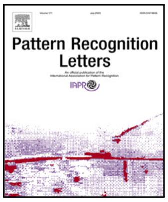
Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Uncovering the authorship: Linking media content to social user profiles

Daniele Baracchi [a], Dasara Shullani [a,*], Massimo Iuliani [a,b], Damiano Giani [a], Alessandro Piva [a,b]

[a] Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Firenze, Via di S. Marta 3, 50134, Firenze, Italy
[b] Forlab, Multimedia Forensic Laboratory, Piazza Ciardi 25, 59100, Prato, Italy

ARTICLE INFO

Editor: Maria De Marsico

Keywords:
Multimedia forensics
Social media
Dataset
Fake news
User profiling
Disinformation detection

ABSTRACT

The extensive spread of fake news on social networks is carried out by a diverse range of users, encompassing private individuals, newspapers, and organizations. With widely accessible image and video editing tools, malicious users can easily create manipulated media. They can then distribute this content through multiple fake profiles, aiming to maximize its social impact. To tackle this problem effectively, it is crucial to possess the ability to analyze shared media to identify the originators of fake news. To this end, multimedia forensics research has advanced tools that examine traces in media, revealing valuable insights into its origins. While combining these tools has proven to be highly efficient in creating profiles of image and video creators, it is important to note that most of these tools are not specifically designed to function effectively in the complex environment of content exchange on social networks. In this paper, we introduce the problem of establishing associations between images and their source profiles as a means to tackle the spread of disinformation on social platforms. To this end, we assembled *SocialNews*, an extensive image dataset comprising more than 12,000 images sourced from 21 user profiles across Facebook, Instagram, and Twitter, and we propose three increasingly realistic and challenging experimental scenarios. We present two simple yet effective techniques as benchmarks, one based on statistical analysis of Discrete Cosine Transform (DCT) coefficients and one employing a neural network model based on ResNet, and we compare their performance against the state of the art. Experimental results show that the proposed approaches exhibit superior performance in accurately classifying the originating user profiles.

1. Introduction

Misinformation has emerged as an omnipresent and all-encompassing issue in today's digital era. The proliferation of fake news across social media platforms, online news websites, and various online communication channels has become an escalating global concern, resulting in significant consequences and harm in numerous domains, such as public health, political polarization, and social unrest [1]. Indeed, there is a long-standing interest by malicious users and organizations in manipulating visual contents and using them for diffusing unreliable information and fake news, especially images and videos depicting faces [2–4], the so-called deepfakes [5,6]. Consequently, the ability to identify fake news is becoming increasingly vital in order to mitigate the dissemination of misinformation and effectively address the individuals responsible for its creation.

To address this issue, researchers in multimedia forensics have been actively developing various tools aimed at understanding the lifecycle of media [7,8]. These tools enable the characterization of multiple aspects, such as the original brand, model, and device used to capture the media, the subsequent processing chain involved, and the detection of specific manipulations or multiple compressions. By employing these methods, it becomes possible to identify the particular techniques employed by malicious users in the creation and distribution of their deceptive content [9–11].

Despite their effectiveness, it is important to note that these techniques do not directly tackle the challenge of identifying the specific profile (whether it be a user or an organization) responsible for distributing a particular content. This limitation becomes evident when the deception arises from the context in which the media is shared rather than from explicit manipulations performed on the content itself. Indeed, it is important to acknowledge that social networks have the potential to diminish the forensic evidence left by various users due to the application of compression, resizing, and filtering operations on the uploaded media [8]. Additionally, the metadata associated with the

[*] Corresponding author.
E-mail addresses: daniele.baracchi@unifi.it (D. Baracchi), dasara.shullani@unifi.it (D. Shullani), massimo.iuliani@unifi.it (M. Iuliani), damiano.giani@edu.unifi.it (D. Giani), alessandro.piva@unifi.it (A. Piva).

[1] https://blog.hootsuite.com/social-media-image-sizes-guide/

https://doi.org/10.1016/j.patrec.2024.03.008
Received 26 June 2023; Received in revised form 3 December 2023; Accepted 13 March 2024
Available online 16 March 2024

media is often partially erased, and the file container structure generated by the social network typically supersedes the original one [12]. Furthermore, it is crucial to note that each platform employs different coding schemes, thereby broadening the spectrum of traces left by each user that needs to be profiled.[1] This diversity in coding schemes is further exemplified by the fact that new social networks emerge frequently. Consequently, malicious users may begin disseminating their content on a novel and unfamiliar platform, where the media becomes tainted with new and unfamiliar traces stemming from various processing techniques, making existing forensic methods obsolete in a very short time if proper counter measures are not taken [13]. Moreover, it is important to acknowledge that media created by a particular user or company does not necessarily follow a consistent processing pattern, as the handling of the generated content can evolve over time, resulting in distinct characteristics. For instance, let us consider a vlogger who typically utilizes a specific camera and editing software to produce their videos. Over time, there might come a day when they decide to switch to a new camera, consequently altering the life cycle of the videos they create. Similarly, an information agency could experience a change in their social media manager, leading to the implementation of new processing techniques for the images they upload.

We proceed in this paper to study the issue of automatically identifying the profile that is responsible for distributing deceptive images on social platforms. We take a practical approach to this problem by considering a realistic scenario where we analyze real images posted by active social media profiles. To accomplish this, a new large scale dataset was compiled, consisting of media downloaded from diverse profiles and organizations where no control over the acquisition protocol was exerted. The dataset encompasses data obtained from multiple social networks, enabling the examination of the problem even when a monitored profile commences sharing new images on a previously unseen and unfamiliar social network. The collected data is made available to the research community for further studies. We also present two straightforward yet effective user profile identification techniques as benchmarks. The first technique follows the conventional approach of generating distinctive handcrafted features using the distribution of DCT coefficients. In contrast, the second approach harnesses the potential of data-driven methods by employing a convolutional neural network based on ResNet. Furthermore, we define three experimental scenarios with increasing levels of difficulty to serve as benchmarks, showcasing the capabilities and limitations of profiling techniques. The results show that the proposed approaches currently outperform the state of the art, emphasizing the necessity for novel and enhanced techniques to address this challenging problem effectively. The paper is structured as follows. First, we provide an overview of the current state of the art in forensics analysis of media shared on social networks, including currently available datasets (Section 2). Next, we detail the data we gathered specifically for the task at hand (Section 3). We then present the methods we developed to extract distinctive features from images on social networks (Section 4). Then, we describe the scenarios considered, the experimental setup (Section 5), and the results obtained (Section 6). Finally, we conclude by summarizing the results achieved and by discussing possible future works (Section 7).

2. Related work

The field of forensic analysis for social media content has made notable advancements in the past decade. Significant attention has been devoted to discerning the source of media, with a specific focus on identifying the social network from which it originated. Several studies have emphasized the effectiveness of analyzing metadata, coding properties, and file formats to identify the originating social network for both images and videos. Shared images, for instance, usually exhibit distinctive patterns in file names, resolutions, and coding parameters [14,15].

These approaches have demonstrated their effectiveness in determining the most recent stage of a media's life cycle. However, they have limited efficacy when it comes to characterizing any processing that occurred prior to the sharing of the media on a social network. Other approaches, based on the analysis of signal statistics, such as the distribution of discrete cosine transform (DCT), discrete wavelet transform (DWT) coefficients, and noise residuals, have been successful in accurately characterizing the specific social network from which the media originated [11–13,16–18]. The analysis of DCT statistics has also been found to possess the capability of characterizing previous processing steps. This is attributed to the fact that multiple compressions, conducted with varying parameters, generate unique distributions of coefficients [19,20].

By combining container-based and content-based features, promising results have been obtained in the identification of sharing steps that go beyond the last one [21–23]. Recent approaches have managed to reconstruct image sharing chains on social media platforms back to three steps along the sharing chain by employing a cascade of backtracking blocks [24].

The main objective of all the works mentioned above is primarily centered on the exploration of the number of times and the specific social networks on which a media content has been shared. However, to the best of our understanding, no prior research or existing literature has focused on the task of categorizing the profile responsible for sharing the content. A first approach to user profile classification, albeit not based on social networks, has been proposed by Albright et al. [25]. In that study, the researchers utilized file type, compression level, and quantization matrix of images obtained directly from news websites to determine the specific source responsible for publishing each image. They discovered that these features can effectively verify and classify the original news site with encouraging levels of performance. As a result, it can be inferred that each online news platform applies unique processing methods to its published content, which can be discerned by analyzing the media signals. However, we argue that a more realistic scenario would be to identify the user or organization responsible for sharing an image downloaded from a social platform. Most recent contents, in fact, are usually disseminated through social networks instead of the website of the user. This scenario, however, is significantly more challenging due to the aforementioned suppression of traces carried out by social networks.

The task of handling social profiles poses additional challenges primarily because of the absence of a suitable dataset that encompasses media content from diverse agencies and social networks. Indeed, the majority of existing forensic datasets do not include images sourced from social networks [26–29]. Only a few datasets, like VISION [30] and Forchheim [31] datasets, encompass images exchanged on specific platforms. Nevertheless, even in these cases, the data was uploaded by a limited number of users, and there is no indication of the profiles utilized. It is noteworthy that even the dataset proposed in [25], despite being intended for profile classification purposes, solely consists of media directly downloaded from the websites of users and organizations.

3. SocialNews data collection

We collected a dataset comprising images posted by 21 user profiles on 3 prominent social networks: Facebook, Instagram, and Twitter. The choice of user profile was driven by both scientific and practical considerations. In order to obtain a coherent dataset, it was necessary to select users who had profiles on each of the considered social networks. Profiles were chosen for users considered reputable (such as respected news agencies) as well as for users considered unreliable (such as organizations known for spreading propaganda and fake news). Additionally, profiles of public figures famous enough to be considered sources of information or misinformation were taken into account. We sought to partially mitigate the bias arising from all authors being residents in the same country by balancing the geographical origin, resulting in a set of profiles from 13 different countries.

Table 1
SocialNews main features include details on profile name, country of origin, profile type, social network platform (SNP), image resolution, Quality Factor (QF) mode and the total number of images associated with each profile.

Profile	SNP	Image resolution		QF mode	# Images
		Minimum	Maximum		
Ajmubasher	FB	1080 × 1080	1920 × 1920	92	207
Qatar	IN	640 × 640	1349 × 1685	90	259
News Agency	TW	1024 × 1024	1920 × 1920	85	100
ANSA	FB	456 × 288	1982 × 1984	92	196
Italy	IN	320 × 320	1080 × 1350	90	269
News Agency	TW	456 × 288	2048 × 2048	85	239
BBC	FB	256 × 256	2048 × 1646	71	200
UK	IN	1080 × 1080	1440 × 1798	90	264
News Agency	TW	435 × 556	1610 × 2048	85	114
ByoBLU	FB	828 × 474	1920 × 1080	93	178
Italy	IN	720 × 405	1440 × 1440	90	260
News Agency	TW	534 × 300	2048 × 1444	85	223
CNA	FB	1024 × 661	2048 × 1552	90	208
Singapore	IN	720 × 540	1440 × 1352	90	268
News Agency	TW	1200 × 676	1600 × 900	71	216
CNN	FB	460 × 259	2000 × 1472	92	191
USA	IN	720 × 720	1296 × 1595	84	263
News Agency	TW	800 × 450	1920 × 1080	85	51
Dawat-e-Islami	FB	1280 × 623	3019 × 1389	92	220
Pakistan	IN	612 × 612	1080 × 1350	90	265
News Agency	TW	501 × 540	1837 × 2048	92	217
fanpage.it	FB	526 × 526	2048 × 2048	92	190
Italy	IN	1080 × 1080	1440 × 1773	90	270
News Agency	TW	320 × 568	1080 × 1350	85	128
Fox News	FB	400 × 400	2048 × 1365	92	126
USA	IN	720 × 405	1080 × 1350	92	271
News Agency	TW	553 × 556	1080 × 1080	85	229
Joe Biden	FB	680 × 453	2048 × 2048	92	188
USA	IN	639 × 426	1440 × 1804	90	263
Personal Profile	TW	1024 × 577	1638 × 2048	85	47
Joe Rogan	FB	320 × 400	1638 × 2048	74	221
USA	IN	320 × 400	1440 × 1800	74	269
Personal Profile	TW	473 × 1024	2048 × 2048	85	101
MSNBC	FB	1080 × 360	2048 × 1823	92	99
USA	IN	1080 × 607	1440 × 1800	90	259
News Agency	TW	1024 × 512	2048 × 1823	85	126
NY Times	FB	960 × 1200	1638 × 2048	100	193
USA	IN	576 × 720	1440 × 1800	74	260
News Agency	TW	410 × 512	1440 × 1800	100	170
O Globo	FB	481 × 481	1152 × 2048	90	214
Brazil	IN	612 × 612	1080 × 1080	90	269
News Agency	TW	640 × 330	2025 × 2048	85	165
RTnews	FB	540 × 304	1820 × 2048	92	201
Russia	IN	1080 × 1080	1080 × 1080	90	256
News Agency	TW	460 × 258	1920 × 1080	90	180
SCMP	FB	820 × 284	2048 × 2048	92	205
China	IN	720 × 719	1440 × 1440	97	269
News Agency	TW	1080 × 1080	1253 × 2048	85	15
Tehran Times	FB	384 × 384	1440 × 1800	74	221
Iran	IN	359 × 201	1440 × 1801	74	265
Magazine	TW	612 × 612	1024 × 576	75	23
The Australian	FB	500 × 500	2044 × 2048	71	142
Australia	IN	640 × 640	1440 × 1800	97	276
News Agency	TW	306 × 201	2042 × 2048	85	208
The Guardian	FB	512 × 268	2048 × 2048	84	162
UK	IN	1000 × 1000	1440 × 1800	90	262
News Agency	TW	820 × 1020	2048 × 2048	85	144
The Namibian	FB	502 × 268	3329 × 1259	94	200
Namibia	IN	320 × 166	1440 × 1800	80	247
News Agency	TW	214 × 216	2048 × 2048	85	242
WION	FB	552 × 296	2048 × 2048	87	199
India	IN	612 × 407	1080 × 1080	90	259
News Agency	TW	918 × 506	1920 × 1080	98	75
12517 images: **3961** from Facebook – **5543** from Instagram – **3013** from Twitter					

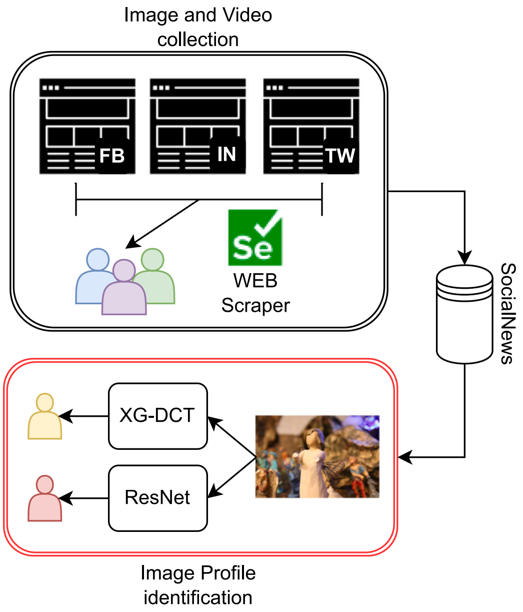


Fig. 1. *SocialNews* collection and evaluation pipeline. Media contents are downloaded from Facebook (FB), Instagram (IN) and Twitter (TW) using a web scraper to build the dataset. Then, the two benchmark methods are trained to identify the profile of origin.

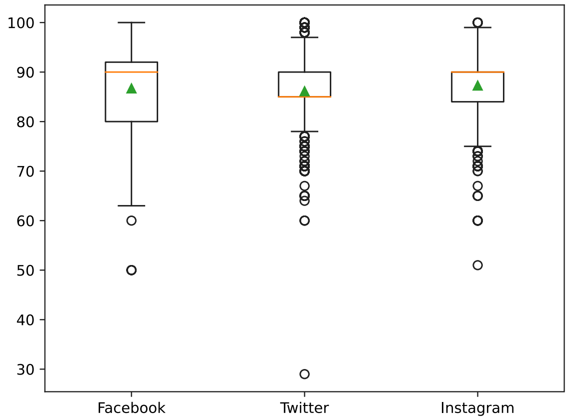


Fig. 2. Quality Factors range in each Social Network Platform. The green triangle corresponds to the average value. Best viewed in colors.

To ensure the inclusion of the same type of images typically encountered by regular web users, we obtained the contents by parsing the web pages of each profile and extracting the URLs linked to the images. To automate the process, a Python script was developed using the `selenium` library.[2] The custom software visited the selected profiles, examined the HTML code to identify the source links related to the media, and subsequently downloaded each content utilizing the `requests` library.[3] Furthermore, the software collected any available post titles to provide supplementary context for the downloaded content, in addition to the media files themselves. A pictorial representation of the process is reported in Fig. 1.

The comprehensive list of users, along with their country of origin, profile type (news agency, personal profile), and key statistics for each social network (such as the number of downloaded images), can be found in Table 1. The media contents were downloaded between the end of 2022 and the beginning of 2023 in reverse chronological order, starting with the most recent ones at that time for each chosen profile. Overall, the dataset includes 12,517 JPEG-encoded images, with an average of 600 images per profile. The minimum image resolution available is 214 × 216 pixels, while the maximum is 3329 × 1259 pixels. We present in Fig. 2 the distribution of the estimated JPEG quality factor (QF) for the gathered images. The predominant estimated

[2] https://www.selenium.dev/

[3] https://requests.readthedocs.io/en/latest/

(a) Facebook

(b) Instagram

(c) Twitter

Fig. 3. Examples of images extracted from the ANSA profile.

QF values are approximately 87 for Facebook, 86 for Instagram, and 87 for Twitter.

Since the images were directly collected from authentic accounts, a manual content analysis was conducted to identify any anomalies or biases in the acquired media. The manual verification of content revealed that each profile generally publishes distinct content on each platform. Additionally, a significant proportion of images posted by news agencies were found to feature the agency's logo, typically positioned in one of the corners. To provide an example, we showcase some media content published by ANSA on the three social networks in Fig. 3.

The accumulated data has been effectively arranged into a novel dataset named *SocialNews*. This dataset is made available for research purposes upon request[4] and encompasses not only the collected data but also videos associated with the same set of profiles, acquired using the same technique described above.

4. Methods

In the task at hand, our focus lies in identifying unique traces left by user processing prior to uploading images on social networks. To do so, our investigation delved into two methods: (i) a traditional approach leveraging the distribution of discrete cosine transform (DCT) coefficients, and (ii) a deep learning model built upon a ResNet backbone. In the subsequent paragraph, we offer a brief overview of the key aspects of these two approaches.

DCT coefficients distribution classifier. Most image editing and processing techniques, including social network sharing, generally involve a series of JPEG compression steps, possibly integrated with other image processing operations such as resizing. It is well known that sequences of multiple JPEG compressions with different settings leave distinctive traces in the distribution of the DCT coefficients [19,32–34]. For this purpose, we examine the histogram of DCT coefficients through the following process. First, the DCT is applied to 8×8 blocks throughout the entire image. Afterward, normalized histograms are formed from the dequantized DCT coefficients, focusing on the first 9 AC frequencies in zigzag sequence. Specifically, we considered only the 41 bins in the range $[-20, 20]$. These histograms are then combined to construct a 369-dimensional feature vector. Finally, the feature vectors thus created were then used to train an XGBoost [35] classifier (from this point onward marked as XG-DCT) to determine the origin profile.

ResNet-based classifier. Convolutional Neural Networks (CNNs) have demonstrated remarkable effectiveness across various tasks, showcasing their ability to extract meaningful features from images. Their success can be attributed to their hierarchical architecture, which allows them to learn and capture complex patterns, making them well-suited for tasks such as image classification, object detection, and even natural language processing. Among the various proposed convolutional architectures, ResNet [36] has consistently exhibited excellent performance across a wide range of computer vision problems. Moreover, the research community has developed a number of ResNet models pre-trained on large datasets, which can be used as powerful feature extractors. These models can be further fine-tuned to generate task-specific classifiers, enhancing their effectiveness in specific applications. Therefore, we built a profile classifier based on a ResNet18 architecture pre-trained on Imagenet. The used backbone extracts, for each image, a feature vector of 1000 elements. This vector is subsequently fed to a multi-layer perceptron which produces the final decision on the profile of origin. The whole architecture is fine-tuned to produce the final classifier.

Given that, as previously observed, various news agencies often include a unique logo in their images, a network that operates on resized images could potentially identify the source by examining the logo. However, this capability poses a significant threat to the model's usefulness, as an attacker could easily deceive the classifier by cropping the image and eliminating the distinctive logo. To address this issue, we made the decision to divide our images into non-overlapping patches measuring 256×256 pixels. By doing so, only a small portion of these patches would contain a logo, thereby preventing the neural network from relying on the logo's presence to differentiate between profiles. During the testing phase, all the patches associated with a particular image would be processed by the network. The final decision regarding the entire picture would then be determined through majority voting.

5. Experimental setting

To evaluate the performance of the proposed methods in identifying the social profile of origin for the images, experiments were conducted in three increasingly challenging scenarios whose graphical representation is shown in Fig. 4.

In the first scenario (Single-Platform), we focus on content originating from a single, known social platform. For each social network (Twitter, Instagram, Facebook), we conducted training and testing of the classifiers using two disjoint sets of images exclusively from that particular platform. In order to avoid introducing bias due to the

[4] Please refer to https://lesc.dinfo.unifi.it/ for instructions on how to gain access to the SocialNews dataset and the related source code.

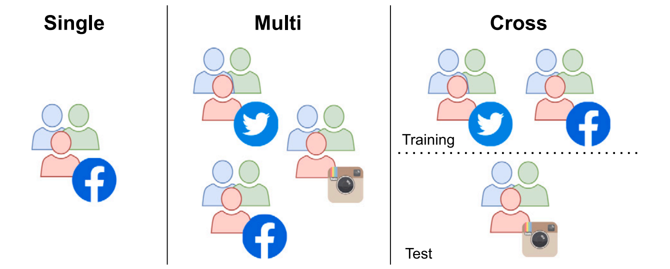


Fig. 4. Pictorial representation of the three experimental scenarios. The Single-Platform scenario focuses on images exclusively sourced from a single social network. In the Multi-Platform scenario, images can originate from any of the selected social networks. The Cross-Platform scenario involves query images generated by the same profiles encountered during training but shared on a new social network that has not been previously encountered.

specific subdivision of the data into two subsets, we have adopted a stratified 10-fold cross-validation strategy. As a result, all data from the chosen social platform were divided into 10 disjoint folds, ensuring to maintain the same proportions among the various profiles. Thus, each experiment was repeated 10 times, each time using a different fold as the training data and the remaining nine as the test data.

In the second scenario (Multi-Platform), we introduce a reduction in the available background information by considering that the images can belong to any of the available social networks. In this case, the classifiers were trained and tested on two separate sets of images from any of the available platforms. Similarly to the Single-Platform case, we have adopted a 10-fold cross-validation protocol to avoid introducing bias. In this scenario, however, data from all platforms are considered in each individual experiment. With this experimental setup, the task is anticipated to be more challenging as the traces of each user profile are mixed due to the utilization of multiple social networks. Additionally, there is no guarantee that users apply the same processing pipeline for content shared on different social media platforms, resulting in higher variability in the data.

Lastly, in the third scenario (Cross-Platform), we address the challenging setting where a user begins sharing images on a new social platform that has not been encountered before. This scenario introduces the possibility of unknown compression and coding schemes being employed. To simulate this circumstance, we implemented a leave-one-social-out strategy. This involved constructing multiple classifiers, each one of them trained on content from all the social networks except one and then tested on content belonging to the omitted platform. Since in this case the data split between training and testing was enforced by the definition of the scenario, it was not necessary to adopt a cross-validation strategy.

Furthermore, in all three scenarios, the training data was used to generate both a training set and a validation set. In the case of XG-Boost, we performed an internal 5-fold cross-validation grid search to determine the best hyperparameters (implicitly generating a validation set equal to 20% of the training data each time). During this process we searched for the optimal number of estimators (between 10 and 200), the optimal subsampling percentage (between 0.5 and 1), and the optimal learning rate η (between 0.1 and 0.5). These parameters were then used to train the final model on all the training data. In the case of ResNet, we divided the training data into a training set and a validation set with a ratio of 90/10. The validation set was used to monitor the model's performance during training and for early stopping. In all cases, we kept the test set segregated until the final performance evaluation, preventing it from influencing the training process.

For each scenario, the performance achieved by the proposed methods was compared with that obtained using the approach by Albright et al. [25]. This approach involves extracting information such as file type, compression level, and quantization matrix from the images to create a unique feature vector (*QM-feature*) for each image. Subsequently, a categorical Naive Bayes classifier is applied to the feature vector to make a decision regarding the content. Moreover, to evaluate the statistical significance of our results, we compared the accuracy obtained by the proposed approaches with the one obtained by QM by means of a two-tailed Welch's t-test [37] to ensure that their difference is statistically significant.

Table 2
Accuracies obtained on the Single-Platform scenario. We report, for each social platform, the performance for the method by Albright et al. [25] (**QM**), the proposed method based on the distribution of DCT coefficients (**XG-DCT**), and the proposed ResNet-based method (**ResNet**). The final row presents the *p*-value for the comparison between QM and the proposed methods.

Profile	Facebook			Instagram			Twitter		
	QM	XG-DCT	ResNet	QM	XG-DCT	ResNet	QM	XG-DCT	ResNet
ANSA	0.148	0.551	**0.890**	0.007	**0.532**	0.515	0.000	0.615	**0.667**
Ajmubasher	0.372	0.541	**0.757**	0.000	0.676	**0.957**	0.290	0.570	**0.900**
BBC	**0.805**	0.775	0.672	0.000	0.564	**0.777**	0.000	**0.833**	0.827
ByoBLU	0.691	0.770	**0.878**	0.292	0.646	**0.787**	0.641	0.731	**0.782**
CNA	0.591	**0.779**	0.701	0.000	0.653	**0.785**	**0.986**	0.968	0.769
CNN	0.131	0.236	**0.537**	0.498	0.494	**0.828**	0.000	0.157	**0.581**
Dawat-e-Islami	0.700	0.750	**0.772**	0.000	0.487	**0.731**	**0.889**	0.885	0.816
fanpage.it	0.237	0.474	**0.732**	**0.985**	0.641	0.841	0.039	0.242	**0.735**
FOX News	0.000	0.349	**0.683**	0.620	0.705	**0.733**	**0.969**	0.734	0.853
Joe Biden	0.229	0.447	**0.512**	0.015	0.559	**0.678**	0.000	0.043	**0.429**
Joe Rogan	**0.792**	0.710	0.729	0.123	0.520	**0.601**	**0.960**	0.941	0.691
O Globo	0.042	0.631	**0.799**	0.208	0.625	**0.736**	0.000	0.321	**0.589**
MSNBC	0.000	0.495	**0.789**	0.035	0.788	**0.863**	0.000	0.556	**0.691**
NY Times	0.674	0.679	**0.799**	**0.927**	0.608	0.613	0.800	**0.853**	0.739
RTnews	0.109	0.418	**0.845**	0.000	0.707	**0.855**	**0.844**	0.717	0.735
SCMP	0.063	0.332	**0.502**	0.372	0.665	**0.943**	0.000	0.733	**0.929**
The Australian	0.472	0.507	**0.772**	0.696	0.681	**0.802**	0.274	0.423	**0.446**
The Guardian	0.451	0.716	**0.802**	0.000	0.401	**0.719**	0.076	0.688	**0.839**
The Namibian	0.575	0.720	**0.846**	0.316	0.506	**0.907**	0.231	0.537	**0.681**
Tehran Times	0.213	0.670	**0.703**	0.362	0.566	**0.617**	0.000	0.652	**0.765**
WION	0.724	0.709	**0.849**	0.000	0.568	**0.870**	0.720	**0.773**	0.742
Average	0.398	0.593	**0.729**	0.262	0.600	**0.765**	0.454	0.652	**0.722**
p-value	–	8.17e–03	1.18e–05	–	1e–04	3.23e–07	–	0.0226	7.99e–04

6. Results and discussion

In this section, for each of the analyzed scenarios, we report and discuss the results obtained from the proposed approaches (XG-DCT and ResNet) in relation to the current state of the art.

Single-platform. We report the obtained results for the Single-Platform scenario in Table 2. The average accuracy obtained by QM (0.37) is significantly lower than the one reported by Albright et al. [25], and both XG-DCT (0.61) and ResNet (0.73) outperform it. This outcome is expected because the QM method relies on traces left by the most recent compression applied to the image. While this method might prove effective for images directly obtained from the author's website, it is not tailored to function optimally in situations where the content is shared on a social network. Indeed, in these cases, the multimedia content undergoes a second compression that tends to weaken or remove the traces used. ResNet consistently outperforms other methods across all social networks, showcasing superior performance. Additionally, its average accuracy remains stable irrespective of the originating platform. Finally, we compare the results obtained by QM with those obtained by XG-DCT and ResNet by means of a two-tailed Welch's t-test. According to the test, we have enough empirical evidence to reject the null hypothesis with a confidence level of 95%. As a consequence, the two proposed approaches significantly improve over the state of the art.

Multi-platform. In this case, the data taken into consideration include the content produced by all profiles on all social media platforms. This results in a higher variability of media content across the three platforms, as some profiles tend to post distinct types of media on different social networks. The accuracies obtained by the methods are reported in Table 3. The average accuracy achieved by the QM method is 0.30, which is again significantly lower than the ones obtained by the XG-DCT (0.59) and ResNet (0.77) approaches. Moreover, the

Table 3
Accuracies obtained on the Multi-Platform scenario. We report the performance for the method by Albright et al. [25] (**QM**), the proposed method based on the distribution of DCT coefficients (**XG-DCT**), and the proposed ResNet-based method (**ResNet**). The final row presents the *p*-value for the comparison between QM and the proposed methods.

Profile	QM	XG-DCT	ResNet
ANSA	0.055	0.551	**0.736**
Ajmubasher	0.080	0.583	**0.870**
BBC	0.228	0.592	**0.754**
ByoBLU	0.467	0.667	**0.825**
CNA	0.298	0.766	**0.798**
CNN	0.287	0.341	**0.738**
Dawat-e-Islami	0.536	0.665	**0.761**
fanpage.it	0.014	0.512	**0.749**
FOX News	0.430	0.639	**0.885**
Joe Biden	0.092	0.494	**0.640**
Joe Rogan	**0.814**	0.763	0.744
O Globo	0.031	0.486	**0.819**
MSNBC	0.006	0.669	**0.824**
NY Times	0.504	0.639	**0.746**
RTnews	0.765	0.575	**0.920**
SCMP	0.233	0.526	**0.669**
The Australian	0.396	0.540	**0.655**
The Guardian	0.016	0.590	**0.811**
The Namibian	0.357	0.588	**0.844**
Tehran Times	0.281	0.601	**0.721**
WION	0.362	0.625	**0.834**
Average	0.306	0.595	**0.774**
p-value	–	1.61e–05	4.92e–09

Table 4
Accuracies obtained on the Cross-Platform scenario. We report, for each left-out social platform, the performance for the method by Albright et al. [25] (**QM**), the proposed method based on the distribution of DCT coefficients (**XG-DCT**), and the proposed ResNet-based method (**ResNet**). The final row presents the *p*-value for the comparison between QM and the proposed methods.

Profile	Facebook			Instagram			Twitter		
	QM	XG-DCT	ResNet	QM	XG-DCT	ResNet	QM	XG-DCT	ResNet
ANSA	0.015	0.168	**0.361**	0.000	0.052	**0.366**	0.138	0.293	**0.690**
Ajmubasher	0.179	0.024	**0.902**	0.000	0.031	**0.516**	0.060	0.150	**0.803**
BBC	0.000	0.000	0.000	0.000	0.000	**0.014**	**0.605**	0.000	0.013
ByoBLU	0.652	0.489	**0.613**	0.004	0.035	**0.539**	0.534	0.556	**0.680**
CNA	0.000	0.154	**0.263**	0.000	0.295	**0.677**	0.000	0.000	**0.126**
CNN	0.005	0.037	**0.479**	0.000	0.027	**0.752**	0.000	0.078	**0.200**
Dawat-e-Islami	**0.536**	0.423	0.535	0.000	0.004	**0.512**	0.014	0.300	**0.769**
fanpage.it	0.016	0.079	**0.669**	0.004	0.111	**0.583**	0.039	0.109	**0.478**
FOX News	0.127	**0.540**	0.484	0.026	0.303	**0.780**	0.004	0.197	**0.949**
Joe Biden	0.037	0.356	**0.413**	0.057	0.285	**0.529**	0.000	**0.255**	0.207
Joe Rogan	0.100	**0.701**	0.561	**0.844**	0.755	0.459	0.000	0.000	**0.312**
O Globo	0.005	0.234	**0.400**	**0.784**	0.632	0.622	0.012	0.182	**0.397**
MSNBC	0.000	0.343	**0.704**	0.000	0.498	**0.864**	0.000	0.167	**0.632**
NY Times	**0.674**	0.575	0.348	0.000	0.004	**0.457**	**0.812**	0.465	0.712
RTnews	0.303	0.159	**0.694**	0.000	0.492	**0.925**	0.006	0.300	**0.528**
SCMP	0.073	0.210	**0.807**	0.000	0.212	**0.378**	0.000	0.000	0.000
The Australian	0.472	**0.535**	0.467	0.014	0.486	**0.607**	0.019	0.038	**0.212**
The Guardian	0.000	0.537	**0.642**	0.164	0.183	**0.646**	0.000	0.167	**0.456**
The Namibian	0.300	0.565	**0.702**	0.300	0.449	**0.766**	0.194	0.331	**0.381**
Tehran Times	0.195	0.371	**0.670**	0.257	0.355	**0.473**	0.000	0.000	**0.025**
WION	0.653	0.548	**0.797**	0.000	0.131	**0.866**	0.493	**0.533**	0.520
Average	0.210	0.328	**0.503**	0.117	0.255	**0.577**	0.154	0.227	**0.486**
p-value	–	0.078	2e–05	–	0.0689	6.26e–08	–	0.395	8e-04

results from XG-DCT and ResNet are very similar to the average mean accuracies obtained in the first scenario across the three social networks (Facebook, Instagram, and Twitter). This suggests that the variability in media content across platforms does not significantly impact the capabilities of these features. On the contrary, presenting the same profile across various social networks could prove beneficial. This is because classifiers have greater opportunities to identify the distinctive traits displayed by a user rather than those that might occur by chance on a single platform. Furthermore, we can reach the same conclusions as the Single-Platform scenario regarding the significance of the results, as the *p*-value obtained from the Welch's t-test is below the threshold of 0.05.

Cross-platform. In this scenario, we simulate the introduction of a new social network by training our methods on two out of the three available platforms and testing them on the remaining one (as exemplified in Fig. 4). We report the obtained results in Table 4. As anticipated, the accuracy performance decreases for all three methods. However, the decline in performance is more pronounced for QM (0.16) and XG-DCT (0.27) approaches, while ResNet maintains a considerable level of discriminative power (0.52). This could be attributed to the fact that the traces identified by the CNN method are less reliant on the most recent processing step and, as a result, exhibit greater resilience to variations introduced by different social platforms. Consequently, the ResNet-based approach demonstrates more robust performance in scenarios involving unknown social networks. Finally, the Welch's t-test confirms that XG-DCT is not significantly better than the baseline, whereas ResNet does significantly improve over QM with a confidence level of 95%.

7. Conclusions

In this paper we focused on the task of profile identification for social network images. We collected a substantial dataset consisting of images from 21 real profiles associated with relevant agencies and newspapers across three distinct social networks. Two benchmark approaches were introduced to address the profile classification task: one utilizing classical handcrafted features in the frequency domain, and the other employing convolutional neural networks. Three scenarios of increasing difficulty were considered, and a comparison was made with a state-of-the-art method in all tests conducted. The results demonstrated that the data-driven approach yielded more effective and consistent results compared to other methods. Additionally, this approach maintained the ability to distinguish between profiles even in more challenging scenarios where unknown social networks were involved.

While the results we have obtained are promising, we want to emphasize some limitations of the proposed approaches that could be addressed in future research. First and foremost, the classifiers developed are based on an assumption that the content to be examined must necessarily belong to one of the considered profiles. However, the real-world application of a method for profile attribution must necessarily account for the possibility that the content may have been generated by a previously unseen user. Additionally, social profiles associated with large organizations could, in practice, correspond to multiple users with distinct pipelines, as the same account might be used by several individuals or editorial teams to disseminate their content. This possibility is not currently explicitly handled but would certainly be beneficial for accurately attributing content to its authors. Finally, as a suggestion for future work, it is recommended to merge the features extracted by the two approaches in order to create a more comprehensive and powerful descriptor. By leveraging the complementary nature of the extracted information, a merged descriptor has the potential to enhance the accuracy and robustness of the profile identification task.

CRediT authorship contribution statement

Daniele Baracchi: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Dasara Shullani:** Conceptualization, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Massimo Iuliani:** Conceptualization, Writing – original draft, Writing – review & editing. **Damiano Giani:** Data curation, Software. **Alessandro Piva:** Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alessandro Piva reports financial support was provided by Defense Advanced Research Projects Agency. Alessandro Piva reports financial support was provided by Government of Italy Ministry of Education University and Research.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the Italian Ministry of Universities and Research (MUR) under Grant 2017Z595XS, and in part by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112090136. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government. The authors would also like to thank Chiara Camerota for her advice on the statistical tests employed in this paper.

References

[1] J. Hendrix, D. Morozoff, Media forensics in the age of disinformation, in: H.T. Sencar, L. Verdoliva, N. Memon (Eds.), Multimedia Forensics, Springer Singapore, 2022, pp. 7–40.
[2] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Niessner, Face2Face: Real-time face capture and reenactment of RGB videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[3] M. Ngo, S. Karaoglu, T. Gevers, Self-supervised face image manipulation by conditioning GAN on face decomposition, IEEE Trans. Multimed. (2021) 1.
[4] Q. Xu, H. Wang, L. Meng, Z. Mi, J. Yuan, H. Yan, Exposing fake images generated by text-to-image diffusion models, Pattern Recognit. Lett. 176 (2023) 76–82.
[5] S. Parkin, The rise of the deepfake and the threat to democracy, The Guardian (2019) URL https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.
[6] Y. Liang, M. Wang, Y. Jin, S. Pan, Y. Liu, Hierarchical supervisions with two-stream network for Deepfake detection, Pattern Recognit. Lett. 172 (2023) 121–127.
[7] L. Verdoliva, Media forensics and DeepFakes: An overview, IEEE J. Sel. Top. Sign. Proces. 14 (5) (2020) 910–932.
[8] C. Pasquini, I. Amerini, G. Boato, Media forensics on social media platforms: a survey, EURASIP J. Inf. Secur. (1) (2021) 1–19.
[9] A. Piva, An overview on image forensics, Int. Scholarly Res. Notices 2013 (2013).
[10] P. Bestagini, M. Fontani, S. Milani, M. Barni, A. Piva, M. Tagliasacchi, S. Tubaro, An overview on video forensics, in: European Signal Processing Conference, 2012, pp. 1229–1233.
[11] K. Rana, G. Singh, P. Goyal, SNRCN2: Steganalysis noise residuals based CNN for source social network identification of digital images, Pattern Recognit. Lett. 171 (2023) 124–130.
[12] D. Shullani, D. Baracchi, M. Iuliani, A. Piva, Social network identification of laundered videos based on DCT coefficient analysis, IEEE Signal Process. Lett. 29 (2022) 1112–1116.
[13] S. Magistri, D. Baracchi, D. Shullani, A.D. Bagdanov, A. Piva, Towards continual social network identification, in: 11th International Workshop on Biometrics and Forensics, IEEE, 2023, pp. 1–6.
[14] P. Mullan, C. Riess, F. Freiling, Forensic source identification using JPEG image headers: The case of smartphones, Digit. Investig. 28 (2019) S68–S76.
[15] O. Giudice, A. Paratore, M. Moltisanti, S. Battiato, A classification engine for image ballistics of social data, in: S. Battiato, G. Gallo, R. Schettini, F. Stanco (Eds.), Image Analysis and Processing, Springer, 2017, pp. 625–636.
[16] R. Caldelli, R. Becarelli, I. Amerini, Image origin classification based on social network provenance, IEEE Trans. Inf. Forensics Secur. 12 (2017) 1299–1308.
[17] I. Amerini, T. Uricchio, R. Caldelli, Tracing images back to their social network of origin: A CNN-based approach, in: IEEE Workshop on Information Forensics and Security, 2017, pp. 1–6.
[18] Manisha, A. Karunakar, C.-T. Li, Identification of source social network of digital images using deep neural network, Pattern Recognit. Lett. 150 (2021) 17–25.
[19] T. Bianchi, A. Piva, Detection of nonaligned double JPEG compression based on integer periodicity maps, IEEE Trans. Inf. Forensics Secur. 7 (2) (2011) 842–848.
[20] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, S. Tubaro, Aligned and non-aligned double JPEG detection using convolutional neural networks, J. Vis. Commun. Image Represent. 49 (2017) 153–163.
[21] Q. Phan, C. Pasquini, G. Boato, F.G.B. De Natale, Identifying image provenance: An analysis of mobile instant messaging apps, in: IEEE International Workshop on Multimedia Signal Processing, 2018, pp. 1–6.
[22] Q. Phan, G. Boato, R. Caldelli, I. Amerini, Tracking multiple image sharing on social networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 8266–8270.
[23] N. Siddiqui, A. Anjum, M. Saleem, S. Islam, Social media origin based image tracing using deep CNN, in: Fifth International Conference on Image Information Processing, 2019, pp. 97–101.
[24] S. Verde, C. Pasquini, F. Lago, A. Goller, F. De Natale, A. Piva, G. Boato, Multi-clue reconstruction of sharing chains for social media images, IEEE Trans. Multimed. (2023) 1–15.
[25] M. Albright, N. Menon, K. Roschke, A. Basharat, Source attribution of online news images by compression analysis, in: IEEE International Workshop on Information Forensics and Security, IEEE, 2021, pp. 1–6.
[26] C. Galdi, F. Hartung, J.-L. Dugelay, SOCRatES: A database of realistic data for SOurce camera REcognition on smartphones, in: Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, SCITEPRESS, 2019.
[27] H. Tian, Y. Xiao, G. Cao, Y. Zhang, Z. Xu, Y. Zhao, Daxing smartphone identification dataset, IEEE Access 7 (2019) 101046–101053.
[28] S. Taspinar, M. Mohanty, N. Memon, Source camera attribution of multi-format devices, 2020, arXiv:1904.01533.
[29] B.C. Hosler, X. Zhao, O. Mayer, C. Chen, J.A. Shackleford, M.C. Stamm, The video authentication and camera identification database: A new database for video forensics, IEEE Access 7 (2019) 76937–76948.
[30] D. Shullani, M. Fontani, M. Iuliani, O.A. Shaya, A. Piva, VISION: a video and image dataset for source identification, EURASIP J. Inf. Secur. 2017 (1) (2017) 15.
[31] B. Hadwiger, C. Riess, The Forchheim image database for camera identification in the wild, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI, Springer, 2021, pp. 500–515.
[32] T. Bianchi, A. Piva, Image forgery localization via block-grained analysis of JPEG artifacts, IEEE Trans. Inf. Forensics Secur. 7 (3) (2012) 1003–1017.
[33] X. Liu, W. Lu, Q. Zhang, J. Huang, Y.-Q. Shi, Downscaling factor estimation on pre-JPEG compressed images, IEEE Trans. Circuits Syst. Video Technol. 30 (3) (2019) 618–631.
[34] W. Lu, Q. Zhang, S. Luo, Y. Zhou, J. Huang, Y.-Q. Shi, Robust estimation of upscaling factor on double JPEG compressed images, IEEE Trans. Cybern. 52 (10) (2021) 10814–10826.
[35] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[37] B.L. Welch, The generalization of 'student's' problem when several different population variances are involved, Biometrika 34 (1/2) (1947) 28–35.