arXiv:2506.16742v2 [cs.CV] 22 Sep 2025

UNCERTAINTY-AWARE INFORMATION PURSUIT FOR INTERPRETABLE AND RELIABLE MEDICAL IMAGE ANALYSIS

Md Nahiduzzaman
School of Computing Technologies
RMIT University
Melbourne VIC 3000, Australia
s4045807@student.rmit.edu.au

Steven Korevaar
School of Computing Technologies
RMIT University
Melbourne VIC 3000, Australia
steven.korevaar@rmit.edu.au

Zongyuan Ge
Faculty of Engineering
Monash University
Clayton VIC 3800, Australia
zongyuan.ge@monash.edu

Feng Xia
School of Computing Technologies
RMIT University
Melbourne VIC 3000, Australia
feng.xia@rmit.edu.au

Alireza Bab-Hadiashar
School of Engineering
RMIT University
Melbourne VIC 3000, Australia
alireza.bab-hadiashar@rmit.edu.au

Ruwan Tennakoon
School of Computing Technologies
RMIT University
Melbourne VIC 3000, Australia
ruwan.tennakoon@rmit.edu.au

September 24, 2025

ABSTRACT

To be adopted in safety-critical domains like medical image analysis, AI systems must provide human-interpretable decisions. Variational Information Pursuit (V-IP) offers an interpretable-by-design framework by sequentially querying input images for human-understandable concepts, using their presence or absence to make predictions. However, existing V-IP methods overlook sample-specific uncertainty in concept predictions, which can arise from ambiguous features or model limitations, leading to suboptimal query selection and reduced robustness. In this paper, we propose an interpretable and uncertainty-aware framework for medical imaging that addresses these limitations by accounting for upstream uncertainties in concept-based, interpretable-by-design models. Specifically, we introduce two uncertainty-aware models, EUAV-IP and IUAV-IP, that integrate uncertainty estimates into the V-IP querying process to prioritize more reliable concepts per sample. EUAV-IP skips uncertain concepts via masking, while IUAV-IP incorporates uncertainty into query selection implicitly for more informed and clinically aligned decisions. Our approach allows models to make reliable decisions based on a subset of concepts tailored to each individual sample, without human intervention, while maintaining overall interpretability. We evaluate our methods on five medical imaging datasets across four modalities: dermoscopy, X-ray, ultrasound, and blood cell imaging. The proposed IUAV-IP model achieves state-of-the-art accuracy among interpretable-by-design approaches on four of the five datasets, and generates more concise explanations by selecting fewer yet more informative concepts. These advances enable more reliable and clinically meaningful outcomes, enhancing model trustworthiness and supporting safer AI deployment in healthcare. Our code and models are available at: [GitHub to be released upon acceptance]

Keywords Concept-based models · Explainable AI · Implicit uncertainty handling · Medical imaging · Variational information pursuit

1 Introduction

Given the growing prevalence of deep learning (DL) in safety-sensitive domains like medical diagnosis, ensuring model interpretability is crucial for cultivating trust and facilitating responsible decision-making. Despite their impressive performance, DL models often function a "black box" [1, 2], making it difficult for users to fully trust their outputs, or use them in collaborative settings. To address this challenge, a substantial body of research has focused on developing methods to explain how DL models make decisions, particularly through post-hoc (PH) analysis applied after training. A key advantage of PH methods is that they preserve the original performance. However, PH explanations often fall short of being truly meaningful or faithful to the model's internal reasoning [3, 4, 5, 6]. As a result, interpretable-by-design (IBD) models have recently gained significant attention in the research community. These IBD models are designed such that the final model is fully human-interpretable, decision making processes are transparent and understandable while maintaining strong performance.

One of the most prominent IBD models in medical applications is the Concept Bottleneck Model (CBM) [7]. CBMs work by mapping the extracted features from the input image to an intermediate set of human-understandable concepts such as "blue-whitish veil" or "regression structures" in the context of skin cancer diagnosis. These concepts are then passed through a single linear layer to produce the final prediction. The use of this intermediate concept space offers a transparent explanation of the model's reasoning, as the concepts are directly aligned with clinical knowledge. However, to maintain interpretability, CBMs [7, 8, 9] that employ linear classifiers as their final prediction layers, can be limiting in domains such as medical imaging, where diagnostic features often interact in complex, non-linear ways [10]. Furthermore, as discussed in Section 2.1.2, CBMs are susceptible to information leakage and rely on a fixed set of concepts for decision-making, even when some concepts may be irrelevant or unreliable for specific data points.

The Variational Information Pursuit (V-IP) framework [10, 11, 12, 13, 14], which emulates the human decision-making process, addresses some of the above limitations. Rather than processing all information at once or combining concepts linearly, V-IP observes parts of the input sequentially, selecting each based on its expected informativeness. In skin cancer example, it poses clinically meaningful queries, such as "Is a blue-whitish veil present?" or "Are there regression structures?," sequentially, and uses the accumulated history of query-answer pairs to decide what concept to examine next. This iterative process allows the model to refine its prediction step-by-step and make decisions only based on a subset of concepts that are more relevant for a specific input without any human intervention.

A key limitation of the original V-IP framework lies in its inability to dynamically adapt to upstream uncertainties[1] *specific to individual inputs or samples at test time.* While V-IP can model *dataset-level uncertainty* and prioritise queries (or concepts) that are generally more informative and less uncertain across the training distribution, it does not account for uncertainty specific to *individual samples*. For example, in bowel cancer prediction, "abnormal lymph node appearance" may be a highly predictive concept within the training dataset, and V-IP may learn to prioritise it. However, for a specific test case where lymph nodes are not visible in the patient's imaging data, relying on this feature becomes unreliable. In such cases, the corresponding question–answer pair contributes no diagnostic value, and a human expert would not use it. To make matters worse, the concept predictor model, which can be a DL model, still produces a binary yes/no output for this concept, irrespective of the actual visibility in the input data.

In this paper, we address the previously identified limitation of the V-IP framework and investigate how sample-level uncertainties can be effectively integrated into the V-IP framework. In our approach, we use concept-level uncertainty estimates for each sample to guide the "next-best" query selection process and to update V-IP's belief about a given class. We explore two key strategies for integrating uncertainty: 1) Explicit Uncertainty Handling via *Explicit Uncertainty-Aware V-IP (EUAV-IP)*: Here, highly uncertain concept predictions are excluded through a masking mechanism. 2) Implicit Uncertainty Integration via *Implicit Uncertainty-Aware V-IP (IUAV-IP)*: Rather than masking uncertain queries, IUAV-IP feeds both the concept prediction and its associated uncertainty into the querier. This enables the model to reason about uncertainty directly, instead of ignoring it. Our approach is agnostic about the uncertainty quantification method used. It can leverage many established methods to quantify upstream uncertainties, allowing it to handle both *aleatoric* and *epistemic* uncertainties. Furthermore, uncertainty aware V-IP models (UAV-IP) can provide a classification based on a subset of more certain concepts and ignore the uncertain ones without any expert intervention unlike CBMs [9, 8]. Together, these contributions enable UAV-IP to deliver more efficient, flexible, and clinically realistic interpretations. The main contributions of this work are summarized as follows:

- To the best of our knowledge, this is the first work to incorporate sample-level uncertainty into the V-IP framework, enabling models to make decisions based on a **more reliable subset of concepts per sample**, without requiring human intervention, while maintaining interpretability.

[1]The uncertainty primarily stems from the concept predictor (CP) (i.e., bottleneck model) used by V-IP to generate query answers.

- In addition to the explicit EUAV-IP approach that skips uncertain queries, we propose an implicit integration of uncertainty into the V-IP framework, wherein the querier is modified to consider both the *query history* (i.e., previously asked queries and their answers) and the *uncertainty* of all candidate queries when selecting the next one. Furthermore, we introduce a **penalty-based loss function** that encourages the model to prioritize high-certainty, informative concepts early in the reasoning process. Finally, we incorporate a **hybrid stopping criterion** that adaptively determines when to terminate the querying process, thereby reducing the number of queries while maintaining or improving predictive performance.
- Our results on five public medical imaging datasets with four different modalities demonstrate that UAV-IPs generate more concise and meaningful explanations while maintaining high performance compared to state-of-the-art methods.

2 Related Work

2.1 Interpretable-by-Design (IBD) Methods

2.1.1 Prototype-Based Models

Prototype-based models learn parts of an image that resemble human-understandable patterns. ProtoPNet [15] compares image patches to learned prototypes for each class, offering intuitive visual explanations. However, it needs many prototypes and is slow. ProtoTree [16] makes this more efficient by organizing prototypes in a tree structure. ProtoAttend [17] adds attention to focus on relevant regions. Despite being transparent, these models sometimes learn parts that don't match how humans perceive a concept [18] and defining prototypes may be irrelevant in some medical imaging tasks.

2.1.2 Concept Bottleneck Models (CBMs)

CBMs predict a set of human-labeled concepts before making the final decision. Koh et al. [7] introduced this idea by adding a concept layer between the input and output. Changes to these concept values directly affect the prediction, offering clear insight into the model's reasoning. Later works improved this by ensuring unique concept mappings [18], focusing on relevant image regions [19], and adding decoders to reconstruct the input image [20]. Some works aimed to reduce labeling costs. Yuksekgonul et al. [21] proposed post-hoc CBMs that only train the final layer using concept activation vectors. Others [22, 23] used Generative Pretrained Transformer (GPT) and Contrastive Language–Image Pretraining (CLIP) to generate concepts automatically.

Despite recent advances, many CBMs continue to suffer from limited expressivity due to their reliance on linear mappings from concepts to predictions [7], which may be inadequate for modeling complex decision boundaries [24]. Additionally, CBMs are vulnerable to *information leakage*[25], where latent features inadvertently carry spurious or non-interpretable signals into the prediction layer. This undermines their goal of human-understandable reasoning. Some works attempt to address these issues, for example, logic-based approaches like concept rule learning [26] enforce discrete boolean operations between concepts and labels, offering stronger safeguards against information leakage and enforcing fully-interpretable classification.

2.1.3 Active Testing-Based Models

To move beyond the linear additivity concern in CBMs, Chattopadhyay et al. [10] proposed the Information Persuit (IP), which selects only the most informative queries to classify an input. This follows the idea of active testing [27], similar to how a doctor might ask follow-up questions before making a diagnosis. However, IP relies on complex generative models, making it hard to scale. Later, V-IP model [11] simplified this by jointly learning the query selection (querier) and classifier using cross-entropy loss. Kolek et al. [13] made it more scalable by learning binary queries directly from data using CLIP features. Chattopadhyay et al. [12] further advanced this by integrating GPT and CLIP to generate and answer semantic queries, removing the need for manual concept labels for model training.

2.2 Uncertainty in IBD Models

Uncertainty modeling is widely used to improve the interpretability, robustness, and trustworthiness of DL models [28, 29]. In general, uncertainty in machine learning is categorized into two types: *epistemic uncertainty* (also known as model uncertainty), which arises from limited training data or model capacity, and *aleatoric uncertainty* (data uncertainty), which stems from inherent noise or ambiguity in the input data [30].

Kim et al. [8] introduced Probabilistic Concept Bottleneck Model (ProbCBM), an extension of traditional CBMs in which each concept is modeled as a distribution rather than a single deterministic output. This probabilistic formulation captures the uncertainty in concept predictions and propagates it to the final classifier, leading to more interpretable outputs. In contrast, Evidential Concept Embedding Model (evi-CEM) by Gao et al. [9] models uncertainty with the goal of improving concept prediction quality, particularly in weakly supervised settings, by identifying and correcting misaligned concepts using a small set of labeled examples and Concept Activation Vectors (CAVs).

Despite these strengths, these models inherit a key limitation from standard CBMs: the linear additivity constraint and the issue of information leakage [25, 26]. Additionally, these methods require all concepts to make a decision. When some concepts are uncertain, they depend on human experts to review and adjust them [9], or substitute them with class averages [8], which overlooks sample-specific nuances.

3 Methodology

3.1 Problem Definition and Preliminary

The *supervised query-based classification* problem consists of two primary components. The first involves predicting responses to a set of human-interpretable binary queries for a given input image $x \in \mathcal{X}$. Each query Q_m is a function $Q_m : \mathcal{X} \to \{-1, +1\}$ that maps the input to a binary response $q_m(x)$. For instance, if x represents a dermoscopic image and Q_m refers to the question "Is the lesion palpable/elevated?", then $q_m(x)$ would be $+1$ if the feature is present, and -1 otherwise. The complete set of such queries is denoted by $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_M\}$. Typically, queries, $\mathcal{Q}$ are predefined and curated by domain experts. Alternatively, large language models (LLMs) can be leveraged to derive concept-based queries automatically [31, 22, 23, 12]. The second involves predicting the class label $y \in \mathcal{Y}$ based on the responses $\{q_m(x)\}_{m=1}^{M}$. It is assumed that these query-answer pairs are sufficient for classification, i.e., for all $(x, y) \in (\mathcal{X}, \mathcal{Y})$, the following holds: $P(y \mid x) = P(y \mid x' \in \mathcal{X} : \{q_m(x') = q_m(x)\}_{m=1}^{M})$.

Variational Information Pursuit (V-IP) framework [11] relies on a **concept predictor (CP)**, $Q_\theta : \mathcal{X} \to \{-1, +1\}^M$, which estimates responses to all queries and focuses primarily on predicting the class label (the second component described above). Q_θ can be instantiated using any model, including deep neural networks. To predict the class label, the V-IP applies the following two trainable modules iteratively:

1. *Querier* $g_\gamma : \mathbb{H} \to \mathcal{Q}$: a sequential decision function that, at each step t, selects the most informative query $Q_t \in \mathcal{Q}$ based on the accumulated query-answer history $\mathcal{H}_{[0:t-1]}(x; \gamma) := \{(Q_i, q_i(x))\}_{i=1}^{(t-1)} \in \mathbb{H}$ where $\mathcal{H}_{[0:0]} = \{\}$.
2. *Probabilistic Classifier* $f_\eta : \mathbb{H} \to \mathcal{Y}$: a model that predicts the class distribution.

Both components are jointly trained by minimising the KL divergence between the true label distribution and the model prediction:

$$\mathcal{L}_{\text{VIP}} = \min_{\gamma, \eta} \mathbb{E}_{X, \mathcal{H}} \left[D_{\text{KL}} \left(P(Y \mid X) \parallel f_\eta\big(\mathcal{H}_{[0:t]}(x; \gamma)\big) \right) \right] \tag{1}$$

During inference, queries are selected sequentially until the classifier's confidence surpasses a threshold (MAP) or there is no additional information (Stability)[10]. The collected query-answer pairs up to this point, $\mathcal{H}_{[0:t]}$, serve as the model's explanation.

3.2 Proposed Method

The V-IP framework described above relies on a CP, $Q_\theta(x)$, to generate answers to queries selected by the querier g_γ. However, in real-world scenarios, particularly in domains such as medical imaging, these predictions are often imperfect due to factors like noise, limited supervision, or model uncertainty. Consequently, the querier may prioritise concept queries that appear informative at a dataset level but are unreliable for individual samples, potentially leading to incorrect classifications.

To mitigate this issue, we take inspiration from how human experts navigate uncertainty in clinical decision-making. When faced with ambiguous results from a clinically relevant test, a practitioner may choose to rely instead on other, more reliable sources of information. Following this intuition, we propose a simple yet effective modification to the V-IP framework: the classifier f_η is designed to ignore uncertain concept predictions using a masking strategy. We refer to this approach as *Explicit Uncertainty-Aware Variational Information Pursuit (EUAV-IP)*.

While EUAV-IP introduces a straightforward mechanism to exclude unreliable information during classification, it does not modify the querier's decision process in selecting the "next best" query. Unlike this rigid approach, human

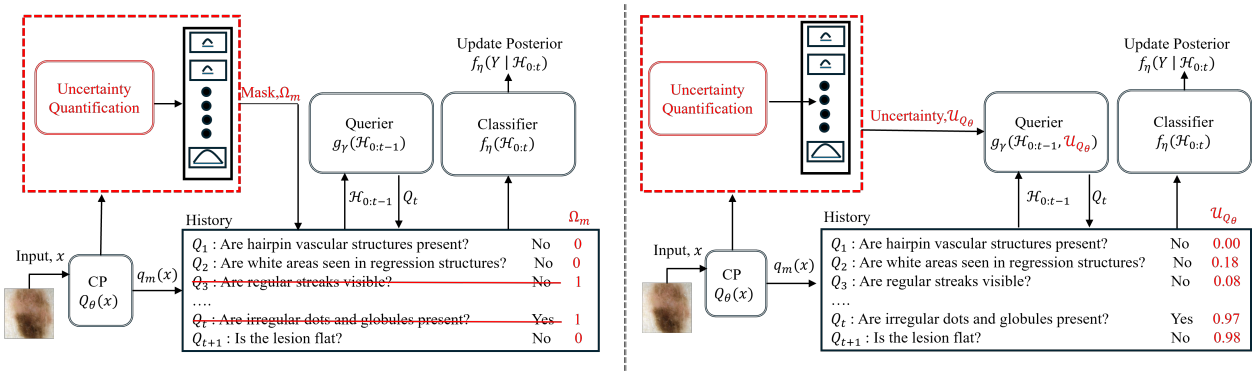


Figure 1: Comparison between the EUAV-IP (left) and IUAV-IP (right) frameworks. While EUAV-IP applies an explicit mask (Ω_m) to discard uncertain concepts based on uncertainty quantification, IUAV-IP introduces an implicit uncertainty-aware approach. Instead of masking, IUAV-IP passes a normalized uncertainty vector $\mathbf{U}_{Q_\theta}$ to the querier, enabling it to learn to select queries that are both informative and reliable. Red highlights indicate the components and signals introduced in EUAV-IP and IUAV-IP.

decision-makers not only disregard unreliable evidence but also adjust future decisions based on their awareness of uncertainty. Furthermore, experts may still consider ambiguous but potentially valuable tests when no clearly reliable alternatives exist.

To emulate this nuanced reasoning, for the first time, we introduce *Implicit Uncertainty-Aware Variational Information Pursuit (IUAV-IP)*. In IUAV-IP, the querier g_γ is conditioned on the estimated uncertainties and learns to prioritize more trustworthy concept signals adaptively. This implicit integration allows the model to reason more flexibly in complex or borderline cases where uncertain concepts may still hold significant predictive value.

An overview of the both UAV-IP architectures, with the modified components highlighted, is shown in Figure 1. The following sections describe our approach to quantifying uncertainty and detail how the query selection and loss functions are modified to support explicit and implicit uncertainty integration.

3.2.1 Explicit Uncertainty Integration in EUAV-IP

EUAV-IP requires *sample-level* uncertainty estimates for each concept prediction q_m produced by the CP $Q_\theta(x)$. In principle, any machine learning model can serve as the CP, along with an appropriate uncertainty quantification method. In this work, we focus specifically on deep learning-based CPs and uncertainty estimation techniques, as described in Section 3.2.3.

For a given input sample x, the uncertainties associated with all concept predictions are denoted by the vector

$$\mathcal{U}_{Q_\theta}(x) = [\mathcal{U}_{Q_1}(x), \ldots, \mathcal{U}_{Q_M}(x)] \,.$$

These uncertainties are used to define a binary mask $\Omega_m(x) \in \{0, 1\}$ for each query m, which indicates whether the corresponding concept prediction should be masked out due to high uncertainty. The mask is defined as:

$$\Omega_m(x) = \begin{cases} 1 & \text{if } \mathcal{U}_{Q_m}(x) \geq \mathcal{T}_\mathcal{U} \quad \text{(uncertain)}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, $\mathcal{T}_\mathcal{U}$ is a predefined or learned threshold that determines whether a concept is considered uncertain. This threshold can be set manually or optimized using a validation set, for example, via a lightweight classifier trained to distinguish between correct and incorrect concept predictions.

The binary uncertainty masks are incorporated into the probabilistic classifier during both training and inference to ignore unreliable concept predictions selectively. The masked classifier is defined as:

$$f_\eta \left(\mathcal{H}_{[0:t]}(x; \gamma) \odot \Omega_{[0:t]}(x) \right) . \tag{3}$$

3.2.2 Implicit Uncertainty Integration in IUAV-IP

While EUAV-IP excludes high-uncertainty concepts using binary masks, the IUAV-IP framework adopts a softer approach. IUAV-IP provides the querier with both the concept history, $\mathcal{H}_{[0:t-1]}$, and its associated uncertainty, $\mathcal{U}_{Q_\theta}(x)$,

enabling it to learn a query selection policy that balances informativeness and reliability that is adapted to the current sample. The modified Querier is defined as:

$$Q_{[t]} = g_\gamma \left(\mathcal{H}_{[0:t-1]}(x;\gamma), \mathcal{U}_{Q_\theta}(x) \right).$$

The classifier retains its original form as V-IP and operates exclusively on the history, $\mathcal{H}_{[0:t]}$.

Updated loss function: While the above model can be directly optimized using the V-IP loss function in (1), our ablation study (Sec. 4.2.4) revealed that the model does not adequately prioritize certain queries over uncertain ones. To address this and to further encourage the querier to select reliable queries, we introduce an additional penalty term that is applied when a selected concept prediction is incorrect:

$$\mathcal{L}_{\text{penalty}} = \mathbb{E}_X \left[\sum_{i=1}^{t} \mathbb{I}[q_{[i]}(x) \neq \tilde{q}_{[i]}(x)] \right], \tag{4}$$

where $q_{[i]}(x)$, $\tilde{q}_{[i]}(x)$ denote the predicted and ground-truth responses for query $Q_{[i]}$, respectively, and $\mathbb{I}[\cdot]$ is the indicator function, which equals 1 if the condition inside the brackets is true and 0 otherwise. The overall loss function is:

$$\mathcal{L}_{\text{IUAV-IP}} = \mathcal{L}_{\text{V-IP}} + \lambda \cdot \mathcal{L}_{\text{penalty}}, \tag{5}$$

where λ is a hyperparameter we tune using a validation set.

Hybrid Stopping Criterion: Since the V-IP framework selects queries sequentially, a critical design decision is determining when to stop querying. Prior work has utilized threshold- or stability-based criteria, which halt the process once a predefined threshold is reached or when no additional information is gained. While intuitive, this approaches can either terminate too early, missing informative concepts, or continue unnecessarily, resulting in longer explanations without meaningful performance gains. As shown in Table 3, this often leads to an increased number of queries without corresponding improvements in accuracy.

To address these limitations, we adopt a hybrid confidence-plateau criterion that jointly considers both prediction certainty and stability. Specifically, we monitor the classifier's predicted confidence scores across steps and stop querying when the following condition is met:

$$[c_t \geq \tau] \wedge \left[\frac{1}{w-1} \sum_{j=t-w+1}^{t-1} |c_{j+1} - c_j| < \delta \right].$$

Here $c_t = \max f_\eta(\mathcal{H}_{[1:t]})$ denote the classifier's confidence at step t, w is the window length and δ, τ are the thresholds.

3.2.3 Query-Wise Uncertainty Quantification

Uncertainty estimation is a critical component in safety-sensitive domains like medical imaging. Prior work has proposed various approaches for quantifying uncertainty in deep neural networks [32]. In this study, we adopted two widely used techniques to quantify uncertainty in the outputs of the CP Q_θ.

Entropy-based Uncertainty: Given an input image x and a query Q_m, the concept predictor outputs a predictive distribution $\Phi^{(m)} \in \mathbb{P}(\mathcal{Y})$, where each element $\phi_k^{(m)} = P(Y = y_k \mid x, \theta)$, and $k \in \{-1, +1\}$ due to binary outputs. The entropy of this distribution captures uncertainty as:

$$\mathcal{U}_{Q_m}^{\mathcal{H}}(x) = - \sum_{k \in \{-1,+1\}} \phi_k^{(m)} \log_2(\phi_k^{(m)}). \tag{6}$$

Monte Carlo Dropout-based Uncertainty: Following Sale et al. [29], we also estimate uncertainty using a second-order distribution (i.e., $\mathbb{P}(\mathbb{P}(\mathcal{Y}))$) over predicted probabilities. However, unlike their approach, which uses an ensemble of models to approximate this distribution, we adopt MC dropout [28] to capture both aleatoric and epistemic uncertainty in a more computationally efficient manner. Let $\Pi^{(m)} \in \mathbb{P}(\mathbb{P}(\mathcal{Y}))$ be the distribution over predictive probabilities obtained via multiple stochastic forward passes. The total uncertainty is:

$$\text{TU}(Q_m) = \underbrace{\mathbb{E}\left[\Phi^{(m)}(1 - \Phi^{(m)})\right]}_{\text{Aleatoric}} + \underbrace{\text{Var}\left(\Phi^{(m)}\right)}_{\text{Epistemic}}. \tag{7}$$

where $_{\Phi^{(m)} \sim \Pi_m}$. This can be approximated with finite number of MC samples $\mathcal{S}$:

$$\mathcal{U}_{Q_m}^{\text{MC}}(x) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \Phi_m^{(s)}(1 - \Phi_m^{(s)}) + \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(\Phi_m^{(s)} - \bar{\Phi}_m \right)^2. \tag{8}$$

4 Experiments and Results

4.1 Experimental Setup

Datasets: We evaluate our framework on five publicly available medical imaging datasets spanning four modalities: dermoscopy (PH2 [33], Derm7pt [34]), ultrasound (BrEaST [35]), microscopy (White Blood Cell (WBCatt) [36]), and knee X-ray (Osteoarthritis Initiative (OAI) [37]). **PH2** and **Derm7pt** comprise 200 and 827 dermoscopic images respectively, both labeled as *Nevus* or *Melanoma*, and annotated with 8 and 38 binary clinical concepts. **BrEaST** includes 256 breast ultrasound images labeled as *Benign* or *Malignant*, with 7 binary BI-RADS descriptors. **WBCatt** comprises 10,198 peripheral blood smear images from five white blood cell types, each annotated with 29 binary attributes related to morphology, granularity, and cytoplasmic and nuclear features. **OAI** includes X-ray images from 4,506 patients with 4-level Kellgren-Lawrence Grades (KLG) and 10 ordinal concepts, which we model as a classification task, whereas [7] regress to the final KLG.

Evaluation metrics: To evaluate the performance of our models, we report standard classification metrics including area under the ROC curve (AUC), accuracy (Acc), and the average number of queries needed to reach a decision (Q_{avg}). In addition to these standard metrics, we introduce two reliability-focused evaluation metrics to assess the quality and impact of query selection:

Step-wise Error Rate (SER) Let $\tilde{q}_m$ be the ground-truth concept value for query Q_m, and q_m the predicted value. Then SER is computed over all samples as: $\text{SER}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[q_t^{(i)} \neq \tilde{q}_t^{(i)}]$, where N is the number of test samples.

Global Error Rate (GER) This represents the average number of wrong queries selected by the model: $\text{GER} = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{t=1}^{T_i} \mathbb{I}[q_t^{(i)} \neq \tilde{q}_t^{(i)}]}{T_i} \times Q_{avg}$, where T_i is the total number of queries used by sample i.

Implementation details: Our framework is trained in two stages. First, a *CP* with a ResNet-101 backbone is trained using concept supervision. Then, the *querier* and *classifier* are trained using the now-trained and static CP's concept predictions and their associated uncertainties. While training the EUAV-IP and IUAV-IP models, we follow the same configuration and hyperparameter settings as in the original V-IP framework [11]. For our proposed IUAV-IP model, we introduce an additional penalty loss term (Eq. 5) to discourage the querier from selecting incorrect concept predictions. We tune the penalty coefficient λ per dataset based on validation performance. The hybrid stopping criterion (3.2.2) is used during inference for all models. The confidence threshold τ is set to 0.90 for all binary classification and WBCatt datasets and 0.70 for the other multi-class datasets. The window size w is fixed at 3, and the confidence change threshold δ is set to 0.004 across all datasets. All implementation details and code are available at: [GitHub repository to be released upon acceptance of the paper].

4.2 Experimental Results

4.2.1 Quantative Evaluation

The performance of the proposed EUAV-IP and IUAV-IP models across five datasets is summarised in Table 1. All methods were run 10 times with different random seeds and train-validation-test splits; mean and standard deviation are reported.

We compare our proposed methods against a range of baselines, including a BlackBox classifier (ResNet101) and a decision tree-based interpretable model (CBM-DT [38]), where a decision tree is used to classify the outputs from a CP. We also include comparisons with the standard CBM [7] and its recent uncertainty-aware variants, ProbCBM [8] and evi-CEM [9].

Among V-IP-based methods, we evaluate the original V-IP [11], two explicit uncertainty-aware variants: EUAV-IP (Entropy) and EUAV-IP (Monte Carlo), two implicit uncertainty-aware variants: IUAV-IP (Entropy) and IUAV-IP (Monte Carlo), and an oracle-based version, UAV-IP (Oracle), which uses ground-truth concepts to mask incorrect answers. While the oracle provides an upper bound on performance, it is not feasible in real-world scenarios where concept labels are unavailable at inference time.

Among the four variants of our proposed method, the IUAV-IP (Monte Carlo) variant consistently achieves high accuracy across all five datasets while maintaining shorter query lengths. This demonstrates that implicit uncertainty integration is more effective than explicit integration — partly because explicit methods rely on thresholds, whereas implicit methods can leverage uncertainty without such constraints. Additionally, the results indicate that Monte Carlo–based uncertainty estimation is more effective than entropy-based estimation. Further discussion of these findings is provided in Section 5.1. The proposed variants significantly outperform the original V-IP [11] method, demonstrating

Table 1: Comparative performance of different state-of-the-art models across five datasets. **Bold** indicates the best non-oracle result, underline indicates the second-best. FWQ is the First Wrong Query.

Dataset	Model	Accuracy ↑	AUC ↑	Queries ↓	FWQ ↑	GER ↓
Derm7pt	CBM-DT [38]	78.74 ± 2.65	82.05 ± 2.09	38.00 ± 0.00	-	-
	CBM [7]	82.01 ± 1.27	87.60 ± 2.68	38.00 ± 0.00	-	-
	ProbCBM [8]	82.89 ± 2.11	85.01 ± 3.21	38.00 ± 0.00	-	-
	evi-CEM [9]	79.70 ± 2.06	83.95 ± 2.50	38.00 ± 0.00	-	-
	V-IP [11]	81.81 ± 1.38	87.90 ± 1.33	26.81 ± 0.93	4.75 ± 0.32	4.29 ± 0.27
	EUAV-IP (Entropy)	81.75 ± 2.33	87.91 ± 2.07	24.68 ± 3.38	6.92 ± 0.67	5.18 ± 0.28
	EUAV-IP (Monte Carlo)	83.49 ± 1.26	89.99 ± 1.49	23.65 ± 0.52	5.84 ± 0.38	6.15 ± 1.62
	IUAV-IP (Entropy)	82.55 ± 1.98	90.47 ± 3.25	28.36 ± 0.66	18.10 ± 0.50	3.97 ± 0.33
	IUAV-IP (Monte Carlo)	**90.36 ± 3.88**	**93.59 ± 4.14**	**23.52 ± 0.70**	**20.26 ± 0.79**	**2.12 ± 0.25**
	UAV-IP (Oracle)	92.29 ± 1.70	97.18 ± 0.64	18.44 ± 0.86	0.00	0.00
	BlackBox	83.80 ± 1.43	86.34 ± 1.78	-	-	-
PH2	CBM-DT [38]	90.50 ± 3.26	85.71 ± 6.24	8.00 ± 0.00	-	-
	CBM [7]	92.00 ± 4.11	**96.50 ± 2.74**	8.00 ± 0.00	-	-
	ProbCBM [8]	82.00 ± 2.09	83.53 ± 2.52	8.00 ± 0.00	-	-
	evi-CEM [9]	90.50 ± 3.26	93.77 ± 3.79	8.00 ± 0.00	-	-
	V-IP [11]	88.00 ± 4.47	89.77 ± 4.88	7.96 ± 0.02	2.27 ± 0.73	0.24 ± 0.16
	EUAV-IP (Entropy)	91.30 ± 3.21	93.75 ± 4.20	**6.13 ± 0.28**	1.80 ± 1.10	0.18 ± 0.12
	EUAV-IP (Monte Carlo)	92.10 ± 3.16	95.40 ± 1.29	6.29 ± 0.16	3.40 ± 0.89	0.25 ± 0.10
	IUAV-IP (Entropy)	92.00 ± 3.25	95.29 ± 3.39	7.37 ± 0.21	5.91 ± 0.92	0.18 ± 0.16
	IUAV-IP (Monte Carlo)	**92.50 ± 3.95**	95.37 ± 2.52	7.62 ± 0.39	**5.97 ± 0.97**	**0.15 ± 0.12**
	UAV-IP (Oracle)	92.50 ± 3.18	96.23 ± 2.21	6.13 ± 0.20	0.00	0.00
	BlackBox	95.00 ± 3.06	97.88 ± 1.78	-	-	-
BrEaST	CBM-DT [38]	84.70 ± 3.56	85.91 ± 3.82	7.00 ± 0.00	-	-
	CBM [7]	85.88 ± 4.22	91.10 ± 3.20	7.00 ± 0.00	-	-
	ProbCBM [8]	79.41 ± 1.64	84.08 ± 5.09	7.00 ± 0.00	-	-
	evi-CEM [9]	80.78 ± 3.17	85.34 ± 3.21	7.00 ± 0.00	-	-
	V-IP[11]	87.57 ± 5.09	94.28 ± 2.00	**6.25 ± 0.33**	2.58 ± 0.48	0.31 ± 0.13
	EUAV-IP (Entropy)	88.64 ± 1.86	94.40 ± 1.64	6.51 ± 0.09	2.08 ± 0.61	0.27 ± 0.10
	EUAV-IP (Monte Carlo)	88.14 ± 3.97	95.05 ± 2.29	6.57 ± 0.05	2.82 ± 0.82	0.26 ± 0.10
	IUAV-IP (Entropy)	88.62 ± 3.30	94.94 ± 2.16	6.26 ± 0.14	4.27 ± 1.44	**0.21 ± 0.15**
	IUAV-IP (Monte Carlo)	**89.30 ± 3.20**	**95.83 ± 1.18**	6.96 ± 0.03	**4.49 ± 0.31**	0.35 ± 0.14
	UAV-IP (Oracle)	89.44 ± 4.36	95.98 ± 1.79	6.20 ± 0.10	0.00	0.00
	BlackBox	88.45 ± 2.21	92.65 ± 3.21	-	-	-
WBCatt	CBM-DT [38]	95.44 ± 0.001	97.90 ± 0.001	29.00 ± 0.00	-	-
	CBM [7]	**96.92 ± 0.001**	**99.82 ± 0.001**	29.00 ± 0.00	-	-
	ProbCBM [8]	95.60 ± 0.24	95.60 ± 0.24	29.00 ± 0.00	-	-
	evi-CEM [9]	96.20 ± 0.06	99.37 ± 0.06	29.00 ± 0.00	-	-
	V-IP [11]	96.11 ± 0.12	99.04 ± 0.14	8.98 ± 0.22	5.11 ± 0.34	0.58 ± 0.05
	EUAV-IP (Entropy)	96.29 ± 0.07	99.08 ± 0.04	8.41 ± 0.12	5.70 ± 0.18	0.24 ± 0.01
	EUAV-IP (Monte Carlo)	95.38 ± 0.25	99.24 ± 0.03	8.90 ± 0.37	5.30 ± 1.02	0.13 ± 0.01
	IUAV-IP (Entropy)	96.77 ± 0.20	99.43 ± 0.15	8.88 ± 0.27	6.97 ± 0.12	0.15 ± 0.01
	IUAV-IP (Monte Carlo)	96.83 ± 0.23	99.47 ± 0.15	**8.32 ± 0.68**	**7.11 ± 0.21**	**0.11 ± 0.02**
	UAV-IP (Oracle)	98.97 ± 0.02	99.95 ± 0.01	5.94 ± 0.41	0.00	0.00
	BlackBox	99.00 ± 1.00	99.98 ± 0.01	-	-	-
OAI	CBM-DT [38]	54.61 ± 0.00	59.64 ± 0.00	10.00 ± 0.00	-	-
	CBM [7]	54.50 ± 0.01	70.47 ± 0.001	10.00 ± 0.00	-	-
	ProbCBM [8]	57.48 ± 0.83	71.53 ± 1.20	10.00 ± 0.00	-	-
	evi-CEM [9]	58.17 ± 0.02	72.72 ± 0.18	10.00 ± 0.00	-	-
	V-IP[11]	52.57 ± 0.36	68.80 ± 0.62	8.24 ± 0.10	2.30 ± 0.15	1.92 ± 0.07
	EUAV-IP (Entropy)	51.81 ± 0.17	65.89 ± 0.16	8.27 ± 0.23	2.07 ± 0.08	1.98 ± 0.02
	EUAV-IP (Monte Carlo)	51.33 ± 0.44	65.87 ± 0.65	8.16 ± 0.13	2.32 ± 0.20	1.80 ± 0.07
	IUAV-IP (Entropy)	58.44 ± 1.97	73.24 ± 0.86	8.43 ± 0.65	**5.15 ± 0.38**	1.35 ± 0.30
	IUAV-IP (Monte Carlo)	**60.17 ± 0.14**	**75.14 ± 0.09**	**7.15 ± 0.16**	4.24 ± 0.07	**0.97 ± 0.10**
	UAV-IP (Oracle)	71.79 ± 0.35	88.39 ± 0.70	8.07 ± 0.01	0.00	0.00
	BlackBox	87.22 ± 0.01	95.38 ± 0.01	-	-	-

the effectiveness of incorporating uncertainty. On average, IUAV-IP (Monte Carlo) achieves around a 5% improvement in accuracy while using 20% fewer queries.

IUAV-IP (Monte Carlo) achieves near-oracle performance on most datasets, despite not having access to ground-truth concepts. The only exception is the OAI dataset, where limited CP (see Table 2) quality hinders accurate uncertainty estimation. While the BlackBox model achieves high accuracy, it lacks transparency and interpretability, providing no explanation for its decisions.

Furthermore, IUAV-IP (Monte Carlo) outperforms CBM [7] and its uncertainty-aware variants (ProbCBM [8] and evi-CEM [9]) across all datasets, while also reducing the number of queries required to reach a decision.[2] For example, on the Derm7pt dataset, CBM-based approaches require all 38 concepts for each prediction, whereas IUAV-IP, on average, uses only 23 informative queries. This selective querying reduces explanation length and better mirrors clinical reasoning, where decisions are often based on a few key observations rather than exhaustive assessments. Thus, IUAV-IP provides more concise and intuitive explanations that are easier for domain experts to interpret while also being more accurate.

4.2.2 Reliability Evaluation

To assess the robustness of query selection, we use two additional metrics: Step Error Rate (SER) and First Wrong Query (FWQ). The FWQ asked metric tracks when a model incorporates a concept with a wrong answer. Ideally, models that utilize uncertainty will be able to determine the likelihood of a concept being wrong and postpone relying on it, thus having a higher FWQ. As shown in Table 1, in all four datasets, IUAV-IP (Monte Carlo) utilizes uncertain queries much later than V-IP, demonstrating the effectiveness of conditioning the querier with uncertainty values and the proposed loss function.

Figure 3 shows the SER per query index across all test samples in the Derm7pt dataset. SER measures how often the model selects an incorrect concept at each step. A lower SER indicates more reliable queries are being selected. However, given that our method does not improve the accuracy of the CP, the total number of errors will stay the same between methods, however by prioritizing correct queries at the start and stopping early, we can assure that more reliable queries are used for classification. As shown in the right plot, IUAV-IP exhibits a steadily increasing SER curve that remains relatively low and stable until around step 16-18. This behavior shows that IUAV-IP prioritizes the most certain (and thus reliable) queries early in the decision process. This corresponds with our quantitative results in Table 1, where IUAV-IP achieves high accuracy using fewer queries. In contrast, V-IP shows noisy SER trends, especially in the early stages, reflecting an inability to determine query reliability.

4.2.3 Qualitative Analysis

We also present a qualitative case study from the BrEaST dataset in Figure 2. The input image, labeled as *Malignant*, was evaluated using the V-IP, EUAV-IP, and our proposed IUAV-IP models. First, we computed the total uncertainty per query using the Monte Carlo-based uncertainty formulation in Eq. 8, which is shown in panel (A).

V-IP Behavior. Since the original V-IP framework does not consider uncertainty, it selects the next query solely based on informativeness. As shown in panel (B-left), it initially selects an informative concept (*Query 6: Answer = Yes*) and slightly improves its confidence. However, in subsequent steps, it selects high-uncertainty queries (e.g., *Query 1, 2*) whose answers are incorrect. These errors significantly reduce its belief in the correct class. Consequently, the model's confidence fluctuates and ultimately results in a misclassification, labeling a malignant case as benign.

EUAV-IP Behavior. In contrast, EUAV-IP explicitly masks high-uncertainty queries and avoids using them. As seen in panel (B-middle), it skips uncertain queries (e.g., Query 1, 2) and focuses on more reliable ones like *Query 6* and *Query 5*. After just five steps, the model confidently predicts *Malignant* with over 90% confidence.

IUAV-IP Behavior. IUAV-IP achieves a similar outcome by utilizing concept predictions and their uncertainties. Panel (B-right) illustrates this reasoning process: IUAV-IP begins with a highly certain and relevant concept (*Query 5*). It then selects another informative and confident query (*Query 6*) and so on. As more queries are selected, the model builds confidence gradually. By step 5, its confidence stabilizes above the decision threshold, and the querying process stops with an accurate final prediction *Malignant*. This example illustrates how V-IP may fall prey to misleading high-uncertainty answers, while EUAV-IP succeeds by skipping them. IUAV-IP, without discarding any queries, achieves the same level of reasoning by learning when to trust or ignore uncertain inputs.

[2]While models like ProbCBM and evi-CEM are designed to incorporate human feedback or interventions to adjust concept responses, we did not utilise such interventions in our experiments to ensure a fair comparison across methods.

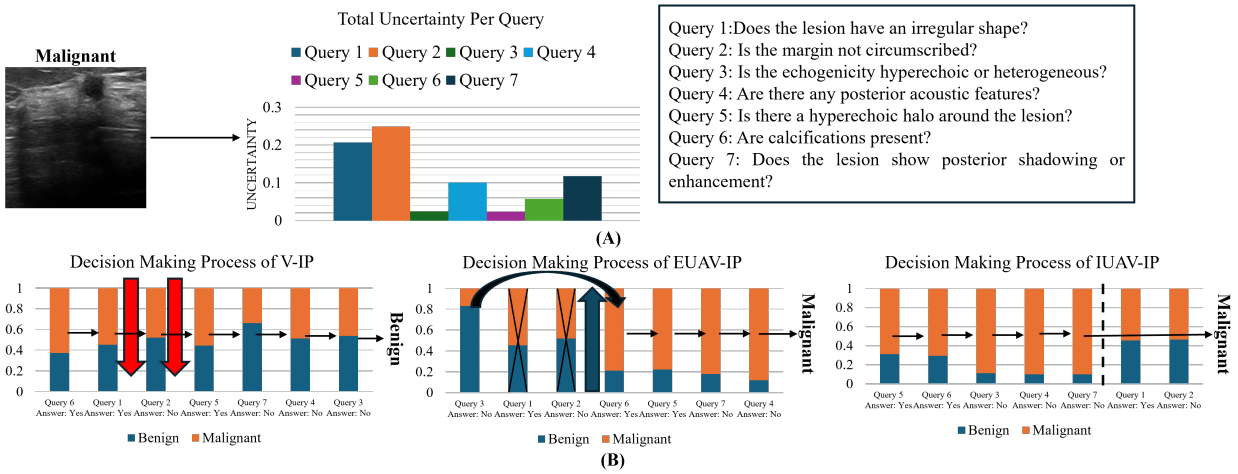


Figure 2: **Qualitative comparison of query-based reasoning across models for a malignant ultrasound image from the BrEaST dataset. (A)** Total uncertainty per query computed using Monte Carlo Dropout (Eq. 8). Queries 1 and 2 are highly uncertain, whereas Queries 5 and 6 are more confident. **(B)** Decision-making trajectories for three models: V-IP, EUAV-IP and IUAV-IP. V-IP selects highly uncertain and incorrect queries (red arrows), leading to an incorrect prediction (*Benign*), where EUAV-IP skips the uncertain queries. Our IUAV-IP model reaches the correct prediction (*Malignant*) with stable confidence, without relying on query masking.

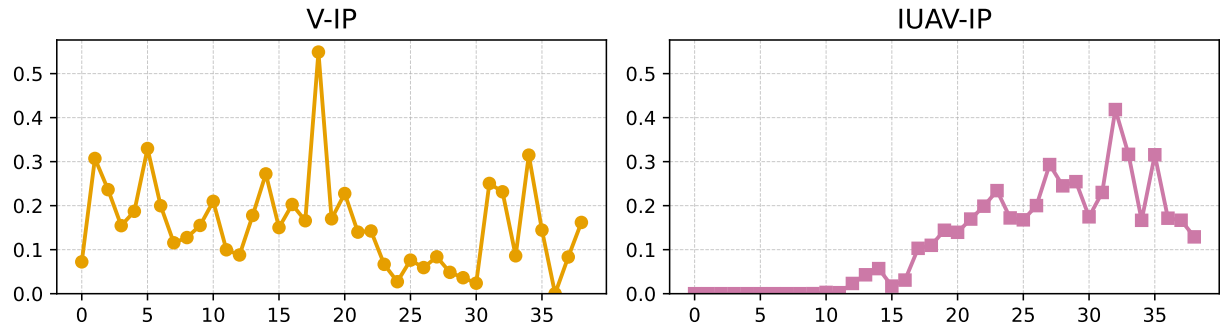


Figure 3: **Step-wise Error Rate (SER) comparison across models on the Derm7pt dataset.** IUAV-IP (right) maintains a low SER in early steps, with errors rising only in later queries-indicating more informed and cautious reasoning. In contrast, V-IP shows earlier spikes in SER, reflecting unstable or premature query decisions. This highlights the effectiveness of penalty-guided querying in delaying incorrect answers and improving explanation reliability.

4.2.4 Ablation Study

To assess the impact of different uncertainty estimation methods and our two novel contributions: (1) a loss penalty that prioritizes more certain queries, and (2) a novel stopping criterion. We analyze the model's performance with different configurations in an ablation study.

As shown in Table 1, the second-order Monte Carlo-based uncertainty estimates yielded slightly higher AUCs compared to first-order entropy-based estimates. This improvement may be attributed to the Monte Carlo method's ability to capture both epistemic (model) and aleatoric (data) uncertainties, whereas entropy primarily reflects epistemic uncertainty. Accurately estimating both types of uncertainty is crucial in medical imaging applications to enhance model reliability and trustworthiness.

Table 3 presents the effectiveness of incorporating the penalty term (P) and the hybrid stopping criterion (HS) in the IUAV-IP (Monte Carlo) model. The results indicate that the IUAV-IP model, when augmented with both the penalty term

Table 2: Effectiveness of uncertainty estimates in classifying query-answer correctness, measured by average AUC using ground-truth annotations. ConceptAUC is the AUC of concept predictor model.

Dataset	Monte Carlo	Entropy	ConceptAUC
PH2	95.75%	91.19%	99.48%
Derm7pt	84.59%	74.64%	70.71%
BrEaST	91.00%	71.30%	99.96%
WBCatt	90.39%	90.19%	95.09%
OAI	79.84%	77.00%	73.68%

and the hybrid stopping criterion, achieves higher performance compared to configurations where these components are omitted. Notably, the inclusion of the penalty term contributes more significantly to the overall performance enhancement. Figure 4 further substantiates these findings. In the Derm7pt dataset (with similar trends observed in other datasets), the IUAV-IP (Monte Carlo) model with the penalty term initiates with confident queries and gradually transitions to more uncertain ones. Conversely, the variant without the penalty term displays erratic certainty patterns, indicating less effective prioritization. Collectively, these quantitative and qualitative insights validate our hypothesis that effective uncertainty-aware reasoning necessitates both input-awareness and penalty-based supervision.

The results in Table 3 consistently demonstrate that the hybrid stopping approach reduces the average number of queries and improves accuracy and reliability.

Table 3: Comparison of IUAV-IP across combinations of Penalty (P) and Hybrid Stopping (HS). **Bold** indicates best performance per dataset.

Dataset	P	HS	Accuracy ↑	AUC ↑	Queries ↓
Derm7pt	✗	✗	82.16 ± 2.21	89.98 ± 2.14	27.10 ± 0.21
	✗	✓	84.68 ± 2.50	91.62 ± 3.28	26.78 ± 0.46
	✓	✗	81.33 ± 1.75	88.26 ± 1.81	26.45 ± 0.70
	✓	✓	**90.36 ± 3.88**	**93.59 ± 4.14**	**23.52 ± 0.70**
BrEaST	✗	✗	88.03 ± 2.53	94.22 ± 1.43	6.89 ± 0.03
	✗	✓	89.00 ± 2.65	95.70 ± 1.82	6.69 ± 0.43
	✓	✗	88.12 ± 2.82	94.34 ± 1.87	**6.45 ± 0.72**
	✓	✓	**89.60 ± 3.20**	**95.83 ± 1.18**	6.96 ± 0.03
OAI	✗	✗	53.14 ± 1.57	70.19 ± 0.21	9.88 ± 0.13
	✗	✓	54.12 ± 1.22	72.14 ± 0.08	9.32 ± 0.68
	✓	✗	53.57 ± 0.28	70.36 ± 0.23	9.81 ± 0.28
	✓	✓	**60.17 ± 0.14**	**75.14 ± 0.09**	**7.15 ± 0.16**

5 Discussion

5.1 Uncertainty as a Proxy for Reliability

A key hypothesis of this work, which explains the superior performance of our proposed method, is that concept-level uncertainty, derived using either entropy or Monte Carlo Dropout, serves as a reliable heuristic for the correctness of predicted concept values. Table 2 presents the average AUC for binary classification of correct vs. incorrect concept answers using uncertainty estimates. Both Monte Carlo and Entropy estimates achieve strong discrimination of concept correctness. However, the second-order Monte Carlo-based estimates consistently outperform entropy-based ones. For example, in the Derm7pt dataset, Monte Carlo achieves an AUC of 84.59%, compared to 74.64% for entropy. Furthermore, as shown in Table 1, the more reliable the uncertainty estimator, the better the performance.

5.2 Dynamic Subset Selection for Explainability and Uncertainty Handling

A key idea in the V-IP framework is that basing predictions on a subset of concepts improves explainability, reflected in the evaluation of query length, where shorter is better.

This is particularly relevant to medical imaging for two reasons. First, it mirrors how clinicians make decisions: when reporting a diagnosis, they focus on the most relevant concepts rather than listing all possible ones. Second, by dynamically selecting a subset of concepts for each data point, the model can disregard those that are uncertain and base its decisions solely on the most reliable information available.

While V-IP captures the first behaviour, it lacks the ability to address the second. Our method adds this capability.

Table 4: Validation that querying a subset of concepts yields strong performance. **Bold** indicates the best non-oracle result.

Dataset	Model	Accuracy ↑	AUC ↑	Queries ↓
Derm7pt	V-IP (NS)	81.21 ± 1.00	87.40 ± 2.17	38.00 ± 0.00
	EUAV-IP	83.49 ± 1.26	89.99 ± 1.49	23.65 ± 0.52
	IUAV-IP	**90.36 ± 3.88**	**93.59 ± 4.14**	**23.52 ± 0.70**
	UAV-IP (Ora-NS)	92.29 ± 1.61	97.09 ± 0.57	32.56 ± 0.22
BrEaST	V-IP (NS)	86.49 ± 5.29	94.98 ± 2.49	7.00 ± 0.00
	EUAV-IP	88.14 ± 3.97	95.05 ± 2.29	**6.57 ± 0.05**
	IUAV-IP	**89.30 ± 3.20**	**95.83 ± 1.18**	6.96 ± 0.03
	UAV-IP (Ora-NS)	89.44 ± 4.13	95.57 ± 2.38	6.94 ± 0.03
OAI	V-IP (NS)	52.57 ± 0.36	68.80 ± 0.62	10.00 ± 0.00
	EUAV-IP	51.33 ± 0.44	65.87 ± 0.65	8.16 ± 0.13
	IUAV-IP	**60.17 ± 0.14**	**75.14 ± 0.09**	**7.15 ± 0.16**
	UAV-IP (Ora-NS)	71.76 ± 0.39	88.39 ± 0.70	8.14 ± 0.00

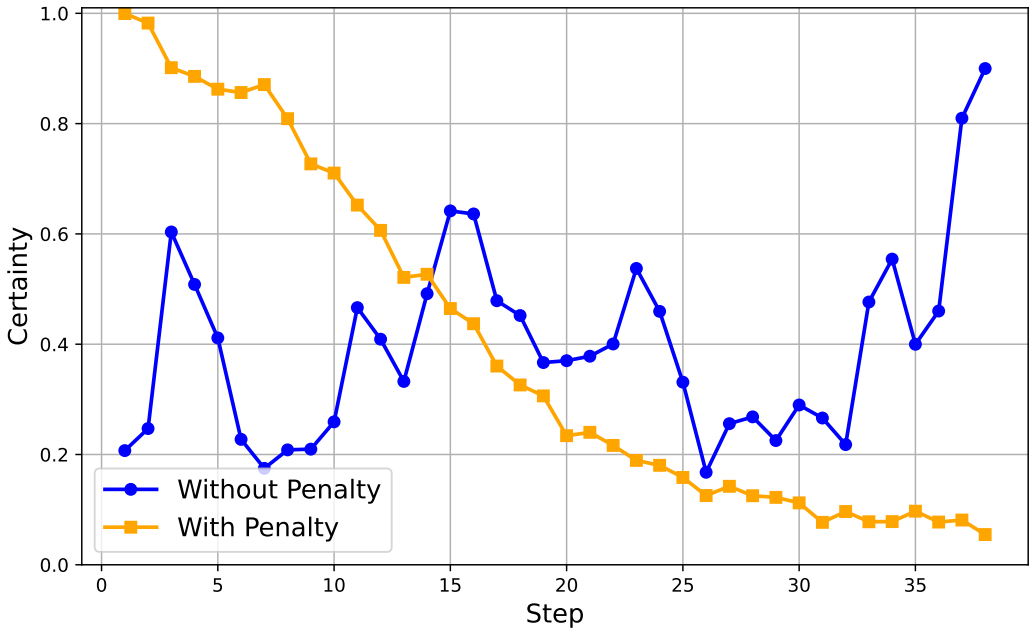


Figure 4: Query certainty trends over steps in Derm7pt dataset. IUAV-IP (Monte Carlo) with penalty (orange) begins with high-certainty queries and gradually transitions to more uncertain ones, while the version without penalty (blue) exhibits noisier and less structured query selection. This demonstrates the effectiveness of penalty-guided querying in prioritizing reliable information first.

To demonstrate that shorter query histories lead to better decisions, we conducted an experiment with the V-IP framework. We removed its stopping criterion, creating two variants: V-IP(NS), which uses all concepts, and UAV-IP(Ora-NS), which uses only correctly predicted concepts. As shown in Table 4, UAV-IP(Ora-NS), using only correctly classified concepts, achieves the best performance. However, this is unrealistic since concept labels are not available at inference. In contrast, our proposed method achieves performance close to UAV-IP(Ora-NS) without using any concept annotations. This highlights the benefits of selecting relevant subsets for decisions and doing so dynamically based on uncertainty.

6 Conclusion

In this work, we addressed a critical gap in interpretable AI for medical imaging by introducing uncertainty-aware extensions to the V-IP framework. We propose two models: EUAV-IP, which skips uncertain concepts, and the novel IUAV-IP, which integrates uncertainty estimates into the query selection process to enable more informed, cautious, and clinically aligned decision-making. By learning to prioritise reliable concepts and deferring uncertain ones, the model mirrors the diagnostic reasoning of medical professionals, balancing informativeness with trustworthiness. Through extensive evaluation across diverse medical imaging datasets, we demonstrated that IUAV-IP not only improves

predictive performance but also produces shorter, more meaningful explanations. These improvements are not merely technical, they are essential for real-world deployment in clinical settings, where interpretability, reliability, and efficiency are paramount. The ability to delay or avoid uncertain queries without sacrificing accuracy enhances the model's safety and transparency that are important for responsible use of AI in healthcare.

While our approach still relies on concept-level supervision during training, it lays the groundwork for future systems that can autonomously generate and reason over clinically relevant concepts using vision-language models. Ultimately, this work moves us closer to AI systems that can collaborate with clinicians, not just as accurate tools, but as interpretable and trustworthy partners in decision-making. On top of that guiding the querier to avoid unreliable queries requires access to concept-level ground truth during training, which may not always be feasible in practice. In future work, we aim to explore more flexible loss functions that can enable effective query selection without requiring explicit concept labels. Additionally, we plan to integrate IUAV-IP with VLMs and LLMs by enabling them to both generate and answer concept-based queries under uncertainty.

References

[1] Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[2] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *Information Flow in Deep Neural Networks*, page 24, 2022.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[4] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[5] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[6] Robert Geirhos, Roland S Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un) reliability of feature visualizations. In *Proceedings of the 41st International Conference on Machine Learning*, pages 15294–15330, 2024.

[7] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[8] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *International Conference on Machine Learning*, pages 16521–16540. PMLR, 2023.

[9] Yibo Gao, Zheyao Gao, Xin Gao, Yuanye Liu, Bomin Wang, and Xiahai Zhuang. Evidential concept embedding models: Towards reliable concept explanations for skin disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 308–317. Springer, 2024.

[10] Aditya Chattopadhyay, Stewart Slocum, Benjamin D Haeffele, Rene Vidal, and Donald Geman. Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7430–7443, 2022.

[11] Aditya Chattopadhyay, Kwan Ho Ryan Chan, Benjamin David Haeffele, Donald Geman, and Rene Vidal. Variational information pursuit for interpretable predictions. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Aditya Chattopadhyay, Kwan Ho Ryan Chan, and Rene Vidal. Bootstrapping variational information pursuit with large language and vision models for interpretable image classification. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Stefan Kolek, Aditya Chattopadhyay, Kwan Ho Ryan Chan, Hector Andrade-Loarca, Gitta Kutyniok, and René Vidal. Learning interpretable queries for explainable image classification with information pursuit. *arXiv preprint arXiv:2312.11548*, 2023.

[14] Yuyan Ge, Kwan Ho Ryan Chan, Pablo Messina, and René Vidal. Ip-crr: Information pursuit for interpretable classification of chest radiology reports. *arXiv preprint arXiv:2505.00191*, 2025.

[15] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[16] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14933–14943, 2021.

[17] Sercan O Arik and Tomas Pfister. Protoattend: Attention-based prototypical learning. *Journal of Machine Learning Research*, 21(210):1–35, 2020.

[18] Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. Comprehensible convolutional neural networks via guided concept learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[19] Cristiano Patrício, João C Neves, and Luis F Teixeira. Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3799–3808, 2023.

[20] Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N Balasubramanian. A framework for learning ante-hoc explainable models via concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10286–10295, 2022.

[21] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

[22] Tuomas P Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2023.

[23] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

[24] Theo MV Janssen and Barbara H Partee. Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier, 1997.

[25] Milad Mahinpei, Yahav Bechavod, and Mukund Srivastava. Promises and pitfalls of black-box concept learning models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 23061–23073, 2021.

[26] Yibo Gao, Hangqi Zhou, Zheyao Gao, Bomin Wang, Shangqi Gao, Sihan Wang, and Xiahai Zhuang. Learning concept-driven logical rules for interpretable and generalizable medical image classification. *arXiv preprint arXiv:2505.14049*, 2025.

[27] Donald Geman and Bruno Jedynak. An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):1–14, 1996.

[28] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[29] Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise aleatoric and epistemic uncertainty quantification. *arXiv preprint arXiv:2406.02354*, 2024.

[30] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

[33] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[34] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[35] Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1):148, 2024.

[36] Satoshi Tsutsui, Winnie Pang, and Bihan Wen. Wbcatt: a white blood cell dataset annotated with detailed morphological attributes. *Advances in Neural Information Processing Systems*, 36:50796–50824, 2023.

[37] Michael Nevitt, David Felson, and Gayle Lester. The osteoarthritis initiative. *Protocol for the cohort study*, 1(2), 2006.

[38] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.