Combining Visual and Contextual Information for Fraudulent Online Store Classification

Wouter Mostard
mostard@dataprovider.com
Dataprovider.com
Groningen, The Netherlands

Bastiaan Zijlema
zijlema@dataprovider.com
Dataprovider.com
Groningen, The Netherlands

Marco Wiering
m.a.wiering@rug.nl
University of Groningen
Groningen, The Netherlands

ABSTRACT

Following the rise of e-commerce there has been a dramatic increase in online criminal activities targeting online shoppers. Considering that the number of online stores has risen dramatically, manually checking these stores has become intractable. An automated process is therefore required. We approached this problem by applying machine learning techniques to extract and detect instances of fraudulent online stores. Two sources of information were used to determine the legitimacy of an online store. First, contextual features extracted from the HTML and meta information were used to train various machine learning algorithms. Second, visual information, like the presence of social media logos, was added to make improvements on this baseline model. Results show a positive effect for adding visual information, increasing the F1-score from 0.93 to 0.98 over the baseline model. Finally, this research shows that visual information can improve recall during web crawling.

CCS CONCEPTS

• **Information systems** → *Web mining*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

fraud detection, machine learning, object detection, website classification

1 INTRODUCTION

Over the past decade there has been a rapid increase in e-commerce activities.[1] As a result, large companies are transitioning from physical stores to online equivalents. However, the increase in e-commerce also paved the way for novel criminal activities. This includes fraudulent online stores not delivering ordered products or stealing credit card information given by the consumer during check-out. This type of criminality has a negative influence on the e-commerce market, potentially leading to a diminishing consumer trust and brand reputation. How these fraudulent online stores can be automatically detected will be the main focus of this research.

Fraudulent online stores start looking more like their legitimate counterparts, e.g. by adding an address line or social media button. This makes fraudulent stores difficult to discriminate from their legitimate counterparts. Figure 1 shows an example of a fraudulent online store showing various payment method logos.

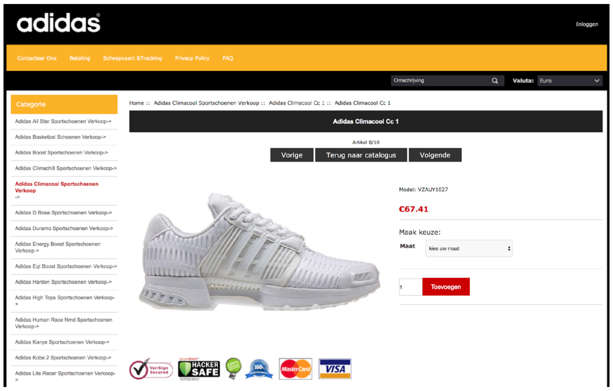
adidas
Inloggen
Contacteer Ons
Betaling
Scheepvaart &Tracking
Privacy Policy
FAQ
Omschrijving
Valuta: Euro
Categorie
Adidas All Star Sportschoenen Verkoop->
Adidas Basketbal Schoenen Verkoop->
Adidas Boost Sportschoenen Verkoop->
Adidas Climachill Sportschoenen Verkoop->
Adidas Climacool Sportschoenen Verkoop ->
Adidas D Rose Sportschoenen Verkoop->
Adidas Duramo Sportschoenen Verkoop->
Adidas Energy Boost Sportschoenen Verkoop->
Adidas Eqt Boost Sportschoenen Verkoop->
Adidas Harden Sportschoenen Verkoop->
Adidas High Tops Sportschoenen Verkoop->
Adidas Human Race Nmd Sportschoenen Verkoop->
Adidas Kanye Sportschoenen Verkoop->
Adidas Kobe 2 Sportschoenen Verkoop->
Adidas Lite Racer Sportschoenen Verkoop->
Home :: Adidas Climacool Sportschoenen Verkoop :: Adidas Climacool Cc 1 :: Adidas Climacool Cc 1
Adidas Climacool Cc 1
Artikel 8/10
Vorige
Terug naar catalogus
Volgende
Model: VZAUY1027
€67.41
Maak keuze:
Maat
kies uw maat
1
Toevoegen
VeriSign Secured
HACKER SAFE
MasterCard
VISA

Figure 1: Example of a Fraudulent Online Store

A model is proposed using various machine learning techniques and is laid out in two parts. First, a baseline model is put forward based on contextual features, such as whether a link to social media or a phone number is present. Second, visual information is added to improve classification results. The main contributions are as follows:

- A baseline system based on page, company, and website-level features is developed. Different machine learning techniques are compared based on this information. Furthermore, the coefficients of a logistic regression model are analyzed to determine feature importance.
- A computer vision system is created for extraction and classification of payment and social media logos in online stores, which yields production worthy performance on high resolution screenshots.
- An analysis has been conducted how visual information can both improve classification results for fraudulent online stores, as well as enrich contextual information. This includes looking at correlations to find discrepancies between visual cues and information found in the HTML, along with how this information can be used to increase web crawling performance.

The rest of the paper is organized as follows. Section 2 presents related work in classification of fraudulent online stores and methods for extracting visual information. In section 3 data collection,

[1] https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf

WI '19, October 14–17, 2019, Thessaloniki, Greece
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6934-3/19/10…$15.00
https://doi.org/10.1145/3350546.3352504

feature selection, and model selection for both the contextual and visual pipeline is described. Section 4 provides more details about the proposed implementation. Section 5 then presents the results and discussion of the performed experiments. The paper concludes with a conclusion and future work in section 6.

2 RELATED WORK

Although the classification of malicious online behaviour has been well studied [1][8], classification of fraudulent online stores is relatively new. In [17] the authors propose an approach where websites that sell counterfeit goods are represented by three types of information: URL-level, page-level, and website-level features. URL-level features contain information such as the presence of the word "replica" in the URL, indicating that an online store sells fake products. Page-level features are based on the HTML content, while website-level features are based on meta-information. For example the Alexa ranking, which is an estimate of a website's popularity, is used. Notably, the website-level features, e.g. presence of large I-Frames and percentage of savings on items, seem to be discriminative features. In [17] a mention about generating a screenshot of all visited URLs is being made, but there is no further discussion as to what has subsequently been done with it. In [5] the authors propose a system where the trustworthiness of an online store is established, based on the search results of a particular search engine query. The paper gives a list of relevant features that are also used in this research, such as the presence of an email address or social media details, but lacks information regarding the importance of features. Furthermore, more than half of the used contextual features are different and no visual information is used. Lastly, results are focused on the query results give certain familiar clothing brands, while our approach is brand independent. A Natural Language Processing approach is proposed in [13] by applying sentiment analysis on a feature vector generated from the body of textual content on the website. In [18] the authors show that the presence of social media has a significant effect on classification performance. An important conclusion is that social media information is more prevalent on legitimate websites, compared to fraudulent online stores. Fraudulent online stores try to mimic this effect by placing social media buttons. Currently, no research has been conducted in finding these fake social media buttons and the interplay with information found in the HTML.

At present, no empirical research has been conducted that combines visual features with meta-features when determining the legitimacy of an online store. Thus, the added value of this particular research is two-fold. First, it shows that using another modality in combination with the page-level features can greatly improve classification results of fraudulent online stores. Second, visual information can be used to enrich the information gathered by the web crawler. An example is the presence of a Western Union payment method as suggested in [5]. It is understood that a large portion of online stores using Western Union as a payment method only give notification about this by showing a button on the front page. For a HTML based web crawler it would be nearly impossible to detect these instances considering the filenames of these images are quite often described with a generic term, such as "paymentmethods.png". For this reason a method for extracting visual information from the online stores is required. In practice *convolutional neural network* (CNN) [11] are applied, considering its known satisfactory results in the area of logo detection [2][15].

3 METHOD

3.1 Data Collection

A set of positive, i.e. fraudulent, and negative, i.e. legitimate, online stores is required for building the classifier. In May 2018 the Dutch consumers association, Consumentenbond, published an article regarding fraudulent online stores[2]. This article enclosed a publicly available list of 1833 domains, these have been added to the fraudulent store data set. Furthermore, local law enforcement added another 43 examples to this set of fraudulent stores.

Thuiswaarborg[3] is a Dutch e-commerce association which provides a trust mark for online stores in the Netherlands. Considering all online stores that are listed on the website are manually checked for legitimacy, this list can serve as the set of negative examples. A list of 1499 registered domain names was extracted from the Thuiswaarborg website and added to the set of legitimate online stores.

3.2 Feature Selection

3.2.1 Contextual features used for online store classification.

(1) *Page-level features:* Page-level features refer to features that can be found using the HTML shown on the first 50 crawled pages of a domain. It is expected that fraudulent online stores will only use a select number of payment methods [5] and lack valid links to social media platforms [18]. Other features include: the number of pages and products, HTML size, and the number of payment methods found in image tags or the textual body. Where legitimate online stores have few products available it is expected that fraudulent online stores show a larger collection of items.

(2) *Company-level features:* A legitimate appearance is of particular importance for fraudulent online stores. This is pursued by adding information that suggests the stores' legitimacy. This includes information such as the presence of a *business registration number* (brn), phone number, or physical address.

(3) *Website-level features:* Website-level features relate to information that cannot be directly observed through the HTML. The hypothesis is that legitimate stores will have a larger number of incoming and outgoing links, given that fraudulent stores will have a closed environment and are less often referred to by other websites. Legitimate online stores will take measures to make their website safe for customers. Having a large open port count can be a serious risk for visiting customers. It is thus expected that legitimate online stores will have fewer open ports than their fraudulent counter parts.

Using data available at Dataprovider.com, the contextual features shown in Table 1 were retrieved for each domain. The domains that

[2] https://www.consumentenbond.nl/nieuws/2018/consumentenbond-laat-850-foute-webwinkels-offline-halen

[3] https://www.thuiswinkel.org/english

Table 1: Contextual Features

Feature name	Description
PAGE-LEVEL FEATURES	
html_size	size homepage
pages	number of pages found
products	number products
num_social_html	# social found in HTML
num_payment_html	# payments found in HTML
shopping_cart_system	shopping cart system found
copyright	copyright found
analytics_id	Google analytics ID found
COMPANY-LEVEL FEATURES	
company_name	company name found
brn	whether a brn is found
tax_number	tax number presence
phone_number	phone number presence
email_address	presence email
address	presence address
WEBSITE-LEVEL FEATURES	
incoming_links	# incoming hyperlinks
outgoing_links	# outgoing hyperlinks
open_ports_count	number of open TCP ports

were not indexed have been removed from the data set. Furthermore, online stores that did not return a 200 OK HTTP response have been removed. This would imply that the online store was already taken down or redirected and thus could not serve as a training example, considering that the data could not be verified. All the features were normalized using MinMax normalization. The equation for MinMax normalization is as follows:

$$\text{normalize}(x) = \frac{x - min(x)}{max(x) - min(x)}$$

Missing values were imputed using the mean of the train set during 10-fold cross validation. For the binary variables, zeros were imputed for every missing value. In total there were 2000 online stores selected, 1000 legitimate and 1000 fraudulent.

3.2.2 Visual information. A $1920 \times n$ pixels screenshot of the front page is generated for each of the domains retrieved. In order to generate a representative image of the front-page, all content is loaded before generating a screenshot. As a result, the screenshot algorithm will wait until all elements of the page are loaded before generating a screenshot. Some domains have implemented an auto-scroll functionality. Consequently, the height of the screenshot is initially unknown.

To generate a preliminary training set the coordinates of logos have been annotated using an open source library called sloth[4]. *Regions of Interest* (RoIs) have been cropped from the screenshots in order to learn the background class as proposed in [2]. Only RoIs that have an *Intersection over Union* (IoU) of 0 with any of the annotations are added to the data set. The IoU is the intersection divided by the union of two RoIs. This ensures that no parts of the logos are added to the background class.

In order to improve logo classification, two additional data sets have been generated with 1535 and 16999 screenshots respectively. After classification of each set the false positives and true positives are added to the data set and the model is retrained. This process is repeated until satisfactory results on both the training and test data were achieved in all three data sets. To make sure that the results are representative, no domain that was used for training the logo detection classifier was used in the test set.

Table 2: Distribution of Logo Dataset

Class	# Training examples
PAYMENT METHODS	
American Express	38
MasterCard	1137
Maestro	543
PayPal	2175
Visa	846
SOCIAL MEDIA	
Facebook	292
Instagram	136
Twitter	82
BACKGROUND	
background	5572

3.2.3 Visual Features. The 5 most frequently occurring payment methods and 3 social media platforms are selected as the target class for logo classification. As can be seen from Table 2, there is a large class imbalance. During the process of iteratively adding the false positives and true positives as described in section 3.1 it was found that some classes showed more class variance, and thus were harder to learn. For this reason, more examples have been added for these classes. One example is the MasterCard class, of which an excerpt is shown in Figure 2.

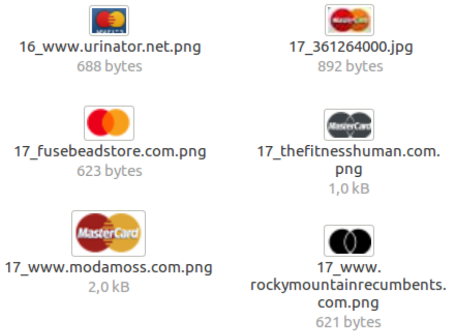

Figure 2: Excerpt of MasterCard Training Examples

The most extreme case is the background class, where each image patch was unique. Other classes, such as American Express, showed little class variance and thus fewer examples sufficed.

[4] https://github.com/cvhciKIT/sloth

3.3 Model Selection

The classification system has been developed in Python using scikit-learn. A support vector machine (SVM) [6], random forest [3], and k-nearest neighbors (KNN) [7] classifier were trained with their best found hyperparameters to determine the classification performance of using only the contextual features. The hyperparameters were determined using a 10-fold cross validated grid search on the contextual features data set as described in section 3.1. For the SVM an RBF kernel was applied and the parameters C and γ were determined. Maximum depth and minimum sample leave size were determined for the random forest as well as the optimal number of neighbors for the KNN classifier.

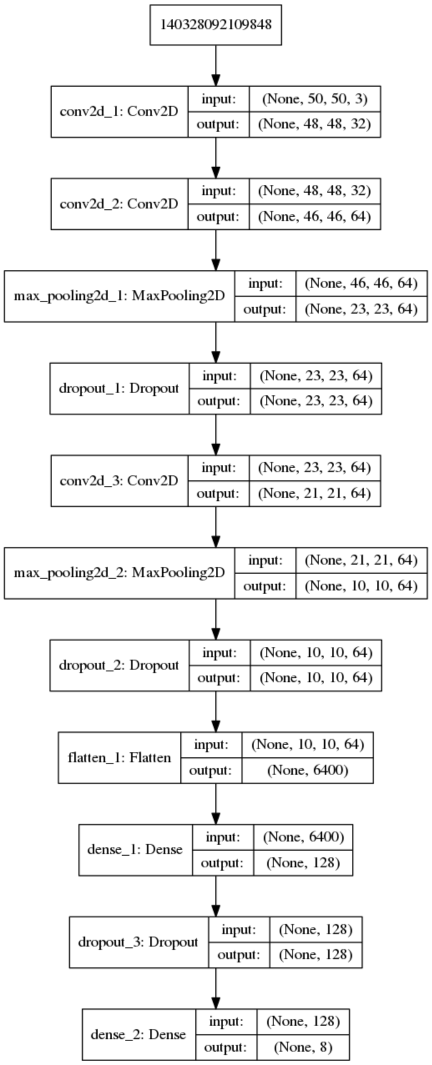
140328092109848
conv2d_1: Conv2D
input: (None, 50, 50, 3)
output: (None, 48, 48, 32)
conv2d_2: Conv2D
input: (None, 48, 48, 32)
output: (None, 46, 46, 64)
max_pooling2d_1: MaxPooling2D
input: (None, 46, 46, 64)
output: (None, 23, 23, 64)
dropout_1: Dropout
input: (None, 23, 23, 64)
output: (None, 23, 23, 64)
conv2d_3: Conv2D
input: (None, 23, 23, 64)
output: (None, 21, 21, 64)
max_pooling2d_2: MaxPooling2D
input: (None, 21, 21, 64)
output: (None, 10, 10, 64)
dropout_2: Dropout
input: (None, 10, 10, 64)
output: (None, 10, 10, 64)
flatten_1: Flatten
input: (None, 10, 10, 64)
output: (None, 6400)
dense_1: Dense
input: (None, 6400)
output: (None, 128)
dropout_3: Dropout
input: (None, 128)
output: (None, 128)
dense_2: Dense
input: (None, 128)
output: (None, 8)

Figure 3: CNN Architecture for Logo Classification

For the visual features pipeline a CNN has been developed in Keras. Its model architecture is depicted in Figure 3. The 8 output classes correspond to the 7 target classes and a background class. A 25% drop-out rate was applied to prevent overfitting [16]. All layers were initialized using the Xavier uniform initializer [10]. The convolutional layers apply a ReLU activation, while the last fully connected layer applies softmax. The weights were trained using a batch size of 64 50×50 input images for 40 epochs. Accuracy increase was flattened at around 35 epochs, with an average validation accuracy of 96% on the logo annotations.

4 IMPLEMENTATION

The contextual features are retrieved by doing an API call to the Dataprovider.com database. Missing values are imputed and normalized as described in section 3.

Object detection is a two phase problem based on the Region-Based Convolutional Neural Network proposed in [9]. First, the image needs to be segmented into RoIs. However, instead of using the selective search algorithm as described in [9] this research employs a simpler method, using the Canny edge detection algorithm implemented in OpenCV [4]. The reason for using this method is the fact that online stores tend to have low variability in color and texture, thus a simpler and faster segmentation method would suffice. The input RGB image is converted into grey-scale. The image is pre-processed with a 3×3 Gaussian kernel to remove noise and improve edge detection. An automated Canny Edge segmentation method[5] is applied to automatically determine the upper and lower hysteresis and a dilatation kernel of 3×3 and $\sigma = 0.33$ is used.

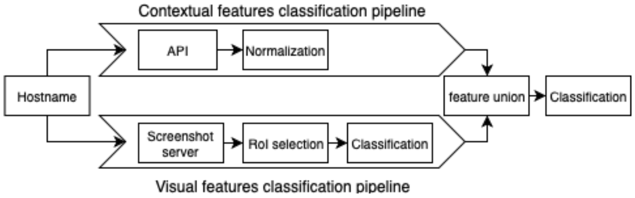

Figure 4: Proposed Model

Contours from the binarized image are found using the OpenCV findContours function. Considering that all logos will adhere some aspect ratio and size to remain recognizable, only RoIs with a ratio of $1.0 < ratio < 4.0$ and an area of $w * h > 400$ pixels are taken into account. These remaining RoIs are fed into the classifier. A probability threshold has been set at 0.90 to prevent false positives when classifying RoIs as one of the target classes. Any classified RoI with a probability of < 0.90 will be classified as background. The final proposed model is shown in Figure 4.

5 RESULTS AND DISCUSSION

5.1 Baseline Results

Table 3 shows the baseline results and standard deviations based on contextual features using 10-fold cross validation and the corresponding best found hyperparameters for each classifier. The random forest classifier was the best performing classifier scoring an F1-score of 0.93, while having the lowest standard deviation. The best found hyperparameters for the random forest classifier were: no specified *maximum depth*, *minimum sample* size of leaves is 1.

[5] https://www.pyimagesearch.com/2015/04/06/zero-parameter-automatic-canny-edge-detection-with-python-and-opencv/

Table 3: Baseline Performance for SVM, Random Forest, and KNN Using Contextual Features

Class	Precision	Recall	F1-score
SVM	0.83 ±0.24	**0.94** ±0.05	0.88 ±0.15
Random Forest	**0.94** ±0.09	0.91 ±0.11	**0.93** ±0.05
KNN	0.91 ±0.14	0.93 ±0.09	0.92 ±0.09

Table 4 shows the coefficients, standard errors, and p values of a logistic regression classifier trained using the entire data set with ground truth labels. Features with $p < 0.001$, i.e. statistically significant features, are bold faced. As expected, the presence of social media and address information showed to be of significant importance. Interestingly, the presence of a shopping cart system was found to be statistically significant with a large positive weight. Possibly, this was due to the fact that criminals use a standard embedded code with a well known shopping cart system, while legitimate stores would invest in developing their own shopping cart system.

Table 4: Coefficients, Standard Error, and p Value for the Contextual Features Using a Logistic Regression Classifier

feature name	coef	std err	p-value
PAGE-LEVEL FEATURES			
html_size	-0.014	0.116	0.9017
pages	**0.159**	**0.036**	**0.0000**
products	0.201	0.104	0.0536
count_payment_html	-0.009	0.01	0.3507
count_social_html	**-0.111**	**0.012**	**0.0000**
shopping_cart_system	**0.45**	**0.021**	**0.0000**
copyright	**0.324**	**0.021**	**0.0000**
analytics_id	-0.042	0.028	0.1383
COMPANY-LEVEL FEATURES			
company_name	0.013	0.028	0.6396
brn	-0.011	0.046	0.8148
tax_number	-0.032	0.062	0.6040
phone_number	**-0.108**	**0.025**	**0.0000**
email_address	**-0.126**	**0.03**	**0.0000**
address	**-0.142**	**0.034**	**0.0000**
WEBSITE-LEVEL FEATURES			
incoming_links	0.415	0.325	0.2022
outgoing_links	-0.237	0.199	0.2347
open_ports_count	**1.556**	**0.122**	**0.0000**

Handcrafted pieces of code are harder to detect than their third party counterparts due to the variety in code. A large number of open ports showed to be a negative indicator of legitimacy. While legitimate online stores had on average 6.75 open ports, the fraudulent online stores had on average 12.23 open ports. This 5.48 increase could be due to the fact that criminals do not care about the increased risk that a large number of open ports brings to the customer. Lastly, the number of social media accounts found showed a significant negative effect on the probability of an online store being fraudulent. This is explained by the fact that fraudulent online stores usually do not link to social media. A total of 412 links to social media were found on the legitimate websites, while 0 links were found in any of the fraudulent online stores. This was also observed in [18].

5.2 Benefits of Adding Visual Information for Fraudulent Online Store Classification

To calculate how often the logos of various payment and social methods co-occurred with contextual features the Matthews correlation coefficient (MCC) was calculated. The Matthews correlation coefficient indicates the correlation between two binary sets and is given by the following equation:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

A high MCC would imply that a link or reference to social media would co-occur frequently with its corresponding logo on the front page.

Table 5: Matthews Correlation Coefficient Between the Mentioning of Link to Social Media and Payment Methods Found on the Website and Classified Logos

	Legitimate	Fraudulent
American Express	0.14	0.03
MasterCard	0.11	0.01
Maestro	0.38	0.09
PayPal	0.26	0.22
Visa	0.16	0.13
Facebook	0.05	0.00
Twitter	0.13	0.00

Table 5 shows the Matthews correlation coefficient for each payment method and social media profile, indexed for both the legitimate and fraudulent online stores. The *Instagram* class has been removed from these results because it generated too many false positives. All other logos showed a precision score of larger than 0.90 and were thus specified as giving a reliable prediction. All showed a higher correlation coefficient for the legitimate class, thus implying that the visual and contextual features of the payment and social media classes co-occur more frequently in legitimate online stores. This strengthens our hypothesis that fraudulent online stores try to look legitimate by adding various logos.

This relationship was exploited to improve classification. For each of the classified logos a separate feature called *<logo>_no_html* was developed and added to the contextual feature vector. This Boolean feature depicted the discrepancy between visual and contextual information and was given a 1 if a logo was found by the logo detector but no reference was found in the HTML. In all other cases this feature was set to 0. This resulted in 7 new features. This information was added to the contextual feature vector by simple concatenation. Adding the co-occurrence between a button and the HTML reference to the contextual feature vector increased performance, as can be seen in Table 6. The best performing classifier is the random forest classifier with an F1-score of 0.98. For the random forest classifier the precision and recall increased by 5% and 7% respectively. The F1-score increased with 5%, while showing a decrease in standard deviation.

Table 6: Classification Performance for SVM, Random Forest, and KNN Using Contextual and Visual Features

Class	Precision	Recall	F1-score
SVM	0.89 ±0.03	**0.98** ±0.02	0.93 ±0.02
Random Forest	**0.99** ±0.005	**0.98** ±0.02	**0.98** ±0.01
KNN	0.98 ±0.03	0.96 ±0.01	0.97 ±0.01

5.3 Improve Information Retrieval Using Visual Information

As an ultimate test for the benefit of adding visual information we examined the output of the contextual model on unseen online stores. To make sure that each class would have approximately the same number of instances, classification continued until 5000 instances were classified for each class. Removing the domains that did not return a HTTP 200 code resulted in a data set with 9145 domains. 4778 of these were classified as legitimate and 4376 as fraudulent. Next, a screenshot was generated of each hostname in the final data set. Payment and social logos were retrieved using the computer vision pipeline. Table 7 shows the number of instances found using the HTML and visual information. The increase in instances found when adding hostnames where a visual cue is found but no reference in the HTML are in bold face.

Table 7: Increase of Instances Found by Adding Visual Information

	American Express	MasterCard	Maestro	PayPal	VISA	Facebook	Twitter
Legitimate							
page	983	1486	336	1959	1741	3377	1959
+ visual	1218	2146	483	2089	2439	3425	2301
increase	**24%**	**44%**	**43%**	**12%**	**40%**	**1%**	**18%**
Fraudulent							
page	43	312	68	666	336	1	0
+ visual	363	1275	710	807	1156	60	239
increase	**774%**	**309%**	**944%**	**21%**	**215%**	**5900%**	**x**

The number of payment logos found using visual cues in the legitimate class increased on average with 26%. For the fraudulent class recall increased on average with 450%. This means that the average increase of finding logos in the absence of a mention in the HTML is 424% higher in the fraudulent class. This enforces our hypothesis that online criminals show various payment methods to seem more legitimate, while only supporting few. Furthermore, it suggests that criminals tend to use improper filenames. If no mention of the logo is made on the website it could very well be that all the logos are placed under a single image named for example *payment_methods.png*. On legitimate websites this phenomenon is less common.

It was observed that a large portion of legitimate online stores had Scalable Vector Graphics (SVG) in XML format. These graphics load dynamically and are used to allow for different screen resolutions. For a static HTML crawler it is impossible to determine the presence of payment logos in these cases. This problem is alleviated when a fully rendered screenshot is used. This result is depicted by the increased number of instances found in for example the MasterCard class in the legitimate online stores.

On average, a 9.45% increase in recall is found for the social media profiles for the legitimate online stores. Considering that links to social media profiles are virtually non-existent on the fraudulent online stores these statistics are excluded. These statistics show two things. First, the absolute number of social media profiles found in legitimate online stores is much larger than in fraudulent online stores, an observation also made by [18]. Second, the increase of social media logos found is much larger in the fraudulent class. This again suggests that criminals try to mimic legitimate online stores by placing fake social media buttons.

6 CONCLUSION AND FUTURE WORK

This research demonstrates a robust method for improving classification performance of counterfeit websites. First, it shows that features other than standard HTML crawled features, for example the number of open ports, significantly improve the classification results. Second, it validates that logos can be extracted from online stores using a simple segmentation technique and classified using a standard CNN. This research also reveals that visual cues, such as the presence of social media buttons, hold valuable information, both for the classification of counterfeit online stores and for enriching information that is found during web crawling. This allows for improvements to the quality of the data and invites novel research regarding the interplay between HTML and what is actually shown on a webpage.

Future research should be conducted regarding the relationship between contextual and visual features for fraudulent online store classification and other applications. This includes research in determining the type of products sold on an online store by analyzing the product image and description. Furthermore, it is the intention to improve the object detection pipeline by comparing contemporary classification systems such as Faster-RCNN [14] and Single Shot Detectors [12] for faster and more precise classification of high resolution website screenshots. Additionally, criminals are constantly improving the quality of fraudulent online stores in order to increase resemblance to legitimate online stores. This form of *concept drift* requires further research.

REFERENCES

[1] Sushma Nagesh Bannur, Lawrence K. Saul, and Stefan Savage. 2011. Judging a Site by Its Content: Learning the Textual, Structural, and Visual Features of Malicious Web Pages. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (AISec '11)*. ACM, New York, NY, USA, 1–10.

[2] Simone Bianco, Marco Buzzelli, Davide Mazzini, and Raimondo Schettini. 2017. Deep Learning for Logo Recognition. *Neurocomput.* 245, C (July 2017), 23–30.

[3] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (2001), 5–32.

[4] John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8, 6 (Nov 1986), 679–698.

[5] Claudio Carpineto and Giovanni Romano. 2017. Learning to Detect and Measure Fake Ecommerce Websites in Search-engine Results. In *Proceedings of the International Conference on Web Intelligence (WI '17)*. ACM, New York, NY, USA, 403–410.

[6] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (Sept. 1995), 273–297.

[7] Thomas Cover and Peter Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theor.* 13, 1 (1967), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[8] Matthew F. Der, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2014. Knock It off: Profiling the Online Storefronts of Counterfeit Merchandise. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1759–1768.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2016. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 1 (Jan. 2016), 142–158.

[10] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Yee Whye Teh and Mike Titterington (Eds.), Vol. 9. PMLR, Chia Laguna Resort, Sardinia, Italy, 249–256.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (Nov 1998), 2278–2324.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 21–37.

[13] Mahdi Maktabar, Anazida Zainal, Mohd Aizaini Maarof, and Mohamad Nizam Kassim. 2018. Content Based Fraudulent Website Detection Using Supervised Machine Learning Techniques. In *Hybrid Intelligent Systems*, Ajith Abraham, Pranab Kr. Muhuri, Azah Kamilah Muda, and Niketa Gandhi (Eds.). Springer International Publishing, Cham, 294–304.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (June 2017), 1137–1149.

[15] Jürgen Schmidhuber. 2015. Deep Learning in Neural Networks. *Neural Netw.* 61, C (Jan. 2015), 85–117.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 1929–1958.

[17] John Wadleigh, Jake Drew, and Tyler Moore. 2015. The E-Commerce Market for "Lemons": Identification and Analysis of Websites Selling Counterfeit Goods. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1188–1197.

[18] KuanTing Wu, ShingHua Chou, ShyhWei Chen, ChingTsorng Tsai, and Shyan-Ming Yuan. 2018. Application of Machine Learning to Identify Counterfeit Website. In *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services (iiWAS2018)*. ACM, New York, NY, USA, 321–324.