(10) Publication No: **IN202641117426 A1**
(43) Publication Date: 09-10-2026
(22) Date of Filing: 30-09-2026
Journal No: 41/2026

(19) Intellectual Property Office, India

(12) INDIAN PATENT APPLICATION

(54) Title: **"a forensic image analysis system for detection, localization and source attribution of artificial intelligence-generated or artificial intelligence-manipulated content in a digital image"**

(51) International Classification:
G06V 10/764, G06N 3/08, G06K 9/62, G06V 10/774 G06N 3/04

(21) Application No: **202641117426**

(31) Priority Document No: --

(32) Priority Date: --

(86) International Application No: --
Filing Date: --

(87) International Publication No: --

(61) Patent of Addition to Application No: --
Filing Date: --

(62) Divisional to Application No: --
Filing Date: --

(71) Name of Applicant(s):
Vellore Institute Of Technology,

(72) Name of Inventor(s):
1. Dr. E.s. Madhan
2. Mr. Rahul Yadav
3. Mr. Aashish Kumar Mahato
4. Mr. Bibek Gami

(57) Abstract:

The present invention relates to a forensic image analysis system and method for detecting, localizing and attributing artificial intelligence-generated or artificial intelligence-manipulated content. The system employs a microservice-based orchestration architecture configured to dispatch image data to a plurality of heterogeneous detector modules and receive respective detection scores, confidence information, operational-status information and processing metadata. A dynamic reliability-weighted fusion engine adjusts contributions of the detector modules according to runtime reliability, while a disagreement-aware calibration engine reduces confidence when detector outputs conflict. The system further performs frequency-domain and spatial-domain analysis and dynamically combines corresponding anomaly representations based on inter-domain consistency. An explainability module generates region-specific artifact maps and structured forensic explanations. An uncertainty assessment module separately characterizes data-associated and model-associated uncertainty. A generator attribution module compares composite spectral, chromatic, texture and noise fingerprints with stored generator signatures. Candidate artifact regions are independently verified using localized frequency and spatial analysis, and a forensic output comprising a calibrated verdict, trust information, artifact localization, uncertainty assessment and source-attribution information is generated. Fig 1 to 10.

FORM 2

THE PATENTS ACT, 1970

(39 of 1970)

&

THE PATENTS RULES, 2003

COMPLETE SPECIFICATION

(See Section 10; rule 13)

Title of the Invention

"A forensic image analysis system for detection, localization and source attribution of artificial intelligence-generated or artificial intelligence-manipulated content in a digital image"

APPLICANTS:

Name : **VELLORE INSTITUTE OF TECHNOLOGY**

Nationality : Indian

Address : Vellore Institute of Technology, Tiruvalam Road, Katpadi, Vellore, Tamil Nadu, India -632014.

The following specification particularly describes the invention and the manner in which it is performed.

TECHNICAL FIELD

[0001] The present invention generally relates to the field of digital image forensics, multimedia authentication and detection of artificial intelligence-generated or manipulated visual content. More particularly, the invention relates to a modular forensic image analysis system employing multiple heterogeneous image detectors coordinated through a service-based orchestration architecture.

[0002] The invention particularly pertains to frequency-domain and spatial-domain image analysis, dynamic reliability-based fusion of detector outputs, disagreement-aware confidence calibration, explainable artifact localization, uncertainty quantification, generator fingerprint identification and region-specific verification of suspected image artifacts. The invention further relates to examination of content credentials, invisible watermarks and image metadata for determining provenance information associated with a digital image.

BACKGROUND ART

[0003] Background description includes information that may be useful in understanding the present invention. It is not an admission that any of the information provided herein is prior art or relevant to the presently claimed invention, or that any publication specifically or implicitly referenced is prior art.

[0004] The rapid development of generative artificial intelligence has enabled the creation and manipulation of highly realistic digital images using generative adversarial networks, diffusion models, vision-language models and other image-synthesis architectures. Modern image generators are capable of reproducing realistic textures, lighting, facial structures, backgrounds and object relationships, thereby making artificially generated images increasingly difficult to distinguish from images captured using conventional cameras.

[0005] Artificial manipulation may also be restricted to a selected portion of an image. For example, an authentic photograph may include an artificially generated face, an inpainted object, a synthetically extended background or another locally modified region. In such cases, the authentic portion of the image may conceal or dilute the characteristics of the manipulated region when the

complete image is processed using only a global image classifier.

[0006] The increasing availability and quality of synthetic-image-generation tools have created challenges for digital-media authentication, forensic investigation, journalism, legal proceedings, cybersecurity, social-media moderation and identity verification. Reliable forensic examination therefore requires more than a binary determination of whether an image is artificial or authentic. It may additionally require localization of suspicious regions, assessment of the reliability of the detection, identification of contradictory detector outputs, differentiation between image-related and detector-related uncertainty, and attribution of the image to a possible generative source.

[0007] Conventional artificial-image detection systems commonly employ a single classification model trained to distinguish artificially generated images from authentic images. Such systems may perform satisfactorily on images produced by generators represented within their training data but may exhibit reduced accuracy when presented with content generated by a different or previously unseen image-generation architecture. A detector trained substantially on generative adversarial network outputs may therefore fail to reliably identify images produced using diffusion-based generators.

[0008] One example identified in the prior-art review is **US 10,803,391 B2**, relating to deepfake detection using deep neural networks. According to the prior-art assessment contained in the disclosure, the reference principally employs a single-model binary classification arrangement and does not provide multimodal detector fusion, frequency-domain analysis, uncertainty quantification or generator attribution. The presently disclosed system differs by coordinating multiple heterogeneous detectors and combining their outputs with frequency-spatial analysis, runtime reliability assessment, disagreement calibration, uncertainty decomposition and source-fingerprint matching.

[0009] Single-model detection systems also generally produce a binary output or an artificial-image probability without indicating the operational reliability of the detector. A numerical score may appear highly confident even where the input

image is blurred, compressed, cropped, reprocessed or materially different from the data used for training the detector. Such systems may therefore fail to distinguish a stable prediction from an unreliable prediction caused by insufficient image evidence or limited model knowledge.

[0010] Certain existing arrangements employ multiple detectors to improve classification performance. However, outputs from the detectors are commonly combined using predetermined or static weights. Each detector may therefore retain the same influence irrespective of whether the detector operated normally, used a fallback model, encountered an error or returned a delayed response.

[0011] For example, **US 11,132,529 B2**, identified in the disclosure as relating to detection of manipulated media content, is distinguished in the prior-art review as providing binary detection without explainable spatial artifact localization and employing static-weight fusion without runtime reliability adjustment. The presently disclosed system instead determines a runtime reliability for each detector based on detector confidence, operational status and processing behaviour and dynamically reduces the contribution of a detector operating in a fallback or error state. The system further generates localized artifact information and structured forensic explanations.

[0012] Static fusion also fails to adequately represent disagreement between individual detector outputs. A first detector may indicate a high likelihood of artificial generation, while a second detector may strongly indicate authenticity. A simple weighted average may produce an apparently moderate-confidence score without indicating that the underlying detectors are in direct conflict.

[0013] A combined score generated from mutually consistent detector outputs should not be treated as having the same reliability as an identical numerical score generated from contradictory outputs. There therefore exists a need for disagreement-aware calibration capable of measuring inter-detector dispersion, identifying conflicting outputs and reducing the apparent confidence of a combined result when the detectors materially disagree.

[0014] Known image-analysis arrangements may further operate in only one image domain. Spatial-domain detectors evaluate pixel-level characteristics including texture, edges, object structure, anatomical geometry, lighting and semantic relationships. Such detectors may identify visible inconsistencies but fail to detect hidden frequency patterns created by image synthesis, upsampling or abnormal noise generation.

[0015] Frequency-domain detectors examine transform coefficients, spectral-energy distributions, periodic structures and compression-related artifacts. While such detectors may identify characteristics not visible in the spatial domain, frequency-only analysis may fail to identify anatomical abnormalities, structural distortions, inconsistent lighting, unnatural boundaries and localized semantic defects.

[0016] The prior-art review identifies **US 2022/0198818 A1** as relating to artificial-image detection using frequency analysis. According to the disclosure, the reference is limited to frequency-domain analysis and does not provide spatial-domain fusion, cross-domain attention, detector-disagreement handling or generator fingerprint identification. The presently disclosed system differs by processing the image through frequency and spatial branches and dynamically combining the corresponding anomaly representations according to their relative strengths and inter-domain consistency.

[0017] Simple concatenation or averaging of spatial and frequency features may also be inadequate because such processing does not determine whether the two domains confirm or contradict one another. A strong anomaly from one domain may dominate a detection even where the anomaly results from ordinary compression, resizing or natural texture characteristics. There is therefore a need for consistency-gated fusion capable of modifying the relative contribution of each domain based on cross-domain agreement.

[0018] Another limitation of known artificial-image detectors is the absence of meaningful explainability. Many detectors provide only a numerical score or binary verdict without identifying the image region responsible for the result.

Certain explanation techniques generate attention maps or heatmaps, but such visual outputs may not indicate the category of the detected artifact or explain why the region influenced the final result.

[0019] A heatmap may indicate that a model responded to a facial region without explaining whether the response resulted from anatomical inconsistency, texture smoothness, lighting mismatch, frequency irregularity or boundary distortion. There therefore exists a need for an artifact-localization mechanism capable of converting region activations and attribution information into structured forensic explanations identifying both the location and nature of a suspected artifact.

[0020] For example, **US 2023/0114584 A1**, identified in the disclosure as relating to synthetic-image detection using vision transformers, is described as employing a single vision-transformer classifier without multi-detector fusion, disagreement-aware calibration, region-level frequency cross-verification or comprehensive forensic report generation. The presently disclosed system differs by subjecting candidate regions identified by a vision-based detector to independent localized frequency and spatial verification before retaining the regions in the final artifact map. The system further associates the verified regions with detector-trust, uncertainty and source-attribution information in a consolidated forensic output.

[0021] Global image classifiers are also generally unsuitable for detecting partial manipulation. Where a synthetic region occupies only a small portion of an otherwise authentic image, the authentic characteristics of the surrounding image may dominate the global classification result. Accordingly, a manipulated region may remain undetected.

[0022] There is therefore a need for differential regional analysis in which the image is partitioned into multiple regions, a dominant image baseline is determined, and individual regions are compared with the baseline to identify localized divergence in noise, texture, edges and frequency characteristics. Such an arrangement can distinguish a uniformly synthetic image from a substantially

authentic image containing a locally manipulated portion.

[0023] A further technical problem concerns false-positive localization. A vision-based detector may identify a suspicious region because of an unusual but authentic object, compression artifact, photographic effect or lighting condition. Accepting every visually identified region without independent verification may therefore reduce forensic reliability.

[0024] Independent region-level verification is required to confirm whether a candidate region also exhibits signal-level characteristics associated with artificial generation or manipulation. Such verification may examine localized entropy, frequency-energy distribution, noise response, edge consistency and transition behaviour before retaining or rejecting the candidate region.

[0025] Conventional detection systems also commonly generate a single confidence value without distinguishing between uncertainty arising from the image and uncertainty arising from the detection models. Image-related uncertainty may result from blur, noise, low resolution, cropping or severe compression. Detector-related uncertainty may result from inconsistent model outputs, limited training coverage or an unseen image-generation technique.

[0026] These uncertainty sources require different responses. Image-related uncertainty may require acquisition or enhancement of better-quality evidence, whereas detector-related uncertainty may require additional detectors, human review or model retraining. There is therefore a need for decomposed uncertainty assessment capable of separately characterizing data-associated uncertainty and model-associated uncertainty.

[0027] The prior-art review further identifies **WO 2023/034521 A1** as relating to content-authenticity verification using metadata or watermark information. According to the disclosure, the reference does not provide pixel-level forensic analysis, dynamic reliability adjustment or decomposed uncertainty assessment. Although metadata and watermark indicators can provide useful provenance information, such information may be absent or removed through recompression,

screenshot capture, editing or social-media processing. The presently disclosed system therefore employs metadata and watermark examination as one component of a broader cascade that also includes pixel-level, frequency-level and region-level forensic analysis.

[0028] Known systems may further determine only whether an image is synthetic without identifying the likely generator responsible for creating the image. Generator attribution may be useful in forensic investigation, misinformation analysis, intellectual-property disputes and tracing misuse of a particular image-generation service.

[0029] Certain attribution techniques rely primarily on residual-noise fingerprints. However, residual-noise signatures may be weakened or destroyed by compression, scaling, cropping, screenshot capture or subsequent image editing. A more robust attribution arrangement therefore requires a composite source fingerprint combining multiple independent image characteristics.

[0030] There accordingly exists a need for a generator-attribution mechanism capable of combining spectral-energy characteristics, chromatic distributions, texture patterns and residual-noise information into a composite source representation. The composite representation may be compared with stored reference fingerprints associated with known generators and with a real-image baseline, while dynamically adjusting the relative contribution of individual signal components when compression or other image degradation is detected.

[0031] Existing detection systems also lack an integrated operational architecture for coordinating multiple independent detection, fusion, explanation, uncertainty, attribution and verification services. A monolithic architecture may be difficult to update, scale or maintain, and failure of a single component may interrupt the complete forensic analysis.

[0032] A modular service architecture permits individual detectors to be independently deployed, updated or replaced. However, merely separating detectors into services does not solve problems relating to service availability,

fallback operation, delayed responses, normalization of heterogeneous outputs or failure isolation. A reliable forensic architecture must therefore evaluate the runtime state of each detector and adapt the fusion process according to the actual condition of the detector during analysis of a particular image.

[0033] The present invention is directed toward overcoming one or more of the foregoing limitations by providing a modular forensic image analysis system that coordinates heterogeneous detectors and dynamically modifies their contributions according to runtime confidence, operational status and processing behaviour. The system measures detector disagreement and calibrates the combined result towards a neutral state when detector outputs materially conflict.

[0034] The system further performs coordinated frequency-domain and spatial-domain analysis, generates localized and human-interpretable artifact information, separates image-associated uncertainty from detector-associated uncertainty, constructs a composite source fingerprint and independently verifies candidate artifact regions using localized signal analysis.

[0035] Accordingly, the disclosed system produces a forensic result supported by multiple independent forms of evidence rather than by a single classification model or a single image-analysis domain. The resulting forensic output may include a calibrated verdict, detector-trust information, verified artifact regions, uncertainty information, provenance indicators and likely generator-attribution information.

SUMMARY

[0036] The present invention provides a digital forensic image analysis system for detecting, localizing, verifying and attributing artificial intelligence-generated or manipulated image content. The system uses a modular microservice architecture in which multiple heterogeneous detectors analyse the image in parallel and provide detection scores, confidence information, operational status and processing metadata.

[0037] A dynamic reliability-weighted fusion module adjusts the contribution

of each detector according to its runtime confidence, service status and response time. A disagreement-aware calibration module evaluates conflicts between detector outputs and reduces the confidence of the combined result when substantial disagreement is detected. The system further performs dual-domain frequency and spatial analysis and combines the resulting anomaly information according to inter-domain consistency.

[0038] The invention also includes explainable artifact localization, uncertainty assessment, generator fingerprint identification and region-level verification. Suspicious image regions are mapped and classified, uncertainty is separated into image-related and model-related components, and spectral, chromatic, texture and noise characteristics are compared with stored generator signatures. Candidate regions identified by a vision-based model are independently verified using localized frequency and spatial analysis.

[0039] The system generates a forensic output comprising a calibrated detection verdict, detector trust information, localized artifact findings, uncertainty information, provenance indicators and likely generator-attribution information. Accordingly, the invention provides an explainable, reliability-aware and multi-domain framework for forensic examination of artificial and manipulated visual content.

[0040] These and other features will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings. While the invention has been described and shown with reference to the preferred embodiment, it will be apparent that variations might be possible that would fall within the scope of the present invention.

BRIEF DESCRIPTION OF DRAWINGS

[0041] The accompanying illustrations are incorporated into and form a part of this specification in order to aid in comprehending the current disclosure. The pictures demonstrate exemplary implementations of the current disclosure and,

along with the description, assist to clarify its fundamental ideas.

[0042] Fig.1 shows the three-tier architecture of the forensic image analysis system (100) comprising a client interface (102), an orchestration layer (104) and backend analysis services (106).

[0043] Fig 2 shows the parallel dispatch of image data from the orchestration controller (104) to an image-classification detector (108), a multimodal visual-artifact detector (110), a frequency-domain detector (112) and a watermark and metadata cascade (114).

[0044] Fig 3 shows a flow diagram of the dynamic reliability-weighted fusion engine (116) determining a runtime reliability of each detector module from confidence, operational status and processing-time information and generating a fused detection score (146).

[0045] Fig 4 shows a flow diagram of the disagreement-aware confidence calibration engine (118) determining inter-detector disagreement, identifying conflicting detector outputs and calibrating the fused detection score towards a neutral detection state.

[0046] Fig 5 shows the dual-domain image analysis engine (120) comprising a frequency-domain analysis branch (122), a spatial-domain analysis branch (124) and a consistency-gated fusion module (126).

[0047] Fig 6 shows the interconnected forensic processing pipeline in which the orchestration controller (104) controls the dynamic reliability-weighted fusion engine (116), the disagreement-aware confidence calibration engine (118), the dual-domain image analysis engine (120) and the region verification engine (134) and a forensic output generator (136) generates a forensic result (150).

[0048] Fig 7 shows the artifact localization engine (128) partitioning the digital image into a plurality of image regions and generating an artifact map (198).

[0049] Fig 8 shows the uncertainty assessment engine (130) separately characterizing data-associated uncertainty (202) and model-associated uncertainty (204) and assigning a reliability category (206).

[0050] Fig 9 shows the generator attribution engine (132) combining a

spectral signature (210), a chromatic profile (212), a texture profile (214) and a residual-noise profile (216) into a composite source fingerprint (218) for comparison with stored reference fingerprints (220).

[0051] Fig 10 shows a flow diagram of the region verification engine (134) independently analysing frequency and spatial characteristics of a candidate artifact region (226) identified by a vision-based detection model (224) to confirm or reject the candidate artifact region.

[0052] It should be noted that the figures are not drawn to scale, and the elements of similar structure and functions are generally represented by like reference numerals for illustrative purposes throughout the figures. It should be noted that the figures do not illustrate every aspect of the described embodiment sand do not limit the scope of the present disclosure.

[0053] Other objects, advantages, and novel features of the invention will become apparent from the following detailed description of the present embodiment when taken in conjunction with the accompanying drawings.

DETAILED DESCRIPTION OF THE INVENTION

[0054] While the present invention is described herein by way of example using embodiments and illustrative drawings, those skilled in the art will recognize that the invention is not limited to the embodiments of drawing or drawings described and are not intended to represent the scale of the various components. Further, some components that may form a part of the invention may not be illustrated in certain figures, for ease of illustration, and such omissions do not limit the embodiments outlined in any way. It should be understood that the drawings and the detailed description thereto are not intended to limit the invention to the particular form disclosed, but on the contrary, the invention is to cover all modifications, equivalents, and alternatives falling within the scope of the present invention as defined by the appended claim.

[0055] As used throughout this description, the word "may" is used in a permissive sense (i.e. meaning having the potential to), rather than the mandatory sense, (i.e. meaning must). Further, the words "a" or "an" mean "at least one" and the word "plurality" means "one or more" unless otherwise mentioned.

Furthermore, the terminology and phraseology used herein are solely used for descriptive purposes and should not be construed as limiting in scope. Language such as "including," "comprising," "having," "containing," or "involving," and variations thereof, is intended to be broad and encompass the subject matter listed thereafter, equivalents, and additional subject matter not recited, and is not intended to exclude other additives, components, integers, or steps. Likewise, the term "comprising" is considered synonymous with the terms "including" or "containing" for applicable legal purposes. Any discussion of documents acts, materials, devices, articles, and the like are included in the specification solely for the purpose of providing a context for the present invention. It is not suggested or represented that any or all these matters form part of the prior art base or were common general knowledge in the field relevant to the present invention.

[0056] In this disclosure, whenever a composition or an element or a group of elements is preceded with the transitional phrase “comprising”, it is understood that we also contemplate the same composition, element, or group of elements with transitional phrases “consisting of”, “consisting”, “selected from the group of consisting of, “including”, or “is” preceding the recitation of the composition, element or group of elements and vice versa.

[0057] The present invention is described hereinafter by various embodiments with reference to the accompanying drawing, wherein reference numerals used in the accompanying drawing correspond to the like elements throughout the description. This invention may, however, be embodied in many different forms and should not be construed as limited to the embodiment set forth herein. Rather, the embodiment is provided so that this disclosure will be thorough and complete and will fully convey the scope of the invention to those skilled in the art. In the following detailed description, numeric values and ranges are provided for various aspects of the implementations described. These values and ranges are to be treated as examples only and are not intended to limit the scope of the claims. In addition, several materials are identified as suitable for various facets of the implementations.

[0058] The present invention relates to a digital forensic image analysis

system, hereinafter referred to as “OriginLayer,” for detecting, localizing, verifying and attributing artificial intelligence-generated or artificial intelligence-manipulated content in a digital image. The system analyses the digital image through multiple independent detection paths, evaluates the reliability and agreement of the respective detection paths, identifies suspicious image regions, estimates uncertainty associated with the resulting detection, and generates an explainable forensic output.

[0059] The invention addresses limitations associated with conventional image-authenticity detectors that rely upon a single classification model or a single type of image feature. A detector trained using outputs of a particular generative model may fail to identify content produced by a different generative architecture. Similarly, analysis limited to pixel-level characteristics may fail to detect frequency-domain artifacts, while frequency-only analysis may fail to identify anatomical, semantic, textural or structural inconsistencies visible in the spatial domain. The present invention therefore combines heterogeneous detector outputs with frequency-domain analysis, spatial-domain analysis, source fingerprinting, artifact localization, uncertainty assessment and region-level verification.

[0060] The system is arranged as a modular service architecture having decoupled image-receiving, orchestration, detection, fusion, verification and reporting functions. This arrangement permits individual detection services to be independently deployed, updated, replaced or scaled without requiring replacement of the entire forensic analysis architecture. It further permits continued generation of a forensic output when one or more detector services become unavailable, operate using a fallback mechanism or return an error.

[0061] In an embodiment, the system comprises a client interface, an orchestration layer and one or more backend analysis services. The client interface receives a digital image and presents the resulting forensic analysis. The orchestration layer validates the received image, coordinates parallel and sequential analysis operations, exchanges data with the analysis services and

assembles a combined forensic result. The backend analysis services host one or more image classifiers, vision-language models, frequency analysers, vision-prompt-tuned models, fingerprint matching modules and localized verification modules.

[0062] The three-tier arrangement and associated processing pipeline are illustrated in the disclosure, in which detector operations are initially dispatched in parallel and the resulting outputs are subsequently processed through fusion, calibration, explainability, uncertainty, attribution and verification stages.

[0063] Image Reception and Initial Processing

[0064] When a digital image is submitted, an image input interface receives image data corresponding to the digital image. The image may be provided in a compressed or uncompressed image format, including JPEG, PNG or WebP. The system validates the image format, image size, file integrity and availability of sufficient image data for further analysis.

[0065] In an exemplary embodiment, an upper file-size limit of 10 MB is applied. However, the limit may be increased or decreased according to available processing resources, network bandwidth and intended deployment conditions. The file-size limit is therefore not intended to restrict the scope of the invention.

[0066] The received image may be encoded into a transport-compatible data representation for communication between the orchestration layer and the analysis services. In one implementation, the image is converted into Base64-encoded image data. In another implementation, binary image data, an image object, a storage reference or a secure resource locator may be exchanged between the respective services.

[0067] A content identifier is generated from the image data. The content identifier may comprise a cryptographic hash, such as a SHA-256 hash, generated from the entire image or a normalized representation thereof. The identifier enables the system to determine whether the same image has previously been analysed.

[0068] The content identifier is compared with identifiers stored in a cache. When a valid cached forensic result corresponding to the content identifier is available, the stored result may be returned without repeating the complete forensic analysis. When a result is unavailable or has exceeded an expiry period, the image is transmitted to the detector services for analysis. The cache may be maintained as a least-recently-used cache, distributed cache, database-backed cache or another storage arrangement.

[0069] **Parallel Detection Stage**

[0070] The orchestration layer dispatches the image data to a plurality of heterogeneous detector modules. The detector modules analyse fundamentally different characteristics of the image and generate respective detector scores. The detector modules may be executed asynchronously and substantially simultaneously to reduce overall analysis time.

[0071] Each detector output comprises at least a score representing an estimated likelihood that the image includes artificially generated or manipulated content. The detector output may further include a confidence value, an operational status, a processing duration, identified artifact information, region-specific findings and explanatory metadata.

[0072] In an embodiment, the plurality of detector modules comprises an image-classification detector, a multimodal visual-artifact detector and a frequency-domain detector. A watermark and metadata detection cascade may operate in parallel with the detector modules.

[0073] The image-classification detector processes the image through a trained visual classification model. The model may comprise a convolutional neural network, vision transformer, prompt-tuned visual model or another architecture trained to distinguish artificial images from camera-originated images. The classifier produces probability values associated with an artificial class and an authentic class.

[0074] In an exemplary implementation, a locally hosted vision-transformer

model is used as the primary classification model. When the primary model is unavailable, the orchestration layer may direct the image to one or more remotely hosted classification models. When both the primary and remotely hosted models are unavailable, a fallback analyser may examine image entropy, pattern diversity, data repetition and file-size-to-resolution characteristics.

[0075] The operational status of the image-classification detector is recorded as a successful status, fallback status or error status. Recording the operational status enables the subsequent fusion stage to distinguish a primary-model prediction from a less reliable fallback output.

[0076] The multimodal visual-artifact detector analyses both the image content and contextual relationships between objects represented in the image. The detector may examine anatomical anomalies, physical inconsistencies, text artifacts, texture irregularities, structural distortions, symmetry anomalies and boundary defects.

[0077] Anatomical analysis may identify additional fingers, missing body portions, inconsistent joints, impossible poses or disproportionate facial and body geometry. Physical consistency analysis may identify incorrect reflections, inconsistent shadow directions, floating objects or violations of expected object interaction. Text analysis may identify malformed characters, inconsistent fonts or semantically meaningless text.

[0078] Texture analysis may identify unnaturally smooth skin, repetitive texture patches or absence of expected surface details. Structural analysis may identify warped buildings, irregular backgrounds or distorted object boundaries. Symmetry analysis may identify facial or object symmetry that is inconsistent with naturally captured imagery. Boundary analysis may identify merging between hair and skin, fabric and background, or adjacent objects.

[0079] The multimodal detector may return an overall detector score together with a list of identified artifacts. Each identified artifact may include an artifact category, a description, a confidence value and region information. The detector

may further divide the image into a plurality of regions, including upper-left, upper-right, lower-left, lower-right and central regions, and generate region-specific findings.

[0080] The frequency-domain detector analyses spectral and statistical characteristics of the image. The detector may examine Shannon entropy, transitions between adjacent data values, periodic block-boundary patterns and high-frequency energy characteristics.

[0081] Entropy analysis determines the distribution and unpredictability of values within the image data. Artificially generated images may exhibit entropy characteristics different from those of naturally captured photographs. Transition analysis identifies smooth and sharp transitions between adjacent image-data values and determines whether the transition distribution exhibits excessive regularity.

[0082] Block-boundary analysis identifies periodic patterns associated with fixed-size transform blocks. Such patterns may indicate unusual image processing, recompression or synthesis. Spectral analysis determines the distribution of energy across frequency bands and identifies abnormal concentration or suppression of high-frequency components.

[0083] The frequency-domain detector may operate directly on encoded or decoded image data. In one embodiment, selected frequency characteristics are derived from byte-level data to provide a rapid preliminary analysis. In another embodiment, the image is decoded and transformed into frequency coefficients using a discrete cosine transform or Fourier transform.

[0084] **Watermark and Metadata Analysis**

[0085] The system may perform a cascade of provenance and metadata checks in parallel with the primary detectors. The cascade may include content-credential analysis, invisible-watermark analysis, exchangeable-image-file metadata analysis and extensible-metadata-platform analysis.

[0086] Content-credential analysis determines whether the image contains

embedded provenance information, including information conforming to a content-authenticity standard. Such information may identify an image source, capture device, editing event, signing authority or processing history.

[0087] Invisible-watermark analysis examines the image for watermarks embedded in spatial or frequency components. The watermark analysis may include detection of generator-associated watermarks, spread-spectrum marks, discrete-wavelet marks or combined wavelet-transform and cosine-transform watermarks.

[0088] Metadata analysis examines fields relating to a camera model, capture time, geographic location, image-editing software, processing history and other image properties. Extensible metadata may be examined for identifiers associated with known generative tools or editing applications.

[0089] The absence of a watermark or metadata indicator is not treated as conclusive evidence of authenticity because such information may be removed by recompression, screenshots, cropping, social-media processing or deliberate metadata stripping. The watermark and metadata results are therefore used as additional provenance signals rather than as the sole basis of the forensic verdict.

[0090] Dynamic Reliability-Weighted Fusion

[0091] Following completion of the initial detector operations, a dynamic reliability-weighted fusion module evaluates the runtime reliability of each detector. The module prevents a degraded, fallback or erroneous detector from retaining the same influence that it would have during successful primary-model operation.

[0092] For each detector, the runtime reliability is determined from a combination of self-reported confidence, operational status and processing-time information. A detector producing a high-confidence result through its intended primary model is assigned greater reliability. A detector operating through a fallback model is substantially penalized. A detector returning an error is assigned a minimal contribution.

[0093] Processing time may also be used as an indicator of operational health. A response obtained within an expected time period may retain its normal contribution, whereas a response approaching a timeout condition may be assigned reduced reliability. The weighting assigned to confidence, operational status and response time may be configured according to deployment requirements.

[0094] The system maintains a predetermined base weight for each detector. In an exemplary embodiment, a multimodal visual detector is assigned a greater base weight than a local classification detector, while the frequency analyser is assigned a supplementary base weight. The base weights are adjusted in accordance with the calculated runtime reliabilities.

[0095] The adjusted weights are normalized so that the combined contribution of the active detectors corresponds to a common total. The detector scores are thereafter combined using the normalized weights to generate a reliability-weighted fused score.

[0096] In one embodiment, the fusion module automatically selects between multiple fusion modes. Static fusion may be selected when all detectors operate successfully and exhibit high reliability. A single-detector mode may be selected when only one detector remains sufficiently reliable. Dynamic reliability-weighted fusion may be selected when two or more detectors are available but exhibit unequal runtime reliability.

[0097] This adaptive selection allows the system to operate under normal and degraded service conditions. The system therefore provides graceful degradation rather than terminating the entire forensic analysis when an individual detector fails.

[0098] **Disagreement-Aware Confidence Calibration**

[0099] A reliability-weighted score may still be misleading when the individual detector outputs strongly disagree. The system therefore applies disagreement-aware confidence calibration after reliability-weighted fusion.

[00100] The calibration module determines the dispersion of the detector scores relative to their weighted mean. It may calculate weighted variance, weighted standard deviation and a difference between the greatest and smallest detector scores. These characteristics are combined to determine a disagreement level.

[00101] The disagreement level may be classified as agreement, mild disagreement, moderate disagreement or severe disagreement. A conflict condition may be separately identified when a first detector strongly indicates artificial generation while another detector strongly indicates authenticity.

[00102] When the detectors substantially agree, the reliability-weighted fused score is retained with little or no modification. When disagreement increases, the fused score is progressively moved towards a neutral detection state. Moving the score towards the neutral state prevents the system from presenting an apparently confident verdict when the underlying detectors have produced contradictory evidence.

[00103] A trust score is generated from the detector agreement. A higher trust score represents substantial agreement between the detectors, while a lower trust score represents considerable dispersion or conflict.

[00104] The calibration module may generate an associated recommendation. A high-confidence recommendation may be generated when detector agreement is high. A moderate-confidence or low-confidence recommendation may be generated when disagreement exceeds corresponding thresholds. A human-review recommendation may be generated when a conflict is identified or when the trust score falls below a predetermined level.

[00105] The disclosure describes this calibration as one of seven interacting modules, together with dynamic reliability fusion, dual-domain analysis, explainability, uncertainty decomposition, source attribution and localized manipulation detection.

[00106] Frequency-Spatial Fusion Network

[00107] The system further comprises a frequency-spatial fusion network configured to analyse the image through a frequency branch and a spatial branch. The branches operate on different image characteristics and are fused according to their anomaly strengths and mutual consistency.

[00108] The frequency branch examines low-frequency, mid-frequency and high-frequency characteristics. The low-frequency portion may represent overall structure and large-scale repetition. The mid-frequency portion may represent recurring texture patterns. The high-frequency portion may represent fine detail, sensor noise and abrupt transitions.

[00109] Low-frequency analysis may employ a sliding window over encoded or decoded image data. Adjacent or spaced windows are compared to determine repetitive patterns. A frequency anomaly is generated when repetition exceeds a corresponding threshold.

[00110] Mid-frequency analysis may divide a selected portion of the image data into a plurality of chunks and determine similarity between neighbouring chunks. Repeated similarity may indicate texture patterns associated with generative processing.

[00111] High-frequency analysis may determine differences between adjacent data values and evaluate variation of the resulting differences. Artificial images may exhibit a more uniform noise distribution than camera-originated images because they do not contain natural sensor noise produced during image capture.

[00112] The frequency branch may further generate a multi-band spectral fingerprint. The image data is distributed across a plurality of frequency bands, and the normalized energy or occurrence within each band is stored as a spectral representation.

[00113] The spatial branch examines edge density, edge consistency, texture smoothness and colour distribution. Edge density may be determined from the proportion of transitions exceeding an edge threshold. Edge consistency may be determined from variation between successive edge magnitudes.

[00114] Artificial images may exhibit excessive edge regularity or boundaries that are unnaturally smooth. Texture smoothness may be determined by partitioning the image into local portions and comparing local-value distributions. Excessive uniformity may indicate artificial generation.

[00115] Colour-distribution analysis may examine hue, saturation, intensity or encoded value distributions. An image having an unusually uniform colour distribution or a generator-associated chromatic bias may be assigned a corresponding colour anomaly value.

[00116] The frequency and spatial branches produce respective anomaly scores and confidence values. An inter-domain agreement is determined by comparing the two branch outputs. A high inter-domain agreement indicates that spectral and spatial evidence support a similar conclusion. A low agreement indicates that one domain contradicts the other.

[00117] Dynamic attention weights are assigned to the two branches. The branch exhibiting a stronger anomaly signal may be assigned greater influence, subject to the inter-domain agreement. When both branches strongly indicate artificial generation, the combined result may be increased. When the branches significantly disagree, the combined result may be reduced.

[00118] In another embodiment, query, key and value attention representations are generated for the frequency and spatial features. Bidirectional attention is applied so that frequency features influence the interpretation of spatial features and spatial features influence the interpretation of frequency features. A consistency gate suppresses a branch when its output statistically contradicts the corresponding evidence from the other branch.

[00119] The resulting fusion vector therefore does not represent a simple concatenation or average of two modalities. Instead, the contribution of each domain is controlled according to the relative anomaly strength and cross-domain consistency.

[00120] Explainable Artifact Detection

[00121] The system comprises an explainability module configured to convert internal image-analysis outputs into spatial and textual forensic evidence. The explainability module produces an attention map, identifies suspicious image regions, classifies artifact types and generates a structured explanation.

[00122] The image may be divided into a fixed or adaptive grid. In an exemplary embodiment, the image is divided into twenty-five cells arranged as a five-by-five grid. Each cell represents a spatial region of the image and is assigned an attention value.

[00123] The attention value may be derived from the combined detection score, detected artifact regions, gradient information, feature activations and proximity to previously identified anomalies. The highest-ranking cells are designated as artifact hotspots.

[00124] The distribution of attention values may be classified as concentrated, distributed or substantially uniform. A concentrated distribution indicates that anomalies are localized in one or more specific regions. A distributed pattern indicates that anomalies occur across multiple image regions. A uniform pattern may indicate that the entire image shares similar generative characteristics.

[00125] The explainability module identifies artifact regions and classifies them into one or more categories. The categories may include anatomical, texture, lighting, frequency, semantic, edge, background and colour artifacts.

[00126] Each artifact region is associated with coordinates, a confidence value and an explanatory description. For example, the system may report that a high-confidence spectral inconsistency is present in an upper image region, that facial texture is excessively smooth, or that a boundary between an object and its background exhibits unnatural blending.

[00127] A semantic artifact translation engine may receive an activation matrix and feature-attribution vectors. The activation matrix identifies regions having elevated model response. The attribution vectors indicate whether the corresponding feature contributes towards an artificial or authentic classification.

[00128] The translation engine maps combinations of activation location, artifact category and contribution direction to a structured natural-language statement. The statement identifies what was detected, where it was detected and why the identified feature influenced the verdict.

[00129] The module may further rank key factors according to their contribution to the detection result. Each factor may be classified as artificial-indicative, authentic-indicative or neutral. The resulting explanation therefore extends beyond a heatmap and provides an interpretable evidentiary summary.

[00130] Ensemble Uncertainty Quantification

[00131] The system comprises an uncertainty assessment module configured to distinguish a detection probability from the reliability of that probability. The module determines a confidence interval, separates uncertainty arising from image-data ambiguity from uncertainty arising from limited detector knowledge, and assigns a reliability classification.

[00132] The uncertainty assessment may operate on outputs of multiple detector modules, repeated inference operations or both. Repeated inference may be carried out using stochastic perturbation, including random deactivation of selected model elements, to determine whether the prediction remains stable.

[00133] A weighted ensemble output is determined from the detector results. Variance and standard deviation are calculated from the distribution of the outputs. A confidence interval is generated around the weighted ensemble result.

[00134] The module identifies outlier detector outputs that differ substantially from the ensemble mean. An outlier may indicate a detector-specific failure, generator-specific blind spot or incompatibility between the input image and the corresponding model.

[00135] Data-associated uncertainty, also referred to as aleatoric uncertainty, represents uncertainty arising from the image itself. Such uncertainty may result from low resolution, blur, noise, severe compression, cropping or limited visible evidence.

[00136] Model-associated uncertainty, also referred to as epistemic uncertainty, represents uncertainty arising from disagreement, insufficient model knowledge or an image characteristic not adequately represented in the detector training data.

[00137] The two uncertainty components may be combined into a total uncertainty value while remaining separately available in the forensic result. Separation of the uncertainty components allows the system to distinguish poor-quality evidence from an inadequate detector.

[00138] The module assigns a reliability tier according to the uncertainty characteristics. The reliability tiers may include a high-reliability tier, a review tier, a cautionary tier and a human-assessment-required tier. In another embodiment, the tiers may be represented as court-ready, investigative lead, cautionary and inadmissible.

[00139] Different remediation actions may be associated with the separate uncertainty components. High data-associated uncertainty may trigger image enhancement, acquisition of a higher-quality image or rejection of inadequate evidence. High model-associated uncertainty may trigger quarantine of the sample, additional detector analysis or inclusion of the sample in a model-retraining workflow.

[00140] **Generator Fingerprint Identification**

[00141] The system further comprises a generator fingerprint identification module configured to determine a likely source of an artificial image. The module constructs a composite source fingerprint from multiple independent signal layers rather than relying solely upon a noise pattern.

[00142] The composite source fingerprint may include a spectral signature, chromatic profile, texture profile and residual-noise profile. Each profile represents a different characteristic associated with image generation.

[00143] The spectral signature may be obtained using a cosine transform, Fourier transform or another frequency decomposition. The signature may represent high-frequency roll-off, periodic energy distribution, upsampling

artifacts and generator-associated frequency patterns.

[00144] The chromatic profile may be generated from hue, saturation and value distributions. Certain generators or training datasets may exhibit characteristic saturation, colour-temperature or hue biases.

[00145] The texture profile may represent repetition frequency, smoothness, local detail density and spatial texture relationships. The residual-noise profile may be generated by applying a denoising filter and analysing the remaining image residual. The residual may include patterns associated with a generator architecture, initialization process or image-synthesis pipeline.

[00146] The separate profile vectors are combined into a multi-dimensional source representation. The source representation is compared with reference fingerprints stored in a generator-signature database.

[00147] The database may include reference signatures associated with diffusion models, generative adversarial networks, multimodal image generators and a camera-originated image baseline. The reference database may be updated as additional generators become available.

[00148] Similarity between the image fingerprint and each stored reference fingerprint may be determined using cosine similarity, Euclidean distance, learned metric comparison or another vector-matching operation. A source probability distribution is generated from the resulting similarities.

[00149] When compression is detected, the relative contribution of the spectral and chromatic components may be increased because compression may destroy or distort residual noise. When the image is substantially uncompressed, the noise component may receive a greater contribution.

[00150] The module identifies the most likely generator when a matching score exceeds a predetermined attribution threshold and is sufficiently distinguishable from the real-image baseline. When no reference signature satisfies the attribution requirements, the system may indicate that the image is likely artificial but the specific generator is unknown.

[00151] Differential Composite Anomaly Detection

[00152] The system comprises a spatial artifact mapping module configured to detect partially manipulated or composite images. Such images may include inpainted, outpainted, spliced or regionally generated content combined with authentic image regions.

[00153] The image is divided into a plurality of patches. The patch dimensions may be dynamically selected according to the image resolution so that a relatively small manipulated object is not obscured by a larger authentic background.

[00154] For each patch, the module determines local noise, texture, edge, frequency and artifact-density characteristics. Dominant characteristics of the entire image are used to establish a global image baseline.

[00155] The local characteristics of each patch are compared with the global baseline. A statistical divergence value is generated to represent the extent to which a local patch differs from the dominant image characteristics.

[00156] A patch exhibiting a high divergence is identified as a candidate manipulated region. However, the system does not automatically classify the entire image as composite merely because a local anomaly is present. Composite decision logic examines relationships between neighbouring patches.

[00157] A composite-manipulation verdict may be generated when one or more patches having a high artificial-content confidence are surrounded by or spatially associated with patches having a high authenticity confidence. In contrast, when substantially all patches exhibit similar artificial characteristics, the image may be classified as uniformly generated rather than composite.

[00158] The resulting output comprises a deviation matrix, tamper map and segmentation mask. The tamper map identifies regions likely to contain artificial or manipulated content and may separately indicate authentic and artificial portions.

[00159] Vision-Prompt-Tuned Analysis and Region-Level Verification

[00160] The system may employ a vision-prompt-tuned model to perform detailed visual analysis and identify candidate artifact regions. The model generates bounding-box coordinates, artifact categories and explanatory findings.

[00161] In an exemplary implementation, a multimodal vision-language model adapted using prompt tuning or a low-rank adapter is hosted on a local graphics-processing resource. The named model and adaptation technique are merely examples and may be replaced with another region-capable visual analysis model.

[00162] Each candidate region identified by the vision model is independently examined by a localized signal-analysis module. This cross-verification prevents acceptance of a candidate artifact merely because it was produced by the vision model.

[00163] The region verification module extracts image data corresponding to the bounding-box coordinates. The extracted image portion is analysed independently of the global image.

[00164] Localized analysis may determine entropy, frequency-energy distribution, high-frequency noise, edge consistency and transition smoothness. The module may apply a localized discrete cosine transform and examine the distribution of transform coefficients.

[00165] The module may further determine a Laplacian response to estimate high-frequency detail and gradient characteristics to evaluate edge consistency. Pixel-transition analysis may determine whether the local region is unnaturally smooth or excessively regular.

[00166] The localized characteristics are combined to generate a verification result. A candidate region satisfying a verification threshold is retained as a confirmed artifact region. A candidate region failing the verification threshold is discarded or marked as unverified.

[00167] The final artifact map distinguishes vision-model detections confirmed by signal-level analysis from unverified detections. This two-stage process reduces false positives caused by semantic or visual misinterpretation by the

vision model.

[00168] Forensic Verdict Generation

[00169] Following completion of the detection, fusion, calibration, dual-domain analysis, explainability, uncertainty, attribution and verification operations, a forensic output generator assembles the results.

[00170] The output may comprise the reliability-weighted score, calibrated score, detector trust score, individual detector scores, detector operational statuses, disagreement classification, conflict indicators, uncertainty values, reliability tier, attention map, artifact hotspots, confirmed artifact regions, tamper map, source-attribution result, watermark findings and metadata findings.

[00171] The calibrated score may be mapped to a plurality of verdict categories. In an exemplary embodiment, a score greater than 0.75 is mapped to "AI GENERATED," a score from 0.55 to 0.75 is mapped to "LIKELY AI," a score from 0.45 to 0.55 is mapped to "UNCERTAIN," a score from 0.25 to 0.45 is mapped to "LIKELY REAL," and a score below 0.25 is mapped to "REAL."

[00172] The score ranges are exemplary and may be modified after calibration using a validation dataset. The selected thresholds may vary according to whether the system is deployed for social-media moderation, journalistic verification, legal evidence assessment, cybersecurity investigation or another forensic application.

[00173] A natural-language summary is generated according to the verdict and supporting evidence. The summary may identify the strongest artificial-indicative factors, the most relevant authentic-indicative factors, suspicious image regions, the degree of detector agreement and any requirement for human review.

[00174] The output may be displayed through an interactive interface having confidence gauges, detector breakdowns, spatial heatmaps, bounding boxes, provenance information and source-attribution results. A downloadable forensic report may also be produced.

[00175] Further, the operations need not be performed in the disclosed order, although in some examples, an order may be preferred. Also, not all functions

need to be performed to achieve the desired advantages of the disclosed system and method, and therefore not all functions are required.

[00176] Various modifications to these embodiments are apparent to those skilled in the art from the description and the accompanying drawings. The principles associated with the various embodiments described herein may be applied to other embodiments. Therefore, the description is not intended to be limited to the embodiments shown along with the accompanying drawings but is to be providing the broadest scope consistent with the principles and the novel and inventive features disclosed or suggested herein. Accordingly, the invention is anticipated to hold on to all other such alternatives, modifications, and variations that fall within the scope of the present invention and appended claims.

CLAIMS

I/We Claim:

1. A forensic image analysis system for detection, localization and source attribution of artificial intelligence-generated or artificial intelligence-manipulated content in a digital image, the system comprising:

i) an image input interface configured to receive the digital image and generate image data corresponding to the digital image;

ii) an orchestration controller configured to asynchronously distribute the image data to a plurality of heterogeneous detector modules and receive, from each detector module, a detector score and detector metadata comprising a confidence value, an operational status and processing-time information;

iii) a dynamic reliability-weighted fusion engine configured to determine a runtime reliability of each detector module based on the detector metadata, dynamically modify a relative contribution of each detector module and combine the detector scores using normalized detector weights to generate a fused detection score;

iv) a disagreement-aware confidence calibration engine configured to determine an inter-detector disagreement based on dispersion and range of the detector scores, identify conflicting detector outputs and calibrate the fused detection score towards a neutral detection state in proportion to the determined inter-detector disagreement;

v) a dual-domain image analysis engine comprising a frequency-domain analysis branch and a spatial-domain analysis branch configured to respectively generate a frequency anomaly representation and a spatial anomaly representation, and a consistency-gated fusion module configured to dynamically combine the frequency anomaly representation and the spatial anomaly representation based on an inter-domain agreement;

vi) an artifact localization engine configured to partition the digital image into a plurality of image regions, determine region-specific anomaly information and generate an artifact map identifying one or more candidate artifact regions;

vii) an uncertainty assessment engine configured to determine uncertainty associated with outputs of the plurality of heterogeneous detector modules and separately characterize data-associated uncertainty and model-associated uncertainty;

viii) a generator attribution engine configured to generate a composite source fingerprint from a plurality of signal characteristics of the digital image and compare the composite source fingerprint with stored reference fingerprints associated with a plurality of image generators;

ix) a region verification engine configured to receive a candidate artifact region identified by a vision-based detection model and independently analyse frequency and spatial characteristics of the candidate artifact region to confirm or reject the candidate artifact region; and

x) a forensic output generator configured to generate a forensic result comprising a calibrated detection score, a detection verdict, detector trust information, localized artifact information, uncertainty information and generator-attribution information,

wherein the orchestration controller controls execution of the dynamic reliability-weighted fusion engine, the disagreement-aware confidence calibration engine, the dual-domain image analysis engine and the region verification engine in an interconnected forensic processing pipeline and dynamically reduces contribution of a detector module operating in a fallback or error state.

2. The forensic image analysis system as claimed in claim 1, wherein the dynamic reliability-weighted fusion engine is configured to determine the runtime reliability of each detector module based on a combination of the confidence value generated by the detector module, a status score corresponding to successful operation, fallback operation or error operation, and a normalized processing-time

characteristic, and adjust a predetermined base weight of the detector module according to the determined runtime reliability before combining the detector scores.

3. The forensic image analysis system as claimed in claim 1, wherein the disagreement-aware confidence calibration engine is configured to determine a weighted dispersion of the detector scores, determine a difference between a maximum detector score and a minimum detector score, classify the inter-detector disagreement into a plurality of disagreement levels comprising agreement, mild disagreement, moderate disagreement and severe disagreement, and identify a conflict when at least two detector modules generate detection scores representing opposing detection states relative to the neutral detection state.

4. The forensic image analysis system as claimed in claim 1, wherein the frequency-domain analysis branch is configured to determine low-frequency pattern repetition, mid-frequency texture repetition and high-frequency variation characteristics of the image data, and wherein the spatial-domain analysis branch is configured to determine edge consistency, texture smoothness and colour-distribution characteristics, the consistency-gated fusion module being configured to assign dynamically determined relative weights to the frequency-domain analysis branch and the spatial-domain analysis branch based on respective anomaly strengths and the inter-domain agreement.

5. The forensic image analysis system as claimed in claim 1, wherein the uncertainty assessment engine is configured to perform repeated inference operations or ensemble-output analysis to determine a distribution of detection outputs, derive a confidence interval corresponding to a combined detection output, separate uncertainty associated with image-data ambiguity from uncertainty associated with disagreement or limited knowledge of the detector modules, and assign a reliability category to the forensic result based on the separated uncertainty.

6. The forensic image analysis system as claimed in claim 1, wherein the generator attribution engine is configured to extract:

i) a spectral signature representing frequency-energy characteristics;

ii) a chromatic profile representing colour-distribution and saturation characteristics;

iii) a texture profile representing texture smoothness, repetition and detail characteristics; and

iv) a noise profile representing residual noise characteristics,

and combine the spectral signature, the chromatic profile, the texture profile and the noise profile into the composite source fingerprint for similarity comparison with the stored reference fingerprints, wherein a relative contribution of the spectral signature is increased upon detection of compression of the digital image.

7. A method for forensic analysis of artificial intelligence-generated or artificial intelligence-manipulated content in a digital image, the method comprising:

i) receiving and validating the digital image;

ii) generating a content identifier corresponding to the digital image and performing a cache lookup for a previously generated forensic result;

iii) asynchronously supplying image data corresponding to the digital image to a plurality of heterogeneous detector modules configured to analyse different image characteristics;

iv) receiving a detector score, a confidence value, an operational status and processing-time information from each detector module;

v) determining a runtime reliability of each detector module from the confidence value, operational status and processing-time information;

vi) dynamically modifying relative detector weights based on the determined runtime reliability and combining the detector scores to generate a reliability-weighted fused detection score;

vii) determining a disagreement characteristic from the detector scores, identifying conflicting detector outputs and calibrating the reliability-weighted fused detection score according to the disagreement characteristic;

viii) separately analysing the digital image in a frequency domain and a spatial domain and dynamically combining frequency-domain anomaly information and spatial-domain anomaly information based on an agreement between the frequency-domain anomaly information and the spatial-domain anomaly information;

ix) partitioning the digital image into a plurality of image regions and determining region-specific anomaly characteristics to generate an artifact map;

x) determining data-associated uncertainty and model-associated uncertainty corresponding to the detection output;

xi) extracting a plurality of image signal characteristics and generating a composite source fingerprint for comparison with stored source fingerprints;

xii) identifying one or more candidate artifact regions using a vision-based detection model and independently verifying the one or more candidate artifact regions using localized image-signal analysis; and

xiii) generating a forensic output comprising a detection verdict, calibrated detection information, detector trust information, localized artifact information, uncertainty information and source-attribution information.

8. The method as claimed in claim 7, wherein partitioning the digital image comprises dynamically determining a grid size based on a resolution of the digital image, generating a plurality of image patches, establishing a global image baseline from dominant noise and texture characteristics of the digital image, comparing local artifact characteristics of each image patch with the global image baseline, and generating a composite manipulation indication when an image patch having an artificial intelligence-associated anomaly is spatially associated with one or more image patches exhibiting authentic image characteristics.

9. The method as claimed in claim 7, wherein independently verifying the one or more candidate artifact regions comprises receiving bounding-box coordinates generated by a vision-prompt-tuned image analysis model, extracting image data within each bounding box, analysing entropy, localized frequency characteristics, edge consistency and pixel-transition characteristics of the extracted image data, generating a region verification output and retaining or rejecting the candidate artifact region based on the region verification output; and wherein the method further comprises sequentially examining the digital image for content credentials, an invisible watermark, exchangeable image file metadata and extensible metadata platform information and associating detected provenance information with the forensic output.

Dated this 25th day of September 2026



Somya Kaushik
AGENT FOR THE APPLICANT
IN/PA/5019

ABSTRACT

“A forensic image analysis system for detection, localization and source attribution of artificial intelligence-generated or artificial intelligence-manipulated content in a digital image”

The present invention relates to a forensic image analysis system and method for detecting, localizing and attributing artificial intelligence-generated or artificial intelligence-manipulated content. The system employs a microservice-based orchestration architecture configured to dispatch image data to a plurality of heterogeneous detector modules and receive respective detection scores, confidence information, operational-status information and processing metadata. A dynamic reliability-weighted fusion engine adjusts contributions of the detector modules according to runtime reliability, while a disagreement-aware calibration engine reduces confidence when detector outputs conflict. The system further performs frequency-domain and spatial-domain analysis and dynamically combines corresponding anomaly representations based on inter-domain consistency. An explainability module generates region-specific artifact maps and structured forensic explanations. An uncertainty assessment module separately characterizes data-associated and model-associated uncertainty. A generator attribution module compares composite spectral, chromatic, texture and noise fingerprints with stored generator signatures. Candidate artifact regions are independently verified using localized frequency and spatial analysis, and a forensic output comprising a calibrated verdict, trust information, artifact localization, uncertainty assessment and source-attribution information is generated.

Fig 1 to 10.

Dated this 25th day of September 2026



Somya Kaushik
AGENT FOR THE APPLICANT
IN/PA/5019

5