Robust Detection of AI-Generated Claim Images Using Hybrid Ensembles in Motor Insurance

Data Analytics in Applications, WS 2025/2026
Research Institute for Management, Logistics and Production
TUM School of Management

Ufkun Özalp
Technical University Munich
Munich, Germany
ufkun.oezalp@tum.de

Efecan Murat Öksüz
Technical University Munich
Munich, Germany
efecan.oeksuez@tum.de

Nicolas Siebold
Technical University Munich
Munich, Germany
nicolas.seibold@tum.de

Abstract—**This study presents an approach for detecting AI-generated and manipulated images in the context of auto insurance claims. Based on the challenges of our industry partner Allianz, the project focuses on the risk posed by fraudulent damage images and their financial and procedural implications. We started with the creation of a synthetic dataset of manipulated automotive images using a state-of-the-art text-to-image and inpainting model. Further we evaluate several state-of-the-art image detectors using this data. Based on the observation that individual detectors make to many mistakes, we design and evaluate a hybrid detection pipeline that combines multiple models, deciding based on a majority voting principle. The results show that such a hybrid system can improve robustness and detection performance compared to single-model baselines. In addition to the direct business benefits for Allianz, this initiative also contributes to the rapidly growing field of AI image forensics research and demonstrates how ensemble detection architectures can support reliable fraud detection in industrial insurance applications.**

I. Introduction

The steady advancement of artificial intelligence (AI) has paved the way for a new or improved form of fraud. A wave of artificially generated or AI-altered images is currently flooding not only public networks but also government agencies and insurance companies [1]–[3].

While these models are an incredible asset to the creative industries, design workflows, and content production, simultaneously they also bring with them significant problems. These problems are primarily ethical and security risks associated with their often unrestricted use. The images generated by the models are almost impossible to distinguish from real images with the naked eye. This problem has been recognized more in the area of deepfakes, which are mostly images or videos of well-known personalities that are distributed on the internet. Nowadays, one has to look closely to see whether an online report or post is an authentic video or a very good fake [1], [2], [4], [5].

This problem is now spreading from online media to many other areas, one of which is the insurance industry, as mentioned above. This problem has also been recognized by Allianz Insurance Company, our industry partner for this project, and classified as a major risk but also as an area with potential for improvement [1], [3], [6]–[8].

Recent industry reports highlight first documented cases and realistic scenarios in which policyholders or organized fraudsters use AI tools to generate or manipulate claim photos, for example by exaggerating existing damage, fabricating additional impact areas, or reusing images from unrelated incidents [9], [10]. At the same time, specialized vendors are beginning to offer AI-based authenticity checks for claim images, indicating that insurers are actively seeking technical countermeasures to this emerging threat. For large insurance groups processing thousands of claims per day, even a small fraction of successful AI-based fraud attempts can translate into substantial financial losses and increased operational overhead, as suspicious cases require manual review and cross-checking against additional evidence [6]–[9].

In parallel, the regulatory environment in Europe is tightening. The EU Artificial Intelligence Act introduces transparency obligations for certain AI systems and explicit rules for so-called deepfake content, including requirements to clearly disclose synthetically generated or manipulated media in specific contexts [11]. For financial institutions and insurers operating in the EU, these developments add a compliance dimension to the technical and business challenges: they must not only defend against fraud, but also demonstrate appropriate controls, risk management, and transparency in the deployment of AI systems that process potentially synthetic visual evidence [11], [12].

Against this background conditions, there is a growing need for robust, scalable, and domain-specific AI-image detection systems that can reliably distinguish genuine claim images from synthetically altered ones under real-world conditions.

Our work responds to this need by investigating an detection pipeline tailored to motor insurance claims at Allianz, combining multiple detection models in a hybrid architecture to increase robustness and practical applicability. By grounding the technical design in the specific operational requirements and constraints of an industry partner, the project aims to bridge the gap between academic research on AI-image forensics and its deployment in mission-critical industrial insurance workflows.

Our project was structured using the cross-industry standard process for data mining, the CRISP-DM model [13]–[15]. This model helped us keep track of our data, the progress, and the ultimate goal of our work. We also structured our report according to this model: First, we present our business understanding, which forms the basis for the challenge posed by Allianz. We then discuss our understanding of the data before describing our data creation process in more detail in the following chapter. [13]–[15].

II. Business Understanding

In order to deliver a successful project, a solid foundation must first be established so that our partner's requirements can be implemented effectively. To analyze and understand the situation, we first focused on the business level and clarified how AI-generated and manipulated images affect the risk landscape and value creation of one of the biggest insurers worldwide [16]–[19]. As part of the project, we used a business model canvas, which we refined iteratively based on our findings, to establish a framework for our work and align our technical focus with the business findings.

Our understanding of the business case comes primarily from our own research and publicly available industry sources, as internal data from the Allianz Group were not fully accessible due to privacy restrictions. Therefore, we relied on published estimates of claim volumes, average claim costs, and known typologies of image-based insurance fraud to approximate the potential financial impact of AI-generated claim images [20]–[22]. On this basis, we modeled a scenario in which even a small fraction of manipulated or fully synthetic images among thousands of daily claims can accumulate to substantial annual losses and additional investigative effort.

From these analyses, we derived our central objective for the project: to support Allianz in reliably detecting ai-generated and manipulated images, while reducing losses and claim handling time, with open source based models that check authenticity. While working on the model, we came across another area that is of great interest to the Allianz Group, bulk image manipulation through inpainting. To be more precise about the envisioned model, it should flag suspicious images early, integrate into existing claims handling workflows with minimal friction, and provide outputs that are interpretable and actionable for human claims handlers. At the same time, the solution should avoid excessive false positives to not hurt the image and trust in Allinaz.

To provide a brief insight, we have compiled various calculations, which will now be outlined briefly. Firstly, the amount and costs of possible fraud cases. Based on assumptions supported by literature and available data, it can be assumed that up to 10% of claims are fraudulent. We have based our calculation on a conservative estimate of approximately 2%. With an estimate of 2%, the losses for an insurance giant can run into the millions. However, not only can these potential losses be avoided, but the processing time from receipt to settlement of a claim can also be significantly reduced. According to Allianz they have an apprximate claim handling time of about 10 days, recent studies and first tests have shown that this time can be reduced by about 80%. This would not only create a monetary value for Allianz but allso a competitive edge against other competitors [23]–[25].

Based on these conclusions the requirements for the technical work where set. First, the detection approach must be robust against a diverse set of generative models and manipulation techniques, rather than overfitting to a single generator. Second, it must be scalable enough to handle high claim volumes without introducing prohibitive latency. Third, it must operate under realistic data constraints, such as limited access to labeled insurance images and the need to protect customer privacy. Altogether, this business understanding phase provided the boundary conditions within which our data strategy, modeling choices, and evaluation criteria were defined.

III. Data Preparation and Data Understanding

Following the CRISP-DM methodology, this section describes our data sources, assesses their quality, and documents the construction of the benchmark dataset used for evaluating detection models.

A. Data Sources and Selection Rationale

Our benchmark dataset draws on two publicly available sources. For the *real* class, we use 374 images from the test split of the **CarDD** (Car Damage Detection) dataset [26], which contains genuine photographs of damaged vehicles annotated in COCO format with bounding boxes for various damage types such as dents, scratches, and cracks. These images serve as ground truth for legitimate insurance claim photographs.

For the source material used to create manipulated images, we use the **Stanford Cars** dataset [27], which contains 16,185 clean, undamaged car images spanning 196 classes organized by make, model, and year. From this pool, we selected 101 source images using stratified sampling with a fixed random seed (42) for reproducibility. The selection criteria ensured a minimum resolution of 400×300 pixels and coverage of at least 70 unique car makes, providing diversity in vehicle appearance, color, and body shape.

The choice of these two datasets is motivated by the realistic fraud scenario: a fraudster would photograph their undamaged vehicle and use AI tools to add convincing fake damage, then submit the manipulated image as a claim. By pairing real damaged images (CarDD) against real cars with synthetically added damage (Stanford Cars + inpainting), our benchmark directly tests this attack vector.

B. Data Quality Assessment

We assessed both datasets along four data quality dimensions (intrinsic, contextual, representational, and accessibility) widely used in data quality research [**?**] and documented this assessment within the CRISP-DM data understanding phase [14]:

Intrinsic quality. Both datasets originate from established academic sources with consistent labeling and curation.

CarDD images are real photographs with verified damage annotations. Stanford Cars images are correctly classified by make, model, and year. No systematic mislabeling was identified in either dataset.

Contextual quality. A key limitation is that neither dataset contains actual Allianz claims photographs, which were inaccessible. CarDD originates from an Asian automotive context, while Stanford Cars reflects primarily North American vehicle markets. Consequently, the distribution of vehicle types, damage patterns, and photographic conditions in our benchmark may differ from Allianz's real claims pipeline. This limits the external validity of absolute performance numbers, though relative model comparisons remain meaningful.

Representational quality. Images in both datasets vary in resolution, aspect ratio, lighting conditions, and camera angles. This variability is realistic for insurance claims, where customers submit photographs from diverse devices and environments. The minimum resolution filter (400×300) ensures sufficient image detail for the inpainting model to produce convincing manipulations.

Accessibility quality. Both datasets are publicly available under academic licenses and use standard formats (JPEG images, COCO/CSV annotations), enabling straightforward integration into our processing pipeline.

C. Synthetic Dataset Construction

The most critical component of our data preparation is the automated pipeline for generating manipulated images. Rather than testing against fully AI-generated car images we focus on the harder and more realistic scenario of *inpainting-based manipulation*, where real car photographs are altered by adding synthetic damage to specific body panels.

We use **Flux Fill Dev** [28] by Black Forest Labs, an inpainting model based on rectified flow transformers. In our setting, Flux Fill Dev produced photorealistic edits with seamless blending at masked boundaries, representing a challenging manipulation scenario for detection systems. The model was deployed on an NVIDIA RTX 5090 GPU (32 GB VRAM) via ComfyUI, using the official Flux Fill inpaint workflow.

The generation pipeline, illustrated in Fig. 1, consists of five automated stages:

1) **Car detection:** YOLOv8 [29] localizes the vehicle bounding box in each source image.
2) **Panel zone masking:** The bounding box is divided into semantic zones (hood, doors, fenders, bumpers, quarter panel). A random zone is selected, and an irregular blob mask (8–18% of the car bounding box area) with Gaussian-blurred edges is generated within it. The mask is stored as the alpha channel of an RGBA PNG.
3) **Prompt matching:** A damage description prompt is selected from a catalog of 14 templates, matched to the mask location for coherence (e.g., hood masks receive hood-specific prompts such as impact damage, while door masks receive parking-lot door dings).
4) **Inpainting:** The masked image and prompt are submitted to ComfyUI's API. Flux Fill Dev generates the fake damage within the masked region while preserving the surrounding image.
5) **Parameter study:** Each source image is processed at three diffusion step counts (20, 30, and 40 steps), yielding $101 \times 3 = 303$ manipulated images with full provenance metadata (source, mask zone, prompt, seed, step count).

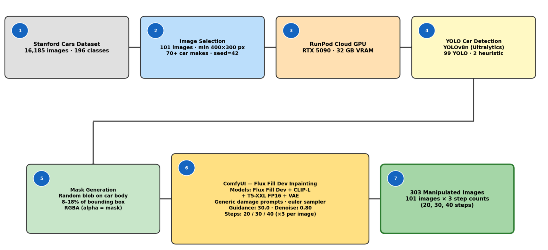

Fig. 1. Automated pipeline for generating manipulated car damage images via Flux Fill Dev inpainting.

Table I summarizes the final benchmark dataset composition.

TABLE I
BENCHMARK DATASET COMPOSITION

Category	Source	Count
Real (damaged)	CarDD test set	374
Manipulated	Stanford Cars + Flux Fill	303
– 20 steps		101
– 30 steps		101
– 40 steps		101
Total		**677**

D. Preprocessing and Evaluation Protocol

All detection models are evaluated in *zero-shot* mode without any fine-tuning on our dataset. This design choice reflects the intended production scenario: an off-the-shelf detector must generalize to unseen manipulation techniques without requiring retraining on domain-specific data. Consequently, no train/validation/test split is applied; the full 677-image dataset serves as the evaluation set.

Each model applies its own internal preprocessing (e.g., CLIP-based models use 224×224 center crops; SigLIP-based models use their respective image processors). No dataset-level resizing, normalization, or augmentation is performed, ensuring that models are tested under realistic conditions. A fixed decision threshold of 0.5 is used for binary classification: images scoring ≥ 0.5 are classified as fake, and those below as real.

IV. MODELING

This section describes our detection model selection, the rationale behind each choice, and the design of hybrid ensemble strategies. All quantitative results are presented in Section V.

A. Model Selection and Architecture

To evaluate whether current AI-image detectors can identify inpainting-based manipulations, we selected six open-source models spanning three distinct architectural families. The goal was to test diverse detection approaches rather than optimizing a single architecture, reflecting the uncertainty about which features (frequency artifacts, semantic inconsistencies, or statistical patterns) are most informative for detecting Flux-generated manipulations.

Table II provides an overview of the selected models.

TABLE II
DETECTION MODELS: ARCHITECTURE AND PROVENANCE

Model	Backbone	Source	Original Task
Umm_Maybe	ViT	HF	AI image det.
Ateeqq	SigLIP	HF	AI vs. human
Deepfake_v1	SigLIP	HF	Face deepfake
DeepFake_v2	ViT	HF	Face deepfake
UniversalFake-Detect	CLIP ViT-L/14	CVPR'23	Cross-gen. det.
CLIP_GripUnina	open_clip	CVPR'24W	Synthetic det.

HF = HuggingFace community model

ViT-based models. Umm_Maybe is a Vision Transformer fine-tuned for binary AI-image classification. DeepFake_v2 uses a similar ViT backbone but was originally trained for face deepfake detection. Both models output a direct probability score via softmax.

SigLIP-based models. Ateeqq and Deepfake_v1 use SigLIP (Sigmoid Loss for Language-Image Pre-training) [30] as their backbone. Ateeqq produces highly polarized probability distributionswhich contributes to its characteristically low false positive rate. Deepfake_v1 was included to test whether face-deepfake detectors transfer to non-face manipulation scenarios.

CLIP-based models. UniversalFakeDetect [31] uses a CLIP ViT-L/14 backbone [32] with a trained linear classifier head, following the approach of Ojha et al. for cross-generator generalization. It was specifically designed to detect manipulations from unseen generative models, making it theoretically well-suited for our Flux-based test set. CLIP_GripUnina [33] is a CLIP-based detector from Cozzolino et al. In our inference pipeline, its raw output is treated as a Log-Likelihood Ratio (LLR) and mapped with a sigmoid for probability scoring.

Scoring verification. During initial benchmarking, we discovered that two academic models (UniversalFakeDetect and CLIP_GripUnina) produced inverted classification scores due to incorrect sigmoid polarity in their published inference code. We developed a dedicated verification procedure that tests each model's scoring direction against a small set of known real and fake images before running the full benchmark. After correction, UniversalFakeDetect's detection rate increased from near-zero to over 90%, confirming the importance of this validation step.

B. Ensemble Detection Pipeline

A central finding from our individual model analysis is that no single detector achieves both high detection rates and low false positives simultaneously. However, the models exhibit *complementary error patterns*: models that excel at detecting manipulations tend to produce more false positives, while conservative models miss more fraud but rarely flag legitimate images. This observation motivates combining multiple detectors into ensemble strategies [34].

We designed and tested nine ensemble strategies using the three strongest individual models - Ateeqq, Umm_Maybe, and UniversalFakeDetect - as the ensemble core. DeepFake_v2 was excluded due to an unusable 92% false positive rate, and Deepfake_v1 and CLIP_GripUnina offered no complementary advantage over the top three.

The strategies fall into four categories:

Staged pipeline. A sequential architecture where Ateeqq serves as a high-precision first filter (when it flags an image as fake, it is almost certainly correct). Images that pass the first stage are evaluated by a dual-model consensus of Umm_Maybe and UniversalFakeDetect, both of which must agree before flagging. Disagreements are routed to a manual review queue. This design minimizes false positives at each stage.

Majority voting. Two variants were tested: a 2-of-3 vote (at least two models must flag an image) and a 3-of-4 vote (using all three core models plus Deepfake_v1). Each model casts a binary vote based on the 0.5 threshold.

Weighted score averaging. The continuous probability scores from the three core models are combined as a weighted average and compared against a 0.5 threshold. Four weighting schemes were tested: equal weights, AUC-proportional, precision-proportional, and recall-proportional.

Threshold-optimized ensemble. Per-model decision thresholds are individually optimized to maximize each model's Youden's J statistic (sensitivity + specificity -1), followed by a majority vote on the optimized binary predictions. This allows each model to operate at its own optimal operating point rather than sharing a uniform threshold.

Additionally, a **confidence cascade** strategy was tested, which uses high-confidence gates at both extremes (very high scores trigger immediate flagging, very low scores trigger immediate clearance) with a weighted middle zone for ambiguous cases.

Fig. 2 illustrates the staged pipeline architecture, which represents the most interpretable and deployment-ready design among the strategies tested.

C. Inference Protocol

All models are used in zero-shot mode without fine-tuning, simulating production deployment where detectors must generalize to unseen manipulation techniques. Each image is scored independently by all six models; no batch normalization or cross-image calibration is applied. For individual model evaluation, the fixed threshold of 0.5 is used uniformly. For ensemble strategies, per-model optimized thresholds are additionally tested as described above. All scoring is deterministic:

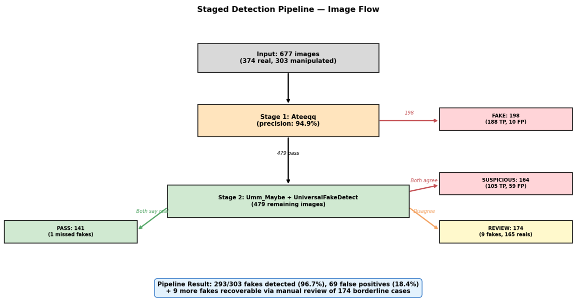
Staged Detection Pipeline — Image Flow
Input: 677 images
(374 real, 303 manipulated)
Stage 1: Ateeqq
(precision: 94.9%)
198
FAKE: 198
(188 TP, 10 FP)
479 pass
Stage 2: Umm_Maybe + UniversalFakeDetect
(479 remaining images)
Both agree
SUSPICIOUS: 164
(105 TP, 59 FP)
Both say real
PASS: 141
(1 missed fakes)
Disagree
REVIEW: 174
(9 fakes, 165 reals)
Pipeline Result: 293/303 fakes detected (96.7%), 69 false positives (18.4%)
+ 9 more fakes recoverable via manual review of 174 borderline cases

Fig. 2. Staged ensemble detection pipeline: Ateeqq pre-filter followed by dual-model consensus and manual review queue.

given the same image and model, the same score is produced. The full set of comparative results is presented in Section V.

V. Evaluation

In the following section, we refer to the findings we have obtained through our tests and analyses. Our test is based on images that were created using the model described above. The dataset consists of 677 images, of which 374 are real images from the carDD dataset and 303 are images that were generated by inpainting. With this, we extracted the best-performing models and detection pipelines as mentioned in the section above. First, all relevant calculations are briefly outlined. For the analysis we used different metrics including: Accuracy, Precision, Recall, F1, AUC, ROC, False-negative rate and False-Positive rate [35]–[38]. In the following i will briefly explain the different measurements:

Accuracy displays how correct the models decisions are overall while Precision shows how often the model is correct when predicting the target class (in our case AI-images). Recall displays how many of the AI pictures we actually found while the The F1 score describes the harmonic mean of Precision and Recall. When speaking about false positives (FP) we talk about images that are authentic but are classified as AI, while false negatives (FN) are images that are generated or manipulated by AI but where classified as authentic. The ROC curve shows the relationship between the true positive rate and the false positive rate for varying threshold values. The AUC summarizes the ROC curve as a threshold-independent measure of discrimination (0.5 = random, 1.0 = perfect) [35]–[38].

A. Individual models

We first conducted an individual analysis where we performed the same test for of each detector in isolation. Figure 4 represents the percentages of the samples predicted as fake, for the CarDD images that corresponds to FP, whereas for Fake_Manipuleted it reflects the TP rate. The results are mixed there is a clear sensitivity trade off, some models have a high detection rate for fakes but at the cost of increased a FP rate.

The Box-plots in 3 show the distribution of predicted fake scores for Real (CarDD) and Fake (Manipulated) samples for each model. The red dashed line indicates the fixed decision threshold used for converting scores into binary predictions (fake vs. real). Here we can indicate, that some models are conservative (low FP, higher FN) vs others sensitive (high TPR, higher FP).

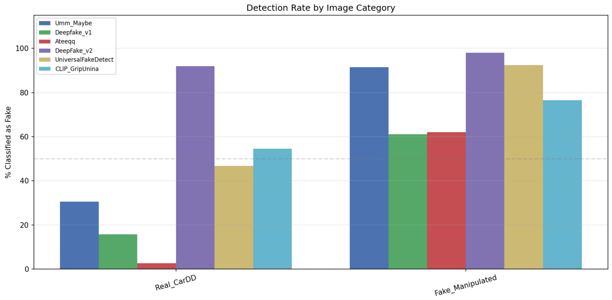

Fig. 3. Score distributions (boxplots)

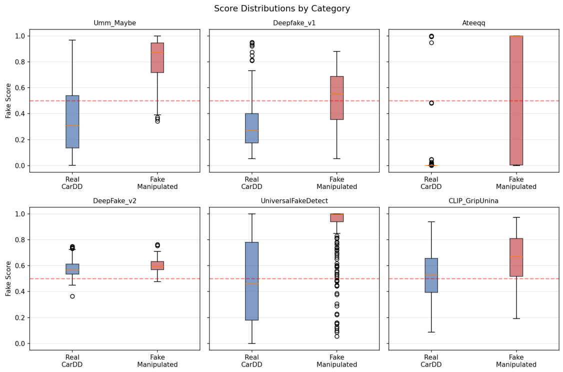

Fig. 4. Detection Rate by Image Category

Combining the test we did on the individual level we visualized using the metrics of ROC and AUC in Figure 5. The curves illustrate the trade-off between true positives and false negatives: to achieve many true positives, the models accept a relatively high number of false negatives, which results in an average AUC of 0.801.

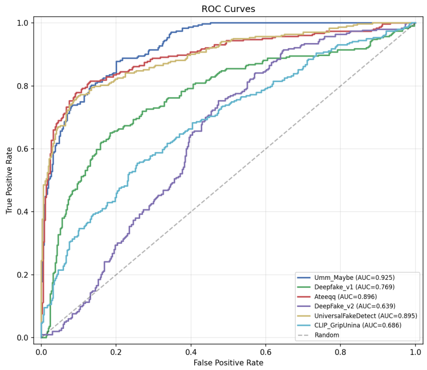

Fig. 5. Individual ROC curves

While conducting the tests, we discovered that some detectors are particularly good at detecting real images, while others are superior at detecting AI-altered and generated images. As mentioned in the section above, we concluded that combining the different detectors would be the most efficient and promising approach for reliable detection. One concern had about the method was, to make sure, the models would not make the same mistakes. In other words, one model (good with real images) says it's not real, and the other model (good with fake images) says it's fake when it's actually a real image. Using the most promising detectors, for detecting fakes and/or real images, we archived a small number of shared errors across all core models in Figure 6.

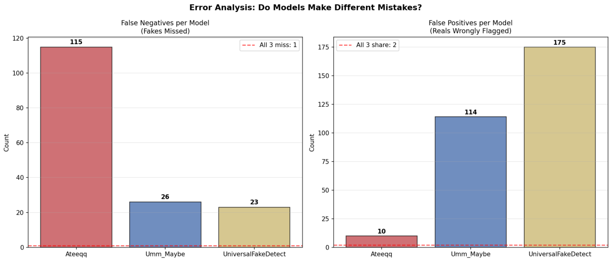

Fig. 6. Error overlap

We conducted further tests in addition to this experiment, these will not be deeply discussed here. One of which tested the different damage types. The information we could gathered was used for our image generation models in order to obtain the most convincing images possible.

B. Ensemble models

Based on the findings mentioned above, the models were then combined into various hybrid models. In order to ensure a model that was as accurate as possible, we tested various combinations and approaches. The Figure 7 shows the performance of the combination models in comparison to the individual detectors. It is evident that some individual models perform similarly well in certain areas, but perform significantly worse in the overall comparison.

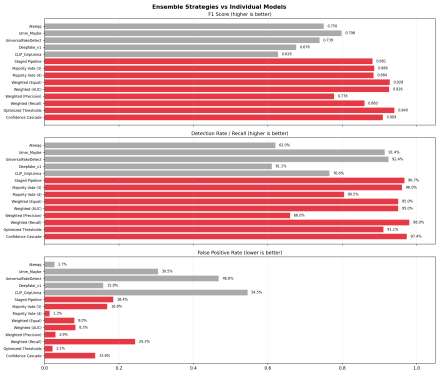

Fig. 7. Ensemble Strategies vs Individual Models

In our context at Allianz, the test results show three things in particular that we consider important.

Firstly, AI-generated and manipulated damage images can be detected with a high degree of accuracy, meaning that automated verification can directly contribute to reducing successful fraud cases [2], [3]. Second, hybrid and ensemble models offer a better balance between precision and recall than individual detectors, which means that fewer fraud attempts are overlooked (FN) without too many FP or too much investigation [3] Third, a scalable detection pipeline can be realistically integrated into existing claims processing workflows, supporting end-to-end processing of genuine claims and targeted investigation of suspicious cases, thereby strengthening Allianz's strategic position in digital fraud management and compliance [16], [25].

In Figure 8 you can see the visualization of the model we selected. With optimized threshold majority voting (Ateeqq = 0.05, Umm_Maybe = 0.58, UFD = 0.93), 276/303 fakes are detected (91.1% TPR) with 8/374 false positives (2.1% FPR). Overall, this results in an accuracy of 94.8% and an F1 score of 0.9404 (TP=276, FP=8, TN=366, FN=27).

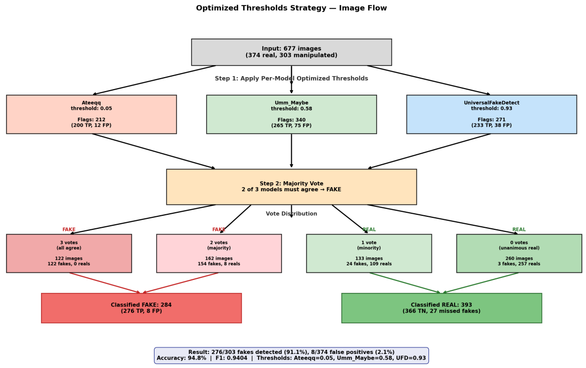

Fig. 8. Ensemble Flow

Now that we have decided on a model, we have considered how best to present it. In the following section, we describe our deployment.

VI. Deployment

A. System Architecture of FraudLens

FraudLens is deployed as a multi-layer AI application for claim-image fraud analysis. The architecture includes a web frontend, an API backend, a persistence layer, and a hybrid AI inference core.

Frontend Layer: The frontend provides adjuster-facing workflows for image upload, model selection, result review, and feedback capture. It orchestrates image submission and result rendering while preserving interactive response times.

Backend Layer: The backend exposes REST APIs for user/session state, gallery/history persistence, image storage, and exact-model inference. In the current implementation, exact detector execution is exposed through `/api/detect-exact`, which delegates inference to a Python worker process.

Database Layer: MongoDB is used for structured metadata (users, claim analysis history, settings) and image payload references. The deployment supports cloud and local storage modes, enabling controlled data locality per user profile.

AI Core: The AI core is hybrid:

1) Browser-local models for low-latency screening,
2) Backend-managed exact detectors for model-faithful execution,
3) External AI services for additional model coverage.

This structure follows practical MLOps deployment patterns for heterogeneous serving [44]–[46].

Allianz Integration Points (Target Enterprise Deployment): The integration points are:

1) Claim intake/FNOL systems for image and claim context ingestion,
2) Core claims platform for embedding fraud scores and explanations into adjuster worklists,
3) Identity and access systems (SSO, RBAC) for controlled analyst access,
4) SIU/case-management tools for escalation of high-risk claims,
5) Governance/model-risk systems for model approval and audit export.

These are deployment interfaces for enterprise operation and governance, not public API changes.

B. Operational Pipeline

The end-to-end claim-image processing pipeline is as follows:

1) An adjuster uploads a claim image from the frontend.
2) The frontend normalizes and submits the image to the selected detection pipeline.
3) The detection router dispatches to local, exact backend, or external detectors based on model ID.
4) For exact detectors, the backend invokes a Python worker, which loads/caches model artifacts and returns label, confidence, and class predictions.
5) The backend stores image references and analysis results, applying hash-based deduplication to avoid redundant binary storage.
6) The frontend presents explanations, confidence scores, and latency to the adjuster.
7) Adjuster feedback is recorded and can be used for quality monitoring and future model governance.

C. Non-Functional Requirements

Scalability: API services should be horizontally scalable, while heavyweight exact-model execution should run in isolated worker pools with queue-based back-pressure [44].

Latency: Interactive claim triage requires bounded inference latency. Caching, asynchronous heavy-model execution, and parallel detector calls are needed to stabilize tail latency.

Monitoring and Logging: Production monitoring should track model latency, timeout/failure rates, detector drift indicators, and end-to-end pipeline health. Audit logging must preserve user, claim, and model-version traceability [43], [46].

Model Updates: Model lifecycle management should include versioned release, offline validation, shadow/canary deployment, and rollback on performance regression [44], [45].

D. Deployment Architecture Diagram

Figure 9 presents the end-to-end deployment flow of FraudLens, from Allianz claim intake channels to backend orchestration, detector execution paths, persistence layers, and human-in-the-loop feedback.

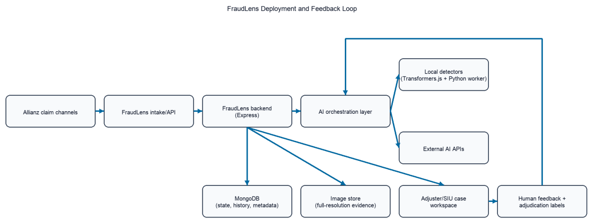
FraudLens Deployment and Feedback Loop
Allianz claim channels
FraudLens intake/API
FraudLens backend (Express)
AI orchestration layer
Local detectors (Transformers.js + Python worker)
External AI APIs
MongoDB (state, history, metadata)
Image store (full-resolution evidence)
Adjuster/SIU case workspace
Human feedback + adjudication labels

Fig. 9. FraudLens deployment and feedback loop across intake, orchestration, storage, and human adjudication.

E. Compliance and Governance

EU AI Act: The AI Act (Regulation (EU) 2024/1689) entered into force on August 1, 2024, with phased applicability. FraudLens deployments should implement transparency and human oversight controls relevant to synthetic/deepfake media analysis and insurance decision support [39], [40].

Data Sovereignty and Security: Customer claim images require strict confidentiality and controlled residency. Deployment should enforce encryption in transit/at rest, regional data processing boundaries, access minimization, and auditable processing records to align with GDPR principles [41].

Financial-Sector AI Governance: Insurance deployment should apply proportional risk controls, accountability structures, and third-party model governance consistent with sector guidance [42].

VII. Conclusion and Outlook

In this work, we set ourselves the goal of investigating how AI-generated and manipulated damage images can be detected in a realistic insurance context. We wanted to apply these findings to our partner Allianz and tailor them to their needs. We created a synthetic dataset with images of vehicle damage by editing real vehicle photos with inpainting-based manipulations. This allowed us to simulate and test possible fraud attempts. On this basis, we evaluated several detectors and showed that although each model has different strengths and weaknesses, combining several models brings about an improvement.

In the context of Allianz, our results show that automated verification of damage patterns for AI manipulation is fundamentally possible and applicable, but depends heavily on the selected operating point (threshold) and the error cost assessment. Individual detectors either deliver a high detection rate for manipulated images or a high detection rate for genuine images, but rarely achieve both at the same time. Combining several models in a majority vote ensemble with optimized thresholds can significantly mitigate this trade-off: In our setup, 276/303 manipulated images are detected (91.1% TPR) with 8/374 genuine images incorrectly marked (2.1% FPR), which corresponds to an accuracy of 94.8% and F1 = 0.9404. For Allianz, this means that the approach is designed so that only cases above a defined risk threshold are escalated, while unremarkable cases continue without friction enabling a human-in-the-loop setup rather than fully automated rejection.

A look into the future reveals several possibilities for how we can further develop the project. First of all, the entire project should be tested with a larger, alliance-specific damage dataset. This dataset can focus more on different markets, vehicle types, and recording conditions in order to confirm its robustness under real-world fluctuations. With our proposed image processing approach, Allianz could create the user-defined model. In addition, the ensemble can be expanded to include additional sources of information such as text descriptions of claims, metadata, and provenance signals from images. Finally, introduction into the claim handling process at Allianz requires continuous monitoring and regular recalibration of thresholds to account for new generators and evolving fraud patterns. This has to be done while being aligned with the EU AI Act and internal model risk policies. By combining these measures, the current prototype is transformed into a scalable and business-critical component within the digital fraud management strategy of Allianz.

Acknowledgment

We would like to thank Allianz for their support and the weekly calls, which were an essential part of our understanding of the problem and the requirements. We greatly appreciate the help and can say with conviction that we have grown through the task at hand and the intensive monitoring. We would also like to express our sincere thanks to Benedikt Blümelhuber and Sebastian Junker for their support, who assisted our work with helpful advice and interesting lectures.

References

[1] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, T. S. Bonaventura, V. Bruni, R. Caldelli, F. G. B. Natale, R. Nicola, L. Guarnera, S. Mandelli, G. Marcialis, M. Micheletto, A. Montibeller, G. Orrù, A. Ortis, P. Perazzo, D. Salvi, and D. Vitulano, "Deepfake Media Forensics: State of the Art and Challenges Ahead," 10.48550/arXiv.2408.00388, August 2024.

[2] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media," PeerJ Comput. Sci., vol. 10, p. e2037, May 2024.

[3] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera, S. Mandelli, T. Majid, G. L. Marcialis, M. Micheletto, A. Montibeller, G. Orrù, A. Ortis, P. Perazzo, G. Puglisi, and D. Vitulano, "Deepfake Media Forensics: Status and Future Challenges," J. Imaging, vol. 11, no. 3, p. 73, March 2025.

[4] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media," PeerJ Comput. Sci., vol. 10, p. e2037, May 2024.

[5] G. Bendiab, H. Haiouni, I. Moulas, and S. Shiaeles, "Deepfakes in digital media forensics: Generation, AI-based detection and challenges," J. Inf. Secur. Appl., vol. 88, p. 103935, February 2025.

[6] A. Patzer, "Insurance fraud through AI generated images: Challenge and solutions for the insurance industry," Vaarhaft, August 21, 2024. [Online]. Available: https://www.vaarhaft.com/post/insurance-fraud-through-ai-generated-images-challenge-and-solutions-for-the-insurance-industry

[7] Salviol Global Analytics, "Insurance Digital Transformation and AI Fraud Detection: Key Insights from insurance events," Salviol, December 19, 2025. [Online]. Available: https://www.salviol.com/post/insurance-digital-transformation-and-ai-fraud-detection

[8] Salviol Global Analytics, "Is That Claim Photo Real? How AI Image Fraud Detection Is Changing Insurance," Salviol, June 24, 2025. [Online]. Available: https://www.salviol.com/post/ai-image-fraud-detection-insurance

[9] B. Garlick, "Allianz turns to AI to tackle huge spike in insurance fraud," Insurance Business Mag., September 22, 2025. [Online]. Available:https://www.insurancebusinessmag.com/uk/news/breaking-news/allianz-turns-to-ai-to-tackle-huge-spike-in-insurance-fraud-550492.aspx

[10] Allianz Trade, "Deepfake fraud: How does AI make it easier for fraudsters?," Allianz Trade, 2025. [Online]. Available: https://www.allianz-trade.com/en_BE/news/fraud/deepfake-fraud-ai.html

[11] European Parliament and Council of the European Union, "Article 50: Transparency obligations for providers and users of certain AI systems," Regulation (EU) 2024/1689 (Artificial Intelligence Act), July 12, 2024. [Online]. Available: https://artificialintelligenceact.eu/article/50

[12] European Commission, "Regulatory framework for AI," European Commission, August 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[13] I. Amerini, G. Cassano, R. Caldelli, A. Ortis, F. Stanco, and S. Battiato, "Deepfake Detection: A Study on Human Appearance and Scene Context," Procedia Comput. Sci., vol. 192, pp. 4442–4451, 2021 [25th Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst., pp. 4442–4451, 2021].

[14] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc., pp. 1–78, 2000 [Technical Report, CRISP-DM Consortium].

[15] P. Larrañaga, D. Atienza, J. Diaz-Rozo, A. Ogbechie, C. E. Puerto-Santana, and C. Bielza, "Industrial Applications of Machine Learning," 1st ed., CRC Press, 2018.

[16] S. Philip, "2025: The year US P&C insurers must modernize fraud detection–here's why," Shift Technology, September 5, 2025. [Online]. Available: https://www.shift-technology.com/resources/reports-and-insights/modernize-fraud-detection

[17] K. Kamalapurkar, N. Sharma, and M. Canaan, "Property and casualty carriers can win the fight against insurance fraud," Deloitte Insights, April 24, 2025 [FSI Predictions 2025]. [Online]. Available: https://www.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2025/ai-to-fight-insurance-fraud.html

[18] A. Kilroy, "Insurance Fraud Statistics 2026," Forbes Advisor, January 2, 2026. [Online]. Available: https://www.forbes.com/advisor/insurance/fraud-statistics/

[19] M. Smith, "Synthetic Identity Fraud: The next generation of identity theft," J. Insur. Fraud Am., vol. 14, no. 2, pp. 12–17, 2023 [Coalition Against Insurance Fraud]. [Online]. Available: https://insurancefraud.org/publications/jifa-synthetic-fraud/

[20] Innoveo, "How GANs and AI-Generated Images are Transforming Insurance Fraud Detection," Innoveo, December 14, 2023. [Online]. Available: https://innoveo.com/blog/how-gans-and-ai-generated-images-are-transforming-insurance-fraud-detection/

[21] Debevoise & Plimpton LLP, "Debevoise Discusses Use of AI-Generated Images for Fake Insurance Claims and Other Frauds," The CLS Blue Sky Blog, January 28, 2026 [Columbia Law School's Blog on Corporations and the Capital Markets]. [Online]. Available:https://clsbluesky.law.columbia.edu/2026/01/28/debevoise-discusses-use-of-ai-generated-images-for-fake-insurance-claims-and-other-frauds/

[22] SAS Institute, "Can you spot the fake claim? New images show how convincing AI-generated insurance fraud has become," SAS, February 11, 2026. [Online]. Available: https://www.sas.com/en_gb/news/press-releases/2026/february/can-you-spot-the-fake-claim-new-images-show-how-convincing-ai-generated-insurance-fraud-has-become.html

[23] M. Sweitzer, “AI in Insurance Claims for Faster Processing and Increased Accuracy,” Shift Technology, January 15, 2025. [Online]. Available: https://www.shift-technology.com/resources/reports-and-insights/ai-in-insurance-claims-for-faster-processing-and-increase-accuracy

[24] Insurance Asia, “Allianz deploys AI agents for motor and health claims,” Insurance Asia, February 12, 2026. [Online]. Available: https://insuranceasia.com/insurance/news/allianz-deploys-ai-agents-motor-health-claims

[25] Allianz SE, “When the storm clears, so should the claim queue,” Allianz Media Center, November 3, 2025. [Online]. Available:https://www.allianz.com/en/mediacenter/news/articles/251103-when-the-storm-clears-so-should-the-claim-queue.html

[26] X. Wang, W. Li, and Z. Wu, “CarDD: A New Dataset for Vision-Based Car Damage Detection,” IEEE Trans. Intell. Transp. Syst., vol. 24, no. 7, pp. 7202–7214, Jul. 2023, doi: 10.1109/TITS.2023.3258480.

[27] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D Object Representations for Fine-Grained Categorization,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), 2013, pp. 554–561, doi: 10.1109/ICCVW.2013.77.

[28] Black Forest Labs, “FLUX.1 Fill [dev],” 2024. [Online]. Available: https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev

[29] G. Jocher, J. Qiu, and A. Chaurasia, “Ultralytics YOLO,” Version 8.0.0, Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[30] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid Loss for Language Image Pre-Training,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 11975–11986.

[31] U. Ojha, Y. Li, and Y. J. Lee, “Towards Universal Fake Image Detectors that Generalize Across Generative Models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 24480–24489.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in Proc. 38th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 139, pp. 8748–8763, 2021.

[33] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, “Raising the Bar of AI-generated Image Detection with CLIP,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2024, pp. 4356–4366.

[34] Z.-H. Zhou, “Ensemble Methods: Foundations and Algorithms,” 1st ed., Chapman and Hall/CRC, 2012.

[35] D. Mahendris, “Auto-Insurance-Fraud-Detection” GitHub repository, October 2023. [Online]. Available: https://github.com/DandiMahendris/Auto-Insurance-Fraud-Detection/blob/main/01_data_pipeline.ipynb

[36] 26Klu, “Accuracy, Precision, Recall, and F1 Score,” Klu Technical Glossary, 2024. [Online]. Available: https://klu.ai/glossary/accuracy-precision-recall-f1

[37] S. Galli, “AUC-ROC Analysis: What is it and how to use it for model evaluation,” Train in Data Blog, October 10, 2024. [Online]. Available: https://www.blog.trainindata.com/auc-roc-analysis/

[38] S. Narkhede, “Understanding ROC Curves and the AUC-ROC Score,” Built In, June 15, 2024. [Online]. Available: https://builtin.com/data-science/roc-curves-auc

[39] European Union, “Regulation (EU) 2024/1689 (Artificial Intelligence Act),” 2024. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[40] European Commission, “Regulatory framework proposal on artificial intelligence.” Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[41] European Union, “Regulation (EU) 2016/679 (GDPR),” 2016. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[42] EIOPA, “Opinion on Artificial Intelligence Governance and Risk Management,” 2021. Available: https://www.eiopa.europa.eu/publications/opinion-artificial-intelligence-governance-and-risk-management_en

[43] NIST, “AI Risk Management Framework (AI RMF 1.0),” 2023. Available: https://www.nist.gov/itl/ai-risk-management-framework

[44] Google Cloud, “MLOps: Continuous delivery and automation pipelines in machine learning.” Available: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

[45] M. Zinkevich, “Rules of Machine Learning.” Available: https://developers.google.com/machine-learning/guides/rules-of-ml

[46] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J. Crespo, and D. Dennison, “Hidden Technical Debt in Machine Learning Systems,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Declaration of Authorship

We hereby declare that we have authored this thesis independently and have not used any sources or aids other than those explicitly indicated. All passages, whether quoted verbatim or paraphrased from published or unpublished sources, have been clearly identified and properly cited in accordance with academic standards. This work has not been submitted, either in its current form or in any similar version, to any other examination board or for any other course credit.

We hereby grant the examiners and the associated institution the right to use this work exclusively for evaluation and grading purposes within the framework of the course ”Data Analytics in Applications”. Any further use, reproduction, distribution, or publication, in whole or in part, requires our explicit prior written consent. All copyrights remain with the authors.