

ORIGINAL RESEARCH

Open access

Deep Learning-Based Detection of Manipulated Fundus Images: A CNN Concatenation Approach for Medical Image Integrity

James O'Leary^{1*}, Aisling Dunne¹, Sean O'Brien²

Abstract

The fraudulent application of image transformation methods to clinical data represents a severe hazard that can jeopardize patient safety. Consequently, identifying artificial alterations using deep learning frameworks is vital for mitigating the challenges posed by modified medical imagery in clinical environments. The experimental data were partitioned into an authentic fundus dataset and a manipulated dataset. The proposed framework for detecting fundus image modification utilizes a deep learning architecture rooted in a Convolutional Neural Network (CNN) structure, incorporating a concatenation mechanism to achieve rapid processing speeds and minimize the attenuation of input image weights. When evaluating authentic data, the system attained an average sensitivity of 0.08, precision of 1.00, F1-score of 0.99, and an AUC of 0.988. On the manipulated dataset, the framework achieved a sensitivity of 1.00, precision of 0.84, F1-score of 0.92, and an AUC of 0.988. In comparison, a panel of five ophthalmologists demonstrated lower average metrics on the altered images, yielding a sensitivity of 0.71, precision of 0.61, F1-score of 0.65, and an AUC of 0.822. This investigation highlights the potential for mitigating and preventing the complications arising from fabricated medical imagery within healthcare systems. The introduced methodology for identifying altered fundus photographs via a deep learning model outperforms human specialists, proving to be an efficient diagnostic strategy.

Keywords Deep learning, Fundus image, Manipulation, Concatenate, Medical image

*Correspondence:

James O'Leary
james.oleary@tcd.ie

¹ Department of Ophthalmology and Medical Image Forensics, School of Medicine, Trinity College Dublin, Dublin, Ireland

² Department of Fundus Image Manipulation Detection, School of Engineering, University College Cork, Cork, Ireland

Introduction

The progression of image processing methodologies powered by deep learning has initiated a major wave of academic investigation across multiple disciplines [1]. Within these approaches, image transformation systems employ artificial intelligence models to alter visual data through the generation, modification, and integration of distinct features [2]. This technology can assist in reconstructing heavily corrupted photographs or establishing digital cartography from satellite views, thereby offering a viable path for expanding constrained training repositories for deep learning architectures. Furthermore, an investigation team at the University of Lübeck in Germany documented an AI framework that exhibited exceptional capability in oncology diagnostics by utilizing image transformation paradigms [3]. In an investigation conducted by Frid-Adar *et al.* [4], a Generative Adversarial Network (GAN) was implemented to synthesize artificial clinical images to circumvent the difficulties associated with gathering authentic medical data. Through the utilization of GAN-driven data augmentation, the diagnostic capability of a Convolutional Neural Network (CNN) regarding hepatic

lesion classification was enhanced. The investigators contrasted the diagnostic accuracy of the CNN when trained using both original and synthesized datasets, ultimately validating the artificial hepatic lesion photographs via expert clinical reviews. Correspondingly, in a study by Lee *et al.* [5], research was executed to synthesize mammographic imagery using GANs to overcome common constraints in clinical inquiries, such as the anonymization of private patient credentials and prohibitive expert annotation expenses. Consequently, image modification methodologies have been deployed for productive applications—particularly within medical disciplines constrained by limited data access—sparking numerous investigations into the fabrication or generation of clinical imagery [4-7].

While diverse strategies have been introduced to alleviate the hurdles of obtaining clinical datasets, minimal attention has been directed toward preventing the fraudulent exploitation of these tools. Recently, image transformation paradigms have sparked significant public debate under the classification of "Deep-Fake" phenomena [8, 9]. This developing risk of exploitation is uniquely alarming within the clinical sector, where the deceptive deployment of transformation frameworks presents a profound threat to patient well-being. Additionally, the unauthorized exploitation of image fabrication technologies can trigger a range of critical dilemmas.

As of 2017, the economic valuation of the precision medicine market reached 47.47 billion USD, maintaining a compound annual growth rate of 13.3%. Parallel to this market expansion, the volume of clinical trial clearances necessary to confirm the therapeutic validity of novel medical products is consistently escalating [10]. Image transformation frameworks could be deceitfully employed to forge patient enrollment criteria or clinical scenarios, thereby dodging transparent experimental and verification protocols [11]. Furthermore, given that Institutional Review Board (IRB) authorization is an absolute requirement for the majority of clinical investigations, fabricated medical scans could potentially be deceitfully submitted to obtain institutional approval. With the swift transition toward digital infrastructure in social and healthcare ecosystems, including medical billing networks, the likelihood of altering or fabricating diagnostic images to secure illicit insurance disbursements also rises [12]. Indeed, sophisticated image transformation models that obscure the boundary between genuine and synthesized clinical data can be leveraged to forge or manipulate formal medical records [11, 13].

These deceptive technologies impact not only the healthcare sector but any industry dependent on image-centric data, thereby introducing substantial societal dilemmas. Although extensive academic literature concerning image modification has emerged, a relative paucity of research has focused on the identification of these forged assets. In light of the expanding volume of fabricated visuals and the potentially lethal ramifications of medical image forgery, prioritizing research on the detection of manipulated visual data is of paramount importance.

In contrast to conventional studies that primarily concentrate on image synthesis, a few contemporary projects have introduced architectures to identify falsified clinical images. For example, Reichman *et al.* [14] presented a CNN-centric methodology to detect GAN-mediated modifications in CT imaging, demonstrating high capability in identifying both synthetic lesion insertions and deletions. Similarly, Zhang *et al.* [15] introduced a two-tier cascade architecture merging patch-scale anomaly identification with frequency-centric global classification to accurately detect localized GAN alterations in clinical datasets.

These inquiries highlight the accelerating academic interest in medical image forensics. Nevertheless, the majority of historical efforts have focused on CT or MRI modalities. Conversely, our architecture is specifically tailored for fundus photography and establishes its real-world utility by direct benchmarking against practicing ophthalmologists.

The integration of deep learning architectures within clinical imaging has been thoroughly investigated, particularly regarding classification and semantic segmentation procedures utilizing convolutional neural networks (CNNs) [16, 17]. For instance, Krizhevsky *et al.* highlighted the capacity of CNNs in extensive visual classification tasks utilizing the ImageNet repository [16], while Litjens *et al.* provided a thorough synthesis of deep learning implementations in medical image diagnostics, emphasizing its efficacy across diverse analytical procedures [17]. Our investigation expands upon this established foundation by focusing directly on the identification of falsified fundus photographs.

In this paper, we introduce an analytical model structured to recognize altered clinical images within healthcare ecosystems. Specifically, we concentrate on detecting variations in fundus images generated by a previously established deep learning framework and comparing the accuracy of our detection approach with formal diagnostic assessments performed by practicing ophthalmologists.

Previous Research

In an earlier investigation [18], we introduced a medical image generation framework driven by deep learning to mitigate the challenges of acquiring sufficient clinical data, specifically targeting the relative shortage of pathological lesion samples relative to healthy controls, by focusing on fundus photography.

This preceding study utilized the "Ocular Disease Intelligent Recognition" repository from Kaggle [19] to assemble training samples distributed across four fundus classifications: normal, diabetic retinopathy, glaucoma, and macular degeneration. To guarantee data uniformity and standards, a structured filtration and preprocessing pipeline was executed, culminating in a final collection of 356 premium-grade fundus photographs.

The image synthesis network presented in that preliminary study leveraged a Res U-Net architecture, which integrates a standard U-Net framework with Residual Blocks (Figure 1). This methodology successfully synthesized images that closely mirrored the authentic data, surpassing the capabilities of a standard U-Net. In the quantitative appraisal of image correspondence (Table 1), the Res U-Net framework outperformed the baseline U-Net across three evaluation metrics—Root Mean Square Error (RMSE), Structural Similarity Index Map (SSIM), and Fréchet inception distance (FID)—exhibiting a highly notable eightfold advancement in FID, which signifies a substantial improvement in visual fidelity.

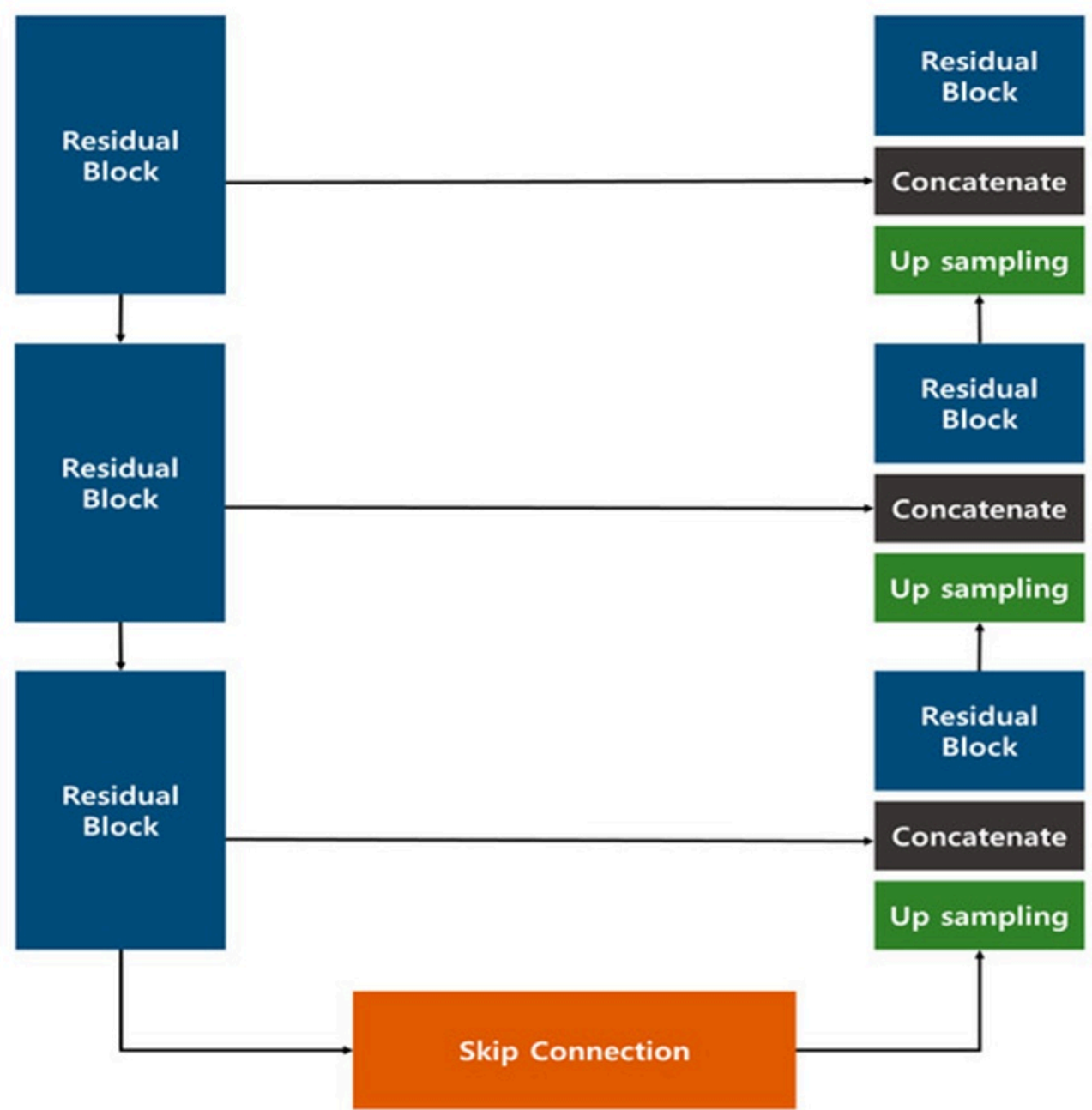


Figure 1. Res U-Net Model Configuration proposed in the previous study [18].

Table 1. Image similarity evaluation from the previous study [18].

Res U-Net		FID	SSIM	RMSE
	Normal	6.37	0.94	15.38
	Glaucoma	63.48	0.75	27.94
	Diabetic Retinopathy	20.66	0.90	20.09

	Macular Degeneration	38.09	0.84	30.4
U-Net	Normal	254.32	0.65	38.58
	Glaucoma	310.90	0.65	39.93
	Diabetic Retinopathy	284.28	0.61	36.76
	Macular Degeneration	253.49	0.64	37.78

Furthermore, to establish the clinical validity of the synthesized photographs, practicing ophthalmologists were recruited to evaluate them. The imagery generated by the Res U-Net framework appeared visually indistinguishable from authentic fundus scans, to such a degree that approximately 30% of the evaluations could not accurately separate genuine samples from synthesized ones. These observations indicate that the Res U-Net paradigm presents a more pronounced potential than a conventional U-Net for practical deployment as a dependable medical image synthesis model.

Although our preliminary investigation validated the practicality of synthesizing fundus photographs via deep learning, it left the potential for malicious exploitation of these artificial images unexplored. In light of growing societal apprehensions regarding deepfake tools and their deceptive application within healthcare environments, we expanded our research to develop strategies for identifying deep learning–altered fundus images.

Data Configuration

To facilitate the identification of altered fundus photographs in the current study, we utilized genuine and manipulated fundus images derived from more than 6000 raw source graphics within the "Ocular Disease Intelligent Recognition" repository [19] that was utilized in our earlier work. Within this framework, a total of 4 clinical states were chosen, encompassing three primary blinding ocular conditions—diabetic retinopathy, glaucoma, and macular degeneration—along side normal controls. The protocols for data selection and the alteration methodologies employed in this research were detailed in our preceding publication [18]. Furthermore, to construct an identification model capable of recognizing diverse fabrication strategies, we integrated altered graphics synthesized via a Cycle GAN into our training dataset. This measure was implemented to avoid an artificially elevated detection accuracy that could occur if the manipulated images only retained characteristics specific to the original generation network. Additionally, because actual clinical environments may present a wide array of falsified materials, we aggregated the framework's repository from multiple data sources. For training the Cycle GAN, we leveraged Kaggle's "MESSIDOR-2 DR Grades" and "Glaucoma Detection" repositories, both containing fundus photographs paired with professional clinical annotations [20, 21].

Table 2 details the dataset distribution utilized for training and assessing our deep learning architecture designed to detect manipulated fundus images. Within the Normal classification, we deployed 350 genuine graphics and 214 altered graphics; for Glaucoma, 203 genuine graphics and 125 altered graphics; for Diabetic Retinopathy, 398 genuine graphics and 147 altered graphics; and for Macular Degeneration, 217 genuine graphics and 129 altered graphics. In aggregate, the dataset comprised 1168 authentic photographs and 615 altered photographs. From this pool, the validation cohort was formed using randomly partitioned cases: 90 genuine graphics and 10 altered graphics for every individual category, spanning the Normal group and all three pathological cohorts.

Table 2. Composition of the fundus image manipulation detection dataset.

	Manipulation Data	Real Data
Normal	214 (Train: 204, Test: 10)	350 (Train: 260, Test: 90)
Glaucoma	125 (Train: 115, Test: 10)	203 (Train: 113, Test: 90)
Diabetic Retinopathy	147 (Train: 137, Test: 10)	398 (Train: 308, Test: 90)
Macular Degeneration	129 (Train: 119, Test: 10)	217 (Train: 127, Test: 90)
Total	615 (Train: 575, Test: 40)	1168 (Train: 808, Test: 360)

Manipulated Fundus Image Detection Model

Our framework for identifying altered fundus photographs accepts as input both authentic fundus scans, scaled to $256 \times 256 \times 3$, and manipulated counterparts produced by an independent modification network. The data input resolution was established at $256 \times 256 \times 3$ to facilitate rapid classification of alterations while maintaining the computational throughput of the deep learning system. The underlying architecture is a CNN that implements a Concatenate mechanism to realize rapid inference while mitigating the degradation of feature weight parameters from the source data. Its streamlined topology ensures high-speed yet highly accurate classification.

Figure 2 demonstrates the layout of the classification framework. Excluding the initial input layer and the terminal flattening matrix, the network is built from four distinct processing blocks, each incorporating a pair of 2D convolutional layers, a concatenate layer, and a down-sampling pooling layer. As source visuals pass through the pooling and convolutional modules, the network isolates dominant features within the fundus scans. Analogous to the encoding pathway of a U-Net, the spatial dimensions of the feature map are halved at every progressive tier, whereas the channel volume is doubled. Deviating from the standard U-Net configuration, which links encoding and decoding features, our approach executes feature merging directly inside the encoder stage to restrict data loss by maintaining high-resolution details during early feature extraction. To counteract the degradation of feature weights that compounds with increasing layer depth, the architecture incorporates a concatenate layer that links directly back to the preceding pooling output.

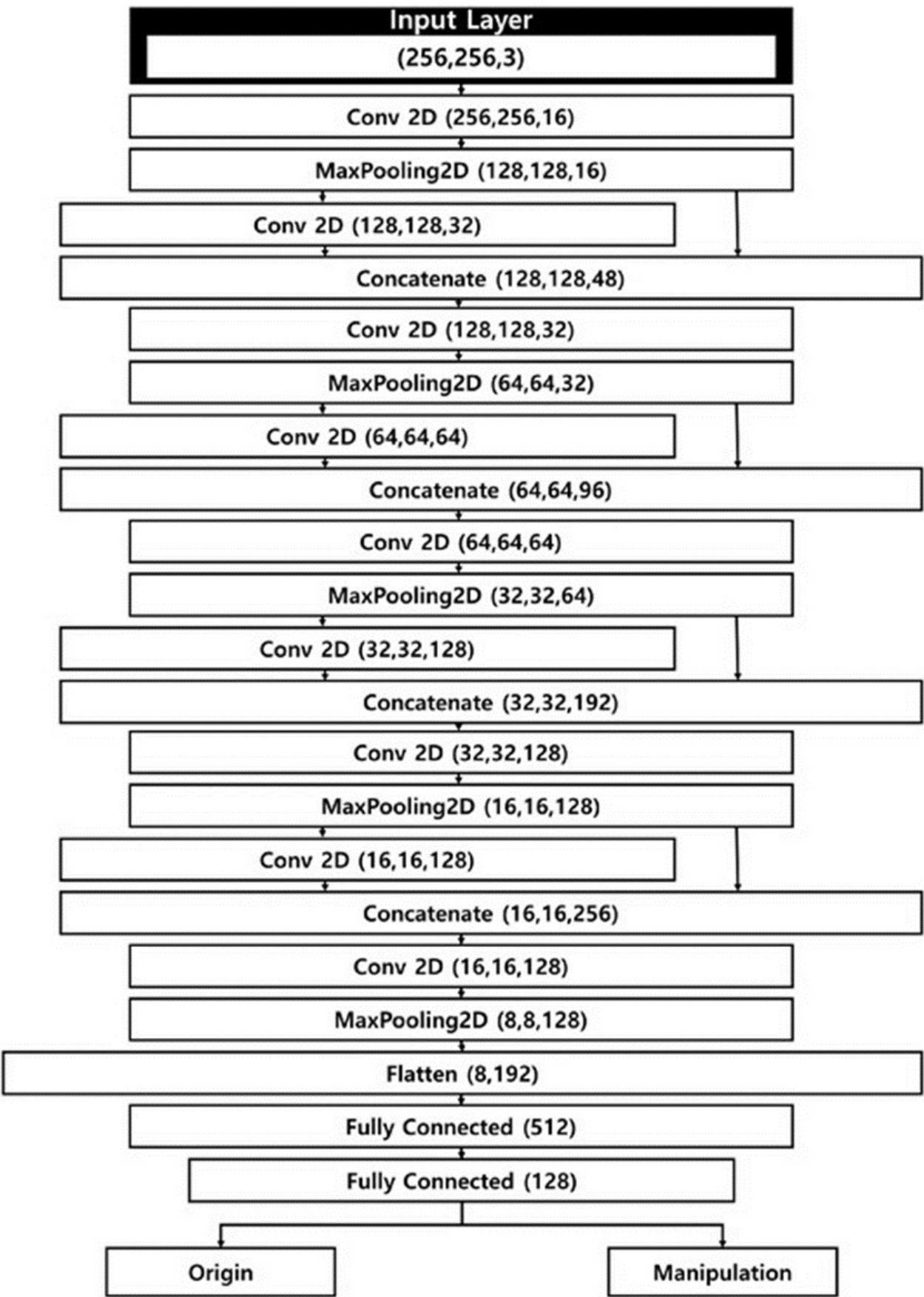


Figure 2. Structure of the fundus image manipulation detection model.

Upon executing the computations within the fifth layer, the tensor transitions to dimensions of $8 \times 8 \times 128$ before entering the flattening stage. In this step, the multi-dimensional structure is transformed into a single-dimensional vector, yielding a flat matrix comprising 8192 components. Through subsequent dense fully connected layers, the dimensionality of this flat matrix is step-wise minimized. Ultimately, the system categorizes each fundus photograph as either authentic or altered. The framework incorporates five convolutional layers using 3×3 kernel structures and a stride value of 1, each paired with batch normalization and a ReLU activation function. A dropout probability of 0.3 was integrated to control overfitting.

Our classification framework is engineered to suppress feature degradation during convolutional operations, an aspect of critical importance since the differentiation of forged imagery depends heavily on both baseline pixel properties and their localized variances. Consequently, the concatenate architecture serves to preserve these mandatory feature values.

Created to counteract the deceptive exploitation of image synthesis models in medical fields, this classification framework can quickly ascertain whether clinical records have been tampered with, and it is tailored to address a diverse spectrum of forged elements that might be encountered in operational clinical settings.

Results and Discussion

Performance of the manipulated fundus image detection model

In this investigation, we utilized 1168 authentic fundus photographs and 615 altered fundus photographs to validate the diagnostic capabilities of the fundus image manipulation detection model. From this distribution, 808 authentic fundus photographs and 575 altered fundus photographs were assigned for network training, while 360 authentic fundus photographs and 40 altered fundus photographs were reserved for network evaluation. Regarding the validation subset, an identical sample volume was maintained across all clinical conditions to facilitate direct cross-comparison with the diagnostic reader study performed by clinical ophthalmologists detailed hereafter.

To execute a quantitative assessment of the fundus image manipulation detection model, four analytical parameters derived from the confusion matrix were implemented: sensitivity, precision, F1-score, and Area Under the Curve (AUC). Sensitivity reflects the capacity of an AI model to correctly pinpoint true positive classifications; in an inquiry of this nature, focused on uncovering fabricated data, misclassification errors can lead to critical clinical ramifications, rendering sensitivity an essential measure. Precision illustrates how reliably the model's positive classifications correspond to actual positive occurrences, and by determining precision, we can verify the statistical trustworthiness of the AI framework's classification output. This parameter must be monitored to refine and optimize the model's overall diagnostic performance. The F1-score, serving as the harmonic mean of precision and sensitivity, is deployed to counter statistical skews that can emerge between these two metrics. In this research, because the volume of fabricated data entries is lower than that of genuine data entries, a balanced group-level quantitative analysis is required, dictating the use of the F1-score. Lastly, the AUC is commonly utilized in diagnostic classification research as it delivers an all-inclusive performance summary across every potential classification threshold.

Table 3 displays the sensitivity, precision, and F1-score outputs recorded by the fundus image manipulation detection model. For authentic data, the architecture achieved an average sensitivity of 0.98, precision of 1.00, and an F1-score of 0.99 across the four clinical categories (Normal, Glaucoma, Diabetic Retinopathy, and Macular Degeneration). These outcomes verify that the framework operates with high accuracy when identifying authentic, untampered fundus photographs.

Table 3. Results of the fundus image manipulation detection model (Sensitivity, Precision, F1-Score).

		F1-Score	Precision	Sensitivity

Real Data	Normal	0.99	1.00	0.98
	Glaucoma	0.99	1.00	0.99
	Diabetic Retinopathy	0.98	1.00	0.96
	Macular Degeneration	0.99	1.00	0.99
Manipulation Data	Normal	0.91	0.83	1.00
	Glaucoma	0.95	0.91	1.00
	Diabetic Retinopathy	0.83	0.71	1.00
	Macular Degeneration	0.99	0.91	1.00

When assessing altered data, the architecture registered a sensitivity of 1.00, precision of 0.84, and an F1-score of 0.92. Looking across the four clinical conditions, the identification accuracy for the Normal and Diabetic Retinopathy cohorts exhibited marginally lower precision and F1-score markers, an effect likely induced by the strong structural similarities between the fabricated and genuine photographs. None the less, all diagnostic metrics remained robust, with Glaucoma and Macular Degeneration yielding high classification performance.

Table 4 outlines the recorded AUC values, indicating 0.989 for Normal, 0.978 for Diabetic Retinopathy, and 0.994 for both Glaucoma and Macular Degeneration. Notwithstanding minor variations among the diagnostic classifications, the model uniformly achieved AUC scores exceeding 0.97, verifying its proficiency in correctly separating authentic fundus photographs from their manipulated counterparts.

Table 4. Results of the fundus image manipulation detection model (AUC).

	Macular Degeneration	Diabetic Retinopathy	Glaucoma	Normal
AUC	0.994	0.978	0.994	0.989

To improve the transparency and explainability of the introduced fundus image manipulation detection model, we utilized Grad-CAM (Gradient-weighted Class Activation Mapping) to map the specific regions of interest within the source image that the network prioritizes during inference. **Figure 3** illustrates the Grad-CAM maps generated from altered fundus photographs representing the four distinct diagnostic cohorts: diabetic retinopathy (DR), glaucoma, macular degeneration (MD), and normal controls.

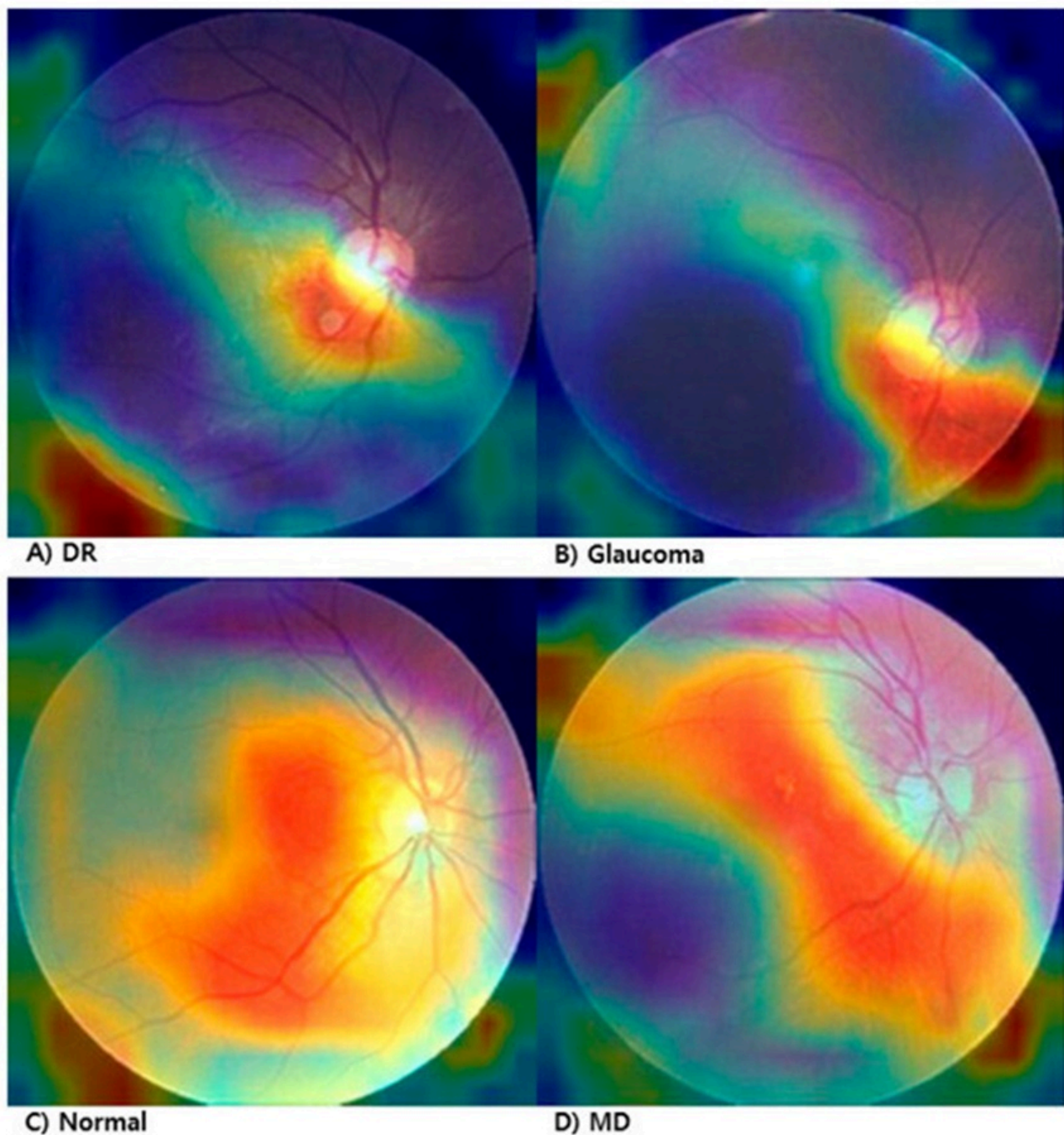


Figure 3. Grad-CAM visualization of manipulated fundus images. (A) Diabetic Retinopathy (DR), (B) Glaucoma, (C) Normal, (D) Macular Degeneration.

In the DR scan (**Figure 3A**), the network concentrates its attention heavily on the inferior central zone, presumably reacting to anomalies in tissue topography and microvascular architecture. In the glaucoma photograph (**Figure 3B**), the system fixes on the optic nerve head region, implying a pronounced sensitivity to anatomical warping or regional luminance variations. In the normal photograph (**Figure 3C**), despite the absence of clinical pathology, the model targets a specific localized structural irregularity, demonstrating high responsiveness to minute textural disparities generated during the manipulation process. In the MD scan (**Figure 3D**), the diagnostic focus is oriented primarily on the macular field.

These visualization outcomes illustrate that the framework successfully isolates regions indicative of fabrication relative to the underlying clinical state, reinforcing its diagnostic transparency and interpretability in clinical practice.

Evaluation of human expert diagnostic performance on forged imagery

To ascertain the practical utility of the automated classification network within operational clinical environments, a comparative diagnostic assessment was executed involving practicing ophthalmologists. The evaluation repository for each specific pathological condition comprised 45 authentic fundus photographs interspersed with 5 digitally altered samples; the participating clinicians were tasked with unmasking the forged assets from a randomized, blinded sequence. Five independent ophthalmologists, each affiliated with a distinct medical center, took part in the study. Utilizing readers from diverse institutional backgrounds allowed for a more rigorous and generalized validation of the automated framework's clinical efficacy across varying diagnostic settings.

Due to geographical separation, the evaluation materials were distributed via electronic mail, with clinicians recording their diagnostic determinations within a standardized spreadsheet containing 200 mixed authentic and fabricated cases. The diagnostic task was conducted using an image resolution of $256 \times 256 \times 3$ pixels—matching the exact input geometry of the deep learning model—with altered images designated by a "0" index within the scoring matrix. No arbitrary time constraints were imposed on the experts, resulting in an average completion window of 7 days. To eliminate diagnostic bias and control variables, no patient demographic data or medical histories were provided for either the authentic or the manipulated cohorts. Evaluating the aggregated average performance of the five clinicians was prioritized over grading individual diagnostic acumen, as the primary objective was to establish a representative baseline for human expert detection capability.

Table 5 outlines the average sensitivity, precision, and F1-score documented by the clinical experts. For unaltered fundus images, the medical specialists achieved an average sensitivity of 0.93, precision of 0.96, and an F1-score of 0.95 across all ocular conditions. Although these statistics indicate robust diagnostic capability, they fall short of the metrics demonstrated by the automated network.

Table 5. Detection results of manipulated fundus images by ophthalmologists (Sensitivity, Precision, F1-Score).

		F1-Score	Precision	Sensitivity
Real Data	Normal	0.94	0.98	0.91
	Glaucoma	0.93	0.95	0.92
	Diabetic Retinopathy	0.94	0.96	0.93
	Macular Degeneration	0.97	0.96	0.97
Manipulation Data	Normal	0.76	0.67	0.88
	Glaucoma	0.51	0.45	0.60
	Diabetic Retinopathy	0.61	0.61	0.64
	Macular Degeneration	0.72	0.74	0.72

When assessing the falsified dataset, the clinicians' diagnostic efficacy diminished sharply, yielding a mean sensitivity of 0.71, precision of 0.61, and an F1-score of 0.65. Notably, human detection capability collapsed most severely when evaluating glaucoma and diabetic retinopathy cases, underscoring the extreme difficulty of identifying these sophisticated digital alterations without computer-aided diagnostic software.

Table 6 details the Area Under the Curve (AUC) statistics for the human readers. The experts achieved an AUC of 0.895 for normal scans, 0.788 for diabetic retinopathy, 0.844 for macular degeneration, and a study-minimum of 0.762 for glaucoma. These values are uniformly lower than the corresponding model metrics, demonstrating a clear imperative for deep learning-driven forensics to expose micro-level image manipulation.

Table 6. Detection results of the fundus image manipulation detection model by ophthalmologists (AUC).

	Macular Degeneration	Diabetic Retinopathy	Glaucoma	Normal
AUC	0.844	0.788	0.762	0.895

This investigation explored the viability of identifying fraudulent modifications in clinical datasets using an advanced deep learning framework, aiming to mitigate the rising threat of medical document forgery within modern healthcare systems. The experimental pipeline was validated using altered fundus photographs sourced from an established generative network, and the model's diagnostic accuracy was directly benchmarked against a panel of 5 practicing ophthalmologists.

The introduced classification framework demonstrated exceptional and uniform performance across both authentic and falsified image repositories. On unmanipulated data, the network achieved a mean sensitivity of 0.98, precision of 1.00, F1-score of 0.99, and an AUC of 0.988 across the four diagnostic cohorts. On the tampered dataset, the system maintained a sensitivity of 1.00, precision of 0.84, F1-score of 0.92, and a stable AUC of 0.988, consistently exceeding a 97% classification accuracy across all clinical categories.

While the framework experienced marginal declines in precision and F1-scores for the normal and diabetic retinopathy altered subsets, the AUC remained universally high. This stability indicates that the network maintains structural robustness irrespective of the underlying generative architecture used to forge the image (e.g., CycleGAN versus Res U-Net), confirming its generalized utility as an image forensics tool.

To confirm its clinical relevance, the network's metrics were contrasted with human diagnostic scores. For authentic imagery, the clinicians recorded a mean sensitivity of 0.93, precision of 0.96, F1-score of 0.95, and an AUC of 0.822. However, when confronted with manipulated images, their scores dropped to a sensitivity of 0.71, precision of 0.61, F1-score of 0.65, and an AUC of 0.822. While human performance on genuine scans was relatively competitive with the AI, the expert panel exhibited distinct performance deficits on altered data, trailing the automated model by margins of −0.29 in sensitivity, −0.23 in precision, −0.27 in F1-score, and −0.166 in AUC. These discrepancies clearly establish the technical superiority of the proposed framework.

Table 7 displays the statistical outputs of a bootstrap simulation comparing the human readers and the automated classification system. We implemented bootstrap resampling (n = 10,000) to calculate 95% confidence intervals for sensitivity, precision, and F1 parameters. This iterative resampling methodology serves as a robust non-parametric statistical framework to derive reliable population estimations from constrained or mathematically uncertain distributions.

Table 7. Comparison of bootstrap between ophthalmologist and manipulated image detection model (Sensitivity, Precision, F1-Score).

Dataset Category	Statistical Measure	F1-Score	Precision	Sensitivity

Real Data	Average Difference	0.0425	0.0375	0.0474
	95% Confidence Interval	0.0250–0.0550	0.0250–0.0475	0.0200–0.0700
Manipulated Data	Average Difference	0.2698	0.2228	0.2901
	95% Confidence Interval	0.1600–0.3800	0.0925–0.3550	0.1800–0.3800

The outcomes of the bootstrap analysis demonstrate a statistically superior performance by the automated framework compared to the clinical specialists across all evaluated diagnostic dimensions (sensitivity, precision, and F1-score) within the altered dataset. These findings substantiate the enhanced efficacy of the proposed model in uncovering manipulated digital materials.

Furthermore, the calculation of the Fleiss' kappa coefficient among the participating ophthalmologists yielded a value of 0.0399, indicating a negligible degree of inter-rater consensus. This metric reveals that the medical evaluators held highly divergent perspectives when analyzing identical diagnostic samples. This lack of alignment appears to stem from confounding variables, such as the total absence of patient records and the compromised visual fidelity characteristic of certain fundus graphics. Additionally, since the clinical evaluation was conducted without pre-existing foreknowledge regarding deep learning-driven image alteration, the absence of standardized diagnostic criteria for spotting digital forgery likely contributed to the depressed human detection rates.

In parallel with our introduced architecture, contemporary academic literature has begun exploring forgery detection within clinical imaging. Reichman *et al.* introduced a CNN-centric infrastructure designed to unmask GAN-driven modifications in CT volumes, registering an AUC of 0.941 when identifying both inserted and extracted anomalies. Zhang *et al.* presented a two-tier cascade framework that pairs localized patch-scale anomaly mapping with global frequency-domain classification, documenting an accuracy exceeding 99% in exposing highly localized GAN alterations. Correspondingly, Sharafudeen *et al.* [22] engineered a deep learning pipeline tasked with identifying altered pulmonary CT scans synthesized via CT-GAN networks; their system achieved a diagnostic accuracy of 91.57%, a sensitivity of 91.42%, and a specificity of 97.20% when discriminating between forged and authentic scans. Moreover, Albahli *et al.* [23] introduced MedNet—an architecture leveraging the EfficientNetV2-B4 backbone—to recognize fraudulent pulmonary CT scans, achieving an overall classification accuracy of 85.49% on tampered datasets.

While the preponderance of these forensic methodologies centers on CT and MRI modalities, our framework uniquely addresses the problem of fundus image fabrication. The proposed model achieved a robust AUC of 0.988 across diverse pathological states, substantially exceeding the diagnostic performance of practicing ophthalmologists (who recorded an AUC of 0.822) under simulated real-world validation constraints. This discrepancy highlights the domain-specific utility and forensic capability of our approach.

Although this investigation confirms that the fundus image manipulation detection model operates with exceptional precision, several technical limitations persist, which we intend to resolve in subsequent iterations. First, the model was trained using three public repositories containing imagery altered specifically via Cycle GAN and Res U-Net frameworks. While utilizing three independent datasets and two separate generation models helps simulate clinical variability, future iterations must incorporate multicenter clinical data collected from authentic operational environments rather than relying exclusively on open-source platforms, alongside a more heterogeneous array of fabrication techniques. Second, this inquiry remains confined to identifying modifications exclusively within fundus photography. Given that clinical records encompass numerous imaging modalities, data fabrication risks threaten every medical specialty; thus, establishing forensic detection systems across a wider spectrum of diagnostic imaging is imperative. Third, the depressed diagnostic accuracy exhibited by the ophthalmologists may be partially attributable to the degraded resolution of specific altered fundus samples. This implies that future synthetic imagery must be generated with structural and visual fidelity that

perfectly matches the quality of authentic clinical data to ensure unbiased performance comparisons. Fourth, this study did not benchmark large-scale, state-of-the-art image classification architectures such as ResNet, EfficientNet, or Vision Transformers. Though widely deployed, these heavy architectures impose substantial computational and memory demands, which diverges from our core objective of designing a lightweight forensic tool. We also recognize the scientific value of ablation analyses to isolate the contributions of distinct structural components. While such validation experiments were omitted here, the integration of concatenation layers and specific preprocessing pipelines was guided by empirical performance gains noted during preliminary exploratory testing. Moving forward, we intend to execute formal quantitative comparisons against high-capacity classification networks and supply rigorous ablation studies tracking architectural modifications.

Conclusion

This study introduced a lightweight, CNN-driven deep learning framework engineered to detect fraudulent alterations in medical fundus photography. The system was trained and validated utilizing a heterogeneous dataset comprising authentic photographs alongside synthetic manipulations produced via CycleGAN and Res U-Net architectures. The framework achieved superior diagnostic performance across all target metrics, yielding a mean AUC of 0.988 and demonstrating pronounced resilience when exposing tampered visual assets.

Direct benchmarking against five practicing ophthalmologists indicated that the automated network significantly outperformed human specialists in identifying forged images, particularly in highly complex diagnostic categories such as glaucoma and diabetic retinopathy. While the medical experts exhibited comparable diagnostic proficiency when evaluating authentic clinical data, their diagnostic sensitivity and accuracy decayed substantially when presented with manipulated images.

These insights indicate that the introduced model provides an effective means of differentiating authentic fundus images from altered variations, irrespective of the underlying synthesis method. The proposed framework represents a highly viable solution for reinforcing clinical data integrity and preventing the fraudulent exploitation of synthetic medical imaging. Future research directions will focus on broadening the validation repository through multicenter, real-world data collection and comparing the network's efficiency against larger, state-of-the-art classification architectures including ResNet, EfficientNet, and Vision Transformers.

Acknowledgements

None

Conflict of interest

None

Financial support

This research was supported by a grant of the Information and Communications Promotion Fund through the National IT Industry Promotion Agency (NIPA), funded by the Ministry of Science and ICT (MSIT), Republic of Korea.

Ethics statement

None

Received: 12 Nov 2022

Revised: 29 Dec 2022

Accepted: 16 Jan 2023

Published online: 15 March 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Dhamo H, Farshad A, Laina I, Navab N, Hager GD, Tombari F, et al. Semantic image manipulation using scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA. p. 5213-22. <https://doi.org/10.1109/CVPR42600.2020.00526>.
- Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017; Venice, Italy. p.2223-32. <https://doi.org/10.1109/ICCV.2017.244>.
- Uzunova H, Ehrhardt J, Jacob F, Frydrychowicz A, Handels H. Multi-scale GANs for memory-efficient generation of high-resolution medical images. In: Med Image Comput Comput Assist Interv (MICCAI). Shenzhen, China; 2019. p. 112-20. https://doi.org/10.1007/978-3-030-32226-7_13.
- Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*. 2018;321:321-31. <https://doi.org/10.1016/j.neucom.2018.09.013>.
- Lee HS, Song HJ, Her YD. A study on the image generation with residual U-Net generator of CycleGAN in digital breast tomography. *J Knowl Inf Technol Syst*. 2023;18:1633-41.
- Iqbal T, Ali H. Generative adversarial network for medical images (MI-GAN). *J Med Syst*. 2018;42(11):231. <https://doi.org/10.1007/s10916-018-1072-9>.
- Fan C, Lin H, Qiu Y. U-Patch GAN: a medical image fusion method based on GAN. *J Digit Imaging*. 2023;36(1):339-55. <https://doi.org/10.1007/s10278-022-00696-7>.
- Westerlund M. The emergence of deepfake technology: a review. *Technol Innov Manag Rev*. 2019;9(11):39-52.
- Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: a survey of face manipulation and fake detection. *Inf Fusion*. 2020;64:131-48. <https://doi.org/10.1016/j.inffus.2020.06.014>.
- World Health Organization. Number of clinical trial registrations by location, disease, phase of development, age and sex of trial participants (1999-2024) [Internet]. Geneva: World Health Organization; 2025 [cited 2025 Apr 22].
- Kim H, Jung DC, Choi BW. Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks. *J Korean Soc Radiol*. 2019;80:259-73. <https://doi.org/10.3348/jksr.2019.80.2.259>.

National Health Insurance Service. Significance and characteristics of health insurance system [Internet]. Wonju: National Health Insurance Service; 2025 [cited 2025 Apr 22].

Mirsky Y, Mahler T, Shelef I, Elovici Y. CT-GAN: malicious tampering of 3D medical imagery using deep learning. In: Proceedings of the 28th USENIX Security Symposium (USENIX Security 19); 2019; Santa Clara, CA. p.461-78.

Reichman B, Jing L, Akin O, Tian Y. Medical image tampering detection: a new dataset and baseline. In: Pattern Recognition. ICPR International Workshops and Challenges. Cham: Springer; 2021. p.266-77.
https://doi.org/10.1007/978-3-030-68763-2_20.

Zhang J, Huang X, Liu Y, Han Y, Xiang Z. GAN-based medical image small region forgery detection via a two-stage cascade framework. PLoS One. 2024;19(1):e0290303.
<https://doi.org/10.1371/journal.pone.0290303>.

Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84-90.
<https://doi.org/10.1145/3065386>.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
<https://doi.org/10.1016/j.media.2017.07.005>.

Kim YS. A study on the performance comparison of the fundus image generation model. Math Stat Eng Appl. 2022;71:527-37.

Ocular Disease Recognition: Right and Left Eye Fundus Photographs of 5000 Patients [Internet]. Kaggle; [cited 2025 Apr 22].

MESSIDOR-2 DR Grades: Adjudicated DR Severity, DME, and Gradability for the MESSIDOR-2 Fundus Dataset [Internet]. Kaggle; [cited 2025 Apr 22].

Glaucoma Detection: OCT Scans [Internet]. Kaggle; [cited 2025 Apr 22].

Sharafudeen MJ, Chandra Sekaran VS. Leveraging vision attention transformers for detection of artificially synthesized dermoscopic lesion deepfakes using Derm-CGAN. Diagnostics (Basel). 2023;13(5):825.
<https://doi.org/10.3390/diagnostics13050825>.

Albahli S, Nawaz M. MedNet: medical deepfakes detection using an improved deep learning approach. Multimed Tools Appl. 2024;83(16):48357-75.
<https://doi.org/10.1007/s11042-023-17562-5>.