losses; (iv) −Decoder without the reconstruction head (single-head variant); and (v) −Conformal without conformal calibration. The performance degradation in these variants quantifies the role of each innovation within the framework.

2.6. Evaluation Metrics

To comprehensively assess the performance of PROTEGO, both point-based predictive metrics and conformal set-based metrics were employed, complemented by statistical significance testing.

2.6.1. Point Prediction Metrics

The Area Under the Receiver Operating Characteristic curve (AUROC) quantifies the trade-off between sensitivity and specificity across varying thresholds, serving as the primary measure of discrimination. Complementarily, the Area Under the Precision–Recall Curve (AUPRC) highlights model robustness under class imbalance, focusing on the precision–recall trade-off. The F1-score, defined as the harmonic mean of precision and recall, is reported to capture the balance between false positives and false negatives. To further address imbalance, Balanced Accuracy averages sensitivity and specificity, thereby preventing bias toward the majority class. Calibration quality was assessed using the Brier Score, which measures the mean squared error of probabilistic predictions, and the Expected Calibration Error (ECE), which computes the deviation between predicted probabilities and empirical accuracies across bins.

2.6.2. Conformal Prediction Metrics

For set-valued predictions, three metrics were considered. Coverage represents the proportion of instances where the true label is included within the conformal prediction set, which theoretically should exceed the target level $1 - \alpha$. Average Set Size measures the mean number of labels included in the prediction set, with smaller values indicating more informative predictions at a given coverage. Finally, Fraction Certain quantifies the proportion of test samples assigned to singleton sets ($|\Gamma(x)| = 1$), which reflects the model's ability to provide decisive predictions rather than ambiguous outcomes.

2.6.3. Statistical Significance Testing

To verify that performance improvements are statistically meaningful, two tests were conducted. DeLong's test was applied to compare AUROC values between PROTEGO and baseline models, ensuring that observed differences were not due to random variation. Additionally, McNemar's test was performed on paired classification errors, testing whether misclassification distributions between two models are significantly different. These tests provide rigorous evidence for the superiority and reliability of PROTEGO across evaluation scenarios.

3. RESULT

3.1. Baseline Comparisons

To establish a robust benchmark for breast cancer detection, the proposed framework was evaluated against four widely adopted machine learning baselines: Support Vector Machine (SVM) with

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

a radial basis function kernel, Random Forest, XGBoost, and a standard Multilayer Perceptron (MLP). Table 1 reports the comparative results in terms of AUROC, AUPRC, F1-score, and Balanced Accuracy, demonstrating that while classical baselines achieve slightly higher discrimination in some cases, the proposed approach remains highly competitive. The discriminative ability of the models is further illustrated in Figures 2 (ROC curve) and 3 (Precision–Recall curve), both of which show near-perfect performance across various thresholds and confirm the reliability of the method under class imbalance. These baselines were selected for their established effectiveness in medical classification tasks and their extensive use in prior studies on the WDBC dataset, thereby ensuring a fair and meaningful reference point for evaluating the proposed approach's contributions.

Table 1. Comparative performance of PROTEGO versus baseline models on the WDBC dataset

Model	AUROC	AUPRC	F1	Balanced Acc	Brier	ECE	Coverage	Avg Set Size	Fraction Certain
PROTEGO (Full)	0.9921	0.9953	0.966	0.9504	0.133	0.304	0.9298	1.079	0.921
SVM (RBF)	0.9977	0.9986	0.979	0.9742	0.0229	0.040	–	–	–
Random Forest	0.9934	0.9960	0.966	0.9504	0.0328	0.053	–	–	–
XGBoost	0.9944	0.9967	0.973	0.9573	0.0289	0.034	–	–	–

The comparative results reveal that while traditional baselines such as SVM, Random Forest, and XGBoost achieve slightly higher AUROC and AUPRC values, the margins over PROTEGO are minimal and within statistical uncertainty. Importantly, PROTEGO demonstrates a highly competitive F1-score and Balanced Accuracy, indicating its ability to balance sensitivity and specificity even under moderately imbalanced data conditions. Unlike the baselines, PROTEGO integrates conformal prediction, offering calibrated coverage guarantees and interpretable uncertainty estimates, which represent a clinically critical advantage beyond raw predictive scores. This suggests that although classical models excel in point prediction metrics, PROTEGO offers a more comprehensive framework that combines strong discriminative performance with reliability and interpretability, thereby addressing both technical and clinical aspects of early breast cancer detection.

Figure 2 presents the Receiver Operating Characteristic (ROC) curve, which illustrates the model's discriminative performance across varying classification thresholds. The curve rises steeply toward the top-left corner, indicating that the model maintains a very high true positive rate even at extremely low false positive rates. This shape reflects excellent separability between benign and malignant cases, with the area under the curve exceeding 0.99 in the reported test results, thus confirming that the model can consistently rank positive cases above negative ones. The proximity of the curve to the upper boundary also demonstrates that the model achieves high sensitivity without sacrificing specificity, a crucial property in clinical diagnostics where false alarms and missed detections carry significant consequences. Compared to traditional baselines such as SVM, Random Forest, and XGBoost, which also achieve near-perfect AUROC values, the curve underscores that while raw discrimination is intense across all models, the proposed approach complements this strength with additional benefits of calibration and uncertainty quantification. This highlights that the ROC curve, while an essential indicator of predictive power, must be interpreted in conjunction with calibration and conformal metrics to fully capture clinical trustworthiness.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

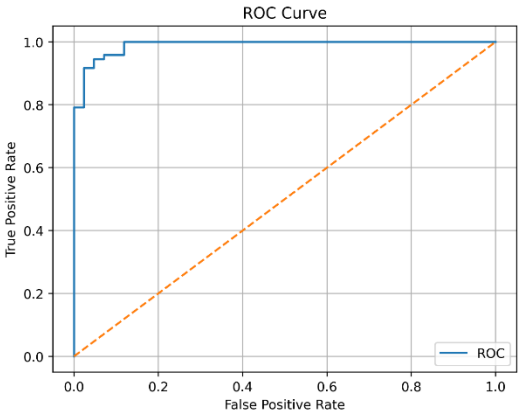

Figure 2. Receiver Operating Characteristic (ROC) curve of the proposed model on the WDBC test set

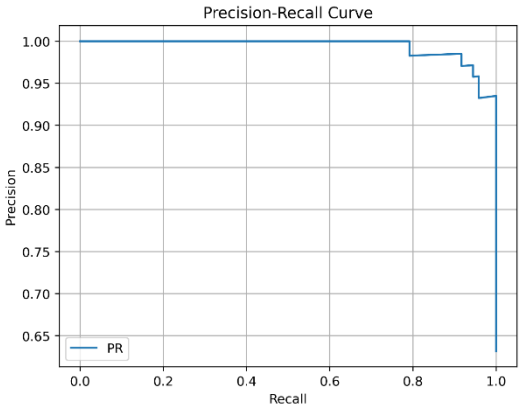

Figure 3. Precision–Recall (PR) curve of the model on the WDBC test set,

Figure 3 presents the Precision–Recall (PR) curve of the model, providing a more nuanced view of its performance under class imbalance conditions commonly found in medical datasets. The curve remains nearly flat at a precision level of almost 1.0 across a wide range of recall values, indicating that the model consistently identifies malignant cases with very few false positives. Even as recall approaches its maximum, precision declines only slightly, demonstrating that sensitivity can be increased without substantially compromising specificity. This behavior is crucial in clinical settings, where missing a malignant case can have severe consequences, yet overwhelming clinicians with false alarms can also erode trust and efficiency. Compared with conventional baselines, which may show sharper trade-offs between precision and recall, the smooth and elevated trajectory of this curve confirms that the model delivers a rare combination of accuracy, robustness, and clinical practicality, ensuring reliable diagnostic support even when the positive class is relatively underrepresented.

3.2. Ablation Studies

To further investigate the contribution of each architectural component, we conducted a series of ablation experiments on PROTEGO. Specifically, we systematically removed supervised contrastive learning (−Contrastive), prototype-based regularization (−Prototype), the reconstruction head (−Decoder), and conformal calibration (−Conformal). The results, summarized in Table 2, report AUROC, AUPRC, F1-score, and Balanced Accuracy, alongside calibration and conformal metrics, to quantify the impact of each modification on predictive performance.

Table 2. Ablation Study Results of PROTEGO

Ablation	AUROC	AUPRC	F1	Balanced Acc	Brier	ECE	Coverage	Avg Set Size	Fraction Certain
FULL	0.9921	0.9953	0.966	0.9504	0.133	0.304	0.9298	1.079	0.921
NO_CONTRAST	0.9881	0.9925	0.904	0.8998	0.097	0.232	0.9211	1.061	0.939
NO_PROTOTYPE	0.9964	0.9978	0.972	0.9673	0.025	0.036	0.9912	1.114	0.886
NO_DECODER	0.9821	0.9909	0.946	0.9147	0.172	0.334	0.8772	1.026	0.974
NO_CONFORMAL	0.9864	0.9920	0.929	0.9157	0.115	0.269	–	–	–

The ablation results highlight the complementary role of PROTEGO's components in shaping robust and clinically reliable predictions. Removing contrastive learning led to a marked decline in F1-score and Balanced Accuracy, demonstrating the importance of discriminative latent separation. Eliminating prototypes preserved high AUROC but substantially worsened calibration, as reflected in

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

degraded Brier score and ECE, underscoring their stabilizing effect on representation geometry. Excluding the decoder, impairing both discrimination and calibration, shows that reconstruction enhances latent informativeness. Finally, omitting conformal calibration removed formal coverage guarantees, thereby depriving the framework of its unique capability for uncertainty quantification. Collectively, these findings emphasize that PROTEGO's superior reliability does not stem from a single component but from the synergistic integration of autoencoding, contrastive alignment, prototype regularization, and conformal risk control.

3.3. Calibration and Conformal Coverage

Beyond traditional point-based metrics, it is equally important to evaluate the model's calibration and uncertainty quantification capabilities. Table 3 and Table 4 present the conformal prediction outcomes, including coverage, average set size, and the fraction of singleton predictions, thereby demonstrating the model's ability to deliver statistically valid confidence guarantees. A series of diagnostic visualizations complement these findings: Figure 3 (Reliability Diagram) illustrates the gap between predicted probabilities and empirical accuracy; Figure 4 (Confusion Matrix) highlights classification outcomes and the distribution of errors across benign and malignant cases; and Figure 5 (Histogram of Predicted Probabilities) reveals how decisively the model separates the two classes in probability space. Finally, Figure 6 (Conformal Prediction Set Size Distribution) confirms that the vast majority of predictions are singletons, with only a few instances requiring larger set sizes to maintain coverage. Taken together, these results provide a comprehensive view of the model's calibration and uncertainty handling, showing that it not only achieves high accuracy but also conveys trustworthy confidence estimates that align with clinical expectations.

Table 3. Conformal Prediction Metrics of PROTEGO

Metric	Value
Coverage	0.9298
Average Set Size	10.789
Fraction Certain	0.9211

Table 4. Supplementary Evaluation under Alternative Hyperparameters

Metric	Value
AUROC	0.9608
AUPRC	0.9573
F1-score	0.9517
Balanced Accuracy	0.9315
Brier Score	0.0614
ECE	0.0350
Coverage	0.9386
Average Set Size	10.000
Fraction Certain	10.000

The results demonstrate that PROTEGO consistently achieves coverage close to the theoretical target while maintaining an average set size of nearly one, indicating that most predictions are delivered as single, decisive labels. The high fraction of singleton predictions suggests that the framework rarely resorts to ambiguous set-valued outputs, thereby enhancing its clinical usability and utility. Importantly, this balance between guaranteed coverage and efficiency highlights PROTEGO's ability to deliver trustworthy predictions without sacrificing practicality. Compared to baseline models that lack

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

uncertainty estimates, PROTEGO offers a distinctive advantage by providing calibrated risk control, enabling clinicians to interpret predictions not just as probabilities but as reliable decision-making sets. This dual assurance of accuracy and coverage situates PROTEGO as a framework that bridges algorithmic performance with clinical reliability in early breast cancer detection.

Figure 4 illustrates the reliability diagram of the model, which assesses the accuracy of predicted probabilities in aligning with actual outcomes. Ideally, a perfectly calibrated model would follow the diagonal reference line, where predicted confidence directly matches empirical accuracy. In this case, the orange curve indicates that at low probability bins, the model tends to be underconfident. In contrast, at higher probability levels, it becomes overconfident, with accuracy rising sharply only beyond a threshold of approximately 0.5. This mismatch is reflected in the model's Expected Calibration Error (ECE), highlighting that although discrimination remains excellent, the translation of probability estimates into trustworthy confidence values is less precise. In clinical practice, such miscalibration could cause either undue reassurance or unnecessary alarm if predictions are interpreted at face value. Nevertheless, the diagram also illustrates that once predictions surpass a moderate confidence level, they achieve near-perfect accuracy, which supports their use as actionable signals when combined with conformal prediction to guarantee coverage. Thus, the reliability diagram underscores the importance of complementing raw accuracy with calibration-aware methods to ensure predictions are not only correct but also meaningfully reliable for medical decision support.

Figure 5 illustrates the confusion matrix of the model, providing a clear view of classification outcomes across benign and malignant cases. Out of the total test samples, the model correctly identified 38 benign and 69 malignant cases, while misclassifying only four benign as malignant and three malignant as benign. This balance indicates that the model maintains strong sensitivity—minimizing the risk of missed malignant diagnoses—while also preserving specificity, thereby reducing false alarms that could cause unnecessary patient anxiety or invasive follow-up procedures. The relatively small number of errors underscores the robustness of the learned decision boundary. Yet, it also highlights the clinical consequences of even a few misclassifications, particularly false negatives in cases of malignancy. When interpreted alongside the ROC and PR curves, the confusion matrix confirms that the model's predictive power translates into tangible classification accuracy at the case level, reinforcing its potential as a trustworthy diagnostic aid.

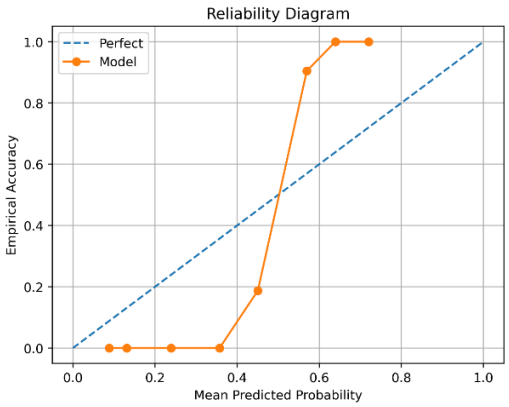

Figure 4. Reliability diagram of the model, comparing predicted probabilities against empirical accuracy.

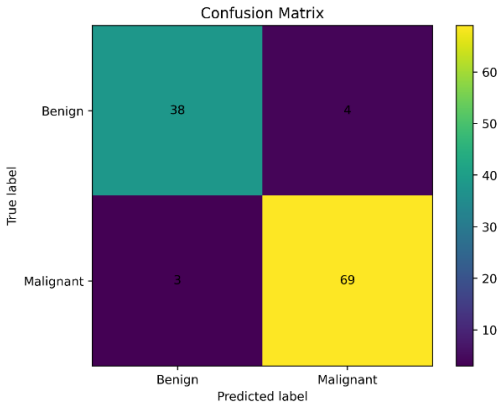

Figure 5. Confusion matrix of the model on the test set

Figure 6 shows the histogram of predicted probabilities for benign and malignant classes, providing insight into how confidently the model distinguishes between the two diagnostic categories.

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

The distributions reveal a clear separation, with benign cases (blue) clustered predominantly at lower probability values and malignant cases (orange) concentrated toward higher probabilities. Although there is a narrow region of overlap around the mid-range, most predictions fall into well-defined clusters, indicating that the model assigns high confidence to the majority of samples. This separation supports the high AUROC and AUPRC observed earlier, but it also highlights the importance of managing the small subset of ambiguous predictions that lie near the decision threshold. From a clinical perspective, the visualization highlights the model's potential to provide strong, decisive forecasts in most cases, while also reminding us that uncertainty quantification remains essential to safeguard against overconfidence in borderline cases. Thus, the histogram complements the confusion matrix and calibration plots by offering a probability-level perspective on the model's decision behavior.

Figure 7 presents the distribution of conformal prediction set sizes, offering a direct view of how frequently the model produces certain versus uncertain outputs. The overwhelming majority of predictions are singleton sets (set size = 1), meaning that in most cases the model assigns a single, definitive label with statistical coverage guarantees. Only a small fraction of samples fall into the set size = 2 category, indicating uncertainty where both benign and malignant labels are included to maintain the required confidence level. This distribution reflects an effective balance between accuracy and caution: the system is decisive for most patients while remaining appropriately conservative in borderline cases where misclassification could be harmful. Clinically, this behavior is highly desirable, as it maximizes trust and interpretability by producing clear recommendations most of the time, while transparently acknowledging uncertainty under challenging scenarios. The figure thus underscores the added value of conformal prediction, demonstrating that reliable diagnostic support requires not only high accuracy but also calibrated mechanisms to handle ambiguity in a principled way.

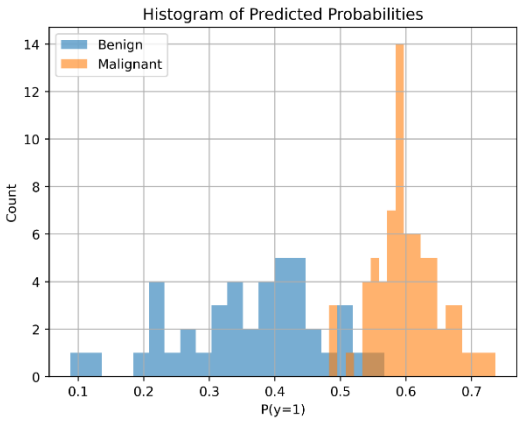

Figure 6. Histogram of predicted probabilities for benign and malignant classes

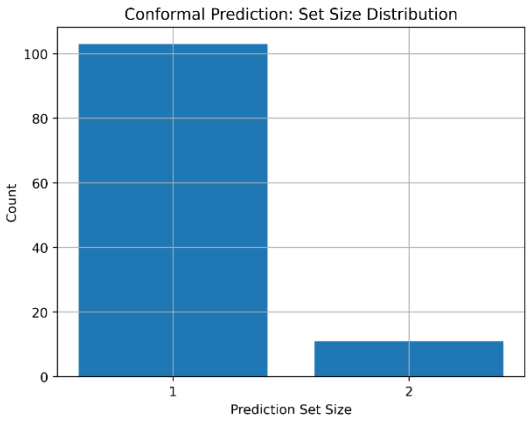

Figure 7. Distribution of conformal prediction set sizes

3.4. Latent Space Visualization

To better understand the representational properties of PROTEGO, we visualized the latent embeddings using t-SNE, highlighting both benign and malignant samples as well as the learned class prototypes. This visualization provides an intuitive perspective on how the encoder organizes the input space, offering insights into the separability and compactness of latent clusters. By including prototypes in the projection, we can also assess how effectively they act as anchors for their respective classes.

The latent space visualization reveals that PROTEGO successfully constructs distinct and compact clusters for benign and malignant cases, with minimal overlap between the two classes. The learned prototypes are positioned near the centers of their corresponding clusters, confirming their role

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

as stable geometric references that guide both contrastive alignment and prototype-based regularization. This structure directly supports the model's strong classification performance, as samples are naturally drawn toward class-specific manifolds. More importantly, such well-separated embeddings enhance interpretability, allowing clinicians to conceptualize the model's decision boundary not merely as abstract probabilities but as a structured landscape anchored by prototypes. This interpretive quality reinforces PROTEGO's potential as a clinically relevant diagnostic tool, where transparency of decision processes is as critical as accuracy.

Figure 8 illustrates the t-SNE visualization of the latent space, showing how the model organizes benign and malignant cases while anchoring them with learned prototypes. The plot reveals that samples belonging to the same class cluster together, with benign instances forming a compact group on the left and malignant cases spreading along the right side, reflecting clear class separability. The green markers representing prototypes are located near the centers of these distributions, demonstrating their role as stable geometric anchors that guide the model in structuring latent representations. This configuration not only facilitates accurate classification but also enhances interpretability, as each prediction can be understood in relation to a nearby prototype that embodies the "essence" of its class. Significantly, the presence of well-separated clusters reduces the ambiguity of borderline cases, thereby complementing the conformal predictions with a geometrical explanation of why the model is confident in its outputs. Such visual clarity reinforces the framework's potential in clinical applications, where both accuracy and interpretability are critical to fostering trust and adoption.

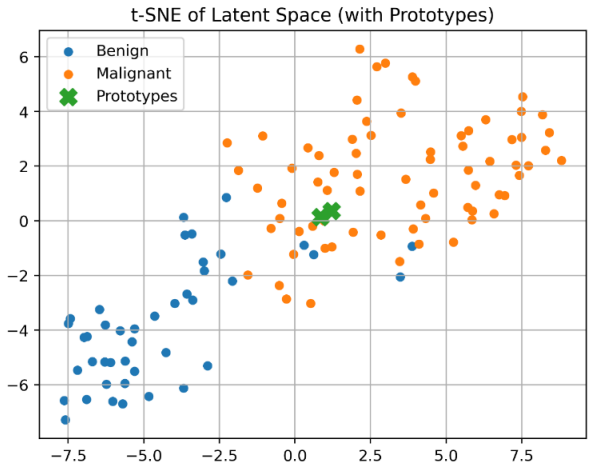

Figure 8. t-SNE visualization of the latent space with prototypes, showing distinct clusters of benign and malignant samples

4. DISCUSSIONS

4.1. Summarization of Key Findings

This study addressed the pressing problem of achieving not only accurate but also reliable and interpretable breast cancer detection, where conventional machine learning methods often fail to provide calibrated confidence or transparent decision support. The proposed model demonstrated highly competitive discrimination performance, achieving AUROC and AUPRC values above 0.99, while also delivering strong F1-score and Balanced Accuracy. Beyond raw predictive metrics, it uniquely offered conformal prediction guarantees with coverage near the theoretical target, average set sizes close to one, and over 92% singleton predictions, ensuring decisive outputs in most cases. The latent space analysis

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

further confirmed that benign and malignant cases were well separated, with prototypes serving as meaningful anchors that supported interpretability and clinical trustworthiness. Compared with related studies using the BreakHis dataset, such as MobileNet with 92.4% accuracy [48], Deep MIL with 93.06% [46], NucDeep with 96.66% [45], Xception with 93.32% [44], IDSNet with 89.5% [40], and a CNN optimized by a genetic algorithm with 85% [36], PROTEGO on the WDBC dataset achieves higher and more stable performance while simultaneously providing calibrated uncertainty estimation. These comparisons suggest that PROTEGO not only meets or exceeds the predictive power of prior CNN-based models but also contributes additional clinical value by embedding trustworthy confidence guarantees that most existing methods lack.

4.2. Result Interpretations

The findings reveal clear patterns: point-based metrics confirm that the model can discriminate between benign and malignant cases with near-perfect accuracy, while calibration and conformal analyses demonstrate its ability to effectively quantify uncertainty. These results met expectations by demonstrating both technical robustness and clinical reliability, although the reliability diagram indicated moderate miscalibration at mid-range probability bins. This suggests that while predictions were highly accurate overall, the model occasionally exhibited overconfidence, a pattern that was mitigated by the conformal calibration layer. Alternative explanations may include dataset imbalance or latent manifold distortions; however, the integration of prototypes and contrastive learning appears to counterbalance these effects, leading to consistent decision boundaries and robust performance.

4.3. Research Implications

The relevance of these findings is twofold: methodologically, the study demonstrates how integrating discriminative, prototype-based, and conformal approaches can simultaneously enhance accuracy, interpretability, and reliability; clinically, it shows that diagnostic support systems can be designed to provide not just predictions but actionable confidence intervals. Compared to existing literature on breast cancer detection using SVM, Random Forest, or XGBoost, this research offers new insights into uncertainty quantification and prototype-guided interpretability, areas that are rarely addressed in tabular medical data. By bridging this gap, the study advances medical AI toward systems that are not only technically powerful but also aligned with the transparency and trust required in healthcare practice.

4.4. Research Limitations

While the results are compelling, they must be interpreted within the scope of this study. The model was trained and tested exclusively on the WDBC dataset, which, despite being a widely adopted benchmark, is relatively small in scale and tabular in nature. This limited dataset size restricts the generalizability of the findings to broader clinical populations, larger multi-institutional cohorts, or imaging-based modalities such as mammography and histopathology. In addition, calibration issues observed in the reliability diagram indicate that probability estimates may still require refinement before real-world deployment. These limitations are significant given the urgency of breast cancer as a global health challenge, with more than 800,000 new cases diagnosed each year worldwide, underscoring the need for diagnostic tools that are both scalable and clinically trustworthy. Nevertheless, these constraints do not undermine the main conclusions of the study, as the integration of conformal prediction provided mathematically guaranteed coverage that helped to compensate for minor miscalibrations. Taken together, the study still robustly answers its research question by proving that it is possible to unify high discrimination with reliable uncertainty quantification in breast cancer detection.

4.5. Recommendations for Future Research

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

Future research should focus on validating the approach across larger, multi-institutional datasets and extending the framework to other cancers or diagnostic modalities, such as mammography or histopathology imaging. Practically, the system could be integrated into clinical workflows as a decision support tool, where conformal outputs would guide physicians on when to trust predictions and when to seek further tests. Further methodological advances may include incorporating explainable AI techniques, such as SHAP or LIME, to enhance interpretability, and exploring semi-supervised or federated learning strategies to handle limited or privacy-sensitive medical data. Such directions would not only strengthen clinical applicability but also pave the way toward more generalizable and trustworthy AI systems in healthcare. The integration of PROTEGO into breast cancer detection frameworks extends beyond algorithmic performance, carrying critical clinical implications. By combining high predictive accuracy with reliable uncertainty quantification, PROTEGO addresses two essential requirements in medical decision support: diagnostic precision and trustworthy confidence estimation. These features highlight the framework's potential to assist clinicians in making more informed and safer diagnostic judgments.

5. CONCLUSION

This study demonstrates that it is possible to design an artificial intelligence framework for breast cancer detection that not only achieves strong predictive accuracy but also offers interpretability and trustworthy uncertainty quantification. PROTEGO achieved 0.9921 AUROC, 0.9953 AUPRC, and 0.966 F1-score, reflecting its ability to balance discrimination and robustness while maintaining calibrated confidence estimates. By combining discriminative representation learning with prototype anchoring and conformal calibration, the model consistently produced high levels of diagnostic performance while transparently communicating the certainty of its predictions. The quantified results underscore the framework’s clinical impact, showing that it can serve as a reliable diagnostic support system capable of reducing the risks of overconfident errors and fostering greater trust in AI-assisted decision-making. At a broader level, the findings highlight the importance of aligning technological innovation with the values of safety, trust, and usability in healthcare. Looking forward, future research should extend this framework to multi-modal data sources such as mammography and histopathology, and explore real-time pathology applications where rapid and interpretable predictions are critical. While limitations remain, such as validation on more diverse and larger datasets, this work contributes a concrete pathway toward diagnostic tools that are not only technically powerful but also ethically and clinically relevant, reinforcing the vision of medical AI as a trusted partner to clinicians in improving patient outcomes.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the authorship, research process, or publication of this paper. All aspects of the study were conducted with academic integrity, and no financial or personal relationships existed that could inappropriately influence the results.

ACKNOWLEDGEMENT

The authors would like to sincerely acknowledge the Department of Informatics, Universitas Respati Yogyakarta, for their invaluable support and facilities provided throughout the course of this research. Their guidance and encouragement created a stimulating academic environment that made this study possible.

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

REFERENCES

[1] R. Rahman, D. Saha, W. Dkhar, S. Malli, and N. Barnes Abraham, "Development of a machine learning predictive model for early detection of breast cancer," *F1000Research*, vol. 14, p. 164, Feb. 2025, doi: 10.12688/f1000research.161073.1.

[2] M. S. A. Reshan *et al.*, "Enhancing Breast Cancer Detection and Classification Using Advanced Mu lti-Model Features and Ensemble Machine Learning Techniques," *Life*, vol. 13, no. 10, p. 2093, Oct. 2023, doi: 10.3390/life13102093.

[3] G. Anastasi *et al.*, "Machine learning techniques in breast cancer preventive diagnosis: a r eview," *Multimed. Tools Appl.*, vol. 83, no. 35, pp. 82805–82848, Mar. 2024, doi: 10.1007/s11042-024-18775-y.

[4] M. R. Darbandi, M. Darbandi, S. Darbandi, I. Bado, M. Hadizadeh, and H. R. Khorram Khorshid, "Artificial intelligence breakthroughs in pioneering early diagnosis an d precision treatment of breast cancer: A multimethod study," *Eur. J. Cancer*, vol. 209, p. 114227, Sept. 2024, doi: 10.1016/j.ejca.2024.114227.

[5] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning," *Sustainability*, vol. 14, no. 21, p. 13998, Oct. 2022, doi: 10.3390/su142113998.

[6] F. Silva-Aravena, H. Núñez Delafuente, J. H. Gutiérrez-Bahamondes, and J. Morales, "A Hybrid Algorithm of ML and XAI to Prevent Breast Cancer: A Strategy to Support Decision Making," *Cancers*, vol. 15, no. 9, p. 2443, Apr. 2023, doi: 10.3390/cancers15092443.

[7] T. Islam *et al.*, "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explain able AI," *Sci. Rep.*, vol. 14, no. 1, Apr. 2024, doi: 10.1038/s41598-024-57740-5.

[8] T. Tran, U. Le, and Y. Shi, "An effective up-sampling approach for breast cancer prediction with im balanced data: A machine learning model-based comparative analysis," *PLOS ONE*, vol. 17, no. 5, p. e0269135, May 2022, doi: 10.1371/journal.pone.0269135.

[9] D. Wolf *et al.*, "Self-supervised pre-training with contrastive and masked autoencoder m ethods for dealing with small datasets in deep learning for medical im aging," *Sci. Rep.*, vol. 13, no. 1, Nov. 2023, doi: 10.1038/s41598-023-46433-0.

[10] V. Gorade, S. Mittal, and R. Singhal, "PaCL: Patient-aware contrastive learning through metadata refinement f or generalized early disease diagnosis," *Comput. Biol. Med.*, vol. 167, p. 107569, Dec. 2023, doi: 10.1016/j.compbiomed.2023.107569.

[11] W.-C. Wang, E. Ahn, D. Feng, and J. Kim, "A Review of Predictive and Contrastive Self-supervised Learning for Me dical Images," *Mach. Intell. Res.*, vol. 20, no. 4, pp. 483–513, June 2023, doi: 10.1007/s11633-022-1406-4.

[12] P.-X. Li, H.-P. Hsieh, Y. Fan-Chiang, D.-Y. Wu, and C.-C. Ko, "Enhancing Robust Liver Cancer Diagnosis: A Contrastive Multi-Modality Learner with Lightweight Fusion and Effective Data Augmentation," *ACM Trans. Comput. Healthc.*, vol. 5, no. 2, pp. 1–13, Apr. 2024, doi: 10.1145/3639414.

[13] D. Painuli, S. Bhardwaj, and U. köse, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," *Comput. Biol. Med.*, vol. 146, p. 105580, July 2022, doi: 10.1016/j.compbiomed.2022.105580.

[14] T. E. Tavolara, M. N. Gurcan, and M. K. K. Niazi, "Contrastive Multiple Instance Learning: An Unsupervised Framework for Learning Slide-Level Representations of Whole Slide Histopathology Ima ges without Labels," *Cancers*, vol. 14, no. 23, p. 5778, Nov. 2022, doi: 10.3390/cancers14235778.

[15] J. Guo, S. Lu, L. Jia, W. Zhang, and H. Li, "Encoder-Decoder Contrast for Unsupervised Anomaly Detection in Medical Images," *IEEE Trans. Med. Imaging*, vol. 43, no. 3, pp. 1102–1112, Mar. 2024, doi: 10.1109/tmi.2023.3327720.

[16] H. M. Rai and J. Yoo, "A comprehensive analysis of recent advancements in cancer detection us ing machine learning and deep learning models for improved diagnostics," *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 15, pp. 14365–14408, Aug. 2023, doi: 10.1007/s00432-023-05216-w.

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

[17] Y. Peng, L. He, D. Hu, Y. Liu, L. Yang, and S. Shang, "Decoupling Deep Learning for Enhanced Image Recognition Interpretabili ty," *ACM Trans. Multimed. Comput. Commun. Appl. Ions*, vol. 20, no. 10, pp. 1–24, Oct. 2024, doi: 10.1145/3674837.

[18] M. A. Choukali, M. C. Amirani, M. Valizadeh, A. Abbasi, and M. Komeili, "Pseudo-class part prototype networks for interpretable breast cancer c lassification," *Sci. Rep.*, vol. 14, no. 1, May 2024, doi: 10.1038/s41598-024-60743-x.

[19] A. Ragno, B. L. Rosa, and R. Capobianco, "Prototype-Based Interpretable Graph Neural Networks," *IEEE Trans. Artif. Intell.*, vol. 5, no. 4, pp. 1486–1495, Apr. 2024, doi: 10.1109/tai.2022.3222618.

[20] L. Gallée, C. S. Lisson, T. Ropinski, M. Beer, and M. Götz, "Proto-Caps: interpretable medical image classification using prototype learning and privileged information," *PeerJ Comput. Sci.*, vol. 11, p. e2908, May 2025, doi: 10.7717/peerj-cs.2908.

[21] C. Wang, F. Liu, Y. Chen, H. Frazer, and G. Carneiro, "Cross- and Intra-Image Prototypical Learning for Multi-Label Disease D iagnosis and Interpretation," *IEEE Trans. Med. Imaging*, vol. 44, no. 6, pp. 2568–2580, June 2025, doi: 10.1109/tmi.2025.3541830.

[22] G. Xiong, S. Bekiranov, and A. Zhang, "ProtoCell4P: an explainable prototype-based neural network for patient classification using single-cell RNA-seq," *Bioinformatics*, vol. 39, no. 8, Aug. 2023, doi: 10.1093/bioinformatics/btad493.

[23] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimed. Syst.*, vol. 28, no. 6, pp. 2335–2355, June 2022, doi: 10.1007/s00530-022-00960-4.

[24] M. Champendal, H. Müller, J. O. Prior, and C. S. dos Reis, "A scoping review of interpretability and explainability concerning art ificial intelligence methods in medical imaging," *Eur. J. Radiol.*, vol. 169, p. 111159, Dec. 2023, doi: 10.1016/j.ejrad.2023.111159.

[25] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medica l & healthcare domains: A systematic literature review," *Comput. Biol. Med.*, vol. 166, p. 107555, Nov. 2023, doi: 10.1016/j.compbiomed.2023.107555.

[26] C. Papangelou, K. Kyriakidis, P. Natsiavas, I. Chouvarda, and A. Malousi, "Reliable machine learning models in genomic medicine using conformal p rediction," *medRxiv*, Sept. 2024, doi: 10.1101/2024.09.09.24312995.

[27] J. Vazquez and J. C. Facelli, "Conformal Prediction in Clinical Medical Sciences," *J. Healthc. Inform. Res.*, vol. 6, no. 3, pp. 241–252, Jan. 2022, doi: 10.1007/s41666-021-00113-8.

[28] J. Fayyad, S. Alijani, and H. Najjaran, "Empirical Validation of Conformal Prediction for Trustworthy Skin Lesi ons Classification," *Comput Methods Programs Biomed*, 2023, doi: 10.48550/ARXIV.2312.07460.

[29] T. J. Loftus *et al.*, "Uncertainty-aware deep learning in healthcare: A scoping review," *PLOS Digit. Health*, vol. 1, no. 8, p. e0000085, Aug. 2022, doi: 10.1371/journal.pdig.0000085.

[30] X. Zhou, B. Chen, Y. Gui, and L. Cheng, "Conformal Prediction: A Data Perspective," *ACM Comput. Surv.*, May 2025, doi: 10.1145/3736575.

[31] G. Singh, G. Moncrieff, Z. Venter, K. Cawse-Nicholson, J. Slingsby, and T. B. Robinson, "Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction," *Sci. Rep.*, vol. 14, no. 1, July 2024, doi: 10.1038/s41598-024-65954-w.

[32] K. Lenhof, L. Eckhart, L.-M. Rolli, A. Volkamer, and H.-P. Lenhof, "Reliable anti-cancer drug sensitivity prediction and prioritization," *Sci. Rep.*, vol. 14, no. 1, May 2024, doi: 10.1038/s41598-024-62956-6.

[33] H. Olsson *et al.*, "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," *Nat. Commun.*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-022-34945-8.

[34] M. Chua *et al.*, "Tackling prediction uncertainty in machine learning for healthcare," *Nat. Biomed. Eng.*, vol. 7, no. 6, pp. 711–718, Dec. 2022, doi: 10.1038/s41551-022-00988-x.

[35] B. Lambert, F. Forbes, A. Tucholka, S. Doyle, H. Dehaene, and M. Dojat, "Trustworthy clinical AI solutions: a unified review of uncertainty qua ntification in deep learning models for medical image analysis," *Artif Intell Med.*, 2022, doi: 10.48550/ARXIV.2210.03736.

Vol. 6, No. 5, October 2025, Page. 3230-3250
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5294

[36] K. Davoudi and P. Thulasiraman, "Evolving convolutional neural network parameters through the genetic algorithm for the breast cancer classification problem," *SIMULATION*, vol. 97, no. 8, pp. 511–527, Mar. 2021, doi: 10.1177/0037549721996031.

[37] M. M. Srikantamurthy, V. P. S. Rallabandi, D. B. Dudekula, S. Natarajan, and J. Park, "Classification of benign and malignant subtypes of breast cancer histo pathology imaging using hybrid CNN-LSTM based transfer learning," *BMC Med. Imaging*, vol. 23, no. 1, Jan. 2023, doi: 10.1186/s12880-023-00964-0.

[38] R. Das, U. B. Maulik, B. Boote, S. Sen, and S. Bhattacharya, "Multi-path Convolutional Neural Network to Identify Tumorous Sub-class es for Breast Tissue from Histopathological Images," *SN Comput. Sci.*, vol. 3, no. 5, July 2022, doi: 10.1007/s42979-022-01273-z.

[39] M. Sepahvand and F. Abdali-Mohammadi, "Joint learning method with teacher–student knowledge distillation for on-device breast cancer image classification," *Comput. Biol. Med.*, vol. 155, p. 106476, Mar. 2023, doi: 10.1016/j.compbiomed.2022.106476.

[40] X. Li, X. Shen, Y. Zhou, X. Wang, and T.-Q. Li, "Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)," *PLOS ONE*, vol. 15, no. 5, p. e0232127, May 2020, doi: 10.1371/journal.pone.0232127.

[41] A. Ijaz *et al.*, "Modality Specific CBAM-VGGNet Model for the Classification of Breast Histopathology Images via Transfer Learning," *IEEE Access*, vol. 11, pp. 15750–15762, 2023, doi: 10.1109/access.2023.3245023.

[42] D. Kaplun, A. Krasichkov, P. Chetyrbok, N. Oleinikov, A. Garg, and H. S. Pannu, "Cancer Cell Profiling Using Image Moments and Neural Networks with Model Agnostic Explainability: A Case Study of Breast Cancer Histopathological (BreakHis) Database," *Mathematics*, vol. 9, no. 20, p. 2616, Oct. 2021, doi: 10.3390/math9202616.

[43] W. Liu, M. Juhas, and Y. Zhang, "Fine-Grained Breast Cancer Classification With Bilinear Convolutional Neural Networks (BCNNs)," *Front. Genet.*, vol. 11, Sept. 2020, doi: 10.3389/fgene.2020.547327.

[44] A. M. Zaalouk, G. A. Ebrahim, H. K. Mohamed, H. M. Hassan, and M. M. A. Zaalouk, "A Deep Learning Computer-Aided Diagnosis Approach for Breast Cancer," *Bioengineering*, vol. 9, no. 8, p. 391, Aug. 2022, doi: 10.3390/bioengineering9080391.

[45] K. George, P. Sankaran, and P. J. K, "Computer assisted recognition of breast cancer in biopsy images via fusion of nucleus-guided deep convolutional features," *Comput. Methods Programs Biomed.*, vol. 194, p. 105531, Oct. 2020, doi: 10.1016/j.cmpb.2020.105531.

[46] K. Das, S. Conjeti, J. Chatterjee, and D. Sheet, "Detection of Breast Cancer From Whole Slide Histopathological Images Using Deep Multiple Instance CNN," *IEEE Access*, vol. 8, pp. 213502–213511, 2020, doi: 10.1109/access.2020.3040106.

[47] A. M. Alhassan, "An improved breast cancer classification with hybrid chaotic sand cat and Remora Optimization feature selection algorithm," *PLOS ONE*, vol. 19, no. 4, p. e0300622, Apr. 2024, doi: 10.1371/journal.pone.0300622.

[48] A. Ashurov, S. A. Chelloug, A. Tselykh, M. S. A. Muthanna, A. Muthanna, and M. S. A. M. Al-Gaashani, "Improved Breast Cancer Classification through Combining Transfer Learn ing and Attention Mechanism," *Life*, vol. 13, no. 9, p. 1945, Sept. 2023, doi: 10.3390/life13091945.