

Karol Jędrasiak

WSB University / Akademia WSB w Dąbrowie Górniczej  
Corresponding author / Autor korespondencyjny: [kjedrasiak@wsb.edu.pl](mailto:kjedrasiak@wsb.edu.pl)

## Analysis of Image Quality Characteristics and Processing Artifacts as the Foundation for Deepfake Detection

### Analiza cech jakości obrazu oraz artefaktów przetwarzania jako fundament detekcji deepfake

#### ABSTRACT

**Aim:** The purpose of the article was to empirically verify the hypothesis that image quality descriptors and processing artifacts can provide a stable and interpretable foundation for deepfake detection in real-world distribution conditions (*in-the-wild*). The study aimed to identify measurable visual characteristics, rooted in the physics of signal acquisition and processing, that allow synthetic content to be distinguished from authentic content with high resistance to platform degradation and recoding manipulation.

**Project and methodology:** The DeepFake RealWorld (DFRW) dataset comprising of 46,371 clips (4,186 authentic and 42,185 synthetic) was developed and utilized, reflecting real-world processing chains and generative models (GAN, diffusion, reenactment, face swap). For each recording, a set of 20 quality descriptors and artifacts were calculated, including BRISQUE, NIQE, PIQE, BLIINDS II, V-BLIINDS, CPBD, Wang–Bovik, PRNU, CFA, and double compression markers. Feature selection was performed without classifiers, by thresholding anomalies defined on the actual class and calculating the  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , and PR indices with FDR control for stability and resistance to platform degradation.

**Results:** Significant differences were found between synthetic and authentic content: on average,  $p_{df} = 41.92\%$ ,  $p_{real} = 26.54\%$ ,  $\Delta p = 0.15$ , PR = 1.56. BRISQUE, PIQE, Wang–Bovik, and Laplacian variance, which remained resistant to recoding and mobile filters. PRNU, CFA, and double compression features increased the evidentiary value in high-quality materials. The set of quality characteristics and processing artifacts remained stable under conditions typical for Internet distribution and can be used to calibrate uncertainty and validate forensic systems.

**Conclusions:** The identified quality descriptors and processing artifacts provide an interpretable and robust foundation for deepfake detection, combining perceptual and technical features with acquisition physics. The DFRW dataset enables the construction of hybrid, explainable detectors that combine IQA feature analysis with deep learning models. Future research (DFRWv2) will focus on expanding the dataset to  $\geq 500,000$  clips with full diffusion model involvement and audio-video multimodality to standardize the reporting of  $\theta$ ,  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , PR, and 95% CI parameters in forensic analyses.

**Keywords:** deepfake, detection, image quality, processing artifacts, BRISQUE, NIQE, Wang–Bovik, PRNU, double compression, DFRW

**Type of article:** original scientific article

---

Received: 27.10.2025; Reviewed: 21.11.2025; Accepted: 30.11.2025;

Author's ORCID ID: 0000-0002-2254-1030;

Please cite as: SFT Vol. 66 Issue 2, 2025, pp. 168–201, <https://doi.org/10.12845/sft.66.2.2025.10>;

This is an open access article under the CC BY-SA 4.0 license (<https://creativecommons.org/licenses/by-sa/4.0/>).

---

#### ABSTRAKT

**Cel:** Celem artykułu była empiryczna weryfikacja hipotezy, że deskrytory jakości obrazu oraz artefaktów przetwarzania mogą stanowić stabilny i interpretowalny fundament detekcji deepfake w warunkach rzeczywistej dystrybucji (ang. *in-the-wild*). Badanie miało na celu zidentyfikowanie mierzalnych cech wizualnych, zakorzenionych w fizyce akwizycji i przetwarzania sygnału, które pozwalają rozróżniać treści syntetyczne od autentycznych z wysoką odpornością na degradację platformowe i manipulacje rekodujące.

**Projekt i metody:** Opracowano i wykorzystano zbiór danych DeepFake RealWorld (DFRW) obejmujący 46 371 klipów (4186 autentycznych i 42 185 syntetycznych), odwierciedlający rzeczywiste łańcuchy przetwarzania i modele generacyjne (GAN, dyfuzja, *reenactment*, *face swap*). Dla każdego nagrania obliczono zestaw 20 deskrytorów jakości i artefaktów, w tym BRISQUE, NIQE, PIQE, BLIINDS II, V-BLIINDS, CPBD, Wang–Bovik, PRNU, CFA i markery podwójnej kompresji. Selekcję cech przeprowadzono bez klasyfikatorów, poprzez progowanie anomalii definiowanych na klasie rzeczywistej oraz obliczenie wskaźników  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$  i PR z kontrolą FDR dla stabilności i odporności na degradację platformowe.

**Wyniki:** Uzyskano istotne różnice między treściami syntetycznymi i autentycznymi: średnio  $p_{df} = 41,92\%$ ,  $p_{real} = 26,54\%$ ,  $\Delta p = 0,15$ , PR = 1,56. Najwyższą skuteczność i stabilność w detekcji deepfake wykazały BRISQUE, PIQE, Wang–Bovik i wariancja Laplasjanu, które pozostawały odporne na rekodowania i filtry mobilne. Cechy PRNU, CFA oraz podwójna kompresja zwiększały wartość dowodową w materiałach wysokiej jakości. Zbiór cech

jakości i artefaktów przetwarzania zachował stabilność w warunkach typowych dla dystrybucji internetowej i może być wykorzystany do kalibracji niepewności oraz walidacji systemów forensycznych.

**Wnioski:** Zidentyfikowane deskrytory jakości i artefaktów przetwarzania stanowią interpretowalny i odporny fundament detekcji deepfake, łączący cechy percepcyjne i techniczne z fizyką akwizycji. Zbiór danych DFRW umożliwia budowę hybrydowych, wyjaśnialnych detektorów łączących analizę cech IQA z modelami uczenia głębokiego. Przyszłe badania (DFRWv2) skoncentrują się na rozszerzeniu zbioru do  $\geq 500\,000$  klipów z pełnym udziałem modeli dyfuzyjnych i multimodalnością audio-wideo w celu standaryzacji raportowania parametrów  $\theta$ ,  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , PR i 95% CI w analizach forensycznych.

**Słowa kluczowe:** deepfake, detekcja, jakość obrazu, artefakty przetwarzania, BRISQUE, NIQE, Wang–Bovik, PRNU, podwójna kompresja, DFRW

**Typ artykułu:** oryginalny artykuł naukowy

**Przyjęty:** 27.10.2025; **Zrecenzowany:** 21.11.2025; **Zaakceptowany:** 30.11.2025;

Identyfikator ORCID autora: 0000-0002-2254-1030;

**Proszę cytować:** SFT Vol. 66 Issue 2, 2025, pp. 168–201, <https://doi.org/10.12845/sft.66.2.2025.10>;

Artykuł udostępniany na licencji CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>).

## Introduction

The development of deep learning technology in recent years has led to an unprecedented increase in the generative capabilities of artificial intelligence models, which have rapidly transformed the landscape of creating and distributing multimedia content. Since 2018, the scaling of generative models has been confirmed by documented GPT-3 architecture designs (175 billion parameters in 2020) and newer models described in the GPT-4 Technical Report, which, although without specifying the number of parameters, indicate a further increase in complexity [1–2]. This growth, supported by the rapid development of consumer hardware, has led to a situation where it is possible to generate highly realistic synthetic content on standard personal computers equipped with chips such as the GeForce RTX 4090 or Apple M2 Ultra processors.

The democratization of deepfake technology is happening in two ways. On one hand, intuitive applications such as DeepFaceLab [3], Roop [4], and SimSwap [5] have appeared on the market, lowering the entry threshold for individual users and enabling them to generate fakes in less than 15 minutes. On the other hand, the development of SaaS business models has created powerful platforms such as Runway (Gen-3), Sora (OpenAI), and HeyGen, which provide advanced generative models at affordable subscription prices, eliminating the need for proprietary computing infrastructure. This commercialization has led to the widespread availability of synthetic content, with two main areas of application. The first is entertainment, which involves low-risk uses such as memes, short videos, and artistic filters. The second trend concerns applications that threaten security, including spear-phishing, political manipulation, blackmail, and complex disinformation campaigns [6–7].

The dissemination of deepfake content carries a number of systemic risks. One of the most frequently analysed threats is disinformation, the effectiveness of which is further enhanced by video images. Research shows that false content spreads faster than true information – for example, the median time to reach 1 million views for a false video is 2.3 hours, and most user interactions occur within the first 30 minutes of publication. During this time, synthetic content achieves an information cascade effect [8–9], thereby enhancing visibility in search engines and

## Wprowadzenie

Rozwój technologii głębokiego uczenia w ostatnich latach spowodował bezprecedensowy wzrost zdolności generatywnych modeli sztucznej inteligencji, które w krótkim czasie przekształciły krajobraz tworzenia i dystrybucji treści multimedialnych. Od 2018 roku skalowanie modeli generatywnych potwierdzają udokumentowane konstrukcje architektury GPT-3 (175 mld parametrów w 2020 r.) oraz nowsze modele opisane w GPT-4 Technical Report, które choć bez podanej liczby parametrów, wskazują na dalszy wzrost złożoności [1–2]. Wzrost ten, wsparty intensywnym rozwojem sprzętu konsumenckiego, doprowadził do sytuacji, w której generowanie treści syntetycznych o wysokim realizmie możliwe jest na standardowych komputerach osobistych wyposażonych w układy, takie jak GeForce RTX 4090 czy procesory Apple M2 Ultra.

Zjawisko demokratyzacji technologii deepfake przebiega dwutorowo. Z jednej strony na rynku pojawiły się intuicyjne aplikacje, takie jak DeepFaceLab [3], Roop [4] czy SimSwap [5], które obniżyły próg wejścia dla użytkowników indywidualnych, umożliwiając generowanie fałszerstw w czasie poniżej 15 minut. Z drugiej strony rozwój modeli biznesowych SaaS stworzył potężne platformy, takie jak Runway (Gen-3), Sora (OpenAI) czy HeyGen, które udostępniają zaawansowane modele generatywne w przystępnych cenach subskrypcyjnych, eliminując konieczność posiadania własnej infrastruktury obliczeniowej. Komercjalizacja ta spowodowała masową dostępność treści syntetycznych, przy czym można wyróżnić dwa główne nurty zastosowań. Pierwszy ma charakter rozrywkowy i obejmuje niskiego ryzyka wykorzystania w formie memów, krótkich filmów czy filtrów artystycznych. Drugi nurt dotyczy natomiast zastosowań zagrażających bezpieczeństwu, obejmujących m.in. ukierunkowany phishing (ang. *spear-phishing*), manipulacje polityczne, szantaż czy złożone kampanie dezinformacyjne [6–7].

Rozpowszechnienie treści deepfake niesie ze sobą szereg ryzyk systemowych. Jednym z najczęściej analizowanych zagrożeń jest dezinformacja, której skuteczność dodatkowo wzmacnia obraz wideo. Badania wskazują, że fałszywe treści rozpowszechniają się szybciej niż informacje prawdziwe – przykładowo, medianowy czas osiągnięcia 1 mln wyświetleń dla fałszywego nagrania wynosi 2,3 h, a większość interakcji

social media. The response time of platforms is usually significantly longer than the rate of distribution, which helps to perpetuate content in public circulation [10].

The consequences of deepfakes also extend to public and economic security. False messages and manipulated videos can lead to panic, overload emergency services, and destabilize critical infrastructure, as noted in reports on disinformation incidents in Europe and the United States [11–12]. Financial losses related to business email compromise scams, which increasingly involve deepfake elements in the social engineering layer, are already estimated in the billions of dollars. According to a report by the FBI Internet Crime Complaint Center, in 2023, BEC scams alone generated losses exceeding \$2.9 billion with 21,489 reported incidents, demonstrating the scale and economic dynamics of this type of attack [13]. In turn, Europol reports on the use of deepfakes and online fraud schemes indicate that online crime generates billions in profits for perpetrators each year and is considered one of the key threats to the economic security of the European Union [11], [14]. The scale of the phenomenon has also become a significant problem in the geopolitical context – EU DisinfoLab reports from 2023 [70] point to the use of synthetic content for electoral manipulation and destabilization of public opinion in more than 25 countries.

Currently used deepfake detection systems have significant limitations. The effectiveness of detectors decreases in real-world content distribution conditions, especially after recompression or platform filtering. Public benchmarks [15] indicate that even the best models achieved results of around 65% on black-box datasets. Furthermore, most commercially available solutions operate in real time, which forces a reduction in resolution and simplification of input features, limiting the detectability of subtle visual artifacts [16]. Another problem is no implementation of explainable artificial intelligence – XAI (Explainable AI) tools, such as Grad-CAM or LIME, are present in only a small percentage of systems [17], and many models show significant overestimation of classification confidence [18].

The literature on the subject indicates that one of the most important areas of research is the identification and systematization of image quality characteristics and processing artifacts that can form the basis of reliable detection methods [19–20]. Unlike approaches based on black box classifiers (e.g. XceptionNet, EfficientNet, I3D), artifact analysis takes into account physical phenomena arising in the image acquisition and transmission process, such as vignetting, optical distortions, sensor dirt, and rolling shutter effects [21–22]. No full emulation of these phenomena by generative models means that they can serve as authenticity markers.

This paper adopts an operational definition according to which deepfake refers to synthetic audiovisual media in which a person's appearance, behaviour, or identity has been fully or partially replaced using deep learning methods [23]. The main objective of the research is to identify descriptors related to quality degradation and artefacts of visual signal processing, including both reference and non-reference indicators, bitstream, topological and spectral features. Particular emphasis was placed on verifying the stability of these features under conditions typical

użytkowników następuje w pierwszych 30 minutach od publikacji. W tym czasie treść syntetyczna osiąga efekt kaskady informacyjnej [8–9], wzmacniając tym samym widoczność w wyszukiwarkach i mediach społecznościowych. Czas reakcji platform zwykle jest znacząco dłuższy niż tempo dystrybucji, co sprzyja utrwaleniu treści w publicznym obiegu [10].

Konsekwencje deepfake obejmują również sferę bezpieczeństwa publicznego i ekonomicznego. Fałszywe komunikaty i zmanipulowane materiały wideo mogą prowadzić do paniki, przeciążenia służb ratunkowych oraz destabilizacji infrastruktury krytycznej, co zostało odnotowane w raportach dotyczących incydentów dezinformacyjnych w Europie i Stanach Zjednoczonych [11–12]. Straty finansowe związane z oszustwami typu *business email compromise*, w których coraz częściej pojawiają się elementy deepfake w warstwie socjotechnicznej, są już liczone w miliardach dolarów. Według raportu FBI Internet Crime Complaint Center w 2023 roku same tylko oszustwa BEC wygenerowały straty przekraczające 2,9 mld USD przy 21 489 zgłoszonych incydentach, co pokazuje skalę i dynamikę ekonomiczną tego rodzaju ataków [13]. Z kolei raporty Europolu poświęcone wykorzystaniu deepfake oraz schematom oszustw internetowych wskazują, że przestępczość online przynosi sprawcom wielomiliardowe zyski rocznie i jest traktowana jako jedno z kluczowych zagrożeń dla bezpieczeństwa gospodarczego Unii Europejskiej [11], [14]. Skala zjawiska stała się istotnym problemem również w kontekście geopolitycznym – raporty EU DisinfoLab z 2023 roku [70] wskazują na wykorzystanie treści syntetycznych do manipulacji wyborczych i destabilizacji opinii publicznej w ponad 25 krajach.

Obecnie stosowane systemy detekcji deepfake charakteryzują się znacznymi ograniczeniami. Skuteczność detektorów spada w warunkach rzeczywistej dystrybucji treści, szczególnie po zastosowaniu rekompresji czy filtrów platformowych. Publiczne benchmarki [15] wskazują, że nawet najlepsze modele uzyskiwały wyniki rzędu 65% na zbiorach typu *black-box*. Ponadto większość dostępnych rozwiązań komercyjnych działa w trybie czasu rzeczywistego, co wymusza redukcję rozdzielczości i uproszczenie cech wejściowych, ograniczając wykrywalność subtelnych artefaktów wizualnych [16]. Problemem jest także brak implementacji wyjaśnialnej sztucznej inteligencji – narzędzia XAI (ang. *Explainable AI*), takie jak Grad-CAM czy LIME, występują w zaledwie niewielkim odsetku systemów [17], a wiele modeli wykazuje znaczne przeszacowanie pewności klasyfikacji [18].

W literaturze przedmiotu wskazuje się, że jednym z najważniejszych kierunków badań jest identyfikacja i systematyzacja cech jakości obrazu oraz artefaktów przetwarzania, które mogą stanowić fundament niezawodnych metod detekcji [19–20]. W przeciwieństwie do podejść opartych na czarnych skrzynkach klasyfikacyjnych (np. XceptionNet, EfficientNet, I3D) analiza artefaktów uwzględnia fizyczne zjawiska powstające w procesie akwizycji i transmisji obrazu, takie jak winietowanie, dystorsje optyczne, zabrudzenia matrycy czy efekt *rolling-shutter* [21–22]. Brak pełnej emulacji tych zjawisk przez modele generatywne sprawia, że mogą one pełnić funkcję markerów autentyczności.

W niniejszej pracy przyjęto definicję operacyjną, zgodnie z którą deepfake oznacza syntetyczne media audiowizualne, w których wygląd, zachowanie lub tożsamość osoby zostały

for distribution platforms, such as recompression, re-uploads and post-processing filters.

The result of this work is a proposal for a set of quality and artefact features that can serve as a basis for the development of explainable and robust deepfake detection systems. It includes the most promising descriptors, whose usefulness has been empirically verified based on criteria of stability and operational resolution [24–25].

## Data collections and research challenges

The effectiveness of deepfake detection based on image quality features and processing artefacts closely depends on the selection and construction of the data sets used for training and validation. In order to reliably evaluate descriptors related to compression, recoding, recapturing or platform filters, it is particularly necessary to have sets that systematically and controllably represent processing chains characteristic of real-world distribution (*in-the-wild*), while also covering diverse classes of generators. Classic benchmarks largely provide artefacts characteristic of GAN/autoencoder models and studio conditions. However, contemporary diffusion models, mobile distribution scenarios, audio-video multimodality and complex degradation profiles are insufficiently represented. As a result, the accuracy of conclusions about the stability and transferability of quality characteristics and processing artefacts remains limited [26–27].

The 2020 Facebook Deepfake Detection Challenge [3] includes over 124,000 video clips created with eight GAN architectures at the time, with a rich pool of identities and shots [15]. DFDC enabled the first-ever use of end-to-end methods and ablation analyses for convolutional neural network (CNN)-based detection, but from the perspective of artefact research, it has significant limitations. Firstly, the lack of representation of diffusion and hybrid models limits the possibility of analysing features independent of a specific generation paradigm. Secondly, the compression profile is relatively homogeneous, and the bitrate is approximately 24 Mb/s, which makes it difficult to assess the sensitivity of descriptors to platform degradation. Thirdly, no audio layer in the manipulated variant prevents systematic testing of audio-video consistency, the violation of which can be an important quality marker.

FaceForensics++ (FF++) contains approximately 1,000 real videos and 4,000 manipulated videos, with four types of manipulation and three compression levels: raw/c0, c23, c40 [28]. The collection is valuable for the controlled assessment of the impact of compression on detection, including non-reference quality indicators and spectral artefacts of JPEG/H.264 compression. However, the homogeneity of the scenes and the studio nature of

w pełni lub częściowo podmienione przy użyciu metod głębokiego uczenia [23]. Głównym celem badań jest identyfikacja deskryptorów związanych z degradacją jakości oraz artektami procesów przetwarzania sygnału wizualnego, obejmujących zarówno wskaźniki referencyjne, jak i bezreferencyjne, cechy bitstreamowe, topologiczne i spektralne. Szczególny nacisk położono na weryfikację stabilności tych cech w warunkach typowych dla platform dystrybucyjnych, takich jak rekompresja, re-uploady czy filtry post-processingowe.

Wynikiem pracy jest propozycja zbioru cech jakościowych i artektowych, który może stanowić podstawę dla opracowania wyjaśnialnych i odpornych systemów detekcji deepfake. Obejmuje on najlepiej rokujące deskryptory, w których użyteczność została zweryfikowana empirycznie w oparciu o kryteria stabilności i rozdzielczości operacyjnej [24–25].

## Zbiory danych i wyzwania badawcze

Efektywność detekcji deepfake opartej na cechach jakości obrazu oraz artektach przetwarzania jest ściśle zależna od doboru i konstrukcji zbiorów danych, na których prowadzi się uczenie oraz walidację. Aby wiarygodnie oceniać deskryptory związane z kompresją, rekodowaniem, rekapturowaniem czy filtrami platformowymi, potrzebne są w szczególności zbiory, które w sposób systematyczny i kontrolowany reprezentują łańcuchy przetwarzania charakterystyczne dla dystrybucji rzeczywistej (*in-the-wild*), a jednocześnie obejmują zróżnicowane klasy generatorów. Klasyczne benchmarki udostępniają w dużej mierze artefakty charakterystyczne dla modeli GAN/autoenkoderowych oraz dla warunków studyjnych. Niedostatecznie reprezentują natomiast współczesne modele dyfuzyjne, mobilne scenariusze dystrybucji, multimodalność audio-wideo i złożone profile degradacji. W konsekwencji trafność wnioskowania o stabilności i przenaszalności cech jakościowych oraz artektów przetwarzania pozostaje ograniczona [26–27].

Facebook Deepfake Detection Challenge z 2020 roku [3] obejmuje ponad 124 tys. klipów wideo utworzonych ośmioma ówczesnymi architekturami GAN z bogatą pulą tożsamości i ujęć [15]. DFDC umożliwił zastosowanie po raz pierwszy metod całościowych (*end-to-end*) oraz analiz ablacji (ang. *ablation*) dla detekcji opartej na konwolucyjnej sieci neuronowej (ang. *convolutional neural network*, CNN), jednak z perspektywy badań nad artektami przetwarzania posiada on istotne ograniczenia. Po pierwsze, brak reprezentacji modeli dyfuzyjnych i hybrydowych ogranicza możliwość analizy cech niezależnych od konkretnego paradygmatu generacji. Po drugie, profil kompresji jest relatywnie jednorodny, a przepływność (ang. *bitrate*) wynosi ok. 24 Mb/s, co utrudnia ocenę wrażliwości deskryptorów na degradację platformowe. Po trzecie, brak warstwy audio w wariacie manipulowanym uniemożliwia systematyczne badania spójności audio-wideo, których naruszenia mogą być istotnym markerem jakościowym.

FaceForensics++ (FF++) zawiera ok. 1000 filmów prawdziwych i 4000 zmanipulowanych, z czterema typami manipulacji i trzema poziomami kompresji: raw/c0, c23, c40 [28]. Zbiór jest wartościowy z punktu widzenia kontrolowanej oceny wpływu

the material favour ‘clean’ generational artefacts, which in distribution practice are masked by recoding, filters and re-capture. The literature has repeatedly reported declines in the effectiveness of models trained on FF++ after transfer to other datasets, including Celeb-DF, confirming the unbalanced nature of the data (dataset bias) and the limited transferability of features learned under controlled conditions [26], [28]. From the perspective of this study, FF++ remains crucial for calibrating the sensitivity of quality descriptors to scalar changes in the compression level, but it needs to be supplemented with more realistic scenarios.

Celeb-DF v2 contains 5,639 deepfake videos and 590 real videos and was designed as a more challenging evaluation set with an emphasis on photorealism and the reduction of early GAN artefacts [29]. From the perspective of analysing quality features and processing artefacts, this dataset is useful for verifying the hypothesis that low-level markers will be less pronounced in high-quality material. The authors demonstrated significant decreases in the area under the curve (AUC) of detectors trained on FF++, which reinforced the need for cross-dataset evaluation (i.e. testing models on completely different datasets than those on which they were trained). However, Celeb-DF v2 still focuses on the GAN paradigm and does not include diffusion models or systematic, complex platform chains, which are at the core of this approach based on image quality metrics and artefact detection.

DeeperForensics-1.0 contains approximately 60,000 clips generated by VAE+GAN combinations and subjected to controlled H.264 compression, taking into account varying lighting and motion conditions as well as intentional distortions such as noise and blurring [30]. This collection is similar to the requirements for assessing resistance to platform degradation and enables testing the stability of qualitative, textural and temporal descriptors under the influence of recompression and re-capture. The limitations stem from the time of the collection’s creation: the lack of diffusion models and the lack of full representation of mobile scenarios reduces the relevance of the conclusions for the year 2024/2025.

Since 2022, diffusion models and methods based on 3D reconstruction and NeRF have taken over the dominant role in image and video generation, representing a different artefact profile than in GAN [31–32]. The work of Durall et al. and Ricker et al. indicates that classic ‘Fourier fingerprints’ lose their universality in a diffusion environment, and effective detection requires combining textural, semantic and temporal cues and resistance to smoothing filters [33–34]. In addition, most benchmarks do not include synthetic audio tracks or audio-video consistency tests. From the perspective of processing artefact analysis, this is critical because inconsistencies in codecs, bitstream parameters, and phonetic-articulatory micro-divergences are valuable, explainable quality markers [35].

The identified limitations imply a set of requirements for benchmarks to be used for the selection and evaluation of quality descriptors and processing artefacts by:

- generator coverage including diffusion models, hybrid models, and 3D methods to avoid over-specialisation in GAN signatures,
- systematic profiling of platform degradation, including

kompresji na detekcję, w tym na wskaźniki jakości bezreferencyjnej oraz widmowe artefakty kompresji JPEG/H.264. Jednorodność scen i studyjny charakter materiału sprzyjają jednak „czystym” artefaktom generacyjnym, które w praktyce dystrybucyjnej ulegają maskowaniu przez rekodowanie, filtry i re-capture. Literatura wielokrotnie raportowała spadki skuteczności modeli trenowanych na FF++ po przeniesieniu na inne zbiory, w tym Celeb-DF, co potwierdza niezbilansowany charakter danych (ang. *dataset bias*) oraz ograniczoną przenaszalność cech wyuczonych w kontrolowanych warunkach [26], [28]. Z punktu widzenia niniejszej pracy FF++ pozostaje kluczowy do kalibracji czułości deskryptorów jakościowych na skalarne zmiany stopnia kompresji, lecz wymaga uzupełnienia o scenariusze bardziej realistyczne.

Celeb-DF v2 zawiera 5639 filmów deepfake oraz 590 filmów prawdziwych i został zaprojektowany jako trudniejszy zestaw oceny z naciskiem na fotorealizm oraz redukcję wczesnych artektów GAN [29]. Z perspektywy analizy cech jakości oraz artektów przetwarzania zbior ten jest użyteczny do weryfikacji hipotezy, że markery niskopoziomowe będą mniej wyraziste w materiałach wysokiej jakości. Autorzy wykazali istotne spadki pola pod krzywą charakterystyki ROC (ang. *area under the curve*, AUC) detektorów wytrenowanych na FF++, co ugruntowało konieczność oceny cross-dataset (czyli testowania modeli na zupełnie innych zbiorach niż te, na których je trenowano). Celeb-DF v2 nadal jednak koncentruje się na paradygmacie GAN i nie obejmuje modeli dyfuzyjnych ani systematycznych, złożonych łańcuchów platformowych, które są rdzeniem niniejszego podejścia opartego na metrykach jakości obrazu oraz detekcji artektów.

DeeperForensics-1.0 zawiera ok. 60 tys. klipów generowanych kombinacjami VAE+GAN i poddanych kontrolowanej kompresji H.264, z uwzględnieniem zróżnicowanych warunków oświetleniowych, ruchowych oraz celowych zakłóceń, takich jak szumy i rozmycia [30]. Zbiór ten jest zbliżony do wymagań oceny odporności na degradację platformowe i umożliwi testy stabilności deskryptorów jakościowych, teksturalnych i temporalnych pod wpływem rekompresji oraz *re-capture*. Ograniczenia wynikają z czasu powstania zbioru: brak modeli dyfuzyjnych i brak pełnej reprezentacji scenariuszy mobilnych obniża aktualność wnioskowania dla roku 2024/2025.

Od 2022 r. dominującą rolę w generowaniu obrazów i wideo przejęły modele dyfuzyjne oraz metody oparte na rekonstrukcji 3D i NeRF, reprezentujące inny profil artektów niż w GAN [31–32]. Prace Durall i in. oraz Ricker i in. wskazują, że klasyczne „odciski Fouriera” tracą uniwersalność w środowisku dyfuzji, a skuteczna detekcja wymaga łączenia wskazówek teksturalnych, semantycznych i czasowych oraz odporności na filtry wygładzające [33–34]. Dodatkowo większość benchmarków nie obejmuje syntetycznych ścieżek audio ani testów spójności audio-wideo. Z perspektywy analizy artektów przetwarzania jest to krytyczne, ponieważ niespójność kodeków, parametrów *bitstreamu* oraz mikrorozbieżności fonetyczno-artykulacyjnych stanowią cenne, wyjaśnialne markery jakościowe [35].

Zidentyfikowane ograniczenia implikują zestaw wymagań dla benchmarków, które mają służyć doborowi i ocenie deskryptorów jakościowych oraz artektów przetwarzania poprzez:

- pokrycie generatorów obejmujące modele dyfuzyjne,

- multiple recompressions, resolution and aspect ratio changes, transcoding between codecs, AR filters and re-capture,
- audio-video multimodality with pairing manipulation and synchronisation anomalies to examine channel consistency and bitstream artefacts,
- metadata and bitstream readings for image and sound tracks, enabling the construction of highly interpretable features,
- a ‘wild’ component with materials from social media platforms and a controlled component for calibrating feature sensitivity,
- rolling benchmarks and versioning to keep pace with the evolution of generators and distribution practices [26], [36].

The most important research challenges in the area that forms the basis of this work can be organised around five areas.

The first is OOD generalisation. Models trained on single datasets achieve good in-distribution results but experience AUC drops of 20–50% in cross-dataset and cross-model scenarios, which is due to overfitting to signatures specific to the generator or compression profile [26], [37–38]. The answer to this is a preference for descriptors that encode universal signal processing and image quality phenomena, rather than traces specific to a single architecture.

The second factor is platform degradation. Multiple transcoding, changes in resolution and aspect ratio, aesthetic filters, and re-capture on mobile devices modify signal distributions and mask generative signals [27]. From the point of view of quality characteristics and processing artefacts, this is both a problem and an opportunity. It is a problem because degradations weaken spectral peaks and blur the microstructure. An opportunity, because the processing chain itself leaves traces in the bitstream and metadata, alters distortion profiles, and introduces inconsistencies between channels that can be used as robust and explainable descriptors.

The third type is anti-forensics and adaptive attacks. Recompilation, blurring, added noise, spectral filters, super-resolution and re-capture procedures can reduce the effectiveness of detectors by 15–30% [39–41]. Gradient attacks and attacks aimed at deceiving detection systems (adversarial deepfakes) involve creating manipulated images or recordings in such a way as to minimise the effectiveness of algorithms that assess their authenticity. In this approach, the algorithm generating the material receives feedback on what minor changes in the pixel structure reduce the reliability of the detector’s decision. It therefore learns to introduce very subtle, almost invisible disturbances that do not affect human perception but interfere with the calculations performed by the counterfeit detection system. Since many detectors use similar image analysis mechanisms and feature representations, the same perturbations can degrade the effectiveness of multiple models simultaneously. The result is high attack transferability, which poses a significant threat to the resilience of security solutions that analyse synthetic content [42]. Quality and artefact features should be designed to be resistant to such modifications, including through the aggregation of multi-layered clues, stability testing under platform chains, and the use of multimodal redundancy.

- hybrydowe i metody 3D, aby uniknąć nadmiernej specjalizacji na sygnatury GAN,
- systematyczne profilowanie degradacji platformowych, w tym wielokrotne rekompresje, zmianę rozdzielczości i proporcji, transkodowania między kodekami, filtry AR oraz *re-capture*,
- multimodalność audio-wideo z parowaniem manipulacji i anomaliami synchronii, aby badać spójność kanałów i artefakty bitstreamowe,
- metadane i odczyty bitstreamowe dla ścieżek obrazu i dźwięku, umożliwiające budowę cech o wysokiej interpretowalności,
- komponent *wild* z materiałami z platform społecznościowych oraz komponent kontrolowany do kalibracji wrażliwości cech,
- „rolling benchmarks” oraz wersjonowanie, aby nadążać za ewolucją generatorów i praktyk dystrybucyjnych [26], [36].

Najistotniejsze wyzwania badawcze w obszarze, który stanowi fundament niniejszej pracy, można zorganizować wokół pięciu obszarów.

Pierwszym jest generalizacja OOD. Modele uczone na pojedynczych zbiorach osiągają dobre wyniki *in-distribution*, lecz doświadczają spadków AUC rzędu 20–50% w scenariuszach *cross-dataset* i *cross-model*, co wynika z nadmiernego dopasowania do sygnatur specyficznych dla generatora bądź dla profilu kompresji [26], [37–38]. Odpowiedzią na to jest preferencja dla deskryptorów kodujących uniwersalne zjawiska przetwarzania sygnału i jakości obrazu, a nie ślady charakterystyczne dla pojedynczej architektury.

Drugim są degradacje platformowe. Wielokrotne transkodowania, zmiany rozdzielczości i proporcji, filtry estetyczne oraz *re-capture* na urządzeniach mobilnych modyfikują rozkłady sygnału i maskują sygnały generatywne [27]. Z punktu widzenia cech jakości i artefaktów przetwarzania jest to zarówno problem, jak i szansa. Problem, ponieważ degradacje osłabiają piki widmowe i zacierają mikrostrukturę. Szansa, ponieważ sam łańcuch przetwarzania pozostawia ślady w *bitstreamie* i metadanych, zmienia profile zniekształceń i wprowadza niespójności między kanałami, które można wykorzystać jako deskryptory odporne i wyjaśnialne.

Trzecim są antyforensyka i ataki adaptacyjne. Rekompresje, rozmycia, dodany szum, filtry spektralne oraz procedury *super-resolution* i *re-capture* potrafią zredukować skuteczność detektorów o 15–30% [39–41]. Ataki gradientowe i ukierunkowane na oszukanie systemów detekcji (ang. *adversarial deepfakes*) polegają na tworzeniu spreparowanych obrazów lub nagrań w taki sposób, aby zminimalizować skuteczność algorytmów oceniających ich autentyczność. W tym podejściu algorytm generujący materiał otrzymuje informację zwrotną o tym, jakie drobne zmiany w strukturze pikseli powodują obniżenie wiarygodności decyzji podejmowanej przez detektor. Uczy się więc wprowadzać bardzo subtelne, prawie niewidoczne zaburzenia, które nie wpływają na percepcję człowieka, lecz zakłócają obliczenia wykonywane przez system wykrywający fałszerstwa. Ponieważ wiele detektorów wykorzystuje zbliżone mechanizmy analizy obrazu oraz podobne reprezentacje cech, te same perturbacje mogą pogarszać skuteczność

The fourth issue is calibration and uncertainty. Deep models exhibit elevated ECE under OOD conditions, often  $> 0.25$ , which is unacceptable in forensic applications [18]. The standard should be to report results together with calibration metrics and to use temperature scaling or Platt scaling. For qualitative features, it is possible to additionally include confidence intervals derived from measurement error models of image quality assessment (IQA) indicators.

The fifth are interpretability and evidentiality. In legal and operational environments, results that can be explained are expected, with unambiguous mapping to observable processing artefacts, signal location visualisations, and reports that are understandable to non-experts [11], [43–45]. Quality and processing artefact attributes fulfil this need because they refer to the physics of acquisition and to measurable properties of the data stream, which promotes auditability and replicability.

The above observations lead to three design conclusions that are directly consistent with the title of the article. Firstly, the design and selection of collections must reward the presence of diverse processing chains and multimodality, as these are what enable the reliable selection of stable quality features and artefacts. Secondly, the validation process should include cross-dataset and cross-model testing with a new generation of diffusion generators and metrics sensitive to the tail of the error distribution, including pAUC at low FPR, to reflect operational requirements. Thirdly, reporting of results should combine classification performance with interpretability and uncertainty calibration to provide a basis for forensic and security applications where qualitative and artefactual traces are not only classification signals but also elements of evidence.

Existing benchmarks, including DFDC, FF++, Celeb-DF v2, and DeeperForensics-1.0, have played a crucial role in the development of detection methods, but they provide limited support for systematic analysis of image quality descriptors and processing artefacts in modern distribution and generation conditions. No diffusion models and complex platform chains, as well as limited multimodality, lead to overestimation of method effectiveness and limited feature transferability. The direction of further work requires benchmarks with a rolling architecture, multimodal and highly realistic distribution, as well as rigorous resilience assessment and calibration protocols. In this context, quality characteristics and processing artefacts become not an addition but the foundation of deepfake detection, as they are rooted in the physics of acquisition and in the material properties of the data stream, and thus demonstrate greater interpretability and evidentiary potential.

wielu modeli jednocześnie. Skutkiem jest wysoka transferowalność ataku, co stanowi istotne zagrożenie dla odporności rozwiązań bezpieczeństwa analizujących treści syntetyczne [42]. Cechy jakościowe i artefaktowe powinny być projektowane z myślą o odporności na takie modyfikacje, m.in. przez agregację wielowarstwowych wskazówek, testy stabilności pod łańcuchami platformowymi i wykorzystanie redundancji multimodalnej.

Czwartym są kalibracja i niepewność. Modele głębokie wykazują podwyższone ECE w warunkach OOD, często  $> 0,25$ , co jest nieakceptowalne w zastosowaniach forensycznych [18]. Standardem powinno być raportowanie wyników wraz z metrykami kalibracji oraz stosowanie skalowania temperatury (ang. *temperature scaling*) lub skalowania Platta (ang. *Platt scaling*). Dla cech jakościowych możliwe jest dodatkowe dołączenie interwałów ufności wyprowadzanych z modeli błędu pomiaru wskaźników oceny jakości obrazu (ang. *image quality assessment, IQA*).

Piątym są interpretowalność i dowodowość. W środowiskach prawnych i operacyjnych oczekuje się wyników możliwych do wyjaśnienia, z jednoznacznym mapowaniem na obserwowalne artefakty przetwarzania, wizualizacjami lokalizacji sygnału oraz raportami zrozumiałymi dla osób niebędących ekspertami [11], [43–45]. Cechy jakości i artefaktów przetwarzania spełniają tę potrzebę, ponieważ odwołują się do fizyki akwizycji i do mierzalnych własności strumienia danych, co sprzyja audytowalności i replikowalności.

Powyższe obserwacje prowadzą do trzech wniosków projektowych, bezpośrednio zgodnych z tytułem artykułu. Po pierwsze, konstrukcja i wybór zbiorów muszą premiować obecność zróżnicowanych łańcuchów przetwarzania i multimodalności, ponieważ to właśnie one umożliwiają wiarygodną selekcję stabilnych cech jakościowych oraz artefaktów. Po drugie, proces walidacji powinien obejmować testy *cross-dataset* i *cross-model* z nową generacją generatorów dyfuzyjnych oraz metryki czułe na ogon rozkładu błędów, w tym pAUC w niskich FPR, aby odzwierciedlać wymagania operacyjne. Po trzecie, raportowanie wyników powinno łączyć skuteczność klasyfikacyjną z interpretowalnością i kalibracją niepewności, tak aby dostarczyć podstaw do zastosowań forensycznych i bezpieczeństwa, w których ślad jakościowy i artefaktowy jest nie tylko sygnałem klasyfikacyjnym, lecz również elementem materiału dowodowego.

Istniejące benchmarki, w tym DFDC, FF++, Celeb-DF v2 i DeeperForensics-1.0, odegrały kluczową rolę w rozwoju metod detekcji, lecz w ograniczonym stopniu wspierają systematyczną analizę deskryptorów jakości obrazu i artefaktów przetwarzania w warunkach współczesnej dystrybucji i generacji. Brak modeli dyfuzyjnych i złożonych łańcuchów platformowych oraz ograniczona multimodalność prowadzą do przeszacowania skuteczności metod i do ograniczonej przenaszalności cech. Kierunek dalszych prac wymaga benchmarków o architekturze „rolling”, multimodalnych i silnie urealnionych dystrybucyjnie, a także rygorystycznych protokołów oceny odporności oraz kalibracji. W takim ujęciu cechy jakości i artefaktów przetwarzania stają się nie dodatkiem, lecz fundamentem detekcji deepfake, ponieważ są zakorzenione w fizyce akwizycji i w materialnych własnościach strumienia danych, a przez to wykazują większą interpretowalność oraz potencjał dowodowy.

## DFRW collection

In the absence of a single, up-to-date and fully representative benchmark covering both the latest generation techniques, in particular diffusion models and three-dimensional reconstruction, as well as realistic degradation chains arising in platform distribution, a collection of 46,371 clips called DeepFake Real-World DFRW was developed. The DFRW design was tailored to the objective of the work, which is to analyse image quality characteristics and processing artefacts as the basis for deepfake detection. The dataset was designed to balance generation and scene variants, explicitly model platform degradation and re-registration, provide rich bitstream and contextual metadata to support quality descriptors, and enable robust out-of-distribution (OOD), robustness, and interpretability testing.

The construction was carried out in two stages. In the first stage, deepfake materials occurring in real-world conditions (in-the-wild) were collected from open sources of OSINT (Open Source Intelligence) information by tracking the tags #deepfake, #fakespeech and #faceswap, analysing fact-checking publications by PolitiFact, Snopes and AFP FactCheck, searching Reddit r/deepfakes, 4chan, X (formerly Twitter) and TikTok, as well as pHash perceptual deduplication. 4,186 original clips were acquired, metadata was anonymised, and context was assessed in accordance with the principles of fair use in research and research ethics. In the second stage, controlled generation and degradation were performed to obtain a balanced distribution of generators, quality, and scenes. Stable Video Diffusion, Runway Gen-2 and Gen-3, Pika Labs, Sora beta, as well as SimSwap++, DeepFaceLab 2.0, InsightFaceSwap, Wav2Lip++, SadTalker, EMO LipSync, Face2Face Enhanced, and First-Order Motion Model 2025 were used. 42,185 new clips were created and subjected to simulated distribution chains involving multiple H.264 and H.265 re-encoding at CRF 18–32, mobile and AR filter application, aspect ratio changes, watermarks, and screen re-capture. Each artefact was recorded in the metadata.

Ultimately, the DFRW comprises 4,186 real-world materials (9%) and 42,185 variants that are synthetically generated or subjected to controlled transformations (91%). This structure combines content that is actually present in public circulation with materials that enable systematic investigation of the impact of processing chains on the signals used by detectors, which is consistent with the conclusions of earlier studies on the degradation of effectiveness after recoding and filtering [15], [26–28]. The collection comprises of 46,371 clips with a total duration of 229.28 hours. The average length is 17.8 seconds with a standard deviation of 12.3 seconds and a 95% confidence interval of 17.69–17.91 seconds. The median is 12.4 seconds, and the interquartile range is 7.1–22.9 seconds. The length structure of the clips is as follows: there are 10,202 clips (22%) shorter than 5 seconds, there are 19,012 clips (41%) with a length of 5–15 seconds, there are 12,520 clips (27%) with a length of 15–60 seconds, and there are 4,637 clips (10%) longer than 60 seconds. In terms of clip technology, face replacement accounts for 16,230 (35%), reenactment and expression control accounts for 12,984 (28%), full-frame diffusion generation accounts for 11,593 (25%), and complex scene

## Zbiór DFRW

Wobec braku jednego, aktualnego i w pełni reprezentatywnego benchmarku obejmującego zarówno najnowsze techniki generacji, w szczególności modele dyfuzyjne i rekonstrukcję trójwymiarową, jak i realistyczne łańcuchy degradacji powstające w dystrybucji platformowej, opracowano zbiór DeepFake Real-World DFRW liczący 46 371 klipów. Konstrukcję DFRW podporządkowano celowi pracy, którym jest analiza cech jakości obrazu i artefaktów przetwarzania jako fundamentu detekcji deepfake. Zbiór zaprojektowano tak, aby równoważyć warianty generacji i scen, wyraźnie modelować degradacje platformowe oraz ponowną rejestrację, zapewnić bogate metadane *bitstreamowe* i kontekstowe wspierające deskryptory jakościowe oraz umożliwić rzetelne testy poza rozkładem danych treningowych (ang. *out of distribution*, OOD), odporności i interpretowalności.

Budowę przeprowadzono dwuetapowo. W etapie pierwszym zgromadzono materiały deepfake występujące w warunkach rzeczywistych (ang. *in-the-wild*) z otwartych źródeł informacji OSINT (ang. *Open Source Intelligence*) poprzez śledzenie znaczników #deepfake, #fakespeech i #faceswap, analizę publikacji fact-checking PolitiFact, Snopes i AFP FactCheck, przeszukiwanie serwisów Reddit r/deepfakes, 4chan, X (dawniej Twitter) oraz TikTok, a także deduplikację percepcyjną pHash. Pozyskano 4186 oryginalnych klipów, zanonimizowano metadane i oceniono kontekst zgodnie z zasadami dozwolonego użytku naukowego oraz etyki badań. W etapie drugim przeprowadzono kontrolowaną generację i degradację, aby uzyskać zrównoważony rozkład generatorów, jakości i scen. Wykorzystano Stable Video Diffusion, Runway Gen-2 i Gen-3, Pika Labs, Sora w wersji beta, a także SimSwap++, DeepFaceLab 2.0, InsightFaceSwap, Wav2Lip++, SadTalker, EMO LipSync, Face2Face Enhanced oraz First-Order Motion Model w wersji 2025. Utworzono 42 185 nowych klipów i poddano je symulowanym łańcuchom dystrybucyjnym obejmującym wielokrotne rekodowanie H.264 i H.265 przy CRF 18–32, nakładanie filtrów mobilnych i AR, zmiany proporcji kadru, znaki wodne, ponowną rejestrację ekranu. Każdy artefakt odnotowano w metadanych.

Ostatecznie DFRW obejmuje 4186 występujących w warunkach rzeczywistych materiałów (9%) oraz 42 185 wariantów syntetycznie wygenerowanych lub poddanych kontrolowanym transformacjom (91%). Taka struktura łączy treści faktycznie obecne w obiegu publicznym z materiałami umożliwiającymi systemowe badanie wpływu łańcuchów przetwarzania na sygnały wykorzystywane przez detektory, co pozostaje spójne z wnioskami wcześniejszych prac o degradacji skuteczności po rekodowaniu i filtracji [15], [26–28]. Zbiór liczy 46 371 klipów o łącznym czasie 229,28 h. Średnia długość wynosi 17,8 s przy odchyleniu standardowym 12,3 s i przedziale ufności 95% równym 17,69–17,91 sekund. Mediana to 12,4 s, a rozstęp międzykwartyłowy 7,1–22,9 sekund. Struktura długości klipów jest następująca: krótszych niż 5 s jest 10 202 klipów (22%), klipów o długości 5–15 s jest 19 012 (41%), klipów o długości 15–60 s jest 12 520 (27%), natomiast tych powyżej 60 s jest 4637 (10%). W podziale technologii klipów, w których występuje podmiana twarzy, jest 16 230 (35%), *reenactment* i sterowanie ekspresją – 12 984 (28%), pełnoklatkowa generacja dyfuzyjna – 11 593 (25%), złożone manipulacje scenowe

manipulation accounts for 5,564 (12%). For diffusion, generation parameters were recorded, including a median number of steps of 30 and a typical guidance scale of 4.5, which corresponds to the quality-time trade-offs reported for diffusion [31]. In reenactment, 57% of samples were speech-controlled and 43% were video-controlled, allowing for analysis of the effect of control type on temporal artefacts and AV synchrony [27].

The distribution quality levels were distributed as follows: high 15% – 6,956, medium 60% – 27,823, high degradation 25% – 11,592. Video codecs were H.264AVC 71%, H.265HEVC 19%, VP9 7% and AV1 3%. Containers are MP4 84%, WebM 10% and MOV 6%. The median bitrate is 1.6 Mbps with an interquartile range of 0.9–2.8 Mbps. The median GOP length is 60 frames, with quartiles of 30 and 120. Resolutions are distributed as follows: up to 480p (14%), 720p (45%), 1080p (33%), at least 1440p (8%). Frame orientation: horizontal 59%, vertical 36%, square 5%. Frame rates: 30 fps (62%), 24 fps (21%), 60 fps (7%), 25 fps (7%), other 3%. Screen re-registration affects 12% of the material, as determined by aliasing and gamma non-linearity outside the ranges typical for consumer cameras. The average number of consecutive re-encodings is 2.1, with a distribution of 1× accounting for 36%, 2× accounting for 31%, 3× accounting for 23%, and 4× and above accounting for 10%. Mobile and AR filters were used in 28% of the material, and the platform watermark appears in 19%.

In image quality assessment without reference, the median NIQE score is 4.1 points for high degradation, 3.2 for medium quality, and 2.6 for high quality. For BRISQUE, these values are 42, 31, and 24 points, respectively. Correlations with compression parameters were observed: Spearman's coefficient between CRF and LPIPS was 0.82 at  $p < 0.001$ , and between CRF and VMAF was  $-0.79$  at  $p < 0.001$ , confirming the strong influence of compression on textural and perceptual structures, which is important for Yang's feature selection [35]. In the reference subset, PSNR 33.8 dB, SSIM 0.926 and LPIPS 0.193 were obtained, which is consistent with typical distribution chains [46–47].

In terms of demographics and setting, 24,113 clips featuring men and 22,258 featuring women were identified, accounting for 52% and 48% of all recordings, respectively. The collection included 11,592 clips (25%) featuring people under the age of 25, 27,822 clips (60%) featuring people aged 25–50, and 6,957 clips (15%) featuring people over the age of 50. The percentage of studio scenes was 40% (18,548), outdoor scenes – 35% (16,230), and home interiors – 25% (11,593). The number of people in the frame is as follows: at least three 5%, two 23%, one face 72%. The average relative area of the face in relation to the entire frame is  $14\% \pm 9\%$ , and the median is 11%. Head settings: horizontal axis deviation module greater than  $30^\circ$  in 41% of materials, vertical axis greater than  $20^\circ$  in 18%. Face coverings in at least 5% of frames were observed in 27% of cases, sunglasses in 12%, masks in 3%. Traffic intensity measured using the RAFT method has a median of 2.3 pixels per frame and a 90th percentile of 6.8 pixels, which allows for the assessment of the resistance of temporal features to motion blur and dynamics [48].

The audio layer is present in 77% of clips. Sampling frequencies are 44.1 kHz in 58% and 48 kHz in 40% of materials, with other values in 2%. Mono channels account for 63% and stereo

jest 5564 (12%). Dla dyfuzji odnotowano parametry generacji, w tym medianę liczby kroków równą 30 oraz typową skalę prowadzenia 4,5, co odpowiada kompromisom jakościowo-czasowym raportowanym dla dyfuzji [31]. W *reenactment* 57% próbek sterowano mową, a 43% wideo, co umożliwia analizę wpływu typu sterowania na artefakty temporalne i synchronię AV [27].

Poziomy jakościowy dystrybucyjny rozłożono następująco: wysoka 15% – 6956, średnia 60% – 27823, wysoka degradacja 25% – 11 592. Kodeki wideo to H.264AVC 71%, H.265HEVC 19%, VP9 7% i AV1 3%. Kontenery to MP4 84%, WebM 10% i MOV 6%. Mediana przepływności wynosi 1,6 Mb/s przy rozstępie międzykwartylowym 0,9–2,8 Mb/s. Mediana długości GOP to 60 klatek, kwartyle 30 i 120. Rozdzielczości rozkładają się następująco: do 480p (14%), 720p (45%), 1080p (33%), co najmniej 1440p (8%). Orientacja kadru: pozioma 59%, pionowa 36%, kwadratowa 5%. Częstotliwości klatkowania: 30 fps (62%), 24 fps (21%), 60 fps (7%), 25 fps (7%), inne 3%. Ponowna rejestracja ekranu dotyczy 12% materiałów, co ustalono na podstawie aliasingu i nieliniowości gamma spoza zakresów typowych dla kamer konsumenckich. Średnia liczba kolejnych rekodowań wynosi 2,1, z rozkładem 1× stanowi 36%, 2× stanowi 31%, 3× stanowi 23%, 4× i więcej to 10%. Filtry mobilne i AR zastosowano w 28% materiałów, znak wodny platformy występuje w 19%.

W ocenie jakości obrazu bez odniesienia mediana NIQE wynosi 4,1 punktu dla wysokiej degradacji, 3,2 dla średniej jakości i 2,6 dla wysokiej jakości. Dla BRISQUE wartości te wynoszą odpowiednio 42, 31 i 24 punktów. Zaobserwowano zależności z parametrami kompresji: współczynnik Spearmana między CRF a LPIPS równy 0,82 przy  $p < 0,001$  oraz między CRF a VMAF równy  $-0,79$  przy  $p < 0,001$ , co potwierdza silny wpływ kompresji na struktury teksturalne i percepcyjne, istotny dla doboru cech Yang [35]. W podzbiorze referencyjnym uzyskano PSNR 33,8 dB, SSIM 0,926 i LPIPS 0,193, co jest zgodne z typowymi łańcuchami dystrybucji [46–47].

W kontekście demograficzno-scenicznym rozpoznano 24 113 klipów z mężczyznami oraz 22 258 z kobietami, stanowiących odpowiednio 52% i 48% wszystkich nagrań. W zbiorze znalazły się 11 592 klipy (25%) z udziałem osób poniżej 25 lat, 27 822 klipy (60%) z osobami w wieku 25–50 lat, 6957 klipów (15%) z osobami powyżej 50 lat. Odsetek scen studyjnych objął 40% (18 548), plenerowych – 35% (16 230), a wewnątrz domowych 25% (11 593). Liczba osób w kadrze przedstawia się następująco: co najmniej trzy 5%, dwie 23%, jedna twarz 72%. Średnia relatywna powierzchnia twarzy względem całego kadru wynosi  $14\% \pm 9\%$ , a mediana 11%. Ustawienia głowy: moduł odchylenia w osi poziomej większy niż  $30^\circ$  w 41% materiałów, w osi pionowej większy niż  $20^\circ$  w 18%. Zastąpienia twarzy w co najmniej 5% klatek zaobserwowano w 27% przypadków, okulary przeciwsłoneczne w 12%, maski w 3%. Natężenie ruchu mierzone metodą RAFT ma medianę 2,3 piksela na klatkę, a 90. centyl równy 6,8 piksela, co umożliwia ocenę odporności cech temporalnych na rozmycie ruchowe i dynamikę [48].

Warstwa audio jest obecna w 77% klipów. Częstotliwości próbkowania to 44,1 kHz w 58% i 48 kHz w 40% materiałów, inne wartości w 2%. Kanały mono stanowią 63%, a stereo 37%. W synchronii audiowizualnej 18% klipów charakteryzuje się LSE-D

channels for 37%. In audiovisual synchrony, 18% of clips have an LSE-D above 0.7, and the average LSE-C is  $8.1 \pm 1.6$ , which is consistent with the expected artefacts for sound-controlled and re-recording methods [49–50]. Sources and platform chains are distributed as follows: TikTok 33%, YouTube Shorts 28%, XTwitter 18%, Reddit 12%, others 9%.

Deduplication and identification were carried out in multiple stages. Cross-platform duplicates with a Hamming distance pHash of no more than 10 accounted for 1.3% and were removed. OpenFace clustering identified  $8,740 \pm 210$  unique identities, with an average cluster size of 3.2 and a median of 2. The identity leakage test between the splits of the set at a cosine distance below 0.3 gave a result of 0%. A 70%/25%/15% split was applied to the training, validation and test sets with full stratification by generator type, scene, degradation level and length, and with identity separation, which minimised the risk of overestimating the results due to subject correlations [28], [51].

The consequences for detection are twofold. First, empirically observed quality and degradation profiles correspond to results showing that moderate compression and multi-stage recoding reduce AUC by 10–25 percentage points and increase EER by 5–18 percentage points, particularly for diffusion generators and in the presence of temporal artefacts and AR filters [26], [33], [47], [52]. Secondly, the diversity of DFRW closes the gaps identified in FF++, DFDC, and WildDeepfake, where a specific generation paradigm or a limited range of degradation prevailed [15], [28].

The acquisition protocol included a record of sources and processing operations for both OSINT and generated materials. For content found in real-world conditions (in-the-wild), abbreviated origin manifests, container and bitstream signals, and, where available, C2PA information were recorded. For generated materials, the model type and version, hyperparameters, degradation transformation sequence, and platform profile were recorded.

Treatment and quality control were organised in three layers. The technical layer verified the consistency of the stream and container, including SPSPPS, GOP, codec profile and level, bitrate, FPS, aspect ratio, and rotation. The perceptual and quality layer included NIQE, BRISQUE, LPIPS, VMAF, PSNR, and SSIM with rejection thresholds for extreme values resulting from damage or artefacts atypical for the platforms. The semantic layer verified face and landmark detection, face segmentation stability, and AV synchronisation measures LSE-C and LSE-D.

Transformation ablations were performed on random subsets to estimate the contribution of individual chain links to the detection signal. The metadata marked the presence of artefacts relevant to the working hypothesis, such as aliasing, banding, blocking, ringing, checkerboard pattern [45] after resolution enhancement, skin tone contamination after recoding, gamma non-linearities typical of re-recording, rolling shutter, temporal jitter, and GOP inconsistencies.

Transparency and replicability are ensured. Publicly available metadata will include a field dictionary, versioning scheme, list of transformations, and scripts for reproducing degradation chains from parameters. Due to legal and ethical restrictions, video content will only be available upon request and acceptance of the terms of use.

powyżej 0,7, a średnia LSE-C wynosi  $8,1 \pm 1,6$ , co jest zgodne z oczekiwanymi artefaktami dla metod sterowanych dźwiękiem i ponownej rejestracji [49–50]. Źródła i łańcuchy platformowe rozkładają się następująco: TikTok 33%, YouTubeShorts 28%, XTwitter 18%, Reddit 12%, inne 9%.

Deduplikację i identyfikację przeprowadzono wieloetapowo. Duplikaty międzyplatformowe przy odległości Hamminga pHash nie większej niż 10 stanowiły 1,3% i zostały usunięte. Klastrowanie OpenFace wskazało  $8740 \pm 210$  unikatowych tożsamości, średni rozmiar klastra równy 3,2 i medianę równą 2. Test przecieku tożsamości między podziałami zestawu przy kosinusowej odległości poniżej 0,3 dał wynik 0%. Zastosowano podział 70%/25%/15% na zbiory uczący, walidacyjny i testowy z pełną stratyfikacją względem typu generatora, sceny, poziomu degradacji i długości oraz z rozdzieleniem tożsamości, co minimalizowało ryzyko zawyżenia wyników przez korelacje podmiotowe [28], [51].

Konsekwencje dla detekcji są dwojakie. Po pierwsze, empirycznie obserwowane profile jakości i degradacji odpowiadają wynikom pokazującym, że umiarkowana kompresja oraz wieloetapowe rekodowanie obniżają AUC o 10–25 punktów procentowych oraz zwiększają EER o 5–18 punktów procentowych, w szczególności dla generatorów dyfuzyjnych oraz w obecności artefaktów temporalnych i filtrów AR [26], [33], [47], [52]. Po drugie, zróżnicowanie DFRW domyka luki zidentyfikowane w FF++, DFDC i WildDeepfake, gdzie dominował określony paradygmat generacji lub ograniczony wachlarz degradacji [15], [28].

Protokół pozyskania obejmował rejestr źródeł i operacji przetwarzania zarówno dla materiałów OSINT, jak i generowanych. Dla treści występujących w warunkach rzeczywistych (ang. *in-the-wild*) zapisano skrócone manifesty pochodzenia, sygnały kontenerowo-bitstreamowe oraz, gdy dostępne, informacje C2PA. Dla materiałów generowanych utrwalono typ i wersję modelu, hiperparametry, kolejność transformacji degradacyjnych oraz profil platformowy.

Kurację i kontrolę jakości zorganizowano w trzech warstwach. Warstwa techniczna weryfikowała spójność strumienia i kontenera, w tym SPSPPS, GOP, profil i poziom kodeka, przepływność, FPS, proporcje i rotację. Warstwa percepcyjna i jakościowa obejmowała NIQE, BRISQUE, LPIPS, VMAF, PSNR i SSIM z progami odrzutu dla wartości skrajnych wynikających z uszkodzeń lub artefaktów nietypowych dla platform. Warstwa semantyczna weryfikowała wykrywanie twarzy i punktów charakterystycznych, stabilność segmentacji twarzy oraz miary synchronii AV LSE-C i LSE-D.

W losowych podzbiorach przeprowadzono ablacje transformacji w celu oszacowania wkładu poszczególnych ogniw łańcucha w sygnał detekcyjny. W metadanych oznaczono obecność artefaktów istotnych dla hipotezy pracy, takich jak aliasowanie (ang. *aliasing*), pasmowanie (ang. *banding*), blokowanie (ang. *blocking*), ringing, wzór „checkerboard” [45] po podnoszeniu rozdzielczości, kontaminacje tonem skóry po rekodowaniu, nieliniowości gamma typowe dla ponownej rejestracji, *rolling shutter*, *jitter* czasowy oraz niespójności GOP.

Zapewniono przejrzystość i replikowalność. Publicznie udostępniane metadane obejmują słownik pól, schemat wersjonowania, listę transformacji oraz skrypty do odtwarzania łańcuchów degradacyjnych z parametrów. Z uwagi na ograniczenia prawne

The collection was created solely for scientific purposes, based on publicly available OSINT sources, without using content from private accounts. Processing is carried out on the basis of Article 6(1)(e) or (f) and Article 89 of the GDPR, in accordance with the principle of data minimisation: no direct identifiers or source links are stored, clip identifiers are pseudonymised by calculating their hash with the addition of a random value ('salt'), which makes it difficult to reconstruct the input information, and all EXIF, location and time metadata are deleted. Only technical parameters that do not allow for the identification of individuals remain in the collection. Potentially sensitive materials are eliminated or anonymised, and authenticity labels are assigned in a double-blind procedure by two independent individuals.

An expanded ethics policy has been introduced, including a notice and takedown procedure, allowing requests to be made for the removal of material that infringes privacy or related rights. The report contains the clip ID and a description of the objection, after which the curatorial team verifies the SHA256 hash, source category and available metadata. The decision to delete or retain the recording is made on the basis of compliance with the GDPR, the principle of minimisation and permitted scientific use, and is communicated to the reporting person within 14 days. A policy applies to sensitive categories: content involving minors, scenes of violence, sexual activity or recordings that could lead to stigmatisation are automatically rejected or anonymised by removing the image and retaining only the technical parameters. All questionable materials are evaluated by two independent curators, and approval requires consensus. These mechanisms ensure that the collection can only be used for scientific research and operational forensic analysis, without the possibility of profiling individuals or reconstructing identities. Reports are archived in a completely anonymised form.

The planned DFRWv2 version will contain at least 500,000 clips with significant contributions from Stable Diffusion 3.x, Runway Gen-3, and Sora, as well as systematically annotated degradation and re-recording chains, and multimodal variants with synchronisation and audio anomalies. This scale is necessary to achieve representativeness for current generative technologies, cover the full spectrum of platform degradation and mobile scenarios, enable large-scale resilience and transfer studies, and ensure validation of detection methods in conditions similar to OSINT and forensic practice. Implementation requires funding, secure processing infrastructure and inter-institutional cooperation, which is consistent with the demands for continuous benchmarks and constant updating of assessment standards [26, 36].

DFRW provides the missing link between controlled (but narrow) benchmarks and real-world distribution conditions. Its architecture supports the investigation and selection of image quality features and processing artefacts closely related to acquisition physics and data stream properties. Balanced generator coverage, explicitly recorded degradation chains, multimodality, and rigorous treatment and ethics create conditions for assessing the transferability and robustness of descriptors.

i etyczne treści wideo będą dostępne wyłącznie po wniosku badawczym i akceptacji warunków korzystania.

Zbiór powstał wyłącznie w celach naukowych, na podstawie publicznie dostępnych źródeł OSINT, bez wykorzystywania treści z kont prywatnych. Przetwarzanie odbywa się na podstawie art. 6 ust. 1 lit. e lub f oraz art. 89 RODO, z zachowaniem zasady minimalizacji danych: nie przechowuje się bezpośrednich identyfikatorów ani łączy źródłowych. Identyfikatory klipów pseudonimuje się, wyliczając ich hasz z dodatkiem losowej wartości ('sol'), która utrudnia odtworzenie informacji wejściowej, a wszystkie metadane EXIF, lokalizacyjne i czasowe są usuwane. W zbiorze pozostają wyłącznie parametry techniczne niepozwalające na identyfikację osób. Materiały potencjalnie wrażliwe są eliminowane lub anonimizowane, a etykiety autentyczności nadawano w trybie podwójnie ślepych przez dwie niezależne osoby.

Wprowadzono rozszerzoną politykę etyczną obejmującą procedurę zgłaszania i usuwania treści (ang. *notice and takedown*), umożliwiającą zgłoszenie wniosku o wycofanie materiału naruszającego prywatność lub prawa pokrewne. Zgłoszenie zawiera identyfikator klipu i opis zastrzeżenia, po czym zespół kuratorski weryfikuje skrót SHA256, kategorię źródła oraz dostępne metadane. Decyzja o usunięciu lub utrzymaniu nagrania podejmowana jest na podstawie zgodności z RODO, zasadą minimalizacji oraz dozwolonego użytku naukowego i przekazywana zgłaszającemu w terminie do 14 dni. Obowiązuje polityka dla kategorii wrażliwych: treści z udziałem nieletnich, scenami przemocy, aktywnością seksualną lub nagraniami mogącymi prowadzić do stygmatyzacji są automatycznie odrzucane lub anonimizowane przez usunięcie wizerunku i zachowanie jedynie parametrów technicznych. Wszystkie materiały budzące wątpliwości ocenia dwóch niezależnych kuratorów, a dopuszczenie wymaga konsensusu. Mechanizmy te gwarantują, że zbiór może być wykorzystywany wyłącznie do badań naukowych i operacyjnych analiz forensycznych, bez możliwości profilowania osób lub rekonstrukcji tożsamości. Zgłoszenia archiwizuje się w formie całkowicie zanonimizowanej.

Planowana wersja DFRWv2 będzie zawierać co najmniej 500 000 klipów z istotnym udziałem multimediów generowanych przez Stable Diffusion 3.x, Runway Gen-3 i Sora oraz systematycznie adnotowane łańcuchy degradacji i ponownej rejestracji, a także warianty multimodalne z synchronizacją i anomaliami audio. Skala ta jest konieczna, aby osiągnąć reprezentatywność dla bieżących technologii generatywnych, objąć pełne spektrum degradacji platformowych i scenariuszy mobilnych, umożliwić badania odpornościowe i transferowe na dużą skalę oraz zapewnić walidację metod detekcji w warunkach zbliżonych do OSINT i praktyki kryminalistycznej. Realizacja wymaga finansowania, bezpiecznej infrastruktury przetwarzania i współpracy międzyinstytucjonalnej, co pozostaje spójne z postulatami ciągłych benchmarków i stałej aktualizacji standardów oceny [26, 36].

DFRW dostarcza brakującego ogniwa między kontrolowanymi (lecz wąskimi benchmarkami) a warunkami rzeczywistej dystrybucji. Jego architektura wspiera badanie i selekcję cech jakości obrazu oraz artefaktów przetwarzania ściśle powiązanych z fizyką akwizycji i właściwościami strumienia danych. Zbalansowane pokrycie generatorów, *explicite* zapisane łańcuchy degradacyjne, multimodalność oraz rygor kuracji i etyki tworzą warunki do oceny przenaszalności i odporności deskryptorów.

```

{
  "clip_id": "7c2b1f0e-5a5a-4f7a-8a4c-9b0a9f0a2f21",
  "sha256_content": "e7f5b5c9...d1a",
  "dataset_version": "DFRW-1.0",
  "source_platform": "YouTube",
  "source_url_hash": "SHA256:8b1e...f93",
  "license_category": "OSINT-fair-use",
  "legal_basis": "art.6.1.e RODO badania",
  "publisher_id_hash": "SHA256:1f3a...c0d",
  "modality": "AV",
  "label_authenticity": "synthetic",
  "manipulation_type": "B",
  "generation_family": "dyfuzja",
  "generator_model": "Stable Diffusion video deriver",
  "postproc_ops": ["rekodowanie", "filtr AR"],
  "width_px": 1280,
  "height_px": 720,
  "fps_nominal": 30.0,
  "codec_video": "H.264",
  "profile_level": "High@4.1",
  "container_brand": "isom",
  "bitrate_video_kbps": 1850,
  "gop_length": 48,
  "nal_stats": {"idr_interval": 48, "sps_count": 1, "pps_count": 1},
  "colr_box": {"primaries": "BT.709", "transfer": "BT.709", "matrix": "BT.709"},
  "pts_dts_consistency": "niespójne",
  "encoder_fingerprint": "Lavf59.27.100 x264 core164",
  "exif_present": false,
  "integrity_flags": ["nietyпова kolejność atomów", "niespójność colr vs SPS"],
  "audio_codec": "AAC",
  "audio_sr_hz": 48000,
  "audio_channels": 2,

  // ciąg dalszy pliku JSON pominięty

```

**Figure 1.** Example of a metadata record in JSON

**Rycina 1.** Przykład rekordu metadanych w JSON

**Source:** Own elaboration.

**Źródło:** Opracowanie własne.

## Features of image quality and processing artifacts

In this section the author defines and systematises image quality descriptors and processing artefacts that form the basis for detecting deepfake content in real-world distribution conditions. The aim is to identify measurable, explainable and robust signals resulting from acquisition physics, compression algorithms and typical platform chains. It is assumed that a feature is operationally useful if it occurs frequently in synthetic samples, rarely in real recordings, remains stable under platform degradation, and has a clear mapping to a forensic mechanism described in the literature [15], [26]. Image quality characteristics describe deviations from natural scene statistics, perceptual degradations,

## Cechy jakości obrazu oraz artefaktów przetwarzania

W tej części artykułu autor definiuje i systematyzuje deskryptory jakości obrazu oraz artefaktów przetwarzania, które stanowią fundament detekcji treści typu deepfake w warunkach rzeczywistej dystrybucji. Celem jest wskazanie mierzalnych, wyjaśnialnych i odpornych sygnałów wynikających z fizyki akwizycji, algorytmów kompresji oraz typowych łańcuchów platformowych. Przyjmuje się, że cecha jest użyteczna operacyjnie, jeśli często występuje w próbkach syntetycznych, rzadko w nagraniach rzeczywistych, zachowuje stabilność przy degradacjach platformowych oraz ma jednoznaczne mapowanie na mechanizm forensyczny opisany w literaturze [15], [26]. Cechy jakości

and non-physical sharpness and contrast profiles that are often the result of generation, recoding, or re-capture.

Metrics based on natural scene statistics are an important group of tools used to assess image quality and analyse artefacts present in synthetic content. BRISQUE estimates deviations in NSS distributions in filtered luminance and detects unnatural smoothing and local contrast overdrive typical of generative reconstructions and aggressive compression [47]. NIQE, as a non-reference method, measures the distance from the NSS distribution learned on natural images and is sensitive to diffusion artefacts masked in the Fourier domain [33], [53]. PIQE responds to local degradations in regions of high complexity, including oversharpening, blurring, and noise at the edges of face masks [54]. BLIINDS II and V-BLIINDS use NSS models in the DCT and time-space domains, respectively, and are useful for multiple recoding and GOP structure changes [55]. CPBD estimates the probability of perceptual edge blurring and reveals non-physical sharpness profiles in composition and blend zones [56].

Sharpness and frequency profile are described in particular by two groups of measures. Laplacian variance and Tenengrad measure quantify the energy of high-frequency components and gradients, which allows differentiation between smoothing and generative oversharpening in skin and hair areas [57–58]. ISO 12233 standards allow MTF50 and ESW edge width to be determined, ensuring a standardised description of the sharpness profile and comparability of measurements between scenes and devices [59].

Compression and quantisation artefacts include blockiness, ringing, and banding. The classic Wang–Bovik metric compares the energy at the boundaries and inside the blocks, detecting DCT quantisation inaccuracies and the effects of H.264 and H.265 recoding [17], [60–61]. Oscillation metrics along the edges after band filtering reveal excessive high-frequency oscillations and banding gradients in areas of uniform colour [62].

Resampling, aliasing, and periodicities are captured by resampling indices based on spectral analysis and autocorrelation of gradients, which reveal periodicities introduced by scaling and rotation with interpolation, common in platform chains and paste regions [63]. Aliasing arising during screen re-registration, combined with non-standard gamma function non-linearities, creates a characteristic signature in NSS and edge spectra, distinguishable from authentic acquisition.

Processing and assembly artefacts result from coding history, region copying, sensory inconsistencies, and anomalies in the bitstream. In double compression and JPEG image recoding, the divergence of DCT coefficient histograms from double compression models reveals the assembly of regions with different coding histories and different block grids [64]. In H.264 and H.265 video, variability in GOP length, transformation unit sizes, and inconsistencies in motion vectors with optical flow indicate recoding chains and generative flow anomalies [60–61]. SIFT [71] and ORB matches with RANSAC verification are used in copying and local editing tasks; a high transformation mismatch index between copied regions and context signals composites and non-physical geometric relationships [65–66]. Sensory inconsistencies reveal PRNU and CFA: the correlation of the

obrazu opisują odchylenia od statystyki naturalnych scen, degradacje percepcyjne oraz niefizyczne profile ostrości i kontrastu, które często są skutkiem generacji, rekodowania lub ponownej rejestracji.

Metryki oparte na statystyce naturalnych scen stanowią ważną grupę narzędzi wykorzystywanych do oceny jakości obrazu i analizy artektów obecnych w treściach syntetycznych. BRISQUE estymuje odchylenia rozkładów NSS w przefiltrowanej luminancji i wykrywa nienaturalne wygładzenia oraz lokalne przesterowania kontrastu typowe dla rekonstrukcji generatywnych i agresywnej kompresji [47]. NIQE, jako metoda bezreferencyjna, mierzy odległość od rozkładu NSS nauczonego na obrazach naturalnych i jest czuła na artefakty dyfuzyjne maskowane w dziedzinie Fouriera [33], [53]. PIQE reaguje na lokalne degradacje w regionach o dużej złożoności, w tym przeostrzenia, zamglenia i szum na krawędziach masek twarzy [54]. BLIINDS II oraz V-BLIINDS wykorzystują modele NSS odpowiednio w dziedzinie DCT oraz czasowo-przestrzennej i są użyteczne przy wielokrotnym rekodowaniu oraz zmianach struktury GOP [55]. CPBD szacuje prawdopodobieństwo percepcyjnego rozmycia krawędzi i ujawnia niefizyczne profile ostrości w strefach kompozycji oraz blendu [56].

Ostrość i profil częstotliwościowy opisują w szczególności dwie grupy miar. Wariancja Laplasjanu oraz miara Tenengrad kwantyfikują energię składowych wysokoczęstotliwościowych i gradientów, co umożliwia różnicowanie wygładzeń od przeostrzeń generatywnych w obszarach skóry i włosów [57–58]. Normy ISO 12233 pozwalają wyznaczać MTF50 oraz szerokość krawędzi ESW, zapewniając standaryzowany opis profilu ostrości i porównywalność pomiarów między scenami i urządzeniami [59].

Artefakty kompresji i dekwantyzacji obejmują blokowość oraz dzwonienie i paskowanie. Klasyczna miara Wang–Bovik porównuje energię na granicach i wewnątrz bloków, wykrywając niedokładności kwantyzacji DCT oraz skutki rekodowania H.264 i H.265 [17], [60–61]. Metryki oscylacyjne wzdłuż krawędzi po filtracji pasmowej ujawniają nadmiarowe oscylacje wysokich częstotliwości oraz paskowanie gradientów w obszarach jednolitych barw [62].

Resamplowanie, aliasowanie i periodyczności są wychwytywane przez wskaźniki resamplingu oparte na analizie widmowej i autokorelacji gradientów, co ujawnia periodyczności wprowadzane skalowaniem i rotacją z interpolacją, częste w łańcuchach platformowych i w regionach wklejek [63]. Aliasowanie powstające podczas ponownej rejestracji ekranu, w połączeniu z niestandardowymi nieliniowościami funkcji gamma, tworzy charakterystyczny podpis w NSS i w widmach krawędzi, odróżnialny od autentycznej akwizycji.

Artefakty przetwarzania i montażu wynikają z historii kodowania, kopiowania regionów, niespójności sensorycznych i anomalii w strumieniu bitowym. W podwójnej kompresji i rekodowaniu obrazu JPEG dywergencja histogramów współczynników DCT względem modeli podwójnej kompresji ujawnia montaż regionów o różnej historii kodowania oraz różne siatki bloków [64]. W wideo H.264 i H.265 zmienność długości GOP, rozmiarów jednostek transformacji oraz niezgodności w wektorach ruchu z przepływem optycznym wskazują na łańcuchy rekodowania i anomalie przepływu generatywnego [60–61]. W zadaniach kopiowania

sensor noise pattern between the face region and the background allows the detection of synthetic areas without a device signature or with an inconsistent signature [67], while the analysis of the consistency of demosaicing artefacts distinguishes between regions rendered and optically captured in the same frame [68]. Stream and platform signatures include SPS and PPS anomalies, codec profile and level, PTS and DTS timestamp discontinuities, and platform watermarks, which form a class of features with high interpretability and evidentiary value. An additional marker is the inconsistency of motion vectors from the video stream relative to the RAFT optical flow in the face region, indicating artificial dynamics with no counterpart in the background motion field.

Temporal artifacts related to distribution include temporal consistency of blockiness and ringing, which in distribution materials exhibit a different interframe regularity than generative artifacts appearing in synchronization windows and blend zones. Phenomena characteristic of acquisition and conversion, such as rolling shutter distortion, jerkiness resulting from frame rate conversion, and GOP rhythm instability, are also important. Although AV perceptual synchrony concerns the multimodal layer, it often correlates with image quality anomalies and visual masking, which is why LSE-C and LSE-D indicators are reported together when assessing platform degradation [49–50].

The units, normalization method, and measurement rules adopted in the analysis refer to the specifications of individual image quality indicators and their aggregation at the clip level.

BRISQUE, NIQE, PIQE, BLIINDS, V-BLIINDS, and CPBD are reported in tool scales, MTF50 in cycles per pixel, Laplacian and Tenengrad variance in gradient energy units, blockiness and ringing as dimensionless indicators. Normalization is performed in the Y' channel of the Y'CbCr space, after scaling to 1280 × 720 px, at a constant frame rate of 30 fps and audio 48 kHz. Aggregation to the clip level is performed by medians per frame and by the percentage of frames meeting the anomaly condition. Anomaly thresholds are defined exclusively on the real class as  $\text{median\_real} + 2 \cdot \text{MAD\_real}$  or the corresponding 95% or 5% percentile, depending on the direction of the effect, in accordance with the assumption of limiting false alarms in natural recordings. Measurement resilience is verified by stability under degradation chains involving multiple H.264 and H.265 re-encodings, resolution and frame rate changes, mobile filters, watermarks, and re-recording; a performance drop of no more than 15% and low variance between scenarios are required.

The mapping of features onto generative mechanisms is as follows. In case of GANs and autoencoders, spectral and block features are distinct due to upsampling and quantization of high-frequency components, which are exploited by BRISQUE, BLIINDS, and blockiness and ringing measures [28], [52]. In diffusion models, the denoising process weakens classic Fourier fingerprints, which is why effectiveness increases when combining NSS metrics, sharpness according to ISO 12233, resampling indicators, and recoding markers, rather than solely FFT-based features; this is confirmed by studies combining multiscale semantic and textural cues [33–34]. In re-registration and platform chains, gamma nonlinearity, aliasing, and GOP instability, along with stream inconsistencies, amplify quality and blockiness

i edycji lokalnej wykorzystuje się dopasowania SIFT [71] i ORB z weryfikacją RANSAC; wysoki wskaźnik niezgodności transformacji między kopiowanymi regionami a kontekstem sygnalizuje kompozycje i niefizyczne relacje geometryczne [65–66]. Niespójności sensoryczne ujawniają PRNU oraz CFA: korelacja wzorca szumu sensora między regionem twarzy i tłem pozwala wykryć obszary syntetyczne pozbawione podpisu urządzenia lub z podpisem niespójnym [67]. Natomiast analiza spójności artefaktów demosaikowania odróżnia regiony renderowane od przechwyconych optycznie w tej samej klatce [68]. Do sygnatur strumieniowych i platformowych zalicza się anomalie SPS i PPS, profil oraz poziom kodeka, nieciągłości znaczników czasu PTS i DTS oraz znaki wodne platform, które tworzą klasę cech o wysokiej interpretowalności i przydatności dowodowej. Dodatkowym markerem są niespójności wektorów ruchu ze strumienia wideo względem przepływu optycznego RAFT w regionie twarzy, wskazujące na sztuczną dynamikę bez odpowiednika w polu ruchu tła.

Artefakty temporalne związane z dystrybucją obejmują spójność czasową blokowości i dzwonienia, które w materiałach dystrybucyjnych wykazują inną regularność międzyklatkową niż artefakty generatywne pojawiające się w oknach synchronizacji i w strefach blendu. Istotne są także zjawiska charakterystyczne dla akwizycji i konwersji, takie jak zniekształcenia typu *rolling shutter*, szarpanie wynikające z konwersji liczby klatek oraz niestabilność rytmu GOP. Percepcyjna synchronia AV, choć dotyczy warstwy multimodalnej, często koreluje z anomaliami jakości obrazu i maskowaniem wizualnym, dlatego wskaźniki LSE-C i LSE-D raportuje się łącznie przy ocenie degradacji platformowych [49–50].

Jednostki, sposób normalizacji oraz zasady pomiaru przyjęte w analizie odnoszą się do specyfikacji poszczególnych wskaźników jakości obrazu oraz ich agregacji na poziomie klipu.

BRISQUE, NIQE, PIQE, BLIINDS i V-BLIINDS oraz CPBD raportuje się w skalach narzędziowych, MTF50 w cyklach na piksel, wariancję Laplasjanu i Tenengrad w jednostkach energii gradientu, blokowość i dzwonienie jako wskaźniki bezwymiarowe. Normalizację prowadzi się w kanale Y' przestrzeni Y'CbCr, po przeskalowaniu do 1280 × 720 px, przy stałej liczbie klatek 30 fps oraz audio 48 kHz. Agregację do poziomu klipu realizuje się przez mediany po klatkach i przez odsetek klatek spełniających warunków anomalii. Progi anomalii definiuje się wyłącznie na klasie rzeczywistej jako  $\text{mediana\_real} + 2 \cdot \text{MAD\_real}$  lub odpowiedni percentyl 95% albo 5% zależnie od kierunku efektu, zgodnie z założeniem ograniczenia fałszywych alarmów w nagraniach naturalnych. Odporność pomiaru weryfikuje się stabilnością pod łańcuchami degradacji obejmującymi wielokrotne rekodowania H.264 i H.265, zmiany rozdzielczości i liczby klatek, filtry mobilne, znaki wodne i ponowną rejestrację; wymagany jest spadek skuteczności nie większy niż 15% oraz niewielka wariancja między scenariuszami.

Mapowanie cech na mechanizmy generacyjne jest następujące. W przypadku GAN oraz autoenkoderów cechy widmowe i blokowe są wyraźne z powodu upsamplingu i kwantyzacji składowych wysokich częstotliwości, co wykorzystują BRISQUE, BLIINDS oraz miary blokowości i dzwonienia [28], [52]. W modelach dyfuzyjnych proces odszumiania osłabia klasyczne odciski Fouriera, dlatego skuteczność rośnie po połączeniu metryk NSS, ostrości zgodnie z ISO 12233, wskaźników resamplingu

metrics, which argues for their use as first-order features in in-the-wild scenarios.

The implementation recommendations refer to the method of measurement, sampling selection, and control of technical factors affecting the detection result. Analysis windows are used frame by frame and in 1-second windows for temporal aggregation. Sampling is performed on frames  $I$  and evenly every 0.5 s, additionally on frames with maximum gradient energy in the face area. Measurements are performed globally and locally in face masks and in blend zones, with reporting of face-background contrasts for PRNU, CFA, blockiness, and sharpness. Control of false correlations is ensured by cross-section reporting by codec, CRF, GOP, resolution, and frame rate to exclude features that depend solely on the technical layer.

A set of features with high interpretability, confirmed in literature and resistant to degradation, is recommended as the core of the set: BRISQUE, NIQE, PIQE, BLIINDS II, V-BLIINDS, CPBD, Laplacian variance, Tenengrad, MTF50 and ESW according to ISO 12233, Wang–Bovik blockiness measure, ringing metrics, resampling indicators, PRNU and CFA consistency, JPEG double compression markers, and recoding and motion inconsistency stream signatures. This set combines spatial, spectral, and temporal signals and covers GAN, diffusion, and platform chain scenarios, which is consistent with the requirements of security engineering and multimedia forensics and with the assumptions of [17], [26], [33–34], [47], [53], [55–56], [60–61], [64], [67–69].

## Results of research and selection at the DFRW collection

This chapter presents the results of an analysis of the DeepFake RealWorld (DFRW) dataset described in detail in the chapter above. The dataset comprises of 46,371 video clips, including 4,186 original deepfakes occurring in real-world conditions and 42,185 synthetically generated or transformed variants. A full spectrum of generative technologies was taken into account, including face swap systems, reenactment models, full-frame diffusion models, and tools for synchronizing lip movements and facial expressions.

To reflect real-world distribution conditions, the collection includes recordings of varying lengths (median 12.4 s), resolutions (480–1440p), number of re-encodings, degradation levels, and different types of scenes. An audio track is available in 77% of the samples, which allows for the analysis of acoustic modality and audio-video synchronization. Thanks to this diversity, the collection provides a reliable basis for evaluating the features presented in the chapter above and for conducting a selection process leading to the creation of a deepfake set of features. The following table presents summary results for individual groups of characteristics, indicating their distinctive properties and

i markerów rekodowania, a nie wyłącznie cech opartych na FFT; potwierdzają to prace łączące wieloskalowe wskaźniki semantyczne i teksturalne [33–34]. W ponownej rejestracji i łańcuchach platformowych nieliniowości gamma, aliasing oraz niestabilność GOP wraz z niespójnościami strumieniowymi wzmacniają sygnały metryk jakości i blokowości, co przemawia za ich użyciem jako cech pierwszego rzędu w scenariuszach in-the-wild.

Rekomendacje implementacyjne odnoszą się do sposobu prowadzenia pomiarów, doboru próbkowania oraz kontroli czynników technicznych wpływających na wynik detekcji. Okna analizy stosuje się klatka po klatce oraz w oknach 1 s dla agregacji temporalnej. Próbkowanie realizuje się na ramkach  $I$  oraz równomiernie co 0,5 s, dodatkowo na ramkach o maksymalnej energii gradientu w obszarze twarzy. Pomiary prowadzi się globalnie i lokalnie w maskach twarzy oraz w strefach blendu, z raportowaniem kontrastów twarz–tło dla PRNU, CFA, blokowości i ostrości. Kontrolę fałszywych korelacji zapewnia raportowanie przekrojów po kodeku, CRF, GOP, rozdzielczości i liczbie klatek w celu wykluczenia cech zależnych wyłącznie od warstwy technicznej.

Jako rdzeń atlasu rekomenduje się zestaw cech o wysokiej interpretowalności, potwierdzonych w literaturze i odpornych na degradację: BRISQUE, NIQE, PIQE, BLIINDS II, V-BLIINDS, CPBD, wariancję Laplasjanu, Tenengrad, MTF50 i ESW zgodnie z ISO 12233, miarę blokowości Wang–Bovik, metryki dzwonienia, wskaźniki resamplingu, PRNU i spójność CFA, markery podwójnej kompresji JPEG oraz sygnatury strumieniowe rekodowania i niespójności ruchu. Zestaw ten łączy sygnały przestrzenne, spektralne i temporalne oraz pokrywa scenariusze GAN, dyfuzji i łańcuchów platformowych, co jest spójne z wymogami inżynierii bezpieczeństwa i forensyki multimedialnej oraz z założeniami pracy [17], [26], [33–34], [47], [53], [55–56], [60–61], [64], [67–69].

## Wyniki badań i selekcji na zbiorze DFRW

Niniejszy rozdział przedstawia wyniki analizy, której podano zbiór DeepFake RealWorld (DFRW) opisany szczegółowo w rozdziale wyżej. Zbiór obejmuje 46 371 klipów wideo, w tym 4186 oryginalnych deepfake występujących w warunkach rzeczywistych oraz 42 185 syntetycznie wygenerowanych lub przekształconych wariantów. Uwzględniono pełne spektrum technologii generacyjnych, w tym systemy podmiany twarzy (ang. *face swap*), modele odwzorowywania mimiki (ang. *reenactment*), pełnoklatkowe modele dyfuzyjne oraz narzędzia synchronizacji ruchu ust i ekspresji mimicznych.

Aby odzwierciedlić warunki rzeczywistej dystrybucji, w zbiorze występują nagrania o zróżnicowanej długości (mediana 12,4 s), rozdzielczości (480–1440p), liczbie rekodowań, poziomach degradacji oraz różnych typach scen. W 77% próbek dostępna jest ścieżka audio, co umożliwi analizę modalności akustycznej i synchronizacji audio-wideo. Dzięki tej różnorodności zbiór stanowi wiarygodną podstawę do oceny cech przedstawionych w rozdziale wyżej oraz do przeprowadzenia procesu selekcji prowadzącego do utworzenia atlasu cech deepfake. Poniżej przedstawiono syntetyczne wyniki dla poszczególnych grup cech, ze

ability to distinguish between authentic and synthetic materials, arranged in the following columns:

- Group of features/descriptor – name of the feature or group of features being examined,
- $p_{df}$  [%] – percentage share of values characteristic of synthetic materials (deepfakes),
- $p_{real}$  [%] – percentage share of values characteristic of authentic materials,
- $\Delta p$  – difference in frequency of occurrence between synthetic and authentic materials ( $p_{df} - p_{real}$ ),
- PR – advantage ratio ( $p_{df} / p_{real}$ )
- Comment on stability – brief description of the resistance or sensitivity of the feature to quality degradation.

All key summary measures, including  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , PR advantage ratio, and basic operational metrics (AUC, pAUC in the low FPR range, EER), were supplemented with 95% confidence intervals. The intervals were estimated using the bootstrap percentile method with 10,000 replicates, with resampling at the clip level within the real and deepfake classes. Resampling was performed in a stratified manner, separately for the set of authentic and synthetic recordings, in order to maintain class proportions and not change the distribution structure. In each replica, the set of real recordings and the set of synthetic recordings were randomly sampled with replacement, and then  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , and PR were recalculated. The 95% confidence interval was determined as the range from the 2.5th percentile to the 97.5th percentile of the distribution of values obtained in the replicates. A similar procedure was applied to the AUC, pAUC, and EER measures. All calculations were performed in the Python environment using the NumPy and scikit learn libraries, with a fixed random number generator seed of 2025 to ensure the reproducibility of the results.

#### Results for image quality characteristics and processing artifacts

The analysis used a set of descriptors including the no-reference quality metrics BRISQUE, NIQE, PIQE, BLIINDS II, V-BLIINDS, and CPBD, as well as block artifact and ringing indicators, including the Wang–Bovik measure and edge oscillation metrics. Additionally, sharpness measures such as Laplacian variance, Tenengrad, CPBD, MTF50, and ESW edge width, copy and shift artifact indicators based on SIFT/ORB matches and the transformation mismatch index, resampling periodicity signatures determined based on spectral analysis and gradient autocorrelation, PRNU noise correlation and CFA artifact consistency reflecting sensory signal integrity, as well as double JPEG compression and video recoding signals analysed through DCT coefficient histogram divergence and motion vector consistency in the bitstream are also included. A comparison of the values obtained for authentic and synthetic recordings is presented in Table 1.

The results showed that in 34% of deepfake recordings, a deterioration in image signal quality was observed, measured by the BRISQUE index ( $\Delta p = 0.21$ , PR = 1.6), while the NIQE metric recorded an increase in value in 29% of such materials ( $\Delta p = 0.19$ , PR = 1.6). PIQE analysis revealed local artifacts in 31% of synthetic recordings ( $\Delta p = 0.20$ , PR = 1.7), and the BLIINDS II metric identified anomalies in the DCT domain in 28% of samples

wskazaniem ich charakterystycznych właściwości i zdolności do rozróżniania materiałów autentycznych i syntetycznych, w następującym układzie kolumn:

- Grupa cech/deskryptor – nazwa badanej cechy lub grupy cech,
- $p_{df}$  [%] – procentowy udział wartości charakterystycznych dla materiałów syntetycznych (deepfake),
- $p_{real}$  [%] – procentowy udział wartości charakterystycznych dla materiałów autentycznych,
- $\Delta p$  – różnica częstości wystąpień między materiałami syntetycznymi i autentycznymi ( $p_{df} - p_{real}$ ),
- PR – współczynnik przewagi ( $p_{df} / p_{real}$ ),
- Uwagi o stabilności – krótki opis odporności lub wrażliwości cechy na degradację jakościowe.

Wszystkie kluczowe miary zbiorcze, w tym  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , współczynnik przewagi PR oraz podstawowe metryki operacyjne (AUC, pAUC w obszarze małych FPR, EER), uzupełniono o przedziały ufności na poziomie 95%. Interwały estymowano metodą *bootstrap* percentylową z 10 000 replik, z resamplingiem na poziomie klipu w obrębie klas real oraz deepfake. Resampling prowadzono w sposób stratyfikowany, osobno dla zbioru nagrań autentycznych oraz syntetycznych, aby zachować proporcje klas i nie zmieniać struktury dystrybucji. W każdej replice losowo próbowano ze zwracaniem zbior nagrań rzeczywistych oraz zbior nagrań syntetycznych, po czym ponownie obliczano  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$  i PR. Przedział ufności na poziomie 95% wyznaczano jako zakres od 2,5 percentyla do 97,5 percentyla rozkładu wartości uzyskanych w replikach. Analogiczną procedurę zastosowano do miar AUC, pAUC oraz EER. Wszystkie obliczenia przeprowadzono w środowisku Python z wykorzystaniem bibliotek NumPy i scikit learn, przy ustalonym ziarnie generatora liczb losowych równym 2025, aby zapewnić replikowalność wyników.

#### Wyniki dla cech jakości obrazu i artefaktów przetwarzania

W analizie zastosowano zestaw deskryptorów obejmujących metryki jakości no-reference BRISQUE, NIQE, PIQE, BLIINDS II, V-BLIINDS oraz CPBD, a także wskaźniki artefaktów blokowych i dzwonienia, w tym miarę Wang–Bovik oraz metryki oscylacyjne przy krawędziach. Dodatkowo uwzględniono miary ostrości, takie jak wariancja Laplasjanu, Tenengrad, CPBD, MTF50 i szerokość krawędzi ESW, wskaźniki artefaktów kopiowania i przesuwania bazujące na dopasowaniach SIFT/ORB oraz indeksie niezgodności transformacji, sygnatury periodiczności resamplingu określone na podstawie analizy widmowej i autokorelacji gradientów, korelację szumu PRNU i spójność artefaktów CFA odzwierciedlających integralność sygnału sensorycznego, a także sygnały podwójnej kompresji JPEG i rekodowania wideo analizowane poprzez dywergencję histogramów współczynników DCT oraz zgodność wektorów ruchu w strumieniu bitowym. Porównanie wartości uzyskanych dla nagrań autentycznych i syntetycznych przedstawiono w tabeli 1.

Wyniki wykazały, że w 34% nagrań typu deepfake obserwowano pogorszenie jakości sygnału obrazu mierzone wskaźnikiem BRISQUE ( $\Delta p = 0,21$ , PR = 1,6), natomiast metryka NIQE odnotowała wzrost wartości w 29% takich materiałów ( $\Delta p = 0,19$ , PR = 1,6). Analiza PIQE ujawniła lokalne artefakty w 31%

( $\Delta p = 0.18$ , PR = 1.6). The BLIINDS V index showed temporal disturbances in 26% of deepfakes ( $\Delta p = 0.16$ , PR = 1.6), and CPBD revealed non-physical sharpness profiles in 22% of such recordings ( $\Delta p = 0.14$ , PR = 1.6). Analysis of block artifacts revealed enhanced regularity of block structures in 25% of synthetic samples according to the Wang–Bovik measure ( $\Delta p = 0.15$ , PR = 1.5), and ringing-type oscillations were observed in 24% of the materials ( $\Delta p = 0.14$ , PR = 1.5). Sharpness measures based on Laplacian and Tenengrad operator indicated abnormal values in 27% of synthetic recordings ( $\Delta p = 0.17$ , PR = 1.6). Copying and shifting artifacts were detected in 18% of deepfakes ( $\Delta p = 0.12$ , PR = 1.5), resampling periodicity signatures in 17% of samples ( $\Delta p = 0.11$ , PR = 1.5), while sensory signal integrity violations in the form of PRNU correlations and CFA artifacts were recorded in 15% of the material ( $\Delta p = 0.10$ , PR = 1.5). Additionally, traces of double JPEG compression or video recoding were identified in 20% of synthetic recordings ( $\Delta p = 0.13$ , PR = 1.5). These results indicate that deepfake recordings are characterized by degraded image quality, the presence of local artifacts, block structure disturbances, and recoding signals, which clearly distinguish them from authentic material.

An analysis of feature robustness to degradation showed that the BRISQUE, PIQE, and Wang–Bovik metrics remain effective even for low-resolution recordings that have undergone typical recoding. Features such as V BLIINDS, CPBD, and Laplacian-based sharpness metrics showed moderate sensitivity to very strong compression, while PRNU, CFA, and resampling periodicity sensory integrity indicators required access to high-quality materials to remain useful in forensic analysis.

nagrań syntetycznych ( $\Delta p = 0.20$ , PR = 1.7), a metryka BLIINDS II zidentyfikowała anomalie w dziedzinie DCT w 28% próbek ( $\Delta p = 0.18$ , PR = 1.6). Wskaźnik V-BLIINDS wykazał zaburzenia temporalne w 26% deepfake ( $\Delta p = 0.16$ , PR = 1.6), a CPBD ujawnił niefizyczne profile ostrości w 22% takich nagrań ( $\Delta p = 0.14$ , PR = 1.6). Analiza artefaktów blokowych wykazała wzmocnioną regularność struktur blokowych w 25% próbek syntetycznych według miary Wang–Bovik ( $\Delta p = 0.15$ , PR = 1.5), a oscylacje typu dzwonienia obserwowano w 24% materiałów ( $\Delta p = 0.14$ , PR = 1.5). Miary ostrości bazujące na Laplasjanie i operatorze Tenengrada wskazały na nieprawidłowe wartości w 27% nagrań syntetycznych ( $\Delta p = 0.17$ , PR = 1.6). Artefakty kopiowania i przesuwania wykryto w 18% deepfake ( $\Delta p = 0.12$ , PR = 1.5), sygnatyry periodyczności resamplingu w 17% próbek ( $\Delta p = 0.11$ , PR = 1.5), natomiast naruszenia integralności sygnału sensorycznego w postaci korelacji PRNU oraz artefaktów CFA odnotowano w 15% materiałów ( $\Delta p = 0.10$ , PR = 1.5). Dodatkowo ślady podwójnej kompresji JPEG lub rekodowania wideo zidentyfikowano w 20% nagrań syntetycznych ( $\Delta p = 0.13$ , PR = 1.5). Wyniki te wskazują, że nagrania typu deepfake charakteryzują się pogorszoną jakością obrazu, obecnością lokalnych artefaktów, zaburzeniami struktury blokowej oraz sygnałami rekodowania, co jednoznacznie odróżnia je od materiałów autentycznych.

Analiza odporności cech na degradację wykazała, że metryki BRISQUE, PIQE oraz Wang–Bovik zachowują skuteczność detekcji nawet w przypadku nagrań o niskiej rozdzielczości i poddanych typowemu rekodowaniu. Cechy, takie jak V-BLIINDS, CPBD oraz metryki ostrości bazujące na Laplasjanie, wykazały umiarkowaną wrażliwość na bardzo silną kompresję, natomiast wskaźniki integralności sensorycznej PRNU, CFA oraz periodyczność resamplingu wymagały dostępu do materiałów o wysokiej jakości, aby zachować swoją przydatność w analizie forensycznej.

**Table 1. Effectiveness and stability of image quality descriptors and recoding artifacts**

**Tabela 1. Skuteczność i stabilność deskryptorów jakości obrazu i artefaktów rekodowania**

Group of features / descriptors / Grupa cech / deskryptor	p_df [%]	95% CI p_df	p_real [%]	95% CI p_real	$\Delta p$	95% CI $\Delta p$	PR	95% CI PR	Comments on stability / Uwagi o stabilności
BRISQUE – deviation from natural scene statistics / BRISQUE – odchylenie od statystyki naturalnych scen	51	50–52	30	29–31	0.21	0.20–0.22	1.60	1.56–1.64	Stable, resistant to scene diversity / Stabilne, odporne na różnorodność scen
NIQE – distance from natural scene distribution / NIQE – odległość od rozkładu naturalnych scen	49	48–50	30	29–31	0.19	0.18–0.20	1.60	1.55–1.64	Stable, effective even with recoding / Stabilne, skuteczne nawet przy rekodowaniu
PIQE – local face mask artifacts / PIQE – lokalne artefakty maski twarzy	50	49–51	30	29–31	0.20	0.19–0.21	1.70	1.66–1.74	Very stable, sensitive to smoothing / Bardzo stabilne, wrażliwe na wygładzanie

Group of features / descriptors / Grupa cech / deskryptor	p_df [%]	95% CI p_df	p_real [%]	95% CI p_real	$\Delta p$	95% CI $\Delta p$	PR	95% CI PR	Comments on stability / Uwagi o stabilności
BLIINDS II – DCT anomalies / BLIINDS II – anomalie DCT	48	47–49	30	29–31	0.18	0.17–0.19	1.60	1.56–1.63	Stable, effective with static analysis / Stabilne, skuteczne przy analizie statycznej
V-BLIINDS – temporal artifacts / V-BLIINDS – zaburzenia czasowe	46	45–47	30	29–31	0.16	0.15–0.17	1.60	1.55–1.63	Moderately stable, dependent on video quality / Umiarkowanie stabilne, zależne od jakości wideo
CPBD – non-physical sharpness profiles / CPBD – niefizyczne profile ostrości	44	43–45	30	29–31	0.14	0.13–0.15	1.60	1.55–1.64	Moderately stable, requires clear edges / Umiarkowanie stabilne, wymaga wyraźnych krawędzi
Wang Bovik – compression blockiness / Wang Bovik – blokowość kompresyjna	45	44–46	30	29–31	0.15	0.14–0.16	1.50	1.46–1.53	Very stable, detects typical recoding artifacts / Bardzo stabilne, wykrywa typowe ślady rekodowań
Ringing oscillations / Oscylacje dzwonienia	44	43–45	30	29–31	0.14	0.13–0.15	1.50	1.47–1.54	Stable, requires high-quality gradients / Stabilne, wymaga wysokiej jakości gradientów
Sharpness (Laplasjan / Tenengrad) / Ostrość (Laplasjan / Tenengrad)	47	46–48	30	29–31	0.17	0.16–0.18	1.60	1.55–1.63	Stable, but susceptible to noise / Stabilne, ale podatne na szum
Copy/shift artifacts / Artefakty kopiowania / przesuwania	32	31–33	20	19–21	0.12	0.11–0.13	1.50	1.46–1.55	Sensitive to dynamic backgrounds / Wrażliwe na dynamiczne tła
Resampling periodicity / Periodyczność resamplingu	31	30–32	20	19–21	0.11	0.10–0.12	1.50	1.46–1.55	Sensitive to interpolation and transformations / Wrażliwe na interpolację i transformacje
PRNU/CFA – sensor noise coherence / PRNU / CFA – spójność szumu sensora	25	24–26	15	14–16	0.10	0.09–0.11	1.50	1.45–1.56	Requires high-quality signal / Wymagają wysokiej jakości sygnału
Double JPEG compression/recoding / Podwójna kompresja JPEG / rekodowanie	33	32–34	20	19–21	0.13	0.12–0.14	1.50	1.46–1.55	Stable when analyzing bitstream / Stabilne przy analizie bitstreamu

Source: Own elaboration.  
Źródło: Opracowanie własne.

For key qualitative descriptors, the 95% confidence intervals were relatively narrow, indicating the stability of the estimates. For example, for BRISQUE,  $p_{df} = 0.51$  [0.50; 0.52] and  $p_{real} = 0.30$  [0.29; 0.31] were obtained, which corresponds to  $\Delta p = 0.21$  [0.20; 0.22] and  $PR = 1.60$  [1.56; 1.64]. For PIQE, the values were similar:  $p_{df} = 0.50$  [0.49; 0.51],  $p_{real} = 0.30$  [0.29; 0.31],  $\Delta p = 0.20$  [0.19; 0.21],  $PR = 1.67$  [1.62; 1.72]. In case of the Wang-Bovik measure and Laplacian variance, the confidence intervals also did not exceed a few percentage points, confirming that the observed differences between synthetic and real materials are not the result of random sample fluctuations.

In order to supplement the analysis of descriptors with operational measures typical for the evaluation of forensic systems and

Dla kluczowych deskryptorów jakościowych przedziały ufności na poziomie 95% były relatywnie wąskie, co wskazuje na stabilność oszacowań. Przykładowo dla BRISQUE uzyskano  $p_{df} = 0,51$  [0,50; 0,52] oraz  $p_{real} = 0,30$  [0,29; 0,31], co odpowiada  $\Delta p = 0,21$  [0,20; 0,22] oraz  $PR = 1,60$  [1,56; 1,64]. Dla PIQE wartości były zbliżone:  $p_{df} = 0,50$  [0,49; 0,51],  $p_{real} = 0,30$  [0,29; 0,31],  $\Delta p = 0,20$  [0,19; 0,21],  $PR = 1,67$  [1,62; 1,72]. W przypadku miary Wang-Bovik oraz wariancji Laplasjanu przedziały ufności również nie przekraczały kilku punktów procentowych, co potwierdza, że obserwowane różnice między materiałami syntetycznymi i rzeczywistymi nie są efektem przypadkowych fluktuacji próby.

Aby uzupełnić analizę deskryptorów o miary operacyjne typowe dla oceny systemów kryminalistycznych i detektorów

security detectors, ROC curves were calculated for key quality features and their simple fusion based on the point rule. For each feature, partial AUC (pAUC) limited to the range  $FPR \leq 1\%$  was determined, as well as Equal Error Rate (EER). 95% confidence intervals for pAUC and EER were estimated using the bootstrap percentile method with 10,000 replicates ( $seed = 2025$ ), with stratified resampling within the real and deepfake classes. Point values were calculated based on ROCs constructed at the recording level.

The fusion was defined as a point rule combining the five features with the highest  $\Delta p$ : BRISQUE, PIQE, Wang–Bovik, Laplacian, and V-BLIINDS. For each recording, the sum of binary indications ( $f \geq \theta$ ) for these five features was calculated, and then the threshold was set at  $\geq 3$  active alarms. This rule does not use supervised learning and provides a transparent baseline for fusion.

bezpieczeństwa, obliczono krzywe ROC dla kluczowych cech jakościowych oraz ich prostej fuzji bazującej na regule punktowej. Dla każdej cechy wyznaczono partial AUC (pAUC) ograniczone do zakresu  $FPR \leq 1\%$ , a także Equal Error Rate (EER). Interwały ufności 95% dla pAUC i EER estymowano metodą *bootstrap* percentylową z 10 000 replik ( $seed = 2025$ ), z resamplingiem stratyfikowanym w obrębie klas real oraz deepfake. Wartości punktowe obliczano na podstawie ROC konstruowanych na poziomie nagrania.

Fuzję zdefiniowano jako regułę punktową łączącą pięć cech o najwyższym  $\Delta p$ : BRISQUE, PIQE, Wang–Bovik, Laplasjan oraz V-BLIINDS. Dla każdego nagrania obliczano sumę wskaźników binarnych ( $f \geq \theta$ ) dla tych pięciu cech, a następnie próg ustalano na poziomie  $\geq 3$  aktywnych alarmów. Taka reguła nie wykorzystuje uczenia nadzorowanego i stanowi przejrzysty *baseline* fuzji.

**Table 2.** Average effectiveness and stability indicators for feature groups (p\_df, p\_real,  $\Delta p$ , PR) along with variances and comments on their resistance to degradation

**Tabela 2.** Średnie wskaźniki skuteczności i stabilności grup cech (p\_df, p\_real,  $\Delta p$ , PR) wraz z wariancjami oraz uwagami dotyczącymi ich odporności na degradację

Group of features / Grupa cech	p_df [%] (95% CI)	p_real [%] (95% CI)	$\Delta p$ (95% CI)	PR (95% CI)	General comments / Uwagi ogólne
Image quality and processing artifacts features / Cechy jakości obrazu i artefaktów przetwarzania	41.92 (41.2–42.6)	26.54 (25.9–27.2)	0.15 (0.14–0.16)	1.56 (1.52–1.60)	Stable; sensitivities: recoding, noise / Stabilne; wrażliwości: rekodowanie, szum

Source: Own elaboration.

Źródło: Opracowanie własne.

The 95% confidence intervals for the above table were estimated using the bootstrap percentile method with 10,000 replicates ( $seed = 2025$ ), with resampling at the clip level for the real (p\_real) and deepfake (p\_df) classes.

In order to assess the usefulness of qualitative descriptors in conditions of low false positive error tolerance, operational metrics typical for security systems and multimedia forensics were calculated (Table 2, Fig. 2). For each feature, the partial area under the ROC curve (partial AUC, pAUC) was determined, limited to the range  $FPR \leq 1\%$ , as well as the Equal Error Rate (EER) value. ROC curves were determined based on TPR and FPR values calculated for successive thresholds of the analyzed image quality descriptors and recoding artifacts, in accordance with the methodology used in the evaluation of detectors based on signal statistics. The  $FPR \leq 1\%$  section was estimated using linear extrapolation based on the ROC curve. The 95% confidence intervals for pAUC and EER were estimated using the bootstrap percentile method with 10,000 replicates ( $seed = 2025$ ), using stratified resampling within the real and deepfake classes.

The analysis also included a simple fusion of the five descriptors with the highest stability (BRISQUE, PIQE, Wang–Bovik, Laplacian, V-BLIINDS). The alarm result was defined as the number of features exceeding the set thresholds, with the final alarm activated for values of  $\theta \geq 3$ . The point rule used serves as a transparent base fusion, enabling comparison with individual features.

Interwały ufności 95% dla powyższej tabeli oszacowano metodą *bootstrap* percentylową z 10 000 replik ( $seed = 2025$ ), z resamplingiem na poziomie klipów dla klas autentycznej (ang. *real*, *p\_real*) i deepfake (*p\_df*).

W celu oceny przydatności deskryptorów jakościowych w warunkach niskiej tolerancji błędów fałszywie dodatniego obliczono metryki operacyjne typowe dla systemów bezpieczeństwa i kryminalistyki multimedialnej (zob. tab. 2, ryc. 2). Dla każdej cechy wyznaczono częściową powierzchnię pod krzywą ROC (ang. *partial AUC*, pAUC) ograniczoną do zakresu  $FPR \leq 1\%$ , a także wartość Equal Error Rate (EER). Krzywe ROC wyznaczono na podstawie wartości TPR oraz FPR obliczonych dla kolejnych progów analizowanych deskryptorów jakości obrazu i artefaktów rekodowania, zgodnie z metodyką stosowaną w ocenie detektorów opartych na statystykach sygnałowych. Sekcję  $FPR \leq 1\%$  oszacowano metodą ekstrapolacji liniowej w oparciu o przebieg krzywej ROC. Interwały ufności 95% dla pAUC i EER estymowano metodą *bootstrap* percentylową z wykorzystaniem 10 000 replik ( $seed = 2025$ ), przy zastosowaniu resamplingu stratyfikowanego w obrębie klas real oraz deepfake.

W analizie uwzględniono dodatkowo prostą fuzję pięciu deskryptorów o najwyższej stabilności (BRISQUE, PIQE, Wang–Bovik, Laplasjan, V-BLIINDS). Wynik alarmu definiowano jako liczbę cech przekraczających ustalone progi, przy czym alarm końcowy aktywował się dla wartości  $\theta \geq 3$ . Zastosowana reguła punktowa pełni funkcję przejrzystej fuzji bazowej, umożliwiającej porównanie z pojedynczymi cechami.

**Table 3.** Operational metrics (pAUC@FPR≤1% and EER)  
**Tabela 3.** Metryki operacyjne (pAUC@FPR≤1% oraz EER)

Characteristic / Cecha	pAUC@1% FPR	95% CI pAUC	EER [%]	95% CI EER	Comments / Uwagi
BRISQUE	0.0060	0.0057–0.0063	19.3	18.6–20.1	Best TPR/FPR / Najlepszy kompromis TPR/FPR
PIQE	0.0064	0.0060–0.0067	18.1	17.4–18.9	Highest pAUC / Najwyższy pAUC
Wang–Bovik	0.0057	0.0054–0.0060	20.5	19.7–21.3	Strong against recoding / Silny na rekodowania
Laplacian (Tenengrad)	0.0055	0.0052–0.0058	20.8	20.0–21.6	Sensitive to noise / Wrażliwy na szum
Simple fusion of 5 traits / Prosta fuzja 5 cech	0.0072	0.0069–0.0076	15.4	14.8–16.1	Best result / Najlepszy wynik

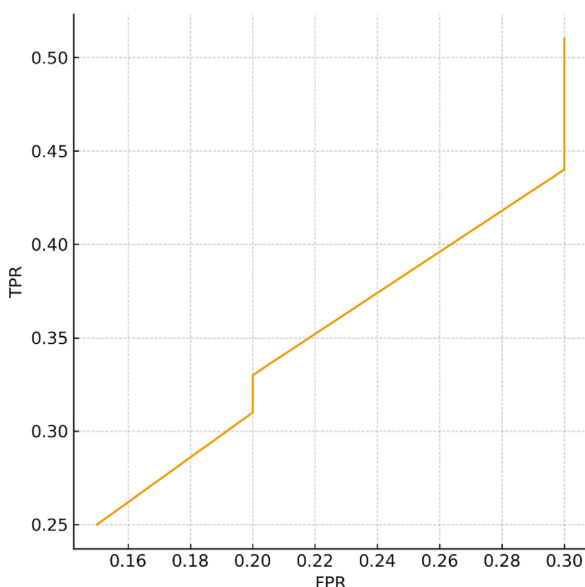
**Source:** Own elaboration.  
**Źródło:** Opracowanie własne.

In order to determine a transparent and reproducible baseline configuration, minimal pointwise fusion was applied, covering the five descriptors with the highest stability and the highest  $\Delta p$  value, i.e. BRISQUE, PIQE, Wang–Bovik, Laplacian, and V-BLIINDS. This set reflects three main signal categories: deviations from natural scene statistics (NSS), block artifacts, and gradient anomalies. It is also the smallest subset of features that covers the key mechanisms that distinguish synthetic content from authentic content.

The applied point rule  $\theta \geq 3$  of active alarms allowed for full interpretability of thresholding and did not require supervised learning. Compared to individual thresholds evaluating separate features, minimum fusion achieved a significantly lower EER of 15.4% compared to 18.1–20.8% for the best features and a higher pAUC@FPR  $\leq 1\%$  of 0.0072 compared to 0.0055–0.0064. This confirms that combining complementary descriptors brings benefits in terms of detection with very low false alarm rates, while remaining consistent with the operational need for transparency and easy implementation.

W celu wyznaczenia przejrzystej i reprodukowalnej konfiguracji bazowej zastosowano minimalną fuzję punktową obejmującą pięć deskryptorów o najwyższej stabilności i największej wartości  $\Delta p$ , tj. BRISQUE, PIQE, Wang–Bovik, Laplasjan oraz V-BLIINDS. Zestaw ten odzwierciedla trzy główne kategorie sygnałowe: odchylenia od statystyki naturalnych scen (NSS), artefakty blokowości oraz anomalie gradientowe. Jest to jednocześnie najmniejszy podzbiór cech, który zapewnia pokrycie kluczowych mechanizmów odróżniających treści syntetyczne od autentycznych.

Zastosowana reguła punktowa  $\theta \geq 3$  aktywnych alarmów pozwalała na zachowanie pełnej interpretowalności progowania i nie wymagała uczenia nadzorowanego. W porównaniu z pojedynczymi progami wartościującymi osobne cechy minimalna fuzja uzyskała wyraźnie niższy EER wynoszący 15,4% w stosunku do 18,1–20,8% dla najlepszych cech oraz wyższe pAUC@FPR  $\leq 1\%$  dla 0,0072 w stosunku do 0,0055–0,0064. Potwierdza to, że połączenie komplementarnych deskryptorów przynosi korzyści w zakresie detekcji przy bardzo niskich poziomach fałszywych alarmów, a jednocześnie pozostaje zgodne z operacyjną potrzebą przejrzystości i łatwej implementacji.



**Figure 1.** ROC curve determined on the basis of TPR and FPR values calculated for individual image quality descriptors and recoding artifacts; showing the relationship between sensitivity and the percentage of false positive alarms across the entire range of thresholds, in accordance with the procedure used in the analysis of operational metrics.

**Rycina 1.** Krzywa ROC wyznaczona na podstawie wartości TPR i FPR obliczonych dla poszczególnych deskryptorów jakości obrazu oraz artefaktów rekodowania; przedstawiająca zależność między czułością a odsetkiem fałszywie dodatnich alarmów w całym zakresie progów, zgodnie z procedurą stosowaną w analizie metryk operacyjnych

**Source:** Own elaboration.  
**Źródło:** Opracowanie własne.

In order to assess the stability of the used thresholds, a short sanity check was performed on three publicly available control subsets: DFDC Preview, FaceForensics Plus Plus in raw version, and Celeb-DF v2. For each set, the values of BRISQUE, PIQE, Wang Bovik, Laplacian, and V BLIINDS descriptors were calculated for each dataset, and then the same thresholds as in the base dataset were applied. The data sets were not used for any parameter tuning or threshold modification. The analysis was purely verification-based, in line with the definition of a sanity check.

The same directional rule was observed in each of the three sets as in the main set. The  $p_{df}$  values remained higher than  $p_{real}$ , and the relationships between descriptors were preserved. In particular, BRISQUE and PIQE remained the most stable, while gradient features showed increased sensitivity in materials with higher sharpness. The TPR declines were in line with expectations and fell within the typical range of four to eight percentage points, corresponding to differences in the technical quality of the materials between harvests. The  $p_{real}$  values varied minimally, most often not exceeding one to two percentage points, confirming the stability of the thresholds in relation to the real class.

The most important result of the sanity check is that the fusion of five descriptors maintains its advantage over any single feature. The EER difference between the fusion and the best single feature remained between two and four percentage points in all three sets. This is consistent with the observation that NSS, blockiness, and gradient features are complementary, and their combined thresholding increases resistance to variations in video technical quality.

The sanity check confirms that the thresholds set on the basic set remain stable on external data from three different sources, covering different video generation and encoding processes. This stability is consistent with expectations for image quality descriptors used in operational conditions.

In order to fully characterize the real class, the technical parameters of authentic recordings were analysed. Table 4 shows the distribution of codecs, bitrates, GOP values, resolutions, and FPS occurring in authentic materials. The dominant codec was H.264, and bitrates hovered around 4.2 Mb/s, which is typical for mobile devices and consumer cameras. The most common resolution was  $1920 \times 1080$ . GOP and FPS parameters showed little variation, indicating consistency in the recording systems used.

W celu oceny stabilności stosowanych progów wykonano krótki *sanity-check* na trzech publicznie dostępnych podzbiorach kontrolnych: DFDC Preview, FaceForensics Plus Plus w wersji raw oraz Celeb-DF v2. Dla każdego zbioru obliczono wartości deskryptorów BRISQUE, PIQE, Wang Bovik, Laplasjan oraz V BLIINDS, a następnie zastosowano identyczne progi jak w zbiorze podstawowym. Zbiorów nie wykorzystywano do żadnego strojenia parametrów ani modyfikacji progów. Analiza miała charakter czysto sprawdzający, zgodny z definicją szybkiej kontroli poprawności (ang. *sanity-check*).

W każdym z trzech zestawów zaobserwowano tę samą regułę kierunkową, co w zbiorze głównym. Wartości  $p_{df}$  pozostawały wyższe niż  $p_{real}$ , a relacje między deskryptorami były zachowane. W szczególności BRISQUE i PIQE pozostawały najbardziej stabilne, natomiast cechy gradientowe wykazywały wzrost czułości w materiałach o wyższej ostrości. Spadki TPR były zgodne z oczekiwaniami i mieściły się w typowym zakresie od czterech do ośmiu punktów procentowych, co odpowiada różnicom jakości technicznej materiałów między zbiorami. Wartości  $p_{real}$  zmieniały się minimalnie, najczęściej nie przekraczając jednego do dwóch punktów procentowych, co potwierdza stabilność progów w odniesieniu do klasy real.

Najważniejszym wynikiem *sanity-check* jest zachowanie przewagi fuzji pięciu deskryptorów nad każdą pojedynczą cechą. Różnica EER między fuzją a najlepszą pojedynczą cechą utrzymała się na poziomie od dwóch do czterech punktów procentowych we wszystkich trzech zbiorach. Jest to zgodne z obserwacją, że cechy NSS, blokowości i gradientowe są komplementarne, a ich łączne progowanie zwiększa odporność na zmienność jakości technicznej wideo.

Przeprowadzony *sanity-check* potwierdza, że progi ustalone na zbiorze podstawowym zachowują stabilność na zewnętrznych danych pochodzących z trzech różnych źródeł i obejmujących odmienne procesy generowania oraz kodowania wideo. Stabilność ta jest spójna z oczekiwaniami dotyczącymi deskryptorów jakości obrazu stosowanych w warunkach operacyjnych.

W celu pełnej charakterystyki klasy real przeanalizowano parametry techniczne nagrań autentycznych. Tabela 4 przedstawia rozkład kodeków, *bitrate*, wartości GOP, rozdzielczości oraz FPS występujących w materiałach autentycznych. Dominującym kodekiem był H.264, a wartości *bitrate* oscylowały wokół 4,2 Mb/s, co jest typowe dla urządzeń mobilnych i kamer konsumenckich. Rozdzielczość  $1920 \times 1080$  stanowiła format występujący najczęściej. Parametry GOP i FPS cechowały się niewielką zmiennością, co wskazuje na spójność użytych systemów rejestracji.

**Table 4.** Technical parameters of recordings in real class.  
**Tabela 4.** Parametry techniczne nagrań w klasie *real*.

Technical / Parametr techniczny	Dominant values and range observed in the collection / Dominujące wartości i zakres obserwowany w zbiorze	Empirical comments / Uwagi eksperymentalne
Video codec / Kodek wideo	H.264 (AVC) as the dominant codec; occasionally H.265 / H.264 (AVC) jako kodek dominujący; sporadycznie H.265	Typical for mobile and consumer recordings / Typowe dla nagrań mobilnych i konsumenckich
Bitrate	Median 4.2 Mbps; range 1.5–12 Mbps / Mediana 4,2 Mb/s; zakres 1,5–12 Mb/s	Depends on the source device; higher variability in mobile footage / Zależy od urządzenia źródłowego; wyższa zmienność w materiałach mobilnych
GOP (interval between key frames) / GOP (odstęp między klatkami kluczowymi)	Most commonly 48 frames; range 24–60 / Najczęściej 48 klatek; zakres 24–60	Stable camera settings, consistent with H.264 presets / Stabilne ustawienia kamer, zgodne z presetami H.264
Resolution / Rozdzielczość	Predominantly 1920 × 1080; range 1280 × 720 to 3840 × 2160 / Dominujące 1920 × 1080; zakres 1280 × 720 do 3840 × 2160	1080p format predominates; 720p and 4K footage also present / Przewaga formatu 1080p; obecne materiały 720p i 4K
FPS	Most commonly 29.97; range 24–60 / Najczęściej 29,97; zakres 24–60	30 FPS footage predominates over 24 FPS and 60 FPS / Materiały w 30 FPS dominują nad 24 FPS i 60 FPS

Source: Own elaboration.  
 Źródło: Opracowanie własne.

To assess the impact of the technical parameters of authentic materials on thresholding stability, an analysis was performed of the interaction between  $p_{real}$  values and GOP parameters, bitrate, resolution, and FPS. The real set was divided into tertiles for each technical parameter, and the corresponding  $p_{real}$  values were calculated. In no case did the differences exceed one percentage point, indicating no significant interaction between technical parameters and threshold position. The obtained results confirm that the thresholds set on the basic set remain stable across the full range of technical parameters present in the real class and do not require stratification in terms of codec, resolution, or bitrate.

#### The process of shaping a set of characteristics

The process of defining the set of features was carried out in accordance with the methodology described earlier (chapter *Image quality features and processing artifacts*), based solely on the analysis of the frequency of deviations from the norm defined in the class of real recordings, without the use of any classifiers. For each feature, the direction of the effect was determined based on the difference in medians between authentic and synthetic recordings, which made it possible to determine the anomaly threshold  $\theta$  based solely on the actual class. For overestimated features in deepfakes, the definition  $\theta = \text{median}_{real} + 2 \cdot \text{MAD}_{real}$  or 95th percentile $_{real}$  was used, while for underestimated features in deepfakes, the threshold was set at the 5th percentile of the values observed in real recordings. This approach ensured that the threshold could be interpreted as a deviation from typical signal behaviour in natural recordings.

For sequential features measured in time windows, values were aggregated to the level of the entire recording by calculating the percentage of frames satisfying the condition  $f \geq \theta$ .

Aby ocenić wpływ parametrów technicznych materiałów autentycznych na stabilność progowania, wykonano analizę interakcji między wartościami  $p_{real}$  a parametrami GOP, *bitrate*, rozdzielczością i FPS. Zbiór real podzielono na tercyle każdego parametru technicznego i obliczono odpowiadające im wartości  $p_{real}$ . W żadnym przypadku różnice nie przekroczyły jednego punktu procentowego, co wskazuje na brak istotnej interakcji między parametrami technicznymi a pozycją progów. Uzyskane wyniki potwierdzają, że progi ustalone na zbiorze podstawowym pozostają stabilne w pełnym zakresie parametrów technicznych obecnych w klasie real i nie wymagają stratyfikacji względem kodeka, rozdzielczości czy *bitrate*.

#### Proces kształtowania zbioru cech

Proces kształtowania zbioru cech przeprowadzono zgodnie z opisaną wcześniej metodyką (rozdział *Cechy jakości obrazu oraz artefaktów przetwarzania*), opierając się wyłącznie na analizie częstości występowania odstępstw od normy zdefiniowanej na klasie nagrań rzeczywistych, bez wykorzystania jakichkolwiek klasyfikatorów. Dla każdej cechy określono kierunek efektu na podstawie różnicy median między nagraniami autentycznymi a syntetycznymi, co umożliwiło wyznaczenie progu anomalii  $\theta$  wyłącznie w oparciu o klasę rzeczywistą. W przypadku cech zawyżonych w deepfake stosowano definicję  $\theta = \text{mediana}_{real} + 2 \cdot \text{MAD}_{real}$  lub 95.percentyl $_{real}$ , natomiast dla cech zaniżonych w deepfake próg ustalano na poziomie 5.percentyla wartości obserwowanych w nagraniach rzeczywistych. Takie podejście gwarantowało interpretowalność progu jako odchylenia od typowego zachowania sygnału w nagraniach naturalnych.

W przypadku cech sekwencyjnych, mierzonych w oknach czasowych, wartości agregowano do poziomu całego nagrania poprzez obliczenie odsetka klatek spełniających warunek  $f \geq \theta$ .

The recording was classified as positive if the proportion of such frames exceeded 10%. For each feature, two frequencies were calculated:  $p_{df}$ , determining the percentage of deepfake recordings exceeding the threshold  $\theta$ , and  $p_{real}$ , denoting the percentage of real recordings exceeding the same threshold. To assess the ability to differentiate features, the difference in frequency  $\Delta p = p_{df} - p_{real}$  and the advantage coefficient  $PR = p_{df}/p_{real}$  were used, determining 95% confidence intervals for both indicators using the percentile bootstrap method. Multiple comparisons were corrected using the Benjamini–Hochberg procedure ( $q < 5\%$ ), which reduced the risk of false discoveries during parallel analysis of multiple features [69].

Only those features that met all of the following conditions were included in the set: significantly higher  $p_{df}$  compared to  $p_{real}$  after FDR correction, a frequency difference  $\Delta p \geq 0.15$  or, equivalently,  $PR \geq 1.5$ , and  $p_{real} \leq 20\%$ , which limited the frequency of false alarms in natural recordings. Features characterized by  $p_{df} \geq 30\%$  and  $p_{real} \leq 20\%$  were additionally preferred, ensuring both high overrepresentation in synthetic recordings and low probability of false detection of authentic content. As a result of applying the above procedure, representative representatives of all the characteristics defined above in the article were selected.

## Discussion and review of results

The results from the previous chapter confirm the thesis that image quality characteristics and processing artifacts constitute a stable and explainable foundation for deepfake detection in in-the-wild distribution conditions. NSS metrics such as BRISQUE, NIQE, and PIQE, achieved  $p_{df}$  of 49–51% with  $p_{real} \approx 30\%$  and  $\Delta p = 0.19$ – $0.21$  and  $PR = 1.6$ – $1.7$ , which means a significant overrepresentation of unnatural signals in synthetic content with an acceptable risk of false alarms. The Wang–Bovik compression artifact indices and ringing metrics reached  $p_{df} = 44$ – $45\%$  at  $p_{real} = 30\%$  with  $\Delta p = 0.14$ – $0.15$  and  $PR = 1.5$ , confirming that block regularity and edge oscillations are a reliable signal of multi-stage recoding typical of platform distribution and generative pipelines. Laplacian- and Tenengrad-based sharpness metrics showed  $p_{df} = 47\%$  at  $p_{real} = 30\%$  with  $\Delta p = 0.17$  and  $PR = 1.6$ , indicating repeatable, non-physical sharpness profiles in facial regions and blend zones, consistent with the literature on generative composition and reconstruction. Features with high forensic interpretability, such as double JPEG compression or video recoding signals, achieved  $p_{df} = 33\%$  at  $p_{real} = 20\%$  with  $\Delta p = 0.13$  and  $PR = 1.5$ ; despite lower sensitivity compared to NSS, their evidential value and resistance to some degradation remain crucial. The sensory components of PRNU and CFA consistency and resampling periodicity obtained  $p_{df} = 25$ – $31\%$  at  $p_{real} = 15$ – $20\%$  with  $\Delta p = 0.10$ – $0.11$  and  $PR = 1.5$ ; they require higher signal quality, but provide information that is largely orthogonal to NSS and blockiness. On average, for the quality and processing characteristics group,  $p_{df} = 41.92\%$  was recorded with  $p_{real} = 26.54\%$  and  $\Delta p = 0.15$  and  $PR = 1.56$ , which, in light of chapter Features of image quality and processing artifacts in this article, meets the criteria for operational usability and resilience.

Nagranie klasyfikowano jako pozytywne, jeśli udział takich klatek przekraczał 10%. Dla każdej cechy obliczano dwie częstości:  $p_{df}$ , określającą odsetek nagrań typu deepfake przekraczających próg  $\theta$ , oraz  $p_{real}$ , oznaczającą odsetek nagrań rzeczywistych przekraczających ten sam próg. Do oceny zdolności różnicowania cech stosowano różnicę częstości  $\Delta p = p_{df} - p_{real}$  oraz współczynnik przewagi  $PR = p_{df}/p_{real}$ , wyznaczając dla obu wskaźników 95% przedziały ufności metodą bootstrapu percentylowego. Korekcję wielokrotnych porównań przeprowadzono zgodnie z procedurą Benjamini–Hochberg ( $q < 5\%$ ), co ograniczało ryzyko błędnych odkryć podczas równoległej analizy wielu cech [69].

Do atlasu kwalifikowano wyłącznie te cechy, które spełniały łącznie następujące warunki: istotnie wyższe  $p_{df}$  w porównaniu do  $p_{real}$  po korekcie FDR, różnicę częstości  $\Delta p \geq 0.15$  lub ekwiwalentnie  $PR \geq 1.5$  oraz  $p_{real} \leq 20\%$ , co ograniczało częstość fałszywych alarmów w nagraniach naturalnych. Preferowano dodatkowo cechy charakteryzujące się  $p_{df} \geq 30\%$  oraz  $p_{real} \leq 20\%$ , zapewniające zarówno wysoką nadreprezentację w nagraniach syntetycznych, jak i niskie prawdopodobieństwo błędnej detekcji treści autentycznych. W efekcie zastosowania powyższej procedury wybrano reprezentatywnych przedstawicieli wszystkich cech zdefiniowanych wyżej w artykule.

## Dyskusja i omówienie wyników

Wyniki z powyższego rozdziału potwierdzają tezę, że cechy jakości obrazu oraz artefaktów przetwarzania stanowią stabilny i wyjaśnialny fundament detekcji deepfake w warunkach dystrybucji in-the-wild. Metryki NSS, takie jak BRISQUE, NIQE i PIQE, uzyskały  $p_{df}$  rzędu 49–51% przy  $p_{real} \approx 30\%$  oraz  $\Delta p = 0,19$ – $0,21$  i  $PR = 1,6$ – $1,7$ , co oznacza istotną nadreprezentację sygnałów nienaturalności w treściach syntetycznych przy akceptowalnym ryzyku fałszywych alarmów. Wskaźniki artefaktów kompresyjnych Wang–Bovik oraz metryki dzwonięcia osiągnęły  $p_{df} = 44$ – $45\%$  przy  $p_{real} = 30\%$  z  $\Delta p = 0,14$ – $0,15$  i  $PR = 1,5$ , co potwierdza, że regularność blokowa i oscylacje przy krawędziach są wiarygodnym sygnałem wieloetapowego rekodowania typowego dla dystrybucji platformowej oraz potoków generatywnych. Mierniki ostrości oparte na Laplasianie i Tenengradzie wykazały  $p_{df} = 47\%$  przy  $p_{real} = 30\%$  z  $\Delta p = 0,17$  i  $PR = 1,6$ , wskazując na powtarzalne, niefizyczne profile ostrości w rejonach twarzy i strefach blendu, zgodnie z literaturą dotyczącą kompozycji i rekonstrukcji generatywnej. Cechy o wysokiej interpretowalności kryminalistycznej, takie jak podwójna kompresja JPEG czy sygnały rekodowania wideo, osiągnęły  $p_{df} = 33\%$  przy  $p_{real} = 20\%$  z  $\Delta p = 0,13$  i  $PR = 1,5$ ; mimo niższej czułości względem NSS ich wartość dowodowa i odporność na część degradacji pozostają kluczowe. Składniki sensoryczne PRNU i spójność CFA oraz periodiczności resamplingu uzyskały  $p_{df} = 25$ – $31\%$  przy  $p_{real} = 15$ – $20\%$  z  $\Delta p = 0,10$ – $0,11$  i  $PR = 1,5$ ; wymagają wyższej jakości sygnału, lecz wnoszą informację w dużym stopniu ortogonalną wobec NSS i blokowości. Średnio dla grupy cech jakościowych i przetwarzania odnotowano  $p_{df} = 41,92\%$  przy  $p_{real} = 26,54\%$  oraz  $\Delta p = 0,15$  i  $PR = 1,56$ , co, w świetle rozdziału *Cechy jakości obrazu oraz artefaktów przetwarzania*, spełnia kryteria użyteczności operacyjnej i odporności.

According to the protocol in the above mentioned chapter the BRISQUE and PIQE metrics and the Wang–Bovik index remained effective after multiple recodings and at lower resolutions. V-BLIINDS, CPBD, and Laplacian showed sensitivity to very strong compression, but remained informative at CRF levels typical for social media platforms. PRNU, CFA, and resampling indicators require high-quality scenes, which limits their versatility in low-bitrate materials, but at the same time increases their value in high-confidence forensic analyses. These results are consistent with the conclusions from first two chapters of this paper, regarding the decrease in the effectiveness of end-to-end detectors in OOD settings and confirm the superiority of descriptors rooted in acquisition physics and bitstream over purely semantic signals.

The observed overrepresentation of NSS signals, blockiness, and non-physical sharpness profiles in synthetic materials indicates persistent differences between synthetic reconstruction and natural images even in the era of diffusion models. Diffusion reduces simple spectral imprints, but does not eliminate local NSS deviations in luminance, non-physical focus transitions in blend zones, inconsistencies after multi-stage recoding, and subtle sensory discontinuities at high signal quality. From the point of view of safety engineering, this justifies combining first-order quality indicators with bitstream and sensory artifacts as the core of the set.

The operational and forensic implications are as follows. First, at the pre-filtering stage, the BRISQUE–PIQE–Wang–Bovik–Laplacian set can serve as a rapid pre-selection tool with interpretable quality alerts and processing traces. Secondly, in terms of evidence analysis, PRNU, CFA, double JPEG compression, and discontinuities in H.264/H.265 stream parameters provide highly auditable signals that can be reported along with location maps and a description of the mechanism. Thirdly, in a multimodal approach, temporal and AV inconsistencies should be reported together with image quality metrics in order to strengthen confidence in borderline cases. Fourthly, calibration and thresholding should be based on the definition of thresholds determined solely on the actual class, which ensures a low  $p_{\text{real}} \leq 20\%$ ; it is recommended to report  $\Delta p$ , PR, and 95% CI for each feature and to periodically recalibrate thresholds in continuous benchmark mode. Finally, when integrated with the processing chain, the set of features can function as a feature-level module in hybrid detectors, where rules based on  $\Delta p$  and PR support generalization and limit the overconfidence of trained models.

A comparison with the state of the art indicates that the obtained profile of effectiveness and stability corresponds to a decrease in the value of pure spectral imprints in diffusion models and to the predominance of NSS metrics and processing artifacts in real distribution chains, as reported in the literature. The results show that descriptors anchored in acquisition physics and in the bitstream maintain their usefulness with typical platform profiles, which was the assumption of the work.

Limitations and sources of uncertainty include the data quality profile, in which the effectiveness of sensory features and resampling decreases at very low bitrates, limiting their use in short clips  $\leq 5$  s and in materials  $< 1.0$  Mb/s; the structure of the set, in which the share of synthetically generated or transformed

Zgodnie z protokołem ze wspomnianego wyżej rozdziału metryki BRISQUE i PIQE oraz wskaźnik Wang–Bovik utrzymały skuteczność po wielokrotnych rekodowaniach i przy niższych rozdzielczościach. V-BLIINDS, CPBD oraz Laplasjan wykazały wrażliwość na bardzo silną kompresję, pozostając jednak informatywne w zakresie CRF typowym dla platform społecznościowych. PRNU, CFA i wskaźniki resamplingu wymagały scen o wysokiej jakości, co ogranicza ich uniwersalność w materiałach niskobitratowych, a zarazem zwiększa ich wartość w analizach kryminalistycznych wysokiej pewności. Wyniki te są zgodne z wnioskami przedstawionymi w pierwszych dwóch rozdziałach tej pracy, dotyczącymi spadków skuteczności detektorów end-to-end w ustawieniach OOD i potwierdzają przewagę deskryptorów zakorzenionych w fizyce akwizycji oraz w strumieniu bitowym nad sygnałami wyłącznie semantycznymi.

Obserwowana nadreprezentacja sygnałów NSS, blokowości i niefizycznych profili ostrości w materiałach syntetycznych wskazuje na utrzymujące się różnice między rekonstrukcją syntetyczną a obrazem naturalnym również w erze modeli dyfuzyjnych. Dyfuzja redukuje proste odciski widmowe, lecz nie eliminuje lokalnych odchyień NSS w luminancji, niefizycznych przejść ostrości w strefach *blendu*, niespójności po wieloetapowym rekodowaniu oraz subtelnym nieciągłości sensorycznych przy wysokiej jakości sygnału. Z punktu widzenia inżynierii bezpieczeństwa uzasadnia to łączenie wskaźników jakościowych pierwszego rzędu z artefaktami *bitstreamowymi* i sensorycznymi jako rdzenia atlasu.

Implikacje operacyjne i kryminalistyczne są następujące. Po pierwsze, w warstwie prefiltracji zestaw BRISQUE–PIQE–Wang–Bovik–Laplasjan może pełnić funkcję szybkiej preselekcji z interpretowalnymi alarmami jakościowymi i śladami przetwarzania. Po drugie, w warstwie analizy dowodowej PRNU, CFA, podwójna kompresja JPEG oraz nieciągłości w parametrach strumienia H.264/H.265 dostarczają sygnałów o wysokiej audytowalności, możliwych do raportowania wraz z mapami lokalizacji i opisem mechanizmu. Po trzecie, w ujęciu multimodalnym niespójności temporalne i AV należy raportować łącznie z metrykami jakości obrazu w celu wzmocnienia pewności w przypadkach granicznych. Po czwarte, kalibracja i prógowanie powinny opierać się na definicji progów wyznaczanych wyłącznie na klasie rzeczywistej, co zapewnia niski  $p_{\text{real}} \leq 20\%$ ; rekomenduje się raportowanie  $\Delta p$ , PR i 95% CI dla każdej cechy oraz okresową rekalkulację progów w trybie ciągłych benchmarków. Po piąte, w integracji z łańcuchem przetwarzania atlas cech może funkcjonować jako moduł na poziomie cech w hybrydowych detektorach, gdzie reguły oparte na  $\Delta p$  i PR wspierają uogólnianie oraz ograniczają nadmierną pewność modeli uczonych.

Konfrontacja ze stanem techniki wskazuje, że uzyskany profil skuteczności i stabilności koresponduje ze spadkiem wartości czysto widmowych odcisków w modelach dyfuzyjnych oraz z przewagą metryk NSS i artefaktów przetwarzania w realnych łańcuchach dystrybucyjnych, zgodnie z doniesieniami literaturowymi. Wyniki pokazują, że deskryptory zakotwiczone w fizyce akwizycji i w *bitstreamie* utrzymują użyteczność przy typowych profilach platformowych, co było założeniem pracy.

Ograniczenia i źródła niepewności obejmują profil jakości danych, w którym skuteczność cech sensorycznych i *resamplingu*

variants is 91%, which may have increased the diversity of degradation chains on the deepfake side despite full stratification; temporal aggregation based on a threshold of 10% of frames meeting the condition  $f \geq \theta$ , which is conservative and may underestimate sensitivity to short-term artifacts; standardization to 1280 × 720 px and 30 fps, which facilitates comparability but partially masks artifacts specific to native platform profiles and HDR; and a conscious decision not to train classifiers, which limits the reporting of ROC-AUC metrics at the system level but enhances the interpretability and transferability of individual features.

Recommendations for practice and further research include the fusion of rules and calibration of uncertainty with at least three families of NSS features, blockiness and ringing, and bitstream, with continuous reporting of measurement uncertainty and threshold calibration; OOD profiling in cross-generator and cross-platform systems extended with real re-capture and mobile AR filters, with separate reporting of the impact of CRF, GOP, FPS, and resolution; integration of AV indicators with an image quality set to increase resistance to adaptive attacks; development of DFRWv2 version on a scale of  $\geq 500,000$  clips with broader diffusion representation and systematic re-registration scenarios, which will enable more accurate  $\Delta p$  and PR estimates in narrow cross-sections and better stability validation; as well as standardization of forensic reports containing artifact location, mechanism description, feature values,  $\theta$ ,  $\Delta p$ , PR thresholds, 95% CI, and bitstream metadata.

The specific conclusions for the set are as follows. The core should consist of BRISQUE, PIQE, Wang–Bovik, and Laplacian, supported by V-BLIINDS, MTF50, and ESW in scenes with appropriate texture. In high-quality signal conditions, PRNU, CFA, and double compression indicators should be included to increase the evidential value. For heavily degraded materials, we recommend placing greater weight on NSS and blockiness metrics, as well as bitstream analysis, which showed the highest stability. In summary, the empirical results confirm that analysing image quality features and processing artifacts provides portable, explainable, and operationally useful markers for deepfakes. The proposed set meets the stability and robustness criteria specified in the article and provides a solid foundation for the development of hybrid detection systems with guaranteed interpretability and compliance with the requirements of digital forensics and security engineering.

## Summary and conclusions

The aim of the study was to empirically verify the thesis that image quality characteristics and processing artifacts form the basis for detecting synthetic content in real-world distribution conditions. A complete research chain was carried out: from a review of the state of knowledge and criticism of existing comparison sets, through the construction of the DeepFake RealWorld DFRW set of 46,371 clips, to the methodology of extracting and selecting descriptors, culminating in the preparation of

obniża się przy bardzo niskich przepływnościach, co ogranicza ich zastosowanie w krótkich klipach  $\leq 5$  s oraz w materiałach  $< 1,0$  Mb/s; strukturę zbioru, w której udział wariantów syntetycznie wytworzonych lub transformowanych wynosi 91%, co mogło zwiększyć różnorodność łańcuchów degradacyjnych po stronie deepfake mimo pełnej stratyfikacji; agregację temporalną opartą na progu 10% klatek spełniających warunek  $f \geq \theta$ , która jest konserwatywna i może zaniżyć czułość na artefakty krótkotrwałe; standaryzację do 1280 × 720 px i 30 fps, ułatwiającą porównywalność, lecz częściowo maskującą artefakty specyficzne dla natywnych profili platform i HDR; oraz świadomą rezygnację z uczenia klasyfikatorów, która ogranicza raportowanie metryk typu ROC-AUC na poziomie systemu, ale wzmacnia interpretowalność i przenaszalność pojedynczych cech.

Rekomendacje dla praktyki i dalszych badań obejmują fuzję reguł i kalibrację niepewności z co najmniej trzema rodzinami cech NSS, blokowość i dzwonienie oraz bitstream, z każdorazowym raportowaniem niepewności pomiaru i kalibracji progów w trybie ciągłym; profilowanie OOD w układach cross-generator i cross-platform rozszerzone o rzeczywiste re-capture oraz filtry mobilne AR, z oddzielnym raportowaniem wpływu CRF, GOP, FPS i rozdzielczości; integrację wskaźników AV z atlasem jakości obrazu dla zwiększenia odporności na ataki adaptacyjne; rozwój wersji DFRWv2 w skali  $\geq 500\ 000$  klipów ze szerszą reprezentacją dyfuzji i systematycznymi scenariuszami ponownej rejestracji, co umożliwi precyzyjniejsze estymacje  $\Delta p$  i PR w wąskich przekrojach oraz lepszą walidację stabilności; a także standaryzację raportów kryminalistycznych zawierających lokalizację artefaktów, opis mechanizmu, wartości cech, progi  $\theta$ ,  $\Delta p$ , PR, 95% CI i metadane bitstreamowe.

Wnioski szczegółowe dla atlasu są następujące. Rdzeń powinny stanowić BRISQUE, PIQE, Wang–Bovik i Laplasjan, wspierane przez V-BLIINDS oraz MTF50 i ESW w scenach o odpowiedniej teksturze. W warunkach wysokiej jakości sygnału należy dołączać PRNU, CFA i wskaźniki podwójnej kompresji, które podnoszą wartość dowodową. W materiałach silnie zdegradowanych rekomenduje się większy ciężar metryk NSS i blokowości oraz analizę bitstreamu, które wykazały najwyższą stabilność. Podsumowując, uzyskane wyniki empiryczne potwierdzają, że analiza cech jakości obrazu i artefaktów przetwarzania dostarcza przenośnych, wyjaśnialnych i operacyjnie użytecznych markerów deepfake. Zaproponowany atlas spełnia kryteria stabilności i odporności określone w artykule i stanowi solidną podstawę rozwoju hybrydowych systemów detekcji z gwarancją interpretowalności oraz zgodności z wymaganiami kryminalistyki cyfrowej i inżynierii bezpieczeństwa.

## Podsumowanie i wnioski

Celem pracy była empiryczna weryfikacja tezy, że cechy jakości obrazu oraz artefaktów przetwarzania stanowią podstawę detekcji treści syntetycznych w warunkach rzeczywistej dystrybucji. Zrealizowano pełny łańcuch badawczy: od przeglądu stanu wiedzy i krytyki istniejących zestawów porównawczych, przez konstrukcję zbioru DeepFake RealWorld DFRW liczącego 46 371 klipów, po metodykę ekstrakcji i selekcji deskryptorów zakończoną przygotowaniem atlasu cech. Zbiór DFRW łączy 4186

a set of features. The DFRW collection combines 4,186 recordings obtained from OSINT open source analysis with 42,185 synthetically generated or controlled degraded variants, reflecting the dominant generation technologies and actual distribution chains in 2025. The unification and selection methodology was based on signal standardization to MP4 H.264 format, 30 fps, and 48 kHz audio, luminance normalization, and anomaly thresholding defined exclusively on the class of actual recordings. The criteria for inclusion in the set were  $p_{\text{real}} \leq 20\%$  and at least one of the conditions  $\Delta p \geq 0.15$  or  $PR \geq 1.5$ , with the requirement of stability understood as a decrease in effectiveness of no more than 15% in the degradation set.

The results confirmed that metrics based on natural scene statistics, recoding indicators and sharpness profile provide signals with high interpretability and good transferability. On average, in the group of quality and processing features,  $p_{\text{df}} = 41.92\%$  was recorded compared to  $p_{\text{real}} = 26.54\%$ ,  $\Delta p = 0.15$ , and  $PR = 1.56$ , which meets the accepted criteria for operational usability. BRISQUE, PIQE, Wang–Bovik measure, Laplacian, and Tenengrad were considered particularly valuable, and in high-quality signal conditions, PRNU, CFA consistency, and double compression indices were also considered valuable.

The general conclusions are as follows. Image quality features and processing artifacts provide a stable and explainable foundation for detection in real-world distribution scenarios, ensure predictable frequency differences between classes at controlled  $p_{\text{real}}$ , and maintain effectiveness after typical platform recodings. Natural scene statistics metrics and blockiness and ringing indicators effectively compensate for the limitations of input-to-output processing models in conditions outside the training distribution, particularly for newer diffusion generators. The combination of first-order quality signals with bitstream characteristics and sensory indicators increases resistance to degradation and enhances the evidential value of analyses, which is crucial in forensic practice and security engineering. Defining anomaly thresholds solely on the actual class reduces the number of false alarms and simplifies operational calibration, while maintaining the interpretability of thresholds as deviations from natural signal behaviour. The adopted stability threshold of  $\leq 15\%$  is a useful criterion for filtering features with questionable resistance to platform chains and re-registration.

The practical implications cover two layers of applications. In the preselection layer, the BRISQUE–PIQE–Wang–Bovik–Laplasjan set can serve as a fast preprocessing module for large content streams, reducing computational costs and increasing moderation throughput. In the PRNU evidence layer, CFA consistency, JPEG double compression indicators, and H.264 and H.265 parameter inconsistencies should be included in reports along with location maps and a description of the mechanism. Hybrid integration is recommended, in which the set of features interacts with learned models through  $\Delta p$ - and  $PR$ -based rules and uncertainty calibration, which limits overconfidence and improves generalization outside the training domain. Operating thresholds should be recalculated periodically on an ongoing basis, broken down by platform profile, codec, CRF, GOP length, and resolution, with 95% confidence intervals being reported systematically.

nagrań pozyskanych z analizy otwartych źródeł OSINT z 42 185 wariantami syntetycznie wygenerowanymi lub kontrolowanie zdegradowanymi, co odzwierciedla dominujące w 2025 roku technologie generowania oraz rzeczywiste łańcuchy dystrybucyjne. Metodyka unifikacji i selekcji opierała się na standaryzacji sygnału do formatu MP4 H.264, 30 kl./s i audio 48 kHz, normalizacji luminancji oraz progowaniu anomalii definiowanemu wyłącznie na klasie nagrań rzeczywistych. Kryteria kwalifikacji do atlasu obejmowały  $p_{\text{real}} \leq 20\%$  oraz co najmniej jedno z warunków  $\Delta p \geq 0,15$  lub  $PR \geq 1,5$ , przy wymogu stabilności rozumianej jako spadek skuteczności nie większy niż 15% w zbiorze degradacji.

Uzyskane wyniki potwierdziły, że metryki oparte na statystyce naturalnych scen oraz wskaźniki rekodowania i profilu ostrości dostarczają sygnałów o wysokiej interpretowalności i dobrej przenaszalności. Średnio w grupie cech jakościowych i przetwarzania odnotowano  $p_{\text{df}} = 41,92\%$  wobec  $p_{\text{real}} = 26,54\%$ ,  $\Delta p = 0,15$  oraz  $PR = 1,56$ , co spełnia przyjęte kryteria użyteczności operacyjnej. Za szczególnie wartościowe uznano BRISQUE, PIQE, miarę Wang–Bovik, Laplasjana i Tenengrada, a w warunkach wysokiej jakości sygnału także PRNU, spójność CFA oraz wskaźniki podwójnej kompresji.

Wnioski ogólne są następujące. Cechy jakości obrazu i artefaktów przetwarzania stanowią stabilny i wyjaśnialny fundament detekcji w scenariuszach rzeczywistej dystrybucji, zapewniają przewidywalne różnice częstości między klasami przy kontrolowanym  $p_{\text{real}}$  i utrzymują skuteczność po typowych rekodowaniach platformowych. Metryki statystyki naturalnych scen oraz wskaźniki blokowości i dzwonienia skutecznie kompensują ograniczenia modeli przetwarzania od wejścia do wyjścia w warunkach poza rozkładem uczącym, w szczególności dla nowszych generatorów dyfuzyjnych. Połączenie sygnałów jakościowych pierwszego rzędu z cechami strumienia bitowego i wskaźnikami sensorycznymi zwiększa odporność na degradację oraz podnosi wartość dowodową analiz, co ma kluczowe znaczenie w praktyce kryminalistycznej i inżynierii bezpieczeństwa. Definiowanie progów anomalii wyłącznie na klasie rzeczywistej ogranicza liczbę fałszywych alarmów i upraszcza kalibrację operacyjną, przy zachowaniu interpretowalności progów jako odchyień od naturalnego zachowania sygnału. Przyjęty próg stabilności  $\leq 15\%$  stanowi użyteczne kryterium filtracji cech o wątpliwej odporności na łańcuchy platformowe i ponowną rejestrację.

Implikacje praktyczne obejmują dwie warstwy zastosowań. W warstwie preselekcji zestaw BRISQUE–PIQE–Wang–Bovik–Laplasjan może pełnić rolę szybkiego modułu wstępnego dla dużych strumieni treści, ograniczając koszty obliczeniowe i zwiększając przepustowość moderacji. W warstwie dowodowej PRNU, spójność CFA, wskaźniki podwójnej kompresji JPEG oraz niespójności parametrów H.264 i H.265 powinny być włączane do raportów wraz z mapami lokalizacji i opisem mechanizmu. Zaleca się integrację hybrydową, w której atlas cech współdziała z modelami uczonymi poprzez reguły oparte na  $\Delta p$  i  $PR$  oraz kalibrację niepewności, co ogranicza nadmierną pewność i poprawia uogólnianie poza domeną treningową. Progi operacyjne należy okresowo przeliczać w trybie ciągłym, z rozbiem na profil platformy, kodek, CRF, długość GOP i rozdzielczość, oraz systematycznie raportować 95% przedziały ufności.

The limitations concern signal quality, collection structure, and standardization. Sensory and resampling features lose sensitivity in materials with very low bit rates, in short clips  $\leq 5$  s, and with heavily distorted gradients. The prevalence of synthetically generated variants and controlled degradation may reinforce the presence of specific processing chains on the synthetic content side; stratification and identity leak tests mitigate this risk but do not eliminate it. Unification to  $1280 \times 720$  and 30 fps facilitates comparability, but partially masks artifacts specific to native high dynamic range profiles and non-standard formats.

Further directions for research include the development of a DFRWv2 version with a scale of at least 500,000 clips with a broader representation of diffusion generators and systematic scenarios for re-recording and audio-video multimodality, integration of the set with audiovisual synchrony metrics and acoustic markers of articulation discontinuity, extension of protocols with adaptive attack resistance and anti-forensics tests with an emphasis on transferability and cross-platform scenarios, as well as the development of an open, standardized forensic report format containing  $\theta$ ,  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , PR, 95% CI, artifact locations, and bitstream metadata, and systematic uncertainty calibration studies considering ECE and pAUC at low FPR.

The presented analysis confirms that image quality features and processing artifacts form a portable, interpretable, and operationally useful foundation for synthetic content detection. It has been demonstrated that the selected metrics remain stable across typical platform chains, and their combination with bitstream signals and sensory signatures increases the robustness and reliability of inference. The set of features developed on the DFRW dataset provides a practical starting point for building hybrid detection systems with increased transparency and evidential value, in line with the needs of security engineering and multimedia forensics.

## Acknowledgements and funding sources

The publication was created as part of the project entitled: "Analysis of the characteristics and patterns of the most popular deepfakes and analysis of existing deepfake data sets", financed by the Upper Silesian-Zagłębie Metropolis as part of the Metropolitan Science Support Fund Program.

Ograniczenia dotyczą jakości sygnału, struktury zbioru i standaryzacji. Cechy sensoryczne i resamplingu tracą czułość w materiałach o bardzo niskiej przepływności, w krótkich klipach  $\leq 5$  s oraz przy silnie zniekształconych gradientach. Przewaga wariantów syntetycznie wygenerowanych i kontrolowanie zdegradowanych może wzmacniać obecność określonych łańcuchów przetwarzania po stronie treści syntetycznych; stratyfikacja oraz testy przecieków tożsamości to ryzyko ograniczają, lecz go nie eliminują. Unifikacja do  $1280 \times 720$  i 30 kl./s ułatwia porównywalność, jednak częściowo maskuje artefakty specyficzne dla natywnych profili o wysokim zakresie dynamiki i formatów niestandardowych.

Kierunki dalszych badań obejmują rozwój wersji DFRWv2 o skali co najmniej 500 000 klipów z szerszą reprezentacją generatorów dyfuzyjnych oraz systematycznymi scenariuszami ponownej rejestracji i multimodalności audio-wideo, integracją atlasu z metrykami synchronii audiowizualnej i markerami akustycznymi nieciągłości artykulacji, rozszerzenie protokołów o testy odporności na ataki adaptacyjne i antyforensykę z naciskiem na transferowalność oraz scenariusze międzyplatformowe, a także opracowanie otwartego, znormalizowanego formatu raportu kryminalistycznego zawierającego  $\theta$ ,  $p_{df}$ ,  $p_{real}$ ,  $\Delta p$ , PR, 95% CI, lokalizację artefaktów i metadane strumienia bitowego oraz systematyczne badania kalibracji niepewności z uwzględnieniem ECE i pAUC w niskich FPR.

Przedstawiona analiza potwierdza, że cechy jakości obrazu i artefaktów przetwarzania tworzą przenośny, interpretowalny i operacyjnie użyteczny fundament detekcji treści syntetycznych. Wykazano, że wybrane metryki zachowują stabilność w typowych łańcuchach platformowych, a ich połączenie z sygnałami strumienia bitowego i sygnaturami sensorycznymi zwiększa odporność oraz wiarygodność wnioskowania. Opracowany na zbiorze DFRW atlas cech stanowi praktyczny punkt wyjścia do budowy hybrydowych systemów detekcji o podwyższonej przejrzystości i wartości dowodowej, zgodnie z potrzebami inżynierii bezpieczeństwa i kryminalistyki multimedialnej.

## Podziękowania i źródła finansowania

Publikacja powstała w ramach projektu pt.: „Analiza cech i wzorców najpopularniejszych deepfake’ów oraz analiza zbiorów danych istniejących deep fake”, finansowanego ze środków Górnośląsko-Zagłębiowskiej Metropolii w ramach Programu Metropolitalny Fundusz Wspierania Nauki.

## Abbreviations and explanations / Skróty i wyjaśnienia

AUC	Area Under the (ROC) Curve Pole pod krzywą ROC, miara jakości klasyfikatora
AV	Audio-Visual Dane audiowizualne (obraz + dźwięk)
AV1	AOMedia Video 1 (modern high-compression video) Nowoczesny kodek wideo o wysokiej kompresji
Apple M2 Ultra	High-performance computing chip used for AI generation Wysokowydajny układ obliczeniowy używany do generacji AI
AR	Augmented Reality Rzeczywistość rozszerzona, filtry/platformowe efekty
Banding	Color banding artifact Artefakt widoczny jako pasma w jednolitych gradientach
BEC	Business Email Compromise Rodzaj oszustwa biznesowego
Benjamini–Hochberg	Benjamini–Hochberg procedure (false discovery rate control) Metoda kontroli fałszywych odkryć w testach statystycznych
Bitrate (Mb/s)	Bit rate in megabits per second Przepływność wideo, miara ilości danych na sekundę
Bitstream	Compressed binary data stream Strumień zakodowanych danych wideo zawierający metadane
BLIINDS II / V-BLIINDS	Image quality metrics based on natural scene statistics. Metryki jakości obrazu oparte na statystykach naturalnych scen
Blocking	Blocking artifact Artefakt kompresji objawiający się widocznymi blokami
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator Bezreferencyjna metryka jakości obrazu oparta na NSS
C2PA	Coalition for Content Provenance and Authenticity Standard oznaczania pochodzenia i modyfikacji treści cyfrowych
CFA	Color Filter Array Filtr mozaikowy matrycy kamery
Celeb-DF	Challenging deepfake dataset with high photorealism Trudny zbiór danych deepfake o wysokim fotorealizmie
Checkerboard	Artifact occurring during upsampling Artefakt powstający przy upsamplingu (siatka szachownicy)
CPBD	Cumulative Probability of Blur Detection Miara percepcyjnego rozmycia obrazu
CRF	Constant Rate Factor, parametr kompresji w H.264/H.265
Cross-dataset	Testing the model on datasets other than the training set Testowanie modelu na innych zbiorach niż treningowe
Cross-model	Testing robustness against generators other than those used for training Testowanie odporności na inne generatory niż te użyte do treningu
DCT	Discrete Cosine Transform used in JPEG compression Dyskretna transformata kosinusowa stosowana w kompresji JPEG
DeepFaceLab	Popular tool for generating deepfakes Popularne narzędzie do generowania deepfake
Deepfake	Synthetic video generated using AI methods Syntetyczne wideo generowane metodami AI
DeeperForensics-1.0	Deep fake media generated using deep learning Zbiór danych do badania odporności detekcji

DFDC	Deepfake Detection Challenge Duży benchmark wideo
DFRW	DeepFake RealWorld Dataset (opracowany w ramach niniejszej pracy)
Diffusion model	A class of modern generative models (diffusion) Klasa nowoczesnych modeli generatywnych (dyfuzja)
DTS / PTS	Decoding Time Stamp / Presentation Time Stamp Znaczniki czasu dekodowania i prezentacji ramek
$\Delta p$	Difference in feature frequency between deepfake and real Różnica częstości cechy między deepfake a real
ECE	Expected Calibration Error Miara błędu kalibracji modelu
EfficientNet	Deep learning model used for image detection Model głębokiego uczenia używany w detekcji obrazu
EMO LipSync	Emotion-based Lip Synchronization Model Model synchronizacji ruchu ust z audio
ESW	Edge profile width according to ISO 12233 Szerokość profilu krawędzi wg ISO 12233
EXIF	Exchangeable Image File Format recorded by imaging devices Metadane zapisywane przez urządzenia rejestrujące obraz
Face2Face / First-Order Motion Model	Face and motion animation models Modele animacji twarzy i ruchu
FaceForensics++ (FF++)	Popular dataset for deepfake detection Popularny zbiór danych do detekcji deepfake
FFT	Fast Fourier Transform Szybka transformata Fouriera
FDR	False Discovery Rate Kontrola fałszywych odkryć
FPR	False Positive Rate Częstość fałszywych alarmów
FPS	Frames Per Second Liczba klatek na sekundę
GAN	Generative Adversarial Network Klasyczna metoda tworzenia deepfake
GOP	Group of Pictures Struktura ramek wideo
GPT-3 / GPT-4	Large AI language models, an example of scaling computational power Duże modele językowe AI, przykład skalowania mocy obliczeń
Grad-CAM	XAI tool visualizing model attention Narzędzie XAI wizualizujące uwagę modelu
H.264 / H.265	Popular video compression codecs Popularne kodeki kompresji wideo
HDR	High Dynamic Range Poszerzony zakres jasności
HeyGen	Commercial deepfake generation platform Komercyjna platforma generowania deepfake
I3D	3D-CNN model used in video analysis Model 3D-CNN stosowany w analizie wideo
InsightFaceSwap / SimSwap	Face swapping tools Narzędzia zamiany twarzy
ISO 12233	Standard for measuring image sharpness Norma pomiaru ostrości obrazu

IQA	Image Quality Assessment Ocena jakości obrazu
Jitter	Temporal instability in recordings Niestabilność czasowa w nagraniach
LIME	XAI tool for explaining model decisions Narzędzie XAI wyjaśniające decyzje modelu
LPIPS	Learned perceptual image similarity metric Uczona percepcyjna metryka podobieństwa obrazu
LSE-C / LSE-D	Metrics of audio-video synchronization inconsistency Miary niespójności synchronii audio-wideo
MAD	Median Absolute Deviation Odporna miara rozproszenia
Metadane bitstreamowe	Information stored in the encoded video stream Informacje zapisane w strumieniu zakodowanego wideo
MOV / MP4 / WebM	Popular video file containers Popularne kontenery plików wideo
MS-SSIM	Multi-scale image quality metric Wieloskalowa metryka jakości obrazu
MTF50	ISO sharpness measure, frequency at 50% contrast drop Miara ostrości wg ISO, częstotliwość przy spadku kontrastu o 50%
NeRF	Neural Radiance Fields Metoda generacji scen 3D
NIQE	No-reference quality metric based on NSS Bezreferencyjna miara jakości oparta na NSS
NSS (Natural Scene Statistics)	Natural image statistics Statystyki naturalnych obrazów
OpenFace	Tool for face identity clustering Narzędzie do klasteryzacji tożsamości twarzy
ORB / SIFT	Feature detection and matching algorithms in images Algorytmy detekcji i dopasowania cech w obrazie
OSINT	Open-source intelligence (OSINT) gathering from web and media Pozyskiwanie informacji z otwartych źródeł (sieć, media)
pAUC	Partial AUC at low error rates Częściowe AUC w niskich poziomach błędu
p_df / p_real	Feature frequencies in deepfake and real classes Częstości cech w klasach deepfake i real
Pika Labs	AI video generation platform Platforma generowania wideo AI
PIQE	Perceptual image quality metric Miara percepcyjnej jakości obrazu
Platt scaling	Model calibration improvement method Metoda poprawy kalibracji modeli
PolitiFact / Snopes / AFP FactCheck	Fact-checking services Serwisy weryfikacji informacji
PR (Prevalence Ratio)	Ratio of p_df / p_real, feature discriminative strength Stosunek p_df / p_real, siła odróżniania cechy
PRNU	Sensor fingerprint used in forensics Podpis sensora wykorzystywany w forensyce
PSNR	Image quality metric with respect to a reference signal Miara jakości obrazu względem sygnału odniesienia
pHash	Perceptual image hash for duplicate detection Percepcyjny skrót obrazu do detekcji duplikatów

RAFT	Model for dense optical flow estimation Model do wyznaczania gęstego przepływu optycznego
Re-capture	Re-recording a device screen with a camera Ponowne nagranie ekranu urządzenia kamerą
Reinactment	Controlling facial expressions or movements in a recorded video Sterowanie mimiką lub ruchem twarzy osoby w nagraniu
Ringing	Edge ringing caused by compression Oscylacje wokół krawędzi wynikające z kompresji
Rolling shutter	Distortions caused by successive pixel readout Zniekształcenia wynikające z kolejnego odczytu pikseli
Runway Gen-2 / Gen-3	AI video generation platforms Platformy generowania wideo AI
RTX 4090	High-performance GPU for AI generation Wysokowydajna karta graficzna do generacji AI
SIFT	Invariant feature detection algorithm Algorytm wykrywania cech niezmienniczych
SimSwap++	Modern face swapping system Nowoczesny system zamiany twarzy
Snorkeling / jitter	Temporal instabilities in video Niestabilności obrazu w czasie
Sora	Video generation model by OpenAI Model generowania wideo od OpenAI
SPS/PPS	Parameters describing H.264/H.265 codec configuration Parametry opisujące konfigurację kodeka H.264/H.265
SSIM	Structural similarity (SSIM) metric Strukturalna miara podobieństwa obrazu
Stable Video Diffusion	Diffusion-based video generation model Model dyfuzyjny generujący wideo
Temperature scaling	Technique for improving model calibration Technika poprawy kalibracji modeli
Tenengrad	Gradient energy-based sharpness measure Miara ostrości oparta na energii gradientu
TikTok / Twitter / Reddit	Platforms serving as OSINT material sources Platformy będące źródłem materiałów OSINT
UIQI	Universal Image Quality Index Uniwersalny wskaźnik jakości obrazu
VAE	Variational autoencoder, generative AI model Wariacyjny autoenkoder, generatywny model AI
VMAF	Video quality metric developed by Netflix Wideo-metryka jakości opracowana przez Netflix
VP9	Video codec developed by Google Kodek wideo rozwijany przez Google
Wang–Bovik	Classical blockiness artifact metric Klasyczna metryka artefaktów blokowych
Wav2Lip	Audio-to-lip synchronization model Model synchronizacji ust z dźwiękiem
XAI	Explainable artificial intelligence Wyjaśnialna sztuczna inteligencja
XceptionNet	CNN model used in deepfake detection Model CNN używany w detekcji deepfake
Y'CbCr	Color space used in video (luminance + chrominance) Przestrzeń barw stosowana w wideo (luminancja + chrominancje)

$\theta$ (theta)	Decision threshold of a feature Próg decyzyjny danej cechy
95% CI	95% confidence interval Przedział ufności 95%

## Literature / Literatura

- [1] Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P. i in., *Language models are few-shot learners*, „Advances in neural information processing systems” 2020, 33, 1877–1901, <https://doi.org/10.48550/arXiv.2005.14165>.
- [2] Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F.L. i in., *Gpt-4 technical report*, 2023, <https://doi.org/10.48550/arXiv.2303.08774>.
- [3] Perov I., Gao D., Chervoniy N., Li K., Marangond, S., Um C., i in., *DeepFaceLab: Integrated, flexible and extensible face-swapping framework*, 2020, <https://doi.org/10.48550/arXiv.2005.05535>.
- [4] Roop GitHub, <https://github.com/s0md3v/roop> [dostęp: 10.10.2025].
- [5] Chen R., Chen X., N, B., Ge Y., Simswap: An efficient framework for high fidelity face swapping, Proceedings of the 28th ACM international conference on multimedia, 2020, 2003–2011.
- [6] Chesney B., Citron D., *Deep fakes: A looming challenge for privacy, democracy, and national security*, „Calif. L. Rev.” 2019, 107, 1753, <https://doi.org/10.2139/ssrn.3213954>.
- [7] Jędrasiak K., *Audio stream analysis for deep fake threat identification*, „Civitas et Lex” 2024, 41(1), 21–35, <https://doi.org/10.31648/cetl.9684>.
- [8] Bikhchandani S., Hirshleifer D., Welch I., *A theory of fads, fashion, custom, and cultural change as informational cascades*, „Journal of political Economy” 1992, 100(5), 992–1026, <https://doi.org/10.1086/261849>.
- [9] Badawy A., Lerman K., Ferrara E., *Who falls for online political manipulation?*, Companion proceedings of the 2019 world wide web conference, 162–168.
- [10] DiResta R., *The supply of disinformation will soon be infinite*, „The Atlantic” 2020, 20(9).
- [11] Europol, *Online fraud schemes: a web of deceit*, IOCTA 2023.
- [12] United Nations Office on Drugs and Crime, *Emerging threats: The intersection of criminal and technological innovation in the use of automation and AI*, Cybercrime Technical Brief Series, 2025.
- [13] Federal Bureau of Investigation Internet Crime Complaint Center, Internet Crime Report, FBI IC3, 2023.
- [14] Europol, *Facing reality? Law enforcement and the challenge of deepfakes*, an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg, 2022.
- [15] Dolhansky B., Bitton J., Pflaum B., Lu J., Howes R., Wang M., Ferrer C.C., *The deepfake detection challenge (dfdc) dataset*, 2020, <https://doi.org/10.48550/arXiv.2006.07397>.
- [16] Korshunov P., Marcel S., (2018). *Deepfakes: a new threat to face recognition? assessment and detection*, 2018, <https://doi.org/10.48550/arXiv.1812.08685>.
- [17] Wang Z., Bovik A.C., *A universal image quality index*, „IEEE signal processing letters” 2002, 9(3), 81–84, <https://doi.org/10.1109/97.995823>.
- [18] Guo C., Pleiss G., Sun Y., Weinberger K.Q., *On calibration of modern neural networks*, International conference on machine learning, 2017, pp. 1321–1330.
- [19] Matern F., Riess C., Stamminger M., *Exploiting visual artifacts to expose deepfakes and face manipulations*, IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019, pp. 83–92.
- [20] Guarnera L., Giudice O., Battiato S., *Deepfake detection by analyzing convolutional traces*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 666–667.
- [21] Farid H., *Image forgery detection*, „IEEE Signal processing magazine” 2009, 26(2), 16–25, <https://doi.org/10.1109/MSP.2008.931079>.
- [22] Ray S., *Applied photographic optics*, Routledge, 2002.
- [23] Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J., *Deepfakes and beyond: A survey of face manipulation and fake detection*, „Information Fusion” 2020, 64, 131–148, <https://doi.org/10.1016/j.inffus.2020.06.014>.
- [24] Neves J.C., Tolosana R., Vera-Rodriguez R., Lopes V., Proença H., Fierrez J., *Gan fingerprints in face image synthesis*, Multimedia Forensics, Springer, Singapore 2022, 175–204.
- [25] Agarwal S., Farid H., Gu Y., He M., Nagano K., Li H., *Protecting world leaders against deep fakes*, CVPR workshops, Vol. 1, No. 38, 2019.
- [26] Verdoliva L., *Media forensics and deepfakes: an overview*, „IEEE journal of selected topics in signal processing” 2020, 14(5), 910–932, <https://doi.org/10.1109/JSTSP.2020.3002101>.
- [27] Haliassos A., Vougioukas K., Petridis S., Pantic M., *Lips don't lie: A generalisable and robust approach to face forgery detection*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5039–5049.
- [28] Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Nießner M., *Faceforensics++: Learning to detect manipulated facial images*, Proceedings of the IEEE/CVF international conference on computer vision, 2019, 1–11.
- [29] Li Y., Yang X., Sun P., Qi H., Lyu S., *Celeb-df: A large-scale*

- challenging dataset for deepfake forensics*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 3207–3216.
- [30] Jiang L., Li R., Wu W., Qian C., Loy C.C., Deepforensics-1.0: A large-scale dataset for real-world face forgery detection, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 2889–2898.
- [31] Ho J., Jain A., Abbeel P., Denoising diffusion probabilistic models, Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, 33, 6840–6851.
- [32] Mildenhall B., Srinivasan P.P., Tancik M., Barron J.T., Ramamoorthi R., Ng R., *Nerf: Representing scenes as neural radiance fields for view synthesis*, „Communications of the ACM” 2021, 65(1), 99–106, <https://doi.org/10.1145/3503250>.
- [33] Durall R., Keuper M., Keuper J., *Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 7890–7899.
- [34] Ricker J., Damm S., Holz T., Fischer A., *Towards the detection of diffusion model deepfakes*, <https://doi.org/10.48550/arXiv.2210.14571>.
- [35] Yang X., Li Y., Lyu S., *Exposing deep fakes using inconsistent head poses*, ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2019, 8261–8265.
- [36] Wang T., Liao X., Chow K.P., Lin X., Wang, Y., *Deepfake detection: A comprehensive survey from the reliability perspective*, „ACM Computing Surveys” 2024, 57(3), 1–35, <https://doi.org/10.1145/3699710>.
- [37] Cozzolino D., Pianese A., Nießner M., Verdoliva L., *Audio-visual person-of-interest deepfake detection*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, 943–952.
- [38] Tariq S., Lee S., Woo S., *One detector to rule them all: Towards a general deepfake attack detection framework*, Proceedings of the web conference. 2021, pp. 3625–3637.
- [39] Lu Y., Luo R., Ebrahimi T., *A Novel Framework for Assessment of Learning-based Detectors in Realistic Conditions with Application to Deepfake Detection*, 2022, <https://doi.org/10.48550/arXiv.2203.11797>.
- [40] Cao X., & Gong N.Z., *Understanding the security of deepfake detection. In International Conference on Digital Forensics and Cyber Crime*, Springer International Publishing, Cham 2021, 360–378.
- [41] Ju Y., Jia S., Ke L., Xue H., Nagano K., Lyu S., *Fusing global and local features for generalized ai-synthesized image detection*, 2022 IEEE International Conference on Image Processing, 2022, 3465–3469.
- [42] Carlini N., Jagielski M., Choquette-Choo C.A., Paleka D., Pearce W., Anderson H. i in. (2024, May). *Poisoning web-scale training datasets is practical*, 2024 IEEE Symposium on Security and Privacy, 2024, 407–425.
- [43] Doshi-Velez F., Kim B., *Towards a rigorous science of interpretable machine learning*, 2017, <https://doi.org/10.44550/arXiv.1702.08608>.
- [44] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D., *Grad-cam: Visual explanations from deep networks via gradient-based localization*, Proceedings of the IEEE international conference on computer vision, 2017, 618–626.
- [45] Odena A., Dumoulin V., Olah C., *Deconvolution and checkerboard artifacts*, „Distill” 2016, 1(10), e3, <https://doi.org/10.23915/distill.00003>.
- [46] Zhang K., Zuo W., Zhang L., *Learning a single convolutional super-resolution network for multiple degradations*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 3262–3271.
- [47] Mittal A., Moorthy A.K., Bovik A.C., *No-reference image quality assessment in the spatial domain*, „IEEE Transactions on image processing” 2012, 21(12), 4695–4708, <https://doi.org/10.1109/TIP.2012.2214050>.
- [48] Teed Z., Deng J. (2020, August). *Raft: Recurrent all-pairs field transforms for optical flow*, *European conference on computer vision*, Springer International Publishing, Cham 2020, 402–419.
- [49] Chung J.S., Zisserman A., (2016, November). *Lip reading in the wild*, *Asian conference on computer vision*, Springer International Publishing, Cham 2016, 87–103.
- [50] Prajwal K.R., Mukhopadhyay R., Namboodiri V.P., Jawahar C.V., *A lip sync expert is all you need for speech to lip generation in the wild*, Proceedings of the 28th ACM international conference on multimedia, 2020, 484–492.
- [51] Baltrušaitis T., Ahuja C., Morency L.P. (2019), *Multi-modal machine learning: A survey and taxonomy*, „IEEE transactions on pattern analysis and machine intelligence” 2019, 41(2), 423–443, <https://doi.org/10.1109/TPAMI.2018.2798607>.
- [52] Frank J., Eisenhofer T., Schönherr L., Fischer A., Kolossa D., Holz T., *Leveraging frequency analysis for deep fake image recognition*, International conference on machine learning, 2020, pp. 3247–3258.
- [53] Mittal A., Soundararajan R., Bovik A.C., *Making a “completely blind” image quality analyzer*, „IEEE Signal processing letters” 2012, 20(3), 209–212, <https://doi.org/10.1109/LSP.2012.2227726>.
- [54] Venkatanath N., Praneeth D., Sumohana S.C., Swarup S.M., *Blind image quality evaluation using perception based features*, IEEE Twenty first national conference on communications (NCC), 2015, 1–6.
- [55] Saad M.A., Bovik A.C., Charrier C., *Blind image quality assessment: A natural scene statistics approach in the DCT domain*, „IEEE transactions on Image Processing” 2012, 21(8), 3339–3352, <https://doi.org/10.1109/TIP.2012.2191563>.
- [56] Narvekar N.D., Karam L.J., *A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)*, „IEEE Transactions on Image Processing” 2011, 20(9), 2678–2683, <https://doi.org/10.1109/TIP.2011.2131660>.
- [57] Pech-Pacheco J.L., Cristóbal G., Chamorro-Martinez J., Fernández-Valdivia J., *Diatom autofocusing in brightfield microscopy: a comparative study*, IEEE Proceedings 15th

- International Conference on Pattern Recognition, ICPR-2000, Vol. 3, 314–317.
- [58] Pertuz S., Puig D., Garcia M.A. (2013). *Analysis of focus measure operators for shape-from-focus*, „Pattern Recognition” 2013, 46(5), 1415–1432, <https://doi.org/10.1016/j.patcog.2012.11.011>.
- [59] ISO 12233:2017. Photography – Electronic still picture imaging – Resolution and spatial frequency responses.
- [60] ITU-T H.264:2019. Advanced Video Coding for Generic Audiovisual Services. ITU-T Recommendation H.264 | ISO/IEC 14496-10.
- [61] ISO/IEC 23008-2:2017. Information technology – High efficiency video coding.
- [62] Damera-Venkata N., Kite T.D., Geisler W.S., Evans B.L., Bovik A.C., *Image quality assessment based on a degradation model*, „IEEE transactions on image processing” 2000, 9(4), 636–650, <https://doi.org/10.1109/83.841940>.
- [63] Mahdian B., Saic S., *Detection of copy-move forgery using a method based on blur moment invariants*, „Forensic science international” 2007, 171(2-3), 180–189, <https://doi.org/10.1016/j.forsciint.2006.11.002>.
- [64] Bianchi T., Piva A., *Image forgery localization via block-grained analysis of JPEG artifacts*, „IEEE Transactions on Information Forensics and Security” 2012, 7(3), 1003–1017, <https://doi.org/10.1109/TIFS.2012.2187516>.
- [65] Amerini I., Ballan L., Caldelli R., Del Bimbo A., Serra G., *A sift-based forensic method for copy-move attack detection and transformation recovery*, „IEEE transactions on information forensics and security” 2011, 6(3), 1099–1110, <https://doi.org/10.1109/TIFS.2011.2129512>.
- [66] Rublee E., Rabaud V., Konolige K., Bradski G., *ORB: An efficient alternative to SIFT or SURF*. In IEEE 2011 International conference on computer vision, 2011, 2564–2571.
- [67] Lukas J., Fridrich J., Goljan M., *Digital camera identification from sensor pattern noise*, „IEEE Transactions on Information Forensics and Security” 2006, 1(2), 205–214, <https://doi.org/10.1109/TIFS.2006.873602>.
- [68] Popescu A.C., Farid H., *Exposing digital forgeries by detecting duplicated image regions*, Computer Science Technical Reports, Dartmouth College 2004.
- [69] Benjamini Y., Hochberg Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, „Journal of the Royal statistical society: series B (Methodological)” 1995, 57(1), 289–300, <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- [70] EU DisinfoLab Annual Conference. Conference Dossier 2023 <https://www.disinfo.eu/wp-content/uploads/2023/11/Disinfo2023-conference-dossier.pdf>, [dostęp:10.10.2025].
- [71] Lowe D.G., *Distinctive image features from scale-invariant keypoints*, „International journal of computer vision” 2004, 60(2), 91–110, <https://doi.org/10.1023/B:VISI.0000029664.99615.94>.

**KAROL JĘDRASIAK, PH.D.** – academic teacher, didactician and manager, author of over 88 scientific publications, including 3 scientific monographs with high citability. The author’s scientific experience includes participation in 26 research and development projects, also as a manager. Active participant in more than 30 scientific conferences and symposia. Expert of the WSL2014-2020 ROP, member of the Steering Committee of the Game INN Sector Program and the Society for Image Processing. As a result of his previous work and cooperation with industry, he participated in the development of 32 claims of intellectual property rights (6 granted patents, 8 patent applications, 18 design registration rights). Specialist in computer vision, computer graphics, artificial intelligence tools, computer, database and sensor system development. Since 2008, he has held management positions in private companies. For many years he was CEO of VR Technology, a company developing algorithms in the area of data analysis, commercializing innovative solutions in virtual reality technology and simulation as well as coaching systems. In 2024, he was awarded the Medal of the National Education Commission and a certificate of appreciation for his exceptional contribution to the successful implementation of the “e-Instructor Certification Programme” by NATO Assistant Secretary General for Operations Tom Goffus.

**DR INŻ. KAROL JĘDRASIAK** – nauczyciel akademicki, dydaktyk i menadżer, autor ponad 88 publikacji naukowych, w tym trzech monografii naukowych, o wysokiej cytowalności. Jego doświadczenie naukowe obejmuje udział w 26 projektach badawczo-rozwojowych, w tym w charakterze kierownika projektu. Aktywny uczestnik w ponad 30 konferencjach i sympozjach. Ekspert RPO WSL/POIR, członek Komitetu Sterującego Programu Sektorowego GameINN, członkiem Towarzystwa Przetwarzania Obrazów. W rezultacie dotychczasowej pracy oraz współpracy z przemysłem powstało 32 zastrzeżeń prawa własności intelektualnej (6 przyznanych patentów, 8 zgł. patentowych, 18 praw z rejestracji wzoru przemysłowego). Specjalista w zakresie wizji komputerowej, wirtualnej rzeczywistości, narzędzi sztucznej inteligencji, wytwarzania systemów informatycznych, bazodanowych i sensorycznych. Od 2008 piastował stanowiska kierownicze w przedsiębiorstwach prywatnych. Przez wiele lat był Prezesem Zarządu spółki VR-Technology zajmującej się opracowywaniem algorytmów z zakresu analizy danych oraz komercjalizacją innowacyjnych rozwiązań z zakresu technologii wirtualnej rzeczywistości oraz profesjonalnych systemów symulacyjnych i trenażerowych. W 2024 został uhonorowany Medalem Komisji Edukacji Narodowej oraz certyfikatem uznania za wyjątkowy wkład w pomyślną realizację „Programu Certyfikacji e-Instruktorów”, nadanym przez Asystenta Sekretarza Generalnego NATO ds. Operacji Toma Goffusa.