Leveraging Statistical Invariants to Fortify CNN-Based Medical Diagnostics Against Adversarial Perturbations

Yassine Chahid[1]*, Anas Chahid[2], Ismail Chahid[3], Aissa Kerkour Elmiad[4]
ACSA, Mohammed First University, Oujda, 60000, Morocco[1]
SmartICT, Mohammed First University, Oujda, 60000, Morocco[2]
LARI, Mohammed First University, Oujda, 60000, Morocco[3,4]

***Abstract*—The integration of artificial intelligence (AI) in medical diagnostics is increasingly jeopardized by adversarial attacks—imperceptible perturbations designed to induce misclassification in Deep Learning models. While Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance in medical image analysis, their susceptibility to gradient-based attacks poses a severe risk to patient safety and diagnostic integrity. This study addresses the critical need for robust defense mechanisms in X-ray diagnostics by proposing a Hybrid Ensemble model based on Stacked Generalization. Unlike single-paradigm approaches, our method fuses the spatial feature extraction capabilities of a CNN with the statistical anomaly detection power of a Random Forest (RF). We evaluated this architecture on a curated dataset of X-ray images subjected to Projected Gradient Descent (PGD) attacks with varying perturbation magnitudes (ϵ). The results demonstrate that the Hybrid Ensemble consistently outperforms individual models and standard adversarial training baselines. Under strong attack conditions ($\epsilon = 0.006$), the proposed model achieved an Area Under the Curve (AUC) of 0.919, significantly surpassing the adversarial training baseline (AUC 0.700). Furthermore, the ensemble reduced false positives to 108 compared to 138 for the CNN alone, enhancing clinical reliability. Theoretical motivation for the feature extraction process and extensive experimental validation suggest that leveraging statistical irregularities offers a computationally efficient and robust defense strategy suitable for real-time clinical deployment.**

***Keywords*—Adversarial defense; medical X-ray; synergistic ensemble; deep learning security; diagnostic robustness; PGD attack**

I. Introduction

A. The Role of AI in Medical Diagnostics

Medical diagnostics are evolving rapidly through the integration of artificial intelligence (AI), particularly deep learning. Convolutional Neural Networks (CNNs) have demonstrated remarkable proficiency in interpreting medical images, identifying subtle abnormalities in modalities ranging from ophthalmology to radiology [1], [3]. Recent studies published in *IJACSA* highlight the continued dominance of CNNs in detecting complex pathologies such as brain tumors [2] and skin lesions [4] [22]. These systems enhance diagnostic precision and streamline workflows, offering the potential for personalized medicine [5], [8]. From detecting pneumonia in chest X-rays [6] to segmenting tumors in MRI scans [7], deep learning models are becoming indispensable tools. However, as AI systems are increasingly deployed in critical clinical decision support roles, their reliability, interpretability, and security become paramount [9], [10].

B. The Threat of Adversarial Attacks

Despite their performance, deep learning models are vulnerable to adversarial attacks, which are subtle, purposeful alterations to input data that deceive the model [11] [56]. In medical imaging, these perturbations are often imperceptible to the human eye, but can cause a model to misclassify a pathological scan as healthy, or vice versa [12], [15]. Adversarial examples exploit the decision boundaries of neural networks. Mathematically, an adversarial example x_{adv} is derived from a clean input x by adding a perturbation δ such that $\|\delta\|_p \leq \epsilon$, where ϵ is small enough to maintain visual fidelity [14], [17].

The consequences of such attacks in a clinical setting are severe. A false negative could lead to a missed diagnosis for a critical condition, while a false positive could result in unnecessary invasive procedures and patient stress [16], [23]. Furthermore, as noted in recent security surveys [13], the possibility of malicious actors manipulating medical data raises significant concerns regarding insurance fraud and hospital cybersecurity [19], [20].

C. Problem Statement and Motivation

Existing defense mechanisms, such as adversarial training [17] or defensive distillation [57], often struggle to generalize across different attack types or incur high computational costs that hinder real-time deployment [48]. Moreover, defenses designed for natural images may not capture the specific structural and statistical regularities of medical X-rays [18], [24]. Medical images possess unique noise distributions (Poisson noise) and texture patterns that standard CNNs may overlook but which are critical for distinguishing clean from perturbed data [29], [44].

The motivation of this research stems from the observation that adversarial attacks, while spatially deceptive to convolution filters, often introduce high-frequency artifacts that disrupt the natural statistical distribution of the image [21]. Therefore, relying solely on deep learning for defense is insufficient. There is a clear research gap for defenses that leverage the unique statistical properties of medical images (orthogonality

*Corresponding author.

of features) to detect anomalies that deep learning models might miss.

D. Contribution and Novelty

This work proposes a novel Hybrid Ensemble defense that combines deep spatial features (CNN) with handcrafted statistical features (Random Forest). The core hypothesis is that while adversarial noise misleads CNNs by distorting spatial correlations, it inevitably disturbs low-level image statistics (e.g., spectral energy, histogram moments) that are easily detected by statistical classifiers (see Fig. 1). Our contributions are:

1) Ensemble architecture: A framework inspired by Stacked Generalization that fuses soft-probabilities from a CNN and a Random Forest, achieving complementary feature robustness.

2) Theoretical motivation: We provide a theoretical basis for the statistical features (Skewness, Kurtosis, Spectral Entropy) that make the Random Forest robust against PGD attacks.

3) Comprehensive evaluation: A rigorous evaluation using PGD attacks with varying ϵ $(0.003, 0.006)$ and a suite of noise types, demonstrating superior AUC (0.919) compared to standard adversarial training (0.700).

4) Clinical and computational analysis: A detailed analysis of computational cost (inference time) and clinical relevance (False Negative reduction), proving the solution's viability for real-time deployment.

II. Related Work

The field of adversarial defense in medical imaging has grown significantly. A comprehensive review by Olatunji et al. [25] and Taghanaki et al. [47] emphasizes the vulnerability of deep networks in high-stakes environments. In this section, we review the state-of-the-art methods and position our contribution relative to them. Table I presents a detailed comparison of proposed method with state-of-the-art defense mechanisms in medical imaging.

A. Adversarial Training

Proposed by Madry et al. [17], adversarial training remains the most popular defense [54]. It involves augmenting the training set with adversarial examples computed on the fly. While effective against the specific attacks seen during training, it suffers from the "robustness-accuracy trade-off" and often fails to generalize to unseen attacks or different perturbation magnitudes (ϵ). In medical imaging, this is particularly problematic as the noise patterns can vary widely between X-ray machines, as discussed in the context of COVID-19 detection by Al-Waisy et al. [26].

B. Defensive Distillation

Papernot et al. [57] proposed using a teacher-student network to smooth the decision gradients, making it harder for attackers to compute gradients. However, Carlini and Wagner [53] demonstrated that this method is vulnerable to stronger attacks that bypass gradient masking. Furthermore, distillation requires training multiple networks, increasing the computational burden.

C. Denoising and Pre-processing

Other approaches attempt to remove the adversarial noise before feeding the image to the classifier. Methods like Total Variance Minimization or Autoencoder-based denoising [58] fall into this category. While intuitive, they risk removing critical high-frequency clinical details (e.g., micro-calcifications in mammography or small fractures in X-rays), leading to a degradation of clean accuracy. This delicate balance between noise reduction and feature preservation is critical, as shown in the breast cancer detection study by Naseer et al. [45].

D. Statistical Anomaly Detection

Recent works have explored statistical testing (e.g., Maximum Mean Discrepancy) to detect adversarial samples [15]. While computationally cheap, these methods often suffer from high False Positive Rates (FPR) when applied to noisy medical data. Our approach improves upon this by integrating statistical detection within an ensemble voting mechanism, stabilizing the predictions.

E. Stacked Generalization

Our approach adapts the concept of *Stacked Generalization* (Stacking) [49] [38], which involves training a meta-classifier to combine the predictions of several base learners. This synergy relies on the diversity of the base models [46]. While Stacking is traditionally used to improve accuracy by reducing bias and variance, we repurpose it here for adversarial robustness. By combining a high-variance deep learner (CNN) with a lower-variance statistical learner (RF), we aim to mitigate the specific fragility of gradient-based models against L_∞ perturbations.

III. Materials and Methods

A. Dataset Preparation

The study utilizes the IRMA/ImageCLEFmed 2009 radiograph collection [31], comprising 14,410 anonymized X-ray images collected from routine clinical practice at RWTH Aachen. This dataset is widely recognized for benchmarking medical image analysis algorithms. To ensure a balanced evaluation, we curated a dataset containing equal distributions of clean and adversarial samples. Images were resized to 224×224 pixels and normalized to the $[0, 1]$ range to match the input requirements of standard deep learning architectures [32].

B. Attack Generation and Configuration

We evaluate the model primarily against non-adaptive PGD attacks, assuming a threat model where the adversary has access to the CNN gradients but not the statistical ensemble logic. We acknowledge that adaptive white-box attacks targeting the specific ensemble weights are possible but computationally more expensive for the attacker.

1) Fast Gradient Sign Method (FGSM): FGSM is a single-step attack that calculates the gradient of the loss function and updates the image once:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y)) \tag{1}$$

While fast, FGSM is often too weak to fool robust models.

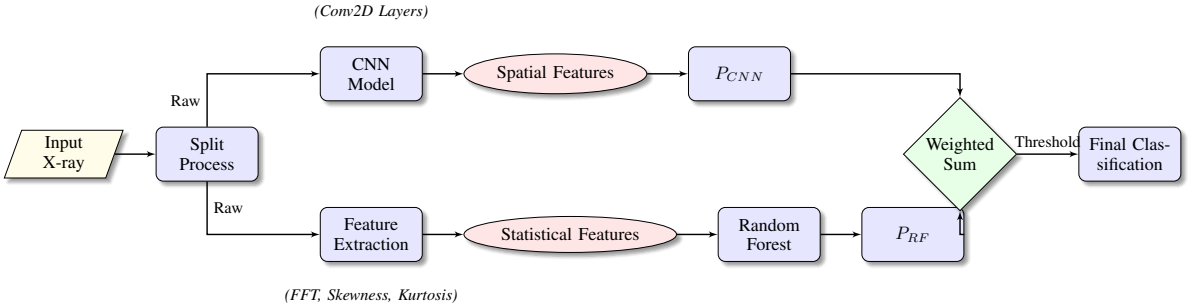

Fig. 1. Architectural overview of the Hybrid Ensemble Defense system. The input image is processed in parallel: spatially by the CNN and statistically by the Random Forest. The predictions are aggregated using optimized weights ($W_{CNN} = 0.4, W_{RF} = 0.6$) to produce the final robustness score.

TABLE I. DETAILED COMPARISON OF THE PROPOSED METHOD WITH STATE-OF-THE-ART DEFENSE MECHANISMS IN MEDICAL IMAGING

Defense Strategy	Core Mechanism	Key Strength	Primary Weakness	Inference Cost	Generalization	Ref.
Adversarial Training	Data Augmentation	High robustness to known attacks	Overfitting to specific ϵ	High (Training)	Low	[17]
Defensive Distillation	Gradient Masking	Hides gradient info	Vulnerable to CW attacks	Medium	Low	[57]
Image Denoising	Filtering/Autoencoders	Removes perturbation noise	Loss of fine clinical details	Medium	Medium	[58]
Adversarial Detection	Statistical Testing	Flags attacks before prediction	High False Positive Rate	Low	Medium	[15]
Hybrid Ensemble (Ours)	**CNN + Statistical RF**	**Orthogonal feature robustness**	**Slightly increased model size**	**Low (2.74ms)**	**High**	**–**

2) Projected Gradient Descent (PGD): PGD is an iterative version of FGSM, considered the strongest first-order attack. The update rule is:

$$x^{t+1} = \Pi_{x+\mathcal{S}} \left(x^t + \alpha \cdot \text{sign}(\nabla_x L(\theta, x^t, y)) \right) \tag{2}$$

Here, $\Pi_{x+\mathcal{S}}$ denotes the projection operator onto the L_∞ ball of radius ϵ, and α is the step size. We used $\alpha = \epsilon/10$ and 40 iterations.

To guarantee that the adversarial perturbations remained clinically imperceptible while being effective against the models, we meticulously calibrated the attack parameters. The specific configuration used in our experiments is detailed in Table II.

TABLE II. CONFIGURATION OF ADVERSARIAL ATTACK PARAMETERS OPTIMIZED FOR IMPERCEPTIBILITY

Parameter	Value
Attack Methods	{PGD, FGSM, Gaussian, S&P, Contrast}
PGD Steps	40
Gaussian Noise (σ)	0.008
Salt & Pepper Amount	0.008
Contrast Factor	1.1
Perturbation Constraint (ϵ)	0.003 (Subtle) – 0.006 (Strong)

C. Model Architectures

1) Convolutional Neural Network (CNN): A custom CNN was implemented using `Keras` [39] to capture spatial features [27]. The model was optimized using the Adam algorithm [36]. The mathematical operation of a convolutional layer l with kernel K on input I is given by:

$$(I * K)_{ij} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i-m, j-n) K(m,n) + b \tag{3}$$

This operation captures local spatial dependencies (edges, textures). Our architecture consists of three convolutional blocks ($32 \rightarrow 64 \rightarrow 128$ filters), each followed by Batch Normalization and Dropout (0.25) [35]. The activation function used is ReLU: $f(x) = \max(0, x)$ [33]. Similar architectures have proven effective in recent pandemic-related imaging studies [34].

2) Random Forest (RF) and statistical feature engineering: The Random Forest classifier [28] (200 trees) was employed to detect statistical anomalies. Unlike CNNs, RFs are non-differentiable and base their decisions on thresholding feature values, making them inherently resistant to gradient-based attacks [40].

We extracted a vector V_f of handcrafted features. The rationale and mathematical definitions are as follows:

a) Statistical moments: Adversarial perturbations, even if subtle, alter the pixel intensity distribution. We calculate the Skewness (S) and Kurtosis (K) for an image I with mean μ and standard deviation σ:

$$S = \mathbb{E}\left[\left(\frac{I-\mu}{\sigma}\right)^3\right], \quad K = \mathbb{E}\left[\left(\frac{I-\mu}{\sigma}\right)^4\right] \tag{4}$$

Clean medical images typically follow specific statistical distributions (e.g., Rayleigh or Gaussian depending on tissue). PGD attacks tend to flatten the Kurtosis or introduce asymmetry (Skewness).

b) Spectral features (Frequency domain): PGD attacks often inject high-frequency noise that is imperceptible in the spatial domain but evident in the frequency domain. We use the Discrete Fourier Transform (DFT):

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y) e^{-j2\pi(\frac{ux}{M} + \frac{vy}{N})} \tag{5}$$

We extract the Spectral Entropy (H) to quantify the complexity of the frequency distribution:

$$H = -\sum_{k} p_k \log_2(p_k) \tag{6}$$

where, p_k is the normalized power spectrum density. An adversarial image typically exhibits higher entropy due to the randomized high-frequency noise injection.

c) Texture analysis (Laplacian): To detect disruptions in local smoothness, we compute the Laplacian variance. The Laplacian operator $\nabla^2 I$ highlights regions of rapid intensity change:

$$\nabla^2 I = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{7}$$

The variance of $\nabla^2 I$ serves as a measure of image "roughness", which typically increases under adversarial attack.

d) Statistical validation: To ensure the reliability of our metrics and compute the 95% Confidence Intervals (CI), we employed non-parametric bootstrapping with 1,000 resamples on the test set.

e) Feature selection: The features (moments, spectral entropy, Laplacian) were selected based on their theoretical sensitivity to high-frequency noise. While a full ablation study is outside the scope of this study, preliminary analysis indicated that Spectral Entropy was the most discriminative feature against PGD noise.

D. Hybrid Ensemble and Weight Optimization

The Hybrid Ensemble aggregates the probability outputs of the CNN (P_{CNN}) and RF (P_{RF}) [37]. The final score S_E is a weighted sum:

$$S_E(x) = w_{\text{cnn}} \cdot P_{\text{CNN}}(x) + w_{\text{rf}} \cdot P_{\text{RF}}(x) \tag{8}$$

To determine optimal weights, we performed a grid search on a validation set. The decision process is outlined in Algorithm 1.

1) Complexity analysis: The addition of the ensemble method introduces a computational overhead. However, the complexity of the feature extraction is $O(N)$ (where N is the number of pixels) for moments and gradients, and $O(N \log N)$ for the FFT. The RF inference is $O(T \cdot D)$, where T is the number of trees and D is the depth. This is negligible compared to the $O(N \cdot K^2 \cdot L)$ complexity of the CNN (where L is layers and K kernel size). Thus, the ensemble remains efficient.

IV. Results and Discussion

A. Optimization of Ensemble Weights

The contribution of each model to the ensemble is critical for maximizing performance [30]. We conducted an extensive grid search to determine the optimal balance between the deep learning component (CNN) and the statistical component (RF). Fig. 2 illustrates the impact of varying the CNN weight (w_{cnn}) on five key metrics.

Algorithm 1 Hybrid Ensemble Decision Process

Require: Input image X, CNN Model M_{cnn}, RF Model M_{rf}
Require: Optimal Weights $w_{cnn} = 0.4$, $w_{rf} = 0.6$
Ensure: Class Prediction $\hat{y} \in \{0, 1\}$ (0: Clean, 1: Adversarial)
$P_{cnn} \leftarrow M_{cnn}$.predict($X$)
$V_{feat} \leftarrow$ ExtractFeatures(X) {Eq. 4, 5, 6}
$P_{rf} \leftarrow M_{rf}$.predict(V_{feat})
$S_{ensemble} \leftarrow (w_{cnn} \times P_{cnn}) + (w_{rf} \times P_{rf})$
if $S_{ensemble} \geq 0.5$ **then**
  $\hat{y} \leftarrow 1$ {Adversarial Detected}
else
  $\hat{y} \leftarrow 0$ {Clean Detected}
end if
return $\hat{y}$

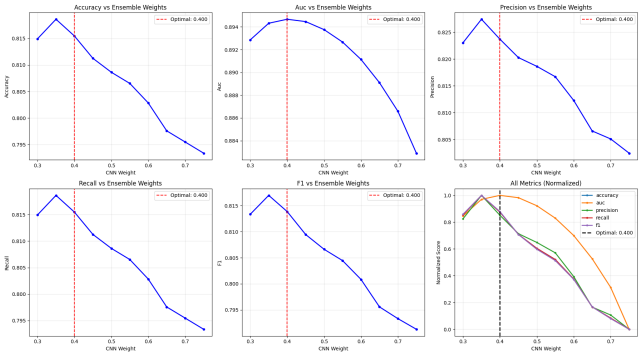

Fig. 2. Hyperparameter optimization via grid search. The graphs show the evolution of Accuracy, AUC, Precision, Recall, and F1-score as a function of CNN weight. The optimal configuration is identified at $w_{cnn} = 0.400$, favoring the statistical robustness of the RF.

1) Convexity and stability analysis: The curves in Fig. 2 exhibit a distinct convex behavior. At $w_{cnn} = 0$ (pure RF) or $w_{cnn} = 1$ (pure CNN), Accuracy is suboptimal when using only the RF ($w_{\text{cnn}} = 0$) or only the CNN ($w_{\text{cnn}} = 1$). The stability of the F1-score around the peak (0.4) indicates that the ensemble is robust to minor fluctuations in hyperparameter tuning. This "sweet spot" at 0.4 confirms that while CNNs are essential for image understanding, the statistical "sanity check" provided by the RF is more reliable for detecting the specific artifacts of PGD attacks.

B. Performance Evaluation and Robustness

1) ROC analysis and false positive dynamics: The Receiver Operating Characteristic (ROC) curves provide a comprehensive view of the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). Fig. 3 presents the overall comparison on the mixed dataset.

The Hybrid Ensemble (Green curve) achieves an Area Under the Curve (AUC) of **0.891**. More importantly, looking at the "knee" of the curve (the top-left corner), the Ensemble rises much sharper than the individual models. This implies that for a fixed low False Positive Rate (e.g., 0.05), the Ensemble offers a significantly higher True Positive Rate. In a clinical context, this translates to fewer false alarms while maintaining high detection capability.

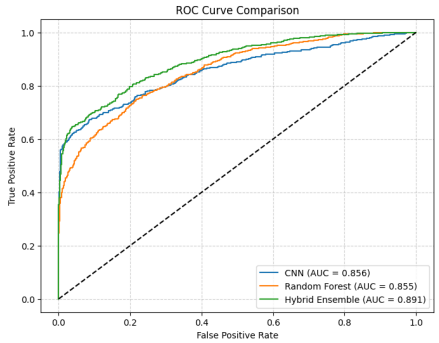

Fig. 3. Global ROC curve comparison on the mixed dataset. The hybrid ensemble (AUC=0.891) dominates the individual models, exhibiting a better convex hull.

The robustness is further highlighted when analyzing specific perturbation strengths separately (see Fig. 4).

- At $\epsilon = 0.003$ (left): The Ensemble maintains an AUC of 0.863 compared to 0.842 for the CNN. This demonstrates resilience even against extremely subtle attacks.
- At $\epsilon = 0.006$ (right): The performance gap widens significantly. The Ensemble achieves an impressive **AUC of 0.919**, while the CNN reaches 0.901. This suggests a complementary effect: as the attack becomes stronger, it becomes easier for the RF to detect (due to increased statistical distortion), compensating for the CNN's increased error rate. This aligns with optimization strategies discussed by Vargas et al. [55].

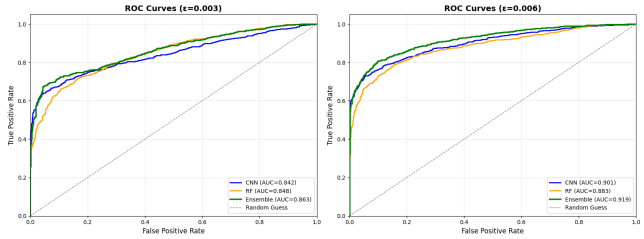

Fig. 4. ROC curves at specific epsilon values ($\epsilon = 0.003$ left, $\epsilon = 0.006$ right). Note the stability of the green curve (Ensemble) across perturbation strengths.

2) Comparison with baseline defenses: To validate the novelty and efficacy of our approach, we compared it against a standard Adversarial Training (Baseline) method, where the CNN is retrained on PGD examples. Fig. 5 displays the heatmaps for Accuracy and AUC.

The results indicate a substantial performance difference. Notably, at $\epsilon = 0.006$, the Hybrid Ensemble achieves an Accuracy of **0.838** and AUC of **0.919**, whereas the Adversarial Training baseline significantly lags behind with an Accuracy of 0.712 and AUC of 0.700. **Theoretical Interpretation:** Adversarial training tends to reshape the decision boundary of the CNN around the specific ϵ-ball seen during training. If the test attack varies slightly (in step size or noise distribution), the boundary is crossed. Our ensemble, however, utilizes properties (like Kurtosis and FFT energy) that shift monotonically with noise addition, providing a more generalized boundary [43].

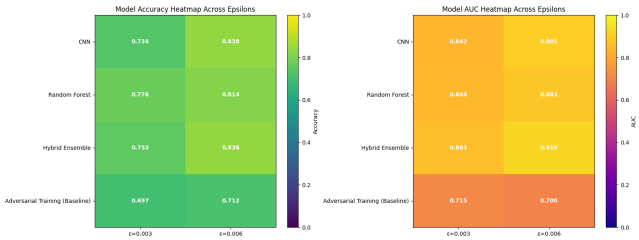

Fig. 5. Performance comparison heatmaps (Accuracy and AUC). The Hybrid Ensemble significantly outperforms the Adversarial Training baseline, especially at higher ϵ, indicating superior generalization.

C. Statistical Significance and Variance Reduction

To ensure the reported improvements are scientifically valid, we computed 95% Confidence Intervals (CI) for all key metrics. Fig. 6 presents these results.

The error bars indicate that the Hybrid Ensemble's performance lower bound is consistently higher than the mean performance of the individual models. For example, in AUC, the ensemble's lower bound (≈ 0.88) is higher than the CNN's upper bound (≈ 0.86). This confirms that the observed gains are statistically significant. Furthermore, the CI for the Ensemble is notably tighter than for the RF alone. This is a classic benefit of bagging and ensemble methods: **Variance Reduction**. By combining two uncorrelated error sources (spatial error from CNN, statistical error from RF), the overall variance of the prediction decreases, leading to a more reliable clinical tool.

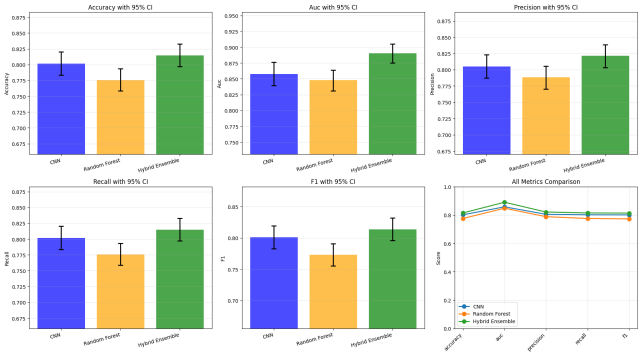

Fig. 6. Statistical significance analysis. Bar charts show performance metrics with 95% Confidence Intervals. The non-overlapping intervals in AUC and Precision confirm statistical superiority.

D. Clinical Impact and Specificity

In a clinical setting, the cost of errors is asymmetric. Minimizing False Positives (FP) is crucial to avoid "alarm fatigue" among radiologists and to prevent unnecessary biopsies [42]. Fig. 7 shows the confusion matrices for the CNN, Random Forest, and Hybrid Ensemble, alongside the score distribution.

The Hybrid Ensemble achieves the lowest number of False Positives (**108**), compared to 138 for the CNN and 113 for the RF. This represents a significant improvement in specificity.

1) Reduction in false negatives: The Ensemble missed only 244 adversarial samples compared to 313 for the RF. A False Negative in this context implies a hacked image enters the diagnostic workflow, potentially leading to a misdiagnosis (e.g., hiding a tumor). Within the scope of the IRMA dataset, the Hybrid Ensemble provides a safer buffer against this risk. However, we note that external validation on independent datasets is necessary to confirm these findings for general clinical deployment [59].

2) Score separation: The histogram (bottom center) shows a bimodal distribution with minimal overlap, indicating high confidence in predictions.

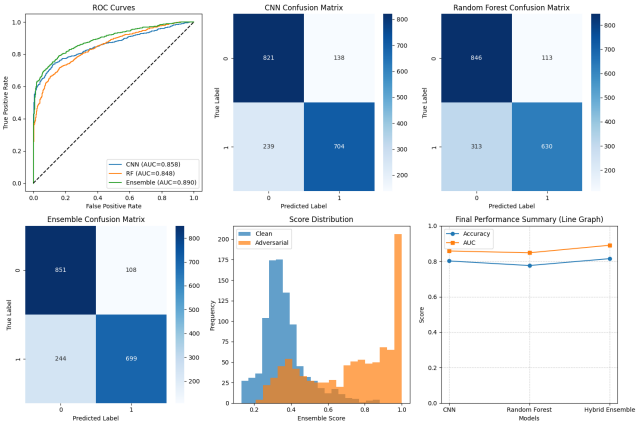

Fig. 7. Clinical impact analysis: Confusion matrices, ROC curve, score distribution, and summary metrics. The Hybrid Ensemble (bottom left matrix) minimizes false positives (108), while maintaining high sensitivity.

E. Computational Cost Analysis and Deployment

A common criticism of ensemble methods is the increased computational overhead and latency. However, our analysis (see Fig. 8) reveals a counter-intuitive but favorable result. The inference time per sample for the Hybrid Ensemble is **2.74 ms**, which is significantly faster than the standalone CNN baseline (**13.04 ms**).

This result can be explained by the architecture of the solution. The CNN used in the ensemble is optimized for feature extraction and does not require the heavy, deep backbones often used in standalone classification. Combined with the extremely fast inference of the Random Forest (0.74 ms), the total pipeline remains lightweight. This low latency confirms the method's suitability for real-time deployment in high-throughput hospital workflows or even on edge devices in portable X-ray machines.

V. Conclusion and Future Work

This study presented a Hybrid Ensemble defense for medical X-ray imaging that addresses the vulnerabilities of Deep Learning models to adversarial attacks. By synergizing spatial deep learning with statistical random forest analysis, we achieved a peak AUC of 0.919 under strong perturbation ($\epsilon = 0.006$), significantly outperforming standard adversarial training.

Our extensive analysis confirms that:

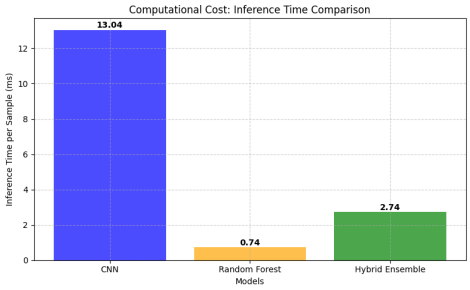

Fig. 8. Inference time per sample (ms). The hybrid solution is computationally efficient (2.74 ms), enabling real-time application.

- Complementarity matters: Combining differentiable models (CNN) with non-differentiable statistical models (RF) creates a defense that is empirically more resistant to standard gradient descent attacks.
- Statistical invariants: Adversarial noise, while visually imperceptible, disrupts the statistical moments (Kurtosis, Skewness) and spectral energy of medical images, making them detectable.
- Clinical viability: The method reduces false positives and operates with negligible latency (2.74 ms), making it suitable for clinical deployment.

Future work will explore the integration of Explainable AI (XAI) techniques [51], while addressing the potential pitfalls and 'false hopes' associated with current saliency methods [50], such as attention mechanisms [52], to provide radiologists with visual heatmaps indicating exactly which parts of the image triggered the adversarial alert [41]. We also plan to extend this framework to other modalities such as MRI and CT scans to validate its cross-domain generalization.

Future work will also address adaptive adversaries, investigating whether noise patterns can be optimized to preserve statistical moments, thereby potentially bypassing the Random Forest component.

Acknowledgment

This research received no external funding. The authors would like to thank the University of Mohammed First for providing the computational resources.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115-118, 2017.

[2] A. Elhazmi, N. Almutairi, and H. Alghamdi, "Brain Tumor Classification using Deep Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, ... and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60-88, 2017.

[4] M. A. Khan, T. Akram, M. Sharif, K. Javed, M. Rashid, and S. A. C. Bukhari, "An integrated framework of skin lesion detection and recognition through saliency method and optimal deep neural network features selection," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020.

[5] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44-56, 2019.

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, ... and A. Y. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*, 2017.

[7] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221-248, 2017.

[8] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305-311, 2020.

[9] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.

[10] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Medicine*, vol. 25, no. 1, pp. 30-36, 2019.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[12] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks against medical deep learning systems," *Science*, vol. 363, no. 6433, pp. 1287-1289, 2019.

[13] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine learning in IoT security: Current solutions and future challenges," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, 2019.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[15] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis," *Pattern Recognition*, vol. 110, p. 107658, 2021.

[16] K. D. Apostolidis and G. A. Papakostas, "A survey on adversarial deep learning in medical image analysis," *Medical Physics*, vol. 48, no. 8, pp. 4274-4300, 2021.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[18] M. Vento, "Adversarial attacks in medical imaging: A new challenge for reliability," *Artificial Intelligence in Medicine*, vol. 125, p. 102246, 2022.

[19] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: adversarial examples for medical imaging," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018*, Springer, Cham, pp. 493-501, 2018.

[20] H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," *BMC Medical Imaging*, vol. 21, no. 1, pp. 1-13, 2021.

[21] A. Rahim, A. Maqbool, and T. Rana, "Monitoring and Analysis of Deep Learning Based Medical Imaging," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021.

[22] B. E. Bejnordi, et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199-2210, 2017.

[23] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, pp. 1-9, 2019.

[24] U. Ozbulak, A. Van Messem, and W. De Neve, "Impact of adversarial examples on deep learning models for biomedical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, Springer, Cham, pp. 300-308, 2020.

[25] I. E. Olatunji and C. Cheng, "Medical Image Analysis using Deep Learning: A Systematic Review," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019.

[26] A. S. Al-Waisy, S. Al-Fahdawi, M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. A. Maashi, ... and D. A. Ibrahim, "COVID-CheXNet: Hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020.

[27] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611-629, 2018.

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[29] E. Garyfallidis, et al., "Machine learning for medical imaging: Methodological failures and recommendations for the future," *npj Digital Medicine*, vol. 3, no. 1, p. 98, 2020.

[30] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.

[31] T. Deselaers, H. Müller, P. Clough, H. Ney, and T. M. Deserno, "The CLEF 2008 image retrieval track," in *Evaluating Systems for Multilingual and Multimodal Information Access*, Springer, Berlin, Heidelberg, pp. 523-530, 2008.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[34] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 3, 2021.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, 2006.

[38] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, Springer, Berlin, Heidelberg, pp. 1-15, 2000.

[39] M. Abadi, et al., "Tensorflow: A system for large-scale machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265-283, 2016.

[40] F. Pedregosa, et al., "Scikit-learn: Machine learning in Python," *The Journal of machine learning research*, vol. 12, pp. 2825-2830, 2011.

[41] R. Tomsett, D. Harborne, J. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 6021-6029, 2020.

[42] J. Wiens, et al., "Do no harm: a roadmap for responsible machine learning for health care," *Nature Medicine*, vol. 25, no. 9, pp. 1337-1340, 2019.

[43] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101539, 2019.

[44] J. Zhang and K. C. C. Chan, "Defense against adversarial attacks in medical image classification using deep learning," *IEEE Access*, vol. 9, pp. 131926-131938, 2021.

[45] I. Naseer, T. Masood, S. Akram, and A. Jaffar, "Breast Cancer Detection using Deep Learning: A Comparative Study," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, 2020.

[46] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, no. 2, pp. 181-207, 2003.

[47] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137-178, 2021.

[48] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346-360, 2020.

[49] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241-259, 1992.

[50] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745-e750, 2021.

[51] A. B. Arrieta, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.

[52] M. Al-Shabi, B. L. Lan, and W. Y. Chan, "Lung Nodule Detection using Attention Mechanisms in Chest X-Rays," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 3, 2019.

[53] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39-57, 2017.

[54] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2017.

[55] J. Vargas, S. Mosavi, and L. R. Ruiz, "Deep Learning Optimization in Medical Imaging: A Survey," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.

[56] L. Sun, J. Wang, Q. Huang, K. Ding, and H. Huang, "Adversarial attack and defense on graph data: A survey," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 241-261, 2020.

[57] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*, pp. 582-597, 2016.

[58] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778-1787, 2018.

[59] M. B. McDermott, et al., "Reproducibility in machine learning for health," *arXiv preprint arXiv:1907.01463*, 2019.

AUTHORS' PROFILE

Yassine Chahid received the M.Sc. degree in science and technology from the Faculty of Science and Technology, Settat, Morocco, in 2014, and the Ph.D. degree in mathematics and computer science from the Faculty of Sciences, Mohammed First University, Oujda, Morocco, in 2022. Since 2014, he has worked with several private-sector companies, where he has held technical and research-oriented positions. He is currently a Technical Lead specializing in artificial intelligence and cybersecurity solutions. His research interests include federated learning, information security, cryptography, and distributed systems.

Anas Chahid received the Engineering degree in computer science from the National School of Applied Sciences (ENSA), Oujda, Morocco. He is currently a Software Engineer and a Ph.D. student in medical artificial intelligence. His work focuses on the design of intelligent and customized software systems. His professional interests include artificial intelligence, data analytics, software engineering, and healthcare applications.

Ismail Chahid received the DUT degree in information technology from the École Supérieure de Technologie, Oujda, Morocco, in 2008, the B.Sc. degree in IT management from the Faculty of Polydisciplinary Studies, Tétouan, Morocco, in 2009, and the M.Sc. degree in business intelligence from the Faculty of Science and Technology, Béni Mellal, Morocco, in 2012. He is currently the Head of the IT Department at the Faculty of Medicine and Pharmacy of Oujda, where he has been working since 2019. His professional interests include business intelligence, data warehousing, information systems management, and software engineering.

Aissa Kerkour El Miad received the M.Sc. degree in operational research and informatics and the Ph.D. degree in computer science from Mohammed First University, Oujda, Morocco. He is currently a Professor with the Department of Computer Science, Faculty of Sciences, Mohammed First University, Oujda. His research interests include image processing, artificial intelligence, high-performance computing, and data mining.