Universiteit Leiden

Measuring Human–Model Alignment of Post-hoc Deepfake Evidence under Controlled Compression

Daria Semenova (s3898296)

Supervisors:
Derya Soydaner & Qianpu Chen

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 9, 2026

Abstract

Deepfake video detectors are commonly evaluated through classification accuracy or confidence scores, but these metrics do not show whether a model bases its decisions on visual artefacts that are also recognisable to humans. This thesis investigates human–model alignment in explainable deepfake video detection under compression degradation. Using GenD, a CLIP-based generalisable deepfake detector, I evaluate whether perturbation-based model evidence maps align with human click-localised artefact annotations from ExDDV. The experiments cover the FaceForensics++, DeeperForensics, and BioDeepAV subsets for which source videos were accessible. For each annotated frame, I evaluate GenD under the original condition and two controlled compression settings, CRF 28 and CRF 35. The results show that compression consistently reduces fake-class confidence and increases prediction instability. However, human-clicked artefact regions remain strongly aligned with model-sensitive regions across datasets and compression levels. A shuffled-click control baseline further shows that this alignment is stronger than expected from the overall spatial distribution of human clicks alone for the two larger datasets. These findings suggest that compression affects model confidence more strongly than it disrupts the spatial structure of model evidence.

Contents

1 Introduction

1.1 Motivation

Synthetic media have become increasingly realistic, accessible, and widely distributed through publicly available generative AI systems and social media platforms. Unlike earlier forms of media manipulation, modern deepfake technologies are no longer limited to specialised organisations or technical experts. Realistic face-swapped and AI-generated videos can now be produced by ordinary users using commercially available applications and online services and subsequently disseminated through social media platforms within minutes. As a result, the challenge of deepfakes is no longer confined to AI developers and technology companies; it increasingly concerns user-generated content that circulates across online platforms [2, 15].

At the same time, regulators have begun responding to the growing societal impact of synthetic media. Article 3(60) of Regulation (EU) 2024/1689, the European Union Artificial Intelligence Act, defines a "deep fake" as AI-generated or manipulated image, audio, or video content that resembles existing persons, objects, places, entities, or events and would falsely appear to be authentic or truthful [6]. Article 50 of the same Regulation, which becomes applicable from 2 August 2026, introduces transparency obligations for certain AI systems, including systems that generate or manipulate content constituting deepfakes [6]. These obligations focus primarily on disclosure, labelling, and transparency for providers and deployers of AI systems. However, they provide limited guidance on how the reliability and human interpretability of deepfake detection systems themselves should be evaluated.

At the same time, much of today's deepfake ecosystem consists of user-generated content distributed through social media platforms. Manipulated videos may be repeatedly uploaded, downloaded, recompressed, screen-recorded, and reshared across different contexts. In such environments, transparency obligations alone may be insufficient, as the practical assessment of potentially manipulated content often relies on automated detection systems. This raises the question of whether the evidence produced by such systems is sufficiently interpretable to support human oversight and accountability.

This creates an important gap between regulatory transparency requirements and technical accountability. In practice, deepfake detection systems may increasingly support content moderation, journalistic verification, abuse reporting, and other high-stakes assessments of online media [5]. Yet many detectors remain black-box systems that produce a confidence score without explaining why a video has been classified as manipulated [1]. A detector may output that a video is fake with high confidence while providing little information about the visual evidence underlying that decision.

This limitation is particularly important because deepfake detection is often applied to content that has been uploaded, downloaded, recompressed, screen-recorded, and redistributed across multiple platforms. Under such conditions, both humans and AI systems may rely on altered or degraded visual cues, and detector confidence may not accurately reflect the quality of its underlying evidence. Therefore, evaluating deepfake detectors solely through classification performance is insufficient. It is also important to understand whether the evidence used by a detector corresponds to artefacts that humans can recognise and inspect.

To investigate this question, this thesis evaluates explanation-level robustness in deepfake detection. Specifically, it studies whether the evidence used by a detector corresponds to manipulation artefacts identified by human annotators and whether this relationship remains stable under realistic

compression conditions. Using the ExDDV dataset [7] and the GenD detector [20], the study examines whether detector evidence remains aligned with human-localised manipulation artefacts under increasing levels of video compression.

1.2 Problem Statement

Existing work on deepfake detection often focuses on classification performance. Benchmark accuracy, AUC scores, and cross-dataset generalisation metrics are important for evaluating whether a model can distinguish manipulated videos from authentic ones. However, classification scores do not directly answer whether a model's evidence corresponds to human-recognisable manipulation artefacts. A detector may achieve strong performance while relying on dataset-specific artefacts, background cues, compression traces, or other spurious signals rather than visually meaningful evidence.

Explainability methods can address this limitation by producing saliency maps, attention maps, or perturbation-based sensitivity maps [1, 18]. However, such explanations are often difficult to evaluate. A heatmap may appear plausible, but visual plausibility alone does not prove that the model is relying on meaningful manipulation artefacts. For this reason, explainability in deepfake detection should be evaluated quantitatively rather than treated only as a qualitative visual add-on.

The ExDDV dataset [7] makes it possible to study this problem more directly because it provides human click-localised artefact annotations. These clicks indicate where annotators perceived suspicious visual evidence in deepfake videos. This creates an opportunity to compare human-localised artefacts with post-hoc model evidence maps and to measure whether a detector is sensitive to regions that humans also identify as suspicious.

At the same time, compression introduces a realistic robustness challenge. Online videos are rarely shared in pristine form: they are often uploaded, downloaded, recompressed, screen-recorded, resized, and redistributed. If compression reduces the model's fake-class confidence, it is important to ask whether it also weakens the spatial alignment between human-clicked regions and model-sensitive regions. This thesis investigates that question using GenD [20], a generalisable deepfake detector, and controlled H.264 compression levels.

1.3 Research Question

This thesis investigates human–model alignment in explainable deepfake video detection under controlled compression degradation. The central research question is:

> To what extent do post-hoc explanation maps derived from GenD align with human click-localised deepfake artefacts on ExDDV, and how does this alignment change under increasing levels of controlled video compression?

This research question is divided into three subquestions:

1. How does controlled video compression affect GenD's fake-class probability and prediction stability?
2. Do human-clicked artefact regions correspond to model-sensitive regions in post-hoc evidence maps?

3. Does human–model alignment remain stable under stronger compression levels?

The thesis focuses on GenD because it is a recent, publicly available, generalisable deepfake detector [20]. It is suitable for this study because it is strong enough to be a credible detector, open enough to be reproducible, and frame-based enough to allow post-hoc spatial analysis. The thesis uses ExDDV because it provides human evidence traces rather than only binary real/fake labels [7]. Together, GenD and ExDDV allow this thesis to move from ordinary detection evaluation toward human-centred evidence alignment.

1.4 Contributions

This bachelor thesis makes five main contributions:

1. It implements a reproducible pipeline for evaluating human–model alignment between ExDDV human click annotations and GenD post-hoc evidence maps.

2. It evaluates GenD under controlled video compression at CRF 28 and CRF 35, allowing explanation-level robustness to be studied under matched degradation conditions.

3. It distinguishes between confidence degradation and spatial evidence alignment, showing that a detector's fake-class confidence and its human-aligned evidence structure can behave differently under compression.

4. It reports cross-dataset results on the available FaceForensics++, DeeperForensics, and BioDeepAV subsets of ExDDV, while documenting the exclusion of the inaccessible DeepfakeDetection/DFDC subset as a data-access limitation.

5. It introduces a shuffled-click control baseline to test whether human–model alignment exceeds what would be expected from generic spatial click distributions.

These contributions position the thesis between deepfake detection, explainable AI, and human-centred evaluation. Rather than proposing a new detector, the thesis evaluates whether a strong open detector produces evidence that is spatially aligned with human-localised artefacts and whether this alignment remains robust under compression.

1.5 Ethical Relevance

The ethical relevance of this thesis lies in the growing role of deepfake detection systems within emerging regulatory and platform-governance frameworks.

Article 50 of Regulation (EU) 2024/1689 introduces transparency obligations for AI-generated and manipulated content, including deepfakes [6]. At the same time, the Digital Services Act (Regulation (EU) 2022/2065) places increasing emphasis on transparency, accountability, and responsible content moderation on online platforms [5]. Together, these developments suggest that automated systems may play an increasingly important role in identifying, verifying, and assessing manipulated media distributed through social media environments.

However, while these regulatory frameworks address transparency, disclosure, and platform accountability, they provide limited guidance on how the outputs of deepfake detection systems

should be interpreted or evaluated. A detector may produce a confidence score indicating that content is likely manipulated, yet provide little insight into the evidence underlying that prediction. As deepfake detectors become increasingly integrated into moderation, fact-checking, abuse reporting, and verification workflows, the question of accountable evidence becomes increasingly important.

A human-centred detector should support inspection rather than replace human judgement entirely. This is particularly important because both false positives and false negatives can cause harm. A false negative may allow manipulated content to circulate, while a false positive may wrongly discredit authentic evidence. For individuals targeted by non-consensual deepfakes, unreliable detection may make it more difficult to demonstrate that harmful content has been fabricated [15]. For journalists, fact-checkers, and political audiences, poorly understood detection systems may contribute to a "liar's dividend", whereby authentic content is dismissed as fake [2].

Explainability does not eliminate these ethical risks by itself. A post-hoc explanation map is not equivalent to forensic proof, nor should it be treated as an objective ground truth. Nevertheless, explanation-level evaluation can improve the inspectability of detection systems [1]. If a model's evidence consistently overlaps with human-localised manipulation artefacts, this provides one indication that the detector relies on visually meaningful signals rather than spurious correlations. Conversely, if compression causes model evidence to diverge from human-recognisable artefacts, this may represent an important warning sign for real-world deployment.

The importance of interpretability is also reflected more broadly in the EU AI Act. While Article 86 concerns explanations in the context of certain high-risk AI systems rather than deepfake detection specifically, it illustrates a broader regulatory recognition that affected individuals may require meaningful information about AI-supported decisions [6]. However, no comparable requirement currently exists for deepfake detection systems used in moderation or verification contexts. This makes explanation-level evaluation especially relevant: if detection systems influence whether content is trusted, removed, or investigated, their evidence should be inspectable rather than limited to an opaque confidence score.

This thesis therefore approaches human–model alignment as a component of responsible and accountable deepfake detection. The goal is not to automate trust, but to evaluate whether the evidence used by a detector can be meaningfully related to human perception. In doing so, the thesis connects deepfake detection with broader discussions in explainable AI, human-centred AI, AI governance, and emerging regulatory frameworks for synthetic media.

1.6 Thesis Overview

This bachelor thesis was conducted at LIACS as part of the Data Science and Artificial Intelligence programme. Section 2 introduces the necessary background on deepfake video detection, generalisation, explainable AI, and video compression. Section 3 discusses related work on deepfake detection benchmarks, generalisable deepfake detection, explainable deepfake detection, human-localised artefact annotations, compression robustness, and the regulatory and ethical context of deepfake detection. Section 4 presents the dataset, model, frame extraction procedure, compression protocol, post-hoc evidence map method, and alignment metrics. Section 5 reports the experimental results, including compression effects and human–model alignment under different compression levels. Section 6 discusses the findings, ethical implications, limitations, and future work. Finally, Section **??** concludes the thesis.

2 Background

This section introduces the technical concepts needed for the thesis. It explains deepfake video detection, the problem of generalisation, post-hoc evidence maps, human–model alignment, and video compression. Unlike the related work section, which discusses prior research, this section defines the concepts used in the experimental design.

2.1 Deepfake Video Detection

Deepfake video detection is the task of distinguishing manipulated videos from authentic videos. In face manipulation settings, deepfakes may be produced through face swapping, facial reenactment, identity transfer, expression manipulation, or generative synthesis. Such manipulations can leave visible or statistical artefacts in the resulting video. These artefacts may appear around the mouth, eyes, facial boundary, skin texture, hairline, lighting, or temporal transitions between frames.

The detection task is usually formulated as a binary classification problem. Given an input frame or video, a detector predicts whether the content is real or fake. In many systems, the model outputs logits or class probabilities, and the fake-class probability can be interpreted as the model's confidence that the input is manipulated. In this thesis, the fake-class probability is used as one measure of detector behaviour under compression.

However, deepfake detection is challenging because manipulation artefacts are not fixed. Different generation methods leave different traces, and newer deepfake techniques may remove or reduce artefacts that were visible in earlier datasets. Moreover, many videos shared online are compressed, resized, or otherwise post-processed, which can hide manipulation traces or introduce new visual artefacts. For this reason, deepfake detection is not only a classification problem, but also a problem of robustness and generalisation.

2.2 Generalisable Deepfake Detection

A central challenge in deepfake detection is generalisation. A detector may perform well on the dataset or manipulation method it was trained on, but fail when tested on a different dataset, generation technique, video quality, or compression level. This can happen when the model learns dataset-specific shortcuts rather than general manipulation evidence. For example, a detector may rely on background patterns, preprocessing artefacts, compression traces, or other incidental features that correlate with the fake label in one dataset but do not generalise to another.

Generalisable deepfake detection aims to reduce this problem by developing models that perform well across different benchmarks and manipulation types. This is important for real-world use because manipulated videos encountered on social media or in verification contexts are unlikely to match the exact conditions of a training dataset. In such settings, a detector should ideally rely on evidence that remains meaningful across datasets and video conditions.

This thesis uses GenD as the detector because it is designed for generalisable deepfake detection across benchmarks [20]. The model is used without fine-tuning. This means that the experiments evaluate GenD as an off-the-shelf detector rather than adapting it to the ExDDV dataset. This choice is important because the thesis focuses on evaluating model evidence, not on training a new detector.

2.3 Explainable AI and Post-hoc Evidence Maps

Explainable artificial intelligence aims to make model behaviour more interpretable to humans. In computer vision, explanation methods often produce spatial maps that indicate which parts of an image influenced a model's prediction. These maps may be produced using gradients, attention, occlusion, perturbation, or other post-hoc methods [1]. Grad-CAM, for example, is a widely used method for producing visual explanations from convolutional neural networks [18].

In deepfake detection, explanation maps are especially relevant because many manipulation artefacts are spatially localised. A human observer may notice a strange mouth boundary, inconsistent eyes, abnormal skin texture, or unnatural blending around the face. If a detector's explanation highlights similar regions, this may indicate that the detector is relying on visually meaningful manipulation evidence. If the explanation highlights irrelevant regions, such as the background or image borders, this may suggest reliance on spurious cues.

However, explanation maps should be interpreted carefully. A visually plausible heatmap does not prove that the model is reasoning correctly, and a post-hoc explanation is not equivalent to forensic ground truth. Explanation methods can also be sensitive to implementation choices and may produce different maps for the same prediction. Therefore, explanation maps should not only be inspected visually but also evaluated quantitatively.

This thesis uses a perturbation-based post-hoc evidence map. Instead of relying on internal model gradients or attention weights, the input frame is systematically modified, and the effect on the model's fake-class probability is measured. Regions that cause larger changes in the fake-class probability when perturbed are treated as more model-sensitive. This produces a spatial evidence map that can be compared with human click annotations.

2.4 Human–Model Alignment

Human–model alignment in this thesis refers to the spatial relationship between human-localised artefact evidence and model-sensitive regions. The human evidence comes from ExDDV, where annotators clicked on regions they perceived as suspicious in deepfake videos [7]. The model evidence comes from perturbation-based sensitivity maps produced for the same annotated frames.

The purpose of human–model alignment is not to claim that the model and the human reason in exactly the same way. A model may use statistical cues that are not consciously visible to a human observer, and a human click is only a sparse indication of a perceived artefact. Nevertheless, if human-clicked regions consistently fall within high-sensitivity regions of the model evidence map, this suggests that the detector is at least partly sensitive to visually meaningful artefacts.

This form of alignment is useful because it moves evaluation beyond binary correctness. A detector may classify a video correctly while relying on irrelevant or dataset-specific evidence. Conversely, a detector may become less confident under degradation while still preserving a spatial evidence structure that overlaps with human-recognisable artefacts. Measuring human–model alignment therefore provides an additional perspective on whether model evidence is inspectable and meaningful.

In this thesis, alignment is measured using point-based metrics. The human click location is mapped to the corresponding cell in the model evidence grid. The sensitivity of that cell is then compared with the sensitivity values of all grid cells. This allows the thesis to measure whether human-clicked artefacts fall in highly model-sensitive regions.

2.5 Video Compression and Robustness

Video compression reduces file size by removing, approximating, or encoding visual information more efficiently. Online platforms commonly recompress uploaded videos to reduce storage and bandwidth requirements. As a result, videos encountered in real-world settings are often not pristine copies of the original file. They may have been uploaded, downloaded, resized, recompressed, screen-recorded, or reshared multiple times.

This thesis uses H.264 compression controlled through the Constant Rate Factor (CRF) parameter. In CRF-based encoding, lower CRF values generally correspond to higher visual quality and larger file size, while higher CRF values correspond to stronger compression and lower visual quality. The original available video condition is treated as the baseline, and two additional compression conditions are used: CRF 28 and CRF 35.

Compression is relevant to deepfake detection because it can affect both the detector's prediction and its explanation map. Compression may remove subtle manipulation artefacts, introduce block artefacts, blur facial details, or change texture statistics. A detector may therefore become less confident, more unstable, or more dependent on different visual cues after compression. Prior work has shown that video processing operations and social media compression can affect deepfake detection performance in realistic deployment scenarios [11, 10, 13].

For this reason, robustness should not be evaluated only through classification accuracy or confidence. It is also important to ask whether the spatial evidence used by a detector remains meaningful under compression. This thesis therefore evaluates both confidence robustness and explanation-level robustness. Confidence robustness is measured through changes in fake-class probability and prediction flips, while explanation-level robustness is measured through the stability of human–model alignment under increasing compression.

3 Related Work

This thesis builds on work in deepfake detection benchmarks, generalisable detection, explainable AI, human-localised artefact annotation, compression robustness, and the regulatory and ethical context of synthetic media. The central gap addressed by this thesis is not the absence of deepfake detectors, but the limited evaluation of whether detector evidence corresponds to human-recognisable manipulation artefacts under realistic video degradation.

3.1 Deepfake Detection Benchmarks

Deepfake detection research has been strongly shaped by benchmark datasets. FaceForensics++ is one of the most widely used benchmarks for facial manipulation detection and includes several manipulation methods such as Face2Face, FaceSwap, DeepFakes, NeuralTextures, and FaceShifter [16]. It is especially relevant for this thesis because it introduced evaluation under different video compression settings, making it an important reference point for studying robustness to video quality degradation.

DeeperForensics-1.0 was introduced as a large-scale dataset for real-world face forgery detection and includes more diverse perturbations and higher-quality forgery generation than earlier benchmarks [9]. The Deepfake Detection Challenge dataset further expanded the scale and diversity of deepfake detection evaluation by introducing a large collection of manipulated and authentic videos

for benchmark comparison [3]. More recently, DeepfakeBench has provided a broader benchmarking framework for comparing deepfake detection methods across datasets and settings [19].

These datasets have been crucial for measuring classification performance and cross-dataset generalisation. However, most benchmark datasets primarily provide binary real/fake labels or video-level labels. Such labels make it possible to evaluate whether a detector is correct, but they do not reveal whether the model relies on human-interpretable evidence. This limitation motivates the use of datasets that include human-localised artefact annotations rather than only classification labels.

3.2 Generalisable Deepfake Detection

A major challenge in deepfake detection is generalisation. Detectors often perform well on the dataset or manipulation method used during training but degrade when tested on unseen datasets, compression levels, or generation techniques. This occurs because models may learn dataset-specific artefacts, compression traces, background cues, or preprocessing signatures rather than general manipulation evidence.

Recent work has therefore focused on improving cross-dataset and cross-manipulation robustness. GenD is particularly relevant to this thesis because it is designed for generalisable deepfake detection across benchmarks [20]. The model builds on strong visual representations and is suitable as an off-the-shelf detector for evaluating whether a generalisable model also produces human-aligned evidence. The use of CLIP-style representations is also relevant because CLIP has shown strong transfer capabilities across visual tasks [14].

This thesis does not propose a new detector or fine-tune GenD for ExDDV. Instead, it uses GenD as an existing generalisable detector and evaluates its evidence structure. This shifts the focus from whether a model can classify deepfakes correctly to whether the spatial evidence behind its predictions corresponds to artefacts identified by human annotators.

3.3 Explainable Deepfake Detection

Explainable AI aims to make model decisions more interpretable by identifying which input regions or features influence a prediction. In computer vision, explanation methods often produce saliency maps, attention maps, gradient-based localisation maps, or perturbation-based sensitivity maps [1]. Grad-CAM is a widely used example of a visual explanation method that highlights image regions contributing to a neural network decision [18].

In deepfake detection, explanation methods are especially relevant because manipulation artefacts are often spatially localised. Human observers may notice unnatural mouth boundaries, inconsistent eye regions, blurred facial contours, abnormal skin texture, or lighting mismatches. A useful detector explanation should therefore ideally highlight regions that correspond to such visible manipulation cues rather than irrelevant background areas.

However, explanation maps are difficult to evaluate. A heatmap may appear visually plausible while still failing to reflect meaningful model reasoning. Conversely, a detector may rely on subtle cues that are difficult for humans to interpret. This makes it important to move beyond qualitative inspection of heatmaps and toward quantitative evaluation of explanation maps. The present thesis contributes to this direction by comparing post-hoc model evidence maps with human click-localised artefact annotations.

3.4 Human-localised Artefact Annotations

Human-localised annotations provide a way to evaluate whether model evidence corresponds to regions that people perceive as suspicious. ExDDV is central to this thesis because it provides human click localisations, textual artefact descriptions, timestamps, and difficulty labels for explainable deepfake detection in video [7]. Unlike datasets that only indicate whether a video is real or fake, ExDDV allows researchers to compare model evidence against human-perceived manipulation artefacts.

The use of click annotations differs from segmentation masks or bounding boxes. A click is sparse and does not define the full spatial extent of an artefact. However, it directly indicates where a human annotator noticed suspicious evidence. This makes click-based evaluation suitable for point-based metrics such as click percentile, top-k hit rate, and distance to the most sensitive model region.

This thesis builds directly on this property of ExDDV. Human clicks are treated as reference points for visible artefact evidence, while GenD blur-grid sensitivity maps are treated as post-hoc model evidence. The goal is not to prove that the model reasons in the same way as a human, but to measure whether the model is sensitive to regions that humans also identify as suspicious.

3.5 Compression Robustness in Deepfake Detection

Compression robustness is an important issue in deepfake detection because online videos are rarely shared in pristine form. They are often uploaded to platforms, recompressed, resized, downloaded, screen-recorded, and reshared. These operations may remove subtle manipulation artefacts, introduce new compression artefacts, or alter the visual statistics that detectors rely on.

Earlier benchmark work already recognised the importance of compression. FaceForensics++ includes different compression levels, and this has made it a common reference point for studying detector performance under video quality degradation [16]. DeeperForensics also emphasises robustness to real-world perturbations and more challenging evaluation conditions [9].

More directly, Lu and Ebrahimi studied the impact of video processing operations on deepfake detection and showed that real-world processing can affect detector performance [11]. Their related assessment framework further argues that deepfake detection should be evaluated under realistic deployment conditions rather than only on clean benchmark data [10]. Montibeller et al. extend this line of work by focusing specifically on social network compression emulation, showing that platform-specific compression and resizing can create a gap between laboratory benchmarks and real-world deployment [13]. The NTIRE 2026 Robust Deepfake Detection Challenge further confirms the current relevance of robustness under common and uncommon degradations [8].

These studies motivate the compression component of this thesis. However, most compression robustness work focuses on classification performance, confidence, or accuracy under degradation. This thesis complements that work by asking whether compression also affects explanation-level robustness: whether the spatial relationship between human-clicked artefacts and model-sensitive regions remains stable under controlled H.264 recompression.

3.6 Regulatory and Ethical Context of Deepfake Detection

Deepfake detection is increasingly relevant not only as a computer vision task, but also as part of broader debates about digital trust, platform governance, and AI regulation. Article 3(60) of Regulation (EU) 2024/1689, the European Union Artificial Intelligence Act, defines a "deep fake" as AI-generated or manipulated image, audio, or video content that may falsely appear authentic or truthful [6]. Article 50 of the same Regulation introduces transparency obligations for certain AI systems that generate or manipulate such content [6].

These legal developments show that synthetic media are now explicitly recognised as a regulatory concern. However, legal transparency obligations do not by themselves solve the technical problem of how manipulated media should be detected, verified, or explained in practice. Meding and Sorge argue that the boundary between legitimate editing and manipulation under the EU AI Act remains conceptually and technically difficult to define [12]. Schmitt et al. further argue that Article 50 creates structural compliance challenges, including difficulties around machine-readable marking and reliable detection across technical environments [17]. The European Commission's 2026 draft guidelines on Article 50 also indicate that the implementation of transparency obligations for synthetic media remains an active and evolving regulatory issue [4].

The Digital Services Act is also relevant because deepfake content often circulates through online platforms. The DSA places increasing emphasis on platform transparency, accountability, and content moderation responsibilities [5]. In this context, deepfake detectors may become part of moderation, verification, and reporting workflows. If such systems influence whether content is removed, trusted, or investigated, their outputs should ideally be interpretable rather than limited to opaque confidence scores.

Several ethical risks make this issue especially important. Deepfakes can contribute to reputational harm, privacy violations, image-based abuse, political manipulation, and democratic distrust. Chesney and Citron describe the broader risks of deepfakes for privacy, democracy, and national security, including the "liar's dividend", where authentic content may be dismissed as fake [2]. Romero Moreno further frames harmful deepfake content through a human-rights perspective, emphasising risks to dignity, privacy, autonomy, and democratic participation [15].

Within this regulatory and ethical context, explainability becomes more than a technical convenience. A detector that outputs only a confidence score may be insufficient for journalists, moderators, investigators, or affected individuals who need to understand why content has been flagged. Although Article 86 of the EU AI Act concerns explanations in the context of certain high-risk AI systems rather than deepfake detection specifically, it reflects a broader regulatory recognition that affected individuals may require meaningful information about AI-supported decisions [6]. This thesis contributes to this discussion by evaluating whether post-hoc detector evidence corresponds to human-localised manipulation artefacts, especially under realistic compression degradation.

3.7 Summary of Research Gap

Existing research has made substantial progress in deepfake detection benchmarks, cross-dataset generalisation, explainable AI, and robustness under compression. However, these lines of work are often evaluated separately. Benchmark studies typically focus on classification performance, explainability studies often rely on qualitative heatmap inspection, and compression studies usually measure accuracy or confidence degradation rather than explanation-level alignment.

This thesis addresses the gap between these areas. It evaluates whether a generalisable deepfake detector produces post-hoc evidence that aligns with human-localised artefacts and whether this alignment remains stable under controlled compression. In doing so, it connects deepfake detection, explainable AI, compression robustness, and responsible deployment under emerging regulatory expectations.

4 Approach

This section describes the experimental approach used to evaluate human–model alignment in deepfake video detection under controlled compression. The approach consists of eight stages: dataset selection, model selection, frame extraction, video recompression, post-hoc evidence map generation, human–model alignment measurement, shuffled-click baseline construction, and statistical evaluation.

4.1 Implementation Environment

The experiments were implemented in Python using a local Conda environment. The GenD and ExDDV repositories were cloned from their public GitHub releases. Video processing was performed with FFmpeg using H.264 encoding. The main Python packages used in the implementation included PyTorch, Transformers, OpenCV, pandas, NumPy, SciPy, scikit-learn, tqdm, and matplotlib.

The experiments were run locally on macOS using PyTorch with Apple Metal Performance Shaders support. The official GenD dependency file could not be installed directly on the local machine because it included a GPU-specific dependency that was not available for the macOS ARM platform. Therefore, the required dependencies were installed manually. The Transformers and Hugging Face Hub versions were fixed to compatible versions after an initial loading issue. After this adjustment, the `GenD_CLIP_L_14` checkpoint loaded successfully.

An alternative `GenD_DINOv3_L` checkpoint was also tested during setup. However, this model required access to a gated DINOv3 backbone on Hugging Face and could not be loaded without authentication. For this reason, the final experiments use only `GenD_CLIP_L_14`. This choice is also consistent with the aim of the thesis: to evaluate a strong, publicly accessible, generalisable detector rather than compare multiple model backbones.

4.2 Dataset

The experiments use the ExDDV annotation file [7]. ExDDV provides human-localised evidence for explainable deepfake detection in video. Each annotation row contains a video filename, source dataset label, manipulation type, textual artefact description, difficulty label, split information, and click locations. The click locations are stored as frame-indexed coordinates, where each annotated frame contains normalised x and y coordinates indicating where a human annotator identified suspicious visual evidence.

The full ExDDV annotation file contains 4347 rows and 9 columns: `id`, `username`, `movie_name`, `text`, `dataset`, `manipulation`, `click_locations`, `difficulty`, and `split`. The split distribution consists of 3562 training rows, 389 validation rows, and 396 test rows. This thesis uses the test split only, because the goal is to evaluate a pretrained detector without training or fine-tuning.

Only rows with non-empty click locations were retained. The final implemented evaluation includes the FaceForensics++, DeeperForensics, and BioDeepAV subsets. In the ExDDV file, the FaceForensics++ label appears as `Farceforensics++`; this label is normalised to FaceForensics++ in the thesis tables and discussion. The DeepfakeDetection/DFDC subset was excluded because the corresponding source videos were not accessible through the available data access route during the experiment.

The final evaluation covers 371 ExDDV test entries from the included subsets. These correspond to 1754 annotated frames: 1245 from FaceForensics++, 481 from DeeperForensics, and 28 from BioDeepAV. The number of unique available source video files differs slightly from the number of ExDDV entries because some entries can refer to the same source video. The final coverage consists of 243 unique FaceForensics++ videos, 90 DeeperForensics videos, and 10 BioDeepAV videos.

4.3 Model

The detector used in the main experiments is `GenD_CLIP_L_14` [20]. GenD is a generalisable deepfake detector designed to perform across different benchmarks. The selected checkpoint uses a CLIP ViT-L/14 visual backbone, which is relevant because CLIP-style models are known for strong transferable visual representations [14].

In this thesis, GenD refers to the complete deepfake detector evaluated in the experiments, while CLIP refers only to the visual backbone used by this specific GenD checkpoint. Therefore, experimental descriptions refer to frames being passed through GenD, GenD output logits, and GenD fake-class probabilities. The term CLIP is used only when describing the underlying backbone architecture.

The model is used without fine-tuning. This is an intentional methodological choice. The thesis does not aim to improve GenD on ExDDV or adapt it to the specific annotation set. Instead, it evaluates whether an existing generalisable detector produces spatial evidence that aligns with human-localised manipulation artefacts.

For each extracted frame, the model outputs logits for two classes. Based on the official GenD usage examples and model behaviour, class 0 is treated as the real class and class 1 as the fake class. The fake-class probability is computed by applying the softmax function to the two output logits:

$$P(\text{fake} \mid x) = \frac{\exp(z_{\text{fake}})}{\exp(z_{\text{real}}) + \exp(z_{\text{fake}})}.$$

Here, x is the input frame, while z_{real} and z_{fake} are the model logits for the real and fake classes. This fake-class probability is used both for measuring prediction confidence and for generating perturbation-based evidence maps.

4.4 Frame Extraction

The evaluation is frame-based. Although the source data are videos, the ExDDV annotations are associated with specific frame indices. For each retained ExDDV row, the `click_locations` field was parsed to identify the annotated frame numbers and their corresponding normalised click coordinates.

For every available source video, the annotated frames were extracted according to the frame indices provided in ExDDV. Each extracted frame was linked to its source dataset, video filename,

manipulation type, difficulty label, frame index, and click coordinate. This design makes it possible to compare the detector's spatial evidence with the exact frame-level human annotation.

The use of frame-level evaluation is appropriate for this thesis because the selected GenD checkpoint operates on individual frames. Temporal information is therefore outside the scope of the implemented experiments. This limitation is discussed later in Section 6.

4.5 Compression Protocol

To evaluate robustness under controlled degradation, each available source video was recompressed using H.264 at two Constant Rate Factor settings: CRF 28 and CRF 35. The original available video condition is treated as the baseline condition. The two compressed conditions represent increasing levels of degradation.

The three evaluation conditions are therefore:

1. baseline / original available video condition;
2. CRF 28 compressed condition;
3. CRF 35 compressed condition.

The same annotated frame indices were extracted from the baseline and compressed versions of each video. This produces matched frame-level samples across compression conditions. The matched design is important because it allows direct comparison of model confidence and human–model alignment for the same annotated visual evidence under different levels of compression.

Compression was introduced after the source videos were obtained rather than by sampling unrelated compressed videos. This ensures that any observed changes in prediction confidence or evidence alignment can be attributed to the controlled recompression step rather than to differences in video content.

4.6 Post-hoc Evidence Maps

To estimate model-sensitive regions, this thesis uses a perturbation-based blur-grid method. This method is post-hoc because it does not modify or retrain the detector. It only measures how the detector's output changes when different spatial regions of the input frame are perturbed.

Each frame is divided into a fixed 7×7 grid, producing 49 cells. For each grid cell, the corresponding image region is blurred while the rest of the frame remains unchanged. The blurred frame is then passed through GenD, and the fake-class probability is recomputed.

Let x be the original frame and let x_{ij}^{blur} be the same frame with grid cell (i, j) blurred. The sensitivity score for cell (i, j) is defined as the absolute change in fake-class probability caused by blurring that cell:

$$S_{ij} = \left| P(\text{fake} \mid x) - P(\text{fake} \mid x_{ij}^{\text{blur}}) \right|.$$

A high value of S_{ij} indicates that perturbing that cell strongly changes the model's fake-class probability. Such a cell is therefore treated as more important for the model's decision. The full set of 49 sensitivity scores forms a coarse spatial evidence map for the frame.

Blur perturbation is used instead of black-box occlusion because it preserves the approximate colour and structure of the local region while reducing detail. This makes the perturbation less visually artificial than replacing a region with a constant colour patch. Since the method only requires repeated forward passes through the model, it is architecture-independent and can be applied without access to internal gradients or attention weights.

4.7 Human–Model Alignment Metrics

The main goal of the alignment analysis is to test whether human-clicked artefact locations fall in regions that are important to the model. Each human click is mapped to the corresponding grid cell in the 7×7 evidence map. The sensitivity value of this clicked cell is then compared with the sensitivity values of all grid cells.

Click percentile. The click percentile measures the percentile rank of the human-clicked cell within the model sensitivity map. A value close to 1 means that the human click falls in one of the most model-sensitive regions. A value close to 0 means that the click falls in a region with low model sensitivity. A value of 0.5 corresponds to the random expectation.

Top-20 hit rate. The top-20 hit rate measures whether the human-clicked cell falls within the top 20% most sensitive grid cells. Under random alignment, the expected hit rate is 20%. Values substantially above 20% indicate that human clicks are more likely than chance to fall in model-sensitive regions.

Top-10 and Top-1 hit rates. The top-10 and top-1 hit rates provide stricter versions of the same measure. The top-10 hit rate checks whether the click falls within the top 10% most sensitive cells, while the top-1 hit rate checks whether the click falls in the single most sensitive grid cell.

Distance to maximum sensitivity. The distance-to-maximum metric measures the normalised spatial distance between the human click location and the most sensitive grid cell. Lower values indicate that the human click is closer to the model's maximum-sensitivity region. This metric complements the percentile and hit-rate metrics by measuring spatial proximity rather than rank alone.

Together, these metrics capture different aspects of human–model alignment. Click percentile evaluates relative sensitivity, top-k hit rates evaluate whether the click falls into highly sensitive regions, and distance to maximum sensitivity evaluates spatial closeness to the strongest model evidence.

4.8 Shuffled-click Control Baseline

To test whether the observed human–model alignment could be explained by a general face-centred or centre-biased prior, an additional shuffled-click control baseline was computed across all included datasets. This baseline keeps the model sensitivity maps fixed but randomly permutes human click locations across frames within the same source dataset and compression condition.

This procedure preserves the overall spatial distribution of human annotations within each dataset-condition group, but breaks the frame-specific correspondence between a human click and the model evidence map. If the actual alignment remains higher than the shuffled-click baseline, this suggests that the observed alignment is not explained only by humans and the model both focusing on similar face-centred regions. Instead, it provides evidence for a stronger frame-specific relationship between human-localised artefacts and model-sensitive regions.

The shuffled-click baseline was computed using 1000 random permutations. For each permutation, the click percentile, top-20 hit rate, top-10 hit rate, and distance to the maximum-sensitivity cell were recomputed. Empirical p-values were calculated as the proportion of shuffled runs that performed at least as well as the actual alignment. For click percentile and top-k hit rates, higher values indicate stronger alignment. For distance to the maximum-sensitivity cell, lower values indicate stronger spatial agreement.

4.9 Statistical Evaluation

Three types of statistical evaluation are performed. First, compression effects on model confidence are evaluated using paired comparisons. For each annotated frame, the baseline fake-class probability is compared with the fake-class probability after CRF 28 and CRF 35 compression. The change in fake-class probability is computed as:

$$\Delta P(\text{fake}) = P(\text{fake} \mid x_{\text{compressed}}) - P(\text{fake} \mid x_{\text{baseline}}).$$

Negative values indicate that compression reduces the model's fake-class confidence. Prediction flip rate is also computed to measure the proportion of frames for which the predicted class changes between the baseline and compressed condition.

Second, human–model alignment is evaluated against random baselines. The mean click percentile is tested against the random expectation of 0.5. The top-20 hit rate is tested against the random expectation of 0.2 using a binomial test. These tests are performed separately for each source dataset and compression condition.

Third, the shuffled-click control baseline is evaluated using empirical permutation p-values. For each dataset-condition group, the observed alignment metrics are compared with 1000 shuffled-click permutations. For click percentile and top-k hit rates, the p-value is computed as the proportion of shuffled runs that achieve alignment at least as high as the actual clicks. For distance to the maximum-sensitivity cell, the p-value is computed as the proportion of shuffled runs that achieve a distance at least as low as the actual clicks.

This evaluation design separates confidence robustness from explanation-level robustness. Confidence robustness is measured through changes in fake-class probability and prediction flips. Explanation-level robustness is measured through the stability of click percentile, top-k hit rates, and distance-to-maximum sensitivity under compression.

5 Experiments

This section reports the empirical results of the proposed evaluation pipeline. The analysis focuses on three questions. First, it examines how controlled compression affects the prediction confidence of GenD. Second, it evaluates whether compression changes the stability of the detector's decisions.

Third, it measures whether model-sensitive regions remain spatially aligned with human-localised deepfake artefacts under increasing compression.

5.1 Dataset Coverage

The implemented experiments use the available ExDDV test samples for which the corresponding source videos could be obtained. The final evaluation includes the FaceForensics++, DeeperForensics, and BioDeepAV subsets. The DeepfakeDetection/DFDC subset is excluded from the implemented analysis because the corresponding original videos were not accessible through the available data access routes at the time of experimentation.

Table 1: Dataset coverage used in the implemented experiments.

Source dataset	Manipulation type(s)	ExDDV entries	Unique videos	Frames	Baseline mean P(fake)	GenD fake prediction rate
FaceForensics++	Face2Face, FaceShifter, FaceSwap	271	243	1245	0.386	35.9%
DeeperForensics	end-to-end	90	90	481	0.518	49.7%
BioDeepAV	videos	10	10	28	0.341	25.0%
Total included	–	371	343	1754	–	–

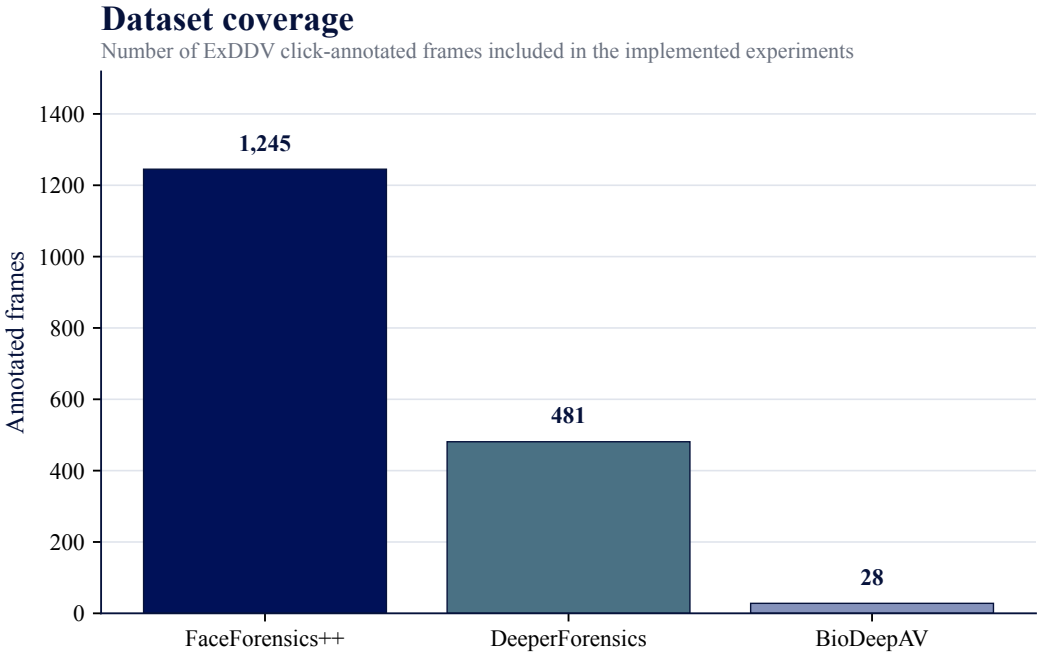

Figure 1: Dataset coverage in the implemented ExDDV evaluation. The figure shows the number of click-annotated frames included for each accessible source dataset.

Overall, the evaluation covers 371 ExDDV test entries and 1754 annotated frames. FaceForensics++ forms the largest subset, while BioDeepAV is included as a smaller external source. Because the subsets differ substantially in size, dataset-level comparisons should be interpreted with this imbalance in mind.

5.2 Effect of Compression on Model Confidence

Table 2 reports the effect of compression on the fake-class probability predicted by GenD. In all three datasets, compression reduces the mean probability assigned to the fake class. This indicates that compression degradation makes the model less confident in detecting manipulated content.

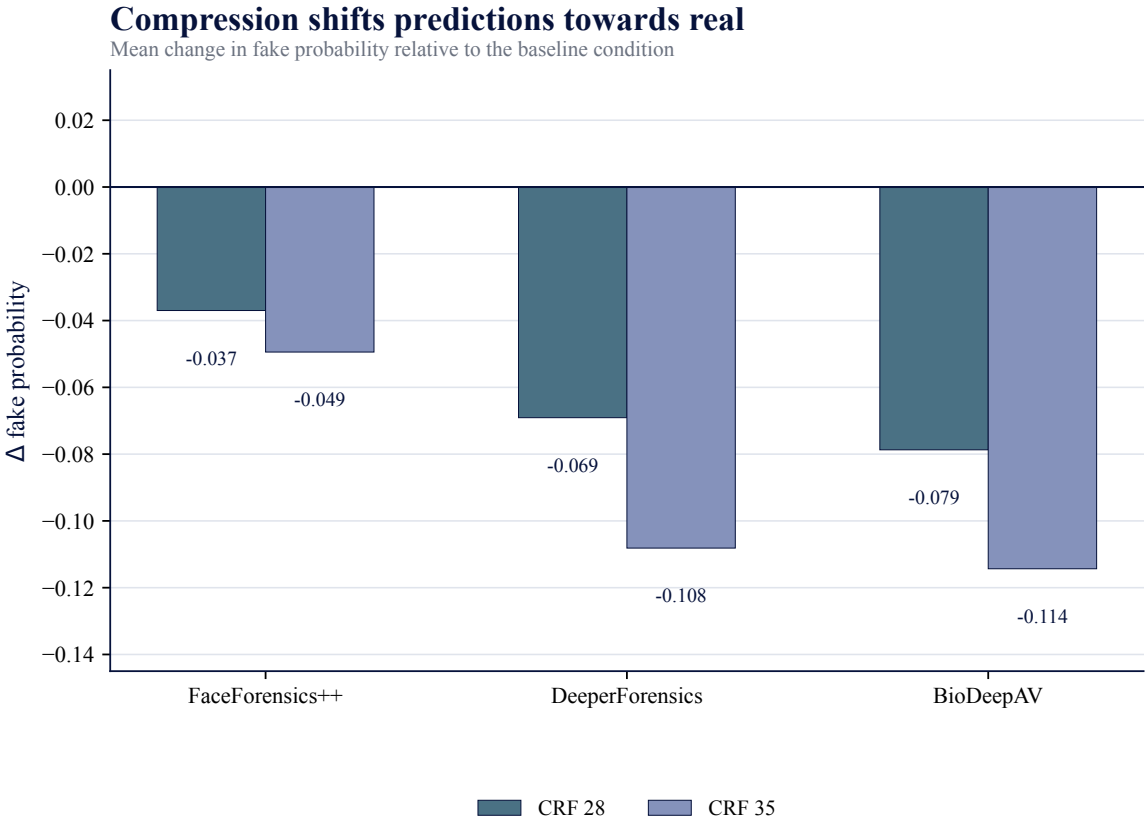

Figure 2: Mean change in GenD fake-class probability under controlled compression. Negative values indicate that compression shifts predictions towards the real class relative to the baseline condition.

Table 2: Effect of controlled compression on GenD fake-class probability.

Source dataset	Compression	Frames	Base P(fake)	Compressed P(fake)	ΔP(fake)	Flip rate	p
FaceForensics++	CRF 28	1245	0.386	0.349	-0.037	9.8%	< 0.001
FaceForensics++	CRF 35	1245	0.386	0.336	-0.049	14.4%	< 0.001
DeeperForensics	CRF 28	481	0.518	0.449	-0.069	14.3%	< 0.001
DeeperForensics	CRF 35	481	0.518	0.410	-0.108	21.8%	< 0.001
BioDeepAV	CRF 28	28	0.341	0.262	-0.079	17.9%	0.037
BioDeepAV	CRF 35	28	0.341	0.227	-0.114	17.9%	0.026

The largest absolute reduction is observed for DeeperForensics at CRF 35, where the mean fake probability decreases from 0.518 to 0.410. BioDeepAV also shows a strong decrease, although

this result should be interpreted cautiously because the subset contains only 28 annotated frames. FaceForensics++ shows a smaller but still statistically significant decrease. Across datasets, the decrease is stronger at CRF 35 than at CRF 28, suggesting that stronger compression produces stronger confidence degradation.

5.3 Prediction Stability under Compression

The flip rate in Table 2 measures the proportion of frames for which the predicted class changes after compression. Across all datasets, compression introduces prediction instability. For FaceForensics++, the flip rate increases from 9.8% at CRF 28 to 14.4% at CRF 35. For DeeperForensics, it increases from 14.3% to 21.8%, making it the most affected subset. BioDeepAV shows a flip rate of 17.9% under both compression levels, although this pattern should again be interpreted cautiously because of the small sample size.

These results show that compression affects not only the magnitude of the fake-class probability, but also the final binary decision for a non-trivial proportion of frames. This directly addresses the first research subquestion: controlled compression reduces GenD's fake-class confidence and can change its predictions.

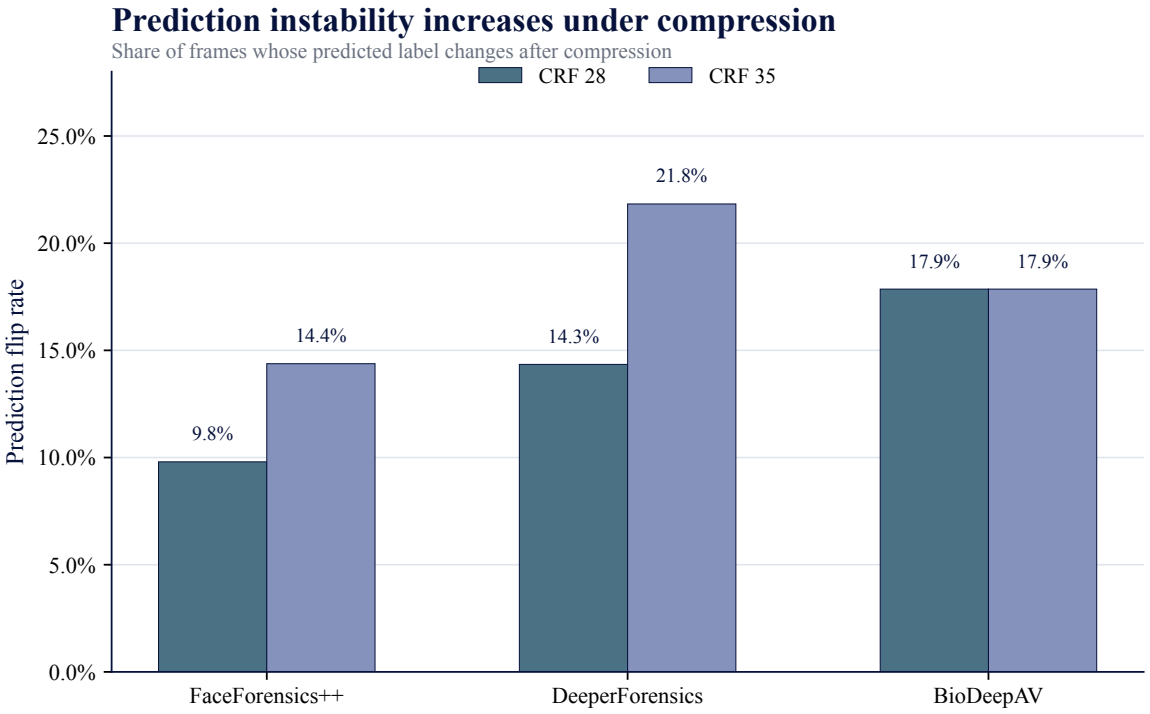

Figure 3: Prediction flip rate under controlled compression. The flip rate measures the share of frames whose predicted class changes between the baseline and compressed condition.

5.4 Human–Model Alignment

The second part of the evaluation measures whether GenD is spatially sensitive to the same regions that human annotators identified as deepfake artefacts. This is evaluated using the blur-grid

alignment procedure described in Section 4. Each frame is divided into grid cells, and the effect of blurring each cell on the model's fake-class probability is measured. Human clicks are then compared against this model sensitivity map.

Table 3 shows that human clicks consistently fall in high-sensitivity regions. The mean click percentile is substantially above the random baseline of 0.5 for all datasets and all compression levels. Similarly, the top-20 hit rate is much higher than the random expectation of 20%.

Table 3: Human–model alignment under baseline and compressed conditions.

Source dataset	Condition	Frames	Mean click percentile	Top-20 hit rate	Top-10 hit rate	Distance to max cell
FaceForensics++	Baseline	1245	0.887	81.3%	67.1%	0.142
FaceForensics++	CRF 28	1245	0.895	83.3%	68.5%	0.141
FaceForensics++	CRF 35	1245	0.888	82.2%	67.1%	0.141
DeeperForensics	Baseline	481	0.869	77.3%	61.5%	0.145
DeeperForensics	CRF 28	481	0.874	80.0%	57.8%	0.145
DeeperForensics	CRF 35	481	0.863	78.0%	58.8%	0.148
BioDeepAV	Baseline	28	0.819	67.9%	25.0%	0.188
BioDeepAV	CRF 28	28	0.824	71.4%	25.0%	0.227
BioDeepAV	CRF 35	28	0.835	64.3%	32.1%	0.193

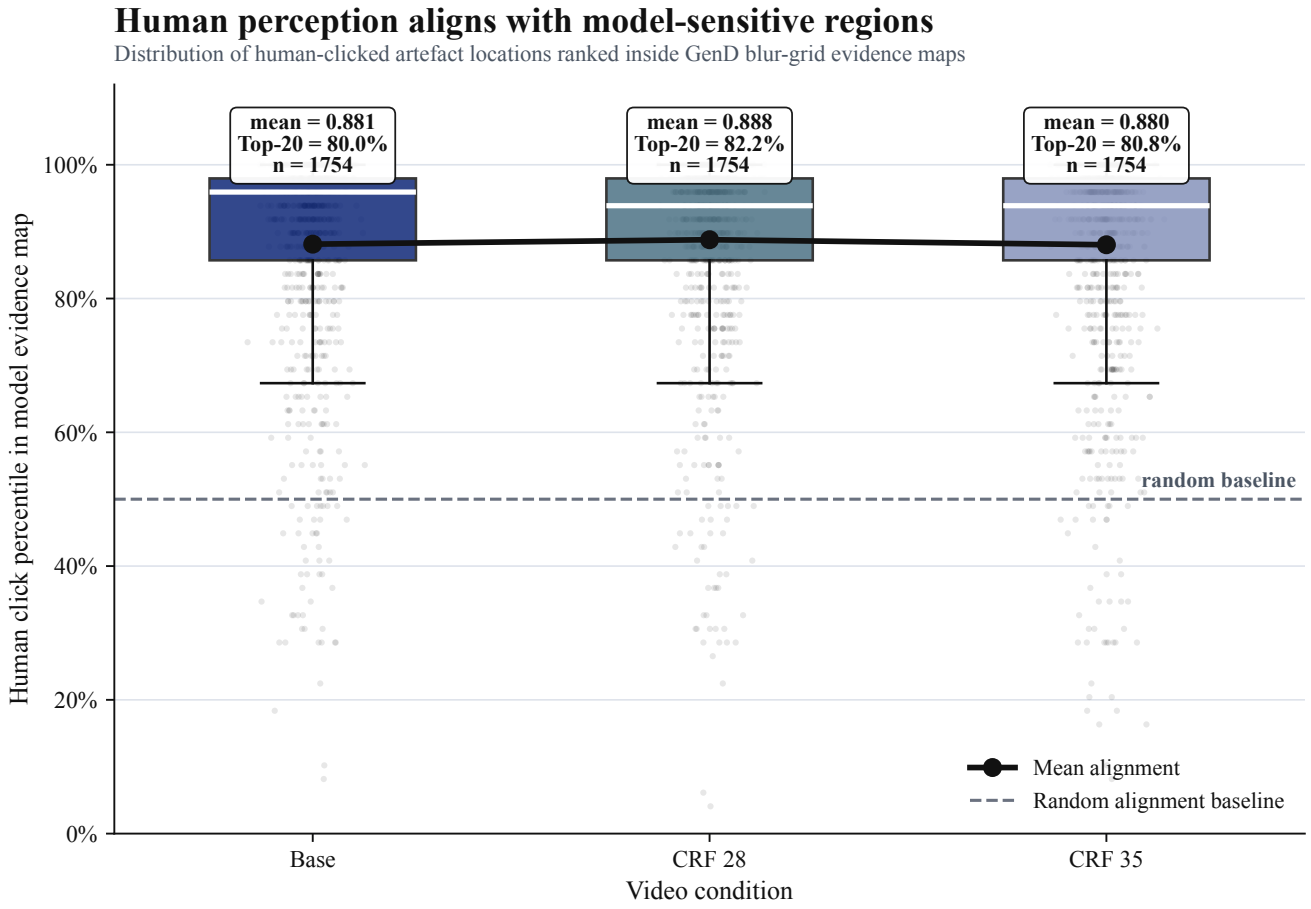

Figure 4: Distribution of human-clicked artefact locations ranked within GenD blur-grid evidence maps across compression conditions. The dashed line indicates the random alignment baseline.

Figure 5: Qualitative examples of GenD blur-grid evidence maps overlaid with ExDDV human click annotations. The heatmap shows perturbation sensitivity, while the cyan marker indicates the human-clicked artefact location. The first row shows a high-alignment example in which model-sensitive regions remain close to the human click across compression conditions. The second row shows a weaker-alignment example in which GenD still assigns a moderate fake-class probability, but the most sensitive regions diverge from the human-clicked location. This illustrates that prediction confidence and human–model spatial alignment are distinct evaluation signals.

The alignment results suggest that even when compression lowers the model's fake-class confidence, the spatial evidence used by the model remains strongly related to the regions selected by human annotators. This distinction is important: compression appears to degrade decision confidence more than it destroys the localisation of model-sensitive evidence.

5.5 Shuffled-click Control Baseline

The alignment results above show that human clicks frequently fall in model-sensitive regions. However, because the task concerns facial manipulation, both human annotations and model evidence may naturally concentrate around the face or central image regions. To test whether the observed alignment can be explained only by this general spatial prior, a shuffled-click control baseline was computed across all included datasets.

In this baseline, human clicks were randomly permuted across frames within the same source dataset and compression condition, while the model sensitivity maps were kept fixed. This preserves the overall spatial distribution of human clicks within each dataset-condition group, but breaks the frame-specific correspondence between the human annotation and the model evidence map.

Table 4: Actual human–model alignment compared with a shuffled-click baseline across datasets. The shuffled baseline preserves the overall spatial distribution of clicks within each dataset and condition, but breaks the frame-specific correspondence between human annotations and model evidence maps.

Dataset	Condition	Actual percentile	Shuffled percentile	Actual top-20	Shuffled top-20	p
BioDeepAV	Baseline	0.819	0.802	67.9%	60.6%	0.256
BioDeepAV	CRF 28	0.824	0.798	71.4%	61.7%	0.144
BioDeepAV	CRF 35	0.835	0.795	64.3%	60.3%	0.060
DeeperForensics	Baseline	0.869	0.763	77.3%	55.9%	$< .001$
DeeperForensics	CRF 28	0.874	0.764	80.0%	56.6%	$< .001$
DeeperForensics	CRF 35	0.863	0.763	78.0%	56.7%	$< .001$
FaceForensics++	Baseline	0.887	0.801	81.3%	63.8%	$< .001$
FaceForensics++	CRF 28	0.895	0.802	83.3%	63.9%	$< .001$
FaceForensics++	CRF 35	0.888	0.806	82.2%	65.3%	$< .001$

Across FaceForensics++ and DeeperForensics, actual human clicks achieve substantially higher click percentiles and top-20 hit rates than the shuffled-click baseline under all compression conditions. These differences are statistically significant with empirical permutation p-values below .001. This strengthens the main alignment result by showing that the observed correspondence is not explained only by a generic spatial prior, such as both humans and the model focusing on central facial regions.

BioDeepAV follows the same directional pattern: actual click percentiles are higher than the shuffled baseline across all conditions. However, these differences are not statistically significant. This is likely due to the small size of the BioDeepAV subset, which contains only 28 annotated frames per condition. Therefore, BioDeepAV is treated as supportive but underpowered evidence in the shuffled-click control analysis.

Overall, the shuffled-click control strengthens the main alignment finding. It shows that, at least for the two larger datasets, the observed human–model correspondence is stronger than would be expected from the overall click distribution alone.

5.6 Spatial Distance to Maximum Sensitivity

In addition to percentile-based and top-k alignment metrics, the analysis also measures the spatial distance between the human click and the most model-sensitive grid cell. This metric captures whether the strongest model evidence is located near the human-localised artefact, even when the click does not fall directly inside the highest-ranked grid cell.

Across datasets and compression conditions, the distance-to-maximum values remain low. FaceForensics++ and DeeperForensics show particularly stable distance values across baseline, CRF 28, and CRF 35. BioDeepAV shows more variation, but this should be interpreted cautiously because the subset contains only 28 annotated frames.

The distance-to-maximum results support the main alignment finding. Human-clicked artefacts remain spatially close to the model's strongest sensitivity regions. This provides additional evidence that the model evidence is not only highly ranked at the click location, but also spatially concentrated near human-recognisable artefacts.

5.7 Confidence and Alignment

The results above show a contrast between confidence degradation and spatial evidence stability. Compression consistently lowers the fake-class probability, but the alignment metrics remain high. This suggests that the model may become less confident under compression while still preserving a spatial evidence structure that corresponds to human-localised artefacts.

This distinction is central to the thesis. If only fake-class probability were reported, compression would appear mainly as a degradation problem. However, the alignment analysis shows that a decrease in confidence does not necessarily imply a complete loss of meaningful spatial evidence.

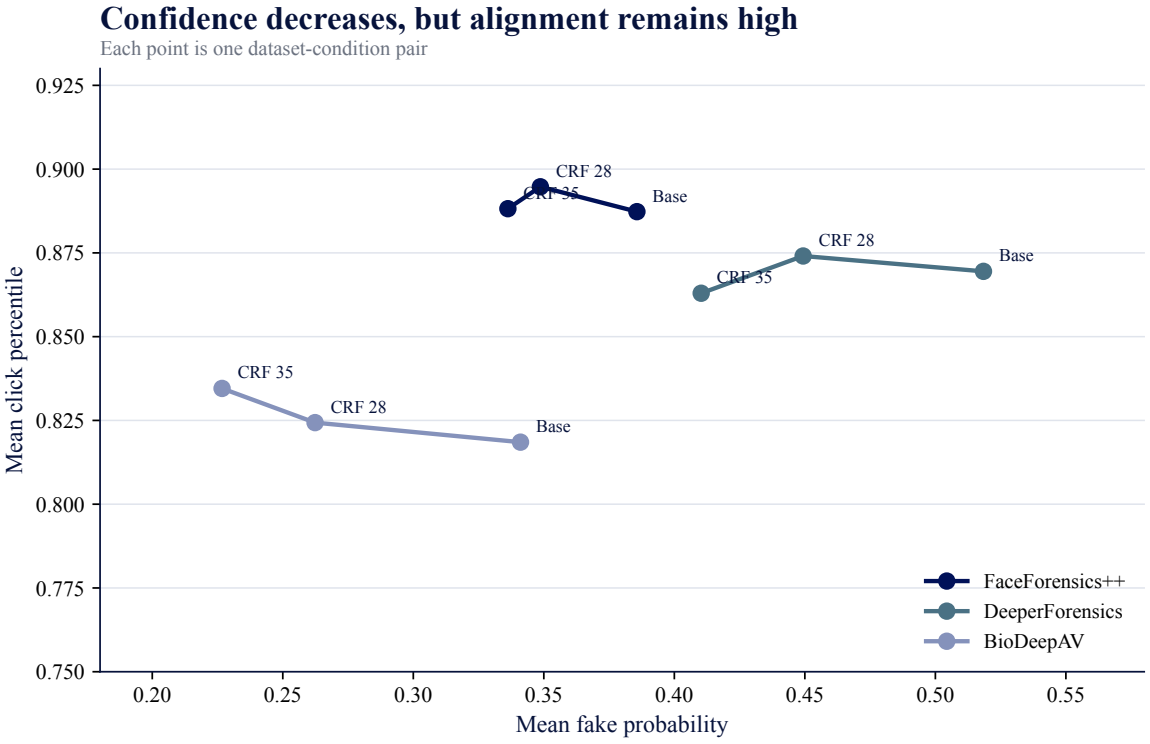

Figure 6: Relationship between mean fake-class probability and mean human–model alignment across datasets and compression conditions. Each point corresponds to one dataset-condition pair.

5.8 Summary of Findings

The experiments lead to three main findings. First, controlled compression systematically reduces GenD's fake-class confidence across all included source datasets. Second, stronger compression increases prediction instability, as shown by higher prediction flip rates. Third, despite this confidence degradation, model-sensitive regions remain strongly aligned with human-localised artefact regions.

The shuffled-click control baseline further shows that this alignment cannot be explained only by the overall spatial distribution of human clicks, with statistically significant improvements over shuffled clicks for FaceForensics++ and DeeperForensics.

For BioDeepAV, the same directional pattern is observed, but the control does not reach statistical significance, most likely because the subset is very small.

Together, these findings support the central argument of this thesis: human–model alignment can remain high even when model confidence degrades under compression. Therefore, evaluating deepfake detectors only through prediction scores is insufficient; explanation-level robustness provides an additional perspective on whether model evidence corresponds to human-recognisable artefacts.

6 Discussion

This section interprets the experimental findings in relation to the research question. The main result of this thesis is that controlled compression reduces GenD's fake-class confidence and increases prediction instability, but it does not substantially disrupt the spatial alignment between model-sensitive regions and human-localised artefacts. This suggests that prediction confidence and explanation-level alignment capture different aspects of detector robustness.

6.1 Confidence Degradation versus Spatial Evidence Stability

The first research subquestion asked how controlled video compression affects GenD's fake-class probability and prediction stability. The results show a consistent decrease in fake-class probability under both CRF 28 and CRF 35. This pattern appears across FaceForensics++, DeeperForensics, and BioDeepAV. The decrease is stronger under CRF 35, which indicates that stronger recompression produces stronger confidence degradation.

This pattern is consistent with previous work showing that video processing operations can substantially affect deepfake detector behaviour in realistic settings [11, 10]. It also relates to work on social-network compression emulation, which shows that platform-specific processing can create a gap between clean laboratory benchmarks and real-world deployment conditions [13]. The confidence degradation observed in this thesis therefore supports the broader finding that post-processing is a central challenge for real-world deepfake detection. However, the present results add an additional layer to this discussion: while compression reduces GenD's fake-class confidence, it does not proportionally reduce the spatial correspondence between human-clicked artefacts and model-sensitive regions.

This finding is consistent with the expectation that compression can remove or distort visual traces used by deepfake detectors. Video recompression can blur fine facial details, alter texture statistics, introduce block artefacts, and reduce the visibility of subtle manipulation cues. As a result, the detector becomes less confident that the frames are fake. The increase in prediction flip rate further shows that compression does not only change probability values, but can also change the final binary prediction for a non-trivial proportion of frames.

However, the alignment results show a different pattern. Despite the confidence decrease, human-clicked artefact regions remain strongly aligned with model-sensitive regions. Mean click percentiles stay far above the random baseline of 0.5, and top-20 hit rates remain much higher than the random expectation of 20%. This means that although compression weakens the strength of the model's fake decision, it does not fully destroy the spatial structure of the evidence used by the model.

This distinction is central to the thesis. If the detector were evaluated only through fake-class probability, the conclusion would be that compression primarily degrades performance. By

contrast, the explanation-level analysis shows that model evidence may remain spatially meaningful even when confidence decreases. Robustness should therefore not be understood only as stable classification confidence. It should also include whether the detector continues to rely on regions that correspond to human-recognisable artefacts.

6.2 Human–Model Alignment as an Additional Evaluation Signal

The second research subquestion asked whether human-clicked artefact regions correspond to model-sensitive regions in post-hoc evidence maps. The results provide strong evidence that they do. Across the included datasets and compression levels, human clicks frequently fall within the most sensitive regions of the blur-grid evidence maps. This suggests that GenD is not only producing predictions, but is also sensitive to regions that human annotators identified as suspicious.

This does not mean that the model and humans reason in the same way. Human clicks are sparse annotations that mark where annotators noticed visible artefacts, while the model may rely on statistical cues that are not consciously visible to humans. Similarly, a perturbation-based evidence map is an approximation of model sensitivity rather than a direct explanation of internal reasoning. Nevertheless, the high alignment scores indicate that the model's evidence is at least spatially related to human-perceived manipulation cues.

The shuffled-click control baseline strengthens this interpretation. Because the shuffled baseline preserves the overall distribution of human click locations while breaking the frame-specific correspondence between clicks and model evidence maps, it controls for a possible face-centred or centre-biased prior. Actual alignment remains significantly higher than the shuffled baseline for FaceForensics++ and DeeperForensics across all compression conditions. This suggests that the observed alignment is not only a consequence of humans and the model both focusing on the face region, but reflects a more specific relationship between human-localised artefacts and model-sensitive regions. For BioDeepAV, the same directional pattern is observed, but the shuffled-click comparison does not reach statistical significance, most likely because the subset is very small.

This result also clarifies the role of human-localised annotations. ExDDV was introduced to move explainable deepfake detection beyond binary labels by providing human descriptions and click locations for visible artefacts [7]. The present thesis uses these clicks not as training supervision, but as an external reference for auditing post-hoc model evidence. The high alignment scores therefore suggest that human-localised artefact annotations can be useful not only for developing explainable detectors, but also for evaluating whether an existing detector produces evidence that is spatially inspectable.

This is important because deepfake detection systems are often evaluated as black boxes. A detector may output a high fake probability, but this score alone does not show whether the model relied on the manipulated face region, irrelevant background information, compression artefacts, or dataset-specific shortcuts. Human–model alignment provides an additional evaluation signal: it tests whether the detector's sensitive regions overlap with human-localised evidence. In this sense, alignment analysis helps distinguish between merely correct predictions and more inspectable evidence.

6.3 Compression Robustness at the Explanation Level

The third research subquestion asked whether human–model alignment remains stable under stronger compression levels. The results suggest that it does. Although CRF 35 reduces fake-class confidence more strongly than CRF 28, the click percentile and top-20 hit rate remain high across conditions. The distance-to-maximum metric also indicates that human clicks remain spatially close to the most model-sensitive grid cells.

The findings should therefore be interpreted as complementary to previous compression-robustness studies. Earlier work mainly evaluates whether detector predictions remain accurate or stable after video processing [16, 11, 13]. This thesis instead shows that prediction confidence and explanation-level alignment can diverge: compression can make a detector less confident while leaving its most sensitive regions spatially close to human-localised artefacts. This does not mean that compression is harmless. Rather, it suggests that robustness has multiple dimensions, and that confidence degradation should be analysed together with evidence-level behaviour.

This finding is relevant because real-world user-generated videos are rarely shared in pristine form. Videos uploaded to social media platforms may be transcoded, resized, recompressed, downloaded, screen-recorded, and reshared. The compression levels used in this thesis are not intended to reproduce the proprietary pipelines of platforms such as Instagram, TikTok, or YouTube exactly. Instead, CRF 28 and CRF 35 serve as controlled proxy conditions for moderate and stronger recompression. This makes it possible to study whether evidence alignment survives degradation in a reproducible way.

The results show that compression affects confidence more strongly than alignment. This may indicate that the detector continues to attend to broadly meaningful facial regions even when compression reduces the strength or clarity of the fake signal. In practical terms, this suggests that a detector's confidence score and its explanation map may carry different information. A lower fake probability under compression does not necessarily mean that the detector's evidence has become meaningless. Conversely, a high confidence score would not be sufficient without evidence-level inspection.

6.4 Dataset-level Differences

The magnitude of the compression effect differs across datasets. DeeperForensics shows the strongest decrease in fake-class probability and the highest flip rate under CRF 35. This suggests that GenD's confidence on DeeperForensics is more sensitive to compression than on FaceForensics++. FaceForensics++ shows a smaller but still consistent confidence decrease. BioDeepAV also shows a large decrease, but this subset contains only 28 annotated frames, so its results should be interpreted cautiously.

The alignment metrics are more stable than the confidence metrics. FaceForensics++ and DeeperForensics both show high click percentiles and high top-20 hit rates across compression levels. BioDeepAV shows lower and less stable top-k values, which may be due to the small sample size and the different nature of the videos. Because BioDeepAV contributes only a small number of annotated frames, it should be treated as an additional exploratory subset rather than as a basis for strong dataset-level conclusions.

The shuffled-click baseline also highlights the effect of dataset size. For FaceForensics++ and DeeperForensics, actual alignment clearly exceeds the shuffled baseline. For BioDeepAV, actual

alignment is directionally higher than shuffled alignment, but the difference is not statistically significant. This supports treating BioDeepAV as an exploratory subset rather than as a basis for strong dataset-level conclusions.

These differences show why cross-dataset reporting is important. If the analysis were limited to one dataset, the findings might appear more uniform than they actually are. The current results suggest that confidence degradation is dataset-dependent, while explanation-level alignment is comparatively more stable across the included subsets.

6.5 Implications for Explainable and Accountable Deepfake Detection

The findings have implications for explainable deepfake detection. Many detection systems provide only a class label or confidence score. In high-stakes contexts, such as content moderation, journalistic verification, abuse reporting, or legal investigation, this may be insufficient. A confidence score can indicate that a video is likely manipulated, but it does not show what evidence supports that decision.

This thesis shows that explanation-level evaluation can provide a more informative view of detector behaviour. If model-sensitive regions consistently overlap with human-localised artefacts, this provides one indication that the detector relies on visually meaningful evidence rather than purely spurious cues. This does not make the detector's output forensic proof, and it does not eliminate the need for human judgement. However, it makes the detector more inspectable.

This point is also relevant to the regulatory and ethical context discussed earlier in the thesis. Article 50 of the EU AI Act introduces transparency obligations for certain AI-generated and manipulated content, while the Digital Services Act places increasing emphasis on platform accountability and content moderation transparency [6, 5]. However, these frameworks do not specify how deepfake detection evidence should be interpreted or explained. The results of this thesis suggest that human–model alignment could be one useful direction for evaluating whether detector evidence is meaningful enough to support human oversight.

6.6 Limitations

This thesis has several limitations. First, the implemented experiments do not include the DeepfakeDetection/DFDC subset of ExDDV because the corresponding source videos were not accessible through the available data access route during the experiment. As a result, the evaluation covers FaceForensics++, DeeperForensics, and BioDeepAV, but not the full ExDDV test set.

Second, the analysis uses only one detector: `GenD_CLIP_L_14`. This choice was appropriate because the model is publicly accessible, generalisable, and suitable for frame-level evaluation. However, the findings may not generalise to other detectors or architectures. An alternative GenD DINOv3 checkpoint was considered during setup, but it required access to a gated backbone and was therefore not used in the final experiments.

Third, the evidence maps are based on a fixed 7×7 blur-grid perturbation method. This method is simple, interpretable, and architecture-independent, but it is spatially coarse. Because each grid cell covers a relatively large part of the frame, the method may miss very fine-grained manipulation artefacts around the eyes, mouth, facial boundary, or skin texture. Smaller grid cells, adaptive regions, segmentation-based perturbations, or gradient-based methods could provide more

detailed localisation. However, these alternatives may also introduce additional computational cost or method-specific assumptions.

Fourth, the evaluation is frame-based. The source data are videos, but GenD is evaluated on extracted annotated frames. This means that temporal artefacts, motion inconsistencies, and frame-to-frame instability are outside the scope of the current experiments. Since some deepfake artefacts are temporal rather than spatial, future work should investigate whether human–model alignment also holds for video-level or temporally aware detectors.

Fifth, the compression protocol uses controlled H.264 recompression at CRF 28 and CRF 35. These conditions are useful for reproducible robustness testing, but they do not exactly reproduce the proprietary compression pipelines of social media platforms. Platforms such as Instagram, TikTok, and YouTube may apply additional resizing, adaptive bitrate encoding, format conversion, and platform-specific processing. Therefore, the results should be interpreted as evidence about robustness under controlled recompression, not as a direct measurement of platform-specific performance.

Sixth, the shuffled-click baseline is also affected by dataset size. While it provides strong evidence against a generic spatial-prior explanation for FaceForensics++ and DeeperForensics, the BioDeepAV subset is too small to provide a statistically strong shuffled-baseline comparison.

Finally, human clicks are sparse annotations. A click indicates where an annotator perceived suspicious evidence, but it does not define the full spatial extent of the artefact. Therefore, high alignment with click locations should be interpreted as evidence of spatial correspondence, not as proof that the model fully captures the same artefact concept as the human annotator.

6.7 Future Work

Future work could extend this thesis in several directions. First, the missing DeepfakeDetection/DFDC subset could be added if the corresponding source videos become accessible. This would make the evaluation closer to the full ExDDV test set and would provide a stronger basis for cross-dataset conclusions.

Second, future work could compare multiple detector backbones. For example, comparing `GenD_CLIP_L_14` with other GenD variants or other deepfake detectors would show whether the observed alignment patterns are specific to one model or more general across architectures.

Third, future work could evaluate platform-specific compression. Instead of using only controlled CRF settings, videos could be uploaded to and downloaded from social media platforms, or processed through social network compression emulators. This would make it possible to test whether human–model alignment remains stable under more realistic distribution pipelines.

Fourth, future work could use more detailed explanation methods. The blur-grid method could be compared with Grad-CAM, occlusion sensitivity, segmentation-based perturbations, or face-region-based analysis. This would help determine whether the observed alignment is robust across explanation methods.

Finally, future research could investigate whether training or fine-tuning with compression augmentation improves both prediction robustness and explanation-level alignment. This would move from evaluating robustness to actively improving it. Importantly, such improvement should be measured not only through classification scores, but also through whether detector evidence remains human-interpretable under degradation.

6.8 Summary

Overall, the discussion shows that the thesis results support the central research argument. Compression makes GenD less confident and less stable in its binary predictions, but model-sensitive regions remain strongly aligned with human-localised artefacts. This means that confidence degradation and explanation-level alignment should be treated as distinct dimensions of robustness. For responsible deepfake detection, it is not enough to ask whether a detector predicts "fake". It is also necessary to ask whether the evidence behind that prediction can be meaningfully inspected and related to human perception.

7 Conclusions and Further Research

This thesis investigated human–model alignment in explainable deepfake video detection under controlled compression degradation. Using GenD and ExDDV human click annotations, I evaluated whether post-hoc blur-grid evidence maps correspond to human-localised deepfake artefacts, and whether this correspondence changes when videos are recompressed.

The first subquestion asked how controlled compression affects GenD's fake-class probability and prediction stability. The results show that compression consistently reduces GenD's fake-class confidence across the included datasets. Stronger compression generally produces a larger decrease, and a non-trivial share of frames changes predicted class after recompression. In plain terms, compression makes the detector less certain and can make it second-guess its final decision.

The second subquestion asked whether human-clicked artefact regions correspond to model-sensitive regions in post-hoc evidence maps. The results show that they do. Human clicks fall substantially above the random percentile baseline and frequently lie within the top-ranked regions of the blur-grid sensitivity maps. The shuffled-click control baseline further supports this conclusion for FaceForensics++ and DeeperForensics, where actual alignment is significantly stronger than expected from the overall spatial distribution of clicks alone.

The third subquestion asked whether this alignment remains stable under stronger compression. The results show that it remains comparatively stable across the baseline, CRF 28, and CRF 35 conditions. The main finding of this thesis is therefore that compression affects GenD's confidence more strongly than it disrupts the spatial structure of its evidence. Even when compression makes the detector less confident in its final prediction, the model-sensitive regions often remain close to the same visual flaws that human annotators noticed.

The answer to the central research question is therefore that GenD's post-hoc evidence maps align strongly with ExDDV human click-localised artefacts, and this alignment remains comparatively robust under controlled compression. This does not mean that GenD and human annotators reason in the same way, nor does it make a post-hoc evidence map equivalent to forensic proof. However, it does show that confidence scores alone provide an incomplete view of detector behaviour. A detector can lose confidence under compression while still preserving human-relevant spatial evidence.

These conclusions should be interpreted with the study's limitations in mind. The DeepfakeDetection/DFDC subset of ExDDV was not included because the corresponding source videos were inaccessible through the available data access route. The BioDeepAV results should be treated as exploratory because this subset contains only 28 annotated frames per condition, and its shuffled-click comparison does not reach statistical significance. The experiments also use one detector, GenD CLIP L 14, and one fixed 7×7 blur-grid explanation method, which provides only coarse

spatial localisation. Finally, the analysis is frame-based and uses controlled H.264 recompression rather than platform-specific social media processing.

Further research should extend the evaluation to the missing DFDC subset, compare multiple detector backbones, and test whether the same alignment patterns hold for more detailed explanation methods such as Grad-CAM, occlusion sensitivity, segmentation-based perturbations, or temporally aware video explanations. Future work could also evaluate platform-specific compression pipelines and investigate whether compression-aware training improves not only prediction robustness, but also explanation-level alignment. Overall, this thesis shows that robust deepfake detection should not only ask whether a model predicts "fake", but also whether the evidence behind that prediction remains inspectable and human-relevant under realistic degradation.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[2] Robert Chesney and Danielle Keats Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6):1753–1820, 2019.

[3] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[4] European Commission. Draft guidelines on the implementation of the transparency obligations for certain ai systems under article 50 of the ai act. European Commission, Shaping Europe's Digital Future, 2026. Published 8 May 2026.

[5] European Union. Regulation (eu) 2022/2065 of the european parliament and of the council on a single market for digital services. Official Journal of the European Union, 2022. Digital Services Act.

[6] European Union. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence. Official Journal of the European Union, 2024. Artificial Intelligence Act.

[7] Vlad Hondru, Eduard Hogea, Darian M. Onchis, and Radu Tudor Ionescu. Exddv: A new dataset for explainable deepfake detection in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

[8] Benedikt Hopf et al. Robust deepfake detection, ntire 2026 challenge: Report. *arXiv preprint arXiv:2604.24163*, 2026.

[9] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020.

[10] Yuhang Lu and Touradj Ebrahimi. Assessment framework for deepfake detection in real-world situations. *arXiv preprint arXiv:2304.06125*, 2023.

[11] Yuhang Lu and Touradj Ebrahimi. Impact of video processing operations in deepfake detection. *arXiv preprint arXiv:2303.17247*, 2023.

[12] Kristof Meding and Christoph Sorge. What constitutes a deep fake? the blurry line between legitimate processing and manipulation under the eu ai act. In *Proceedings of the Symposium on Computer Science and Law*, 2025.

[13] Andrea Montibeller, Dasara Shullani, Daniele Baracchi, Alessandro Piva, and Giulia Boato. Bridging the gap: A framework for real-world video deepfake detection via social network compression emulation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[15] Felipe Romero Moreno. Generative ai and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, 2024.

[16] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[17] Vera Schmitt, Niklas Kruse, Premtim Sahitaj, and Julius Schöning. Transparency as architecture: Structural compliance gaps in eu ai act article 50 ii. *arXiv preprint arXiv:2603.26983*, 2026.

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[19] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[20] Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.