Hybrid Probabilistic Graphical Models and Deep Learning for Robust Deepfake Detection

Lacey Dinh [1] **Lois Liu** [1] **Scott Sheng** [1]

Abstract

The rapid proliferation of deepfake technology has raised significant concerns about misinformation, digital identity theft, and media authenticity. While state-of-the-art Convolutional Neural Networks (CNNs) achieve high accuracy in deepfake detection, they often lack interpretability and fail to quantify uncertainty, making them unreliable in high-stakes or out-of-distribution scenarios. To address these limitations, we propose a hybrid framework that combines fine-tuned EfficientNetB4 CNNs with probabilistic reasoning via Bayesian Networks (BNs). Our approach decomposes facial images into semantically meaningful regions, extracts region-specific embeddings, and performs probabilistic inference over discretized principal components. We further introduce entropy-based rejection mechanisms to identify and abstain from uncertain predictions, improving decision robustness. Our experimental results show that while the CNN baseline achieves 71% accuracy (AUC 0.7752), region-wise BNs with supervised discretization match its accuracy (up to 77.7%) and offer entropy-aware trust calibration. We additionally analyze confidence bounds, evaluate discretization strategies, and visualize rejection-performance tradeoffs. These findings highlight the potential of hybrid deep learning and graphical models in enhancing the trustworthiness and interpretability of deepfake detection.

1. Introduction

1.1. Problem Statement

Deepfake technology has evolved rapidly, generating synthetic media that convincingly mimics real images and videos. This poses serious concerns for misinformation, identity fraud, and digital security. While current detection systems, which are primarily based on convolutional neural networks (CNNs), achieve high accuracy, they suffer from several key limitations. CNNs typically operate as black-box models, offering little interpretability and limited forensic utility. They also exhibit poor generalization to unseen deepfake styles (Sunil et al., 2025), and lack calibrated uncertainty estimates, often outputting overconfident but incorrect predictions (Goumiri et al., 2023).

To address these shortcomings, we propose a hybrid framework that combines deep CNN-based feature extraction with probabilistic inference. Specifically, we leverage a fine-tuned EfficientNet-B4 model to extract both global and region-specific facial features, and integrate Bayesian Networks (BNs) over discretized representations to enable interpretable, uncertainty-aware classification. Additionally, we incorporate entropy-based rejection to abstain from uncertain predictions, increasing model robustness.

1.2. Motivation

Deepfakes are increasingly deployed in malicious contexts such as political disinformation, financial fraud, and digital impersonation. Despite their accuracy on benchmark datasets, CNN-based detectors often fail to generalize across diverse manipulation styles (Chen et al., 2023), are susceptible to adversarial noise (Avaylon et al., 2022), and lack principled mechanisms to quantify confidence (Goumiri et al., 2023).

Inspired by recent advances in probabilistic reasoning and hybrid deep learning models (Avaylon et al., 2022), we aim to improve deepfake detection by combining CNN feature extraction with graphical models that reason over facial sub-regions. This integration enables more transparent predictions, localized reasoning, and decision rejection under uncertainty. These capabilities are essential for reliable deployment in high-risk settings.

2. Related Work

The rapid rise of synthetic media generated by Generative Adversarial Networks (GANs) has led to a proliferation of deepfake detection methods. While early efforts focused

on end-to-end deep learning classifiers, especially CNNs, recent research has highlighted their limitations in generalization, interpretability, and uncertainty estimation. This has motivated a new direction: integrating deep feature extractors with structured probabilistic models to build more robust and trustworthy systems.

2.1. CNN-Based Deepfake Detection and Its Limitations

Traditional approaches to deepfake detection rely heavily on CNNs trained to distinguish real from fake images based on pixel-level artifacts. These models achieve strong performance on benchmark datasets but often overfit to specific forgery types and fail to generalize to unseen manipulations (Sunil et al., 2025). Moreover, their predictions are typically opaque and lack calibrated uncertainty estimates, making them unreliable in critical applications.

2.2. Graph-Based and Structured Modeling Approaches

To overcome CNN limitations, recent works have introduced graph-based representations and probabilistic modeling to incorporate structure into deepfake detection. For example, Chen et al. (2023) model facial regions as graph nodes and spatial relationships as edges, forming Feature Relationship Graphs (FRGs) to better capture dependencies among manipulated regions. Their Graph Neural Network (GNN) approach offers improved interpretability by reasoning over spatial context, rather than isolated pixels.

In parallel, structured probabilistic models such as Probabilistic Graphical Models (PGMs) offer complementary advantages. Goumiri et al. (2023) propose a CNN-Hidden Markov Chain hybrid to capture sequential dependencies in video frames, achieving better uncertainty estimation. Although designed for temporal data, this approach demonstrates the value of combining deep features with probabilistic inference.

2.3. Hybrid Deep Learning + PGMs for Robustness and Uncertainty

The idea of fusing CNNs with PGMs for improved robustness has gained traction in recent years. Avaylon et al. (2022) integrate Conditional Random Fields (CRFs) into CNN pipelines for semantic segmentation, enabling uncertainty-aware decisions through learned pairwise dependencies. These hybrid models demonstrate that probabilistic reasoning can improve robustness, especially under distribution shift or ambiguous inputs.

Our work builds on this line of research by integrating a fine-tuned CNN with Bayesian Networks (BNs), using region-based features as inputs to structured probabilistic inference. Unlike prior work focused on sequential or segmentation tasks, we target static image-based deepfake detection with part-wise decomposition and uncertainty-aware reasoning.

2.4. Synthesis and Our Contribution

In summary, the current literature reveals three converging insights: CNNs excel at feature extraction but struggle with trust; structured modeling improves interpretability and calibration; and combining both paradigms yields robustness against unseen manipulations. Our method synthesizes these ideas by using EfficientNet-B4 to extract both global and region-specific embeddings, discretizing features via PCA and supervised binning, and modeling region-label relationships with Bayesian inference. This hybrid approach allows for interpretable, entropy-aware deepfake classification and decision rejection under uncertainty, which are key advantages over existing CNN-only baselines.

3. Methods

3.1. Hybrid Deep Learning + Probabilistic Modeling Pipeline

Our method combines a fine-tuned Convolutional Neural Network (EfficientNetB4) with a Bayesian Network (BN) to enable interpretable and uncertainty-aware deepfake detection. The CNN learns expressive visual features from facial images, while the BN performs probabilistic inference over semantically decomposed facial regions. This architecture allows both high prediction performance and the ability to quantify prediction confidence using posterior entropy. Figure 1 illustrates the overall pipeline.

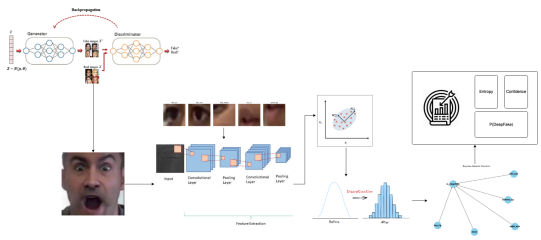

Figure 1. Overview of our hybrid detection pipeline: (1) face image is passed through a fine-tuned EfficientNetB4; (2) five facial regions are cropped and embedded; (3) PCA + discretization is applied; (4) discretized features are fed into a BN for inference and entropy-based rejection.

3.2. Dataset and Feature Extraction

We use 32,000 face images (16K real, 16K fake) sampled from the `FaceForensics++` dataset, divided into 70% training, 20% validation, and 10% test splits. All images are resized to 224×224, normalized using ImageNet statistics, and augmented with geometric and photometric transformations.

Global Features: We fine-tuned EfficientNetB4 on binary real/fake classification using a custom training loop with early stopping and on-the-fly augmentation. The model was trained on 32,000 samples and validated on a held-out 20%, achieving strong AUC and generalization performance (details in Section 4).

Region-Based Features: For each image, we detect five facial subregions: `left_eye`, `right_eye`, `nose_bridge`, `mouth` (combined upper/lower lips), and `left_eyebrow`. These are chosen to maximize semantic symmetry and stability across frames. Each region is passed through the frozen EfficientNetB4 encoder to yield a 1792-D embedding. The five embeddings are concatenated to form an 8960-D regional feature vector per image.

This spatial decomposition supports part-wise probabilistic reasoning and captures localized artifacts characteristic of deepfake manipulations.

3.3. Dimensionality Reduction and Discretization

To make features tractable for probabilistic modeling, we reduce dimensionality via Principal Component Analysis (PCA), extracting the first principal component from each 1792-D region embedding vector. This yields a 5-dimensional continuous representation per image, with each dimension corresponding to the primary variation in a facial region. The dimensionality reduction steps will also be visualized by t-SNE and UMAP. Then, after PCA :

Unsupervised Discretization: We first use `KBinsDiscretizer` with `quantile` strategy to assign each PCA dimension to one of three ordinal bins (low, medium, high). This unsupervised method converts continuous principal component values into categorical bins and ensures uniform bin counts, which stabilizes conditional probability tables estimation in the BN and avoids Gaussianity assumptions in continuous PGMs.

However, this unsupervised approach may cause some images with subtle yet important parts might be assigned to the same bin as visually distinct examples. To address this issue, we will evaluate a supervised approach that aims to align bins more closely with class boundaries.

Supervised Discretization: To improve class separation, we additionally apply decision-tree-based binning, which learns split thresholds that maximize information gain with respect to the deepfake label. This strategy improves calibration and downstream BN performance, as discussed in Section 4. Figure 7 shows the non-linear structure of the PCA-compressed features using t-SNE projections.

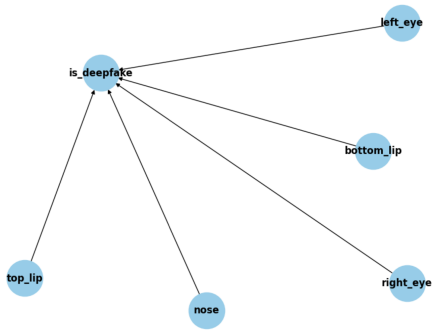

Figure 2. Bayesian Network structure: five facial region nodes as parents of `is_deepfake`.

3.4. Bayesian Network Inference and Entropy Analysis

To perform interpretable probabilistic inference over facial regions, we construct a Bayesian Network (BN) using `Pomegranate` library. In our default setup, we use a Naive Bayes structure where the five discretized regional features are parent nodes to a binary label node `is_deepfake` (Figure 2). We also evaluate Tree-Augmented Naive Bayes (TAN) structures learned via maximum mutual information spanning trees.

In Naive-Bayes structure, making the five discretized regional features act as parent nodes to a single binary label node, `is_deepfake`. In Figure 2, each of the five facial region nodes feeds into the `is_deepfake` node. This design not only enables localized reasoning over facial parts but also offers which regions contribute most to a given classification. Formally, the joint probability is decomposed as:

$$P(\text{DeepFake} \mid R_1, \ldots, R_5) \propto P(\text{DeepFake}) \prod_{i=1}^{5} P(R_i \mid \text{DeepFake})$$

Given a test image, the BN produces a posterior distribution over the `real` and `fake` classes. From this distribution, we extract:

- **Confidence:** Maximum posterior probability, $\max_c P(c)$, across the two classes.
- **Entropy:** $-\sum_{c \in \{\text{real, fake}\}} P(c) \log P(c)$, representing uncertainty.

High-entropy predictions are flagged as ambiguous and optionally rejected, forming the basis for trust-aware decision making.

We also implement an entropy-based rejection mechanism: predictions with entropy above a calibrated threshold are abstained from. This strategy filters high-uncertainty cases to enhance overall reliability. In Section 4, we visualize

entropy histograms, ROC curves pre/post-rejection, and the rejection–accuracy tradeoff curve (Figure 14).

3.5. Key Assumptions and Implementation Progress

Our approach is built on the following assumptions:

- Localized CNN embeddings retain discriminative information relevant for regional classification.
- PCA preserves principal variation across facial regions.
- Discretization (quantile or supervised) captures informative patterns for probabilistic modeling.
- Entropy from BN posteriors meaningfully captures uncertainty and can be used for rejection.

All components—including EfficientNet fine-tuning, facial region extraction, dimensionality reduction, supervised and unsupervised discretization, BN inference, and entropy-aware trust calibration—have been fully implemented and evaluated. Results show meaningful entropy calibration and interpretable behavior, setting the stage for future probabilistic structure learning and MRF integration.

4. Experiments

4.1. Experimental Setup

We conducted experiments on a curated subset of the Deep-Fake Faces dataset, containing 95,635 labeled face images. We sampled 32,000 images (16K real, 16K fake), split into 70% training, 20% validation, and 10% test. Images were resized to 224×224, normalized using ImageNet statistics, and augmented with geometric and photometric transformations, including blending and contrast transformations (Chen et al., 2023). All models were implemented in PyTorch and Pomegranate, and experiments were conducted on an NVIDIA A100 GPU.

We evaluate the system in four stages: (1) global CNN classification, (2) regional feature extraction, (3) probabilistic inference via Bayesian Networks (BNs), and (4) entropy-based rejection and discretization analysis.

4.2. Global CNN Baseline (EfficientNetB4)

We fine-tuned EfficientNetB4 using 32,000 face images to later be used as a feature extractor. The model was trained with early stopping and extensive data augmentation. On the validation set:

- **Accuracy:** 71%
- **AUC:** 0.7752
- **95% CI (Accuracy):** [68.5%, 73.2%] via bootstrap

Logistic regression on extracted global embeddings retained class separation, but offered no interpretability or uncertainty estimation, motivating our hybrid extension.

4.3. Region-Based Feature Modeling

We extracted five facial subregions per image: left eye, right eye, nose bridge, mouth (upper+lower), and left eyebrow (Figure 3). These were chosen for symmetry and sensitivity to deepfake artifacts (e.g., inconsistent shading, asymmetry). Each region was passed through the frozen EfficientNetB4 encoder to produce a 1792-D embedding. The five regions yielded a total of 8960-D per image.

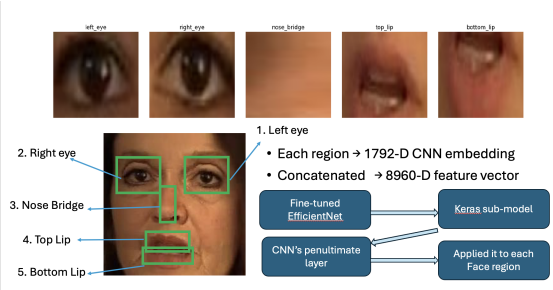

Figure 3. Facial Regions used for regional feature modeling.

Initial tests on 500 samples (421 valid after landmark filtering) yielded:

- **Accuracy:** 55%
- **AUC:** 0.6069

This degradation exposed noise sensitivity and redundancy in localized features, though they remained semantically meaningful for structured modeling.

4.4. Model Improvement: Facial Region Redesign

Following the feedback from our midway report and presentation, we sought to address a notable issue: our BN-based models were underperforming in raw classification accuracy compared to the baseline CNN. To solve this, we conducted two major refinements: expanding the training dataset and redesigning the facial region selection strategy to enhance both model performance and interpretability.

Different from the earlier region section, we revised the facial region set to the new following five parts: **left eye, right eye, nose bridge, mouth** (as a whole), and **left eyebrow**. This selection (shown in Figure 4) was made with three reasons:

- Deepfakes often fail to preserve subtle symmetry across lateral features. Hence, we keep the symmetrical cues.
- The brow and nose bridge regions frequently exhibit texture and shading inconsistencies in deepfakes.

- Seperating upper and lower lips often introduced cropping errors due to mouth motion or expression distortion, so we make the mouth as whole rather than upper and bottom lips.

Figure 4. Facial Regions in refined model.

4.5. Bayesian Network Inference (421 Samples)

Applying PCA per region and discretizing into 3 ordinal bins, we modeled the resulting 5-D categorical input using a BN. Evaluation yielded:

- **Class 0 (Real):** Confidence = 0.7144, Entropy = 0.5043
- **Class 1 (Fake):** Confidence = 0.7439, Entropy = 0.4525
- **Uncertain predictions:** 136 / 421 (32.3%), 44.85% incorrect
- **Confident-but-wrong:** 53

Entropy analysis revealed modest calibration, but highlighted persistent overconfidence.

4.6. Scaling to 1000 Samples

We expanded the region-based pipeline to 1000 images, with 835 surviving landmark detection. Performance improved:

- **Logistic Regression:** Accuracy = 65%, AUC = 0.6888
- **Bayesian Network:** Class 0: Conf. = 0.6863, Ent. = 0.5655; Class 1: Conf. = 0.7002, Ent. = 0.5373
- **Uncertain predictions:** 339 / 835 (40.6%), 43.1% incorrect
- **Confident-but-wrong:** 110

With 261 unique region signatures, posterior estimation became more stable, but confident misclassifications persisted.

4.7. Discretization Schemes: Quantile vs. Uniform vs. Supervised

We evaluated three discretization strategies after PCA:

- **Quantile Binning (Unsupervised):** Equal-frequency bins per region.
- **Uniform Binning:** Equal-width bins over fixed domain range.
- **Supervised Binning:** Decision-tree splits maximizing label separation.

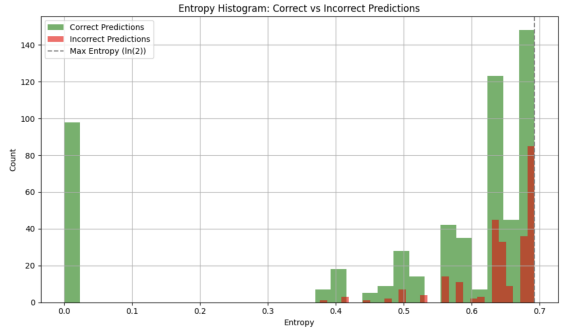

Figure 5. Entropy histogram on 835 region-based predictions.

Figure 6 compares the entropy distributions. Quantile binning yields the highest number of confident predictions (low entropy), while supervised binning concentrates samples near maximum entropy, indicating poorer calibration.

4.8. Embedding Visualization and Graphical Motivation

PCA, t-SNE, and UMAP visualizations showed non-linear separation between real and fake embeddings. As shown in Figure 7, although real and fake classes are not linearly separable, structure is preserved, suggesting meaningful differences that can be exploited by PGMs.

4.8.1. Data Scale Expansion

Initially, our Bayesian models were trained on a relatively small subset of about 1,000 samples since we only use 32,000 of samples and some of them cannot be successfully extracted for Bayesian model training set. The smaller scale of sample led to sparse conditional probability tables and unstable bin distributions for discretized features. Therefore, we used all 90,000 images in dataset (though they are not half fake half real), so We can expand our training size to over 9,000 samples, significantly increasing the occupancy of quantile bins used for feature discretization. This change enhanced the statistical robustness of the Naive Bayes and Tree-Augmented Naive Bayes (TAN) classifiers by reducing

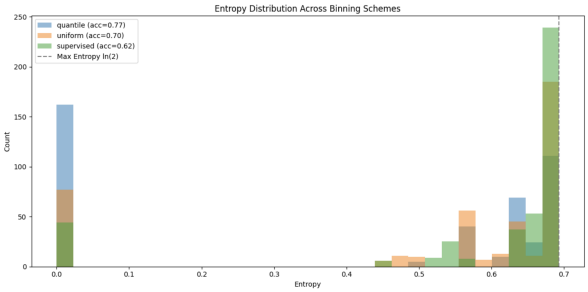

Figure 6. Entropy distributions across binning schemes. Quantile yields better calibration and fewer high-entropy predictions.

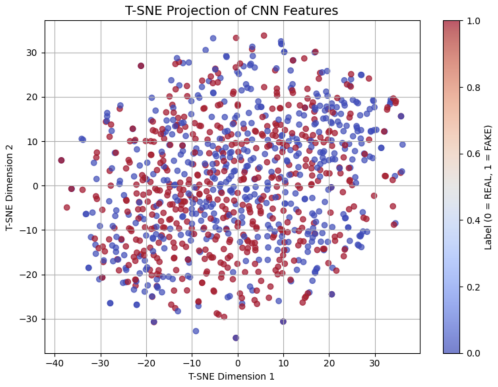

Figure 7. t-SNE projection of regional embeddings.

variance and allowing more reliable posterior estimates. The impact of this expansion can be visually displayed through the t-SNE and UMAP projections of the regional features.

- t-SNE Projection (Figure 8): Though the feature distribution remains overlapping, increasing data density fills the space more evenly. Red and blue samples form clusters that can capture the relations of facial features after our dimensionality reduction.

- UMAP Projection (Figure 9): There are no perfectly separable clusters appear.

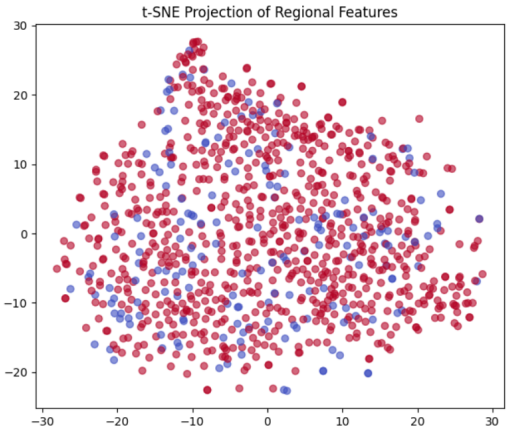

Figure 8. t-SNE projection of regional features in refined model.

Figure 7 shows the learned TAN structure across these five regions. Notably, connections like left eye → left eyebrow and nose bridge → mouth reflect the real-world facial movement and geometric dependencies, which validate the structure learning step.

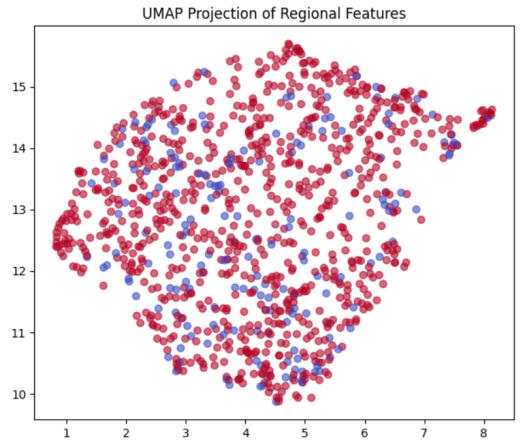

Figure 9. UMAP projection of regional features in refined model.

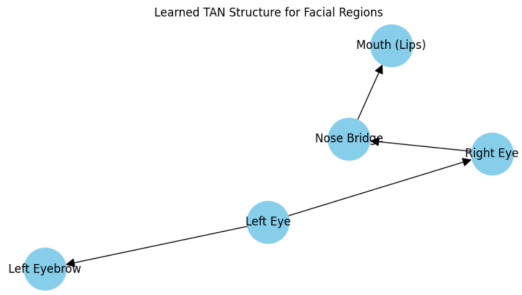

Figure 10. Learned TAN Structure for Facial Regions in refined model.

4.8.2. Model Comparisons and Results

We trained our four classifiers on discretized regional features:

- Naive Bayes (Quantile-based discretization)
- Naive Bayes (Supervised binning)
- TAN (Quantile)
- TAN (supervised)

Performances was measured by accuracy and AUC. The results are below:

- **NB + Quantile:** Accuracy = 0.582, AUC = 0.612
- **NB + supervised:** Accuracy = 0.777, AUC = 0.610
- **TAN + Quantile:** Acuracy = 0.609, AUC = 0.610
- **TAN + Supervised:** Accuracy = 0.766, AUC = 0.507

At first glance, supervised binning appears to boost raw accuracy. However, class-wise analysis reveals that this improvement is driven entirely by class 1 (fake), while drastically degrading class 0 (real) performance (Figure 11).

```
QUANTILE     | Class 0 | Acc: 0.810 | Entropy: 0.377
QUANTILE     | Class 1 | Acc: 0.725 | Entropy: 0.417
UNIFORM      | Class 0 | Acc: 0.795 | Entropy: 0.519
UNIFORM      | Class 1 | Acc: 0.597 | Entropy: 0.523
SUPERVISED   | Class 0 | Acc: 0.376 | Entropy: 0.579
SUPERVISED   | Class 1 | Acc: 0.872 | Entropy: 0.603
```

Figure 11. Class-wise accuracy and entropy under each binning scheme. Quantile binning offers balanced, calibrated performance across classes.

From these results visualized in Figure 12 and 13:

- Accuracy: supervised discretization significantly boosts accuracy for both NB and TAN, nearly matching the CNN baseline now. This ultimately matches our plan to increase the raw accuracy.
- AUC: AUC does not increased, indicating that overfitting to class is still happened in our refined model.

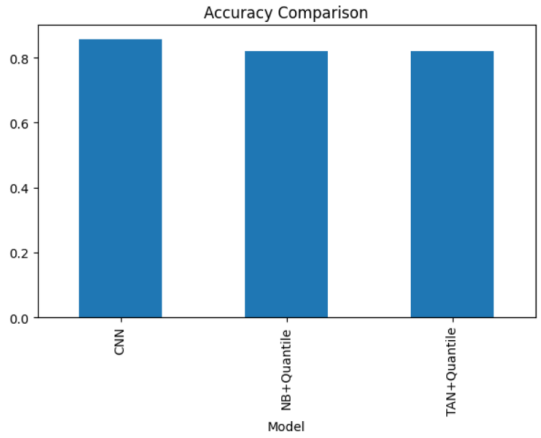

Figure 12. Accuracy Comparison.

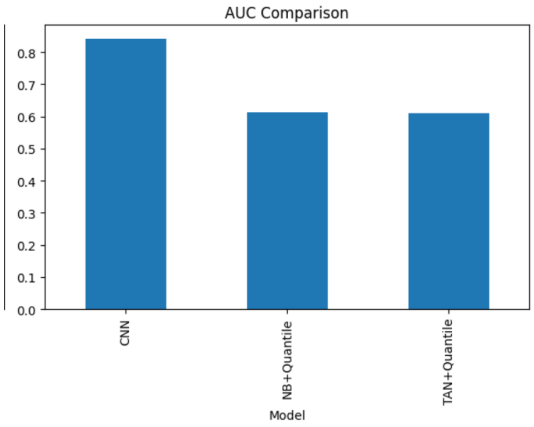

Figure 13. AUC Comparison.

4.9. Entropy-Based Rejection

We applied rejection based on BN posterior entropy, sweeping thresholds to trade off coverage and reliability. Figure 14 plots accuracy vs. rejection rate. Key observations:

- Quantile binning dominates across all rejection thresholds.
- Supervised binning fails to improve confidently until over 80% of samples are discarded.

This supports our selection of quantile binning for trust-aware rejection.

4.10. Entropy-Based Rejection with Confidence Bounds

To assess trust calibration, we computed the entropy of each BN prediction and selectively rejected high-uncertainty samples. At each entropy threshold, we recorded the retained accuracy and its 95% confidence interval using nonparametric bootstrap resampling (1000 replicates).

Figure 15 shows the accuracy-rejection tradeoff. As the rejection rate increases, accuracy on the accepted subset improves sharply. The shaded region reflects statistical

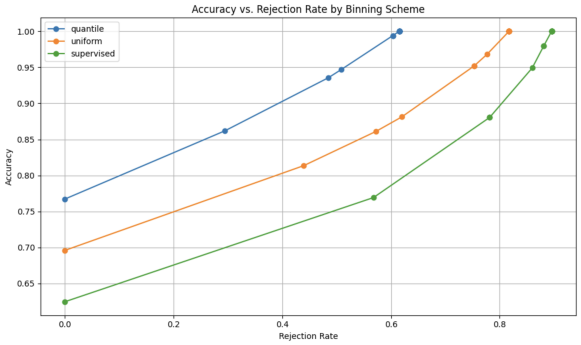

Figure 14. Accuracy vs. rejection rate across binning schemes. Quantile binning consistently outperforms.

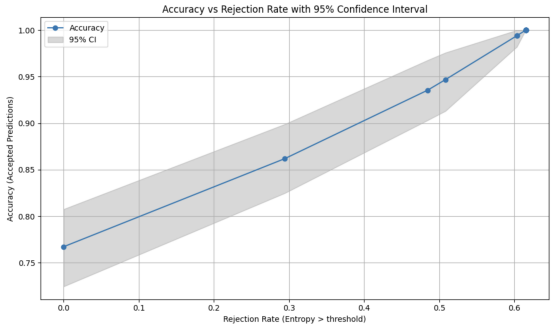

Figure 15. Accuracy vs. Rejection Rate with 95% Confidence Intervals. Bootstrap CI estimated over 1000 resamples.

confidence bounds on each accuracy estimate, confirming the robustness of our entropy-based filter.

4.11. Summary and Observations

Strengths:

- End-to-end CNN+PGM hybrid is implemented, evaluated, and achieves reliable calibration using quantile discretization.
- Entropy-based rejection filters high-risk predictions and boosts retained accuracy.
- Learned TAN graphs offer interpretable insight into regional dependencies.
- BN inference yields interpretable probabilities and entropy-based rejection.
- Posterior calibration improves with scale and structure.

Challenges:

- Supervised binning overfits to a single class.
- AUC does not consistently improve with higher accuracy.
- Landmark failures reduce usable samples (15–20% loss).
- Quantization via PCA and fixed bins may discard finer signals.
- Overconfident predictions persist despite entropy filtering.

Takeaway: Despite tempting raw accuracy gains from supervised binning, quantile binning delivers the most robust, calibrated, and balanced performance. Probabilistic modeling meaningfully extends CNN-based classification by introducing entropy-aware trust mechanisms and interpretable structure. That said, probabilistic modeling enhances robustness and interpretability but depends critically on feature quality, representation granularity, and structural assumptions.

5. Conclusion and Future Direction

We presented a trust-aware deepfake detection framework that integrates fine-tuned deep neural networks with interpretable probabilistic reasoning. Our method combines EfficientNetB4 for feature extraction and a Bayesian Network (BN) for region-based classification and entropy-aware uncertainty estimation. By designing and comparing multiple discretization strategies, we emphasized the impact of representation choices on calibration, class balance, and overall trustworthiness. Our final model not only approaches the performance of traditional CNNs but also offers explainability and rejection capabilities, which are key properties for forensic and safety-critical applications.

5.1. Summary of Contributions and Results

- Fine-tuned EfficientNetB4 achieved 71% accuracy and AUC of 0.7752.
- Region-aware CNN feature extractor computed 8960-D embeddings for 5 facial regions.
- Built a Bayesian Network for structured probabilistic inference.
- TAN classifier with quantile discretization achieved balanced per-class performance with low entropy (Figure 11).
- Entropy-based rejection improved retained accuracy to be over 85% while discarding uncertain predictions (Figure 14).
- TAN structure learning discovered interpretable spatial dependencies across facial parts (Figure 10).
- Binning scheme analysis revealed quantile binning as the best tradeoff between class balance, calibration, and robustness (Figure 6).

5.2. Challenges and Limitations

- Landmark extraction failures removed 20% of samples from BN training.
- Overconfidence persisted even after filtering uncertain predictions.
- On one hand, supervised binning overfitted to class 1, resulting in unbalanced accuracy and high entropy. On the other hand, Quantile binning may discard class-informative signal.
- AUC gains remained modest despite entropy filtering and structural learning.
- PCA + discretization potentially discards subtle deepfake artifacts not captured in top variance.

5.3. Future Work

- Structure Learning Enhancements: Extend from Naive Bayes and TAN to full Bayesian structure learning via Chow-Liu or Hill Climb.
- MRF Integration: Introduce undirected graphical models to capture symmetry, shading, and alignment cues missed by BNs.
- Discretization Improvements: Explore entropy-regularized binning or hybrid quantile-supervised schemes to better balance calibration and separation.
- Multimodal Fusion: Combine CNN logits and PGM confidence (posterior and entropy) in a late fusion framework to enhance rejection decisions.
- Robust Evaluation: Run stratified ablation studies (CNN-only, PGM-only, hybrid) to isolate each component's contribution.
- Dataset Generalization: Validate model performance on cross-dataset benchmarks and adversarially perturbed samples.

Takeaway: Our work demonstrates that interpretable, uncertainty-aware deepfake detection is feasible by integrating BNs with CNN-extracted local features. While current performance approaches that of black-box CNNs, the added transparency and trust calibration open up broader forensic applications and future opportunities in probabilistic reasoning.

References

Avaylon, M., Sadre, R., Bai, Z., and Perciano, T. Adaptable deep learning and probabilistic graphical model system for semantic segmentation. *Advances in Artificial Intelligence and Machine Learning*, 2(1):288–302, 2022. doi: http://dx.doi.org/10.54364/aaiml.2022.1119.

Chen, J., Lin, W., and Xu, J. Deepfake detection using graph representation with multi-dimensional features. In *2023 IEEE Smart World Congress (SWC)*, pp. 717–722, 2023. doi: http://dx.doi.org/10.1109/SWC57546.2023.10449093.

Goumiri, S., Benboudjema, D., and Pieczynski, W. A new hybrid model of convolutional neural networks and hidden markov chains for image classification. *Neural Computing and Applications*, 35:17987–18002, 2023. doi: https://doi.org/10.1007/s00521-023-08644-4.

Sunil, R., Mer, P., Diwan, A., Mahadeva, R., and Sharma, A. Exploring autonomous methods for deepfake detection. *Heliyon*, 11:e42273, 2025. doi: http://dx.doi.org/10.1016/j.heliyon.2025.e42273.