INTERPRETABLE AI FOR ANIMAL HEALTH: BAYESIAN NETWORKS WITH UNCERTAINTY-AWARE DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Timely and trustworthy animal disease detection is critical for animal welfare, food security, and farm profitability, yet existing AI systems struggle with noisy sensors, class imbalance, and limited interpretability. We present *U-DeepBN*, an Uncertainty-Aware Bayesian Network with Deep Learning, that integrates deep feature learning, uncertainty quantification, and Bayesian reasoning for robust and explainable disease prediction. By modeling feature-level uncertainty and explicit dependencies, *U-DeepBN* enables intrinsic interpretability and transparent inference under uncertainty, explanation-targeted perturbations, and data poisoning, supported by theoretical analysis of consistency and robustness properties. Experiments on three real-world animal disease datasets and a cross-domain dataset show that *U-DeepBN* improves accuracy and robustness, achieves competitive or higher explanation fidelity and stability, and substantially reduces runtime compared to strong baselines. Ablation studies reveal that optimal uncertainty types and feature representations are dataset-dependent, with adaptive uncertainty modeling and hybrid features yielding the most reliable performance. These results position *U-DeepBN* as a practical, deployable decision-support tool for precision livestock monitoring.

1 INTRODUCTION

Challenges in Animal Health Monitoring. Monitoring farm animal health is critical for disease control, food security, and economic stability (Arafat et al., 2024; Ghadekar et al., 2024; World Organisation for Animal Health (WOAH), 2025). However, real-world deployment remains challenging due to noisy sensor data, class imbalance, domain shifts, and the limited interpretability of deep learning models (Khandaker et al., 2024). Black-box predictions reduce trust and hinder timely decision-making by farmers and veterinarians, particularly in early intervention scenarios. In addition, variations in animal breeds, housing environments, and sensor configurations can significantly affect model reliability outside controlled settings. These limitations motivate explainable, uncertainty-aware frameworks that provide reliable predictions and transparent reasoning.

Limitations of Existing XAI Methods. Explainability methods remain vulnerable to adversarial manipulation, where distorted inputs or attacks on interpretability mechanisms can compromise predictions and erode user trust (Ghorbani et al., 2019; Slack et al., 2020). Although existing XAI approaches improve transparency, many rely on static feature importance or post-hoc visualizations (Bedi et al., 2024; Khandaker et al., 2024), which often fail to capture uncertainty and variability in real-world environments. Such explanations may appear convincing while masking unreliable or unstable model behavior under distribution shifts and noisy observations. These limitations motivate the development of robust, uncertainty-aware explainability frameworks capable of providing reliable interpretations under noisy and adversarial conditions.

Our Proposal. To address these challenges, we introduce ***an integrated framework that combines uncertainty-aware deep feature extraction with Bayesian reasoning for decision-making***. Unlike Bayesian neural networks or post-hoc XAI, our approach separates representation learning from uncertainty-aware reasoning, allowing uncertainty to guide probabilistic inference over feature-level parameters (see Appendix A for an overview of the training pipeline). This design

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

enables interpretable and robust predictions under data variability and adversarial perturbations, with uncertainty-aware parameter estimation providing a principled foundation for trustworthy AI-assisted animal health monitoring. While the Bayesian network structure captures interpretable dependencies among variables, the uncertainty-weighting mechanism developed in this work applies specifically to parameter estimation rather than to structure learning; we do not claim or analyze an effect on edge formation or structure learning in this version.

Key Contributions: **(1) Novel uncertainty-aware BN framework.** We introduce *U-DeepBN*, a unified framework that models uncertainty at the feature level and propagates it through a Bayesian Network, enabling structured, uncertainty-aware explanations that go beyond prediction-level confidence or post-hoc attributions; **(2) Hybrid deep and classical feature integration.** We introduce a unified modeling approach that combines deep learning–based feature extraction with classical, interpretable image features within a Bayesian Network, supporting probabilistic inference and explicit analysis of feature contributions; **(3) Class-specific uncertainty modeling.** We demonstrate how modeling different types of uncertainty within the Bayesian Network enables more reliable, class-specific inference under data variability and adversarial perturbations; and **(4) Empirical validation on real-world datasets.** We validate *U-DeepBN* on multiple animal disease datasets and a cross-domain dataset, showing improvements in accuracy, robustness, and interpretability over strong baselines. Ablation studies show different types of uncertainty play complementary roles across datasets, while revealing a trade-off between predictive performance and interpretability across feature types.

2 Related Work

Explainable AI (XAI). Adversarial perturbations can distort feature attributions without changing model predictions, undermining trust in XAI systems (Ghorbani et al., 2019). This fragility is exacerbated by high-dimensional inputs and nonlinear models. Prior studies show that explanations strongly influence user trust, especially when sensitive features are involved (Lakkaraju & Bastani, 2020), and adversarial classifiers can mask biases in popular post-hoc methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) (Slack et al., 2020). As a result, many XAI approaches struggle to maintain interpretability and performance under adversarial conditions. Our framework addresses this gap by integrating deep learning with uncertainty-aware BN reasoning.

Recent defenses improve robustness via alternative strategies, including randomized explanations with local differential privacy (XRAND) (Nguyen et al., 2023), aggregation across methods (AGG-Mean) (Rieger & Hansen, 2020), and minimax optimization for adversarial robustness (Lakkaraju et al., 2020). Other approaches, such as Optimal Robust Explanations (OREs) (Malfa et al., 2021) and β-smoothing (Dombrowski et al., 2019), remain limited in cross-domain generalization. In contrast, our uncertainty-aware BN design supports interpretable reasoning across diverse data types.

AI-based Animal Disease Detection. AI-based disease detection shows promise but often underperforms in farm settings due to data imbalance, variability, and computational overhead. Ghadekar et al. (2024) combined YOLO and MobileNet for wildlife disease detection, achieving 91% accuracy but remaining sensitive to data variability. Arafat et al. (2024) used LoRa-based IoT sensors with subspace k-Nearest Neighbors (kNN) for cattle monitoring, facing deployment and data diversity challenges. Deo et al. (2024) applied YOLOv3 and EfficientNetB3 to poultry faecal images at notable computational cost, while Narayanan et al. (2020) used node2vec on simulated bovine metagenomes with limited realism. Our *U-DeepBN* addresses these limitations by modeling uncertainty and fusing multimodal evidence for robust, field-ready performance.

XAI improves model transparency but continues to face challenges related to data scarcity, subjective evaluation, and limited robustness. Prior work has applied XAI to animal health monitoring using wearable sensors (Shi et al., 2024), expert-based validation (Yiğit et al., 2022), post hoc explanations for lumpy skin disease (Jain et al., 2022), and transfer learning for aquaculture data imbalance (Isaac et al., 2024). Transformer-based (Ram Das et al., 2024) and CNN-based XAI models (Khandaker et al., 2024) improve transparency but remain constrained by dataset limitations and subjective interpretation, while post hoc explanations are often inconsistent (Bedi et al., 2024). Common XAI techniques such as Grad-CAM, Score-CAM, and Layer-CAM enhance trust but lack principled defenses or explicit uncertainty modeling. In contrast, *U-DeepBN* integrates

Bayesian reasoning with deep learning to jointly support interpretability and robustness under uncertainty. Unlike prediction-level uncertainty estimators such as MC-Dropout (Gal & Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017), and evidential deep learning (Sensoy et al., 2018), and unlike interpretable-by-design models such as prototype networks (Chen et al., 2019) and concept bottleneck models (Koh et al., 2020), *U-DeepBN* estimates uncertainty for individual input features and uses it to weight the parameters of an explicit probabilistic graph.

To the best of our knowledge, no prior work jointly addresses uncertainty, interpretability, and adversarial resilience in AI-based animal disease detection. *U-DeepBN* bridges this gap through a unified architecture that combines deep uncertainty modeling with explainable AI, providing a trustworthy and deployment-ready solution for animal health monitoring.

3 PRELIMINARIES

Bayesian Network (BN).

A BN is a probabilistic graphical model that represents conditional dependencies among variables using a directed acyclic graph (DAG) (Pearl, 1988). Its compact factorization of joint distributions supports probabilistic inference and uncertainty-aware decisions. In animal disease prediction, BNs combine health indicators such as temperature, behavior, and environmental factors in an interpretable framework. However, manually specified structures and conditional probability tables can limit their scalability and ability to capture complex, nonlinear patterns. Combining BNs with deep learning can address these limits through representation learning with probabilistic reasoning.

Uncertainty-Aware Learning and Reasoning. *Uncertainty quantification* (UQ) is central to reliable machine learning, particularly in safety- and decision-critical applications (Abdar et al., 2021). It characterizes prediction confidence by explicitly representing uncertainty from multiple sources. *Aleatoric uncertainty* captures irreducible data noise (e.g., noise or ambiguous observations), while *epistemic uncertainty* reflects uncertainty in model parameters or structure due to limited or biased training data and can be reduced with additional information. These components form *total uncertainty*, providing a holistic measure of prediction reliability (Hüllermeier & Waegeman, 2021). Subjective Logic (SL) (Jøsang, 2016) explicitly represents uncertainty in probabilistic beliefs, particularly for trust inference. Explicit modeling of these uncertainties helps distinguish reliable from unreliable predictions, improving robustness, interpretability, and reliability in the real world.

Effective decision-making under uncertainty requires principled mechanisms for representing and combining uncertain evidence. In this work, various uncertainty representations are integrated into the proposed *U-DeepBN* framework to model, propagate, and compare uncertainty sources in animal disease prediction, enabling analysis of their impact on predictive reliability and robustness. Unlike other uncertainty estimation methods, our approach models feature-level uncertainty using a single frozen encoder and does not require repeated inference or multiple model instances at test time.

Problem Statement. Accurate and interpretable animal disease prediction is essential for timely diagnosis and treatment planning. However, traditional machine learning models often struggle with uncertainty in noisy, imbalanced, or high-dimensional settings and typically function as black boxes, limiting trust and interpretability. To address these challenges, we propose an ***Uncertainty-Aware Bayesian Network with Deep Learning (U-DeepBN)***, which integrates XAI techniques for principled uncertainty quantification and interpretable inference. *U-DeepBN* fuses outputs from multiple predictive models to improve reliability and accuracy, while XAI components highlight feature contributions for veterinary decision-making. The system's objective is to maximize the posterior predictive probability under quantified uncertainty:

$$\hat{y} = \arg\max_{y} P(Y = y \mid X, \mathcal{M}, u), \tag{1}$$

where X and Y denote the input features and predicted outcomes, $\mathcal{M}$ represents the predictive model, and u captures quantified uncertainty used in decision-making. Here u is computed and fixed during training (Section 4.3) and absorbed into the learned parameters $\mathcal{M}$; it is not recomputed dynamically at inference time. By jointly modeling uncertainty and interpretability, *U-DeepBN* enhances predictive performance while fostering greater trust in AI-assisted animal health monitoring.

4 *U-DeepBN*: UNCERTAINTY-AWARE BN WITH DEEP FEATURES

We propose a hybrid framework integrating deep learning–based image representations with Bayesian Networks (BNs) under explicit uncertainty modeling. The core idea is to leverage the predictive strength of deep neural networks while preserving the interpretability, uncertainty reasoning, and structured dependency modeling of probabilistic graphical models. A key design choice is the joint use of deep and classical image features: deep features provide high-level semantic representations optimized for accuracy, whereas classical features computed from image pixels offer physically meaningful, interpretable cues. Unlike prior hybrids using BNs primarily as classifiers or ensemble components (Friedman et al., 1997; Chipman et al., 2006), *U-DeepBN* treats deep and classical features as uncertain evidence sources whose reliability shapes Bayesian inference. This dual-feature strategy enables flexible trade-offs between predictive performance and interpretability.

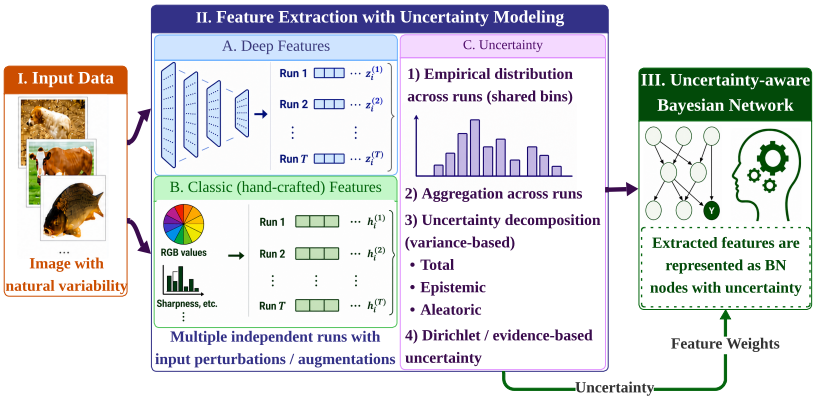

Figure 1: **Framework of *U-DeepBN*.** Illustration of the proposed pipeline, including feature extraction, uncertainty modeling, and uncertainty-aware learning.

The framework proceeds in three stages: **(1)** deep and classical feature extraction from images, **(2)** feature-level uncertainty quantification via stochastic feature modeling, and **(3)** uncertainty-aware Bayesian Network construction and training, where uncertain tabular representations are encoded as BN nodes to enable probabilistic inference and intrinsic interpretability, as shown in Figure 1.

4.1 DEEP AND CLASSICAL IMAGE FEATURE EXTRACTION

Deep Feature Extraction. We employ a pretrained convolutional neural network (CNN) as a fixed feature encoder without task-specific training or fine-tuning. Specifically, a ResNet architecture pretrained on ImageNet is used with the final classification layer removed, and all network parameters are frozen during feature extraction. This design reduces computational cost, memory usage, and training time, avoids reliance on large labeled datasets, and enables the reuse of extracted features across downstream models. Despite the absence of task-specific training, pretrained deep features retain strong representational capacity and provide high-level semantic information that complements classical features, enabling competitive predictive performance with improved efficiency.

Classical Image Features. In parallel, we extract a set of classical, hand-engineered image features directly from pixel intensities to capture low- and mid-level visual characteristics with clear physical interpretations. These features include **(1)** color statistics such as mean Red–Green–Blue (RGB) intensities and normalized color ratios to reduce sensitivity to illumination changes, **(2)** color dominance indicators (e.g., excess red indices) that quantify relative color prevalence, **(3)** hue distribution histograms computed in Hue–Saturation–Value (HSV) space to capture color diversity and variability, and **(4)** image sharpness measures based on the variance of the Laplacian operator as a proxy for blur and noise. Unlike deep features, these classical descriptors are directly interpretable and can be linked to image acquisition conditions and data reliability, supporting transparent and uncertainty-aware reasoning within the proposed framework.

4.2 FEATURE-LEVEL UNCERTAINTY QUANTIFICATION

To quantify uncertainty in extracted image features, we perform R stochastic feature-extraction runs, each applying stochastic preprocessing (perturbed forward passes through the frozen encoder) with a different random seed. Each run yields the same feature schema for the same samples, and we estimate per-feature uncertainty from variability across runs.

We adopt a variance-based decomposition that distinguishes aleatoric, epistemic, and total uncertainty via the law of total variance (Hüllermeier & Waegeman, 2021), which keeps all three quantities on the same scale and avoids distributional assumptions that deep features often violate. We use these terms by analogy: AU_f is between-sample variability, and EU_f is the representation instability of a frozen encoder, which is not reduced by collecting more data. We compute three feature-level uncertainty measures: **(1) Aleatoric uncertainty** captures the variability of feature f across input samples after averaging over stochastic feature-extraction effects, and therefore serves as a proxy for data-dependent or irreducible variability; **(2) Epistemic uncertainty** measures the variability of feature f across stochastic representations of the same input and serves as a proxy for representation instability and sensitivity to stochastic feature extraction; **(3) Total uncertainty** captures the overall dispersion of feature due to data-dependent variability and stochastic representation instability. Details are in Appendix A.1.

Beyond variance-based uncertainty, we consider a Dirichlet-strength formulation inspired by Subjective Logic (SL) (Jøsang, 2016) as an alternative feature-reliability measure. It combines **two dispersion measures**: **(1)** *entropy of the mean distribution*, capturing feature-value spread or concentration, and **(2)** *run-to-run disagreement*, measured by Jensen–Shannon divergence (Menéndez et al., 1997) between each run distribution and the mean. Details are in Appendix A.2.

4.3 UNCERTAINTY-AWARE BAYESIAN NETWORK LEARNING

In the BN, each extracted deep and classical feature is represented as a distinct node, with observed values treated as realizations of corresponding random variables during training. Feature-level uncertainty enables reliability-aware probabilistic modeling by modulating each feature's contribution according to its uncertainty, accounting for instability and variability in image-derived representations. The feature weight is $w_f = \exp(-\lambda u_f)$, where u_f is the normalized uncertainty score and λ controls sensitivity. Table 3's four *U-DeepBN* variants substitute different quantities for u_f: AU_f (Aleatoric), EU_f (Epistemic), TU_f (Total), and the Subjective-Logic uncertainty mass defined in Appendix A.2; each is normalized to $[0, 1]$ before applying $w_f = \exp(-\lambda u_f)$, so that $w_f \in [e^{-\lambda}, 1]$ (hyperparameters in Appendix B.3).

Rather than assuming uniform reliability across features, we adopt a weighted likelihood formulation that modulates each feature's contribution by its associated uncertainty. Specifically, for a feature node f conditioned on its parent variables $\mathrm{Pa}(f)$, the weighted log-likelihood is $\mathcal{L}_f(\theta_f) = \sum_{i=1}^{N} w_f \log p(x_i^{(f)} \mid \mathrm{Pa}(f); \theta_f)$, where $w_f \in [0, 1]$ is a feature-specific weight derived from uncertainty estimates. While constant scaling of the log-likelihood does not change the maximizer under standard Maximum Likelihood Estimation (MLE), this formulation admits a principled interpretation as a Maximum A Posteriori (MAP) estimator with feature-dependent regularization strength. This ensures that higher-uncertainty features contribute less to parameter estimation, reducing overconfident CPDs from unstable representations. At inference, the BN performs probabilistic reasoning over deep and classical features, producing posterior distributions over target variables while accounting for learned feature reliability. A summary of the training pipeline is shown in Appendix A.

Under distribution shift, noisy inputs, or feature ambiguity, down-weighting high-uncertainty features reduces reliance on unstable or spurious features, and the explicit graph lets users inspect feature contributions and their uncertainties to diagnose failures. As an example, Figure 2 illustrates a simplified Bayesian Network highlighting the model's interpretability. The graph contains five nodes, where intermediate variables (e.g., H_hist_02, RG_ratio) mediate the influence of upstream features (e.g., deep_001, lap_var) on the final prediction Y (Diagnosis). Each node represents a semantically meaningful image-derived feature: deep_001 captures high-level patterns, H_hist_02 encodes hue distribution, RG_ratio $= R_{\text{mean}}/G_{\text{mean}}$ measures color dominance, and lap_var

quantifies image sharpness.

The conditional probability table (CPT) explicitly defines how parent variables determine $P(Y = 1)$, while uncertainty-aware adjustment ($P_u = (1-u)P + u/|\mathcal{Y}|$) modulates predictions based on feature reliability. This combination enables transparent, traceable reasoning with uncertainty-adjusted confidence. Node-level uncertainties are aggregated across observed features. In practice, we use the mean uncertainty, $u = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} u_f$, where $\mathcal{F}$ denotes the set of observed feature nodes. To account for uncertainty, each node is associated with an uncertainty score u, and the final probabilities are adjusted via $P_u = (1-u)P + u/|\mathcal{Y}|$, where $|\mathcal{Y}|$ denotes the number of discrete states of the target variable Y. In this example, Y is binary, so $|\mathcal{Y}| = 2$, and $1/|\mathcal{Y}|$ corresponds to a uniform distribution. This adjustment smooths predictions toward an uninformative prior under high uncertainty, reducing confidence from unreliable inputs and providing an interpretable mechanism for uncertainty propagation.

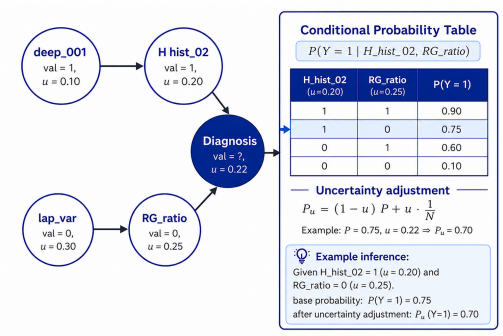

Figure 2: Example uncertainty-aware BN; u smooths $P(Y)$ toward uniform.

Theoretical analysis in Appendix B establishes consistency under reliable features. It shows that uncertainty-aware weighting suppresses noisy features for improved robustness, with weights derived from a principled rather than heuristic formulation.

5 Experimental Setup

Parameterization. All experiments use dataset-specific settings to accommodate differences in size, image quality, and class distribution. A pretrained ResNet-18 backbone is used for deep feature extraction across all experiments. For the Dog Skin Disease dataset, features are reduced to 96 dimensions and combined with classical features, with dependencies modeled using a Chow–Liu tree (Chow & Liu, 1968); features are discretized into 10 bins, a uniform prior with $\alpha = 2.0$ is applied, and two augmentation operations are used. For the Fish Disease and PlantDoc datasets, 128-dimensional features are used with a tree-augmented naive Bayes (TAN) (Friedman et al., 1997) structure, using 15 bins for Fish Disease and 3 bins for PlantDoc. For the Cow Lumpy Skin Disease dataset, a similar TAN structure is used with 128-dimensional features, retaining the top 16 features and discretizing them into 4 bins for the binary setting. To ensure fair comparison, the same ResNet-18 backbone is used for all baselines except Finer-CAM. All experiments follow official splits, with results averaged over multiple runs. Ablation studies use the same settings with targeted variations (Section 6). Additional details are provided in Appendix B.3.

Datasets. We evaluate our framework on three animal disease classification datasets and a cross-domain dataset varying in size, class balance, and disease characteristics: (1) **Cow Lumpy Skin Disease** (Agarwal, 2022), with approximately 800 labeled images for binary classification of healthy and LSD-affected cows; (2) **Dog Skin Disease** (Motiani, 2024), comprising over 4,000 multiclass images of healthy and diseased canine skin; (3) **Freshwater Fish Disease Aquaculture** (Biswas et al., 2024), containing 1,750 images across seven balanced classes spanning bacterial, fungal, viral, parasitic, and healthy fish conditions. and **(4) PlantDoc** (Singh et al., 2020), comprising 2,598 images across 13 plant species and up to 17 disease classes, supporting multi-class plant disease classification under real-world conditions. Because the method targets animal disease diagnosis, we evaluate it primarily on animal disease datasets rather than general-purpose benchmarks, and include PlantDoc to test framework-level generality. Given the scarcity of animal disease data, typically limited in size, this evaluation highlights the method's effectiveness under low-data conditions.

Comparison Schemes. To evaluate our method as an explainable AI approach for animal disease detection, we compare it against representative baselines spanning post-hoc explanation methods and inherently interpretable or causality-aware models, enabling assessment of explanation quality, uncertainty handling, and interpretability. Specifically, we consider: **(1) Integrated Gradients (IG)** (Sundararajan et al., 2017), a gradient-based attribution method that accumulates gradients from a baseline to the input; **(2) SHAP** (Lundberg & Lee, 2017), a game-theoretic approach as-

Table 1: **Model-level predictive performance and robustness.**

Model	Dataset	Acc ↑	Rbst ↑	RT ↓
ResNet-18	Cow Lumpy Skin	0.984 ± 0.005	0.934 ± 0.013	2129.120 ± 523.003
	Dog Skin Disease	0.921 ± 0.022	0.952 ± 0.058	11281.943 ± 2138.041
	Fish Disease	0.996 ± 0.002	0.910 ± 0.004	6109.104 ± 1240.931
	PlantDoc	0.706 ± 0.014	0.851 ± 0.030	1114.012 ± 85.190
Finer-CAM	Cow Lumpy Skin	0.982 ± 0.001	0.952 ± 0.006	3620.001 ± 48.102
	Dog Skin Disease	0.914 ± 0.003	0.928 ± 0.005	19818.397 ± 959.031
	Fish Disease	0.983 ± 0.003	0.956 ± 0.011	9900.093 ± 79.103
	PlantDoc	0.735 ± 0.007	0.913 ± 0.006	1830.623 ± 907.290
U-DeepBN (Ours)	Cow Lumpy Skin	$\mathbf{0.989 \pm 0.001}$	$\mathbf{0.960 \pm 0.001}$	$\mathbf{65.010 \pm 0.058}$
	Dog Skin Disease	$\mathbf{0.927 \pm 0.004}$	$\mathbf{0.972 \pm 0.003}$	$\mathbf{1108.804 \pm 105.719}$
	Fish Disease	$\mathbf{0.999 \pm 0.001}$	$\mathbf{1.000 \pm 0.001}$	$\mathbf{168.833 \pm 2.149}$
	PlantDoc	$\mathbf{0.852 \pm 0.001}$	$\mathbf{0.960 \pm 0.001}$	$\mathbf{538.582 \pm 46.916}$

signing feature importance via Shapley values; **(3) MIP-XAI** (Salih et al., 2024), an iterative proxy method that refines feature rankings by modeling dependencies; **(4) Causality-Aware CNN (CA-CNN)** (Streltsov et al., 2025), which leverages causality maps to weight features by inferred influence; and **(5) Finer-CAM** (Zhang et al., 2025), which suppresses shared features to highlight fine-grained class-discriminative regions. We focus on XAI baselines rather than purely predictive ones, as our primary goal is to propose an explainability framework. Since XAI methods build on predictive models, we also compare predictive performance. These baselines span widely used explanation techniques and interpretable models, allowing us to assess whether uncertainty-aware BN reasoning improves performance and transparency over existing approaches.

Metrics. We evaluate predictive performance and explanation quality using: **(1) Prediction Accuracy (Acc)**, the proportion of correct test predictions; **(2) Robustness (Rbst)**, performance degradation under compromised data, where higher values indicate greater robustness; **(3) Local Fidelity (LF)** (Ribeiro et al., 2016), measuring how well local explanations approximate the model; **(4) Explanation Stability (Sta)** (Alvarez-Melis & Jaakkola, 2018), assessing consistency across similar inputs, with lower values indicating higher robustness; and **(5) Runtime (RT)**, end-to-end execution time including preprocessing, training, and explanation. We report full pipeline runtime for fair comparison, as BN-based models jointly perform prediction and explanation. Runtime is reported in seconds. Metric formulas are in Appendix B.4.

Threat Model. We evaluate threats to explanations and predictive models. **(1) XAI-specific threats** include **(1a)** *fooling post hoc explanations* (Slack et al., 2020), which manipulates feature importance to conceal sensitive features, evaluated by **LF**; and **(1b)** *XAI adversarial perturbations* (Ghorbani et al., 2019), where small input perturbations alter interpretations while preserving predictions, evaluated by **Sta**. The latter is an explanation-oriented threat, distinct from predictor-targeted attacks such as PGD. **(2) Data-poisoning threats** (Chen et al., 2017; Biggio et al., 2011) include **(2a)** backdoor insertion and **(2b)** label flipping. **Acc and Rbst** assess predictive robustness to poisoning; **LF and Sta** (Table 2) assess explanation fidelity and stability of models trained on clean data. The explainer is trained on poisoned data, and perturbations are included when calculating these metrics.

6 Experimental Results & Analyses

Overall Performance and Explanation Quality. Tables 1 and 2 compare schemes across three animal disease datasets and one cross-domain dataset using accuracy, robustness (Rbst), local fidelity (LF), stability (Sta), and runtime (RT), where higher LF and lower Sta indicate better explanation quality and robustness. For runtime (RT), Table 1 reports the model-level execution time, including preprocessing, training, and prediction, whereas Table 2 reports the full-pipeline execution time, including preprocessing, training, prediction, and explanation generation. Overall, *U-DeepBN* achieves near-best performance, providing strong accuracy and robustness while maintaining high-fidelity, stable explanations with lower runtime. We evaluate each uncertainty type independently, with detailed comparisons provided in the ablation study. For *U-DeepBN*, Acc is the clean accuracy of the uncertainty variant with the highest clean accuracy among the four variants in Table 3. Rbst is the ratio of that same variant's compromised accuracy to its clean accuracy. Thus, both quantities are computed from a single uncertainty variant for each dataset; Table 3 reports all four variants.

On the Cow Lumpy Skin Disease dataset, *U-DeepBN* achieves the best accuracy, robustness, and runtime, with explanation fidelity comparable to SHAP (0.759 vs. 0.763) at roughly $40\times$ lower cost.

Table 2: **Explainer-level comparison of explanation quality, robustness, and runtime.**

Explainer / Method	Dataset	LF ↑	Sta ↓	RT ↓
IG	Cow Lumpy Skin	0.590 ± 0.000	0.289 ± 0.059	2131.013 ± 526.124
	Dog Skin Disease	0.321 ± 0.034	0.354 ± 0.041	11288.189 ± 2149.002
	Fish Disease	0.254 ± 0.083	0.617 ± 0.079	6112.790 ± 1241.091
	PlantDoc	0.522 ± 0.090	0.373 ± 0.049	1115.347 ± 82.349
SHAP	Cow Lumpy Skin	**0.763 ± 0.056**	**0.007 ± 0.011**	2697.680 ± 538.092
	Dog Skin Disease	0.866 ± 0.046	0.040 ± 0.007	11890.353 ± 2102.874
	Fish Disease	0.881 ± 0.038	0.019 ± 0.006	6660.395 ± 1223.412
	PlantDoc	0.141 ± 0.013	0.298 ± 0.070	1148.230 ± 75.000
MIP-XAI	Cow Lumpy Skin	0.373 ± 0.192	0.123 ± 0.010	2128.341 ± 531.006
	Dog Skin Disease	0.081 ± 0.061	0.263 ± 0.077	12013.256 ± 2139.329
	Fish Disease	0.039 ± 0.055	0.396 ± 0.107	6107.496 ± 1232.128
	PlantDoc	0.034 ± 0.032	0.398 ± 0.126	1143.612 ± 72.190
CA-CNN	Cow Lumpy Skin	0.230 ± 0.471	0.085 ± 0.033	2121.680 ± 524.120
	Dog Skin Disease	0.429 ± 0.330	0.071 ± 0.023	11278.614 ± 2112.120
	Fish Disease	0.046 ± 0.084	0.088 ± 0.006	6103.555 ± 1237.329
	PlantDoc	0.226 ± 0.298	0.413 ± 0.086	1116.819 ± 76.102
Finer-CAM	Cow Lumpy Skin	0.511 ± 0.044	0.928 ± 0.013	3622.540 ± 49.900
	Dog Skin Disease	0.845 ± 0.002	0.681 ± 0.143	19822.127 ± 957.983
	Fish Disease	0.717 ± 0.035	0.254 ± 0.073	9902.312 ± 78.182
	PlantDoc	0.555 ± 0.088	0.504 ± 0.218	1834.100 ± 905.010
U-DeepBN (Ours)	Cow Lumpy Skin	0.759 ± 0.001	0.030 ± 0.001	**65.130 ± 0.045**
	Dog Skin Disease	**0.919 ± 0.007**	**0.021 ± 0.019**	**1109.065 ± 103.600**
	Fish Disease	**0.932 ± 0.001**	**0.014 ± 0.001**	**169.235 ± 2.700**
	PlantDoc	**0.976 ± 0.001**	**0.261 ± 0.001**	**540.120 ± 48.230**

On the Dog Skin Disease dataset, *U-DeepBN* attains the highest robustness and LF, with accuracy comparable to ResNet-18 (0.927 vs. 0.921, within one standard deviation), while reducing computational cost for imbalanced multiclass classification. On the Fish Disease dataset, *U-DeepBN* attains near-perfect accuracy and the lowest reported Sta, confirming robustness on well-balanced multiclass data. On the PlantDoc dataset, the framework achieves the best performance across all metrics. For the cross-domain PlantDoc dataset, we evaluate framework-level generality rather than zero-shot transfer: the overall U-DeepBN architecture and learning pipeline are kept unchanged, while dataset-dependent parameters are re-parameterized for PlantDoc. Thus, the PlantDoc results demonstrate that the same framework can be applied effectively across domains under dataset-specific parameterization, rather than implying that a single model trained on the animal-disease datasets transfers directly to the plant-disease domain.

For fair comparison, runtime is reported at corresponding evaluation levels: model-level RT includes preprocessing, training, and prediction, while explainer-level RT also includes explanation generation. Our approach achieves significantly lower runtime than baselines, mainly from faster BN training, with feature extraction remaining the dominant cost. A trade-off exists between explanation fidelity and cost: high-fidelity methods such as SHAP incur much higher runtime from iterative or sampling-based procedures, limiting scalability.

Ablation Studies: Impact of Uncertainty Modeling and Feature Representations. We ablate uncertainty and feature types on all datasets under clean and compromised settings.

Table 3: **Ablation study on uncertainty modeling under clean and compromised data.**

Scheme	Cow Lumpy Skin Disease		Dog Skin Disease		Fish Disease		PlantDoc	
	Clean	**Compromised**	**Clean**	**Compromised**	**Clean**	**Compromised**	**Clean**	**Compromised**
No uncertainty	0.988 ± 0.001	0.913 ± 0.002	0.920 ± 0.004	0.886 ± 0.004	0.999 ± 0.001	0.996 ± 0.001	0.850 ± 0.001	0.818 ± 0.001
Epistemic	0.980 ± 0.001	0.944 ± 0.001	**0.927 ± 0.004**	0.901 ± 0.003	0.999 ± 0.001	**0.999 ± 0.001**	**0.852 ± 0.001**	0.818 ± 0.001
Aleatoric	**0.989 ± 0.001**	0.949 ± 0.001	0.920 ± 0.004	**0.907 ± 0.003**	0.999 ± 0.001	**0.999 ± 0.001**	0.843 ± 0.001	**0.819 ± 0.001**
Total	0.980 ± 0.001	0.944 ± 0.001	**0.927 ± 0.004**	0.901 ± 0.003	0.999 ± 0.001	**0.999 ± 0.001**	**0.852 ± 0.001**	0.818 ± 0.001
SL	0.984 ± 0.001	**0.961 ± 0.001**	0.918 ± 0.004	0.893 ± 0.004	0.999 ± 0.001	**0.999 ± 0.001**	0.851 ± 0.001	0.813 ± 0.001

Note: Classification accuracy of an unweighted BN and *U-DeepBN* variants. The best variant improves compromised accuracy on every dataset; the best uncertainty type varies by dataset.

(a) Effect of Various Uncertainty Types. We compare uncertainty types in *U-DeepBN* against a BN without uncertainty modeling (Table 3). On clean data, gains are small because the unweighted BN is already strong: the best variant improves accuracy by at most 0.7 points (Dog Skin, Epistemic/Total), and some variants fall slightly below the baseline (e.g., Epistemic on Cow Lumpy Skin, 0.980 vs. 0.988). Under compromised data (label flipping and backdoor attacks), the best variant improves over the unweighted BN by 4.8 (Cow Lumpy Skin, SL), 2.1 (Dog Skin, Aleatoric), 0.3 (Fish Disease, all variants), and 0.1 points (PlantDoc, Aleatoric), although not every variant

improves on every dataset (e.g., SL on PlantDoc, 0.813 vs. 0.818). Uncertainty weighting therefore helps most where poisoning hurts the unweighted BN most, and the best uncertainty type is dataset-dependent, motivating adaptive selection.

The Epistemic and Total rows in Table 3 have identical accuracy to three decimal places across all four datasets under Clean and Compromised settings. The table reports classification accuracy, not uncertainty magnitude: although the uncertainty estimates differ, their weights $w_f = \exp(-\lambda u_f)$ evidently lead to the same BN decisions at the reported precision. This means the weight differences did not change the highest-scoring BN class for these examples, and the identical accuracies in Table 3 are not an artifact of rounding.

Table 4: **Classification accuracy under clean and compromised data across datasets.**

Feature Type	Cow Lumpy Skin Disease		Dog Skin Disease		Fish Disease		PlantDoc	
	Clean	**Compromised**	**Clean**	**Compromised**	**Clean**	**Compromised**	**Clean**	**Compromised**
Deep	0.978 ± 0.001	0.790 ± 0.006	0.918 ± 0.004	0.890 ± 0.003	$\mathbf{0.999 \pm 0.001}$	$\mathbf{0.999 \pm 0.001}$	0.832 ± 0.006	0.806 ± 0.003
Classical	0.965 ± 0.001	0.925 ± 0.001	0.801 ± 0.005	0.730 ± 0.012	0.954 ± 0.001	0.924 ± 0.001	0.792 ± 0.001	0.781 ± 0.001
Hybrid	$\mathbf{0.989 \pm 0.001}$	$\mathbf{0.961 \pm 0.001}$	$\mathbf{0.927 \pm 0.004}$	$\mathbf{0.907 \pm 0.003}$	$\mathbf{0.999 \pm 0.001}$	$\mathbf{0.999 \pm 0.001}$	$\mathbf{0.852 \pm 0.001}$	$\mathbf{0.819 \pm 0.001}$

Note: For each dataset and setting, the Hybrid-Features value is the best of the four uncertainty variants in Table 3 (an oracle over variants), so Clean and Compromised values may come from different variants.

(b) Effect of Feature Types. Table 4 compares feature types under clean and compromised settings. Hybrid features achieve the highest (or tied-highest)accuracy across datasets, showing that combining pretrained deep and classical features provides complementary benefits for performance and robustness. Deep features perform well on clean data (especially Fish Disease) but degrade under corruption (notably Cow Lumpy Skin), while classical features have lower accuracy and are more sensitive to corruption, particularly on Dog Skin. From an interpretability perspective, classical features are human-understandable (e.g., color, texture), whereas deep features are less transparent. Nevertheless, the CPT-level relationships illustrated in Fig. 2, as described earlier, provide a concrete view of the probabilistic dependencies among learned features and their contributions to the prediction. Hybrid features offer a favorable trade-off, combining strong performance with improved interpretability, demonstrating *U-DeepBN*'s ability to balance accuracy, robustness, and interpretability.

7 Conclusion & Future Work

Summary of Key Contributions & Findings. This paper introduces *U-DeepBN*, an uncertainty-aware and interpretable framework for animal disease detection that integrates deep feature learning with BN-based reasoning. Unlike post-hoc methods, the BN provides intrinsic interpretability by modeling dependencies and propagating uncertainty during inference. Experiments on four real-world datasets show that *U-DeepBN* achieves the best robustness and runtime on all four datasets, with the best or comparable accuracy and explanation fidelity. Results further demonstrate that uncertainty-aware feature weighting improves reliability under data poisoning, and that adaptive uncertainty modeling with hybrid feature representations yields the most robust and generalizable performance across datasets.

Limitations & Future Work. Evaluation is limited to small- and medium-scale benchmarks by the scarcity of annotated animal disease data; scaling BNs to high-dimensional features may need structural constraints; and current explanation metrics do not capture clinical relevance. Future work includes expert-in-the-loop evaluation with veterinarians, user studies, and multi-modal (image, sensor, clinical) extensions. Our evaluation also has specific gaps: *U-DeepBN* uses a frozen encoder whereas the baselines are fine-tuned; we do not compare against UQ baselines (MC-Dropout, deep ensembles, evidential deep learning), discriminative heads on the same features, or alternative BN topologies; headline *U-DeepBN* results are an oracle over uncertainty variants; and Sta normalization is not yet uniform across methods. Appendix D details these limitations.

STATEMENTS

AI USE STATEMENT

In this work, we used generative AI tools solely for proofreading and improving the grammar, clarity, and readability of the manuscript. We did not use generative AI to generate research ideas, conduct the literature review, develop the methodology, analyze data, produce results, or formulate scientific claims. Any other disclosure tasks do not apply to this work. All AI-assisted edits were carefully reviewed by the authors to ensure that they preserved the intended meaning and did not introduce inaccurate information. We take full responsibility for the final content of this work, including any text produced or revised with the aid of generative AI.

BROADER IMPACT STATEMENT

The proposed framework can improve early animal disease detection, supporting better animal welfare, food security, and decision-making through interpretable and uncertainty-aware predictions. However, risks include over-reliance on automated outputs, potential bias from unrepresentative data, and performance degradation under real-world variability. Responsible deployment requires human oversight, validation, and careful data management.

REPRODUCIBILITY STATEMENT

Implementation details are provided in Appendices A.1–A.2 (uncertainty computation), B.3 (parameterization and hyperparameters), and C (threat model). An anonymized code repository is available at `https://anonymous.4open.science/r/U-DeepBN-801C/README.md` for reproducibility.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Shivam Agarwal. Cow lumpy disease dataset. `https://www.kaggle.com/datasets/shivamagarwal29/cow-lumpy-disease-dataset`, 2022. Accessed: 2025-07-10.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Sheik Arafat, C. Priya, R. Premkumar, and S.M. Haji Nishath. Real time cattle health monitoring and early disease detection using iot and machine learning. In *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 450–455, 2024. doi: 10.1109/ICESC60852.2024.10690107.

Punam Bedi, Anjali Thukral, and Shivani Dhiman. Explainable ai in disease diagnosis. In *Explainable AI in Health Informatics*, Computational Intelligence Methods and Applications (CIMA), pp. 87–111. Springer, Singapore, July 8 2024. doi: 10.1007/978-981-97-3705-5_5.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In *Asian conference on machine learning*, pp. 97–112. PMLR, 2011.

Subir Biswas, Debendra Muduli, Md Ariful Islam, Anuradha Shantanu Kanade, Abu Taha Zamani, Shantanu Pandurang Kanade, and Nikhat Parveen. Empirical evaluation of deep learning techniques for fish disease detection in aquaculture systems: A transfer learning and fusion-based approach. *IEEE Access*, 12:176136–176154, 2024. doi: 10.1109/ACCESS.2024.3504283.

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.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Hugh Chipman, Edward George, and Robert McCulloch. Bayesian ensemble learning. *Advances in neural information processing systems*, 19, 2006.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

Ninad Deo, Aagam Bakliwal, Amit D. Joshi, and Suraj T. Sawant. Non-intrusive detection of poultry diseases through faecal analysis using deep learning. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pp. 1–6, 2024. doi: 10.1109/ICITIIT61487.2024.10580821.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bb836c01cdc9120a9c984c525e4b1a4a-Paper.pdf.

Nir Friedman, Dan Geiger, and Moisés Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2–3):131–163, 1997.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Premanand P. Ghadekar, Yash Pramod Gaikwad, Ajinkya Ghansham Ghusarkar, Girish Anil Nikam, Sourabh Bholanath Shintre, and Atharva Rajesh Thokal. Early disease detection in wildlife by analysing behavior and weather conditions using image processing and deep learning. In *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pp. 1–6, 2024. doi: 10.1109/ESCI59607.2024.10497217.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33013681. URL https://doi.org/10.1609/aaai.v33i01.33013681.

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.

Ssekitto Isaac, Oyoka Daniel, Mwebembezi Grivin, Jonah Mubuuke Kyagaba, Emmanuel Lule, and Ggaliwango Marvin. Explainable machine vision techniques for fish disease detection with deep transfer learning. In *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1218–1227. IEEE, 2024.

Anuj Kumar Jain, Raj Gaurang Tiwari, Neha Ujjwal, and Anshbir Singh. An explainable machine learning model for lumpy skin disease occurrence detection. In *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 503–508. IEEE, 2022.

Audun Jøsang. *Subjective Logic*, volume 3. Springer, 2016.

Md. Arafat Alam Khandaker, Ziyan Shirin Raha, Shifat Islam, and Tashreef Muhammad. Explainable ai-enhanced deep learning for pumpkin leaf disease detection: A comparative analysis of cnn architectures. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, December 20–22 2024.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020.

Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 79–85, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375833. URL https://doi.org/10.1145/3375627.3375833.

Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *International conference on machine learning*, pp. 5628–5638. PMLR, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Emanuele La Malfa, Agnieszka M Zbrzezny, Rhiannon Michelmore, Nicola Paoletti, and Marta Z. Kwiatkowska. On guaranteed optimal robust explanations for nlp models. In *International Joint Conference on Artificial Intelligence*, 2021. URL https://api.semanticscholar.org/CorpusID:234336718.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

Yash Motiani. Dogs skin disease dataset. https://www.kaggle.com/datasets/yashmotiani/dogs-skin-disease-dataset, 2024. Accessed: 2025-07-10.

Sai Narayanan, Akhilesh Ramachandran, Sathyanarayanan N. Aakur, and Arunkumar Bagavathi. Genome sequence classification for animal diagnostics with graph representations and deep neural networks, 2020. URL https://arxiv.org/abs/2007.12791.

Truc Nguyen, Phung Lai, Hai Phan, and My T Thai. Xrand: Differentially private defense against explanation-guided attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11873–11881, 2023.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988. ISBN 978-1558604797. URL https://books.google.com/books?id=AvNID7LyMusC.

Athish Ram Das, Nisha Pillai, Bindu Nanduri, Michael J Rothrock Jr, and Mahalingam Ramkumar. Exploring pathogen presence prediction in pastured poultry farms through transformer-based models and attention mechanism explainability. *Microorganisms*, 12(7):1274, 2024.

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*, pp. 1135–1144, 2016.

Laura Rieger and Lars Kai Hansen. A simple defense against adversarial attacks on heatmap explanations. In *5th Annual Workshop on Human Interpretability in Machine Learning*, 2020.

Ahmed M Salih, Ilaria Boscolo Galazzo, Zahra Raisi-Estabragh, Steffen E Petersen, Gloria Menegaz, and Petia Radeva. Characterizing the contribution of dependent features in xai methods. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6466–6473, 2024.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhonghao Shi, Fangyuan Chang, Yuanzheng Jia, Jun Li, Yawei Qiu, Jinfeng Miao, Wei Jiang, Xiaojun Guo, Xiangan Han, and Wei Tang. Classifying and understanding of dairy cattle health using wearable inertial sensors with random forest and explainable artificial intelligence. *IEEE Sensors Letters*, 2024.

Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, pp. 249–253, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450377386. doi: 10.1145/3371158.3371196. URL https://doi.org/10.1145/3371158.3371196.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

Oleksii Streltsov, Vagan Terziyan, and Oleksandra Vitko. Discover and explore weak causality and causal disposition in images for smart manufacturing tasks. *Procedia Computer Science*, 253:187–198, 2025. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2025.01.082. URL https://www.sciencedirect.com/science/article/pii/S1877050925000900. 6th International Conference on Industry 4.0 and Smart Manufacturing.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

World Organisation for Animal Health (WOAH). State of the World's Animal Health 2025. Technical report, World Organisation for Animal Health (WOAH), 2025.

Tuncay Yiğit, Nilgün Şengöz, Özlem Özmen, Jude Hemanth, and Ali Hakan Işık. Diagnosis of paratuberculosis in histopathological images based on explainable artificial intelligence and deep learning. *Traitement du Signal*, 39(3):863–869, 2022.

Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-CAM: Spotting the difference reveals finer details for visual explanation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9611–9620, 2025.

A Illustrative Overview of the Uncertainty-Aware Bayesian Network Training Pipeline

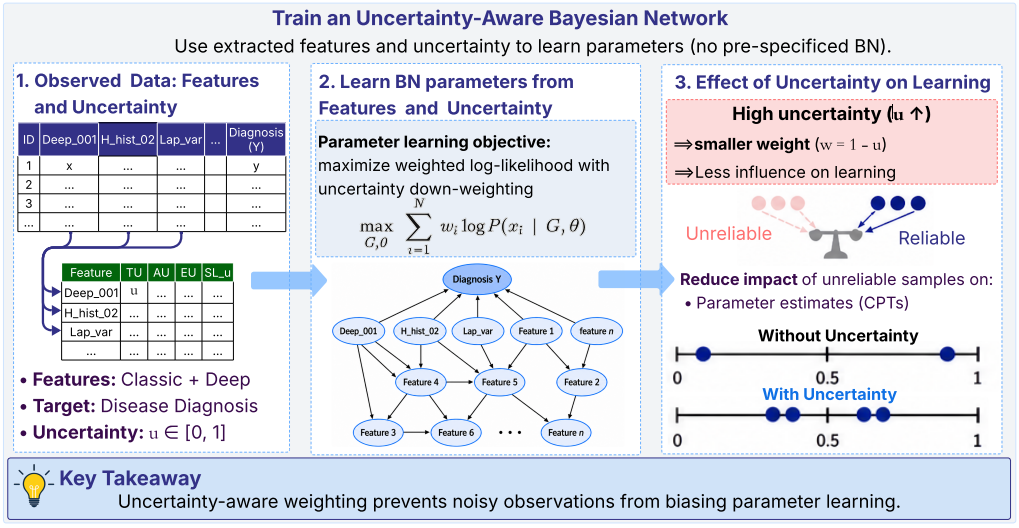

Figure 3: Training pipeline for an uncertainty-aware Bayesian network

Figure 3 provides a visual overview of the proposed uncertainty-aware Bayesian Network (BN) training pipeline. The framework consists of three stages: (1) feature extraction and uncertainty estimation, where deep and classical features are represented as BN nodes with associated uncertainty scores; (2) structure learning, where dependencies among features and the target variable are learned with a standard (unweighted) Chow–Liu or TAN procedure; and (3) uncertainty-aware parameter learning, where feature contributions are modulated by uncertainty to reduce the influence of unreliable observations.

As illustrated, high-uncertainty features receive lower weights, limiting their impact on both parameter estimation (CPTs) and the inference stage. This mechanism improves robustness by preventing noisy or unstable features from dominating the learning process. The figure highlights how uncertainty is propagated from feature-level estimation to probabilistic reasoning, enabling reliable and interpretable inference. Full methodological details are provided in the main paper.

A.1 Feature-Level Uncertainty Quantification

Let $z_{i,f}^{(r)}$ denote the value of feature f for input sample x_i obtained in the r-th stochastic feature-extraction run, where $i = 1, \ldots, N$, $r = 1, \ldots, R$, N is the number of input samples, and R is the number of stochastic runs. We compute three feature-level uncertainty measures as follows.

(1) Aleatoric uncertainty is defined as

$$AU_f = \frac{1}{N-1} \sum_{i=1}^{N} \left(\bar{z}_{i,f} - \bar{z}_f\right)^2, \qquad \bar{z}_{i,f} = \frac{1}{R} \sum_{r=1}^{R} z_{i,f}^{(r)}, \qquad \bar{z}_f = \frac{1}{N} \sum_{i=1}^{N} \bar{z}_{i,f}. \tag{A.1}$$

It captures the variability of feature f across input samples after averaging over stochastic feature-extraction effects, and therefore serves as a proxy for data-dependent or irreducible variability. Features with high aleatoric uncertainty exhibit greater variation across samples, potentially reflecting image ambiguity, noise, or data heterogeneity. When class labels are available, this quantity is computed on class-conditional subsets to avoid conflating between-class differences with within-class data variability.

(2) Epistemic uncertainty is defined as $EU_f = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{R-1} \sum_{r=1}^{R} \left(z_{i,f}^{(r)} - \bar{z}_{i,f} \right)^2 \right]$, which measures the variability of feature f across stochastic representations of the same input and serves as

a proxy for representation instability and sensitivity to stochastic feature extraction. To estimate feature-level epistemic uncertainty, we introduce controlled stochasticity into the otherwise deterministic feature-extraction pipeline. Although the encoder is frozen, multiple perturbed forward passes induce variability reflecting representational sensitivity. This multi-pass procedure is used only for offline uncertainty estimation and BN training (feature preparation); test-time inference uses a single deterministic pass through the frozen encoder, preserving the "no repeated inference or multiple model instances at test time" property stated in the Preliminaries. High epistemic uncertainty indicates features unstable across perturbations and potentially unreliable or corruption-sensitive, whereas low uncertainty indicates more consistent representations.

(3) Total uncertainty is defined according to the law of total variance as $TU_f = AU_f + EU_f$. It captures the overall dispersion of feature f due to data-dependent variability and stochastic representation instability. This variance-based formulation places aleatoric, epistemic, and total uncertainty on the same scale, avoiding mixed variance- and entropy-based quantities. When labels are available, computations are performed class-conditionally, yielding class-specific uncertainty profiles for feature reliability and class-dependent BN parameterization or weighting.

A.2 Subjective Logic-based Uncertainty

We describe the construction of the subjective logic-based uncertainty measure derived from feature distribution statistics.

These quantities are used to define an effective evidence strength:

$$c_f = N_{\text{eff}} \cdot (1 - \hat{H}_f) \cdot \exp(-\gamma JS_f)\,, \tag{A.2}$$

where $\hat{H}_f$ denotes normalized entropy ($\hat{H}_f = H(\hat{\mathbf{p}}_f)/\log K$, i.e., the entropy of the mean distribution divided by its maximum possible value $\log K$ for K histogram bins), JS_f is the average Jensen–Shannon divergence across runs, γ is a scaling parameter, and N_{eff} represents an effective sample size. This formulation ensures that evidence strength decreases when feature distributions are diffuse or inconsistent across runs. Using the resulting evidence strength, we construct Dirichlet parameters as

$$\alpha_f = \mathbf{1} + c_f \hat{\mathbf{p}}_f, \tag{A.3}$$

where $\hat{\mathbf{p}}_f$ denotes the empirical mean distribution obtained by averaging normalized histograms across runs. From these parameters, we compute a normalized uncertainty mass

$$u_f = \frac{K}{\sum_{k=1}^{K} \alpha_{f,k}}, \tag{A.4}$$

where K is the number of histogram bins. The quantity $u_f \in (0, 1]$ increases when evidence is weak (low concentration or high disagreement) and decreases when evidence is strong and consistent.

Importantly, u_f is interpreted as a feature reliability indicator rather than a predictive or belief-theoretic uncertainty. It quantifies the degree of evidence insufficiency associated with feature f, reflecting how confidently the feature representation can be used in downstream probabilistic modeling.

B Theoretical Analysis of Uncertainty-Aware Bayesian Learning

In this section, we provide theoretical justification for the proposed *U-DeepBN*, focusing on the statistical properties of the weighted likelihood formulation introduced in Section 4.3. In particular, we analyze the consistency of parameter estimation under feature-dependent uncertainty weighting.

B.1 Consistency of Weighted Likelihood under Reliable Features

Although weighted maximum likelihood estimation is well studied in statistical learning theory, we include the following result to show that incorporating uncertainty-based feature weights preserves consistency within the proposed *U-DeepBN* framework.

Let f denote a feature node in the Bayesian Network, and let $x_1^{(f)}, \ldots, x_N^{(f)}$ be i.i.d. samples drawn from the true conditional distribution $p^*(x^{(f)} \mid Pa(f))$. The weighted log-likelihood is defined as:

$$\mathcal{L}_N(\theta_f) = \sum_{i=1}^{N} w_{f,N} \log p(x_i^{(f)} \mid Pa(f); \theta_f), \tag{A.5}$$

where $w_{f,N} \in [0, 1]$ is a feature-specific weight derived from uncertainty estimates, with its dependence on the sample size N made explicit.

Theorem 1 (Consistency under reliable features). Assume that:

1. The data are i.i.d. samples from $p^*(x \mid Pa(f))$,
2. The model $p(x \mid Pa(f); \theta_f)$ is correctly specified,
3. The prior $p(\theta_f)$ is continuous and assigns nonzero mass near θ_f^*,
4. The feature weight converges to a constant $w_{f,N} \xrightarrow{p} w_f^* > 0$.
5. Standard regularity and identifiability conditions for consistency of the corresponding M-estimator hold

Then, the uncertainty-aware MAP estimator:

$$\hat{\theta}_f = \arg\max_{\theta_f} \left(\sum_{i=1}^{N} w_{f,N} \log p(x_i^{(f)} \mid \mathrm{Pa}(f); \theta_f) + \log p(\theta_f) \right) \tag{A.6}$$

is a consistent estimator of the true parameter θ_f^*. i.e.,

$$\hat{\theta}_f \xrightarrow{p} \theta_f^* \quad \text{as } N \to \infty. \tag{A.7}$$

Proof. Consider the normalized MAP objective:

$$Q_N(\theta_f) = \frac{1}{N} \sum_{i=1}^{N} w_{f,N} \log p\left(x_i^{(f)} \mid \mathrm{Pa}(f); \theta_f\right) + \frac{1}{N} \log p(\theta_f), \tag{A.8}$$

where $w_{f,N}$ converges in probability to $w_f^* > 0$, i.e.,

$$w_{f,N} \xrightarrow{p} w_f^* > 0. \tag{A.9}$$

Under the i.i.d. assumption and standard regularity conditions, the uniform law of large numbers implies that the empirical log-likelihood converges uniformly in probability to its population counterpart:

$$\frac{1}{N} \sum_{i=1}^{N} \log p\left(x_i^{(f)} \mid \mathrm{Pa}(f); \theta_f\right) \xrightarrow{p} \mathbb{E}_{p^*}\left[\log p\left(X^{(f)} \mid \mathrm{Pa}(f); \theta_f\right)\right]. \tag{A.10}$$

Moreover, the contribution of the prior becomes asymptotically negligible because

$$\frac{1}{N} \log p(\theta_f) \to 0 \quad \text{as } N \to \infty. \tag{A.11}$$

Therefore, together with $w_{f,N} \to w_f^* > 0$, the normalized MAP objective converges to the limiting population objective

$$Q(\theta_f) = w_f^* \mathbb{E}_{p^*}\left[\log p\left(X^{(f)} \mid \mathrm{Pa}(f); \theta_f\right)\right]. \tag{A.12}$$

Since $w_f^* > 0$, multiplication by w_f^* does not change the maximizer of the population log-likelihood. Under correct model specification, the population log-likelihood is maximized at the true parameter θ_f^*. Specifically, the difference between the expected log-likelihood at θ_f^* and that at any candidate θ_f can be expressed as

$$\begin{aligned} &\mathbb{E}_{p^*}\left[\log p\left(X^{(f)} \mid \mathrm{Pa}(f); \theta_f^*\right)\right] \\ &\quad - \mathbb{E}_{p^*}\left[\log p\left(X^{(f)} \mid \mathrm{Pa}(f); \theta_f\right)\right] \\ &= D_{\mathrm{KL}}\left(p^* \,\|\, p_{\theta_f}\right) \geq 0. \end{aligned} \tag{A.13}$$

Under the identifiability assumption, equality holds only when $\theta_f = \theta_f^*$. Therefore, θ_f^* is the unique maximizer of the limiting population objective. Since the normalized empirical MAP objective converges uniformly in probability to the limiting population objective, and the latter has a unique maximizer at θ_f^*, the standard argmax consistency result for M-estimators implies that

$$\hat{\theta}_f \xrightarrow{p} \theta_f^* \quad \text{as } N \to \infty. \tag{A.14}$$

Thus, for reliable features whose uncertainty-based weights converge to a strictly positive constant, the proposed uncertainty-aware weighting mechanism preserves the asymptotic consistency of the MAP estimator. The empirical results in Section 6 provide complementary finite-sample evidence supporting this asymptotic result.

B.2 Robustness to Noisy and Uncertain Features

A key motivation of *U-DeepBN* is to reduce the influence of unreliable features during learning and inference. We formalize this property by quantifying the deviation of the uncertainty-weighted MAP estimator from its prior mode as a function of the feature weight.

Theorem 2 (Prior domination and finite-sample shrinkage for uncertain features). For a feature f, define

$$\ell_{f,N}(\theta_f) = \sum_{i=1}^{N} \log p\Big(x_i^{(f)} \mid \mathrm{Pa}(f); \theta_f\Big), \tag{A.15}$$

and let the uncertainty-aware MAP estimator be

$$\widehat{\theta}_f(w_f) = \arg\max_{\theta_f} \{w_f \ell_{f,N}(\theta_f) + \log p(\theta_f)\}. \tag{A.16}$$

Let

$$\theta_f^0 = \arg\max_{\theta_f} \log p(\theta_f) \tag{A.17}$$

denote the prior mode. Assume that:

1. $\log p(\theta_f)$ is differentiable and μ_f-strongly concave on the parameter domain, with $\mu_f > 0$;
2. $\ell_{f,N}(\theta_f)$ is differentiable and concave; and
3. its score is uniformly bounded as

$$\sup_{\theta_f} \left\| \frac{1}{N} \nabla_{\theta_f} \ell_{f,N}(\theta_f) \right\| \le G_f \tag{A.18}$$

for some finite G_f.

Then the weighted-MAP estimator satisfies

$$\left\| \widehat{\theta}_f(w_f) - \theta_f^0 \right\| \le \frac{w_f N G_f}{\mu_f}. \tag{A.19}$$

Consequently,

$$\widehat{\theta}_f(w_f) \to \theta_f^0 \qquad \text{as } w_f \to 0 \tag{A.20}$$

for fixed N. Thus, the influence of feature f's likelihood on its parameter estimate decreases at least linearly with $w_f N$. In this sense, $w_f N$ controls the effective strength of the data relative to the prior.

Proof. For notational simplicity, write

$$Q_f(\theta; w) = w \ell_{f,N}(\theta) + \log p(\theta). \tag{A.21}$$

By concavity of the likelihood and strong concavity of the log-prior, $Q_f(\theta; w)$ is strongly concave and therefore has a unique maximizer $\widehat{\theta}_f(w)$. Its first-order optimality condition is

$$w \nabla \ell_{f,N}\Big(\widehat{\theta}_f(w)\Big) + \nabla \log p\Big(\widehat{\theta}_f(w)\Big) = 0. \tag{A.22}$$

Since θ_f^0 is the prior mode,

$$\nabla \log p(\theta_f^0) = 0. \tag{A.23}$$

Therefore,

$$\nabla \log p\Big(\widehat{\theta}_f(w)\Big) - \nabla \log p(\theta_f^0) = -w \nabla \ell_{f,N}\Big(\widehat{\theta}_f(w)\Big). \tag{A.24}$$

Because $\log p(\theta)$ is μ_f-strongly concave, its gradient satisfies

$$\Big\langle \nabla \log p(\widehat{\theta}_f(w)) - \nabla \log p(\theta_f^0), \widehat{\theta}_f(w) - \theta_f^0 \Big\rangle \le -\mu_f \left\| \widehat{\theta}_f(w) - \theta_f^0 \right\|^2. \tag{A.25}$$

Substituting the first-order condition and applying the Cauchy–Schwarz inequality gives

$$\mu_f \left\| \widehat{\theta}_f(w) - \theta_f^0 \right\|^2 \le w \left\langle \nabla \ell_{f,N}\Big(\widehat{\theta}_f(w)\Big), \widehat{\theta}_f(w) - \theta_f^0 \right\rangle \tag{A.26}$$

$$\le w \left\| \nabla \ell_{f,N}\Big(\widehat{\theta}_f(w)\Big) \right\| \left\| \widehat{\theta}_f(w) - \theta_f^0 \right\|. \tag{A.27}$$

If $\widehat{\theta}_f(w) \neq \theta_f^0$, division by $\|\widehat{\theta}_f(w) - \theta_f^0\|$ yields

$$\left\| \widehat{\theta}_f(w) - \theta_f^0 \right\| \le \frac{w}{\mu_f} \left\| \nabla \ell_{f,N}\Big(\widehat{\theta}_f(w)\Big) \right\|. \tag{A.28}$$

Using the bounded-score assumption,

$$\left\| \nabla \ell_{f,N}\Big(\widehat{\theta}_f(w)\Big) \right\| \le N G_f, \tag{A.29}$$

and hence

$$\boxed{\left\| \widehat{\theta}_f(w) - \theta_f^0 \right\| \le \frac{w N G_f}{\mu_f}}. \tag{A.30}$$

The same inequality holds trivially when $\widehat{\theta}_f(w) = \theta_f^0$. For fixed N, the right-hand side converges to zero linearly in w, proving

$$\widehat{\theta}_f(w) \to \theta_f^0 \qquad \text{as } w \to 0. \tag{A.31}$$

This completes the proof.

Remark (categorical CPTs on the simplex). For discretized nodes, θ_f lies on a (product of) probability simplices $\mathcal{C}$. The argument above carries over by replacing the first-order conditions with their constrained forms, $\langle w \nabla \ell_{f,N}(\widehat{\theta}_f) + \nabla \log p(\widehat{\theta}_f),\ \theta_f^0 - \widehat{\theta}_f \rangle \le 0$ and $\langle \nabla \log p(\theta_f^0),\ \widehat{\theta}_f - \theta_f^0 \rangle \le 0$; adding them and applying strong concavity gives the same bound. The bounded-score assumption requires restricting to $\mathcal{C}_\varepsilon = \{\theta : \theta_k \ge \varepsilon\}$ (with $\varepsilon < 1/K$), on which $G_f \le 1/\varepsilon$. For a Dirichlet($\boldsymbol{\alpha}$) prior with $\alpha_k > 1$, $\log p$ is μ_f-strongly concave with $\mu_f = \min_k(\alpha_k - 1)$. With the uniform prior $\alpha = 2$ used in our experiments, $\mu_f = 1$ and the bound becomes $\|\widehat{\theta}_f(w_f) - \theta_f^0\| \le w_f N/\varepsilon$.

Interpretation. The theorem provides a finite-sample characterization of the robustness induced by uncertainty weighting. Let the feature set be partitioned into reliable features $\mathcal{F}_r$ and noisy or uncertain features $\mathcal{F}_n$, where

$$w_f \approx 1, \quad f \in \mathcal{F}_r, \qquad w_f \ll 1, \quad f \in \mathcal{F}_n. \tag{A.32}$$

For reliable features, the likelihood contribution remains substantial and the corresponding estimates can be primarily data-driven. For uncertain features, the preceding bound explicitly constrains the displacement of the MAP estimate from the prior mode by a quantity proportional to $w_f N$. Consequently, decreasing w_f limits how strongly an unreliable feature can move its parameter estimate away from the prior-preferred value.

The quantity $w_f N$ should be interpreted as the scaling of the aggregate likelihood contribution, or as an effective likelihood strength. Multiplying the log-likelihood of N observations by w_f gives it the same aggregate scaling as a likelihood containing $w_f N$ equally informative contributions; however, this interpretation does not require $w_f N$ to be an integer or imply exact equivalence to observing $w_f N$ independent samples in arbitrary probabilistic models.

For the fixed-sample result above, $w_f \to 0$ implies prior domination. If both the sample size and the feature weight vary, i.e., $N \to \infty$ and $w_f = w_{f,N}$, a sufficient condition for the same bound to force convergence to the prior mode is

$$w_{f,N} N \to 0. \tag{A.33}$$

Thus, a small weight alone need not guarantee prior domination when the amount of data grows without bound; the relevant quantity is the likelihood scaling $w_{f,N} N$.

Discussion. This result provides theoretical support for the robustness mechanism used in *U-DeepBN*. When the uncertainty estimation mechanism assigns lower weights to corrupted or unreliable features, their likelihood contributions are suppressed and their learned parameters are regularized toward the prior. The theorem therefore establishes robustness conditional on appropriate uncertainty weighting. It does not, by itself, imply that every adversarial perturbation will necessarily be detected or assigned a small weight; such a claim would additionally require assumptions linking perturbation magnitude or feature corruption to the uncertainty-weighting mechanism.

B.3 Detailed Parameterization

All deep-learning-based XAI baselines were trained using the same ResNet-18 backbone to ensure a consistent comparison. Input images were resized to 224×224 pixels and normalized using ImageNet statistics. During training, two augmentation operations were applied: random horizontal flipping and random rotation by 10°. No stochastic augmentation was applied during testing; test images were only resized, converted to tensors, and normalized.

For each dataset, the final fully connected layer of ResNet-18 was replaced to match the number of target classes. Unless otherwise stated, ImageNet-pretrained weights were used and the network was fine-tuned, with an optional frozen-backbone setting fixing all layers except the final classifier. Models were trained using cross-entropy loss and AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 32, and 10 epochs. For *U-DeepBN*, we fixed the uncertainty-weight sensitivity parameter at $\lambda = 1$ to use the uncertainty term at its nominal scale, without additional amplification or attenuation. Feature uncertainty was estimated from $R = 10$ stochastic feature-extraction runs using Monte Carlo dropout with probability $p = 0.1$. The SL measure used $\gamma = 0.8$ and $N_{\text{eff}} = n$, where n is the number of independent, equally weighted observations used to construct each histogram.

To address class imbalance for the Dog Skin Disease dataset, we apply a class-weighted Dirichlet prior during Bayesian Network parameter learning. For Dog Skin, we used training-set class weights $w_k = N/(Kn_k)$ for all models. In *U-DeepBN*, each observation contributed its class weight to the sufficient statistics used to estimate CPTs; the same weights were applied to cross-entropy training of the feature extractor and ResNet baselines. The class weights were set to $[0.9135, 1.3660, 1.0277, 2.2302, 0.8348, 0.6437]$. For the Fish Disease and Cow Lumpy Skin Disease datasets, standard cross-entropy loss was used. Final performance was reported using overall test accuracy and per-class accuracy. We split each dataset into 70% training, 15% validation, and 15% test data. Checkpoint and model selection were performed exclusively using a held-out validation split that was fully disjoint from the reported test set; the test set was used only for final evaluation and was not involved in any model-selection decisions.

For XAI evaluation, the trained ResNet-18 model was used as the common classifier for SHAP, MIP, Integrated Gradients, saliency-map, and Grad-CAM-based explanations. Grad-CAM explanations were computed using the final convolutional block of ResNet-18, i.e., `layer4`. Saliency maps were obtained from the gradient of the target-class score with respect to the input image and normalized to the range $[0, 1]$.

B.4 Details of Evaluation Metrics

This section provides the additional equations of the evaluation metrics used in our experiments.

- **Robustness (Rbst)** measures relative performance degradation when evaluating the model on compromised data compared to healthy (clean) data, defined as

$$Rbst = 1 - \frac{\mathcal{P}_{\text{healthy}} - \mathcal{P}_{\text{compromised}}}{\mathcal{P}_{\text{healthy}}}, \tag{A.34}$$

where $\mathcal{P}_{\text{healthy}}$ and $\mathcal{P}_{\text{compromised}}$ denote performance (e.g., accuracy) on healthy and compromised datasets. Higher values indicate greater robustness. Here $\mathcal{P}_{\text{compromised}}$ is the test accuracy, measured against the true labels, of the model trained on the poisoned training set (Appendix C).

- **Local Fidelity (LF)** (Ribeiro et al., 2016) quantifies how well a simplified explanation $g(x')$ approximates the full model $f(x')$ within a local neighborhood B:

$$\text{LF} = 1 - \frac{1}{|B|} \sum_{x' \in B} |f(x') - g(x')|, \tag{A.35}$$

where g is a local surrogate fitted separately around each instance x to approximate the evaluated model's predicted probability at nearby perturbed points $x' \in B$. We use the same surrogate-fitting procedure for baseline models and *U-DeepBN*. Higher LF indicates closer alignment between local explanations and model behavior.

- **Explanation Stability (Sta)** (Alvarez-Melis & Jaakkola, 2018) evaluates consistency of explanations across similar inputs using a local Lipschitz constant:

$$\hat{L}(x) = \max_{x' \in B_\epsilon(x)} \frac{\|E(x) - E(x')\|_2}{\sqrt{d_E}\,\|x - x'\|_2}, \tag{A.36}$$

where $E(x)$ is the generated explanation or saliency map for input x, and x' represents a perturbed input derived from x. For *U-DeepBN*, $E(x)$ is the vector of per-input posterior effects. For feature f, its effect is the change in the predicted-class posterior probability when that feature is omitted from the evidence: $E_f(x) = P(Y = \hat{y}_x \mid x) - P(Y = \hat{y}_x \mid x_{\setminus f})$, where $\hat{y}_x$ is the class predicted for the original input x. Here d_E is the number of elements in the explanation produced by the evaluated method. We apply this dimension normalization to all six methods before computing Sta, including all schemes. Lower $\hat{L}(x)$ means higher robustness to perturbations.

C Threat Model with Detailed Setup

To evaluate model robustness and explanation stability, we construct compromised training datasets by incorporating traditional adversarial and XAI-specific attacks.

C.1 Backdoor Attack

A subset of training images is modified by injecting a visible square patch trigger. Specifically, a fixed fraction of images (r_{bd}) is randomly selected and altered as follows:

- **Trigger shape:** solid square patch
- **Patch size:** 8% of the shorter image side
- **Patch position:** bottom-right corner (default)
- **Patch color:** white (RGB: 255, 255, 255)

After inserting the trigger, all modified images are relabeled to a single *target class*, which is randomly selected from the set of classes for each dataset. This enforces a strong spurious correlation between the trigger and the target label.

C.2 Label-Flipping Data Poisoning

In addition to the backdoor attack, label-flipping noise is introduced. A separate subset of training images (r_{flip}) is randomly selected and their labels are altered without modifying the image content. Two modes are supported:

- **Random flipping:** labels are reassigned uniformly to any class different from the original label
- **Targeted flipping:** labels are reassigned to a predefined target class

In our experiments, random flipping is used unless otherwise stated.

C.3 Poisoning Procedure

Given a training dataset of size N, we randomly shuffle all samples and partition them into three disjoint subsets:

- $N_{bd} = \lfloor r_{bd} \cdot N \rfloor$ samples for backdoor injection
- $N_{flip} = \lfloor r_{flip} \cdot N \rfloor$ samples for label flipping
- Remaining samples left clean

The two subsets are non-overlapping, ensuring that each sample is affected by at most one type of corruption.

C.4 Feature Masking Mechanism (for Fooling post hoc explanations)

Feature importance is computed by selectively masking regions of the input image. Specifically, we employ an image masker that replaces selected regions with a blurred version of the image, thereby removing local information while preserving global structure. This masking strategy ensures that the model is evaluated on plausible inputs while suppressing the contribution of specific spatial regions.

Formally, given an input image x, a masking operation $M(x, S)$ replaces a subset of features S with blurred values, producing a perturbed input $\tilde{x}$. The model is then evaluated on $\tilde{x}$ to estimate the marginal contribution of the masked features.

C.5 SHAP masking perturbations

Interpretations (i.e., SHAP values) are approximated through repeated perturbations of the input. For each image, multiple masked samples are generated by varying the subset of features being suppressed. In our implementation, we use a blur-based masker with a kernel size of 128×128, and estimate interpretations using a finite number of model evaluations (e.g., up to 10,000 forward passes).

To improve computational efficiency, perturbations are processed in batches, and explanations are computed for the top-K most relevant output classes (with K = number of classes in our experiments).

C.6 Implementation Details

All random selections are controlled by a fixed random seed to ensure reproducibility. The compromised dataset preserves the original directory structure, with modified images saved under their assigned labels. A manifest file is generated to record key parameters, including the selected target class, attack rates, and the number of modified samples.

We use a relatively low backdoor injection rate of 3% to simulate a stealthy attack scenario in which only a small fraction of training samples is compromised. In contrast, a 10% label-flipping rate is used to introduce moderate label noise while preserving the dataset's overall class distribution and visual characteristics.

All experiments were conducted on NVIDIA V100 GPU nodes within a university-managed high-performance computing cluster. Each V100 node consists of 2× NVIDIA V100 GPUs (16 GB memory per GPU), 24 CPU cores based on Intel Xeon Gold 6136 processors, and approximately 384 GB of system memory per node, with up to 368 GB available for jobs. The nodes are connected via standard Ethernet interconnects and include local SSD storage for computation. In addition, we report the approximate runtime of each experiment in the main paper, enabling a clear understanding of execution time alongside hardware specifications. An anonymized code repository is available at `https://anonymous.4open.science/r/U-DeepBN-801C/README.md` for reproducibility.

D Limitations of the Present Evaluation

- *Frozen vs. fine-tuned backbones.* *U-DeepBN* uses a frozen ResNet-18 encoder, whereas the baselines are fine-tuned (10 epochs) on the (possibly poisoned) training data. Part of the robustness and runtime advantage may therefore reflect this asymmetry rather than the uncertainty-aware BN layer. A frozen-backbone baseline and a classifier (e.g., logistic regression or gradient boosting) trained on the same deep+classical feature table would isolate the contribution of the BN layer.
- *UQ and discriminative-head baselines.* We do not compare against MC-Dropout, deep ensembles, evidential deep learning, or adversarially trained CNNs.
- *BN topology.* The Chow–Liu and TAN structures are fixed per dataset; sensitivity to the choice of topology is not evaluated.
- *Identical Epistemic and Total rows.* These rows of Table 3 agree to three decimals on all datasets. We attribute this to identical downstream decisions, but this has not yet been verified against the per-run uncertainty values.