




# Position Paper: Think Globally, React Locally — Bringing Real-time Reference-based Website Phishing Detection on macOS

Ivan Petrukha   
MacPaw  
Kyiv, Ukraine  
petrukha@macpaw.com

Nataliia Stulova   
MacPaw  
Kyiv, Ukraine  
nata.stulova@macpaw.com

Sergii Kryvoblotskyi   
MacPaw  
Kyiv, Ukraine  
krivoblotsky@macpaw.com

## Abstract—

**Background.** The recent surge in phishing attacks keeps undermining the effectiveness of the existing anti-phishing approaches. While state-of-the-art phishing detection solutions can effectively detect phishing across the web, they mainly focus on automated web crawling and updating blacklists.

**Aim.** The time after a phishing website is published on the web and remains unprocessed by detection systems is critical. We aim to reduce this time gap via real-time phishing detection solution and perform continuous background processing without extra user interaction.

**Method.** We propose a real-time on-device solution that identifies phishing sites immediately when encountered by the user. Our reference-based approach analyzes the visual content of webpages, identifying phishing attempts through layout analysis, brand impersonation recognition, and credential input areas detection.

**Results.** Our case study shows that it's feasible to perform background processing on-device continuously. For web browser phishing detection, the process utilizes 16% of a single CPU core and less than 84MB of RAM on an Apple M1 while maintaining a high efficiency with 95.7% precision and 87.7% recall, based on a test dataset of 50K phishing and benign webpages.

**Conclusions.** Our results demonstrate the potential of on-device, real-time phishing detection systems to enhance cybersecurity defensive technologies. We maintained the accuracy of state-of-the-art reference-based phishing detection solutions while bringing this functionality directly to the device.

**Index Terms—**Phishing, Computer Vision, macOS

## 1. Introduction

According to the phishing activity trends reports<sup>1</sup> from the Anti-Phishing Working Group (APWG), the number of phishing attacks keeps increasing year over year, with 2023 being the worst so far, with 4,987,809 unique phishing webpages created. While cloud infrastructure providers like Cloudflare implement phishing detection on their end<sup>2</sup>, these measures are not sufficient to prevent phishing attacks at large. Existing phishing detection methods can be categorized into three categories:

*blacklist-based, classification-based, and reference-based.* Traditional blacklist-based solutions like Google Safe Browsing<sup>3</sup>, although they are effective, do not keep up with the speed of phishing websites spreading [19], whose creation and deployment are typically automated. Classification-based approaches [10], [17], [18] use machine learning algorithms to analyze URL, HTML, or other features to classify webpages as phishing or legitimate, including on-device approaches [2], [7], but lose efficiency against HTML obfuscation techniques and lack accuracy. In contrast, reference-based approaches [1], [13], [14] use computer vision and can effectively analyze webpage appearance to extract information about the displayed content, allowing to detect phishing attempts and form reasonable verdicts.

While the latest reference-based solutions effectively detect phishing across the web, they focus on automated web crawling and filling the blacklists. The time after a phishing webpage is published on the web and remains unprocessed by detection systems is critical. A phishing campaign starts spreading malicious URLs just when they become available, and even within a few minutes, the campaign may affect many users.

In this paper, we propose a real-time on-device unobtrusive anti-phishing solution for macOS and evaluate in on the case of in-browser phishing website detection, improving upon the work of [14]. Our approach relies on macOS-specific system resources and frameworks, which allows us to significantly reduce computational resource demand. Processing live screen capturing requires applying a phishing detection algorithm within a matter of seconds, ensuring users' safety. To achieve this, we rely solely on local machine learning models, eliminating the need for cloud-based tools. Another important aspect of using only local models is privacy concerns. By utilizing local models, user data remains on the device at all times, ensuring enhanced security and peace of mind for the user, which aligns with Apple's approach to privacy protection<sup>4</sup>.

## 2. Threat Model

We consider the threat model of PhishIntention [14], in which the adversary creates a phishing website, impersonating an official website of a known company and demanding the personal data of the website visitor. This

1. <https://apwg.org/trendsreports/>

2. <https://blog.cloudflare.com/2023-phishing-report>

3. <https://safebrowsing.google.com>

4. <https://www.apple.com/ua/privacy/approach-to-privacy/>

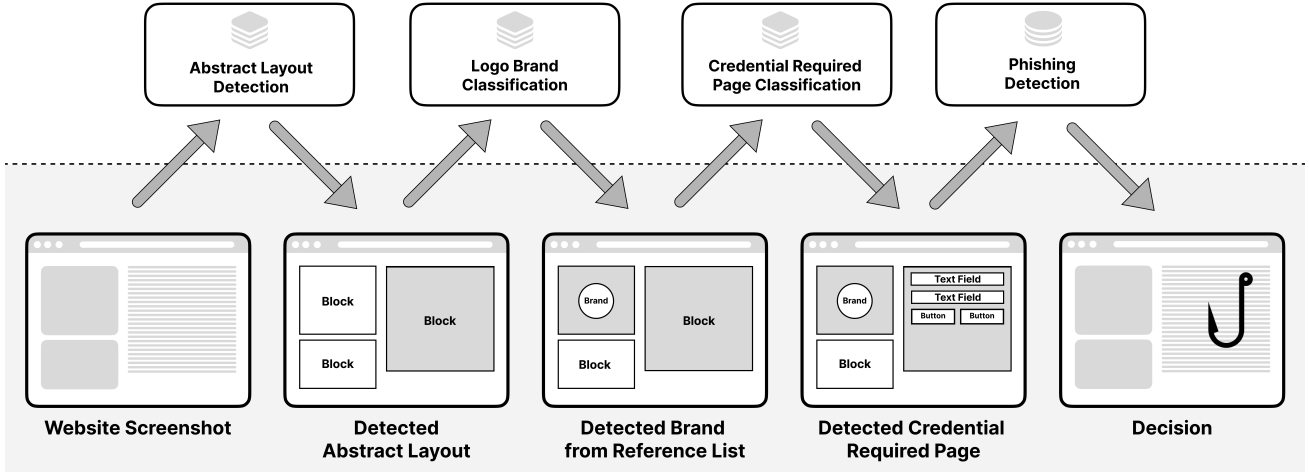


Figure 1. General workflow of our phishing detection approach

form of phishing is executed through social interaction, employing psychological manipulations to trick users into disclosing sensitive security data. Initially, the attacker conducts a thorough analysis to identify the potential vulnerabilities within the target demographic essential for the successful execution of the attack. Subsequently, the phisher endeavors to establish trust with the target. In the final phase, the attacker manipulates the scenario to induce the target to share critical information.

### 3. Solution

Figure 1 presents a high-level overview of our phishing detection approach. Similarly to PhishIntention, we focus on identifying both the intention to impersonate a brand and the intention to capture user personal information. The main algorithm consists of three essential steps. First, we begin with the webpage screenshot analysis (Section 3.1) to identify layout elements and gain rich knowledge about webpage content. Second, we classify the detected logo brand (Section 3.2) that the phishing webpage tries to impersonate. If we find a brand from our reference list, we continue to the next step. Third, we classify the webpage into two categories: if it requires credentials or not (Section 3.3). This step confirms that an impersonated phishing website is trying to steal personal user information. Matching these conditions allows us to decide whether a webpage is phishing or legitimate.

#### 3.1. Abstract Layout Detection

The main purpose of abstract layout detection is to determine whether the currently visible webpage contains a brand logo that may be targeted by hackers and find credential input forms. To achieve this, we analyze screenshots, identifying different layout elements such as logos, buttons, inputs, labels, and blocks. We used the part of the dataset from the section *CRP classifier and AWL detector* of [15]. The dataset contains about 9K webpage screenshots, annotated with regions and types of layout elements, together with the information about

the webpage layout type (credential required or not) and the corresponding brand.

**Layout Elements Extraction.** The DETR (Detection Transformer) [5] architecture revolutionizes object detection by employing transformers, to directly learn the relationships between objects in an image without predefined anchor boxes. Considering its state-of-the-art performance in object detection tasks, we have decided to adopt the DETR model with ResNet-50 [8] backbone. We trained the model using 8,109 labeled webpage screenshots. During the training stage, we conducted a comparative analysis of the DETR model’s performance across varying image sizes. Table 1 shows the mAP (Mean Average Precision) accuracy metric compared with processing speed. The model which works with small images shows good performance, albeit with lower accuracy.

TABLE 1. DETR PERFORMANCE COMPARISON

Image Size	Logo mAP [.5:.95]	Samples per Second
224x386	37.3	10.5
432x768	46.6	3.9

To safeguard user security, it’s imperative that our system processes webpage layout analysis at a minimum rate of once per second. Concurrently, it’s essential to manage the workload on the CPU effectively to prevent overutilization, ensuring the system remains responsive and reliable over extended periods of operation. Achieving an average processing rate of 3.9 samples per second provides the capability to analyze a frame in approximately 1/4 seconds. This rapid processing affords us a substantial interval of 3/4 seconds where the CPU can enter a reduced activity state, thereby mitigating the risk of overheating and sustaining performance over time.

**Reducing Overlapping Boxes.** The NMS (Non-Maximum Suppression) algorithm is used in object detection to eliminate redundant or overlapping bounding boxes. By design, the DETR model doesn’t have a built-in NMS layer. However, given that an abstract webpage

layout should never contain overlapping elements, we integrated the protobuf model from coremltools<sup>5</sup> and merged it with the main model. DETR consistently generates a fixed set of 100 bounding boxes for any input image, which is the default value but can be changed. These bounding boxes are a representation of potential object locations within the image (Figure 2).

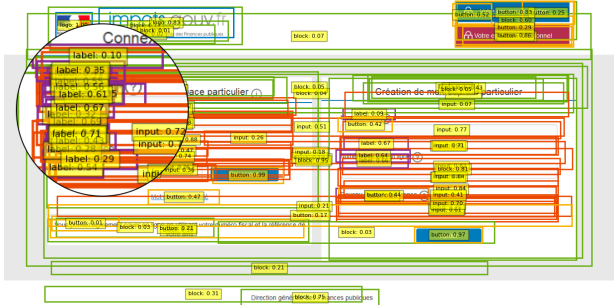


Figure 2. Potential object locations

NMS systematically identifies and keeps only the non-overlapping bounding boxes with the highest confidence scores and makes our outputs flexible, meaning now we can receive from 0 to 100 boxes instead of a fixed count. The remaining bounding boxes represent the most likely object locations within the image (Figure 3).

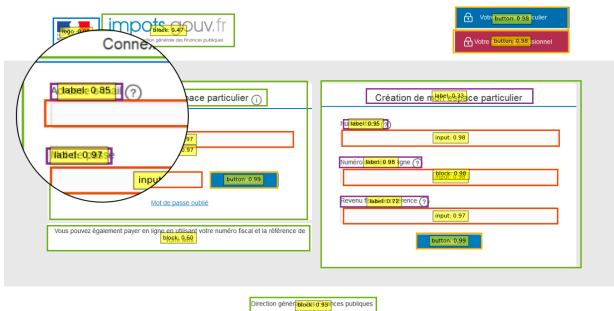


Figure 3. Refined object locations after post-processing

### 3.2. Logo Brand Classification

To train logo brand classification model we used brand logos available in the *OCR-aided Siamese model* dataset part of [15]. This part contains 3061 logo images of 277 brands. We utilized the ResNet-18 model, which was trained on precisely cropped logo layout elements during the training stage. During the evaluation stage, we utilize the layout elements identified by the *LayoutObjectDetector* (DETR) and select the logo element with the highest confidence score. Subsequently, the *LogoBrandClassifier* (ResNet-18) is applied within the bounds of the detected logo’s bounding box. Accuracy metrics details are available in the Section 4.1.2.

### 3.3. Credential Required Page Classification

Credential required page classification is an important step to reduce the number of false positive detections.

5. <https://apple.github.io/coremltools/docs-guides>

To streamline our process and optimize the algorithm we aimed to utilize already extracted features from the Abstract Layout Detection step. The DETR model receives the query image as an input and generates  $N$  output vectors. Because output vectors are used to predict the class and location of the object, they consist of the information about the object and are suitable for the object-level image representation [4]. The object detection model was divided into two models, the first is *LayoutFeatureExtractor*, which generates object-level features of shape  $(B, N, H)$  where  $B$  is batch size,  $N$  is a number of output vectors which is 100 in our case and  $H$  is hidden states for each output which is 256 in our case. The subsequent *LayoutObjectDetector* model takes these raw features as input and transforms it into a tensor of shape  $(B, N, 5)$  where the first 4 elements are coordinates of the bounding boxes and the last element is a layout element class. Then, we employ an additional *LayoutClassifier* model, which repurposes the previously extracted object-level features to transform the task of object detection into image classification. This transition is facilitated by a custom classification head (Figure 4), comprising a streamlined multi-layer perceptron architecture. The classification head processes and flattens the object-level features, applies ReLU activation functions for non-linear processing, and concludes with a Dropout layer to mitigate overfitting. Resulting in the output of shape  $(B, 2)$ , this classification head efficiently provides classification scores, distinctly categorizing pages as credential required or not. Accuracy metrics details about this approach are available in the Section 4.1.3.

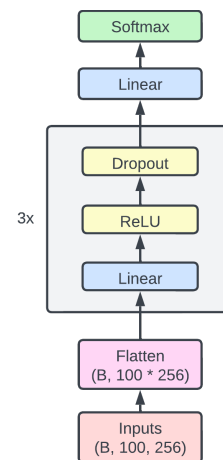


Figure 4. Classification head for DETR object-level features

### 3.4. Models Hierarchy

The resulting hierarchy of the final models is illustrated in Figure 5, showing the structured organization and interrelations among the different models.  $A$ ,  $B$  and  $C$  sections represents different training stages.  $A$  is an initial stage where we are training *LayoutFeatureExtractor* simultaneously with *LayoutObjectDetector* on object detection task.  $B$  and  $C$  can be trained in parallel, but only when the previous stage was fully trained. When we make changes in the  $A$  stage, then stage  $C$ , which includes *LayoutClassifier*, should be retrained because it fully depends on *LayoutFeatureExtractor* outputs.

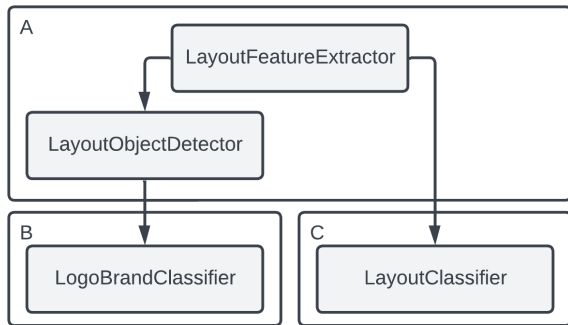


Figure 5. Final Models Hierarchy

All models were converted to the Core ML format to optimize performance through the use of Metal Performance Shaders. Core ML<sup>6</sup> is a built-in high-level macOS native framework to integrate machine learning models into Apple platform applications. It is optimized to perform on-device operations by leveraging the CPU, GPU, and Neural Engine chip while minimizing its memory footprint and power consumption. Metal<sup>7</sup> framework allows applications to directly interact with a device’s GPU. Machine learning applications leverage Metal’s computational acceleration for both training and inference tasks.

### 3.5. macOS Application

We built a native macOS application using Swift programming language and integrated all converted Core ML packages into it. We utilized Screen Capture Kit to access fullscreen frames. Screen Capture Kit<sup>8</sup> is a framework for screen capturing on macOS. This framework has been designed with a focus on performance by GPU utilization. It allows capturing content from various sources such as displays, applications, and windows, along with associated audio. A native macOS application continuously runs in the background, waiting for the active web browser to appear on the screen. While working with different browsers, we need to specify the region of actual web content and ignore the desktop background, address input, or tabs. The Accessibility Framework<sup>9</sup> incorporates an extensive array of tools and functionalities aimed at enhancing the accessibility of Apple devices. For developers, this framework offers the opportunity to delve into the accessibility graph that contains all UI elements and their relations, providing an additional source of information regarding the active application’s state, layout, and controls. In our case, we can inspect the accessibility elements graph of Safari and find an element called *AXWebArea*, which contains the actual webpage. By getting its parent, which is *AXScrollArea*, we can account for the relative coordinates of the web area to the inner page scroll offset. These coordinates are used to crop the fullscreen image and resize it to 432x768 size, making the image acceptable for the machine learning models analysis. After this, the algorithm works exactly as described in Section 3.

6. <https://docs.developer.apple.com/documentation/coreml>

7. <https://docs.developer.apple.com/documentation/metal/>

8. <https://docs.developer.apple.com/documentation/screencapturekit>

9. <https://docs.developer.apple.com/documentation/accessibility/>

## 4. Evaluation

We perform a case study, exploring several use aspects to ensure that our system is optimized for both performance and efficiency:

**Abstract Layout Detection Accuracy.** How does our model’s logo-detecting efficiency compare to the previous solutions in terms of both accuracy and processing speed?

**Logo Brand Classification Accuracy.** How accurately can we classify the brand of detected logo element?

**Credential Required Page Classification Accuracy.** What is the effectiveness of our object-level features-based classification compared to previous work?

**Overall Phishing Detection Performance.** Did we manage to maintain the overall accuracy of phishing webpage detection compared to other solutions?

**System Resources Usage Efficiency.** Can we perform continuous processing in the background without a major impact on user workflow?

### 4.1. Accuracy Metrics

During the accuracy evaluation, we used the same dataset [15] provided by PhishIntention team to compare our solution side by side with competitors. We provide detailed information about each used dataset part at the beginning of corresponding subsection.

**4.1.1. Abstract Layout Detection.** We used 901 labeled webpages from *AWL detector* dataset part of [15] to measure how accurate we can find logo elements. The mAP (Mean Average Precision) measurement is under the IoU (Intersection over Union) thresholds [0.5:0.95] and shows slightly worse accuracy compared to the PhishIntention model, but at the same time, achieves the same results as Phishpedia [13] model (Table 2). We achieved worse results than competitors, but our solution is more efficient in terms of processing speed.

TABLE 2. LOGO DETECTION ACCURACY COMPARISON

Solution	Logo mAP [0.5:0.95]	Samples per Second
PhishIntention	59.5	0.7
Phishpedia	46.6	-
Ours	46.6	3.9

**4.1.2. Logo Brand Classification.** We used 2000 logo images from *OCR-aided Siamese model* dataset part of [15] to measure how accurate we can classify logo brands. Table 3 presents the accuracy metrics achieved by employing the ResNet-18 model for logo brand classification. We achieved the same level of accuracy as PhishIntention and slightly improved accuracy compared to Phishpedia.

TABLE 3. LOGO BRAND CLASSIFICATION ACCURACY

Solution	Test Accuracy
PhishIntention	89.1
Phishpedia	83.5
Ours	90.8

**4.1.3. Credential Required Page Classification.** We used 901 labeled webpages from *CRP classifier* dataset part of [15] to validate the accuracy and robustness of object-level features based classification (Table 4). The increment (+3.1%) in test accuracy is particularly noteworthy as it indicates a reduction in false positive detections, which is clearly observed in Figure 7.

TABLE 4. CRP CLASSIFICATION ACCURACY COMPARISON

Solution	Train Accuracy	Test Accuracy
PhishIntention	99.3	95.0
Ours	99.5	98.1

**4.1.4. Overall Phishing Detection.** Figure 6 shows the ROC (Receiver operating characteristic) of different phishing detection solutions in logarithm scale. We evaluated the algorithm within *Experiment dataset (25K benign and 25K phishing webpages)* part of [15]. This plot shows that our solution maintained phishing detection efficiency while bringing it directly to the device.

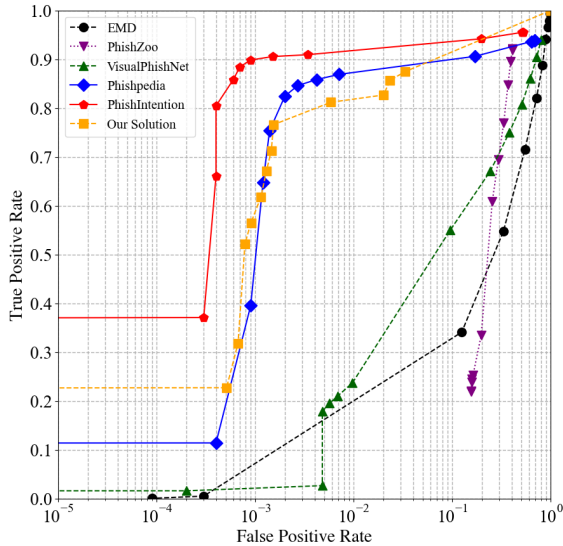


Figure 6. ROC of different phishing detection solutions

Figure 7 shows precision and recall within the same dataset, and additionally false positive rate within *Misleading legitimacy* dataset part of [15] which contains 3049 benign webpages. In this plot we can observe that increased accuracy of credential required page classification reduced the false positive rate from 5.1% to 3.4%.

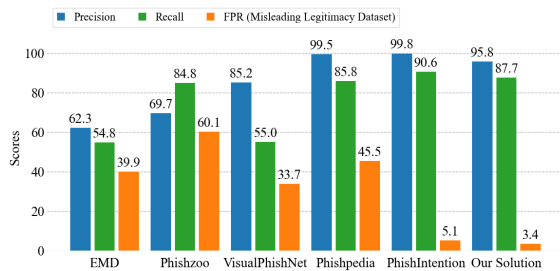


Figure 7. Accuracy metrics of different phishing detection solutions

## 4.2. System Resources Usage

We measured performance on two different devices: MacBook Pro 2020 M1 and 4K iMac 2020 Intel Core i7 3.8 GHz. Initial tests were performed without imposing a frame rate limit, enabling the application to analyze each frame as it becomes available. The results shown in Table 5 highlight the performance difference between Apple’s M1 and Intel’s Core i7 processors, with the former demonstrating better optimization in terms of both CPU usage and RAM consumption.

TABLE 5. SYSTEM RESOURCES USAGE

Device	CPU Usage	RAM Consumption
Apple M1	58%	440 MB
Intel Core i7	74%	860 MB

To meet the criteria for unobtrusive background processing, we restrict the frame rate to a single frame per second. This measure is posited to strike an optimal balance between enhancing user security and minimizing the consumption of system resources. We observe (Table 6) a significant improvement in CPU utilization attributed to increased periods of processor idleness.

TABLE 6. SYSTEM RESOURCES USAGE (FPS LOCK)

Device	CPU Usage	RAM Consumption
Apple M1	20%	440 MB
Intel Core i7	11%	820 MB

Quantization<sup>10</sup> is a process of clamping the model weights. We quantized all models from a default 32-bit floating point to a smaller but less accurate 16-bit floating point. Model quantization yields substantial enhancements in frame processing time, CPU utilization, bundle size, and, most notably, RAM consumption (Table 7).

TABLE 7. SYSTEM RESOURCES USAGE (QUANTIZED MODELS)

Device	CPU Usage	RAM Consumption
Apple M1	16%	84M
Intel Core i7	8%	450M

Subsequently, we checked how quantization affects overall phishing detection efficiency in the same way as described in Section 4.1.4. We found that this optimization technique effectively halves the model’s size while maintaining a minimal impact on its accuracy (Table 8).

TABLE 8. QUANTIZATION IMPACT ON ACCURACY

Floating-point	Precision	Recall
32-bit	95.8%	87.8%
16-bit	95.1%	87.1%

10. [https://huggingface.co/docs/optimum/concept\\_guides/quantization](https://huggingface.co/docs/optimum/concept_guides/quantization)



## 5. Related Work

**Reference-based phishing webpage detection.** State-of-the-art approaches, such as PhishIntention [14], Phishpedia [13], and VisualPhishNet [1], primarily focus on automatically identifying phishing websites on the internet using web crawlers rather than relying on local inference. Solutions, such as DynaPhish [16] and KnowPhish Detector [12], further improve accuracy and extend the applicability of reference-based approaches but still are cloud-first. Our work addresses the performance aspect of such approaches, bringing them locally to the user devices without compromising on phishing detection accuracy.

**On-device phishing detection.** Both industry practitioners and academia researchers have recently begun to explore this attack vector more actively, though, to the best of our knowledge, the current industry focus is on SMS phishing (*smishing*) prevention. Recent works from Samsung R&D Institute focus on smishing: Harichandana *et al.* [3] introduce an on-device pipeline for smishing detection, and J. W. Seo *et al.* [20] present a privacy-preserving SMS classifier. The privacy-preserving aspect of on-device solutions is also explored by researchers in recent works on SMS spam and phishing detection [6], [21].

## 6. Discussion

**Proactive Protection.** Our solution aims to automatically detect phishing attacks, eliminating the need for additional user interaction. There’s no need for users to press extra buttons or copy and paste suspicious URLs for detection. In stressful situations, users may lack the necessary attention to analyze URLs and page content thoroughly. Hence, our solution operates proactively, safeguarding users without their active intervention.

**On-device Processing.** Our final metrics of system resource usage indicate that our solution can operate continuously in the background without a noticeable impact. It is important to highlight that CPU usage represents a fraction of the total available capacity: given that the Apple M1 has 8 cores, this means we are utilizing 16% of the 800% total capacity available across all cores. In comparison to other applications, this level of resource consumption is equivalent to maintaining three active Safari browser tabs (which typically consume about 5% depending on the website) or conducting an active Zoom call (which can use about 22% during a voice call and about 41% during video call).

**Browser Extensions.** In our approach, we have access to the image of the whole screen but select a *region of interest*, which we currently limit to an active browser, performing any screen recognition tasks within this window. In the use case of browser window content tracking, this approach has a direct advantage over relying on browser extensions, as it is browser-agnostic and requires minimal adjustment for working with different browsers: adding Google Chrome support on the Safari-compatible base required up to 50 lines of code only. For the users, the browser-agnostic approach has the direct advantage of lessening the burden of browser extension management.

**Users as Sensors.** Our solution combines with the traditional method of keeping a blacklist of phishing websites. The algorithm can be extended to perform URL matching with a blacklist at the very beginning. If there is no match in the blacklist, we can perform the main part of the algorithm described in this paper. Upon identifying a phishing attempt, both the user and our centralized system are alerted. The primary advantage of this method is that it leverages thousands of users as sensors rather than relying on multiple web crawlers to search for phishing attempts on the web, aligning with the human-as-a-security-sensor (HaaS) [9] approach.

## 7. Limitations

The main limitations of our proposed solution stem from the specifics of the use of computer vision techniques for object detection on the webpage. Our approach only identifies phishing websites that display both a visible brand logo and a credential form on the same page. If either of these elements is missing or becomes hidden when a user scrolls or magnifies, our solution will fail.

Our solution doesn’t protect users from phishing attacks for unknown brands. In case of the need for a well-known brand logos database update, machine learning models should be retrained and updated on the user device. Such an approach also fails when the company rebrands. For example, from Facebook to Meta or from Twitter to X. A recent study [11] has also shown a possible attack on techniques that rely on brand logo recognition, for which we are not robust.

## 8. Conclusions and Future Work

This work argues for the potential of on-device, real-time phishing detection systems to enhance cybersecurity defensive technologies. We demonstrate that modern devices are capable of leveraging machine learning models efficiently, sustaining fast processing speed while conserving system resources. A straightforward future work direction is adapting our solution for iOS devices. This transition is expected to be relatively smooth, as the Core ML packages required for the transfer are fully compatible with the iOS platform.

In our case study, we focused on webpage phishing detection. However, given our capability to capture the entire screen, our potential for advancement is not limited here. A complex solution could orchestrate various algorithms tailored to specific *Regions of Interest*. For example, in web browsers, we can process webpages according to the introduced algorithm, in email clients, we can detect phishing emails, in messengers, we can detect smishing.

## 9. Acknowledgements

We would like to thank the anonymous reviewers for providing valuable feedback and suggestions that helped to improve this paper.

## References

- [1] Sahar Abdelnabi, Katharina Krombholz, and Mario Fritz. VisualPhishNet: Zero-Day Phishing Website Detection by Visual Similarity. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, CCS '20*, pages 1681–1698, New York, NY, USA, November 2020. Association for Computing Machinery.
- [2] Giovanni Armano, Samuel Marchal, and N. Asokan. Real-Time Client-Side Phishing Prevention Add-On. In *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, pages 777–778, June 2016. ISSN: 1063-6927.
- [3] Harichandana B S S, Sumit Kumar, Manjunath Bhimappa Ujji-nakoppa, and Barath Raj Kandur Raja. COPS: A Compact On-Device Pipeline for Real-Time Smishing Detection. In *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, pages 172–179, January 2024.
- [4] Chung-Gi Ban, Dayoung Park, and Youngbae Hwang. Image classification using DETR based object-level feature. In *2022 22nd International Conference on Control, Automation and Systems (ICCAS)*, pages 1297–1300, November 2022. ISSN: 2642-3901.
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers, May 2020. arXiv:2005.12872 [cs].
- [6] J. Dafni Rose, Jennifer Nissitta M, and Jaya Prabha S. Next-Gen Phishing Detection System Based on Federated Learning Integrated CNN-LSTM for SMS Communication. In *2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 367–372, March 2024.
- [7] F.C. Dalgic, A.S. Bozkir, and M. Aydos. Phish-IRIS: A New Approach for Vision Based Brand Prediction of Phishing Web Pages via Compact Visual Descriptors. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–8, October 2018.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs].
- [9] Ryan Heartfield and George Loukas. Detecting semantic social engineering attacks with the weakest link: Implementation and empirical evaluation of a human-as-a-security-sensor framework. *Computers & Security*, 76:101–127, July 2018.
- [10] Hung Le, Quang Pham, Doyen Sahoo, and Steven C. H. Hoi. URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection, March 2018. arXiv:1802.03162 [cs].
- [11] Jehyun Lee, Zhe Xin, Melanie Ng Pei See, Kanav Sabharwal, Giovanni Apruzzese, and Dinil Mon Divakaran. Attacking Logo-Based Phishing Website Detectors with Adversarial Perturbations. In Gene Tsudik, Mauro Conti, Kaitai Liang, and Georgios Smaragdakis, editors, *Computer Security – ESORICS 2023*, pages 162–182, Cham, 2024. Springer Nature Switzerland.
- [12] Yuexin Li, Chengyu Huang, Shumin Deng, Mei Lin Lock, Tri Cao, Nay Oo, Bryan Hooi, and Hoon Wei Lim. KnowPhish: Large Language Models Meet Multimodal Knowledge Graphs for Enhancing Reference-Based Phishing Detection, March 2024. arXiv:2403.02253 [cs].
- [13] Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, and Jin Song Dong. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 3793–3810. USENIX Association, 2021.
- [14] Ruofan Liu, Yun Lin, Xianglin Yang, Siang Hwee Ng, Dinil Mon Divakaran, and Jin Song Dong. Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach. In Kevin R. B. Butler and Kurt Thomas, editors, *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 1633–1650. USENIX Association, 2022.
- [15] Ruofan Liu, Yun Lin, Xianglin Yang, Siang Hwee Ng, Dinil Mon Divakaran, and Jin Song Dong. Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach (Datasets), 2022. <https://sites.google.com/view/phishintention/experiment-structure>.
- [16] Ruofan Liu, Yun Lin, Yifan Zhang, Penn Han Lee, and Jin Song Dong. Knowledge Expansion and Counterfactual Interaction for Reference-Based Phishing Detection. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*, pages 4139–4156, 2023.
- [17] Samuel Marchal, Giovanni Armano, Tommi Grondahl, Kalle Saari, Nidhi Singh, and N. Asokan. Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application. *IEEE Trans. Computers*, 66(10):1717–1733, 2017.
- [18] Samuel Marchal, Kalle Saari, Nidhi Singh, and N. Asokan. Know Your Phish: Novel Techniques for Detecting Phishing Sites and Their Targets. In *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, pages 323–333, June 2016. ISSN: 1063-6927.
- [19] Adam Oest, Yeganeh Safaei, Penghui Zhang, Brad Wardman, Kevin Tyers, Yan Shoshitaishvili, and Adam Doupe. PhishTime: Continuous Longitudinal Measurement of the Effectiveness of Anti-phishing Blacklists. In *29th USENIX Security Symposium (USENIX Security'20) Proceedings*, pages 379–396, 2020.
- [20] Jae Woo Seo, Jong Sung Lee, Hyunwoo Kim, Joonghan Lee, Seongwon Han, Jungil Cho, and Choong-Hoon Lee. On-Device Smishing Classifier Resistant to Text Evasion Attack. *IEEE Access*, 12:4762–4779, 2024.
- [21] Siddharth Sriraman, Sneha Sriram Kannan, Sonali Ravishankar, and B. Bharathi. An On-device Federated Learning System for SMS Spam Classification. In *2022 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–5, September 2022.

## Appendix

**Object-level Features Based Classification.** We tried to make *BrandClassifier* similar to the *LayoutClassifier* and reuse DETR object-level features to transform it into a tensor of shape (B, 16), which should have probabilities of belonging logo to some brand (Other or Apple, Amazon, Netflix, etc.). This approach, while yielding a high precision, suffers from a significantly lower recall rate. To address this issue, we experimented with the Oversampling and Undersampling techniques. Oversampling involves increasing the number of instances in the class with fewer instances, while Undersampling reduces the number of instances in the class with more instances. We observed a significant improvement in the model’s overall performance (Table 9), as evidenced by an F1 score that reached 90%. However, due to limited accuracy, we decided not to use this model in the final solution.

TABLE 9. LOGO BRAND CLASSIFICATION ACCURACY ACROSS DIFFERENT TRAINING TECHNIQUES

Technique	F1	Precision	Recall
Normal	0.83	0.95	0.79
Oversampling	0.79	0.77	0.87
Undersampling	0.90	0.98	0.86

**Webpage Screenshot Analysis Example.** In Figures 8 and 9, we illustrate the analysis of the official company webpage and phishing webpage detection by our solution.

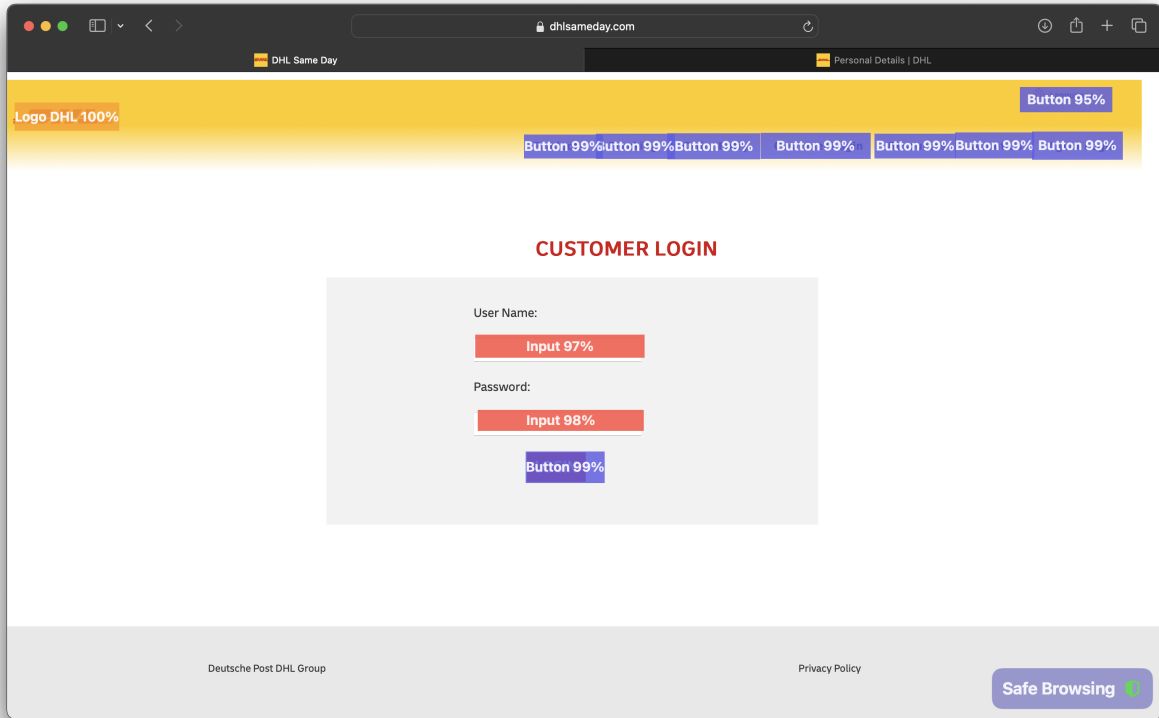


Figure 8. Official DHL website screenshot. This example shows the analysis of the official DHL website, showing the identification of various elements. Every button and input was accurately detected, and the logo was confidently identified as belonging to DHL. The website layout requires credentials, but the authenticity of dhl-sameday.com verifies that it is not a phishing site.

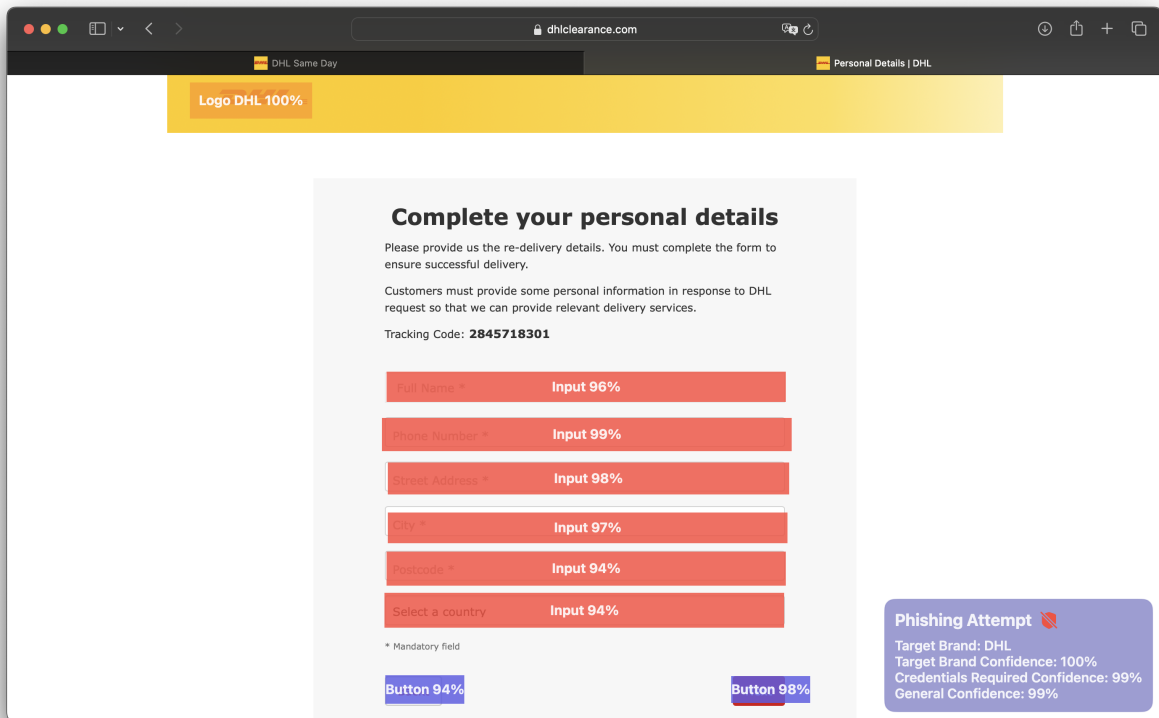


Figure 9. Phishing DHL webpage screenshot. Here, we encountered a layout markedly distinct from the previous example. Nevertheless, our evaluation methodically identified all buttons and inputs, and the layout classifier categorized this page as credential required. The detection of a logo was a key finding, with the classifier confirming with 100% confidence its affiliation with DHL. However, a closer inspection revealed a critical oversight: the site is not the official DHL webpage but a phishing attempt.