instinctools

Computer Vision Deepfake ID Detection For a Dutch Digital Bank

How adding a custom computer vision layer to the existing SaaS ID verification platform empowered a Dutch digital bank to keep up with the evolving threat landscape and save **30%** more AI-generated identity fraud, while reducing false positives for real users by **60%** and decreasing manual document review workload by **40%**.

Business challenge

What are the most common types of eID / digital identity fraud you experience?

#1 Deepfakes

#2 ID document forgery

#3 Account takeover and Synthetic ID fraud

How many successful fraud attempts use AI?

When GenAI emerged as a [game-tech capability](), businesses weren't the only ones who went for it. So did cybercriminals who used the technology to create highly realistic deepfakes and pass Know Your Customer (KYC) onboarding at digital banks. Since the mass adoption of artificial intelligence, the number of AI-generated identity fraud cases in fintech has increased by 233%. Now, 1 in 15 attacks involves deepfakes, and 29% of them succeed.

Our client, a **Netherlands-based digital bank serving customers across the European Economic Area**, chose to be one step ahead of fraudsters. As the bank onboards 4–5 million new users annually, they must protect real customers from account takeover, not to mention safeguarding brand trust in a highly competitive digital banking market.

At the time, the client **relied on a third-party SaaS identity verification (IDV) platform** with OCR and NFC/RFID technologies at its core, which fell short on several fronts:

01 **Failures in detecting AI-generated identity documents.** Passports with AI-altered images or completely fake documents could pass through the checks if they visually match the templates of authentic IDs.

02 **Slow speed of adaptation to fraud patterns.** The SaaS IDV was tied to the vendor's release cycles, meaning weeks of waiting for new rejected fraud cases to be updated in the system's detection logic, while the new attack patterns kept snowballing.

03 **Limited control over the deepfake detection mechanism.** The client's team didn't have direct access to the underlying algorithm and couldn't strengthen the SaaS IDV's forgery detection and analysis by training it on their own fraud samples.

For a bank operating at a multi-million-user scale, the gap between **their needs and software capabilities posed significant financial and reputational risk.** They needed a [top AI-powered system expertise in Finance]() to address the issue.

Solution

Instinctools' [dedicated team]() looked into ways to strengthen the client's identity verification algorithm. Investing in a **custom computer vision-based deepfake detection to run alongside the SaaS IDV proved to be the most beneficial option, as it enabled huffties** and retraining on confirmed internal fraud cases within days.

Our [computer vision]() experts took full ownership of the system's design, training, calibration, and production rollout.

01 Assembling a fraud-realistic dataset

The effectiveness of the custom CV detection mechanism hinged on the quality of the training data. Generic examples of ID fraud, that the off-the-shelf IDV platform had covered, were not enough. Therefore, our team prioritized assembling a solid dataset.

- **Doubling down** on [data preparation]() (exploration, cross-source collection, cleansing) to reduce training time.
- **Defining attack taxonomy** that reflected the most relevant fraud patterns, from partial manipulations of photos, textual and numerical data to fully AI-generated documents.
- **Building a dataset** that included the fake IDs the staff caught during KYC checks, partly and completely synthetic documents, and legitimate but low-quality photos and scans of passports.
- **Increasing the share of hard negatives** in the dataset through data augmentation. Valid documents with blur, glare, or perspective distortion often triggered false alarms in the SaaS IDV platform, and we wanted to avoid this issue by training the custom CV algorithm to accurately distinguish poor capture quality from actual manipulation.
- **Labeling the data** based on different criteria (real document vs. fake, fraud location, etc.).

02 Designing an ID-focused CV pipeline

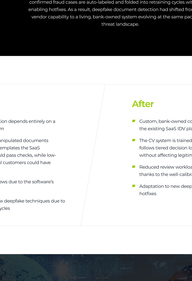
What is YOLO?

YOLO (You Only Look Once) is a computer vision object detection model. It's an example of a leading-edge single-stage detector, known for its ability to locate and classify objects in a single pass through the network and, therefore, is go-to option for real-time object detection.

Our team built the solution around a **single, well-controlled YOLO-based computer vision pipeline,** tailored for identity document analysis. The model owns at:

- Detecting the presence of a document
- Localizing and analyzing in a single pass critical document regions, such as the data page, portrait area, and MRZ code
- Highlighting suspicious area(s) of the document
- Providing a risk score for each problematic element to simplify and speed up the review process

This approach ensured consistency while keeping inference fast enough for real-time KYC flows.



03 Ensuring the detection is reliable beyond known fraud patterns

To avoid building a system that wrote only for already known fraud tricks, we focused on how well the solution would hold up when attackers change their techniques and tools. Therefore, we deliberately checked it against document manipulations and generation methods it hadn't seen before.

The test dataset included:

- Data created by the same top 10 deepfake generation tools we used when assembling the training dataset, but with different faces and parameters
- Data generated by other tools that weren't used for creating the training dataset

This approach ensured the **detection logic wasn't tied to a specific deepfake tool or editing technique, but rather to the visual inconsistencies that tend to appear across AI-generated and manipulated documents,** so the custom CV system remains effective even if fraudsters switch to new image generators.

04 Decision threshold calibration

One of the key challenges with the SaaS IDV platform was the sensitivity of its fraud calibrator, which sometimes caused it to reject legitimate customers due to minor issues (like glare in a photo) while overlooking more serious fraud attempts. To address this problem, we linked the decision thresholds to the risk scores.

The custom CV layer now follows a **tiered decision logic:**

- **Low-risk** cases continue through onboarding automatically, without any additional steps
- **Middle-risk** cases trigger targeted step-ups, such as an NFC scan, requesting an extra selfie or a short video
- **High-risk** cases are flagged for manual review or outright rejection

This approach helped the client strengthen fraud controls and decrease the workload of human reviewers while ensuring fast onboarding for legitimate users.

05 Seamless integration of the CV layer into the client's KYC flow

To keep the onboarding experience unchanged for customers, we deployed the custom computer vision system as a decision microservice within the existing onboarding pipeline.

Still, before letting it influence real-onboarding decisions, the team **ran the custom CV system in observation mode.** At this stage, it processed live KYC traffic alongside the existing systems, but its outputs were not used to approve or reject customers.

Once we validated that the model behaved as expected with real-world data, we rolled it out gradually by feeding it controlled portions of traffic, expanding coverage step by step. This phased approach ensured that the new detection layer strengthened fraud controls without unintended convey-errors, such as over-flagging legitimate users or missing edge cases.

Here's how the client's updated KYC flow works:

01 A user submits their documents (and, if required, a selfie or short video)

02 The existing SaaS IDV platform performs its standard checks, such as OCR, MRZ reading, and NFC validation

03 The custom computer vision system analyzes the same inputs in parallel

04 The higher-level KYC orchestrator then collects and combines the results from both systems and applies the bank's decision logic to determine the next action (approval, step-up verification, or manual review)

Since all checks run behind the scenes and in parallel, **customer experience stays the same, with no extra waiting time or redundant requests.** Meanwhile, under the hood, confirmed fraud cases are auto-labeled and fed back into retraining cycles within days, enabling huffties. As a result, deepfake document detection had shifted from a static vendor capability to a living, bank-owned system evolving at the same pace as the threat landscape.

Before

- Deepfake document detection depends entirely on a third-party SaaS IDV platform
- AI-generated or partially manipulated documents highly resembling valid ID templates the SaaS platform was trained on could pass checks, while low-quality documents from real customers could have been rejected as synthetic
- Time-intensive manual reviews due to the software's unreliability
- Weeks-long response to new deepfake techniques due to vendor-controlled update cycles

After

- Custom, bank-owned computer vision layer added on top of the existing SaaS IDV platform
- The CV system is trained on a high-quality dataset and follows tiered decision logic to perfect deepfake detection without affecting legitimate customers
- Reduced review workload on the humans in the loop thanks to the well-calibrated decision logic
- Adaptation to new deepfake patterns within days thanks to huffties

Business value

- + 30% in detected synthetic document fraud
- -60% falsely flagged legitimate users
- -40% in manual review workload

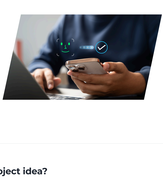
Client's testimonial

> *What matters most to us is staying ahead of a threat that is changing faster than traditional controls can keep up with. The *instinctools team treated this challenge as a* ***risk management problem, not just a technology exercise****, and that approach was reflected in how safely the solution was introduced into production. We now have stronger document checks, clearer signals for our fraud teams, and the confidence that we can respond quickly as new attack patterns emerge.*

Chief Risk Officer,
EEA-operating Digital Bank

Multiplier effect

Computer vision threat detection use cases extend far beyond digital banking. Any company that relies on document checks or user verification, from insurance and healthcare to travel and ecommerce, faces the same growing risk from AI-generated identities and forged documents. A custom computer vision system trained on real-world fraud patterns can help businesses react faster than when relying on generic vendor tools.



Do you have a similar project idea?

Contact us — and we will estimate your projects costs for free!

CONTACT US