UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Aasifar El Ouahabi, Amina. *How Common Are AI-Generated Images on the Web?*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337358

under the terms of the CC BY-SA license.

How Common Are AI-Generated Images on the Web?

Amina Aasifar El Ouahabi

Abstract — The rapid spread of AI-generated Images online raises growing concerns about visual authenticity and trust in digital media. This study measures the prevalence of AI-generated images across publicly accessible websites in five sectors: news and media, art and illustration, education, e-commerce, and tourism and travel. A dataset of 4,520 images was collected from 25 websites using an automated crawler and classified as generated or real using SPAI, a spectral learning classifier and state-of-the-art detection method. Through data analysis and classification, research qüestions are examined: (1) the variation in AI image prevalence across sectors, (2) the concentration of AI-generated images in specific page zones, and (3) the evolution of synthetic image use over time between 2020 and 2025 using Wayback Machine snapshots. The results reveal a global AI rate of 19.2%, led unexpectedly by the news and media sector at 24.7%, with a pronounced concentration in article content zones. In contrast, creative sectors show substantially lower rates. Together, these results expose the uneven adoption of AI-generated imagery across the web and highlight the need for robust detection tools capable of operating beyond controlled lab conditions, as synthetic content becomes an increasingly routine part of the digital landscape.

Index Terms — AI-generated images detection, generative AI, image forensics, synthetic media.

Resum — La propagació ràpida d'imatges generades per IA en línia desperta una preocupació creixent sobre l'autenticitat visual i la confiança en els mitjans digitals. Aquest estudi mesura la prevalença d'imatges generades per IA en llocs web d'accés públic dins de cinc sectors: notícies i mitjans de comunicació, art i il·lustració, educació, comerç electrònic, i turisme i viatges. Es va recopilar un conjunt de dades de 4.520 imatges procedents de 25 llocs web mitjançant un rastrejador automatitzat, i es van classificar com a generades o reals utilitzant SPAI, un classificador d'aprenentatge espectral d'última generació. A través de l'anàlisi de dades i la classificació, s'examinen tres hipòtesis: (1) la variació de la prevalença d'imatges d'IA entre sectors, (2) la concentració d'imatges generades per IA en zones específiques de la pàgina, i (3) l'evolució de l'ús d'imatges sintètiques al llarg del temps entre 2020 i 2025 utilitzant captures de la Wayback Machine. Els resultats mostren una taxa global d'IA del 19,2%, amb el sector de notícies i mitjans al capdavant amb un 24,7%, contràriament al que s'esperava, juntament amb una forta concentració en les zones de contingut dels articles. En canvi, sectors creatius presenten taxes substancialment més baixes. En conjunt, els resultats posen de manifest l'adopció desigual d'imatges generades per IA a través del web i subratllen la necessitat d'eines de detecció robustes que funcionin fora de les condicions controlades de laboratori, a mesura que el contingut sintètic esdevé una part cada cop més normal del paisatge digital.

Paraules clau — detecció d'imatges generades per IA, IA generativa, anàlisi forense d'imatges, mitjans sintètics.

1 Introduction

Are you sure it is not AI-generated? This question has become one of the central reliability concerns of the current decade. As state-of-the-art generative models reach a level of realism that even makes synthetic faces indistinguishable from real photographs [1], the public's ability to tell them apart has been declining rather than improving [2], raising pressing questions about trust in visual media. AI-generated images are now common across many digital environments, and as models improve, they become harder to distinguish from real photographs.

Before automation reached this field, the images published were selected and manually adapted by professionals until they matched the intended function, a process shaped by commercial constraints such as price and licensing [3]. Around 2022 and 2023, modern generative AI such as DALL·E 2, Stable Diffusion, and Midjourney opened a new path: the main change is not only which images are chosen, but that many are now generated directly from text and context. This lowers the cost of visual production, yet it also raises risks of misinformation and tensions over authorship

- *Contact e-mail: 1672430@uab.cat*
- *Supervised by: Alexandra Gomez Villa (Computer Vision Center)*
- *Academic Year 2025/26*

and creative work [4].

This project focuses on the presence of AI-generated images on the web, particularly across publicly accessible websites from different sectors. The potential of generative AI models is so great that the presence of their creations can be either pervasive or virtually absent, depending on the type of website and sector it belongs to. Therefore, the project sets out to understand how widely distributed AI-generated images are and whether their presence varies systematically across different sectors of the web.

Beyond this, the project also examines whether generative AI models have had a major impact since their emergence by analysing earlier snapshots of websites through the Wayback Machine [22].

2 Aim and Objectives

The main aim of this project is to study the presence of AI-generated images on the web, focusing on publicly accessible websites from different sectors. As text-to-image models have become more accessible and capable of producing highly realistic outputs, organisations may potentially be using synthetic images on their websites. However, the extent of this adoption remains unknown.

Hypotheses and planned analyses

Based on these objectives, the study is guided by two working hypotheses and a set of descriptive analyses.

The central hypothesis (Hc) is that the presence of AI-generated images is not uniform across sectors. Websites where creativity and visual production play a central role, such as art and illustration, are expected to show a higher proportion of synthetic imagery than those where content is factual and time-sensitive, such as news and media. This expectation reflects the different incentives that each sector faces when selecting or producing visual content.

Building on this, a more specific hypothesis (H1) is also investigated, concerning the spatial distribution of synthetic content: AI-classified images are expected to concentrate in specific regions of a webpage rather than being allocated uniformly. This hypothesis is treated as an observational relationship rather than a causal effect, given that results are based on classifier predictions rather than manually verified ground truth.

Alongside these hypotheses, the study carries out descriptive analyses, including a comparison of the five sectors according to their proportion of AI-classified images, the examination whether each sector behaves as expected - that is whether a notable presence of synthetic images appears where it would not be anticipated, or the reverse – and lastly, the temporal adoption of synthetic imagery across selected websites from a specific category, relating the observed curve to relevant historical milestones such as the public release of the main generative models.

3 Related works

AI image generation: The development of AI-generated images has its roots in the introduction of Variational Autoencoders (VAEs), proposed by Kingma and Welling as one of the first frameworks capable of learning compact latent representations of images through an encoder-decoder structure [6]. A major step towards photorealistic synthesis came with Generative Adversarial Networks (GANs), proposed by Goodfellow et al., which introduced an adversarial training process between a generator and a discriminator [7]. A further milestone came with latent diffusion models, which became the foundation of modern text-to-image systems such as Stable Diffusion, DALL·E 2, and Midjourney [8]. From 2022 onwards, these tools made AI image generation widely accessible, lowering the cost of visual content production to the point where synthetic images could be integrated into the websites and digital communication materials that this project analyses. More recently, second-generation systems such as Stable Diffusion 3, Flux, and Midjourney v6 have pushed photorealism further, making synthetic images increasingly indistinguishable from real photographs [9].

AI Image detection: As generative models have improved, so has the need to detect their outputs. Research shows that AI-generated images often may contain frequency-domain traces invisible to the human eye, which became the technical basis for automated detection methods [10]. Subsequent work highlighted the importance of media in maintaining trust in digital information, as manipulated visual content becomes increasingly difficult to identify manually [11]. However, most detection methods have only been tested on controlled benchmark datasets, not on real-world web images. This project uses SPAI [12], a state-of-the-art spectral classifier. Rather than learning to detect artifacts from specific generative models, SPAI takes a self-supervised approach: it models the spectral distribution of real images using masked spectral learning - a frequency reconstruction pretext task - and detects AI-generated images as out-of-distribution samples of this learned model. This design makes it generalisable to unseen T2I generators, since the spectral distribution of real images remains invariant regardless of which model produced the synthetic content. Evaluated across 13 generative approaches, including Stable Diffusion, DALL·E, Midjourney, Flux, and Adobe Firefly, it achieves a 5.5% absolute improvement in AUC over the previous state-of-the-art, addressing the gap between laboratory evaluation and real-world application.

While detection of AI-generated images remains challenging, a complementary solution to gain institutional traction is watermarking, embedding provenance metadata directly into content at creation time [13]. The C2PA standard, backed by Adobe, Microsoft, Google, and OpenAI, defines cryptographically signed watermarks - often imperceptible to the human eye - that identify AI-generated outputs [14], and the EU AI Act of 2024 has made such watermarking a legal requirement for synthetic content [15]. Nevertheless, watermarking is not yet a generalized practice, in this work we focus on detecting any AI-generated image, whether it

carries a watermark or not.

Presence of AI-Generated Images in Real Digital Environments: Despite the fast growth of synthetic visual content, empirical evidence on its actual presence online remains limited. Studies on social media suggest that the phenomenon is already significant: around 12% of images shared during the 2024 US election on X were estimated to be generated by AI [16], suggesting that synthetic imagery is no longer marginal in digital communication. The most relevant work is Wei and Tyson's study of AI-generated content on the Pixiv platform [17], which provides direct evidence that synthetic images are transforming how visual material is produced and consumed in specific digital communities. However, existing evidence remains concentrated on social media and creative platforms, leaving the broader web largely unexplored [5].

Overall, previous research shows that AI-generated images are becoming more realistic and that detection methods can identify some of their traces. However, there is still limited empirical evidence on how frequently these images appear across different types of publicly accessible websites, how their presence varies across sectors, and how this use has evolved over time. This project addresses this gap by collecting a real-world web image database and analysing the proportion of synthetic content across website categories.

4 Methodologies

To measure the prevalence of AI-generated images across web sectors, this study combines automated image collection, AI-based classification, temporal dataset and statistical analysis.

4.1 Website Selection and Dataset Construction

Five categories were defined to support a structured comparison across different sectors: news and media, art and illustration, education, e-commerce, and tourism and travel. These categories represent contrasting contexts in which visual content is used on the web, from factual and editorial environments to more commercially driven ones, making them suitable for testing whether AI-generated image prevalence varies across sectors. For instance, news websites have editorial commitments to authenticity, which would suggest a lower presence of synthetic imagery, whereas e-commerce or tourism websites face stronger commercial incentives to use visually appealing content, where AI-generated images might be more common. Categories such as corporate or social media websites were discarded due to stronger access restrictions and a higher proportion of non-relevant visual material.

For each category, five candidate websites were evaluated based on three criteria: (a) stable access to visual content, (b) enough usable images, and (c) low duplication. All 25 candidate websites were retained as final sources, with a collection target of approximately 150 to 200 images per website to limit imbalances between categories.

4.2 Image Collection and Cleaning

Images were collected from each selected website using an automated crawler developed in Python [18], combining BeautifulSoup [19] for static content and Playwright [20] for dynamically rendered pages. The crawler extracts images from standard HTML tags, CSS background properties, and semantic meta tags, navigating up to a defined maximum number of pages per website to ensure broad coverage without excessive requests.

Once collected, images were filtered to retain only content images. First, duplicates were removed using exact content hashing and perceptual hashing. Corrupted files and interface elements such as logos, icons, and banners were discarded. Images below 100x100 pixels or 5KB were also excluded; they were found to correspond mostly to thumbnails and small interface elements that do not carry enough visual information for the classifier to work reliably. In addition, only JPEG, PNG, and WebP formats were retained; formats like SVG do not store photographic content and are therefore not useful. Finally, pages containing explicit or adult content were also excluded, including images of minors, based on keyword filtering applied to URL paths and page titles to keep the dataset clean and handle sensitive content responsibly.

For each retained image, metadata was saved, including source URL, file format, dimensions, HTML tag, parent tag, and CSS class names. The final dataset was organised by category and website to facilitate the per-sector and per-site analyses described in Section 4.5.

4.3 AI-Image Classification

Images were classified using SPAI [12], a pre-trained image classifier that assigns each image a score between 0 (real) and 1 (AI-generated) based on how well its spectral patterns match those of real images. One of its advantages for this project is that it works with images of any resolution without needing to resize them first, which matters when dealing with web images that vary a lot in size and quality.

The operational threshold was selected through three stages of testing. First, an exploratory evaluation on GenAI-Bench samples [21] showed that detection accuracy varied across generators. Second, a balanced benchmark on 120 images from six generators reached 84.2% accuracy at a threshold of 0.30 (see Appendix A, Fig A.1). Third, a mixed-data validation on 200 images (120 AI-generated and 80 real) achieved 82.5% accuracy at a threshold of 0.35 (see Appendix A, Fig. A.2). The difference between 0.30 and 0.35 was small, and 0.35 showed better behaviour on real images, so it was kept as the final working threshold.

The full implementation of the crawler and the classification pipeline is publicly available at [52].

4.4 Temporal Dataset via the Wayback Machine

To study how the presence of AI-generated images has changed over time, a temporal dataset was built using the

Wayback Machine [22]. It is a digital archive of the World Wide Web that allows users to view and interact with past versions of websites. Users can enter a URL and select a date range to see what a website looked like in the past, even if it no longer exists on the live web. Archived snapshots were collected annually from 2020 to 2025, with a target of up to 50 images per website per year, using the same filtering and classification steps described in sections 4.2 and 4.3 with the same operational threshold (0.35).

The news and art categories were initially chosen because they represent two opposite cases: news is expected to show a low AI presence due to editorial standards, while art is one of the sectors where synthetic imagery is most likely to appear early. However, the number and quality of archived snapshots varied considerably between websites, and several sites - particularly in the art category - did not have enough usable images per year for a reliable annual estimate. Because of these limitations, discussed in detail in Section 6.4, the temporal analysis was focused on the News and Media sector and on the three outlets that combined sufficient archival coverage with a measurable AI presence.

The aim of this part of the study is to trace the shape of the adoption curve over time and relate it to known milestones in the public release of the main generative models.

4.5 Statistical Analysis

The statistical analysis is structured around hypotheses and descriptive analyses defined in Section 2. The main metric used throughout is the AI rate, which simply measures the percentage of images classified as AI-generated out of the total classified images for a given website, sector, or page zone.

To address Hc, the AI rate is compared across the five sectors using a descriptive ranking of weighted rates, calculated as total AI-classified images divided by total classified images per sector. Results are also broken down by individual websites within each category to capture within-sector variation. A manual inspection of a sample of images classified as AI-generated was also carried out to check whether the predictions looked visually plausible.

To address H1, each image is assigned to one of six-page zones - Article content, Hero/Banner, Thumbnail gallery, Product catalog, Header/Nav, or Sidebar/Related (Figure 1) - based on a keyword heuristic applied to HTML attributes such as class names, parent tags, and element IDs.

Images without recoverable zone metadata are excluded and reported separately. The relationship between page zone and AI classification is tested using a chi-squared test of independence, with standardised residuals used to identify which zones drive the association. A focused two-proportion test then compares the AI rate of the leading zone against all other zones combined, reporting the 95% confidence interval of the difference.

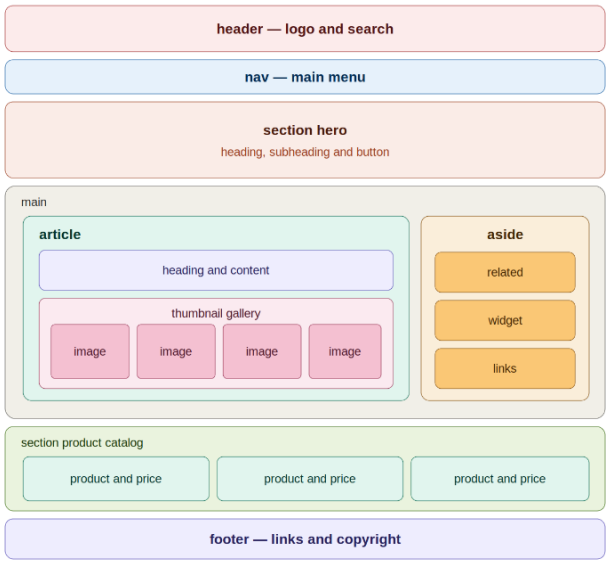


Fig. 1. The six-page zones used to classify each image, mapped to their underlying HTML5 semantic elements [50]

5 Results

This section presents the classification results obtained from the full web image dataset. The analysis is structured in five parts: an overall AI rate baseline, a sector-level comparison addressing the central hypothesis (Hc), a website-level breakdown, a spatial analysis addressing hypothesis H1, and a temporal analysis based on Wayback Machine [22] data.

5.1 Overall AI Rate

Out of the images classified across the 25 websites, 870 were predicted as AI-generated, giving a global AI rate of 19.2%. As shown in Figure 2, the distribution of classifier scores is strongly bimodal: most images concentrate near 0, while a smaller group clusters near 1.0. The y-axis uses a square root scale to make both groups visible.

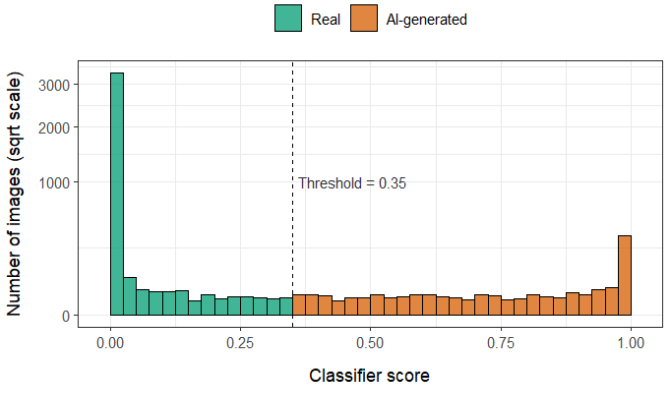


Fig. 2. Distribution of SPAI classifier scores across all 4,520 images collected from the 25 websites.

This figure acts as the baseline for the sector-level, website-

level, and page zone analyses presented in the following sections.

5.2 Sector-level variation (Hc)

The central hypothesis predicts that sectors with a more creative use of visual content, such as art, will show higher AI rates than sectors where content is primarily factual, such as news.

Figure 3 shows the distribution of AI-classified and real images across the five sectors. News/Media presents the highest AI rate at 24.7%, followed by Tourism (21.9%) and Education (21.6%), which are very close to each other. Art (14.1%) and E-Commerce (13.2 %) show the lowest rates, with E-commerce being almost half the rate of News/Media.

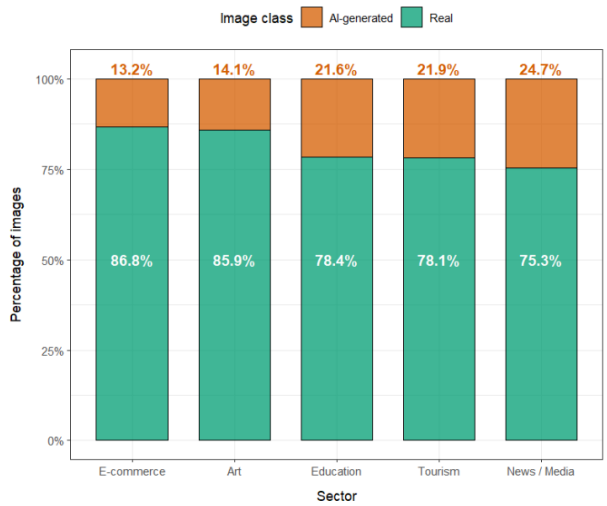


Fig.3. AI rate by sector. Proportion of images classified as AI-generated versus real across the five sectors.

The per-sector score distributions (Figure 4) follow the same bimodal pattern as the global one (Figure 2).

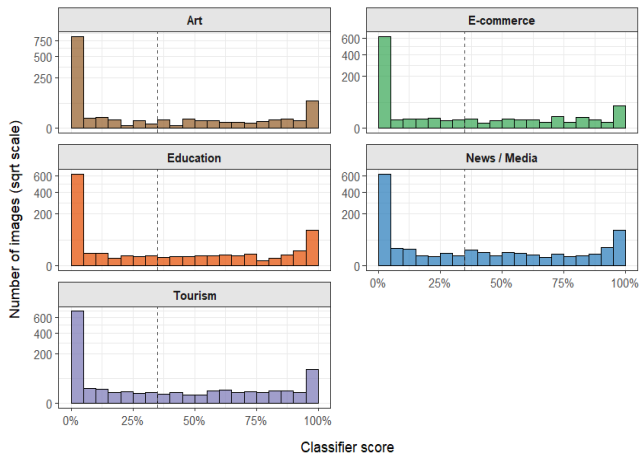


Fig. 4. Distribution of SPAI classifier scores within each of the five sectors.

The relative size of this high-score peak is larger in News/Media and Tourism than in Art and E-Commerce, which is consistent with the differences in AI rate observed in Figure 3.

Overall, the observed ranking places News/Media at the top and Art and E-commerce at the bottom, which differs from the order predicted by Hc.

5.3 AI Rate by Website

While the sector-level analysis provides a general overview, the AI rate varies considerably between individual websites within the same category. The following subsections present the results for each sector separately, identifying the websites with the highest and lowest AI rates.

5.3.1 News and Media

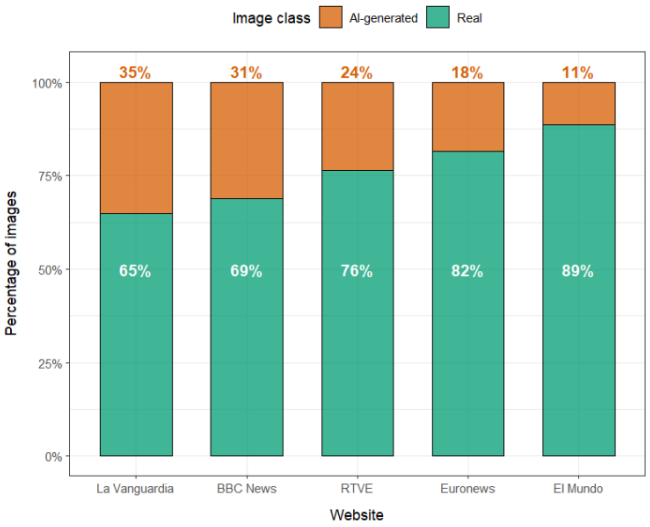


Fig. 5. AI rate per website in the News and Media sector.

Figure 5 shows the widest range of AI rates across websites of all five sectors. La Vanguardia [23] leads at 35%, followed by BBC UK News [24] (31%), RTVE [25] (24%), Euronews [26] (18%), and El Mundo [27] (11%).

The difference between the highest and lowest values is 24 percentage points, indicating considerable variation within the sector despite sharing a common editorial context.

5.3.2 Art and Illustration

As shown in Figure 6, the art and illustration category presents notable variation across websites. Public Domain Review [28] shows the highest AI rate at 30%, followed by DeviantArt [29] (21%) and Rawpixel [30] (20%), which are close to each other. ArtStation [31] (10%) and Old Book Illustrations [32] (4%) show considerably lower rates, with Old Book Illustrations presenting the lowest value across all art websites.

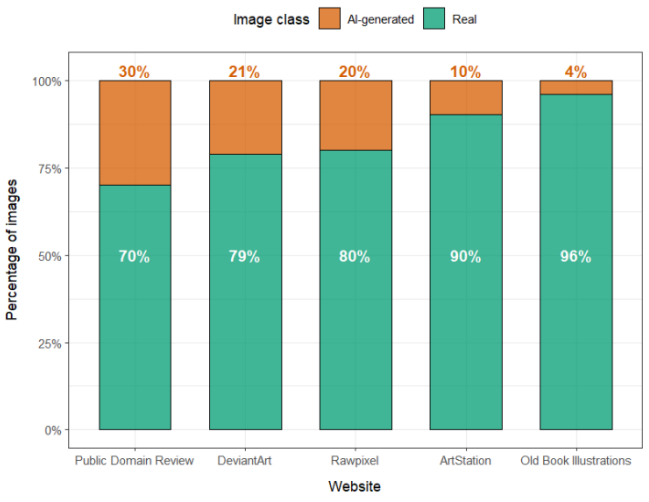


Fig. 6. AI rate per website in the Art and Illustration sector.

5.3.3 Education

The education category presents a relatively consistent pattern, with AI rates ranging from 13% to 31% (Figure 7).

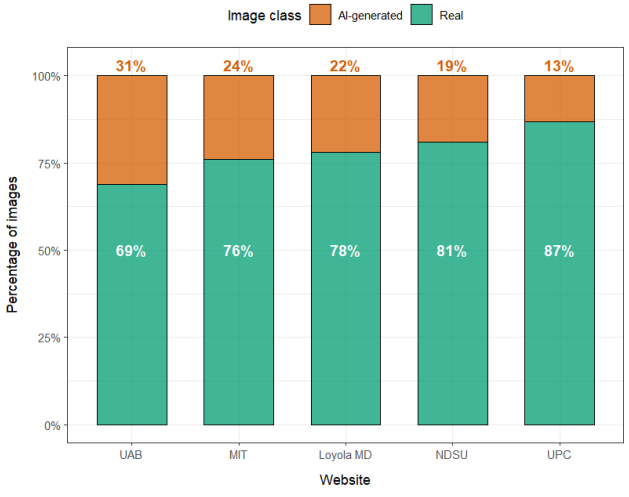


Fig. 7. AI rate per website in the Education sector.

The Universitat Autònoma de Barcelona [33] (UAB) shows the highest rate at 31%, followed by the Massachusetts Institute of Technology [34] (24%), Loyola University Maryland [35] (22%), and North Dakota State University [36] (19%). The Universitat Politècnica de Catalunya [37] presents the lowest rate at 13%. Compared to the news category the spread is narrower at 18 percentage points, suggesting more uniform behaviour across educational institutions.

5.3.4 E-commerce

Within e-commerce, AI rates range from 4% to 27%, one of the widest variations observed in this section (Figure 8). IKEA [38] leads at 27%, followed by SHEIN [39] (17%), while Sotheby's Realty [40] (8%), MediaMarkt [41] (6%), and Amazon [42] (4%) show considerably lower rates. Amazon presents the lowest value across the entire dataset, with a difference of 23 percentage points separating it from IKEA.

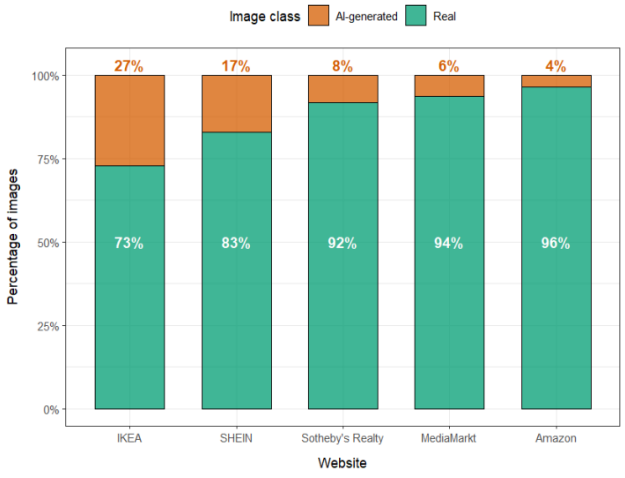


Fig. 8. AI rate per website in the E-Commerce sector.

5.3.5 Tourism and Travel

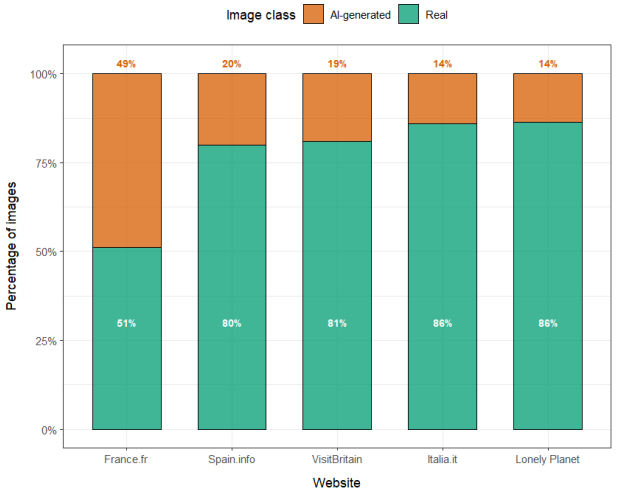


Fig. 9. AI rate per website in the Tourism and Travel sector.

The tourism and travel category is dominated by a single outlier: France.fr [43] presents an AI rate of 49%, the highest value observed across the entire dataset (Figure 9). However, the remaining websites show lower rates: Spain.info [44] (20%), VisitBritain [45] (19%), Italia.it [46] and Lonely Planet [47] (both 14%). Excluding France.fr, the other four websites cluster between 14% and 20%, a much narrower range than the category average suggests.

5.4 Spatial Distribution of AI-Generated Images (H1)

H1 predicts that AI-generated images will concentrate in specific regions of a webpage rather than being distributed uniformly across all zones. To test this, each image was assigned to one of six-page zones based on its HTML metadata. Images without recoverable zone information were excluded from this analysis and represent

approximately 29.6% of the total classified images.

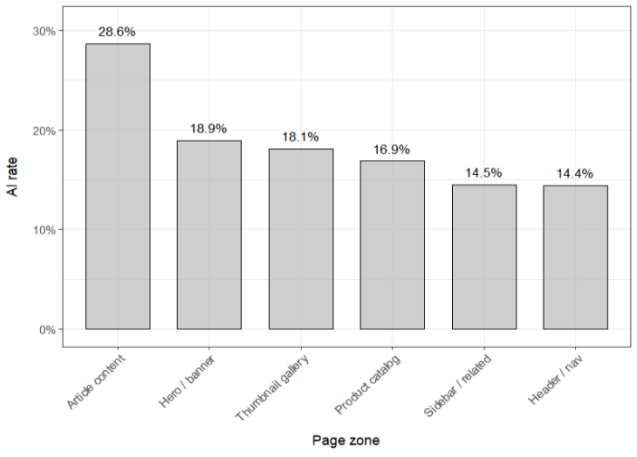


Fig. 10. AI rate by page zone.

As shown in Figure 10, Article content presents the highest AI rate at 28.6%, followed by Hero/Banner (18.9%), Thumbnail gallery (18.1%), Product catalog (16.9%), Sidebar/related (14.5%), and Header/nav (14.4%). The remaining five zones show similar rates, clustering between 14% and 19%. To test whether this distribution differs significantly from what would be expected by chance, a chi-squared test of independence was applied to the zone × class contingency table. The result was significant ($\chi^2(5) = 47.25$, $p \approx 5\times10^{-9}$), indicating that the AI rate is not uniformly distributed across page zones. Standardised residuals confirmed that Article content is the only zone driving the association, with a residual of +6.72 for AI-generated images, while all other zones showed residuals below |2|.

Also, a focused two-proportion test comparing *Article content* against other zones combined confirmed this result: *Article* content shows an AI rate of 28.6% versus 17.9% for the remaining zones, a difference of 10.7 percentage points (95% CI: [7.3%, 14.1%], $p < 0.001$).

Additionally, the breakdown of AI and real images counts per zone and per sector is provided in Appendix B, Table B.1 and Fig. B.1.

5.5 Temporal Analysis (Wayback Machine)

The temporal analysis covers archived snapshots of selected news websites from 2020 to 2025. Table 1 summarises the total images collected, the number classified as AI-generated, and the AI rate per year.

Year	Total images	AI-classified	AI-rate (%)
2020	41	5	12.2
2021	71	17	23.9
2022	78	25	32.1
2023	81	25	30.9
2024	96	25	26.0
2025	123	29	23.6

Table 1. Number of images collected and AI rate per year for news websites via the Wayback Machine [22].

As shown in Figure 11, the AI rate increases from 12.2% in 2020 to a peak of 32.1% in 2022. From 2022 onwards, the rate shows a moderate decline, reaching 23.6% in 2025.

However, the proportion alone does not capture the full picture. The total number of images collected per year increases considerably over the period, from 41 in 2020 to 123 in 2025, which affects the percentage directly. When looking at the absolute number of AI-classified images, the trend is consistently upward: from 5 images in 2020 to 29 in 2025, representing a nearly sixfold increase over the period.

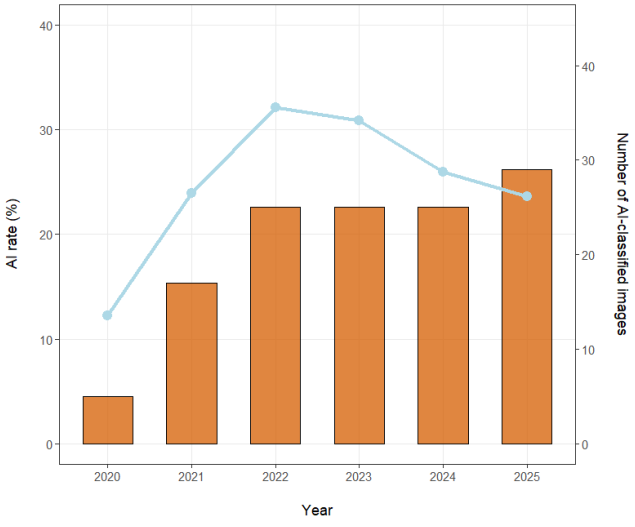


Fig. 11. AI rate (%) per year for news websites from 2020 to 2025 (blue line, left axis) and absolute number of AI-classified images (red bars, right axis).

6 DISCUSSIONS

6.1 Central Hypothesis (Hc): Sector-level Variation

Contrary to Hc, News/Media leads with a 24.7% AI rate, while Art (14.1%) and E-commerce (13.2%) sit at the bottom.

The low AI rates in art and e-commerce are inconsistent with the central hypothesis but can be explained. In art, Old Book Illustrations (4%) stands out as the lowest rate across the entire dataset, which is expected given that it hosts historical pre-digital material that is human made. The remaining art platforms show higher but still moderate rates, consistent with the results of Wei and Tyson [17], who observed that AI-generated content on Pixiv, despite growing rapidly, remained a minority and faced resistance from the human creator community. In e-commerce, Amazon (4%) and MediaMarkt (6%) show very low rates, which may reflect the dominance of standardized product photography where consistency and accuracy take precedence over visual impact. This contrasts sharply with IKEA (27%), whose lifestyle and aspirational imagery create more natural opportunities for synthetic visuals.

The sector-level ranking, however, only tells part of the story: within sectors spread is large (e.g La Vanguardia at 35% vs El Mundo at 11%), and the pattern repeats across all categories. Knowing a website's sector thus says little about its AI use – individual editorial decisions matter as much as sector membership.

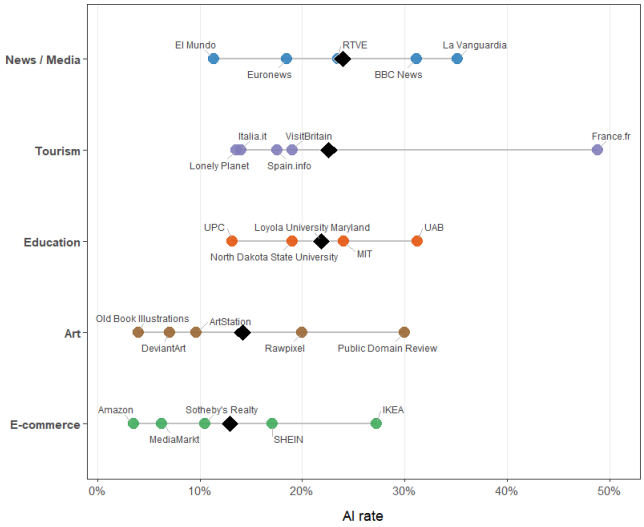


Fig. 12. AI rate per website across all five sectors. The black diamond marks the sector's mean. France.fr (Tourism, 49%) appears as a clear outlier at the far right.

France.fr deserves particular attention. With an AI rate of 49%, it is the single most extreme case in the entire dataset and has a disproportionate influence on the tourism sector average. Removing it would place tourism considerably closer to art and e-commerce in the ranking. The reasons behind this unusually high rate remain speculative without manual verification, but tourism promotion websites rely on aspirational and idealised visual content, where the pressure to present destinations in their best light may create stronger incentives to use synthetic imagery, consistent with the argument made by Feuerriegel et al. [4] that generative AI lowers the cost of producing visually appealing content in commercially driven contexts.

Finally, these results connect with the broader picture of AI adoption on the web. AlDahoul et al. [5] suggest that generative AI is increasingly shaping the visual landscape of the World Wide Web, and the global AI rate of 19.2% observed in this study provides empirical support for this trend. However, the strong within-sector variation and the influence of individual outliers such as France.fr suggest that this adoption is neither uniform nor predictable from sector membership alone, and that website-level analysis is necessary to capture the full picture.

6.2 Hypothesis H1: Spatial Distribution of AI-Generated Images

H1 predicted that AI-generated images would concentrate in specific regions of a webpage rather than distributed uniformly. The results support this hypothesis: the chi-squared test confirmed a significant association between page zone and AI classification. A focused two-proportion test showed that Article content presents a significantly higher AI rate than all other zones combined, with a difference of 10.7 percentage points.

A closer look at the content type of images classified as AI-generated in Article content zones across news websites reveals a clear pattern: feature and long-form content show one of the highest AI rate among editorial categories, at 34.5%, compared to 23.3% for standard news articles and only 6.2% for service or advisory content (see Figure 13).

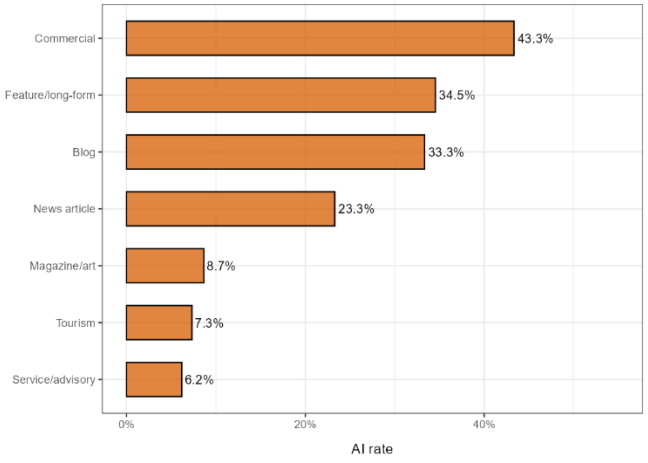


Fig. 13. AI rate by content type within Article content zones across news websites (n = 644 images, manually classified).

Commercial and promotional content shows the highest rate overall at 43.3%, though this figure is driven primarily by La Vanguardia, which has been identified as a high-AI outlier throughout this study. Magazine and art content (8.7%) and tourism-related content (7.3%) show the lowest rate, consistent with their more image-driven and visually controlled content style. This suggests that AI imagery concentrates on content where visual accuracy matters less than visual appeal, while breaking news, portraits, and documentary photography remain predominantly human-captured. If AI images were replacing documentary photography, a stronger signal would be expected in zones such as Hero/banner or Thumbnail gallery, but the data does not show this: Hero/banner (18.9%) and Thumbnail gallery (18.1%) remain below Article content and are not statistically distinguishable from the baseline.

This finding has an important implication. *Article* content is the zone where images are most directly associated with the meaning of the text, whether illustrating a news story, an educational concept, or a product description. The concentration of AI-generated images in this zone suggests that synthetic imagery is being used not only for decorative purposes, but as a direct visual complement to written content. This is consistent with the argument made by Frankowska-Takhari [3] that image selection in professional contexts is shaped by the functional role of the image within the content, and with Feuerriegel et al. [4], who suggest that generative AI lowers the cost of producing visual

content quickly and on demand.

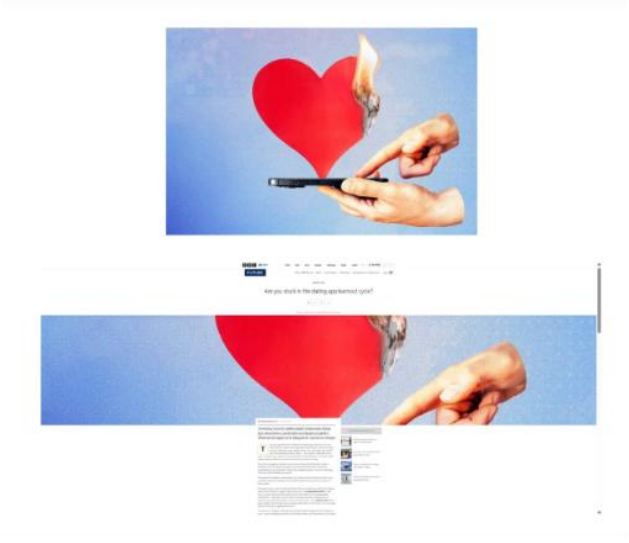


Fig. 14. (a) Image classified as AI-generated (score = 1.0) extracted from the BBC Future website; (b) webpage context showing the same image used as hero visual in an article about dating app behaviour.

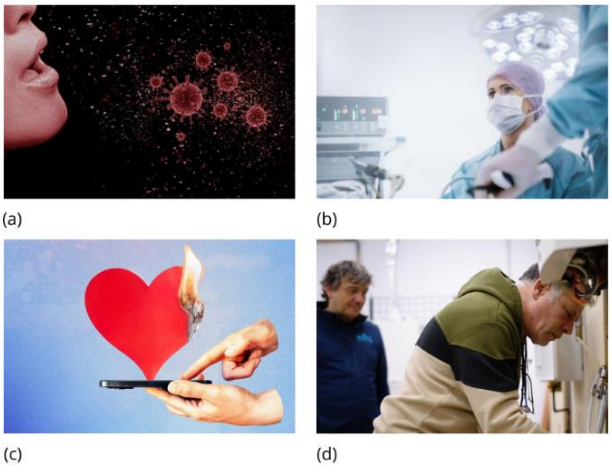


Fig. 15. Sample of images classified as AI-generated from Article content zones across news websites: (a) Euronews (score = 0.9999), (b) La Vanguardia (score = 1.0), (c) El Mundo (score = 0.9938), (d) BBC News (score = 1.0).

Figure 15 shows a sample of images classified as AI-generated at near-maximum confidence from *Article content* zones across news websites. Two of the four images appear conceptual or digitally generated in style, consistent with this interpretation. However, the remaining two appear photographic in nature, which raises the possibility - though it cannot be confirmed without manual verification - that some high-confidence predictions may correspond to heavily processed stock photographs rather than synthetically generated images. This observation is discussed further in Section 6.3.

The remaining zones show AI rates clustering between 14% and 19%, with no zone showing a residual above |2|. Hero/banner (18.9%) and Thumbnail gallery (18.1%) are the closest to Article content, but their difference is not statistically distinguishable from the baseline. This is noteworthy because hero images and banners are often considered the most visually prominent and commercially driven elements of a webpage, where AI adoption might be expected to be highest.

The pattern is consistent across sectors, as shown in Appendix B, Fig. B.1, although the specific zones contributing most to each category differ. In News/Media and Education, Article content is the dominant zone by volume, which may partly explain why these sectors show both high AI rates overall and a strong Article content signal. In E-commerce and Tourism, Product catalog and Hero/banner zones are more prominent, yet their AI rates remain below that of Article content.

One methodological consideration worth noting is that 29.6% of images could not be assigned to a zone due to missing or unrecoverable HTML metadata (see Appendix B, Table B.1). This is a known limitation of heuristic-based zone classification, particularly on modern websites that use obfuscated or dynamically generated CSS class names. While this exclusion does not invalidate the results, it introduces a degree of selection bias that should be acknowledged: the images included in the zone analysis may not be fully representative of the entire dataset.

6.3 Temporal Analysis (Wayback Machine)

The temporal results show two parallel trends that require a separate interpretation: a rate that peaks in 2022 and then declines, and an absolute count of AI-classified images that rises consistently throughout the period. The 2022 peak in the AI rate coincides with public release of the main text-to-image models [8][9], and with the moment when generative AI tools reached mass adoption across many sectors, including journalism [48].

BBC News, the only outlet with enough images every year, shows this most clearly: its rate goes from around 20% in 2020-2021 up to 35% in 2022, with non-overlapping confidence intervals, which means the increase is unlikely to be due to chance (see Appendix C, Fig. C.1). After 2022 the rate declines while the absolute count keeps growing, for two reasons. First, the yearly image total also grows, pulling the percentage down. Second, from 2023 most major newsrooms adopted policies restricting AI images [49] – the Associated Press barred AI-created publishable images, and outlets such as Deutsche Welle and WIRED set explicit limits [48]. These restrictions coincide with the period when the rate begins to fall, though causality cannot be established since this study measures classifier output rather than editorial decisions.

La Vanguardia shows a surprisingly high rate (38%-52%, see Appendix C, Fig. C.2) from 2021 onwards with no valid 2020 baseline and no clear trend in either direction. Since this signal is present even before any public generative

model existed, it is more likely explained by the site's visual style - stock or conceptual photography that SPAI tends to flag as synthetic - than by genuine early adoption, consistent with the classifier false-positive floor discussed in Section 6.4 and 4.

RTVE only becomes readable from 2024, with rates of 9% (n = 11) and 14% (n = 14) in 2024 and 2025 respectively (see Appendix C, Fig. C.3). Its earlier years have fewer than five images per year, making any estimate too imprecise to interpret. Its inclusion in the temporal dataset is justified by its relevance as a major public broadcaster, but its Wayback Machine [22] coverage is too sparse to contribute to the temporal analysis in a meaningful way before 2024.

The absolute count - from 5 to 29 AI-classified images between 2020 and 2025 - is the more robust finding, as it is unaffected by changes in total image coverage, even accounting for the classifier's false-positive rate; the consistent upward direction of this count is hard to attribute entirely to noise. This is consistent with the broader global trend of rapidly growing AI image production [5]. In this context, a sustained increase in AI-classified content across news websites is the expected outcome rather than a surprising one, and the absolute count provides stronger evidence of this trend than the rate alone.

The three outlets behave very differently - BBC News rises around 2022, La Vanguardia stays high throughout, and RTVE is only readable from 2024 – reinforcing the website-level effect noted in Section 6.1 and adding that sites differ not only in how much AI content they show, but in whether a temporal change can be detected at all.

6.4 Reliability and limitations

Reliability: the analysis treats SPAI's output as ground truth, but its 82.5% validation accuracy (Section 4.3) means roughly one in six predictions may be wrong. As noted in Section 6.2, two of four sampled high-confidence images looked photographic, suggesting heavily processed stock photos can be misclassified as AI. All findings should be read as classifier-based estimates, not verified ground truth.

Limitations: Five main limitations should be acknowledged. First, the temporal dataset depends on the Wayback Machine [22], which archives web pages opportunistically rather than systematically, meaning there may be gaps in coverage and missing snapshots that do not necessarily reflect the absence of content during that period. Additionally, the same image quality filters applied in Section 4.3 were also applied to archived snapshots – older snapshots frequently contain degraded or low-resolution images that do not meet the minimum thresholds, further reducing the number of usable images per year. This explains why several websites could not contribute enough images for reliable annual estimates, and why RTVE only becomes readable from 2024. The temporal trends observed should therefore be interpreted as indicative rather than definitive.

Second, this study measures what the classifier detects, not what website editors intended. The fact that AI-generated images concentrate in *Article content* zones suggests they are used to illustrate written content, but this cannot be confirmed without directly examining editorial decisions or content policies.

Third, the zone classification relies on HTML metadata such as class names and element IDs, which could not recover 29.6 % of images. When this metadata is missing or ambiguous – as is common on modern websites that use obfuscated or auto-generated code – neither the HTML attributes nor the surrounding links provide enough information to assign an image to a zone reliably.

Fourth, the dataset covers 25 websites across five sectors. This provides a structured basis for comparison but is not representative of the web, and findings should be read as sector-specific observations rather than universal estimates.

Finally, data collection was constrained by the access restrictions imposed by many websites and APIs. Several platforms either block automated crawlers entirely or impose rate limits that reduce the number of images that can be collected, which may introduce a systematic bias towards websites with more open access policies.

7 Conclusions

Synthetic imagery has quietly become a measurable part of the visual web. Across 4,520 images collected from 25 websites, this study found a global AI rate of 19.2%, a figure that, on its own, answers the central question: AI-generated images are not marginal.

The distribution of this presence, however, defies expectations. News and media – not art or illustration – show the highest AI rate (24.7%) (Hc), and AI-generated images concentrate in *Article content* zones rather than decorative areas (H1), suggesting they are used to illustrate written content on demand. The temporal analysis shows that this adoption peaked around 2022 with the release of major generative models and has since stabilized in rate while growing in absolute numbers.

Taken together, the evidence points to a visual web that has already changed: not through a dramatic shift, but through gradual, sector-specific adoption that standard detection tools are not yet designed to measure at scale.

8 Future Work

While this study provides a first empirical estimate of AI-generated image prevalence across the web, several directions could extend and strengthen the findings.

Several directions could strengthen the findings: expanding the dataset to more websites and categories (e.g social media, corporate communications); manually verifying a representative heuristic with a visual layout analyser to recover the unassigned images; extending the temporal analysis to

better-archived sectors or alternative sources such as Common Crawl or national digital libraries; and combining detection with provenance metadata such as C2PA watermarks [14] to correct misclassification at scale.

References

[1] S. J. Nightingale and H. Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 119, no. 8, e2120481119, 2022. Available: https://www.pnas.org/doi/10.1073/pnas.2120481119

[2] Conjointly, "Can consumers spot AI and real photos in 2024?," Oct. 2024. [Online]. Available: https://conjointly.com/blog/generative-ai-survey-2024/. Accessed: Apr. 2026.

[3] S. Frankowska-Takhari, "Selecting and tailoring of images for online news content: a mixed-methods investigation of the needs and behaviour of image users in online journalism," Ph.D. dissertation, Dept. Comput. Sci., City Univ. London, London, UK, 2018. Available: https://openaccess.city.ac.uk/id/eprint/21947/

[4] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," *Bus. Inf. Syst. Eng.*, vol. 66, no. 1, pp. 111–126, 2024. Available: https://link.springer.com/article/10.1007/s12599-023-00834-7

[5] N. AlDahoul, J. Hong, M. Varvello et al., "Towards a World Wide Web powered by generative AI," *Sci. Rep.*, vol. 15, art. 7251, 2025. Available: https://www.nature.com/articles/s41598-025-90705-w

[6] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013. Available: https://arxiv.org/abs/1312.6114

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *Proc. Neural Information Processing Systems (NeurIPS)*, 2014. Available: https://arxiv.org/abs/1406.2661

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Available: https://arxiv.org/abs/2112.10752

[9] P. Esser, S. Kulal, A. Blattmann et al., "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis (Stable Diffusion 3)," *Proc. International Conference on Machine Learning (ICML)*, 2024. Available: https://arxiv.org/abs/2403.03206

[10] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot... for Now," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Available: https://arxiv.org/abs/1912.11035

[11] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020. Available: https://arxiv.org/abs/2001.06564

[12] D. Karageorgiou, S. Papadopoulos, I. Kompatsiaris, and E. Gavves, "Any-Resolution AI-Generated Image Detection by Spectral Learning," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Available: https://arxiv.org/abs/2411.19417

[13] J. Cao, Q. Li, Z. Zhang, J. Ni, and R. Lu, "Secure and Robust Watermarking for AI-Generated Images: A Comprehensive Survey," *arXiv preprint arXiv:2510.02384*, 2025. Available: https://arxiv.org/abs/2510.02384

[14] Coalition for Content Provenance and Authenticity (C2PA), "C2PA Technical Specification." [Online]. Available: https://c2pa.org/specifications/. Accessed: Apr. 2026.

[15] European Parliament, "Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, 2024. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[16] Z. Chen, J. Ye, B. Tsai, E. Ferrara, and L. Luceri, "From Pixels to Polls: Prevalence, Spreaders, and Responses to AI-Generated Images on X During the 2024 U.S. Election," *arXiv preprint arXiv:2502.11248*, 2025. Available: https://arxiv.org/abs/2502.11248

[17] Y. Wei and G. Tyson, "Understanding the Impact of AI-Generated Content on Social Media: The Pixiv Case," 2024. Available: https://arxiv.org/abs/2402.18463

[18] Python Software Foundation, "Python Language Reference." [Online]. Available: https://www.python.org. Accessed: Apr. 2026.

[19] L. Richardson, "Beautiful Soup Documentation." [Online]. Available: https://www.crummy.com/software/BeautifulSoup/. Accessed: Apr. 2026.

[20] Microsoft, "Playwright: Fast and reliable end-to-end testing for modern web apps." [Online]. Available: https://playwright.dev. Accessed: Apr. 2026.

[21] B. Li, Z. Lin, D. Pathak, J. Li, Y. Fei, K. Wu, T. Ling, X. Xia, P. Zhang, G. Neubig, and D. Ramanan, "GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation," *arXiv preprint arXiv:2406.13743*, 2024. Available: https://arxiv.org/abs/2406.13743

[22] Internet Archive, "Wayback Machine." [Online]. Available: https://web.archive.org. Accessed: Apr. 2026.

[23] La Vanguardia. [Online]. Available: https://www.lavanguardia.com. Accessed: Apr. 2026.

[24] BBC News. [Online]. Available: https://www.bbc.co.uk/news. Accessed: Apr. 2026.

[25] RTVE. [Online]. Available: https://www.rtve.es. Accessed: Apr. 2026.

[26] Euronews. [Online]. Available: https://www.euronews.com. Accessed: Apr. 2026.

[27] El Mundo. [Online]. Available: https://www.elmundo.es. Accessed: Apr. 2026.

[28] The Public Domain Review. [Online]. Available: https://publicdomainreview.org. Accessed: Apr. 2026.

[29] DeviantArt. [Online]. Available: https://www.deviantart.com. Accessed: Apr. 2026.

[30] Rawpixel. [Online]. Available: https://www.rawpixel.com. Accessed: Apr. 2026.

[31] ArtStation. [Online]. Available: https://www.artstation.com. Accessed: Apr. 2026.

[32] Old Book Illustrations. [Online]. Available: https://www.oldbookillustrations.com. Accessed: Apr. 2026.

[33] Universitat Autònoma de Barcelona (UAB). [Online]. Available: https://www.uab.cat. Accessed: Apr. 2026.

[34] Massachusetts Institute of Technology (MIT). [Online]. Available: https://www.mit.edu. Accessed: Apr. 2026.

[35] Loyola University Maryland. [Online]. Available: https://www.loyola.edu. Accessed: Apr. 2026.

[36] North Dakota State University (NDSU). [Online]. Available: https://www.ndsu.edu. Accessed: Apr. 2026.

[37] Universitat Politècnica de Catalunya (UPC). [Online]. Available: https://www.upc.edu. Accessed: Apr. 2026.

[38] IKEA. [Online]. Available: https://www.ikea.com. Accessed: Apr. 2026.

[39] SHEIN. [Online]. Available: https://www.shein.com. Accessed: Apr. 2026.

[40] Sotheby's International Realty. [Online]. Available: https://www.sothebysrealty.com. Accessed: Apr. 2026.

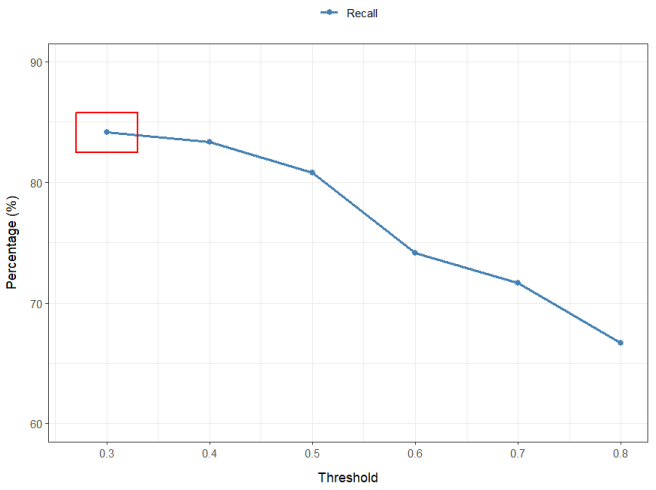
[41] MediaMarkt. [Online]. Available: https://www.mediamarkt.es.

Accessed: Apr. 2026.
[42] Amazon. [Online]. Available: https://www.amazon.es. Accessed: Apr. 2026.
[43] France.fr - Explore France. [Online]. Available: https://www.france.fr. Accessed: Apr. 2026.
[44] Spain.info. [Online]. Available: https://www.spain.info. Accessed: Apr. 2026.
[45] VisitBritain. [Online]. Available: https://www.visitbritain.com. Accessed: Apr. 2026.
[46] Italia.it. [Online]. Available: https://www.italia.it. Accessed: Apr. 2026.
[47] Lonely Planet. [Online]. Available: https://www.lonelyplanet.com. Accessed: Apr. 2026.
[48] T. J. Thomson, S. Boczkowski, and C. Mitchelstein, "Generative Visual AI in News Organizations: Challenges, Opportunities, Perceptions, and Policies," *Journalism Practice*, 2024. Available: https://www.tandfonline.com/doi/full/10.1080/21670811.2024.2331769
[49] C. Merrefield, "Researchers compare AI policies and guidelines at 52 news organizations," *The Journalist's Resource*, Dec. 2023. [Online]. Available: https://journalistsresource.org/home/generative-ai-policies-newsrooms/. Accessed: Apr. 2026.
[50] World Wide Web Consortium (W3C), "HTML5: A vocabulary and associated APIs for HTML and XHTML," W3C Recommendation, 2014. [Online]. Available: https://www.w3.org/TR/html5/
[51] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927. Available: https://doi.org/10.1080/01621459.1927.10502953
[52] aminasifar1, "AI-generated image prevalence on the web — crawler and classification pipeline," GitHub repository, 2026. [Online]. Available: https://github.com/aminasifar1/tfg-how-common-are-ia-images-on-web Accessed: Jun. 2026.

APPENDIX A

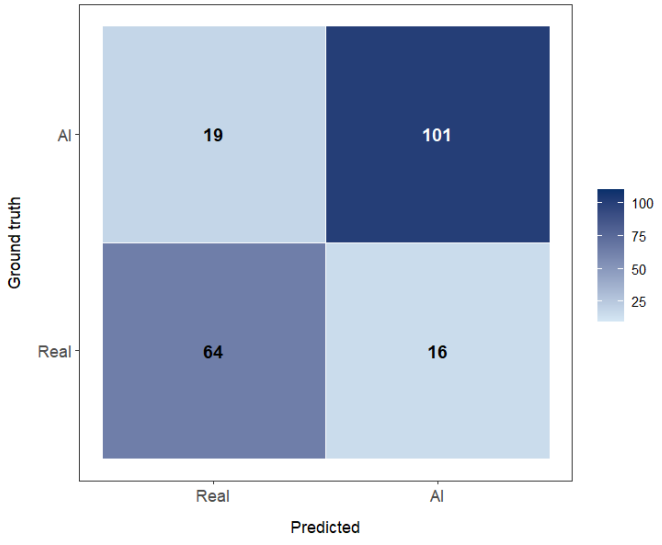
- THRESHOLD ANALYSIS

Fig. A.1. Accuracy across thresholds on the balanced GenAI-Bench subset (120 images, six generators).



This figure shows the classification accuracy and recall of SPAI across threshold values between 0.30 to 0.80, evaluated on a balanced subset of 120 images from six generative models. Accuracy peaks at 84.2% at a threshold of 0.30, after which it decreases consistently as the threshold increases. The red square highlights the best-performing threshold. This evaluation informed the selection of 0.35 as the final working threshold, which showed comparable accuracy while exhibiting better behaviors on real images.

Fig. A.2. Confusion matrix for the mixed-data validation dataset (200 images) at threshold 0.35.

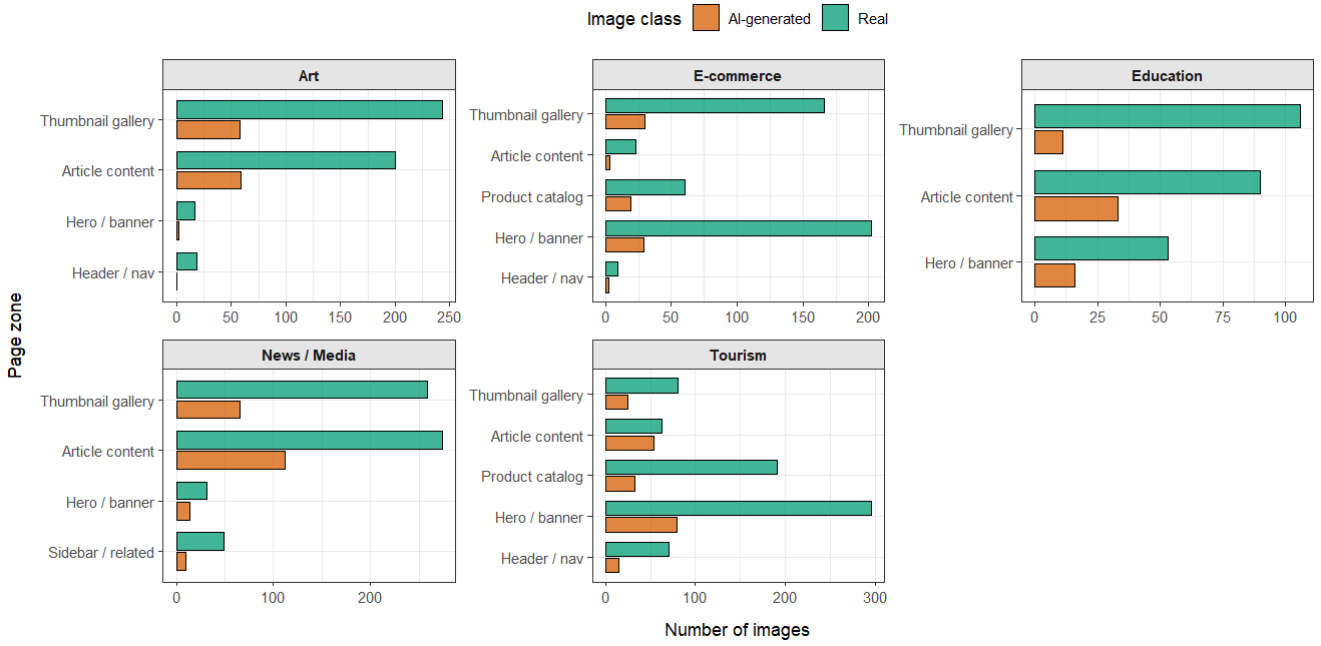


This figure shows the classification results on a mixed dataset of 200 images (120 AI-generated and 80 real) using the final working threshold of 0.35. The model correctly classified 101 AI-generated images and 64 real images, with 19 false negatives and 16 false positives, yielding an overall accuracy of 82.5%. The relatively higher false negative count reflects the conservative nature of the chosen threshold, which prioritizes avoiding false positives on real images.

APPENDIX B

- **SPATIAL DISTRIBUTION OF AI-GENERATED IMAGES**

Fig. B.1. Distribution of real and AI-generated images across page zones and website categories



This figure shows the number of images classified as real and AI-generated in each page zone for the five analyzed website categories: Art, E-commerce, Education, News/Media, and Tourism. The distribution highlights category-specific differences in page structure and image placement. News/Media and Education datasets are dominated by images located within Article content, whereas E-commerce and Tourism contain a larger proportion of images in Product catalog and Hero/banner zones. These differences provide context for the variation in AI-generated image rates discussed in Section 6.2.

Table B.1. Distribution of real and AI-generated images across page zones.

Page zone	Total	AI-generated	Real	AI rate (%)
Article content	911	261	650	28.6
Hero/banner	739	140	599	18.9
Thumbnail gallery	1,044	189	855	18.1
Product catalog	302	51	251	16.9
Sidebar / related	69	10	59	14.5
Header / nav	118	17	101	14.4
Assigned total	**3,183**	**668**	**2,515**	**21.0**
No metadata	1,337			
Grand total	**4,520**			

Article content presents the highest AI rate (28.6%), while the remaining five zones cluster between 14% and 19%. A total of 1,337 images (29.6% of the dataset) could not be assigned to a zone due to missing or unrecoverable HTML metadata and are excluded from the zone-level analysis.

APPENDIX C

- **TEMPORAL ANALYSIS PER OUTLET**

Table A.1. Annual image counts and AI rates per outlet for the Wayback Machine [22] temporal dataset (2020–2025)

	Year	Total images	AI	AI rate (%)
BBC News	2020	25	5	20.0
	2021	46	9	19.6
	2022	46	16	34.8
	2023	39	11	28.2
	2024	46	13	28.3
	2025	46	10	21.7
La Vanguardia	2020	4	0	0.0
	2021	16	8	50.0
	2022	23	9	39.1
	2023	27	14	51.9
	2024	28	11	39.3
	2025	40	15	37.5
RTVE	2020	0	0	0.0
	2021	1	0	0.0
	2022	1	0	0.0
	2023	2	0	0.0
	2024	11	1	9.1
	2025	14	4	14.3

For each of the three News/Media outlets included in the temporal analysis, the table shows the total number of images collected per year, the number classified as AI-generated by SPAI (threshold = 0.35), and the resulting AI rate (%). Years with fewer than five images indicate sparse archival coverage and should be treated as indicative only; the corresponding confidence intervals are too wide to support reliable estimates (see Figures C.1-C.3). RTVE has no archived images for 2020 in the dataset.

Figures C.1-C.3

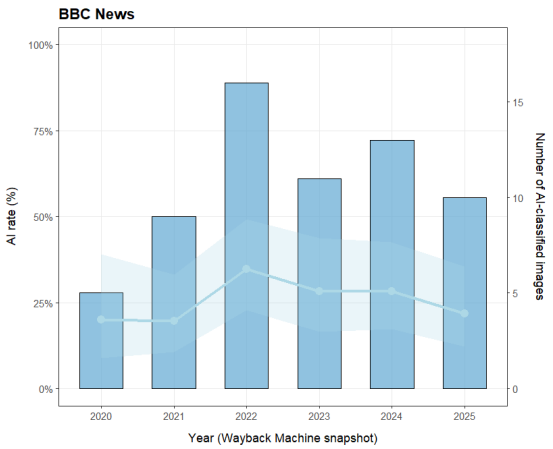
They show the temporal AI rate for each outlet individually. Bars represent the absolute number of AI-classified images per year (right axis) and the line shows the AI rate (left axis). The shaded band around the line is the 95% Wilson score confidence interval.

The Wilson score interval [51] was used instead of the standard normal approximation because it remains statistically valid for small samples and for rates at or near 0% and 100%, both of which occur in this dataset. For a sample of n images with k AI-classified, the 95% Wilson score interval is given by:

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

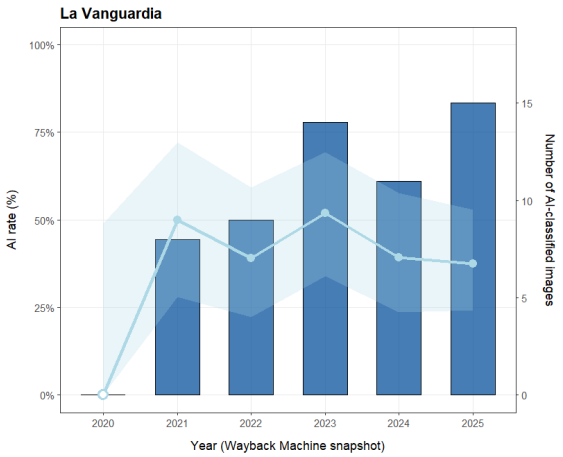
where $\hat{p} = k/n$is the observed AI rate and $z = 1.96$for a 95% confidence level. Unlike the standard normal approximation, this formulation does not assume a large sample size and remains bound within [0, 1] under all conditions. The wider the band, the less reliable the estimate: years with fewer than five images (shown with hollow markers) produce intervals wide enough to span most of the 0-100% range and should not be interpreted as meaningful estimates.

Fig. C.1. Temporal AI rate for BBC News (2020–2025).



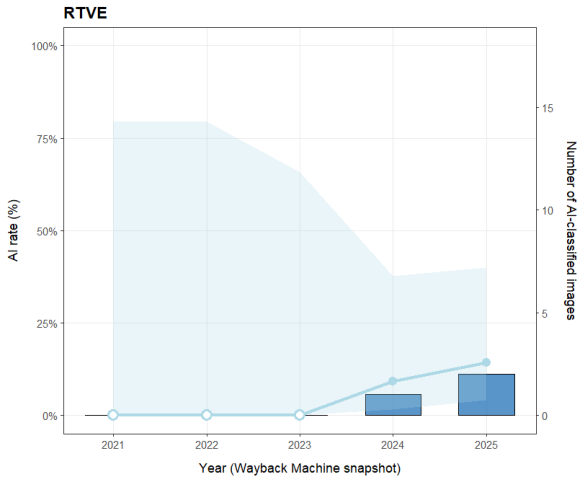
The confidence intervals remain narrow throughout, reflecting consistent annual coverage (n = 25–46 per year). The rate rises from approximately 20% in 2020–2021 to a peak of 35% in 2022, after which it gradually declines to 22% in 2025.

Fig. C.2. Temporal AI rate for La Vanguardia (2020–2025).



The 2020 estimate (0%, n = 4) has a confidence interval spanning almost the full range and should be treated as unreliable. From 2021 onwards, the rate stabilizes at a consistently high level (38–52%) with no clear upward or downward trend.

Fig. C.3. Temporal AI rate for RTVE (2020–2025).



Years 2021–2023 have fewer than five images per year, resulting in very wide confidence intervals. Only the 2024 (9%, n = 11) and 2025 (14%, n = 14) estimates are sufficiently precise to be interpreted.