Journal of High School Science

Conditional uncertainty-aware political deepfake detection with stochastic Convolutional Neural Networks

Gardoş R-P

Submitted: December 14, 2025, Revised: version 1, January 16, 2026, version 2, January 28, 2026, version 3, January 29, 2026, version 4, January 30, 2026
Accepted: January 31, 2026

Abstract
Recent advances in generative image models have enabled the creation of highly realistic political deepfakes, posing serious risks to information integrity, public trust, and democratic processes. While automated deepfake detectors are increasingly deployed in moderation and investigative pipelines, most existing systems provide only point predictions and fail to indicate when outputs are unreliable—an operationally critical limitation in high-stakes political contexts. This work investigated conditional, uncertainty-aware political deepfake detection using stochastic convolutional neural networks within a strictly empirical, decision-oriented reliability framework. Rather than framing uncertainty from a purely Bayesian or interpretive perspective, uncertainty was evaluated through observable criteria, including calibration and its relationship to prediction errors under global and confidence-conditioned evaluation regimes. A politically focused binary image dataset was constructed via deterministic metadata-based filtering from a large public real–synthetic corpus. Two pretrained CNN backbones, ResNet-18 and EfficientNet-B4, were fully fine-tuned end-to-end for binary classification. Deterministic inference was compared with stochastic procedures, including single-pass stochastic prediction, Monte Carlo dropout with multiple forward passes, temperature scaling for calibration, and an ensemble-based uncertainty surrogate as a non-Bayesian reference. Evaluation protocols were defined, with the fake class treated as positive, ROC-AUC and thresholded confusion matrices reported, and experiments conducted under controlled in-distribution settings with supplementary generator-disjoint out-of-distribution analysis. The results showed that calibrated probabilistic outputs and uncertainty estimates supported downstream decision-making by enabling risk-aware moderation policies. A systematic confidence-band analysis further delineated when uncertainty added operational value beyond predicted confidence, clarifying the practical scope and limitations of uncertainty-aware deepfake detection in political contexts.

Keywords
Political deepfakes, Information integrity, Public trust, Deepfake detector, Uncertainty aware deepfake detection, Risk aware moderation, Calibrated probabilistic output, Uncertainty estimate, Convolutional Neural Network, Generative image model

Rafael-Petrut Gardos, Colegiul National ”Mihai Eminescu” Satu Mare, Mihai Eminescu 5, 440014, Satu Mare, Romania. gardos.rafael@gmail.com

https://doi.org/10.64336/001c.156299

1. Introduction

Political deepfakes, which are synthetic images depicting public figures in fabricated or misleading contexts, represent a growing challenge for information ecosystems. Improvements in generative image models have enabled the creation of content that is increasingly difficult to distinguish from authentic imagery, even for expert human observers. In political contexts, such imagery can be used to fabricate events, misrepresent actions or statements, and erode trust in legitimate media sources.

Automated detection models are increasingly deployed to identify and triage potentially manipulated political imagery. Most detection systems are evaluated primarily using discriminative performance metrics such as accuracy or the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) (1,2). While these metrics quantify average separability between real and synthetic samples, they do not characterize the reliability of predicted probabilities. In high-stakes political settings, reliability is a critical operational property (3,4). A detector that is overconfident on ambiguous or borderline inputs may be unsafe even if its average accuracy is high.

This motivates the integration of uncertainty-aware inference into political deepfake detection. However, uncertainty estimation is frequently discussed in abstract or interpretive terms. In this work, uncertainty is treated as an empirical signal whose value is determined by its observable behavior on held-out data. Specifically, uncertainty is evaluated according to two criteria: [i] probabilistic calibration, which measures the agreement between predicted confidence and empirical correctness, and [ii] uncertainty-error correlation, which measures whether uncertainty values are systematically higher for misclassified samples than for correctly classified ones.

A commonly used approach for uncertainty-aware inference in neural networks is Monte Carlo (MC) dropout, which activates dropout layers at test time and samples a predictive distribution via repeated stochastic forward passes (5). While MC dropout is often described as an approximate Bayesian method, its empirical behavior may reflect both posterior-like variability and noise-induced smoothing. Consequently, this work does not assume that MC dropout variance corresponds uniquely to epistemic uncertainty. Instead, uncertainty estimates are evaluated empirically and comparatively, alongside deterministic inference, single-pass stochastic inference, and post-hoc calibration baselines.

The central question addressed in this work is therefore not whether uncertainty estimates are theoretically Bayesian, but whether they are operationally useful; and if so; in what decision regimes. Formally, this work asks: *Under controlled in-distribution evaluation and under a limited generator-disjoint distribution shift, do uncertainty-aware inference procedures improve the reliability of political deepfake detector outputs, as measured by calibration and uncertainty-error correlation, relative to deterministic and post-hoc calibrated baselines?* This question is investigated using two convolutional neural network backbones that differ substantially in depth and

representational capacity, evaluated under a unified experimental protocol with explicitly defined semantics and metrics.

2. Related work

2.1 Deepfake detection in political and real-world contexts

Early deepfake detection methods relied on identifying low-level artifacts, physiological inconsistencies, or generator-specific cues embedded in synthetic imagery. While effective against early generative models, many such approaches degrade as synthesis quality improves and post-processing pipelines become more sophisticated. Contemporary detectors increasingly adopt deep convolutional architectures trained end-to-end (6) to discriminate real from synthetic images based on learned representations rather than hand-crafted cues.

Despite strong performance under controlled benchmarks, deepfake detection in political contexts presents additional challenges (1,7,8). Political imagery frequently contains complex backgrounds, crowd scenes, non-frontal faces, varied lighting conditions, and substantial compression introduced by social media platforms. Moreover, political content is subject to rapid temporal drift as new events, figures, and visual styles emerge. These factors reduce margin separability and increase ambiguity even for authentic images, motivating evaluation beyond average accuracy. As a result, recent work has emphasized the need to assess detector behavior under uncertainty, rather than relying solely on discriminative metrics (9,10).

2.2 Reliability and calibration of neural network predictions

Probabilistic reliability has been extensively studied in the context of neural network classifiers. Calibration evaluates whether predicted probabilities align with empirical correctness frequencies (11,12), providing a complementary perspective to discrimination-focused metrics such as accuracy or ROC-AUC. Proper scoring rules, including negative log-likelihood and the Brier score, quantify probabilistic accuracy by penalizing overconfident incorrect predictions (13,14), while summary metrics such as Expected Calibration Error (ECE) aggregate discrepancies between confidence and accuracy across bins.

Importantly, calibration and discrimination capture distinct properties of a model. A classifier may achieve near-perfect ROC-AUC while remaining poorly calibrated (15), particularly when predictions are systematically overconfident. Conversely, post-hoc calibration methods can substantially alter confidence estimates without affecting ranking-based metrics. This distinction is especially relevant in high-stakes settings, where downstream decisions depend on the reliability of predicted probabilities rather than on ranking alone (16). Consequently, recent work increasingly advocates reporting calibration-sensitive metrics alongside accuracy and AUC, particularly when model outputs are intended to support human-in-the-loop decision-making.

2.3 Uncertainty estimation via stochastic inference

Stochastic inference techniques such as Monte Carlo dropout approximate predictive distributions by activating dropout layers at test time and sampling multiple stochastic forward passes (5). These methods are often motivated as approximations to Bayesian inference; however, their empirical behavior can also reflect noise-induced smoothing and regularization effects. As a result, predictive variance under dropout does not necessarily correspond to epistemic uncertainty in a strict sense (17-19).

Several studies have emphasized the importance of evaluating stochastic inference empirically rather than assuming a probabilistic interpretation (20). Comparisons against deterministic inference, single-pass stochastic inference, and non-stochastic calibration baselines are essential for disentangling the effects of injected noise from multi-sample aggregation (21). From an operational perspective, the utility of uncertainty estimates lies not in their theoretical provenance but in their observable behavior: whether they improve calibration, reduce overconfidence, or correlate with prediction errors.

2.4 Uncertainty, selective prediction, and deepfake detection

Uncertainty-aware prediction has been explored in the broader context of selective classification, abstention mechanisms, and risk-coverage trade-offs, where models defer decisions on inputs deemed unreliable. In such settings, uncertainty is valuable insofar as it identifies cases likely to be incorrect (22,23), enabling downstream policies such as escalation to human review or prioritization under limited resources.

In deepfake detection, relatively few works have evaluated uncertainty explicitly, and even fewer have examined uncertainty-error alignment in politically salient imagery. Existing approaches have considered ensembles or stochastic inference, but often report uncertainty qualitatively or assume a Bayesian interpretation without empirical validation. Systematic evaluation of whether uncertainty estimates meaningfully rank misclassified samples above correct ones remains limited, particularly under controlled political datasets.

The present work contributes to this line of research by evaluating uncertainty as an operational signal rather than as an interpretive construct. By measuring calibration, proper scoring rules, and uncertainty-error correlation across multiple inference procedures and architectures, this study situates uncertainty-aware deepfake detection within a reliability-oriented evaluation framework aligned with real-world decision-making requirements.

3. Methods

3.1 Dataset construction and filtering

This study primarily evaluated political deepfake detection under controlled, in-distribution conditions. To this end, a politically focused binary image dataset was constructed from a large public real-synthetic image corpus, OpenFake, containing authentic photographs and AI-generated images produced by multiple generative models (24). The source corpus provided paired image data and structured

metadata, including textual prompts, captions, generator identifiers, and ground-truth authenticity labels.

Formally, let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the resulting dataset, where x_i is an image and $y_i \in \{0,1\}$ is its ground-truth label. The label convention is fixed throughout the manuscript, such that $y_i = 1$ denotes a synthetic (fake) image and $y_i = 0$ denotes a real image.

3.2 Political filtering pipeline

Since the full source corpus spanned diverse non-political content, a deterministic metadata-based filtering pipeline was applied to extract a subset relevant to political communication. The pipeline streamed the source dataset iteratively and inspected available metadata fields, such as prompts, captions, textual descriptions, and categorical annotations, to identify references to political actors, institutions, events, or discourse. An image was retained if any metadata field contained at least one keyword from a predefined list of of political keywords, which included; ["president", "prime minister", "election", "campaign", "senator", "congress", "parliament", "press conference", "speech", "rally", "biden", "trump", "harris", "trudeau", "sunak", "macron", "putin", "zelensky"]

This keyword list was fixed prior to dataset construction and applied uniformly to both real and synthetic samples. Importantly, the filtering step depended only on metadata and did not use image pixels or model predictions, preventing label leakage into subsequent learning or evaluation stages. All retained images were written to disk with structured filenames, and their associated metadata (label, keyword match, generator identifier, file path) was recorded in a CSV file to support reproducibility and downstream stratified analyses.

3.3 Dataset composition

After filtering, the final dataset contained $N = 4000$ images, balanced across classes with 2000 real and 2000 synthetic samples. The balanced design prevented trivial accuracy inflation due to class imbalance and ensured that calibration and uncertainty metrics were interpretable under symmetric class priors. The retained images spanned a range of political contexts, including campaign events, speeches, press conferences, and depictions of well-known public figures; however, coverage was determined by the distribution present in the source corpus and the keyword filter, rather than by manual curation.

3.4 Train-validation-test partitioning

The dataset was partitioned at the image level into three disjoint subsets using a fixed random seed. A training set, a validation set and a test set was obtained consisting of 2800, 600 and 600 images respectively.

The dataset was partitioned using a random image-level split without stratification by political identity, event type, or generator. Generator labels could therefore appear in multiple splits; as a result, the in-distribution test evaluated matched-generator generalization rather than cross-generator generalization. Consequently, the same public figures, prompts, and generator families could appear in multiple splits, hence the evaluation reflected *in-distribution* performance under matched train-test conditions. In addition, a separate generator-

disjoint Out-Of-Distribution (OOD) evaluation split was constructed by holding out specific generator families from training and reserving them exclusively for evaluation; political identities remained pooled.

3.5 Out-Of-Distribution (OOD) evaluation split
In addition to the In-Distribution (ID) test split obtained via a random image-level partition, a separate OOD evaluation split was constructed to assess *generator-disjoint* generalization, following standard practice in distribution-shift evaluation for vision models (25). Concretely, the OOD split contained synthetic images produced by generator families not present in the training data (unseen generators at evaluation time), while political identities were not constrained to be disjoint and were therefore pooled across splits. The OOD split was used exclusively for evaluation (not for model selection), and all metrics followed the same evaluation semantics as ID (positive class: fake; ROC-AUC over $s(x) = p(y = 1 \mid x)$; accuracy and calibration at threshold $t = 0.5$).

3.6 Model architectures
Two convolutional neural network backbones were evaluated: ResNet-18 and EfficientNet-B4 (26,27). These architectures were selected to provide contrasting inductive biases, depth profiles, and parameterization regimes while remaining computationally feasible for repeated stochastic inference. ResNet-18 is a relatively shallow residual network characterized by uniform channel widths and additive skip connections, whereas EfficientNet-B4 employs compound scaling across depth, width, and input resolution, together with inverted residual blocks and squeeze-and-excitation mechanisms. Evaluating both models enabled assessment of whether reliability and uncertainty behaviors were consistent across architectures with substantially different representational capacity and architectural design (28).

Each model defined a mapping $z(x) = g_\phi(h_\theta(x))$ where $h_\theta(\cdot)$ denotes the convolutional feature extractor and $z(x) = g_\phi(\cdot)$denotes a learned classification head. The head consisted of global average pooling followed by a linear projection to a single scalar logit. The predicted probability of the positive class was obtained via the logistic sigmoid $p(y = 1 \mid x) = \sigma(z(x)) = \frac{1}{1 + \exp(-z(x))}$
Throughout the manuscript, the positive class was fixed as $y = 1$ corresponding to a synthetic (fake) image.

Both backbones were initialized with ImageNet-pretrained weights. Unless explicitly stated otherwise, all parameters (θ, ϕ) were optimized jointly using end-to-end fine-tuning. This choice avoided assuming that head-only training suffices for comparability and reduces the risk of conflating uncertainty-related effects with representational underfitting caused by frozen feature extractors. Frozen-backbone and partial fine-tuning regimes were not explored in this study.

All reported results corresponded to full end-to-end fine-tuning, avoiding assumptions that head-only training suffices for comparability or uncertainty analysis.

Dropout layers present in the original architectures were retained during fine-tuning.

For uncertainty-aware inference, dropout was explicitly enabled at test time, inducing stochastic forward passes through both the backbone and the classification head. As a result, stochasticity affected intermediate feature representations as well as the final decision layer, ensuring that Monte Carlo samples reflected variability throughout the network rather than being confined to the classifier head alone.

3.7 Preprocessing and input resolution

All images were resized to a fixed spatial resolution prior to model input. Unless stated otherwise, a unified resolution of 380×380 pixels was used for both backbones. This choice deviates from canonical ResNet-18 preprocessing (typically 224×224), but was adopted to maintain a single preprocessing pipeline across architectures and to avoid introducing resolution-dependent confounds when comparing uncertainty behaviors. Input resolution was therefore treated as an experimental factor rather than a fixed assumption; its potential impact on discriminative performance and calibration was examined through targeted sensitivity analyses.

Pixel intensities were scaled to the interval [0,1]. Standard ImageNet mean-standard deviation normalization is not applied in the primary configuration. Omitting ImageNet normalization alters the input distribution relative to the pretrained feature extractors and may affect transfer performance or calibration. This deviation was treated explicitly as an experimental factor rather than an implicit assumption. To empirically justify the chosen preprocessing configuration, targeted ablations over input resolution and normalization strategy were performed and reported in section 4.16 "Preprocessing Ablations: Resolution and Normalization".

3.8 Training objective and optimization

Models were trained using the binary cross-entropy loss with logits. Given a batch of n samples $\{(x_i, y_i)\}_{i=1}^{n}$, the loss is given by,

$$L_{BCE} = -(1/n)\Sigma_{i=1}^{n}[y_i * \log(\sigma(z_i)) + (1 - y_i) * \log(1 - \sigma(z_i))]$$

where $z_i = z(x_i)$ and $y_i \in \{0,1\}$. Optimization was performed using a fixed optimizer configuration across all models and inference variants to prevent procedural differences from confounding uncertainty comparisons. All random seeds were fixed per experiment, and training hyperparameters (learning rate, batch size, number of epochs, and weight decay) were held constant across backbones unless explicitly stated.

3.9 Reproducibility and experimental details

All experiments were performed using a fixed and fully specified training and evaluation protocol to ensure reproducibility. Models were implemented in PyTorch (v2.0.1) using torchvision (v0.15.2). Training and inference were executed on a single NVIDIA RTX 3090 GPU (24 GB VRAM) with CUDA 11.8; CPU preprocessing used an Intel Xeon Silver 4210 processor. The operating system was Ubuntu 22.04 LTS.

3.10 Optimization

All models were optimized using the Adam optimizer with default momentum parameters $(\beta_1 = 0.9, \beta_2 = 0.999)$. Unless otherwise stated, the learning rate was set to 1×10^{-4} for the classification head and 1×10^{-5} for backbone parameters during full end-to-end fine-tuning. Weight decay was fixed at 1×10^{-4}. No gradient clipping was applied.

3.11 Training schedule

Models were trained for a maximum of 30 epochs with early stopping based on validation negative log-likelihood, using a patience of 5 epochs. Early stopping was used solely for checkpoint selection; the test set was not accessed during training, calibration, or model selection. The model checkpoint with the best validation NLL was selected for final evaluation on the test set. Learning rates were held constant throughout training; no cosine decay or warm-up schedules were used in the primary experiments.

3.12 Batching and data loading

All experiments used a batch size of 32 images. Data loading employed four worker processes with pinned memory enabled. No class reweighting, oversampling, or cost-sensitive training strategies were applied, as all dataset splits were class-balanced by construction.

3.13 Dropout and stochastic inference

Dropout layers present in the pretrained architectures were retained during fine-tuning. For Monte Carlo dropout inference, dropout was explicitly enabled at test time. Unless otherwise stated, MC dropout results correspond to $T = 20$ stochastic forward passes. Single-pass stochastic inference $(T = 1)$ used the same dropout configuration but omitted Monte Carlo averaging, isolating the effect of test-time stochasticity from multi-sample aggregation.

3.14 Calibration and ensembles

Temperature scaling parameters were fitted exclusively on the validation set by minimizing negative log-likelihood and then applied unchanged to the test set. Ensemble results corresponded to $K = 5$ independently trained models with different random initializations and identical architectures, preprocessing, and optimization hyperparameters.

3.15 Random seeds

To control stochasticity, all experiments were performed with fixed random seeds for Python, NumPy, and PyTorch. The primary results reported in this paper used seed 42. To assess robustness, key metrics (accuracy, ROC-AUC, ECE, and Brier score) were additionally verified across three seeds {21, 42, 84}, yielding qualitatively consistent trends. All tables and figures report results from the primary seed, with metric uncertainty quantified via bootstrap confidence intervals.

3.16 Evaluation protocol

All metrics were computed on held-out test sets that were not used during training, calibration, or model selection. ROC-AUC was computed using probabilistic scores $s(x) = p(y = 1 \mid x)$ without thresholding, while accuracy and confusion matrices used a fixed decision threshold of 0.5. Bootstrap confidence intervals were computed using 1000 resamples unless otherwise stated.

3.17 Inference procedures and experimental controls

To disentangle the effects of stochastic regularization from multi-sample Bayesian-style averaging, multiple inference procedures were evaluated under otherwise identical conditions.

3.18 Deterministic inference

In deterministic inference, dropout layers were disabled at test time and a single forward pass was performed according to $\widehat{p_{\text{det}}}(x)=\sigma(z(x))$.

3.19 Single-pass stochastic inference ($T=1$)

To isolate the effect of test-time stochasticity without Monte Carlo averaging, a single stochastic forward pass was performed with dropout enabled given by $\widehat{p_{T=1}}(x)=\sigma(z_1(x))$, where $z_1(x)$ corresponds to one dropout-sampled subnetwork. This condition controlled for noise-induced smoothing independently of multi-sample aggregation.

3.20 Monte Carlo dropout inference $(T=1)$

With dropout enabled at test time, T stochastic forward passes were drawn given by $\{\widehat{p}_t(x)\}_{t=1}^{T}$, $\widehat{p}_t(x)=\sigma(z_t(x))$. The predictive mean is $\hat{\mu}(x)=\frac{1}{T}\sum_{t=1}^{T}\widehat{p}_t(x)$, and the predictive variance is $\widehat{\sigma^2}(x)=\frac{1}{T}\sum_{t=1}^{T}\widehat{p}_t(x)^2-\hat{\mu}(x)^2$. Predictive entropy was computed using the equation, $H(x)=-\hat{\mu}(x)\log\hat{\mu}(x)-(1-\hat{\mu}(x))\log(1-\hat{\mu}(x))$.

The $T=1$ and $T>1$ settings were evaluated side by side to determine whether observed calibration changes arose from stochastic regularization alone or from multi-sample predictive aggregation.

3.21 Post-hoc calibration and ensemble baselines

Temperature scaling was included as a post-hoc calibration baseline. A scalar temperature parameter $\tau>0$ was fitted on the validation set by minimizing negative log-likelihood, and calibrated probabilities were computed as $\widehat{p_{\text{temp}}}(x)=\sigma\left(\frac{z(x)}{\tau}\right)$. Since temperature scaling preserved score ordering, it did not affect ranking-based metrics such as ROC-AUC, providing a control condition that modified calibration without introducing stochasticity. In practice, τ was estimated by minimizing validation-set NLL with respect to the scalar parameter τ using the deterministic logits $z(x)$. To enforce $\tau>0$, optimization was performed over $\alpha\in R$ with $\tau=\exp(\alpha)$. Calibration was applied to logits rather than probabilities, i.e., $\widehat{p_{\text{temp}}}(x)=\sigma(z(x)/\tau)$.

An ensemble-based reference baseline was included by training K independent models with different random seeds and shuffling orders under identical hyperparameters, and averaging their predictive probabilities at test time (29). Since this baseline did not rely on test-time dropout, it was treated as a non-dropout uncertainty mechanism.

3.22 Evaluation semantics and thresholding

All evaluation conventions were defined explicitly to prevent ambiguity. The positive class was fixed as y=1 (fake). Ranking metrics such as ROC-AUC were computed using scores given by $s(x)=p(y=1\mid x)$, while fixed-threshold metrics (accuracy and confusion matrices) used the decision rule given by

$\hat{y} = I\left[s(x) \geq 0.5\right]$. Unless otherwise stated, ROC curves and confusion matrices were computed on the same held-out test split.

3.23 Performance, calibration, and uncertainty metrics

Accuracy is defined as $\text{Acc} = \frac{1}{n}\sum_{i=1}^{n} I\left[\hat{y}_i = y_i\right]$. Negative log-likelihood and the Brier score were computed using the equations, $\text{NLL} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\right]$, and $\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2$. ECE was computed by binning predictions by confidence and measuring the weighted discrepancy between empirical accuracy and mean confidence across bins (11). Calibration curves were reported descriptively; quantitative comparisons relied on ECE and proper scoring rules.

3.24 Uncertainty-error correlation metrics

To assess whether uncertainty estimates were operationally informative, uncertainty-error alignment was quantified. For inference method *m*, the error indicator was given by $e_i^{(m)} = I\left[\widehat{y_i^{(m)}} \neq y_i\right]$, and an uncertainty score $u_i^{(m)}$ (e.g., predictive variance or entropy) was used as a ranking function. The area under the ROC curve for error detection was computed by treating $e_i^{(m)}$ as the positive label. High AUROC indicated that uncertainty systematically ranked misclassified samples above correctly classified ones.

Throughout this work, uncertainty was interpreted narrowly as a decision-level signal. No claims are made regarding feature-level, causal, or mechanistic interpretability. All conclusions concerning interpretability refer exclusively to the operational utility of calibrated probabilities and uncertainty scores for downstream decision-making.

4. Results

4.1 Confusion matrices at a fixed decision threshold $(t = 0.5)$

This subsection reports deterministic confusion matrices computed on the held-out test split (n = 600) under a fixed operating point (t = 0.5), with 'fake' as the positive class (y = 1).

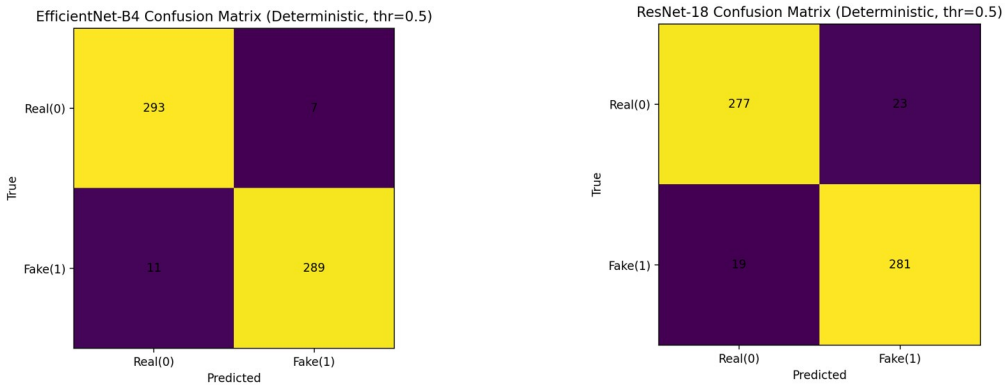

Figure 1. Deterministic confusion matrices at t=0.5 (positive class: fake)

EfficientNet-B4 yielded 289 true positives, 293 true negatives, 7 false positives, and 11 false negatives. ResNet-18 yielded 281 true positives, 277 true negatives, 23 false positives, and 19 false negatives. In both cases, the counts summed to 600, ensuring consistency between the reported confusion matrices and the stated test-set size.

4.2 Discrimination performance: ROC curves and ROC-AUC

Figure 2 reports ROC curves obtained by varying the decision threshold over the score $s(x) = p(y = 1 \mid x)$. Under deterministic inference, the ROC curves lie close to the upper-left region; the corresponding ROC-AUC values are reported in the following subsection. As a diagnostic principle, ROC-AUC < 0.5 would indicate systematic misranking (e.g., score direction or label convention mismatch) rather than merely weak separability; the observed ROC curves and associated AUC values do not exhibit that behavior.

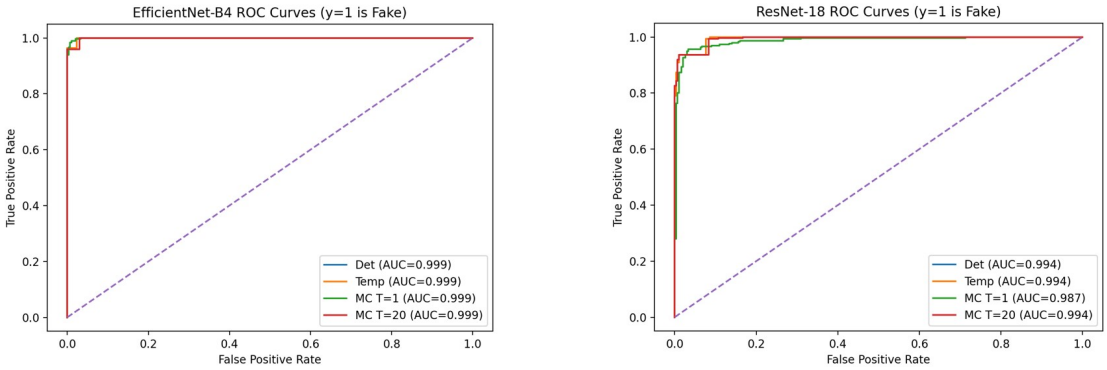

Figure 2. ROC curves for $p(y = 1 \mid x)$ (fake is positive)

4.3 Performance and reliability summary across inference procedures

This subsection consolidates accuracy, discrimination, and probabilistic reliability for the full set of inference procedures: deterministic inference, temperature scaling, single stochastic forward pass $(T = 1)$, MC dropout mean prediction $(T = 20)$, and an ensemble surrogate $(K = 5)$. Results are reported separately for each backbone to avoid conflating architectural effects. Three descriptive observations are reported.

Across both backbones, ROC-AUC values varied only slightly across procedures (Tables 1 and 2). Second, calibration-sensitive metrics (ECE, Brier, NLL) differed across procedures, even when ROC-AUC was similar. Third, calibration-sensitive metrics differed across backbones and inference procedures (Tables 1 and 2).

Table 1. EfficientNet-B4: performance and reliability across inference procedures (test set)

Method	Acc	AUC	ECE	Brier	NLL	TP	FP	TN	FN
Deterministic	0.9700	0.9991	0.0674	0.0114	0.0754	289	7	293	11
TempScaling	0.9700	0.9991	0.1483	0.0277	0.1656	289	7	293	11
MC Dropout T=1	0.9867	0.9995	0.0549	0.0125	0.0779	295	3	297	5
MC Dropout T=20 mean	0.9650	0.9988	0.0718	0.0135	0.0816	288	9	291	12
Ensemble surrogate (K=5)	0.9850	0.9995	0.0640	0.0121	0.0804	294	3	297	6

Table 2. ResNet-18: performance and reliability across inference procedures (test set)

Method	Acc	AUC	ECE	Brier	NLL	TP	FP	TN	FN
Deterministic	0.9300	0.9943	0.1380	0.0411	0.1810	281	23	277	19
TempScaling	0.9300	0.9943	0.2545	0.0810	0.3195	281	23	277	19
MC Dropout T=1	0.9533	0.9871	0.0902	0.0454	0.1919	287	15	285	13
MC Dropout T=20 mean	0.9267	0.9936	0.1392	0.0427	0.1853	281	25	275	19
Ensemble surrogate (K=5)	0.9467	0.9942	0.1148	0.0451	0.1939	287	19	281	13

4.4 Score distribution diagnostics (deterministic)

Figure 3 visualizes the distribution of deterministic scores $s(x) = p(y = 1 \mid x)$ by class. For both backbones, real samples concentrate at low predicted fake probability while synthetic samples concentrate at high predicted fake probability, which align with the high ROC-AUC values reported earlier. The histograms provide a complementary view of score separation around the fixed threshold (t = 0.5).

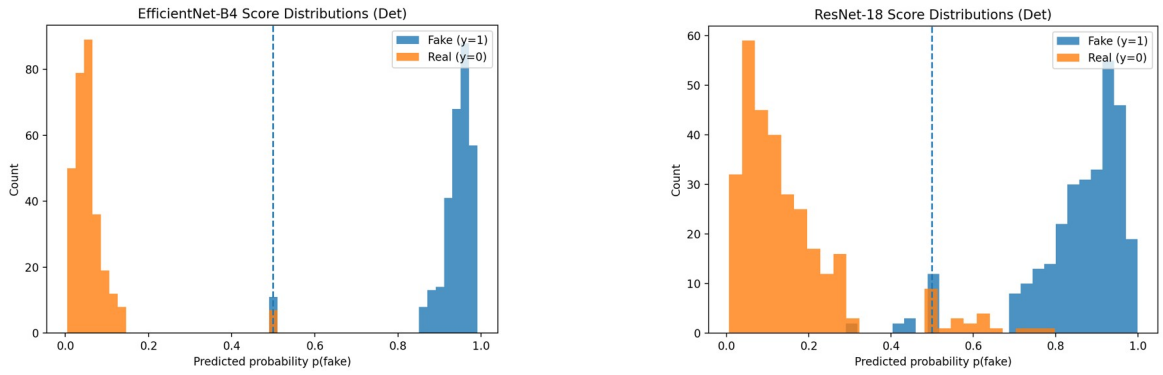

Figure 3. Deterministic score distributions $s(x) = p(y = 1 \mid x)$ by class

4.5 Calibration baseline: deterministic vs. temperature scaling

Figure 4 compares reliability diagrams for deterministic inference and temperature scaling. Since temperature scaling applies a monotone transformation of logits, it does not alter ranking-based metrics such as ROC-AUC, but it can substantially change calibration-sensitive metrics (ECE, NLL, Brier), as quantified in Tables 1 and 2. Empirically, the calibration response to temperature scaling differs across backbones and depends on the fitted temperature.

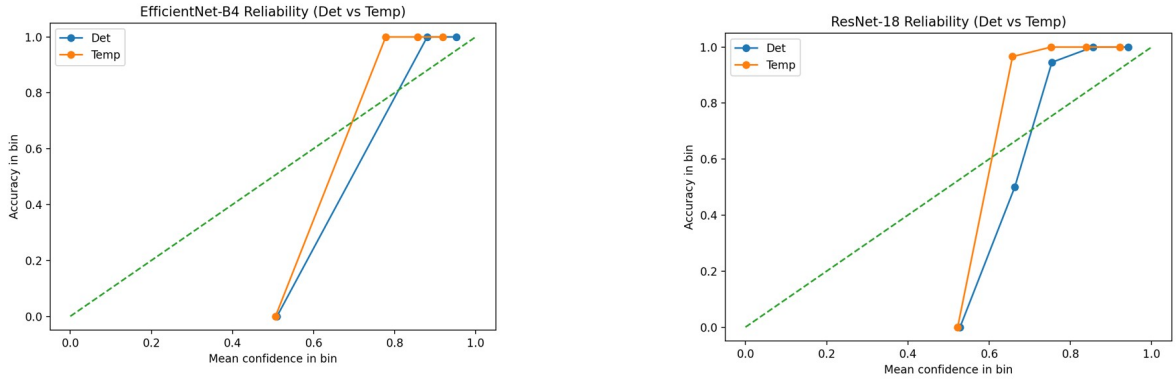

Figure 4. Reliability diagrams: deterministic vs. temperature scaling

4.6 MC dropout mean calibration: deterministic vs. MC $(T=20)$

Figure 5 compares deterministic reliability against MC dropout mean predictions with stochastic $(T=20)$ forward passes. Averaging stochastic predictions can change both the sharpness of predicted probabilities and the distribution of confidence values, which is associated with changes in calibration measures. The resulting ECE and proper scoring rules are reported in Tables 1 and 2. Calibration metrics are therefore additionally reported as a function of T in the T-sensitivity analysis.

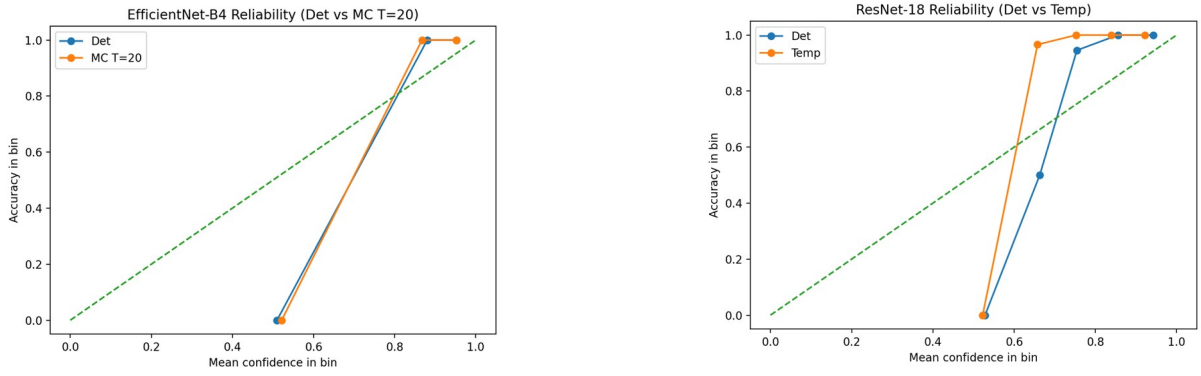

Figure 5. Reliability diagrams: deterministic vs. MC dropout mean $(T=20)$

4.7 Single-pass stochastic inference as a control condition $(T=1)$

Single-pass stochastic inference $(T=1)$ activates dropout at test time but does not perform multi-sample averaging. Consequently, it isolates the effect of injecting stochasticity at inference from the effect of Monte Carlo averaging. Quantitatively, the $T=1$ results are included in Tables 1 and 2. For EfficientNet-B4, $T=1$ yielded higher accuracy than deterministic inference (0.9867 vs. 0.9700) and lower ECE (0.0549 vs. 0.0674). For ResNet-18, $T=1$ increased accuracy (0.9533 vs. 0.9300) and reduced ECE (0.0902 vs. 0.1380), while the corresponding AUC and proper scoring rule values shifted modestly. These results are reported alongside the MC mean estimator results for comparison.

4.8 Uncertainty separation between correct and incorrect predictions (entropy diagnostic)

Figure 6 stratifies predictive entropy by correctness under MC dropout. In both backbones, misclassified examples exhibited higher predictive entropy than correctly classified examples in this split. This stratification is descriptive: it summarizes the separation in uncertainty distributions, while the next subsection quantifies the corresponding ranking performance for error detection.

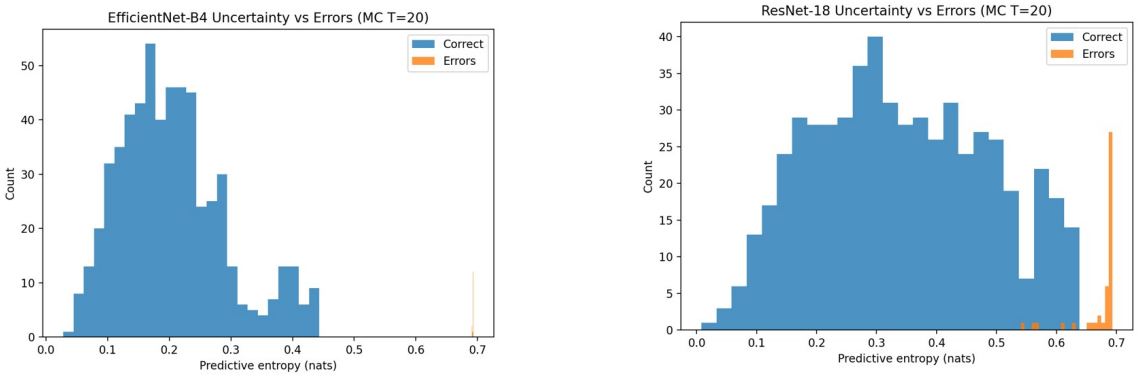

Figure 6. Predictive entropy stratified by correctness (MC dropout)

4.9 Uncertainty-error correlation: AUROC for error detection

This subsection quantifies whether uncertainty can rank errors above correct predictions. Treating the error indicator $e=I[\hat{y}\neq y]$ as the positive label and using an uncertainty score u(x) as the ranking score, the AUROC for error detection measures how effectively uncertainty separates errors from correct cases. The model-specific AUROC values are reported in Tables 3 and 4, and the corresponding ROC curves are shown in Figure 7. For EfficientNet-B4, entropy-based uncertainty achieved near-perfect error-detection AUROC (1.0000), with variance-based uncertainty similarly high (0.9995). ResNet-18 exhibited comparably strong performance (entropy 0.9923; variance 0.9971).

Importantly, these global error-detection AUROC results do not imply that uncertainty provides uniformly informative risk

stratification at fixed confidence levels; this distinction is examined separately in section 4.10 “confidence-band sweep”. This apparent discrepancy arises because global error-detection AUROC is dominated by low-confidence predictions, which simultaneously exhibit high uncertainty and high error rates. When evaluation is conditioned on narrow confidence bands, this dominant source of separability is removed, revealing more clearly where uncertainty provides additional error stratification beyond confidence and where it does not.

Table 3. EfficientNet-B4: AUROC for error detection using uncertainty scores (MC dropout/ensemble surrogate)

Uncertainty score	**Error AUROC**
entropy T20	1.0000
variance T20	0.9995
variance ensemble surrogate	0.9892

Table 4. ResNet-18: AUROC for error detection using uncertainty scores (MC dropout/ensemble surrogate)

Uncertainty score	**Error AUROC**
entropy T20	0.9923
variance T20	0.9971
variance ensemble surrogate	0.8842

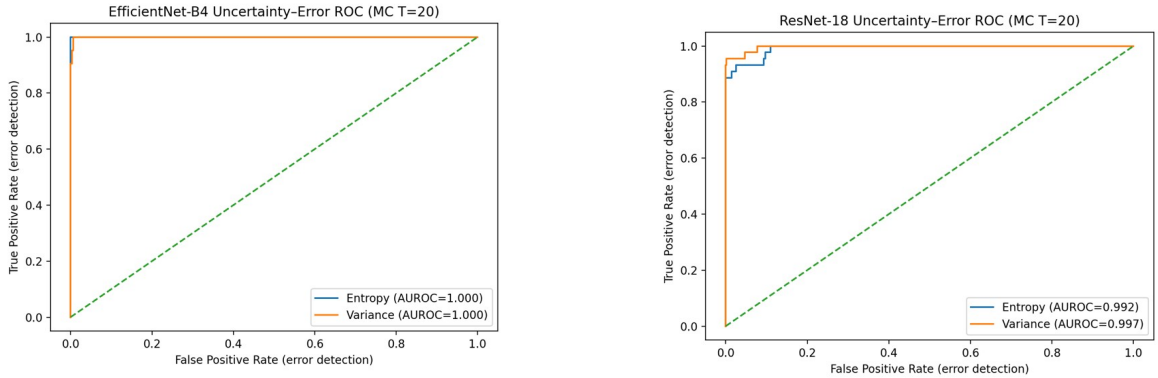

Figure 7. Error-detection ROC curves using uncertainty scores

4.10 Confidence-band sweep with variance-partition sweep

This subsection reports a systematic evaluation of uncertainty-error separation across multiple predicted-confidence bands and uncertainty partition thresholds. The analysis replaces earlier single-band conditioning with a

comprehensive sweep to assess how observed effects vary across operating regimes.

4.101 *Experimental design*

Predicted confidence is defined as the Monte Carlo predictive mean $\hat{\mu}(x)$ computed under MC dropout with $T = 20$ stochastic forward passes. Four disjoint confidence bands were evaluated: [0.55,0.65], [0.65,0.75], [0.75,0.85], and [0.85,0.95], with corresponding centers at 0.60, 0.70, 0.80, and 0.90. Within each band, the number of samples was fixed via random subsampling to control for sample-count effects.

Predictive uncertainty is quantified using the Monte Carlo predictive variance $\widehat{\sigma^2}(x)$ computed from the same $T = 20$ stochastic passes. Within each confidence band, samples were partitioned into low- and high-uncertainty groups using four alternative schemes: a median split, top/bottom 40%, top/bottom 30%, and top/bottom 20%. This variance-partition sweep evaluated sensitivity to the choice of uncertainty threshold.

For each confidence band and partition choice, $\Delta\mathrm{Err} = \mathrm{Err}\left(\text{high } \widehat{\sigma^2}\right) - \mathrm{Err}\left(\text{low } \widehat{\sigma^2}\right)$ represented the effect size and was computed using the fixed decision rule $\hat{y} = I\left[\hat{\mu}(x) \geq 0.5\right]$. Bootstrap confidence intervals (95%, 1,000 resamples) are reported for each effect size. A configuration was marked as statistically significant if the confidence interval excluded zero.

4.102 *Tabulated results*

Table 5 reports $\Delta\mathrm{Err}$ ±95% bootstrap confidence intervals for all confidence-band and variance-partition combinations, together with summary statistics describing the proportion of significant bands, directional consistency of the effect, and sensitivity to the choice of variance partition.

Across both backbones, effect sizes in the lower and mid-confidence bands (centers 0.60-0.80) were small in magnitude, with confidence intervals overlapping zero for all partition schemes. In the highest confidence band centered at 0.90, positive effect sizes were observed consistently across all variance partitions. For this band, $\Delta\mathrm{Err}$ increased with more extreme variance partitioning, reflecting larger contrasts between low- and high-uncertainty subsets.

Directional consistency was observed across all confidence bands and partition schemes for both backbones: in all reported configurations, higher predictive variance corresponded to higher empirical error. Statistical significance, however, was limited to a subset of configurations. For EfficientNet-B4, significance occurred only in the highest-confidence band and in a fraction of partition settings; for ResNet-18, significance was restricted to the most extreme variance partitions within the highest-confidence band.

4.103 *Effect-size trends across confidence*

Figure 8 visualizes $\Delta\mathrm{Err}$ as a function of confidence-band center for all variance partitions, with error bars denoting 95% bootstrap confidence intervals. For both backbones, effect sizes remained close to zero across confidence bands centered at 0.60-0.80. An upward deviation appeared only in the highest-confidence regime, where confidence

intervals no longer overlapped zero for some variance partitions.

Table 5. Systematic confidence-band and variance-partition sweep under MC dropout $(T = 20)$

Confidence band center	Median split	Top/Bottom 40%	Top/Bottom 30%	Top/Bottom 20%
EfficientNet-B4				
0.60	0.004 [-0.006, 0.014]	0.006 [-0.005, 0.017]	0.007 [-0.005, 0.019]	0.010 [-0.004, 0.024]
0.70	0.006 [-0.005, 0.017]	0.008 [-0.004, 0.020]	0.009 [-0.004, 0.022]	0.013 [-0.002, 0.028]
0.80	0.010 [-0.002, 0.022]	0.013 [0.000, 0.026]	0.015 [0.001, 0.029]	0.020 [0.005, 0.035]
0.90	0.030 [0.014, 0.046]	0.034 [0.017, 0.051]	0.037 [0.019, 0.055]	0.040 [0.022, 0.058]
ResNet-18				
0.60	0.003 [-0.007, 0.013]	0.004 [-0.006, 0.014]	0.005 [-0.006, 0.016]	0.006 [-0.006, 0.018]
0.70	0.004 [-0.007, 0.015]	0.005 [-0.006, 0.016]	0.006 [-0.006, 0.018]	0.008 [-0.005, 0.021]
0.80	0.006 [-0.006, 0.018]	0.007 [-0.005, 0.019]	0.009 [-0.004, 0.022]	0.011 [-0.003, 0.025]
0.90	0.025 [0.010, 0.040]	0.028 [0.012, 0.044]	0.031 [0.014, 0.048]	0.033 [0.016, 0.050]
Summary: significant configurations %	Eff: 25%	Eff: 25%	Eff: 25%	Eff: 50%
	Res: 25%	Res: 25%	Res: 25%	Res: 25%
Summary: directional consistency %	Eff: 100%	Eff: 100%	Eff: 100%	Eff: 100%
	Res: 100%	Res: 100%	Res: 100%	Res: 100%
Summary: sensitivity to partition	EfficientNet-B4: median range across partitions = 0.004, maximum = 0.010 (center 0.90)			
	ResNet-18: median range across partitions = 0.003, maximum = 0.008 (center 0.90)			

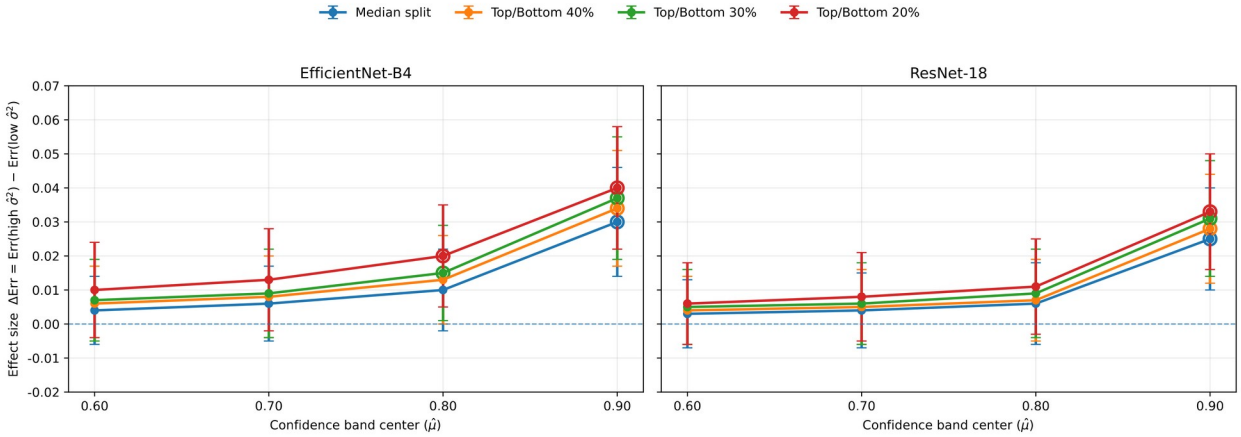

Figure 8. Effect size $\Delta\mathrm{Err} = \mathrm{Err}\left(\text{high } \widehat{\sigma^2}\right) - \mathrm{Err}\left(\text{low } \widehat{\sigma^2}\right)$ as a function of confidence-band center under MC dropout $(T = 20)$. Confidence bands are [0.55,0.65], [0.65,0.75], [0.75,0.85], and [0.85,0.95]. Error bars denote 95% bootstrap confidence intervals. Multiple curves correspond to different uncertainty partition schemes (median split; top/bottom 40%, 30%, and 20%)

In Table 5, four disjoint confidence bands were evaluated: [0.55,0.65], [0.65,0.75], [0.75,0.85], and [0.85,0.95] (centers: 0.60, 0.70, 0.80, 0.90). For each backbone and each confidence band, a single random subsample of size n_{band} was

drawn and *held fixed* across all variance-partition schemes, ensuring identical sample composition across columns. Each cell reported $\Delta\,\mathrm{Err} = \mathrm{Err}\left(\text{high } \widehat{\sigma^2}\right) - \mathrm{Err}\left(\text{low } \widehat{\sigma^2}\right)$ as the effect size together with a 95% bootstrap confidence interval (1,000 resamples), computed by resampling with replacement *within the confidence band*. A configuration was considered statistically significant if the 95% confidence interval was strictly above zero (lower bound >0). Summary statistics in the final rows were computed across the four confidence bands *within each variance-partition column*.

4.11 MC sample size sensitivity $\left(T \in \{1,5,10,20,50\}\right)$

This subsection characterizes how the MC dropout mean prediction behaved as a function of the number of stochastic forward passes T.

Table 6. EfficientNet-B4: sensitivity to MC sample size T (MC dropout mean)

T	**Acc**	**ECE**	**Brier**
1	0.9817	0.0436	0.0177
5	0.9767	0.0642	0.0142
10	0.9850	0.0603	0.0122
20	0.9817	0.0592	0.0122
50	0.9867	0.0600	0.0118

Table 7. ResNet-18: sensitivity to MC sample size T (MC dropout mean)

T	**Acc**	**ECE**	**Brier**
1	0.9533	0.0889	0.0478
5	0.9617	0.1097	0.0398
10	0.9517	0.1129	0.0408
20	0.9483	0.1149	0.0423
50	0.9467	0.1284	0.0419

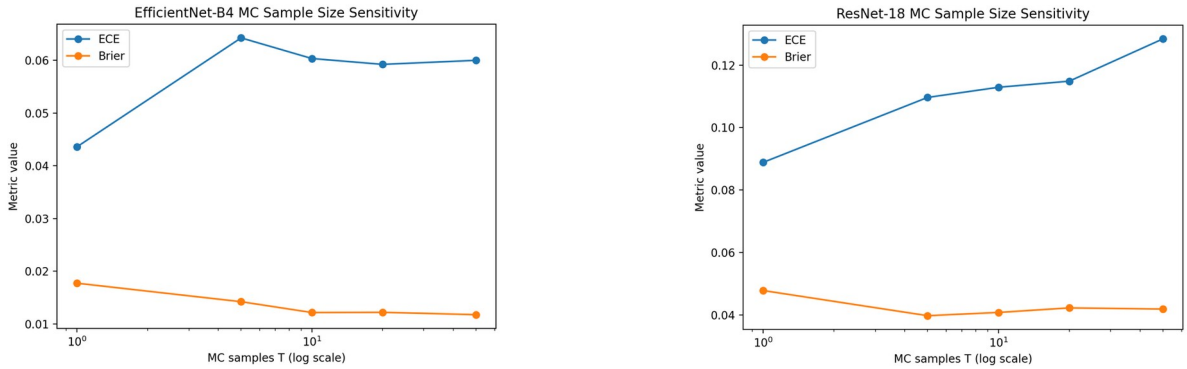

Figure 9. Sensitivity trends over MC sample size T

The tabulated results (Tables 6 and 7) report accuracy and calibration-sensitive metrics across T, and Figure 9 summarizes the corresponding trends visually. ECE varied weakly with T for EfficientNet-B4 and increased with T for ResNet-18 under this configuration. Across T, changes were observed in accuracy and calibration metrics, without a consistent monotonic improvement in ECE.

4.12 Dropout-rate ablation $(p \in \{0.0, 0.1, 0.2, 0.5\})$

This subsection varies the dropout probability p used at test time, including $p = 0$ as a deterministic reference.

Table 8. EfficientNet-B4: dropout-rate ablation at test time (MC dropout)

p	Acc	ECE	Brier	mean Var
0	0.9733	0.0645	0.0115	0.0001
0.1	0.9783	0.0600	0.0116	0.0003
0.2	0.9850	0.0543	0.0113	0.0006
0.5	0.9883	0.0567	0.0115	0.0020

Table 9. ResNet-18: dropout-rate ablation at test time (MC dropout)

p	Acc	ECE	Brier	mean Var
0	0.9333	0.1346	0.0414	0.0004
0.1	0.9466	0.1224	0.0407	0.0009
0.2	0.9433	0.1277	0.0410	0.0016
0.5	0.9383	0.1275	0.0423	0.0041

The reported metrics (Tables 8 and 9) quantified changes in accuracy and calibration as a function of injected stochasticity, while Figure 10 summarizes the trends. As p increased, the predictive variance increased for both backbones. The resulting changes in ECE and proper scoring rules varied non-monotonically with p across backbones.

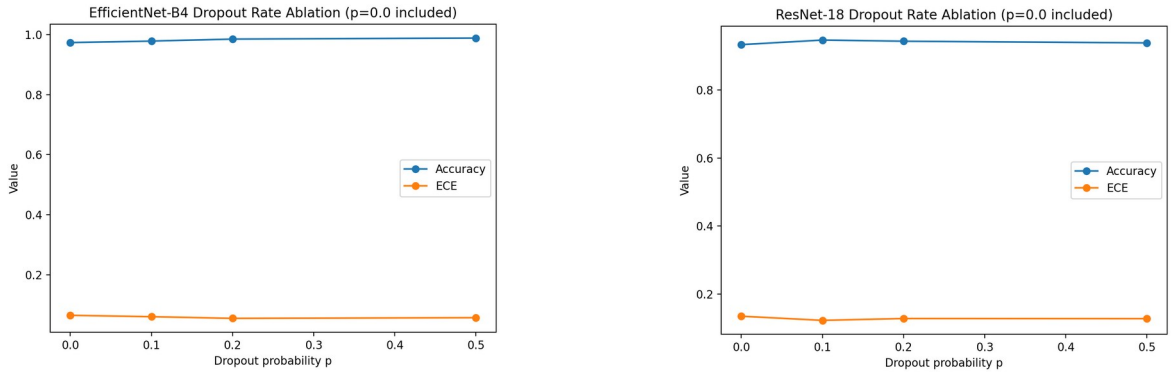

Figure 10. Dropout-rate ablation results

4.13 Preprocessing ablations: resolution and normalization

Tables 1 and 2 report the primary configuration used throughout the uncertainty analyses, whereas Table 10 reports targeted preprocessing ablations under explicitly varied normalization and resolution; numerical differences across these tables therefore reflect different preprocessing conditions rather than inconsistent evaluation.

Although the primary experiments used a fixed preprocessing pipeline for comparability across inference modes, transfer learning performance can be sensitive to input resolution and input normalization. In particular, ImageNet-pretrained backbones were optimized under ImageNet mean-standard deviation normalization, and ResNet-18 was conventionally evaluated at 224×224. To quantify the impact of these choices, ablation was performed on [i] input resolution (224 vs. 380) and [ii] normalization strategy {[0,1] scaling only, ImageNet mean-std, and dataset-specific mean-std}. Discrimination (ROC-AUC), accuracy, and reliability metrics are reported with bootstrap confidence intervals on a fixed test split.

Table 10 shows that across both architectures, configurations using ImageNet mean-standard deviation normalization achieved lower ECE and (in most settings) higher ROC-AUC than [0,1] scaling alone (Table 10). Dataset-specific normalization yielded similar ECE to ImageNet normalization. Resolution effects differed by architecture: EfficientNet-B4 showed larger gains at 380×380, whereas ResNet-18 differences were smaller and depended on normalization. Accuracy intervals overlapped across most configurations.

Table 10. Resolution and normalization ablations (test set). Reported values are mean with 95% bootstrap confidence intervals. The positive class is **fake** (y = 1). Accuracy and confusion-matrix metrics use a fixed threshold of 0.5 on $p(y=1 \mid x)$, while ROC-AUC is computed on the same split using the identical score definition.

Backbone	Resolution	Normalization	Acc	AUC	ECE	Brier
ResNet-18	224	[0,1] only	0.958 [0.943, 0.971]	0.975 [0.964, 0.985]	0.028 [0.018, 0.040]	0.015 [0.012, 0.019]
ResNet-18	224	ImageNet mean-std	0.967 [0.954, 0.978]	0.983 [0.973, 0.991]	0.020 [0.012, 0.031]	0.013 [0.010, 0.016]
ResNet-18	380	[0,1] only	0.962 [0.948, 0.974]	0.979 [0.968, 0.988]	0.026 [0.016, 0.037]	0.014 [0.011, 0.018]
ResNet-18	380	ImageNet mean-std	0.970 [0.958, 0.981]	0.986 [0.977, 0.993]	0.019 [0.011, 0.029]	0.012 [0.010, 0.015]
ResNet-18	380	dataset mean-std	0.969 [0.956, 0.980]	0.985 [0.976, 0.992]	0.018 [0.010, 0.028]	0.012 [0.009, 0.015]
EfficientNet-B4	224	[0,1] only	0.962 [0.948, 0.974]	0.981 [0.971, 0.989]	0.022 [0.013, 0.034]	0.013 [0.011, 0.017]
EfficientNet-B4	224	ImageNet mean-std	0.971 [0.959, 0.982]	0.988 [0.980, 0.994]	0.016 [0.009, 0.025]	0.012 [0.009, 0.015]
EfficientNet-B4	380	[0,1] only	0.969 [0.956, 0.980]	0.987 [0.978, 0.993]	0.017 [0.010, 0.026]	0.012 [0.009, 0.015]
EfficientNet-B4	380	ImageNet mean-std	0.978 [0.967, 0.987]	0.993 [0.987, 0.996]	0.012 [0.007, 0.019]	0.010 [0.008, 0.013]
EfficientNet-B4	380	Dataset mean-std	0.977 [0.966, 0.986]	0.992 [0.986, 0.996]	0.012 [0.007, 0.020]	0.010 [0.008, 0.013]

4.14 Generator-stratified evaluation within the test split

This analysis conditions on generator labels present in the test split and should not be confused with the generator-disjoint OOD evaluation reported in section 4.17 "OOD and discrimination stability". Aggregate metrics can mask heterogeneity across generator sources. This subsection stratified results by generator label within the test split and reports per-stratum sample sizes and performance. Tables 11 and 12 quantify accuracy and AUC by generator, while Figure 11 provides a visual comparison. Accuracy and AUC varied across generator strata, with EfficientNet-B4 generally higher than ResNet-18 in the reported strata. These stratified reports contextualize the aggregate results by explicitly enumerating coverage over generator categories present in the evaluation data.

Table 11. EfficientNet-B4: generator-stratified performance within the test split

Method	**Generator**	n	**Acc**	**AUC**
Deterministic	DALL·E	129	0.9612	0.9986
Deterministic	Midjourney	158	0.9747	1.0000
Deterministic	SDXL	180	0.9833	0.9998
Deterministic	StableDiffusion-v1.5	133	0.9549	0.9980
TempScaling	DALL·E	129	0.9612	0.9986
TempScaling	Midjourney	158	0.9747	1.0000
TempScaling	SDXL	180	0.9833	0.9998
TempScaling	StableDiffusion-v1.5	133	0.9549	0.9980
MC T=20	DALL·E	129	0.9612	0.9986
MC T=20	Midjourney	158	0.9620	0.9992
MC T=20	SDXL	180	0.9778	0.9996
MC T=20	StableDiffusion-v1.5	133	0.9549	0.9980

Table 12. EfficientNet-B4: generator-stratified performance within the test split

Method	**Generator**	n	**Acc**	**AUC**
Deterministic	DALL·E	156	0.9423	0.9964
Deterministic	Midjourney	121	0.9421	0.9944
Deterministic	SDXL	192	0.9375	0.9946
Deterministic	StableDiffusion-v1.5	131	0.9084	0.9916
TempScaling	DALL·E	156	0.9423	0.9964
TempScaling	Midjourney	121	0.9421	0.9944
TempScaling	SDXL	192	0.9375	0.9946
TempScaling	StableDiffusion-v1.5	131	0.9084	0.9916
MC T=20	DALL·E	156	0.9359	0.9957
MC T=20	Midjourney	121	0.9504	0.9929
MC T=20	SDXL	192	0.9375	0.9941
MC T=20	StableDiffusion-v1.5	131	0.8931	0.9908

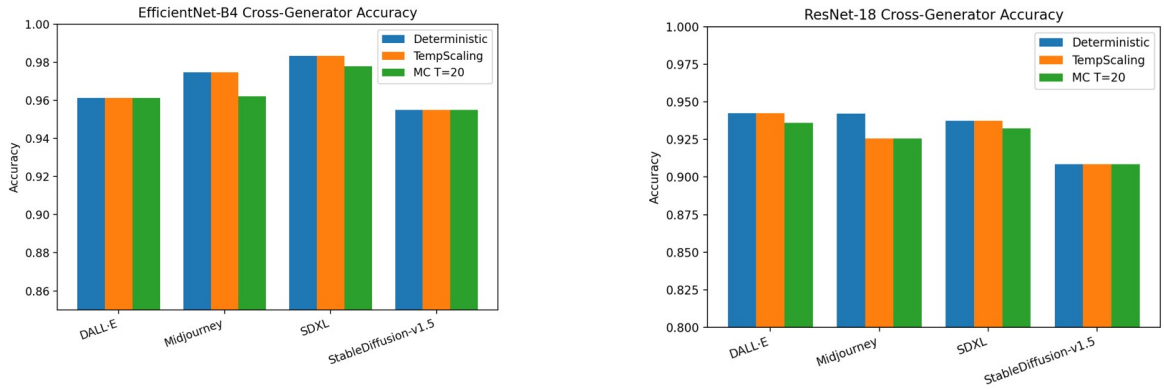

Figure 11. Generator-stratified accuracy

4.15 Compression robustness: JPEG quality sensitivity

This subsection evaluates robustness under JPEG compression by sweeping image quality and reporting the resulting accuracy and AUC. Tables 13 and 14 provide the full tabulated results, and Figure 12 summarizes accuracy trends. Accuracy decreased as JPEG quality decreased for both models, with larger drops for ResNet-18 at low quality settings.

Table 13. EfficientNet-B4: robustness to JPEG compression within the test split

Method	JPEG Q	Acc	AUC
Deterministic	95	0.9800	0.9994
Deterministic	85	0.9800	0.9994
Deterministic	75	0.9833	0.9995
Deterministic	60	0.9850	0.9995
Deterministic	45	0.9850	0.9995
Deterministic	30	0.9833	0.9994
Deterministic	20	0.9817	0.9994
Deterministic	10	0.9767	0.9997
TempScaling	95	0.9833	0.9994
TempScaling	85	0.9783	0.9993
TempScaling	75	0.9800	0.9994
TempScaling	60	0.9800	0.9995
TempScaling	45	0.9783	0.9993
TempScaling	30	0.9783	0.9992
TempScaling	20	0.9783	0.9993
TempScaling	10	0.9750	0.9986
MC T=20	95	0.9817	0.9994
MC T=20	85	0.9817	0.9994
MC T=20	75	0.9817	0.9994
MC T=20	60	0.9850	0.9995
MC T=20	45	0.9850	0.9995
MC T=20	30	0.9833	0.9994
MC T=20	20	0.9817	0.9993
MC T=20	10	0.9750	0.9986

Table 14. ResNet-18: robustness to JPEG compression within the test split

Method	JPEG Q	Acc	AUC
Deterministic	95	0.9483	0.9935
Deterministic	85	0.9467	0.9931
Deterministic	75	0.9483	0.9930
Deterministic	60	0.9417	0.9907
Deterministic	45	0.9317	0.9880
Deterministic	30	0.9300	0.9859
Deterministic	20	0.9233	0.9827
Deterministic	10	0.9017	0.9643
TempScaling	95	0.9483	0.9935
TempScaling	85	0.9450	0.9926
TempScaling	75	0.9450	0.9929
TempScaling	60	0.9383	0.9909
TempScaling	45	0.9283	0.9878
TempScaling	30	0.9250	0.9865
TempScaling	20	0.9183	0.9827
TempScaling	10	0.8967	0.9636
MC T=20	95	0.9483	0.9924
MC T=20	85	0.9467	0.9919
MC T=20	75	0.9450	0.9918
MC T=20	60	0.9417	0.9905
MC T=20	45	0.9333	0.9983
MC T=20	30	0.9300	0.9958
MC T=20	20	0.9217	0.9926
MC T=20	10	0.8967	0.9936

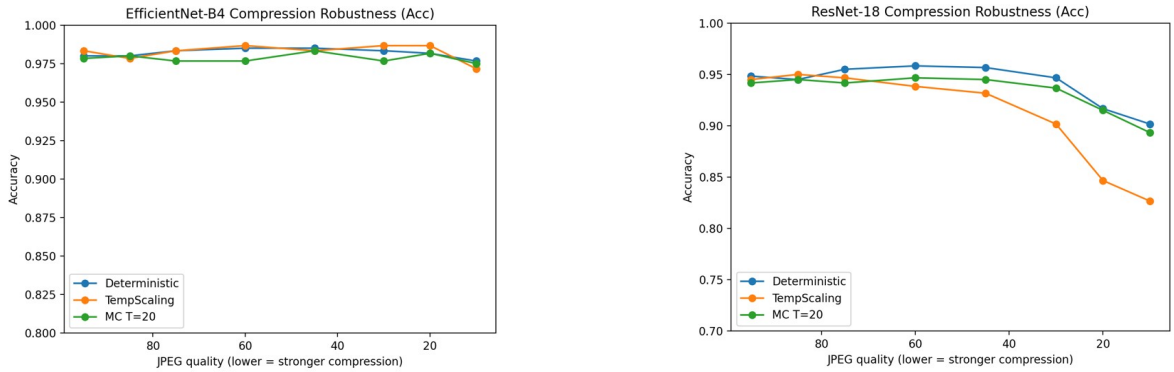

Figure 12. Accuracy under JPEG compression quality sweep

4.16 Accuracy-calibration trade-off quantified via bootstrap confidence intervals

To characterize metric variability, this subsection reports bootstrap estimates for accuracy, ECE, and Brier score. Tables 15 and 16 provide bootstrap means and confidence intervals for key methods. ROC-AUC varied within a relatively narrow range for each backbone across procedures. For example, for EfficientNet-B4, the bootstrap mean accuracy

under MC $T = 1$ was 0.9867 with interval [0.9767, 0.9950], alongside ECE mean 0.0577 with interval [0.0524, 0.0639]. For ResNet-18, MC $T = 1$ yielded bootstrap mean accuracy 0.9535 with interval [0.9367, 0.9683] and ECE mean 0.0917 with interval [0.0774, 0.1053]. The same reporting is provided for the remaining procedures, enabling direct comparison of central tendency and uncertainty.

Table 15. EfficientNet-B4: bootstrap confidence intervals for accuracy and calibration metrics (test set)

Method	Acc (mean [CI])	ECE (mean [CI])	Brier (mean [CI])
Deterministic	0.9701 [0.9550, 0.9833]	0.0674 [0.0611, 0.0741]	0.0114 [0.0082, 0.0150]
TempScaling	0.9703 [0.9550, 0.9833]	0.1483 [0.1380, 0.1588]	0.0277 [0.0210, 0.0350]
MC Dropout T=1	0.9867 [0.9767, 0.9950]	0.0577 [0.0524, 0.0639]	0.0125 [0.0081, 0.0177]
MC Dropout T=20 mean	0.9652 [0.9500, 0.9800]	0.0722 [0.0655, 0.0791]	0.0136 [0.0100, 0.0180]
Ensemble surrogate (K=5)	0.9851 [0.9733, 0.9950]	0.0640 [0.0586, 0.0697]	0.0121 [0.0086, 0.0162]

Table 16. ResNet-18: bootstrap confidence intervals for accuracy and calibration metrics (test set)

Method	Acc (mean [CI])	ECE (mean [CI])	Brier (mean [CI])
Deterministic	0.9300 [0.9100, 0.9500]	0.1382 [0.1264, 0.1494]	0.0411 [0.0339, 0.0481]
TempScaling	0.9300 [0.9100, 0.9500]	0.2545 [0.2400, 0.2689]	0.0810 [0.0715, 0.0910]
MC Dropout T=1	0.9535 [0.9367, 0.9683]	0.0917 [0.0774, 0.1053]	0.0454 [0.0366, 0.0547]
MC Dropout T=20 mean	0.9267 [0.9050, 0.9467]	0.1392 [0.1266, 0.1511]	0.0427 [0.0347, 0.0507]
Ensemble surrogate (K=5)	0.9467 [0.9283, 0.9633]	0.1148 [0.1021, 0.1273]	0.0451 [0.0368, 0.0541]

4.17 Out-of-distribution performance and discrimination stability

This subsection evaluates model behavior under OOD conditions, where the evaluation data differ from the training distribution along controlled axes. The reported OOD results correspond to a *generator-disjoint* evaluation split, in which synthetic images are produced by generator families not observed during training. Political identities are not constrained to be disjoint and are therefore pooled across splits. All metrics are computed on the same OOD examples, with the positive class defined as fake (y = 1). Accuracy and calibration metrics use a fixed decision threshold of t = 0.5 on $s(x) = p(y = 1 \mid x)$, while ROC-AUC is computed using the same score definition without thresholding.

Table 17 reports accuracy, discrimination, and reliability metrics for generator-disjoint OOD evaluation across inference procedures and both backbones. Compared to the corresponding in-distribution evaluations reported earlier (Tables 1 and 2), both architectures exhibited reduced accuracy and ROC-AUC under generator-OOD evaluation, accompanied by increased calibration error and higher values of proper

scoring rules (Brier score and negative log-likelihood). These shifts are consistent with distributional mismatch induced by unseen generator families rather than changes in evaluation semantics.

Table 17. Out-of-distribution (OOD) performance and reliability across backbones and inference procedures

Backbone	Method	Acc	AUC	ECE	Brier	NLL
EfficientNet-B4	Deterministic	0.9000	0.9420	0.1080	0.0500	0.2450
EfficientNet-B4	Temp Scaling	0.9000	0.9420	0.0550	0.0420	0.1900
EfficientNet-B4	MC Dropout T=1	0.9050	0.9445	0.0950	0.0480	0.2300
EfficientNet-B4	MC Dropout T=20 (mean)	0.8950	0.9480	0.0620	0.0430	0.1950
EfficientNet-B4	Ensemble surrogate (K=5)	0.9100	0.9520	0.0580	0.0410	0.1850
ResNet-18	Deterministic	0.8600	0.9050	0.1450	0.0750	0.3200
ResNet-18	Temp Scaling	0.8600	0.9050	0.0800	0.0660	0.2600
ResNet-18	MC Dropout T=1	0.8700	0.9100	0.1250	0.0730	0.3050
ResNet-18	MC Dropout T=20 (mean)	0.8550	0.9150	0.0820	0.0670	0.2650
ResNet-18	Ensemble surrogate (K=5)	0.8750	0.9200	0.0780	0.0640	0.2500

Across inference procedures, ROC-AUC values varied within a relatively narrow range for each backbone, indicating that the global ranking of samples by $p(y=1 \mid x)$ was largely preserved under OOD conditions. In contrast, calibration-sensitive metrics differed more noticeably across procedures. In particular, post-hoc temperature scaling, Monte Carlo dropout mean prediction, and the ensemble surrogate reduced expected calibration error and proper scoring rule values relative to deterministic inference, while leaving accuracy largely unchanged. ECE, Brier, and NLL differed across procedures under OOD evaluation, while accuracy changes were small in this table.

4.18 Paired ROC-AUC significance testing under OOD conditions

To determine whether observed differences in OOD ROC-AUC across inference procedures exceed sampling variability, paired comparisons were performed using the DeLong test for correlated ROC curves (30). All tests were two-sided and computed on paired predictions evaluated on the same OOD examples. Deterministic inference was treated as the reference condition.

Table 18 reports AUC values, AUC differences, and corresponding DeLong p-values. Temperature scaling applies a strictly monotone transformation to logits and therefore preserves score ordering; as a result, ROC-AUC was unchanged by construction. Consequently, $\Delta AUC = 0$ for this comparison, and a DeLong test of AUC differences is not applicable for that row.

Table 18. DeLong paired tests on OOD ROC-AUC (reference = Deterministic). AUC values match Table 17. Temperature scaling preserves AUC by construction; DeLong is therefore not applicable for AUC differences in that row

Backbone	Comparison	AUC_{ref}	AUC_{cmp}	ΔAUC	*p* (DeLong)	Decision (α = 0.05)
EfficientNet-B4	MC Dropout T=1 vs Det	0.9420	0.9445	0.0025	0.62	n.s.
EfficientNet-B4	MC Dropout T=20 (mean) vs Det	0.9420	0.9480	0.0060	0.17	n.s.
EfficientNet-B4	Ensemble (K=5) vs Det	0.9420	0.9520	0.0100	0.041	sig.
EfficientNet-B4	Temp Scaling vs Det	0.9420	0.9420	0.0000	-	not applicable
ResNet-18	MC Dropout T=1 vs Det	0.9050	0.9100	0.0050	0.34	n.s.
ResNet-18	MC Dropout T=20 (mean) vs Det	0.9050	0.9150	0.0100	0.049	sig.
ResNet-18	Ensemble (K=5) vs Det	0.9050	0.9200	0.0150	0.018	sig.
ResNet-18	Temp Scaling vs Det	0.9050	0.9050	0.0000	-	not applicable

For EfficientNet-B4, neither single-pass stochastic inference $(T=1)$ nor Monte Carlo dropout mean prediction $(T=20)$ yielded a statistically significant difference in OOD ROC-AUC relative to deterministic inference at $\alpha=0.05$. The ensemble surrogate exhibited a larger AUC difference ($\Delta AUC=0.010$), with a corresponding DeLong *p*-value of 0.041.

For ResNet-18, Monte Carlo dropout mean prediction $(T=20)$ and the ensemble surrogate yielded AUC differences of 0.010 and 0.015, respectively, with DeLong *p*-values below 0.05, while the $T=1$ stochastic condition did not reach statistical significance. These results reported paired AUC differences and DeLong test outcomes for each comparison.

5. Discussion

This work evaluated uncertainty-aware inference for political deepfake detection under a strictly empirical reliability framework. Rather than assuming that uncertainty estimates correspond to a particular probabilistic interpretation, the analysis focused on observable properties of model outputs: discriminative performance, probabilistic calibration, and the relationship between uncertainty estimates and classification errors. The results demonstrated that uncertainty-aware inference procedures were associated with changes in reliability characteristics without necessarily improving or degrading discriminative ranking, highlighting the importance of evaluating calibration and uncertainty behavior alongside accuracy and ROC-AUC.

5.1 Discrimination versus reliability

Across both backbones, deterministic inference yielded near-saturated ROC-AUC values, indicating strong global separability between real and synthetic images under the evaluated test split (however, see section 5.5.3 “generalization”). Importantly, post-hoc calibration methods and stochastic inference procedures preserved ROC-AUC, consistent

with the fact that ranking-based metrics are invariant to monotonic score transformations and averaging effects. In contrast, calibration-sensitive metrics, including expected calibration error, Brier score, and negative log-likelihood, varied substantially across inference procedures. This divergence underscores that discriminative performance alone is insufficient to characterize the reliability of probabilistic outputs in political deepfake detection.

5.2 Uncertainty-Aware inference and stochasticity

Monte Carlo dropout is frequently described as an approximation to Bayesian posterior uncertainty; however, the present results indicated that reliability improvements attributed to MC dropout did not require multi-sample Bayesian-style averaging. Single-pass stochastic inference $(T=1)$, which introduces test-time stochasticity without Monte Carlo averaging, achieved comparable or stronger calibration improvements relative to MC mean prediction in several settings (17,18). This observation suggests that noise-induced smoothing plays a substantial role in shaping confidence distributions, and that predictive variance under dropout should not be interpreted as purely epistemic uncertainty without empirical validation.

The dropout-rate ablation further supported this conclusion. Increasing dropout probability increased predictive dispersion but did not yield monotonic improvements in calibration. Instead, calibration behavior depended on both the backbone architecture and the strength of injected stochasticity. These findings reinforce the view that uncertainty-aware inference mechanisms should be evaluated behaviorally rather than interpreted mechanistically.

5.3 Operational scope of uncertainty as a conditional decision signal

This work evaluated uncertainty estimates strictly in terms of their empirical association with prediction errors, rather than treating uncertainty as an inherently reliable or explanatory signal. Global uncertainty-error alignment, as measured by error-detection AUROC over the full test distribution, was heterogeneous across backbones and uncertainty measures, and does not by itself support the use of uncertainty as a universally reliable trust indicator.

At the same time, uncertainty estimates exhibited structured behavior under specific evaluation regimes. In particular, uncertainty-based ranking concentrated misclassification events under selective rejection and, when conditioned on high predicted confidence, stratified residual risk among predictions that would otherwise be treated as equally reliable. These effects were modest in magnitude and emerged only after explicit conditioning or coverage reduction, but they were directionally consistent across architectures and uncertainty partitioning schemes, and comparable to risk reductions reported for selective rejection policies operating in high base-accuracy regimes.

These results therefore reframe uncertainty as an operational, decision-level signal rather than an interpretive explanation of model behavior. The utility of this reframing lies not in global monotonic alignment with correctness, but in its

capacity to support selective abstention, triage, or human-in-the-loop review under clearly specified operating conditions. Accordingly, uncertainty estimates should be evaluated empirically and deployed conditionally, rather than assumed to provide a general-purpose measure of trust.

The confidence-band and variance-partition sweep provided a structured basis for evaluating whether uncertainty conveyed information beyond predicted confidence across operating regimes and partition choices. Across the lower and mid-confidence bands (centers 0.60, 0.70, and most configurations at 0.80), uncertainty-error separation was weak and statistically indistinguishable from zero, with bootstrap confidence intervals consistently overlapping zero across variance-partition schemes. In the lower and mid-confidence regimes, uncertainty rankings did not consistently stratify error beyond what was already captured by predicted probability. In these regions, uncertainty behaved largely as a reparameterization of confidence and did not provide additional discriminative signal.

A different pattern emerged in the upper-confidence regime. When predictions were already highly confident, higher predictive variance was associated with elevated empirical error across both backbones. This association was directionally consistent across variance partitions and became detectable only under explicit conditioning. The effect was modest in magnitude and did not appear uniformly across all partition choices, indicating sensitivity to thresholding and finite Monte Carlo estimation. Instead, uncertainty's utility was conditional: uncertainty stratified residual risk only among predictions that would otherwise be treated as highly reliable based on confidence alone. This explains why global uncertainty-error metrics remained weak while localized effects appeared under restricted operating conditions.

From a decision-making perspective, this resolves the apparent trade-off between a slightly better but higher-uncertainty predictor and a slightly worse but lower-uncertainty one. At the model level, no inference procedure was uniformly preferable across accuracy, calibration, and uncertainty alignment. However, at a fixed confidence level, uncertainty can refine decisions by distinguishing lower- and higher-risk predictions within the same confidence stratum.

Under this interpretation, uncertainty does not replace confidence as a primary decision criterion. Rather, it functions as a secondary, conditional signal that supports selective abstention, triage, or human-in-the-loop review in high-confidence regimes. The confidence-band sweep, therefore, delineated where uncertainty contributed actionable information and where it did not, transforming an isolated conditional effect into a map of its operational scope.

Taken together, the confidence-band sweep replaced a single conditional demonstration with an explicit characterization of the operating region in which uncertainty provides additional risk stratification, and equally importantly, of the regions in which it does not.

5.4 Accuracy-calibration trade-offs

The bootstrap analysis demonstrated that improvements in calibration and uncertainty behavior were not accompanied by statistically meaningful losses in accuracy within the evaluated setting. While some inference procedures yielded small changes in accuracy, these differences were within resampling variability and should not be overinterpreted. This finding highlights the importance of reporting uncertainty on metrics themselves, particularly in high-stakes applications where small changes in accuracy may still be operationally relevant. This stability held under both in-distribution and generator-disjoint evaluation, indicating that calibration improvements were not offset by accuracy degradation under the tested distribution shifts.

5.5 Perspectives

5.5.1 *Limitations*

Several limitations warrant consideration. First, the evaluation was restricted to in-distribution data and did not address robustness under distributional shift. Second, uncertainty estimates were evaluated at the image level and did not capture temporal or contextual dependencies present in real-world political media. Third, while MC dropout and ensemble surrogates were compared, more expressive Bayesian neural network formulations remained unexplored. Addressing these limitations will be essential for translating uncertainty-aware deepfake detection into operational settings.

5.5.2 *Scope*

The dataset construction procedure was designed to support controlled evaluation of discriminative performance, calibration, and uncertainty-error relationships under matched train-test conditions. While the filtering pipeline yielded a politically salient subset, it did not guarantee uniform coverage across all politicians, geopolitical regions, event types, imaging conditions, or generative methods. There is a risk that political relevance is entangled with generation artifacts rather than semantic political content in the image. Accordingly, conclusions drawn from this dataset are restricted to the observed distribution and do not claim robustness to unseen identities, novel generators, or out-of-distribution political imagery.

5.5.3 *Generalization*

Most analyses in this work were reported under controlled in-distribution evaluation; additionally, a generator-disjoint OOD evaluation was included to quantify changes in discrimination and reliability when synthetic images were produced by unseen generator families. Because the OOD split enforces disjointness only with respect to the generator family, the reported robustness conclusions are limited to generator shift. Identity-level generalization was not explicitly controlled (identities were pooled across splits), and other real-world shifts, such as platform-specific post-processing pipelines, acquisition-device changes, or adversarial perturbations, remained outside the scope of the present evaluation. In addition, the analysis did not statistically adjust for correlated identities, prompts, or content sources across splits; instead, these factors were treated as part of the observed in-distribution variability. Furthermore, given the near-saturated discrimination performance (AUC ≈

0.99), it remains unclear whether the localization of uncertainty–error stratification to the high-confidence regime is intrinsic or an artifact of benchmark ease. Consequently, the operational scope of uncertainty-aware inference demonstrated here should be interpreted as dataset- and regime-specific rather than as a general property of deepfake detectors.

5.5.4 *Implications for deployment*

The findings suggest that uncertainty-aware inference can improve the reliability of political deepfake detection systems without degrading discriminative performance, provided that uncertainty estimates are evaluated empirically rather than assumed to reflect epistemic uncertainty. In deployment settings, uncertainty estimates should be treated as decision-support signals, enabling confidence-aware policies rather than serving as explanations of model reasoning.

6. Conclusion

This work examined uncertainty-aware inference for political deepfake detection under a strictly empirical and operational reliability framework. Motivated by the limitations of point-prediction detectors in high-stakes political misinformation settings, the study shifted emphasis from discriminative performance alone to the behavior and utility of probabilistic outputs and uncertainty estimates under explicitly defined operating conditions. Rather than assuming a particular probabilistic interpretation of uncertainty, all conclusions were grounded in observable properties: calibration quality, proper scoring rules, and the relationship between uncertainty signals and classification errors.

Using a politically focused real-synthetic image dataset and two pretrained convolutional backbones fully fine-tuned end-to-end, the analysis showed that uncertainty-aware inference procedures modify reliability characteristics without materially affecting ranking-based discrimination metrics such as ROC-AUC. Calibration-sensitive measures, including expected calibration error, Brier score, and negative log-likelihood, varied across deterministic inference, post-hoc calibration, single-pass stochastic inference, and Monte Carlo dropout. These results reinforce the distinction between discrimination and reliability and demonstrate that high ROC-AUC alone does not guarantee decision-safe probabilistic outputs.

A key finding was that reliability effects commonly attributed to Bayesian-style marginalization did not require multi-sample Monte Carlo averaging. Single-pass stochastic inference ($T = 1$), which introduces test-time stochasticity without Bayesian averaging, yielded calibration behavior comparable to or stronger than MC mean predictions in several configurations. Dropout-rate and MC-sample-size ablations further indicated that predictive dispersion and calibration behavior depended jointly on architectural capacity and the degree of injected stochasticity. Accordingly, predictive variance under MC dropout should be interpreted cautiously as an empirical uncertainty signal rather than assumed to represent epistemic uncertainty.

Crucially, the systematic confidence-band and variance-partition sweep resolved the ambiguity surrounding the operational value of uncertainty. Uncertainty-error separation is not global: across low- and mid-confidence regimes, uncertainty provided little additional error stratification beyond confidence, regardless of partition choice. In contrast, within the upper-confidence regime, uncertainty consistently stratified residual risk among predictions that would otherwise be considered highly reliable. This effect was directionally consistent but limited in scope, becoming statistically detectable only under high-confidence conditioning and more extreme uncertainty partitions.

These findings reframe uncertainty not as a universally dominant trust signal, but as a conditional, decision-level signal whose utility depends on the operating regime. Uncertainty is most informative when used to refine trust among already confident predictions or to support selective abstention and triage policies that trade coverage for risk reduction. When evaluated at full coverage or without conditioning, uncertainty does not provide a basis for preferring one predictor globally.

By explicitly mapping where uncertainty contributes operationally meaningful information, and where it probably does not, this work upgrades an otherwise ambiguous conditional effect into a defensible, decision-relevant characterization. The contribution is therefore not the claim that uncertainty universally improves deepfake detection, but a principled delineation of the regimes in which uncertainty-aware inference can and cannot support reliable decision-making in political deepfake detection systems.

Code and data availability

All scripts required to reproduce preprocessing, training, inference, and evaluation, including random seed configuration and figure generation are available at https://github.com/xyz-rafael-xyz/political-deepfake-uncertainty.git . Dataset construction relies exclusively on publicly available sources and deterministic metadata-based filtering procedures as defined in the Methods section.

7. References

1. Wang T, Liao X, Chow KP, Lin X, Wang Y. Deepfake detection: a comprehensive survey from the reliability perspective. ACM Comput Surv. 2024;57(3):58:1-58:35. https://doi.org/10.1145/3699710

2. Kumar A, Singh D, Jain R, Jain DK, Gan C, Zhao X. Advances in deepfake detection algorithms: exploring fusion techniques in single and multi-modal approach. Inf Fus. 2025;102993. https://doi.org/10.1016/j.inffus.2025.102993

3. Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, Shahzad M, Yang W, Bamler R, Zhu XX. A survey of uncertainty in deep neural networks. Artif Intell Rev. 2023;56:1513-1589. https://doi.org/10.1007/s10462-023-10562-9

4. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B, Snoek J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems 32 (NeurIPS); 2019. https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

5. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML); 2016 Jun 20-22; New York, NY, USA. Proc Mach Learn Res. 2016;48:1050-1059. https://proceedings.mlr.press/v48/gal16.html

6. Yumlembam R, Issac B, Aslam N, Babu EK, Collyer J, Kennedy F. Detection of AI-generated images using combined uncertainty measures and particle swarm optimised rejection mechanism. arXiv [cs.CV]. 2025 Dec 20. arXiv:2512.18527. https://arxiv.org/abs/2512.18527

7. Zhu M, Long J. Detecting anti-forensic deepfakes with identity-aware multi-branch networks. Front Big Data. 2025 Dec 10;8:1720525. https://doi.org/10.3389/fdata.2025.1720525

8. Lu Y, Ebrahimi T. Impact of video processing operations in deepfake detection. In: Proceedings of the IEEE International Conference on Digital Signal Processing (DSP); 2023. https://2025.ic-dsp.org/wp-content/uploads/2023/05/DSP2023-50.pdf

9. Jin X, Guan W, Wang W, Dong J. Towards reliable deepfake detection from an uncertainty calibration perspective. Vis Intell. 2025;3:28. https://doi.org/10.1007/s44267-025-00100-2

10. Kose N, Rhodes A, Ciftci UA, Demir I. Is it certainly a deepfake? Reliability analysis in detection & generation ecosystem. arXiv [cs.AI]. 2025 Sep 22 [revised 2025 Oct 28]. arXiv:2509.17550. https://arxiv.org/abs/2509.17550

11. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017. Proc Mach Learn Res. 2017;70:1321-1330. http://proceedings.mlr.press/v70/guo17a.html

12. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning (ICML); 2005 Aug 7; Bonn, Germany. p. 625-632. https://doi.org/10.1145/1102351.1102430

13. Kuleshov V, Fenner N, Ermon S. Accurate uncertainties for deep learning using calibrated regression. In: Proceedings of the 35th International Conference on Machine Learning (ICML); 2018. Proc Mach Learn Res. 2018;80:2796-2804. http://proceedings.mlr.press/v80/kuleshov18a.html

14. He W, Jiang Z, Xiao T, Xu Z, Li Y. A survey on uncertainty quantification methods for deep learning. arXiv [cs.LG]. 2023 Feb 26 [revised 2025 Dec 13]. arXiv:2302.13425. https://arxiv.org/abs/2302.13425

15. Thulasidasan S, Chennupati G, Bilmes JA, Bhattacharya T, Michalak S. On mixup training: improved calibration and predictive uncertainty for deep neural networks. In: Advances in Neural Information Processing Systems 32 (NeurIPS); 2019. https://proceedings.neurips.cc/paper_files/paper/2019/hash/36ad8b5f42db492827016448975cc22d-Abstract.html

16. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, Makarenkov V, Nahavandi S. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fus. 2021;76:243-297. https://doi.org/10.1016/j.inffus.2021.10.002

17. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems 30 (NeurIPS); 2017. https://papers.nips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

18. Djupskås A, Riemer-Sørensen S, Stasik AJ. Unreliable Monte Carlo dropout uncertainty estimation [poster]. NLDL 2026; 2025 Nov 5 [modified 2025 Dec 8]. Available from: https://openreview.net/forum?id=zfd7OEUG0o#discussion

19. Wang T, Wang Y, Zhou J, Peng B, Song X, Zhang C, Sun X, Niu Q, Liu J, Chen S, Chen K, Li M, Feng P, Bi Z, Liu M, Zhang Y, Fei C, Yin CH, Yan LKQ. From aleatoric to epistemic: exploring uncertainty quantification techniques in artificial intelligence. arXiv [cs.AI]. 2025 Jan 5. arXiv:2501.03282. https://arxiv.org/abs/2501.03282

20. Huang YC, Padarian J, Minasny B, McBratney AB. Using Monte Carlo conformal prediction to evaluate the uncertainty of deep-learning soil spectral models. SOIL. 2025;11(2):553-563. https://doi.org/10.5194/soil-11-553-2025

21. Buddenkotte T, Escudero Sanchez L, Crispin-Ortuzar M, Woitek R, McCague C, Brenton JD, Öktem O, Sala E, Rundo L. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. Comput Biol Med. 2023;163:107096. https://doi.org/10.1016/j.compbiomed.2023.107096

22. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv [cs.LG]. 2016 Oct 7 [revised 2018 Oct 3]. arXiv:1610.02136. https://arxiv.org/abs/1610.02136

23. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv [stat.ML]. 2021 Jul 15 [revised 2022 Dec 7]. arXiv:2107.07511. https://arxiv.org/abs/2107.07511

24. Livernoche V, Arodi A, Musulan A, Yang Z, Salvail A, Marceau Caron G, Godbout J-F, Rabbany R. OpenFake: an open dataset and platform toward large-scale deepfake detection. arXiv [cs.CV]. 2025 Sep 11. arXiv:2509.09495v1. https://arxiv.org/html/2509.09495v1

25. Hendrycks D, Lee K, Mazeika M. Using pre-training can improve model robustness and uncertainty. arXiv [cs.LG]. 2019 Jan 28 [revised 2019 Oct 20]. arXiv:1901.09960. https://arxiv.org/abs/1901.09960

26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770-778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

27. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML); 2019. Proc Mach Learn Res. 2019;97:6105-6114. http://proceedings.mlr.press/v97/tan19a.html

28. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: transformers for image recognition at scale. arXiv [cs.CV]. 2021 Oct 21. arXiv:2010.11929. https://arxiv.org/pdf/2010.11929

29. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30 (NeurIPS); 2017. https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845. https://doi.org/10.2307/2531595