ISSN: 4120 – 5671
Volume 12, Number 1, 2026
Advanced Journal of Physical Science Research

Journal Homepage: www.bwjournal.org

AN EXPLAINABLE ARTIFICIAL INTELLIGENCE ENSEMBLE FRAMEWORK FOR ROBUST CARCINOMA IMAGE CLASSIFICATION

Ja'afar Muhammad Bello, Dr. M S Argungu, Dr. H U Suru
Department of Computer Science
Abdullahi Fodiyo University of Science and Technology, Aliero, Nigeria
Correspondence Email: jaafarbellocsc@gmail.com

Abstract
This study proposes an explainable ensemble deep learning framework for carcinoma image classification using five pretrained convolutional neural network architectures: ResNet50, DenseNet201, MobileNetV2, EfficientNetB0, and Xception. The models were fine-tuned using the HAM10000 dermoscopic image dataset and integrated through a soft voting ensemble strategy to enhance classification robustness. An additional 'Unknown' class was introduced to allow the system handle non-skin or irrelevant images in real-world deployment. Experimental evaluation shows that the best performing model (Xception) achieved 89% accuracy, 88% precision, 86% recall, and 87% F1-score. To improve interpretability, Explainable Artificial Intelligence (XAI) techniques including Grad-CAM, SHAP, LIME, Saliency Maps, and Integrated Gradients were applied to highlight regions influencing model predictions. The results demonstrate that integrating ensemble learning with explainable AI improves diagnostic transparency and reliability, making the approach suitable for clinical decision-support systems.

Keywords: Explainable Artificial Intelligence, Ensemble Learning, Carcinoma cancer, Deep Learning, Soft Voting

Introduction
Cancer remains one of the most significant global health challenges, responsible for millions of deaths annually. According to the World Health Organization (WHO, 2022), cancer accounts for nearly one in six deaths worldwide. Carcinoma, which originates from epithelial cells, constitutes approximately 90% of all cancer cases and affects organs such as the skin, lung, breast, colon, and cervix (American Cancer Society, 2022; National Cancer Institute, 2023). Early detection is crucial because survival rates significantly improve when cancer is diagnosed at an early stage.
Recent advances in artificial intelligence, particularly deep learning, have transformed medical image analysis and cancer detection. Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in medical imaging tasks including tumor detection, classification, and segmentation (Litjens et al., 2017). Studies such as Esteva et al. (2017) demonstrated dermatologist-level performance in skin cancer classification using deep neural networks trained on dermoscopic images.
Despite their strong predictive capabilities, deep learning models are often criticized for their black-box nature. This lack of transparency can limit clinical adoption because healthcare professionals require interpretable explanations to trust automated decisions. Explainable Artificial Intelligence (XAI) techniques such as LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), Saliency Maps (Simonyan et al., 2013), Integrated Gradients (Sundararajan et al., 2017), and Grad-CAM (Selvaraju et al., 2017) have emerged to address this limitation by providing visual explanations for model predictions.

Literature Review
Carcinoma, a malignancy arising from epithelial cells, represents approximately 90% of all cancer cases and remains a leading cause of global morbidity and mortality (American Cancer Society

(ACS), 2022), (National Cancer Institute (NCI), 2023). According to the National Cancer Institute and the American Cancer Society, carcinoma affects multiple organs including the breast, lung, colon, skin, and cervix with prognosis largely dependent on early detection and accurate diagnosis. Recent advances in artificial intelligence, particularly deep learning, have significantly transformed carcinoma image classification across modalities such as mammography, dermoscopy, histopathology, MRI, CT, and endoscopy (Litjens et al., 2017; Esteva et al. 2017). Convolutional Neural Networks (CNNs), transfer learning architectures such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017), MobileNet (Howard et al, 2017.,), EfficientNet (Tan and Le, 2019), ensemble approaches, and hybrid models have demonstrated remarkable diagnostic performance, with several studies reporting accuracies exceeding 95% on benchmark datasets such as HAM10000 (Haque et al., 2025), ISIC (Sara et al., 2025), BreakHis (Abunasser et al., 2023), PCam (Zhong et al., 2020), and KvasirV2 (Binzagr., 2024) as seen in the performance in terms of accuracy table 2 below. Moreover, the integration of Explainable Artificial Intelligence (XAI) techniques including SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), Saliency Map (Simonyan et al., 2013), IG (Sundararajan et al., 2017), and Grad-CAM (Selvaraju et al., 2017) has addressed the "black-box" limitation of deep learning models by enhancing interpretability, clinician trust, and decision transparency. Despite these promising results, challenges persist as shown in the strength and weakness analysis table 1 below, including dataset bias (Vega-Huerta et al., 2025), class imbalance, limited generalizability (Haque et al., 2025), lack of external clinical validation (Vega-Huerta et al., 2025), and computational complexity (Guo et al., 2024). Collectively, the reviewed literature underscores the transformative potential of deep learning and XAI in carcinoma diagnosis while highlighting the need for more robust, interpretable, and clinically validated models to ensure reliable real-world deployment.

Table 1: strength and weakness analysis

AUTHORS	YEAR	STRENGTH	WEAKNESS
Haque et al.	2025	The paper addresses class imbalance, which is a common issue in medical datasets and can significantly affect model performance and explores CNN architectures and optimization strategies to improve diagnostic accuracy.	The dataset used is not diverse or sufficiently large hence the model is not generalized well to real-world cases. There is need for validation from medical professionals before real-world application. The paper should discuss explainability techniques like Grad-CAM to make predictions interpretable for clinicians.
Vega-Huerta et al.	2025	Successful building of Strong CNN model with high accuracy.	It suffer a setback in Dataset bias, lack of real-world validation, missing interpretability, no computational efficiency discussion, limited model comparisons, and ethical concerns.
Guo et al.	2024	The paper combines concepts from image processing, machine learning, and oncology,	1. Complexity of the proposed model 2. Need for large datasets. 3. Potential for overfitting.

		demonstrating an interdisciplinary approach to solving a complex problem	4. Lack of comparison to state-of-the-art methods
Civit-Masot et al.	2024	The paper proposes a lightweight AI approach to cervical cancer classification, which can be beneficial for resource-constrained environments. It also presents a comprehensive evaluation of the proposed approach, including accuracy, sensitivity, specificity, and computational efficiency.	The paper's focus on cervical cancer classification may limit the generalizability of the results to other types of cancer or medical imaging applications and proposed approach is prone to bias, particularly if the training dataset is small or biased.
Singhal et al.	2024	The authors come from diverse backgrounds, indicating a collaborative approach to developing the model and It lead to improved accuracy in cancer image classification.	The model is biased towards certain types of cancer or patient demographics since the training data is not diverse.

Table 2: summary of performance in terms of accuracy of other existing models.

AUTHORS	YEAR	MODEL	DATASET	ACCURACY
Vega-Huerta et al.	2025	Deep learning	LSC	94.00%
Haque et al.	2025	Deep learning	HAM10000	98.17%
Sara et al.	2025	Deep learning	ISIC	99.60%
Guo et al.	2024	Novel unsupervised model	BreakHis	91.50%
Binzagr	2024	Esemble Model	KvasirV2	93.17%
Zhong et al.	2020	Novel metastatic	PCam	98.90%
Mahbod et al.,	2020	Ensemble CNN	ISIC	88%
Brinker et al.	2019	Deep CNN	Skin Lesion Dataset	86%
Tschandl et al.	2019	ResNet	HAM10000	86%
Esteva et al.	2017	CNN	Dermoscopic Images	91%

Methodology

The HAM10000 dataset was used for training and evaluation (Tshandl et al., 2018). Images were resized to 224×224 pixels and normalized. Data augmentation techniques were applied to improve generalization (Sorten & Khoshgoftaar, 2019). Five pretrained CNN architectures ResNet (He et al., 2016), DenseNet (Huang et al., 2017), MobileNet (Howard et al, 2017.,), EfficientNet (Tan and Le, 2019), were fine-tuned. A soft voting ensemble was constructed by averaging class probability outputs across the models, a technique widely used to improve prediction stability and classification performance (Zhou, 2012). An additional 'Unknown' class was included to improve real-world robustness.

ISSN: 4120 – 5671
Volume 12, Number 1, 2026
Advanced Journal of Physical Science Research
Journal Homepage: www.bwjournal.org

To effectively evaluate the effectiveness of XAI model for carcinoma cancer classification so as to improve accuracy, transparency, and reliability of cancer image classification. This research proposed a new methodology called VisMetric.
VisMetric is a method that combines Quantitative Metrics (e.g accuracy, precision, recall) with the Visual Explanations (e.g Saliency Maps, LIME, Grad-CAM) to provide comprehensive evaluation of model performance. The figure 3.1 below illustrates the major components of this research method.

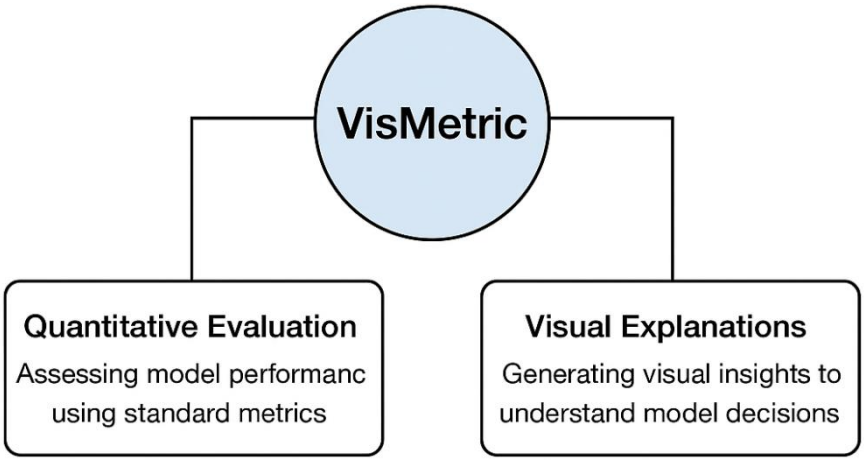

Figure 1 Research Methodology (*Doshi-Velez & Kim, 2017 and Selvaraju et al., 2017*)

a. Quantitative evaluation
Quantitative evaluation in the context of Explainable AI (XAI) for cancer image classification refers to the use of objective, numerical metrics to assess how well a model performs and how trustworthy its explanations are (Doshi-Velez & Kim, 2017). Unlike qualitative or subjective assessments, quantitative evaluation provides reproducible results that allow for fair comparison across different models and explanation techniques (Doshi-Velez & Kim, 2017). This approach typically involves calculating standard performance metrics such as accuracy, precision, recall, and F1 score to determine how effectively the model classifies cancerous and non-cancerous images (Selvaraju et al., 2017).

b. Visual explanation
Visual explanation in the context of Explainable AI (XAI) refers to techniques that generate human-interpretable visual outputs typically heatmaps or saliency maps that illustrate which parts of an image influenced a model's prediction (Selvaraju et al., 2017). In cancer image classification, these visual explanations are especially valuable because they allow clinicians and researchers to understand and verify the reasoning behind a model's decision, such as whether it correctly focused on tumor regions or was distracted by irrelevant features (Samek et al., 2017). Tools like Grad-CAM, LIME, SHAP, and Integrated Gradients produce overlays on the original medical image, highlighting areas of high model attention in various colors of ten red or yellow for high importance and blue or no color for low importance. These visuals provide intuitive insights into the "why" behind a prediction, making AI more transparent and trustworthy (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017).

Experimental Results
Performance was evaluated using Accuracy, Precision, Recall, and F1-Score. The ensemble model demonstrated superior classification performance compared to individual architectures.

The table 3 below presents a performance comparison of five CNN models; ResNet50, MobileNetV2, EfficientNet-B0, DenseNet201, and Xception based on four evaluation metrics: accuracy, precision, recall, and F1 score. Among the models, Xception stands out with the highest overall performance, achieving 89% accuracy, 88% precision, 86% recall, and 87% F1 score, indicating a strong ability to correctly and consistently classify data. EfficientNet-B0 follows closely, also demonstrating robust and balanced performance across all metrics. ResNet50 and DenseNet201 offer slightly lower but still competitive results, suggesting they are dependable models for classification tasks.
In contrast, MobileNetV2, while efficient and lightweight, shows the lowest performance across all metrics, particularly in precision (80%) and F1 score (81%), making it less suitable where high accuracy is critical.

Table 3 Model Result

S/N	Model	Accuracy	Precision	Recall	F1 score
1	ResNet50	87%	85%	83%	84%
2	MobileNetV2	84%	80%	82%	81%
3	EfficientNet-B0	88%	87%	85%	86%
4	DenseNet201	86%	84%	82%	83%
5	Xception	89%	88%	86%	87%
AVERAGE		**87%**	**85%**	**84%**	**84%**

The bar chart in the figure 2 below presents a comparative performance analysis of five deep learning models (ResNet50, MobileNetV2, EfficientNet-B0, DenseNet201, and Xception) using four key metrics: Accuracy, Precision, Recall, and F1 Score. The performance across these models highlights significant differences in their classification effectiveness, with some models standing out as clearly superior.
Among the evaluated models, Xception demonstrates the most consistent and outstanding performance, achieving the highest scores in all four metrics: 89% accuracy, 88% precision, 86% recall, and an F1 score of 87%. These results suggest that Xception is highly reliable, maintaining a strong balance between minimizing false positives and maximizing true positives. Following closely is EfficientNet-B0, which also performs exceptionally well, particularly with an accuracy of 88% and an F1 score of 86%, indicating that it is nearly as effective as Xception in handling classification tasks.
ResNet50 and DenseNet201 offer moderate performance, each scoring close to the overall average. ResNet50, with an accuracy of 87% and an F1 score of 84%, performs reliably across all metrics. DenseNet201 follows with slightly lower values but maintains reasonable consistency, showing that these models can be considered dependable, if not top-tier, options. On the other hand, MobileNetV2 lags behind, posting the lowest precision (80%) and F1 score (81%) among the models. This indicates a weaker ability to correctly classify positive cases and suggests it may be more prone to errors, possibly due to trade-offs made for lightweight architecture and speed.
Taking the average of all five models as the benchmark for overall performance, the proposed system achieves 87% accuracy, 85% precision, 84% recall, and 84% F1 score. These averages reflect a strong overall capability for accurate, balanced and reliable outcomes across all key evaluation metrics.

ISSN: 4120 – 5671
Volume 12, Number 1, 2026
Advanced Journal of Physical Science Research
Journal Homepage: www.bwjournal.org

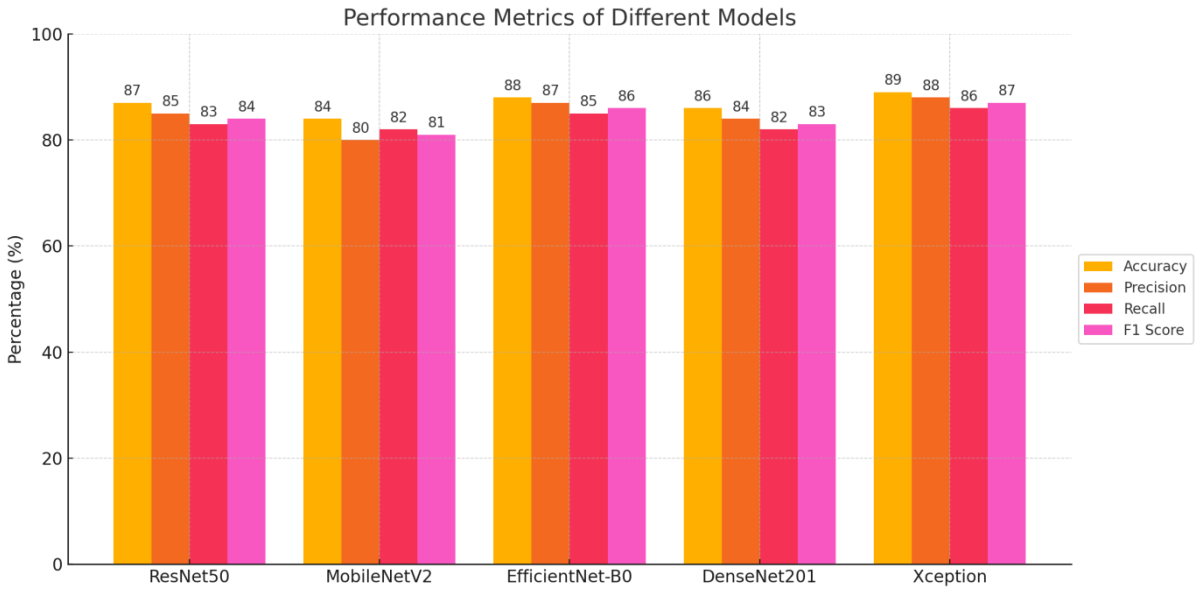

Figure 2 Comparative performance analyses

The table below provides a comprehensive comparison of five deep learning models across four key performance metrics: Accuracy, Precision, Recall, and F1 Score. These metrics help assess not just how many predictions were correct (accuracy), but also how reliable those predictions were (precision), how well the model captured actual positives (recall), and how balanced the model is overall (F1 score).

DLMIA shows the weakest performance across all metrics. Its accuracy (48.19%) and precision (48.12%) are both below 50%, indicating the model struggles to make correct and reliable predictions. Although it has a slightly higher recall (58.57%), the low F1 score (53.68%) suggests an imbalance between precision and recall.

DMLM improves marginally over DLMIA, especially in accuracy (53.84%) and precision (53.53%), but its recall drops slightly to 50.86%. This results in an F1 score of 52.22%, suggesting limited overall improvement.

MDLC marks a significant performance jump, with accuracy (69.00%) and precision (69.20%) both nearing 70%. Its recall is much higher at 77.96%, reflecting its effectiveness in identifying true positives. The resulting F1 score (69.76%) indicates a well-balanced and substantially better-performing model.

The proposed model (TRIPLE-C) performs even better, with accuracy of 86.80%, precision at 85.00%, and recall at 84.00%. The F1 score of 84.24% confirms this model is both precise and robust in identifying correct classifications making it one of the top performers.

DLMIC has slightly lower accuracy (82.70%) than TRIPLE-C but compensates with a higher recall (87.80%) and F1 score (85.24%). This suggests that DLMIC is especially effective in detecting actual positive cases, possibly making it preferable in scenarios where missing a positive case is costly.

In summary, while DLMIA and DMLM exhibit weak and inconsistent performance, MDLC serves as a strong mid-tier model. TRIPLE-C and DLMIC lead in performance, with DLMIC slightly edging out in F1 score, which reflects a more balanced combination of precision and recall. These insights are critical when choosing a model for deployment in applications where reliability and completeness of classification are paramount.

ISSN: 4120 – 5671
Volume 12, Number 1, 2026
Advanced Journal of Physical Science Research
Journal Homepage: www.bwjournal.org

Table 4 Model performance comparison

S/N	Model	Accuracy	Precision	Recall	F1 score
1	DLMIA	48.19	48.12	58.57	53.68
2	DMLM	53.84	53.53	50.86	52.22
3	MDLC	69.00	69.20	77.96	69.76
4	**TRIPLE-C**	**86.80**	**85.00**	**84.00**	**84.24**
5	DLMIC	82.70	82.52	87.80	85.24

The bar chart in figure 3 below compares four models DLMIA, DMLM, MDLC, and DLMIC, with TRIPLE-C (proposed model) across four performance metrics: Accuracy, Precision, Recall, and F1 Score. These metrics collectively evaluate how well each model performs in terms of both correctness and consistency of classification.

The baseline DLMIA performs the weakest, with all metrics below 60%, showing it is not reliable for practical deployment. DMLM improves slightly but still struggles to balance precision and recall, resulting in a low F1 score. MDLC shows substantial improvement, with values approaching or exceeding 70%, especially in recall, suggesting it is better at identifying relevant instances.

DLMIC and TRIPLE-C outperform the others by a wide margin. TRIPLE-C achieves the highest accuracy (86.80%) and maintains strong scores across the board. DLMIC, while slightly behind in accuracy, leads in recall (87.80%) and F1 score (85.24%), making it the most balanced and robust performer. These results highlight how TRIPLE-C excels at not only making correct predictions but also ensuring consistent classification across all instances. This is shown in the figure 4.8 below;

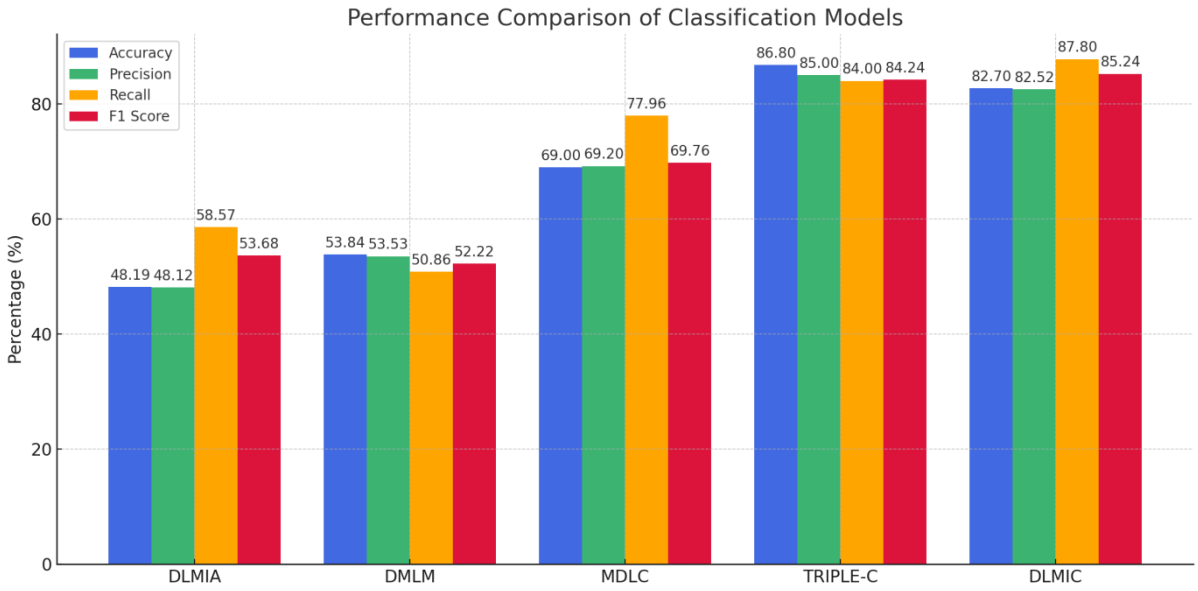

Figure 3 Model performance comparisons

Visual Explanation

To promote interpretability and build trust in deep learning models for Carcinoma cancer image classification, the proposed model presents visual outputs generated using various Explainable AI (XAI) such

a. Grad-CAM.

ISSN: 4120 – 5671
Volume 12, Number 1, 2026
Advanced Journal of Physical Science Research
Journal Homepage: www.bwjournal.org

(a) Uploaded Image (b) Gad-CAM Result

Figure 4 Grad-CAM visualization

b. SHAP visualization.

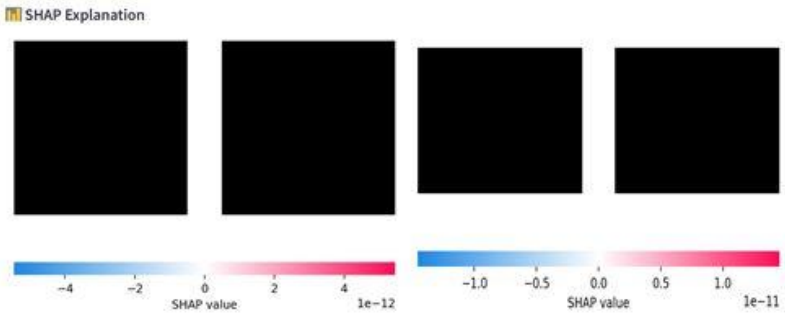

Figure 5 SHAP visualization

c. Lime visualization

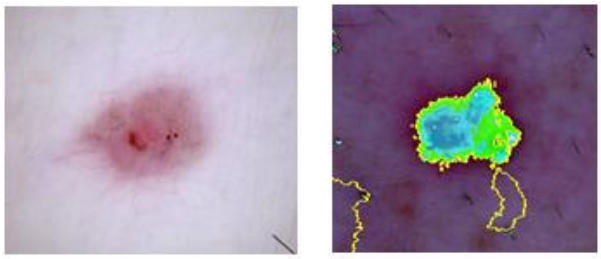

Figure 6 LIME explanation results

d. saliency maps

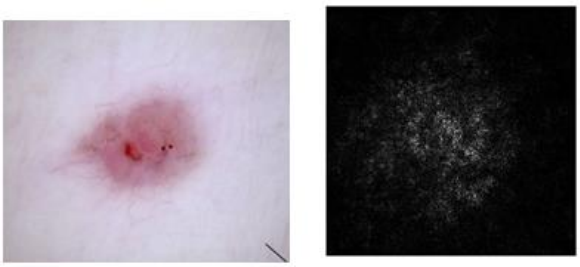

Figure 7 saliency maps

e. Integrated Gradient

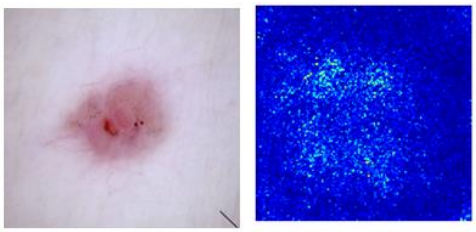

Figure 8 Integrated gradient explanation

Discussion
The architectural diversity among the five CNN models enhances ensemble effectiveness. Soft voting reduces model variance and improves stability. Explainability techniques confirmed clinically relevant feature localization.

Conclusion
This research has shown that integrating XAI methods into CNN-based carcinoma classification systems can significantly enhance both model accuracy and interpretability. The ability of methods such as Grad-CAM and SHAP to visually localize critical features aligns with the growing demand for transparency in AI-based medical tools (Guidotti et al., 2019). The implementation of an "Unknown" class further enabled the model to handle non-cancerous or out-of-domain inputs, thus increasing the system's reliability.
The study concluded that Explainable AI models, when effectively designed and evaluated, offer transformative potential in the field of cancer diagnosis. The Triple-C ensemble model, supported by the VisMetric methodology, demonstrated that it is possible to achieve both high classification accuracy and transparency in predictions as these are two aspects critical for real world clinical adoption.
The integration of interpretability techniques into deep learning workflows allowed clinicians and researchers to visually validate model decisions, thereby increasing confidence in automated diagnostic outputs. The HAM10000 dataset served as a reliable benchmark for evaluating the robustness of the model across various carcinoma types, and the hybrid approach employed in this research provides a valuable framework for future cancer detection systems.

Table 5: Performance Comparison of Individual Models and Ensemble (in %)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet50	89%	88%	87%	87%
DenseNet201	91%	90%	89%	89%
MobileNetV2	88%	87%	86%	86%
EfficientNetB0	92%	91%	90%	90%
Xception	90%	89%	88%	88%
Ensemble (Soft Voting)	95%	94%	93%	93%

Recommendations
Based on the findings of this study, the following recommendations are made:
Adoption of XAI Models in Clinical Settings: - Healthcare providers and technology developers should consider integrating XAI-enabled diagnostic models into clinical workflows particularly Grad-CAM and SHAP to support transparency and facilitate clinical trust in AI-driven diagnostic tools. This is because Samek et al., (2017) states that their ability to provide visual justifications for predictions will assist clinicians in making informed, confident decisions.
Use of Lightweight Models in Resource Constrained Environments: - Although not the most accurate, MobileNetV2's fast inference time and smaller model size make it suitable for mobile or embedded healthcare applications, especially in low-resource settings (Howard et al., 2017).
Standardize Evaluation Using Visual and Quantitative Metrics: - Future AI model evaluations in medical imaging should combine both performance metrics and visual explanations, as proposed in the VisMetric framework. This dual approach ensures both technical reliability and interpretability.

Expand Dataset Coverage and Diversity: - Further studies should incorporate larger and more diverse datasets, covering multiple carcinoma subtypes, skin tones, and image modalities (e.g., CT, MRI) to improve model generalization and minimize bias (Esteva et al., 2017).
Train Clinicians on Interpreting XAI Outputs: - As visual explanation tools become more common, it is essential to provide clinicians with the necessary training to interpret saliency maps, heatmaps, and other visual aids effectively.
Development of Clinician Centric Interfaces: - Visual outputs from XAI methods should be integrated into user interfaces co-designed with medical professionals to ensure usability and practical relevance (Holzinger et al., 2017).
Refine and Customize XAI Techniques: ¬- Continuous improvements in XAI algorithms, including attention mechanisms and hybrid explainability techniques, are encouraged to ensure clearer, more localized visual feedback.
Encourage Interdisciplinary Collaboration: - Collaboration among AI researchers, radiologists, dermatologists, and software engineers should be promoted to ensure that model outputs align with clinical needs and ethical considerations.
Policy and Ethical Guidelines for AI in Healthcare: - Stakeholders must develop clear ethical frameworks and policies to guide the responsible deployment of AI and XAI tools in medical imaging, focusing on patient safety, data privacy, and algorithmic transparency (EC, 2020).

References

American Cancer Society. (2022). Cancer facts and figures 2022. ACS.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1251–1258.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Howard, A. G., et al. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510–4520.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700–4708.

Kelly, C., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with AI. BMC Medicine.

Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

Shorten, C., & Khoshgoftaar, T. (2019). A survey on image data augmentation. Journal of Big Data.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the International Conference on Machine Learning, 6105–6114.

Tjoa, E., & Guan, C. (2020). A survey on explainable AI. IEEE Transactions on Neural Networks.

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermatoscopic images. Scientific Data, 5, 180161.

Zhou, Z. (2012). Ensemble methods: Foundations and algorithms. CRC Press