IJIRIS: International Journal of Innovative Research in Information Security
Volume 12, Issue 07, October 2026
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2026.v1207.02

Medical Deepfake Detection Using GAN-Assisted Training, CNN Classification, YOLO Localization, and NLP-Based Report Consistency Analysis

Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Department of Information Science and Engineering
Vemana Institute of Technology (Autonomous), Bangalore, India
Snehas.is2023@vemanait.edu.in, sonusrisamyuktha.is2023@vemanait.edu.in,
suprajabn.is2023@vemanait.edu.in, vaishnavis.is2023@vemanait.edu.in
Dr.Aruna Reddy H
Associate Professor and Head,
Department of Artificial Intelligence & Machine Learning
Vemana Institute of Technology (Autonomous), Bangalore, India
arunareddyh@vemanait.edu.in https://ror.org/00ha14p11
https://orcid.org/0009-0009-5596-2403

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.12/Issue07/OCIS10081
Research Article | Open Access | Double-Blind Peer-Reviewed| Article ID: IJIRAE/RS/Vol.12/Issue07/OCIS10081
Received:02,September 2026,Revised:28,September 2026,Accepted:03,October 2026,Published Online:30,October2026
https://www.ijiris.com/volumes/Vol12/iss-07/02.OCIS10081.pdf
Article Citation:Dr.Aruna,Sneha,Sonu,Supraja,Vaishnavi(2026),Medical Deepfake Detection Using GAN-Assisted Training, CNN Classification, YOLO Localization, and NLP-Based Report Consistency Analysis. IJIRIS: International Journal of Innovative Research in Information Security, Volume 12, Issue 07 of 2026 pages 343-351
Doi:> https://doi.org/10.26562/ijiris.2026.v1207.02 **BibTeX Key:** Dr.Aruna@2026Medical
IJIRIS papers should be cited as IJIRIS(International Journal of Innovative Research in Information Security, AM Publications, India 2026, ISSN 2349-7017, https://doi.org/10.26562/ijiris.2026.v1207.02 The journal's official abbreviation is IJIRIS. **ORCiD:** https://orcid.org/0009-0004-9398-7488 About the License: Copyright©2026 copyright by the authors. This article is an open access and license under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

Abstract: Manipulated medical images create an integrity problem that cannot be addressed by image classification alone. A useful verification workflow must distinguish suspicious scans, identify potentially altered regions, and present any disagreement with the accompanying report without treating an automated prediction as a diagnosis. This paper presents Medical Deepfake Detector, a proposed framework combining generative adversarial network (GAN)-assisted training, convolutional neural network (CNN) classification, You Only Look Once (YOLO)-based localization, and natural language processing (NLP)-based report analysis. CT and MRI images from the principal scope, while chest X-ray examples are discussed as an additional application requiring separate validation. Synthetic manipulations support offline training; classification and localization provide complementary forensic outputs; and report analysis extracts findings for a conditional consistency check. A chatbot-oriented interface consolidates these outputs for human review. The paper presents the architecture, methodology, implementation specification, threat model, and evaluation protocol. The available project report contains illustrative examples and inference snippets but does not provide dataset counts, trained checkpoints, experiment logs, or numerical benchmarks. Accordingly, the results are presented as qualitative observations rather than validated performance claims. The framework establishes a basis for evaluating multimodal medical-image integrity while making its evidence limitations and clinical boundaries explicit.

Index Terms: medical deepfake detection, medical image integrity, generative adversarial networks, convolutional neural networks, YOLO, natural language processing, report consistency, explainable interface

I. INTRODUCTION

Medical-image integrity is important wherever diagnostic scans are stored, transferred, or interpreted. A manipulated image can retain a plausible appearance while changing the evidence available to a reviewer. CT-GAN demonstrated that deep-learning-based manipulation can insert or remove indications of disease in volumetric CT images, establishing a concrete motivation for medical-image forensics [1]. Detecting such changes is different from identifying whether an image contains an abnormality: a genuine abnormal scan must not be labeled a forgery simply because its anatomy is unusual. The Medical Deepfake Detector project addresses this distinction through a modular verification workflow. The proposed system accepts medical scans and, when available, corresponding reports; prepares the inputs; estimates image-level tampering; highlights suspicious regions; and presents the evidence through an interactive interface. GAN-assisted data generation is confined to model development. CNN and YOLO components address classification and localization, respectively, while an NLP component structures the accompanying report for further comparison. The central research question is whether these complementary signals can be integrated into an interpretable integrity-review process without conflating tampering, disease, and textual disagreement. A report mismatch may indicate an incorrect pairing, an incomplete description, or a model error rather than image forgery. Conversely, a manipulated image and a jointly edited report may agree. The system therefore treats report consistency as supporting evidence, not as an authenticity guarantee.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2026.v1207.02

This manuscript develops the architecture described in the project report into a research-oriented specification. It preserves the principal modules while distinguishing documented design choices from proposed evaluation refinements. It does not claim a new CNN architecture, a new object detector, or experimentally established superiority. Its contribution is the integration specification and the explicit separation of what each module predicts, what evidence is available, and what must be measured before practical deployment.

II. RELATED WORK

A. Medical-image manipulation and forensic detection

Mirsky et al. introduced CT-GAN to demonstrate malicious changes to three-dimensional medical imagery [1]. Their work motivates the threat model, but its attack results are not detection results for the present project. A detector trained on a single manipulation process also requires testing against other processes before broad robustness can be claimed. Zhou et al. studied image manipulation using complementary content and noise features [2]. Their two-stream approach emphasizes that tampering artifacts differ from ordinary semantic objects. This distinction informs the present design: a localizer must be trained on manipulated regions rather than merely on lesions, organs, or other visible structures. Findings from general image-forensics benchmarks do not establish performance on CT or MRI data.

B. Generative learning and CNN feature extraction

Goodfellow et al. introduced adversarial learning with a generator and discriminator [3]. In this project, GANs are proposed as a source of labeled synthetic examples for defensive training, not as an online verification stage. The resulting synthetic data require provenance tracking and separation from the final evaluation set. Residual networks [4] and densely connected networks [5] provide candidate CNN backbones. The project report names ResNet and DenseNet as alternatives but does not identify a selected variant or a reproducible training configuration. Their relevance here is as established feature-extraction architectures; the project cannot inherit their published benchmark performance without task-specific experiments.

C. Region localization

YOLO formulates object detection as prediction of spatial bounding boxes and associated class information [6]. This makes a YOLO-based model a candidate for displaying suspicious image regions. However, using a YOLO implementation does not by itself make a system a forgery detector. The target labels, annotations, training data, and validation procedure must specifically represent medical-image tampering. The original YOLO publication is cited as an architectural foundation, not as documentation of the exact version proposed in the report.

D. Report understanding

BERT provides a contextual language representation that can be adapted to downstream text tasks [7]. CheXbert demonstrates a report-labeling application combining automatic labels with expert annotations in the chest X-ray domain [8]. These studies support treating report interpretation as a separately evaluated NLP problem. They do not establish that a language model can verify image authenticity, or that a chest X-ray report labeler transfers directly to CT and MRI. Table I summarizes the connection between these foundations and the proposed framework. The comparison concerns task relevance rather than a numerical ranking across incompatible datasets.

Table I - Literature Foundations and Their Role in the Proposed Framework

Work	Established contribution	Role in this project	Boundary of applicability
CT-GAN [1]	Demonstrates manipulation of volumetric medical imagery.	Motivates insertion/removal threats and defensive evaluation.	An attack demonstration is not evidence of this detector's accuracy.
Image forensics [2]	Combines content and noise cues for tampered-region detection.	Motivates forensic features and region-level supervision.	General-image results require separate medical-domain validation.
GANs [3]	Introduces adversarial generative learning.	Supports the proposed offline synthetic-data component.	Synthetic examples alone do not establish real-world generalization.
ResNet / DenseNet[4],[5]	Provides residual and dense feature-extraction architectures.	Supplies candidate image-classification backbones.	No project-specific backbone comparison or checkpoint is supplied.
YOLO [6]	Predicts bounding boxes and class information in a unified detector.	Provides a foundation for suspicious-region localization.	Forgery-specific training and annotations remain necessary.
BERT/CheXbert [7],[8]	Supports contextual text modeling and radiology-report labeling.	Informs structured report extraction and comparison.	Text extraction is not visual diagnosis or proof of authenticity.

III. RESEARCH GAP AND CONTRIBUTIONS

The project-level gap is the absence of an explicitly defined connection between a scan-level authenticity score, a localized tampering prediction, and the meaning of the accompanying text. Simply placing these modules in sequence leaves important questions unanswered: what a bounding box represents, how a textual statement is aligned with an image finding, and what happens when one module is unavailable or uncertain. The proposed framework makes four design contributions. First, it separates offline synthetic-data preparation from online inference. Second, it preserves classification and localization as distinct outputs rather than assuming that a global label provides spatial evidence.

IJIRIS: International Journal of Innovative Research in Information Security
Volume 12, Issue 07, October 2026
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2026.v1207.02

Third, it defines report checking as conditional on a matched study and an appropriate source of visual findings. Fourth, it specifies an evaluation plan that separates classification, localization, text extraction, consistency checking, and interface behavior. These are integration and evaluation contributions, not claims of a first-of-its-kind system or of improved accuracy over published methods. Demonstrating an empirical contribution requires the experiments described later in the paper.

IV. SYSTEM ARCHITECTURE

A. Input and data preparation

The principal inputs are CT or MRI images and an optional corresponding radiology report. Each processing request should retain a study identifier, modality, image dimensions, and preprocessing record. A report must be paired with the same study before semantic comparison is attempted. An absent or unverified report should disable only the report-comparison branch, not silently force an inconsistent label. The report's example classifier uses a two-dimensional image resized to 224×224 pixels. This establishes the scope of the supplied inference illustration, not a validated volumetric pipeline. Supporting complete CT or MRI studies requires an explicit choice of slice selection, inter-slice context, and study-level aggregation. Those choices are not documented in the supplied implementation material.

B. Image-analysis branches

The CNN branch outputs an image-level tampering score. The YOLO branch outputs candidate manipulated regions and their scores. Figure 1 represents these as parallel branches after preprocessing. This resolves the report's mixture of sequential block-diagram descriptions and parallel architecture descriptions without requiring one model's prediction to become another model's input. Keeping the branches separate makes disagreement observable. A global classifier may flag an image for which the localizer finds no region. That outcome should trigger review rather than fabrication of a bounding box. Similarly, a localizer may identify a region while the global score remains below its decision threshold.

C. Text-analysis and consistency branch

The text branch extracts findings, anatomical locations, laterality, and assertion status where supported by its label schema. Consistency checking then requires comparable visual findings. A tampering classifier and a forgery localizer do not necessarily predict whether a tumor, lesion, or other pathology is present. The architecture therefore includes a conditional visual-finding source, such as a separately validated finding model or an expert annotation. This dependency is a proposed clarification, not a demonstrated module in the supplied code. When that source is missing, the interface may display extracted report statements but should mark image–report consistency as not assessed. This avoids treating a suspicious-region box as a diagnosis or assuming that every localized alteration corresponds to the same clinical concept.

D. Aggregation and presentation

The aggregator preserves each branch's status and displays a concise explanation. The intended output comprises an image-level prediction, a suspicious-region overlay when available, extracted report findings, the consistency status, and processing limitations. The chatbot is a presentation layer over these structured results. It should not create clinical findings beyond those supported by the modules or the supplied report.

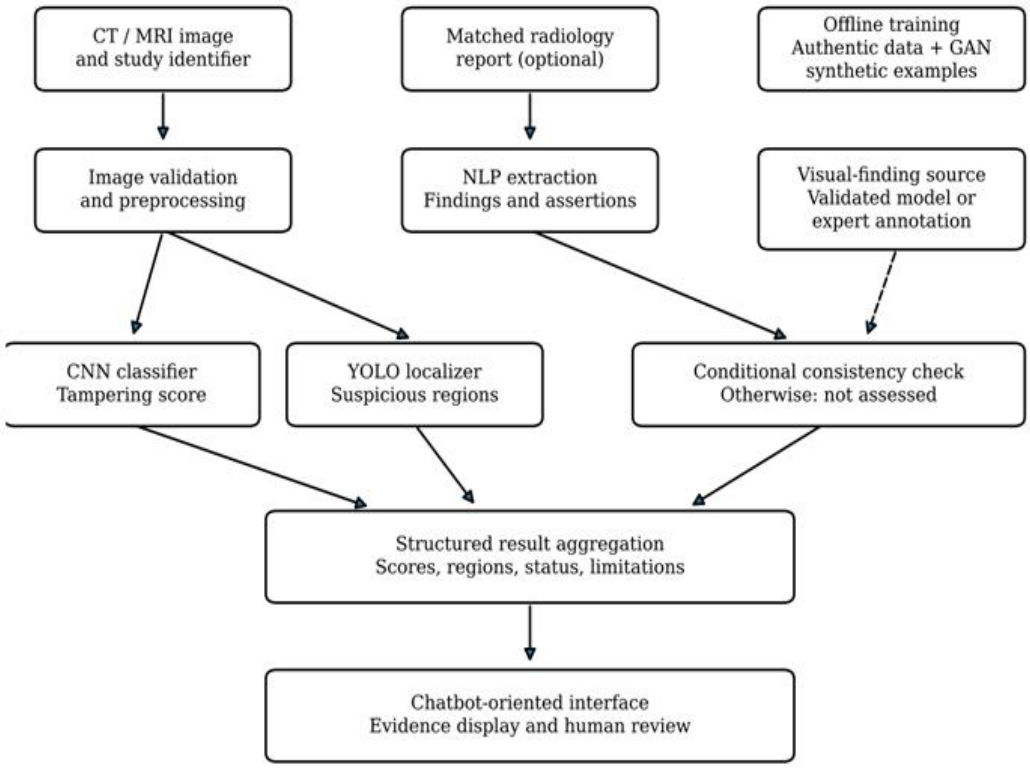


Fig. 1. Proposed verification architecture. GAN-assisted data generation is an offline activity used to develop the image models, not an online inference step. The dashed arrow denotes an additional visual-finding dependency needed for semantic report comparison; this component is not demonstrated by the supplied inference snippets.

IJIRIS: International Journal of Innovative Research in Information Security
Volume 12, Issue 07, October 2026
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2026.v1207.02

V. METHODOLOGY

A. Dataset construction and provenance

The project proposes authentic scans from public repositories, corresponding reports, and synthetic manipulated examples, but supplies no named dataset release or case inventory. Before training, each record should include its source, modality, study grouping, authenticity label, manipulation label where applicable, and permission for research use. Report availability and pairing quality should be recorded separately from image labels. Localization requires annotations describing the altered region rather than an unrelated abnormality. Original scans, derived images, synthetic variants, and slices from the same patient must remain in the same data partition. This is a proposed experimental safeguard, not a description of a completed split. The dataset should include genuine abnormal scans as negative examples for the forgery task. Otherwise, a classifier could learn to associate pathology with manipulation. Cases with normal anatomy, genuine pathology, image artifacts, and different acquisition conditions should be distinguishable in the evaluation inventory.

B. Preprocessing

Preprocessing includes input validation, resizing where appropriate, and intensity normalization. For the supplied two-dimensional inference example, the report describes resizing and division by 255. That operation assumes an image representation compatible with the chosen intensity scale; it is not a complete specification for arbitrary medical-image formats. The implementation must document supported formats and transformations, and apply training-consistent preprocessing at inference. The project also proposes noise removal and Error Level Analysis (ELA). These should be treated as optional experimental branches rather than universally beneficial steps. Any operation that changes the image may also change the traces used by a forensic model. The ELA branch should be evaluated only for explicitly documented image encodings and recompression settings, with a no-ELA baseline. No evidence in the report establishes its utility across all supported modalities.

C. GAN-assisted training

The GAN component is intended to supplement the training set with controlled manipulated examples. Its role is to increase the variety of forensic labels available for model development, not to prove that the training distribution captures all realistic attacks. Generated files should be marked as synthetic and kept separate from clinical records. The generation process should use only the training partition for model fitting. A comparison between training with and without synthetic augmentation is required to establish whether the additional examples help. Testing should also include manipulation sources not used to construct the training data. Generator configuration, source-image grouping, and synthetic-to-authentic proportions remain unspecified in the report and must be recorded before the experiment is reproducible.

D. Image-level classification

Let $s_fake(I)$ denote the CNN score for image I, with larger values indicating greater predicted tampering likelihood. A binary decision can be written as

$$\hat{y}(I) = 1[\, s_fake(I) \geq \tau_cls \,], \qquad (1)$$

where τ_cls is selected using validation data. The report's example uses a threshold of 0.5, but provides no calibration study or class-label mapping artifact. Evaluation should confirm the score interpretation before selecting an operating threshold. A raw model score should not be described as a validated probability of forgery. ResNet and DenseNet are candidate backbones, not simultaneously verified components. Training must specify the selected architecture, initialization, loss, optimizer, learning-rate schedule, augmentation, stopping rule, and random seed. These details should accompany the checkpoint used for all reported test results.

E. Suspicious-region localization

The localization branch returns bounding boxes with associated tampering scores. Training examples must identify altered regions, and negative examples must include authentic scans with legitimate abnormalities. The system should preserve the distinction between a suspected edit and the pathology that the edit might depict. Bounding boxes must be mapped back from model coordinates to the original display coordinates. Resize factors, padding, and cropping parameters therefore belong in the preprocessing record. An empty set of boxes means that no region met the chosen detector threshold; it is not proof that the complete image is authentic. The model variant and its thresholds are configuration details to be validated, not inferred from the name YOLO alone.

F. NLP-based consistency analysis

Report analysis begins by extracting a structured representation of relevant statements. At minimum, it should distinguish a finding from its assertion status. For example, "lesion present," "no lesion," and "cannot exclude a lesion" should not collapse to the same keyword label. Anatomy, laterality, and study context should be preserved where available. The proposed comparison uses three states: consistent, inconsistent, or not assessed. A comparison is assessed only when the report is correctly paired and the visual-finding source supports the same concept and anatomical scope. An unavailable visual finding, unsupported concept, or uncertain pairing produces not assessed. Expert-labeled image–report pairs are required to evaluate this decision process. Consistency does not establish authenticity. A matching image and report may both be wrong, and a mismatch may arise without malicious modification. The interface should expose conflicting statements rather than convert every mismatch into a fake-image label.

VI. IMPLEMENTATION

A. Proposed software organization

The report identifies Python as the main language; TensorFlow or PyTorch for deep learning; OpenCV for image operations; Flask or FastAPI for the backend; and NLP libraries for report processing. It also lists web technologies and Streamlit as interface options. These alternatives describe a development plan rather than a fixed, versioned deployment stack.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2026.v1207.02

A practical implementation can separate ingestion, preprocessing, classification, localization, text extraction, aggregation, and presentation into modules with explicit input and output contracts. Each module should return a status alongside its data, allowing unsupported inputs and missing models to be reported without substituting default success values. Table II distinguishes supplied artifacts from work still required.

Table II - Implementation Evidence in the Supplied Project Report

Component	Available material	Missing evidence
CNN inference	Image loading, preprocessing, and prediction snippet.	Checkpoint, training configuration, labels, and tests.
YOLO inference	Model load, box extraction, and overlay example.	Forgery-trained weights, annotations, version, and tests.
GAN training	Narrative methodology and architectural role.	Training code, data inventory, and generation logs.
NLP analysis	Keyword-based report comparison description.	Parser and labeled paired examples.
Interface	Upload and explanation workflow description.	Runnable service, request traces, and user tests.

B. Scope of the supplied inference examples

The CNN appendix loads medical_deepfake_model.h5, preprocesses test_scan.jpg, and prints a binary message. The PDF does not supply the referenced model or image. The snippet documents an invocation pattern but does not establish the model's training history, accuracy, or supported input distribution. The localization appendix loads yolov8.pt and draws boxes on an image. A checkpoint filename does not identify the classes on which it was trained. The example also does not establish complete handling of multiple detections, empty results, or input errors. These details require a runnable implementation and tests before supporting an end-to-end claim.

C. Interface behavior

The chatbot-oriented interface should confirm that the scan was accepted and report the status of each branch. It should distinguish "not assessed" from "no suspicious region detected," and distinguish both from a processing error. If a report is absent, the interface should display image-analysis outputs without claiming that a consistency check succeeded. Explanations should be assembled from structured outputs. An illustrative template is: "The image model flagged possible manipulation; the highlighted region is a localization prediction; report consistency was not assessed because comparable visual findings were unavailable." This is a proposed message, not an output observed from a deployed system. No explanation should imply a verified diagnosis or replace qualified review.

VII. AI/ML EVALUATION METHODOLOGY

A. Classification metrics

For evaluation, a manipulated image is the positive class. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives. The proposed metrics include

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN), \quad (2)$$

$$\text{Precision} = TP / (TP + FP), \quad (3)$$

$$\text{Recall} = TP / (TP + FN), \quad (4)$$

$$F1 = 2TP / (2TP + FP + FN). \quad (5)$$

The evaluation should state its convention for undefined metrics when a denominator is zero. Accuracy should be accompanied by the confusion matrix, class counts, and selected threshold. Precision–recall and receiver-operating-characteristic curves can describe behavior across thresholds, while calibration should be assessed before presenting scores as probabilities.

B. Localization and language metrics

For a predicted region B_{pred} and a ground-truth region B_{true}, overlap is measured by

$$IoU = |B_{pred} \cap B_{true}| / |B_{pred} \cup B_{true}|. \quad (6)$$

The protocol should report average precision at stated overlap thresholds, sensitivity to manipulated regions, and false-positive regions per authentic image. A bounding-box visualization alone is insufficient to calculate these quantities. Small alterations and images with multiple edits should be reported separately where the sample size permits. The NLP component requires concept-level extraction metrics and an assessment of assertion handling. Consistency checking requires a separately labeled set of matched and mismatched image–report pairs. Evaluation should report both decision quality on assessed cases and coverage, defined as the fraction of eligible requests receiving an assessed comparison. Excluding not-assessed cases without reporting coverage would hide a system limitation.

C. Comparison and ablation design

The proposed comparison includes a CNN-only baseline, classification with localization, and the complete workflow with report analysis. Synthetic augmentation and optional ELA should each be accessed through controlled ablations. Comparisons should use the same patient-level partitions and report variation across repeated runs. The complete workflow should not be assumed to improve scan-level classification merely because it displays more outputs. If scores are later combined into a single classifier, the fusion rule must be specified and fitted using training or validation data only. Otherwise, evaluation should keep forensic and consistency decisions separate and measure the usefulness of their joint presentation.

IJIRIS: International Journal of Innovative Research in Information Security
Volume 12, Issue 07, October 2026
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2026.v1207.02

VIII. SECURITY AND INTEGRITY CONSIDERATIONS

The detector is intended as one component of integrity review, not as a mechanism that prevents all unauthorized edits. The proposed threat model includes modification before upload, replacement of stored files, report mispairing, jointly altered images and reports, and misuse of outputs. Table III summarizes these cases and the limits of the proposed response.

Table III - Proposed Threat Model and Limits of the Verification Workflow

Threat or failure	Proposed response	Residual limitation / evidence status
Alteration before upload	Apply classification and suspicious-region localization.	Unseen manipulations may evade detection; no measured attack coverage is supplied.
Incorrect report pairing	Validate study association before semantic comparison.	Pairing logic and study identifiers require implementation and testing.
Joint image/report alteration	Preserve forensic results independently of textual agreement.	Consistency alone cannot establish that either input is genuine.
Stored-file replacement	Preserve originals, record digests, and maintain an access-controlled audit trail.	These are proposed controls, not demonstrated protection mechanisms.
Sensitive data exposure	Restrict access, minimize retained content, and define deletion and transport controls.	No deployment configuration or security assessment is supplied.
Misleading explanation	Ground explanations in available evidence and expose missing or uncertain outputs.	Human review remains necessary; no user study establishes explanation reliability.

The source report includes a comparison described as protected and unprotected CT scans. It does not specify an implemented encryption, signature, watermarking, or access-control mechanism for that comparison. Visual similarity between scans cannot substantiate a protection claim. Secure storage, access control, and provenance recording are therefore treated as requirements for future implementation rather than measured outcomes. Synthetic examples should be segregated from authentic clinical data. Testing should use appropriately authorized data and avoid exposing identifying information through logs or the conversational interface. Future evaluation involving clinical workflows requires the relevant institutional review and data-governance arrangements; none is claimed for the current evidence.

IX. EXPERIMENTAL SETUP

The supplied report recommends a computer with at least 8 GB RAM and GPU access for training, but does not identify the machine used for an experiment. It provides no dataset size, training duration, number of epochs, optimizer settings, or test-set composition. These omissions prevent reconstruction of a completed experimental setup. The proposed evaluation begins by freezing a data manifest and patient-level partitions. The manifest should enumerate modalities, sources, report availability, genuine abnormalities, and manipulation categories. Training and validation data are used to fit models and select thresholds; the test partition remains reserved for final assessment. Original images and all derived versions stay in one partition. The experiment record should freeze model variants, library versions, checkpoints, random seeds, preprocessing, and hardware. Classification and localization are evaluated first, followed by text extraction, consistency comparison, and the interface. Domain-separated testing should examine unseen manipulation sources and, where data permit, acquisition sites or scanner settings not represented in training. Latency measurements should distinguish loading, preprocessing, model execution, text processing, and rendering. Batch size, input resolution, device, and warm-up policy should accompany timing results. Because no such measurements were supplied, this paper does not describe the system as real-time. Table IV lists the evidence needed for the planned experiments.

Table IV - Proposed Experiments and Required Measurements

Experiment	Required evidence and output
Classification	Labeled test manifest; confusion matrix, F1, sensitivity, specificity, and threshold.
Localization	Ground-truth boxes; stated IoU thresholds, average precision, and false-positive regions.
Report extraction	Expert-labeled text; concept and assertion-level precision, recall, and F1.
Consistency checking	Verified pairs and mismatch labels; decision metrics plus assessment coverage.
Robustness / ablation	Fixed splits; unseen manipulations and matched comparisons with modules removed.
Runtime / interface	Hardware and request logs; latency distribution, failures, and explanation fidelity.

X. RESULTS

A. Qualitative material in the project report

The source report presents four categories of illustrations: original and tampered CT scans, protected and unprotected CT scans, a forged chest X-ray, and a manipulated brain MRI. Their descriptions discuss highlighted suspicious regions and the intended value of visual localization. Table V records what can be concluded without converting illustrations into numerical benchmarks. The examples are not accompanied by a traceable set of image identifiers, trained model outputs, ground-truth annotations, or experimental logs. Their image provenance is also not established in the supplied text. They are therefore summarized rather than reproduced as independently verified detections. The architecture diagram in this paper is a newly drawn schematic, not an experimental result.

Table V- Qualitative Observations from the Source Report and their Evidence Limits

Reported example	Observation described in the report	Interpretation supported by the supplied evidence
Original / tampered CT	Scans compared with regions described as manipulated and highlighted.	Illustrates the intended display; no case-level predictions or ground truth support accuracy or false-positive estimates.
Protected / unprotected CT	Comparison described as preserving appearance while protecting content.	A presentation concept only; no implemented protection method or security test is documented.
Forged chest X-ray	Example described as containing an implanted lesion and a highlighted region.	An additional application not validated X-ray generalization from a CT/MRI model.
Manipulated brain MRI	Example described as showing an artificial tumor region and its localization.	Illustrates the intended MRI use case; no blinded test, overlap metric, or independent annotation is supplied.

Apparently unaltered examples cannot establish a zero false-positive rate, and a highlighted lesion cannot establish that the detector distinguishes a forged lesion from genuine disease. This paper makes no claim of superiority over conventional verification methods or published deepfake detectors. Those comparisons remain experiments to be performed.

B. Quantitative result status

No numerical accuracy, precision, recall, F1, localization average precision, NLP score, or latency measurement is available in the supplied material. There is no test-set denominator from which to derive such values. The report's statements about high accuracy and successful implementation are insufficient to establish those outcomes quantitatively.

C. Functional evaluation status

The available evidence supports an architectural description and illustrative inference paths. End-to-end behavior remains to be verified. Table VI defines functional tests for the implementation; all entries are proposed tests, not recorded passes.

Table VI- Proposed Functional Tests; Execution Evidence is not Supplied

Test case	Expected behavior
Valid scan, no report	Run supported image branches; mark consistency not assessed.
Unreadable input	Reject with a clear error; do not return an authenticity label.
No detected regions	Return an empty list without claiming proven authenticity.
Multiple regions	Display every retained box in original-image coordinates.
Unmatched report	Disable semantic comparison and identify the pairing problem.
Negated / uncertain finding	Preserve assertion status rather than matching keywords alone.
Missing model / failure	Identify the unavailable component and retain valid partial results.

XI. DISCUSSION

The principal value of the architecture is that it makes different forms of evidence visible within one workflow. A scan-level score can support initial review, a region prediction can direct attention, and structured report analysis can expose a disagreement. None of these outputs should silently substitute for another. Their usefulness depends on communicating what was evaluated and what remains uncertain. The most consequential design issue is the boundary between forensic detection and clinical interpretation. A network trained to identify altered pixels does not automatically learn a reliable disease vocabulary. Report checking therefore needs either a separately validated clinical-finding model or expert-provided findings. Making this dependency explicit prevents coherent-looking explanations from exceeding the available evidence. Generalization is another concern. A classifier may respond to preprocessing, export settings, or synthetic-generation characteristics rather than manipulation itself. The proposed ablations and source-separated evaluations are intended to test these alternatives. Results should be interpreted separately for CT, MRI, and additional modalities rather than pooled into an unsupported claim of universal verification. Finally, a practical workflow must balance false alarms with missed manipulations. An operating threshold cannot be chosen from a few illustrations. It should be established using validation data and reviewed against the intended use case. This paper supplies neither a clinically approved operating point nor evidence that the interface reduces diagnostic errors.

XII. LIMITATIONS

The first limitation is evidential: the available artifact is a project report, not a reproducible experimental package. Data manifests, executable training code, checkpoints, ground-truth annotations, and run logs were not supplied. The manuscript therefore describes a proposed framework with qualitative supporting material rather than a validated detector. Second, several choices remain open, including the CNN backbone, GAN configuration, YOLO variant, NLP method, and software stack. Two-dimensional inference examples do not establish complete support for volumetric studies. Optional ELA also lacks a format-specific evaluation. Third, report consistency lacks a demonstrated source of clinical visual findings and a labeled multimodal evaluation set. The conditional comparison specified here is a design refinement requiring implementation. A chatbot explanation cannot compensate for this missing evidence. Fourth, the illustrations do not establish performance across scanners, institutions, patient groups, or unseen manipulation methods. No clinician study, deployment evaluation, or formal security assessment is documented. These limitations preclude claims of clinical readiness, universal protection, or real-time reliability.

XIII. FUTURE WORK

The immediate priority is an authorized, versioned dataset with patient-level partitions, authentic abnormal scans, manipulation annotations, and verified report associations. The next step is a reproducible training and inference configuration, subject to data and institutional restrictions. A CNN-only baseline should precede evaluation of synthetic augmentation and localization. Subsequent work should establish a structured visual-finding source for report comparison and assess extraction, negation, uncertainty, and mismatch detection independently. Full-study CT and MRI support should be evaluated separately from two-dimensional demonstrations. Additional modalities, including chest X-rays, require their own training and test evidence. Future studies can then assess uncertainty handling, domain shift, efficiency, and explanation fidelity. Storage and audit mechanisms should be tested independently of the detector. Evaluation with qualified reviewers should examine whether the interface helps interpret evidence without encouraging overreliance on automated labels.

XIV. CONCLUSION

This paper presents Medical Deepfake Detector as an integrated framework for medical-image integrity review using GAN-assisted training, CNN classification, YOLO localization, and NLP-based report analysis. The architecture separates offline generation from inference and distinguishes forensic outputs from clinical findings. A chatbot-oriented interface is proposed to present scores, suspicious regions, textual evidence, and unresolved cases for human review. The project report provides design descriptions, illustrations, and inference snippets, but not the artifacts needed to verify quantitative performance. Accordingly, this manuscript reports qualitative evidence and specifies the experiments required to evaluate the framework. Its outcome is a structured, testable research design rather than a clinically validated system. Reproducible datasets, trained models, independent evaluation, and bounded explanations are the next requirements for establishing practical effectiveness.

ACKNOWLEDGMENT

The authors thank Dr.Aruna Reddy H, Associate Professor and Head, Department of Artificial Intelligence and Machine Learning (AI&ML), Vemana Institute of Technology, for her guidance and support during the project. The authors also acknowledge the department and institution for supporting the project work.

AUTHOR CONTRIBUTION STATEMENT

Conceptualization: Dr.Aruna Reddy H
Literature Review and Methodology design: Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Software: Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Validation: Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Formal Analysis: Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Investigation: Dr.Aruna Reddy H
Resources: Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Data Curation: Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Writing original draft preparation: Sneha S, Sonu Sri Samyuktha K, Supraja B N, Vaishnavi S
Writing review and Editing: Dr.Aruna Reddy H
Visualization: Dr.Aruna Reddy H
Supervision: Dr.Aruna Reddy H
Conflict of interest: The authors declare no conflicts of interest.
Data availability statement: Data supporting these findings are available within the article, at https://doi.org/10.26562/ijiris.2026.v1207.02, or upon request.
Supplementary materials: The supplementary materials are available at https://doi.org/10.26562/ijiris.2026.v1207.02
Publisher's note: AM Publications, India. IJIRIS (International Journal of Innovative Research in Information Security) Journals stays neutral with regard to jurisdictional claims in published maps and institutional affiliations. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.
https://ampublications-india.com/

REFERENCES

1. Y.Mirsky, T.Mahler, I.Shelef, and Y.Elovici, "CT-GAN: Malicious tampering of 3D medical imagery using deep learning," in Proc. 28th USENIX Security Symp., 2019, pp. 461–478.
2. P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1053–1061. https://doi.org/10.1109/CVPR.2018.00116
3. Manimozhi,Shashanka,Rahul,Rithik,Nandish(2024). A Method of Skin Disease Detection Using Image Processing and Machine Learning. International Journal of Innovative Research in Information Security, Volume 10, Issue 04, Pages 123-128 https://doi.org/10.26562/ijiris.2024.v1004.01
4. Chandramouli,Pavani,Sai,Swathi,Yashosmitha(2024). A Deep Learning Based Soil Type Classification and Recommendation of Crops to the Soil Type. International Journal of Innovative Research in Information Security, Volume 10, Issue 04, Pages 129-132 https://doi.org/10.26562/ijiris.2024.v1004.02
5. I.J.Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, 2014.

6. K.He, X.Zhang, S.Ren, and J.Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
7. G.Huang, Z.Liu, L.van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708. https://doi.org/10.1109/CVPR.2017.243
8. J.Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91
9. J.Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, vol. 1, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
10. Shammi,Poorvika,Pengonda,Rajeshwari,sharmila(2024). Advancing Privacy and Confidentiality in the Digital Healthcare Era: The Image Encryption Solution. International Journal of Innovative Research in Information Security, Volume 10, Issue 04, Pages 133-138 https://doi.org/10.26562/ijiris.2024.v1004.03
11. Smit, S.Jain, P.Rajpurkar, A. Pareek, A. Ng, and M. Lungren, "Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1500–1519. https://doi.org/10.18653/v1/2020.emnlp-main.117
12. Heena,Adarsh,Mohnish,Mukeshreddy,Shaikh(2024).Boosting Network Security: A comparative Analysis of Deep Learning Techniques for Intrusion Detection. International Journal of Innovative Research in Information Security, Volume 10, Issue 04, Pages 139-144 https://doi.org/10.26562/ijiris.2024.v1004.04
13. Jagadevi,Bhavana,Divya,Lavanya,Manasa(2024).Gupta-Yantra: Spying Robot for Defence. International Journal of Innovative Research in Information Security, Volume 10, Issue 04, Pages 145-149 https://doi.org/10.26562/ijiris.2024.v1004.05
14. Kesavan,Debasinha,Benishaw,Ananyagi,Ankitranar(2024). Predicting Indian GDP with Machine Learning: A Comparison of Regression Models. International Journal of Innovative Research in Information Security, Volume 10, Issue 04, Pages 150-155 https://doi.org/10.26562/ijiris.2024.v1004.06