DEGRE: Dynamic Gating Ensembles for Trust-Aware Rejection in Medical Image Diagnostics

Hai Nguyen Hong[*1,2], **Duong Bach**[*2,3], **Nam Phan**[1], **Cuong V. Nguyen**[2], **Cuong Do**[2,4,5†]

[1]FPT University, Ha Noi, Viet Nam
[2]Smart Green Transformation Center (Green-X), VinUniversity, Ha Noi, Viet Nam
[3]Sun Asterisk, Ha Noi, Viet Nam
[4]College of Engineering & Computer Science, VinUniversity, Ha Noi, Viet Nam
[5]Smart Health Center (VISHC), VinUniversity, Ha Noi, Viet Nam
hainhhe171927@fpt.edu.vn, duong.xuan.bach@sun-asterisk.com, namphhe171518@fpt.edu.vn, cuong.nv@vinuni.edu.vn, cuong.dd@vinuni.edu.vn

Abstract

For artificial intelligence to be safely deployed in high-risk domains, it must reliably know its limits. Selective prediction, or learning with a reject option, addresses this by enabling a model to abstain from prediction on inputs it deems unreliable, deferring them to a human expert. While deep ensembles have emerged as a leading approach for uncertainty estimation, their potential is often squandered by rejection methods that rely on static thresholds applied to the mean prediction. In this paper, we propose to learn a dynamic rejection policy directly from the rich behavioral signals of the ensemble itself. Our framework, DEGRE (Dynamic Ensembles Gating for REjection), is a novel meta-learning approach that trains a lightweight gating network on the ensemble's consensus confidence and its internal disagreement (variance)- to explicitly discriminate between correct and incorrect predictions. Through rigorous evaluation across twelve diverse medical imaging benchmarks (MRI, X-ray, CT), DEGRE significantly advances selective prediction, achieving an average risk-coverage (AURC) reduction of 68.2% compared to the standard ensemble baseline. By providing a more reliable method for a model to recognize its own limitations, this learned, adaptive rejection mechanism paves the way for safer and more responsible integration of AI into critical clinical workflows.

Introduction

While deep learning models have achieved expert-level performance in medical imaging, their clinical adoption is hindered by a critical flaw: a tendency to produce overconfident yet incorrect predictions (Guo et al. 2017). This gap between predictive power and reliable self-assessment is a primary barrier to trust, as an erroneous prediction can have severe clinical consequences (Bondi et al. 2022). For AI to be a trustworthy partner, it must possess the ability to "know when it does not know" (Joy et al. 2023).

Selective prediction, or learning with a reject option, directly addresses this by allowing a model to abstain on unreliable inputs, deferring them to a human expert. This enables a safer AI-in-the-Loop (AI²L) paradigm where the human expert remains in control (Natarajan et al. 2025).

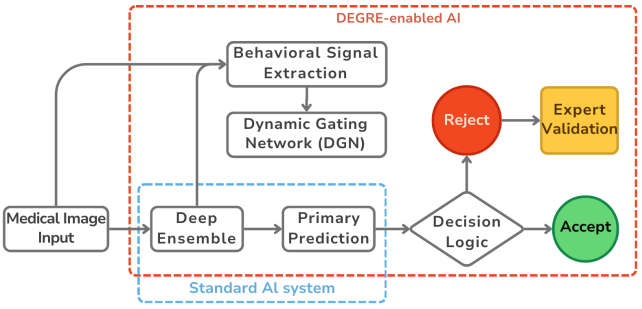

Figure 1: Conceptual overview of the DEGRE framework.

Deep ensembles are a leading technique for improving reliability, boosting both accuracy and providing a robust measure of uncertainty through model disagreement (Lakshminarayanan, Pritzel, and Blundell 2017). However, the standard rejection method—applying a simple confidence threshold—is fundamentally limited. It discards the rich information in the ensemble's disagreement and can fail to detect cases where all models are confidently wrong (Chidambaram and Ge 2024).

To overcome this limitation, this work proposes a shift from a static, heuristic-based rejection to a dynamic, learned rejection policy. We introduce DEGRE (Dynamic Ensemble Gating for Rejection), a novel framework centered around a Dynamic Gating Network (DGN). The DGN is a lightweight meta-model that learns a sophisticated rejection function by analyzing the behavior of a primary prediction ensemble. Instead of relying on a single, manually-tuned threshold on predictive confidence, the DGN takes as input a feature vector capturing both the ensemble's consensus confidence and its internal disagreement. It is then explicitly trained to solve a binary classification task: discriminating between samples the ensemble is likely to classify correctly versus incorrectly. This formulation allows DEGRE to learn a flexible, non-linear decision boundary in the confidence-

[*]These authors contributed equally.
[†]Corresponding author.

disagreement space, enabling a more effective rejection policy.

Key Contributions

1. **A Novel Meta-Learning Framework for Selective Prediction**: We introduce a Dynamic Gating Network (DGN), which learns a data-driven, non-linear rejection policy. By framing rejection as a supervised task on meta-features of ensemble behavior (confidence and disagreement), DEGRE consistently and significantly outperforms static thresholding and other methods, establishing a new benchmark for risk-coverage performance.
2. **Advanced Empirical Results on Diverse Medical Image Benchmarks**: We present a rigorous evaluation across twelve medical imaging datasets spanning CT, X-Ray, and MRI. Our results demonstrate that DEGRE consistently and substantially outperforms all baselines. On average, DEGRE reduces the risk-coverage error by over 68% and, in key cases, improves performance by orders of magnitude, enabling a more robust and trustworthy benchmarking in medical AI.
3. **A Practical Framework for Goal-Driven Rejection**: We introduce a novel, multi-objective optimization method for setting the final rejection threshold. Instead of relying on a single metric, our approach allows clinicians or administrators to define their operational goals-such as target accuracy, desired rejection rate, and acceptable calibration error-and automatically finds the optimal threshold that best balances these competing, real-world constraints.

Related Work

Machine learning systems need reliable uncertainty estimation as their fundamental basis to build trustworthy AI systems which operate in healthcare high-stakes domains. BNNs model weight distributions but MCDO serves as a practical method to generate weights from an approximate posterior through dropout during inference (Gal and Ghahramani 2016). Deep Ensembles provide a scalable solution through independent model training with different initializations while prediction aggregation reveals uncertainty by measuring model disagreements which function as strong indicators of failure (Lakshminarayanan, Pritzel, and Blundell 2017). The empirical strength and simplicity of ensemble-based uncertainty for foundation models and graph neural networks have been demonstrated by recent advances (Rahaman and Thiery 2021; Huo et al. 2025).

Selective prediction represents a method for learning to refuse predictions which controls the trade-off between risk and coverage through the framework established by Chow in 1970 (Chow 1970). Deep learning methods usually use maximum softmax probability (MSP) confidence scores to reject predictions but these approaches are limited by the potential misalignment of the confidence signals (Guo et al. 2017). Adaptive prediction methods in LLMs and segmentation tasks along with one-sided prediction and contextual rejection demonstrate improved approaches yet they face limitations from using static confidence metrics (Joy et al. 2023; Li et al. 2023).

The post-hoc reliability methods consist of calibration techniques which include Temperature Scaling and Isotonic Regression and Beta Calibration to match confidence scores with true correctness probabilities (Guo et al. 2017; Kull, Silva Filho, and Flach 2017). The OOD detection methods such as ODIN and energy-based scores detect distribution departures but these methods fail to handle the ambiguities which occur within clinical data (Liang, Li, and Srikant 2018; Liu et al. 2020). The integration of rejection mechanisms remains absent from recent advancements which include dynamic covariance scaling and feature clipping (Berta, Bach, and Jordan 2024; Chidambaram and Ge 2024).

The reliability of models improves through meta-learning because it teaches models to learn from base learner behaviors. The bilevel optimization method of ReVaR reweights training instances to minimize variance which enhances primary model robustness. Self-assessment frameworks such as‘ for ensemble stacking and few-shot adaptation in AIOps align with adaptive reliability goals (Natarajan et al. 2025; Li et al. 2023).

The combination of gating mechanisms with Mixture-of-Experts (MoE) architectures leads to modular AI systems which use gating networks to direct inputs toward specialized experts for efficient scaling. The performance advantages of MoE algorithms and their training strategies and dynamic routing approaches have been surveyed for large-scale models to develop reliable architectures (Mu and Lin 2025; Bondi et al. 2022).

The DEGRE Framework

This section details the mathematical foundations and architecture of the DEGRE framework. We begin by formally defining the problem of selective prediction in the context of deep ensembles. We then describe our two-phase meta-learning approach: (1) extracting reliability signals from a pre-trained ensemble, and (2) training a lightweight meta-learner, the Dynamic Gating Network (DGN), to predict misclassifications based on these signals. Finally, we introduce a novel goal-driven policy for setting the rejection threshold, enabling practical deployment in clinical settings.

Foundational Principles and Overview

The DEGRE framework enables selective prediction (Chow 1970), allowing a model to reject unreliable inputs and defer them to an expert—a critical capability for trustworthy AI in high-stakes domains (Bondi et al. 2022; Joy et al. 2023).

Our central hypothesis is that an ensemble's collective behavior provides a strong, learnable signal for predicting its own failure. While standard methods use a static threshold on a single metric (Guo et al. 2017; Chidambaram and Ge 2024). DEGRE employs a meta-learner to create a nuanced rejection policy from multiple reliability signals, more accurately discriminating correct from incorrect predictions (Lakshminarayanan, Pritzel, and Blundell 2017; Rahaman and Thiery 2021).

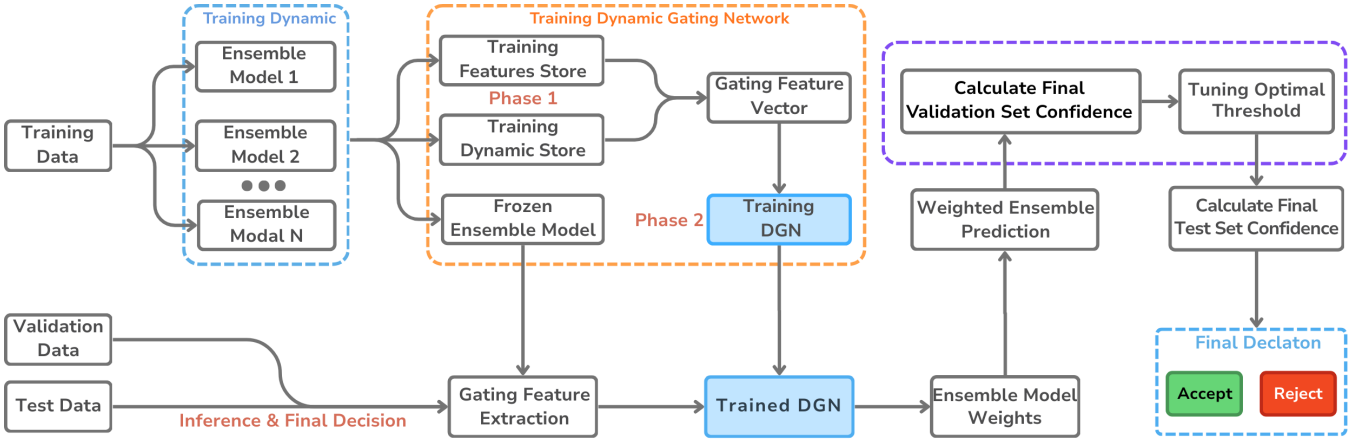

Figure 2: The detailed architecture and workflow of the DEGRE framework.

- **Phase 1: Reliability Signal Extraction.** Given a trained ensemble, we compute instance-wise meta-features for each input. These features are designed to capture two orthogonal aspects of the ensemble's predictive behavior: its consensus confidence and its internal disagreement (Lakshminarayanan, Pritzel, and Blundell 2017; Rahaman and Thiery 2021).
- **Phase 2: Meta-Learning for Rejection.** We train a lightweight meta-classifier, which we term the Dynamic Gating Network (DGN), on a supervised task. The inputs to the DGN are the extracted meta-features, and the target is a binary label indicating whether the base ensemble's prediction for that instance was correct or incorrect. The DGN's output is therefore a direct, learned score of unreliability.

At inference time, this unreliability score is compared against a rejection threshold, τ, to make the final accept-or-reject decision. This threshold is not fixed but is determined via a multi-objective optimization that balances user-defined clinical and operational goals (Joy et al. 2023; Bondi et al. 2022).

While the DGN is architecturally inspired by gating mechanisms found in Mixture-of-Experts (MoE) models (Mu and Lin 2025), its function is distinct. In MoE, a gating network typically routes an input to one of several specialized expert subnetworks for prediction. In contrast, the DGN in DEGRE does not route data; it acts as a meta-learning controller that governs the information workflow, deciding whether the AI's prediction is reliable enough to be used or if the case should be deferred to a human expert. It is a gate on the decision itself, not on the data processing path.

Phase 1: Base Ensemble and Reliability Signal Extraction

Problem Setup Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a dataset of medical images $x_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y} = \{1, \ldots, K\}$. We assume access to a deep ensemble $\mathcal{F} = \{f_{\theta_m}\}_{m=1}^{M}$ of M independently trained classifiers. Each model outputs class probabilities via softmax:

$$
\begin{aligned}
p_{\theta_m}(y \mid x) &= \text{Softmax}(\text{logits}_{\theta_m}(x)) \\
\bar{p}(y \mid x) &= \sum_{m=1}^{M} w_m \, p_{\theta_m}(y \mid x), \\
\text{where} \quad & \sum_{m=1}^{M} w_m = 1, \\
\hat{y} &= \arg\max_k \bar{p}(y = k \mid x)
\end{aligned}
\tag{1}
$$

This formulation allows the ensemble to assign higher importance to more reliable members, providing improved calibration and robustness over uniform averaging (Lakshminarayanan, Pritzel, and Blundell 2017).

Meta-Feature Extraction To summarize the ensemble's behavior, we define a 2D meta-feature vector $\phi(x) = [u_c(x), u_d(x)]$, capturing both consensus and disagreement—key indicators of reliability.

- **Consensus Confidence** (u_c): the maximum of the mean predictive distribution, indicating the ensemble's average confidence:

$$u_c(x) = \max_k \bar{p}(y = k \mid x) \tag{2}$$

- **Ensemble Disagreement** (u_d): the predictive entropy of $\bar{p}(y \mid x)$, measuring uncertainty due to conflicting predictions:

$$u_d(x) = -\sum_{k=1}^{K} \bar{p}(y = k \mid x) \log \bar{p}(y = k \mid x) \tag{3}$$

These two signals are fed into the Dynamic Gating Network (DGN) in Phase 2 to predict reliability.

Phase 2: The Dynamic Gating Network for Misclassification Prediction

The core innovation of DEGRE lies in casting the rejection decision as a supervised meta-learning task. A meta-learner, the Dynamic Gating Network (DGN), is trained to predict the probability that the base ensemble $\mathcal{F}$ misclassifies an input x, using reliability signals $\phi(x) = [u_c(x), u_d(x)]$.

Constructing the Meta-Dataset A meta-dataset $\mathcal{D}_{\text{meta}}$ is built from a held-out validation set $\mathcal{D}_{\text{val}}$, distinct from both the training and test sets. For each sample (x_i, y_i), we define a meta-instance $(\phi(x_i), z_i)$, where $z_i = I(\arg\max_k \bar{p}(y = k \mid x_i) \neq y_i)$ indicates whether the ensemble misclassifies x_i. This reframes the problem as a binary classification task to estimate the likelihood of an error (Lakshminarayanan, Pritzel, and Blundell 2017).

DGN Architecture and Training The DGN, $g_\psi : R^2 \to [0, 1]$, is a lightweight MLP optimized using the Binary Cross-Entropy (BCE) loss:

$$\begin{aligned}\mathcal{L}_{\text{BCE}}(\psi) = E_{(\phi(x),z)\sim\mathcal{D}_{\text{meta}}}\big[&- z \log g_\psi(\phi(x)) \\ &- (1 - z) \log(1 - g_\psi(\phi(x)))\big]\end{aligned} \tag{4}$$

To improve calibration, the Brier score is incorporated, yielding the total loss:

$$\begin{aligned}\mathcal{L}_{\text{DGN}}(\psi) = &\mathcal{L}_{\text{BCE}}(\psi) \\ &+ \lambda \cdot E_{(\phi(x),z)} \left[(g_\psi(\phi(x)) - z)^2\right]\end{aligned} \tag{5}$$

Inference and Goal-Driven Rejection Policy

Rejection Score Generation At test time, the ensemble $\mathcal{F}$ produces predictions for input x_{test}. The DGN computes the rejection score $s(x_{\text{test}}) = g_\psi(\phi(x_{\text{test}}))$, which estimates the probability of misclassification.

Decision Rule and Thresholding The final decision is made via thresholding:

$$\text{Decision}(x) = \begin{cases}\text{reject}, & \text{if } s(x) > \tau, \\ \arg\max_k \bar{p}(y = k \mid x), & \text{otherwise.}\end{cases} \tag{6}$$

Rather than using a fixed τ, we optimize a threshold τ^* based on user-defined goals. Given targets A_{target}, R_{target}, and weights w_A, w_R, w_E, we minimize:

$$\begin{aligned}\tau^* = \arg\min_{\tau \in R} \Big(&w_A \cdot \max(0, A_{\text{target}} - A(\tau)) \\ &+ w_R \cdot |R_{\text{target}} - R(\tau)| \\ &+ w_E \cdot E(\tau)\Big)\end{aligned} \tag{7}$$

This formulation makes the trade-offs between accuracy, coverage, and calibration explicit and tunable to real-world requirements (Guo et al. 2017; Natarajan et al. 2025).

Experiments and Results

Experimental Setup

Environment. Experiments were conducted on an NVIDIA T4 GPU with 16 GB VRAM. Reproducibility was ensured with a fixed random seed (42), deterministic algorithms, Git version control, and Conda dependency management. Results were averaged over three runs to mitigate stochasticity.

Datasets. The evaluation utilized twelve public datasets across CT, MRI, and X-ray modalities, addressing tasks like COVID-19, tuberculosis, hemorrhage, and cancer detection.

Baselines. Methods were benchmarked against a deep ensemble with 3 and 5 members (Lakshminarayanan, Pritzel, and Blundell 2017). Baselines include a Standard Ensemble using maximum softmax probability with temperature scaling (Guo et al. 2017), an Uncertainty-Based approach with Monte Carlo dropout (MCDO) variance (Gal and Ghahramani 2016), post hoc calibration through isotonic regression and beta calibration (Kull, Silva Filho, and Flach 2017), OOD detection using ODIN and Energy Score (Liang, Li, and Srikant 2018; Liu et al. 2020).

Hyperparameters. Hyperparameters were tuned on a 20% validation split using grid search. Deep ensembles used a learning rate of 10^{-4} and batch size of 32 with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) on images resized to 224×224. T. Rejection optimization targeted 99% accuracy on accepted samples and a 5% rejection rate.

Evaluation Metrics. Performance was assessed using multiple metrics. Risk-Coverage Analysis quantified the trade-off between Coverage and Risk using the Area Under the Risk-Coverage Curve (AURC; lower is better), alongside Accuracy on Accepted samples and Rejection Rate, visualized in Risk-Coverage curves (Figure 2) (El-Yaniv and Wiener 2010; Nadeem, Zucker, and Hanczar 2009). Trustworthiness was measured with Expected Calibration Error (ECE; lower is better) via Reliability Diagrams (Naeini, Cooper, and Hauskrecht 2015).. Rejected Case Analysis categorized rejected samples into Failure, OOD, and Ambiguous cases (Hendrycks and Gimpel 2017).

DEGRE Achieves a Superior Risk-Coverage Trade-off in Selective Prediction

Selective prediction aims to balance predictive performance (risk) with the proportion of inputs accepted (coverage). An optimal system minimizes risk while maximizing coverage. The Area Under the Risk-Coverage Curve (AURC) quantifies this trade-off - lower AURC indicates better performance.

Across twelve medical imaging datasets and two ensemble settings ($M = 3$, $M = 5$), DEGRE consistently outperforms all baselines. It achieves higher accuracy on accepted samples while rejecting fewer, thus delivering superior risk-coverage trade-offs.

Detailed metrics, including AURC, accepted accuracy, and rejection rate, are reported in Table 1.

The results in Table 1 and Figure 3 clearly demonstrate DEGRE's superior risk-coverage trade-off. For $M = 5$ ensembles, DEGRE achieves an average AURC of 0.0027 - representing a 61.8% reduction compared to Training Dynamics (0.0071) and 64.5% over post-hoc methods (AURC $\sim$0.0068).

This advantage is especially pronounced on harder datasets. On Head CT ($M = 5$), Energy Score yields a high AURC of 0.2174 due to poor handling of ambiguous in-distribution cases, while DEGRE reduces this to just 0.0089 - a 24x improvement.

Risk-Coverage curves in Figure 3 (six datasets) further illustrate this: DEGRE (blue) consistently achieves lower risk at every coverage level, forming curves closer to the ideal bottom-right corner.

Dataset	Method	fAURC (M=3)	AA (M=3)	RR (M=3)	AURC (M=5)	AA (M=5)	RR (M=5)
COVID-QU-Ex	Baseline 0	0.0028	0.9895	9.52%	0.0009	0.9974	9.52%
	A.1 (MCDO)	0.0076	0.9762	9.52%	0.0029	0.9851	4.29%
	A.2.1 (Isotonic)	0.0027	0.9893	10.71%	0.0013	0.9974	7.62%
	A.2.2 (Beta)	0.0027	0.9893	10.71%	0.0013	0.9974	7.62%
	B.1.2 (ODIN)	0.0057	0.9870	8.33%	0.0009	0.9975	4.52%
	B.2.2 (Energy)	0.0510	0.9474	95.48%	0.0201	**1.0000**	95.95%
	B.3 (Dynamics)	0.0020	0.9922	8.33%	0.0009	0.9973	10.24%
	DEGRE (Ours)	**0.0018**	**0.9922**	**8.10%**	**0.0008**	0.9975	**4.29%**
SARS-COV-2	Baseline 0	0.0045	0.9883	8.04%	0.0058	0.9884	7.77%
	A.1 (MCDO)	0.0066	0.9796	8.04%	0.0066	0.9710	7.51%
	A.2.1 (Isotonic)	0.0034	0.9885	6.97%	0.0058	0.9886	6.17%
	A.2.2 (Beta)	0.0034	0.9885	6.97%	0.0058	0.9886	6.17%
	B.1.2 (ODIN)	0.0130	0.9709	7.77%	0.0081	0.9716	5.63%
	B.2.2 (Energy)	0.1032	0.9552	22.25%	0.1080	0.0000	100.00%
	B.3 (Dynamics)	0.0056	0.9857	6.43%	0.0045	0.9886	5.90%
	DEGRE (Ours)	**0.0027**	**0.9889**	**3.49%**	**0.0020**	**0.9888**	**4.29%**
Intracranial Hemorrhage	Baseline 0	0.0087	0.9794	10.19%	0.0221	0.9482	4.63%
	A.1 (MCDO)	0.0150	0.9759	10.49%	0.0287	0.9111	**2.78%**
	A.2.1 (Isotonic)	0.0087	**0.9804**	21.30%	0.0221	0.9760	61.42%
	A.2.2 (Beta)	0.0087	**0.9804**	21.30%	0.0221	0.9760	61.42%
	B.1.2 (ODIN)	0.0229	0.9448	**4.94%**	0.0282	0.9316	5.25%
	B.2.2 (Energy)	0.1285	0.7500	98.77%	0.1139	**1.0000**	99.07%
	B.3 (Dynamics)	0.0091	0.9771	8.64%	0.0221	0.9539	6.17%
	DEGRE (Ours)	**0.0031**	0.9797	5.56%	**0.0043**	0.9537	7.41%
Head CT Hemorrhage	Baseline 0	0.0450	0.8846	13.33%	0.0334	0.9200	16.67%
	A.1 (MCDO)	0.0369	**1.0000**	56.67%	0.0306	0.9048	30.00%
	A.2.1 (Isotonic)	0.0400	0.8846	13.33%	0.0329	0.9231	13.33%
	A.2.2 (Beta)	0.0400	0.8846	13.33%	0.0329	0.9231	13.33%
	B.1.2 (ODIN)	0.0795	0.8667	**0.00%**	0.1046	0.9000	**0.00%**
	B.2.2 (Energy)	0.2292	0.6667	90.00%	0.2174	0.5000	93.33%
	B.3 (Dynamics)	0.0438	0.9200	16.67%	0.0372	0.9231	13.33%
	DEGRE (Ours)	**0.0182**	0.9286	6.67%	**0.0089**	**0.9286**	6.67%
Brain Tumor	Baseline 0	0.0004	0.9948	**2.76%**	0.0003	0.9974	3.26%
	A.1 (MCDO)	0.0016	0.9919	6.77%	0.0013	0.9902	5.76%
	A.2.1 (Isotonic)	0.0004	0.9948	3.51%	0.0002	1.0000	5.01%
	A.2.2 (Beta)	0.0004	0.9948	3.51%	0.0002	1.0000	5.01%
	B.1.2 (ODIN)	0.0017	0.9922	3.26%	0.0008	0.9921	4.26%
	B.2.2 (Energy)	0.0370	0.8929	92.98%	0.0446	0.0000	100.00%
	B.3 (Dynamics)	0.0005	0.9948	3.01%	0.0002	0.9974	3.01%
	DEGRE (Ours)	**0.0002**	**0.9974**	3.26%	**0.0002**	**1.0000**	**2.76%**
Brain Cancer	Baseline 0	0.0016	0.9964	7.31%	0.0002	1.0000	7.14%
	A.1 (MCDO)	0.0012	0.9912	5.32%	0.0006	0.9965	**5.32%**
	A.2.1 (Isotonic)	0.0007	0.9965	5.65%	0.0002	1.0000	5.48%
	A.2.2 (Beta)	0.0007	0.9965	5.65%	0.0002	1.0000	5.48%
	B.1.2 (ODIN)	0.0061	0.9911	6.81%	0.0007	0.9982	5.65%
	B.2.2 (Energy)	0.0369	0.8649	97.51%	0.0300	0.8889	98.50%
	B.3 (Dynamics)	0.0017	0.9982	7.14%	0.0002	1.0000	6.31%
	DEGRE (Ours)	**0.0003**	**0.9983**	**3.49%**	**0.0001**	**1.0000**	5.98%

*AA: Accepted Accuracy *RR: Rejection Rate

Table 1: Comprehensive Risk-Coverage Performance Comparison

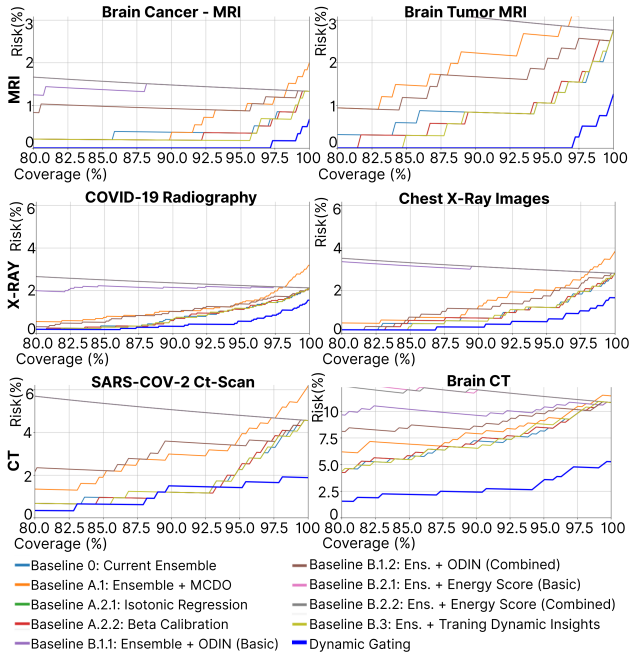

Figure 3: Risk-Coverage Curves on Representative Datasets.

In-Depth Analysis: Deconstructing the Source of DEGRE's Success

To move beyond simply reporting superior metrics, this section presents a series of deeper analyses designed to provide a mechanistic understanding of why DEGRE outperforms existing methods. We demonstrate that its success is not incidental but stems from three core properties: its robustness to diverse and unknown failure modes, its superior efficiency in extracting information from the ensemble, and the fundamentally more discriminative nature of its learned rejection score.

Analysis 1: Robustness to Diverse Failure Modes In real-world clinical deployment, model errors arise from various sources: ambiguous in-distribution (ID) cases, novel out-of-distribution (OOD) inputs, and outright model failures. A reliable safety mechanism must address all these failure types.

Table 2 categorizes rejected samples into #F (failures), #OOD, and #A (ambiguous) across three datasets. Specialized baselines show critical limitations: Temperature Scaling (B.0) identifies ambiguous ID samples but misses OOD, while ODIN (B.1.2) captures OOD but ignores ambiguity. DEGRE, in contrast, provides more balanced rejection, offering broader safety coverage.

DEGRE achieves this by leveraging shared indicators of uncertainty - ensemble consensus confidence (u_c) and disagreement (u_d) - without needing to explicitly differentiate between failure causes. By training a Dynamic Gating Network (DGN) on this space, it learns a unified decision boundary for misclassification detection. As a result, DEGRE functions as a general-purpose failure detector, making it better suited for unpredictable real-world clinical environments.

Dataset	Method	#F	#OOD	#A
Covid-QU-Ex (X-ray)	Baseline 0	5	0	35
	B.1.2	5	14	0
	DEGRE	**3**	**9**	**6**
Brain Cancer (MRI)	Baseline 0	7	0	36
	B.1.2	6	28	0
	DEGRE	**5**	**18**	**13**
Intracranial Hem. (CT)	Baseline 0	8	0	7
	B.1.2	**3**	**14**	0
	DEGRE	9	15	**0**

Table 2: Rejected Sample Composition Analysis ($M = 5$ Ensemble).

Analysis 2: Performance and Efficiency of DEGRE in Ensemble-Based Methods Ensemble-based methods face high computational costs as the number of models (M) increases. DEGRE addresses this by achieving superior performance with smaller ensembles. As shown in Table 3, DEGRE with $M = 3$ outperforms baselines with $M = 5$, offering higher accuracy and lower computational demands. On the Intracranial Hemorrhage dataset, DEGRE ($M = 3$) achieves an AURC of 0.0031, over seven times better than the best $M = 5$ baseline (0.0221), reducing inference-time computation by 40%. Similar trends hold for Head CT and Brain Tumor datasets.

DEGRE's efficiency stems from its Dynamic Gating Network (DGN), which leverages both ensemble consensus and disagreement, treating the latter as a signal of unreliability. This enables DEGRE to extract more information from fewer models, ideal for resource-constrained clinical settings.

The DGN's rejection score excels in distinguishing correct from incorrect predictions, as measured by **AUROC Correct** and **AUPR Correct**. On the Tuberculosis dataset ($M = 3$), DEGRE achieves an AUROC Correct of **0.9880**, surpassing baselines like Training Dynamics (0.9549). This discriminative power ensures a low AURC, optimizing the risk-coverage trade-off. By framing rejection as a supervised learning problem, DEGRE's DGN generates a rejection score optimized for misclassification prediction, outperforming heuristic-based methods.

Ablation Study: Deconstructing the Sources of Performance

To validate our framework, our ablation study addresses two critical questions: (1) How does our direct rejection policy compare to a dynamic re-weighting alternative? (2) What are the individual contributions of the input signals (confidence and disagreement) and the learned gating mechanism?

Architectural Ablation: Rejection Gating vs. Re-weighting Gating To validate our design, we compare DEGRE's direct rejection policy against an alternative that learns to dynamically re-weight ensemble members before

Dataset	Method	AURC	AA	RR
Covid-QU-Ex (X-ray)	Dynamic Re-weighting Gating	0.0025	0.9922	11.19%
	DEGRE (Rejection Gating)	**0.0018**	0.9922	8.10%
Brain Tumor (MRI)	Dynamic Re-weighting Gating	0.0003	0.9973	5.51%
	DEGRE (Rejection Gating)	**0.0002**	0.9974	3.26%

Table 3: Architectural Ablation: Rejection vs Re-weighting (M=3)

applying a standard rejection threshold. This tests whether it is more effective to directly predict misclassifications or to first refine the primary prediction. The results clearly favor DEGRE's direct approach. While both architectures can achieve high accuracy, DEGRE does so more efficiently. For instance, on the Covid-QU-Ex dataset, DEGRE matches the re-weighting gate's accuracy (99.22%) but with a significantly lower rejection rate (8.10% vs. 11.19%). This confirms that for selective prediction, learning a direct rejection policy is a more effective and resource-efficient strategy.

Dataset	B.0	Confidence + Disagreement (Heuristic)	**DEGRE**
Head CT (CT)	0.0334	0.0282	**0.0089**
Sars-cov-2 (CT)	0.0058	0.0081	**0.0020**

Table 4: Architectural Ablation: Confidence vs Disagreement vs Learned Gating (M=3)

Component Ablation: The Value of Disagreement and Learned Gating To prove our hypothesis that DEGRE's superiority stems from using both ensemble disagreement and a learned gating mechanism, we compare three rejection methods: (1) using confidence only (Baseline 0), (2) adding disagreement via a fixed heuristic (ODIN Combined), and (3) our full DEGRE model which learns the combination. The results clearly validate our design. While adding disagreement heuristically provides a modest improvement (e.g., on Head CT, AURC improves from 0.0334 to 0.0282), the performance leap with DEGRE is dramatic. On the same dataset, DEGRE's AURC of 0.0089 is over 3.1 times better than the heuristic approach. This demonstrates that while both signals are crucial, it is the learned combination that drives DEGRE's superior performance over any fixed heuristic.

Dataset	Baseline ($M = 5$)	**DEGRE** ($M = 3$)	Performance
Intracranial Hemorrhage	0.0221	**0.0031**	7.1x
Head CT	0.0334	**0.0182**	1.8x
Brain Tumor	0.0003	**0.0002**	1.5x

Table 5: Ensemble Efficiency Comparison: DEGRE ($M = 3$) vs Baselines ($M = 5$) on AURC

Discussion

Interpretation of Principal Findings

The superior performance of DEGRE results from its approach to treat rejection as a supervised learning problem instead of using fixed heuristics. The Dynamic Gating Network (DGN) generates an extremely discriminative rejection score through its ability to predict misclassifications which surpasses the performance of fixed heuristics (Table 2). The learned policy of DEGRE achieves an excellent risk-coverage trade-off (Figure 3). The ensemble consensus and disagreement mechanism of DEGRE enables it to detect multiple failure modes that include in-distribution ambiguity and out-of-distribution novelty whereas specialized baselines cannot (Table 2). The efficient ensemble signal processing capability of DEGRE enables a 3-model system to outperform 5-model baselines thus making it suitable for clinical environments with limited resources (Table 3).

Broader Implications and Clinical Alignment

The DEGRE system supports the reliable deployment of selective prediction in clinical environments while maintaining human expert oversight through AI assistance. By predicting potential misclassifications, DEGRE enables robust self-awareness that fosters clinical trust and allows for the safe automation of routine cases, while ensuring that complex or ambiguous cases are deferred to clinicians. Furthermore, the goal-driven thresholding mechanism of DEGRE empowers administrators to define target accuracy and rejection rates according to operational requirements, ensuring auditable, policy-driven, and ethically aligned deployment in real-world healthcare workflows.

Limitations and Future Research

Despite its advances, DEGRE has limitations. The computational expense of deep ensembles affects DEGRE and DGN performance depends on the quality of the validation set but distribution shifts can harm results. The research needs to explore additional meta-features and end-to-end training of base models and DGN as well as extend DEGRE to semantic segmentation tasks and develop interactive learning methods to improve DGN performance through human feedback on rejected cases.

Conclusion

The novel meta-learning framework DEGRE tackles deep learning model overconfidence in medical diagnostics through a lightweight Dynamic Gating Network that predicts misclassifications by analyzing deep ensemble consensus confidence and disagreement. The DEGRE framework achieves better risk-coverage trade-offs and well-calibrated predictions through twelve medical imaging benchmark evaluations which enables reliable AI self-awareness. The efficient mechanism enables safe and trustworthy AI integration into clinical workflows which builds trust and promotes responsible AI adoption in medicine.

Acknowledgments

The authors would like to express their sincere gratitude to FPT University, Sun Asterisk, and VinUniversity for their continuous support throughout this research. Special appreciation is extended to the Smart Green Transformation Center (Green-X) and the Smart Health Center (VISHC) at VinUniversity for providing valuable computational resources and research facilities that made this study possible.

References

Berta, E.; Bach, F.; and Jordan, M. I. 2024. Classifier Calibration with ROC-Regularized Isotonic Regression. In *Proceedings of Machine Learning Research*, volume 238.

Bondi, E.; Koster, R.; Sheahan, H.; Chadwick, M.; Bachrach, Y.; Cemgil, T.; Paquet, U.; and Dvijotham, K. 2022. Role of Human-AI Interaction in Selective Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1–9. AAAI Press.

Chidambaram, M.; and Ge, R. 2024. On the Limitations of Temperature Scaling for Distributions with Overlaps. arXiv preprint arXiv:2306.00740.

Chow, C. K. 1970. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1): 41–46.

El-Yaniv, R.; and Wiener, Y. 2010. On the Foundations of Noise-free Selective Classification. In *Journal of Machine Learning Research*, volume 11, 1375–1415.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, 1050–1059.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1321–1330.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. arXiv preprint arXiv:1610.02136.

Huo, J.; Ouyang, X.; Ourselin, S.; and Sparks, R. 2025. Generative Medical Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1–9. AAAI Press.

Joy, T.; Pinto, F.; Lim, S.-N.; Torr, P. H. S.; and Dokania, P. K. 2023. Sample-Dependent Adaptive Temperature Scaling for Improved Calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12859–12867. AAAI Press.

Kull, M.; Silva Filho, T.; and Flach, P. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of Machine Learning Research*, volume 54, 626–635.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30.

Li, B.; Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; Truong, S.; Arora, S.; Mazeika, M.; Hendrycks, D.; Lin, Z.; Cheng, Y.; Koyejo, S.; and Song, D. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Advances in Neural Information Processing Systems*, volume 36.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, volume 33, 2127–2137.

Mu, S.; and Lin, S. 2025. A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications. arXiv preprint arXiv:2503.07137.

Nadeem, M. S.; Zucker, J.-D.; and Hanczar, B. 2009. Accuracy-Rejection Curves (ARCs) for Evaluating Confidence Estimators. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, 159–174. Springer.

Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2901–2907. AAAI Press.

Natarajan, S.; Mathur, S.; Sidheekh, S.; Stammer, W.; and Kersting, K. 2025. Human-in-the-loop or AI-in-the-loop? Automate or Collaborate? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1–9. AAAI Press.

Rahaman, R.; and Thiery, A. 2021. Uncertainty Quantification and Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 34.