Receipt Dataset for Document Forgery Detection

Beatriz Martínez Tornés, Théo Taburet, Emanuela Boros, Kais Rouis, Petra Gomez-Krämer, Antoine Doucet, Nicolas Sidere, Vincent Poulain D'andecy

Receipt Dataset for Document Forgery Detection

Beatriz Martínez Tornés[1[0000−0002−7820−640X]], Théo Taburet[1[0000−0001−8165−6826]], Emanuela Boros[1[0000−0001−6299−9452]], Kais Rouis[1[0000−0002−1709−3683]], Antoine Doucet[1[0000−0001−6160−3356]], Petra Gomez-Krämer[1[0000−0002−5515−7828]] Nicolas Sidere[1[0000−0001−6719−5007]], and Vincent Poulain d'Andecy[2]

[1] University of La Rochelle, L3i, F-17000, La Rochelle, France
firstname.lastname@univ-lr.fr
[2] Yooz, 1 Rue Fleming, 17000 La Rochelle, France
Vincent.PoulaindAndecy@getyooz.com

Abstract. The widespread use of unsecured digital documents by companies and administrations as supporting documents makes them vulnerable to forgeries. Moreover, image editing software and the capabilities they offer complicate the tasks of digital image forensics. Nevertheless, research in this field struggles with the lack of publicly available realistic data. In this paper, we propose a new receipt forgery detection dataset containing 988 scanned images of receipts and their transcriptions, originating from the scanned receipts OCR and information extraction (SROIE) dataset. 163 images and their transcriptions have undergone realistic fraudulent modifications and have been annotated. We describe in detail the dataset, the forgeries and their annotations and provide several baselines (image and text-based) on the fraud detection task.

Keywords: Document forgery · Fraud detection · Dataset.

1 Introduction

Automatic forgery detection has become an inevitable task in companies' document flows, as accepting forged documents can have disastrous consequences. For instance, a fraudster can submit forged proofs leading to identity theft or to obtain a loan for financing criminal activities such as terrorist attacks. However, proposed research works lack generality as they are very specific to a certain forgery type or method and hence the same applies to available datasets.

One of the main challenges in document fraud detection is the lack of freely available annotated data. Indeed, the collection of fraudulent documents is hampered by the reluctance of fraudsters to share their work, as can be expected in any illegal activity, as well as the constraints on companies and administrations to share sensitive information [18, 15, 22, 21]. Moreover, many studies on fraud do not focus on the documents themselves, but on the transactions, such as insurance fraud, credit card fraud or financial fraud [4, 13, 17].

We thus attempt at bridging this gap between the lack of publicly available forgery detection datasets and the absence of textual content, by building a new generic dataset for forgery detection based on real document images without promising data confidentiality. We based the dataset on an existing dataset of scanned receipts (SROIE) that initially was proposed for information extraction tasks, and contains images and text. Furthermore, we, then, tampered the images using several tampering methods (copy and paste, text imitation, deletion of information and pixel modifications) and modified the textual content accordingly. We, thus, provide the textual transcriptions allowing text-only analysis, and we present several baselines for image and text-based analysis for benchmarking.

The rest of this article is organized as follows. Section 2 reviews and analyses existing datasets for document forgery detection. Our new dataset is presented in Section 3. Section 4 presents our experiments and the baselines. Finally, we conclude our work and discuss perspectives in Section 5.

2 Related Work

Since one of the main challenges of investigating fraud in such documents is the lack of large-scale and sensitive real-world datasets due to security and privacy concerns, a few datasets were previously proposed.

The Find it! dataset [1] is freely available and contains 1,000 scanned images of receipts and their transcriptions. However, the images were only acquired by one imaging device, i.e., a fixed camera in a black room with floodlight, so the task of classification into tampered and untampered images in the Find it! competition [2] could be easily solved with image-based approaches. Furthermore, fixed camera acquisition is not a realistic scenario as most documents are acquired today by a scanner or a smartphone. Only one combined text/image approach was submitted.

The Forgery Detection dataset [18] was synthetically built, and it contains 477 altered synthetic payslips in which nearly 6,000 characters were forged. It is mainly conceived for forgery localization. However, the dataset is rather small, and no transcriptions were provided for text analysis. While these transcriptions could be obtained by an OCR, there would be still a need for manual annotations. Also, as the data is synthetically produced, it is difficult to use it in semantic text analysis. As a matter of fact, the different fields (names, companies, addresses etc.) of the payslips have been randomly filled from a database, so the forged documents are no more incoherent than the authentic ones.

Another publicly available dataset is Supatlantique [16], a collection of scanned documents mainly conceived for the problem of scanner identification, containing 4,500 images annotated with respect to each scanner. It also addresses the problem of forgery detection but includes only very few tampered images and no textual transcriptions.

Some other small datasets have been proposed for several specific image-based tasks such as the stamp verification (StaVer) dataset[3], the Scan Distortion

[3] http://madm.dfki.de/downloads-ds-staver

dataset[4] for detecting forgeries based on scan distortions, the Distorted Text-Lines dataset[5] containing synthetic document images where the last paragraph is either rotated, misaligned or not distorted at all and the Doctor bills dataset[6].

Hence, out of all the datasets proposed for forgery detection, most of them are not usable for content-based approaches: not only do they target a specific image detection task, which can lead to duplicated content or synthetically generated content, but they are also limited in size. This dataset addresses these limitations by proposing a pseudo-realistic multimodal dataset of forged and authentic receipts. Compared to the Find it! dataset, our dataset is based on a dataset acquired with several scanners and varying compression factors, which is more realistic. In total, there are 41 different quantization matrices for JPEG compression settings in the original images of the SROIE dataset. Furthermore, our dataset could be used in addition to the Find it! dataset allowing multilingual (English and French) text processing as well as increasing the variety of imaging devices and compression factors for image-based processing.

3 Dataset Building for Forged Receipts Detection

Taking an interest in real documents actually exchanged by companies or administrations is essential for the fraud detection methods developed to be usable in real contexts and for the consistency of authentic documents to be ensured. However, these administrative documents contain sensitive private information and are usually not made available for research [2]. We consider the task of receipt fraud detection, as receipts contain no sensitive information and have a very similar structure to invoices. In that way, realistic scenarios can be associated with receipt forgery, such as reimbursement of travel expenses (earn some extra money, non-reimbursed products), and proof of purchase (for insurance, for warranty).

3.1 SROIE Dataset

The dataset was chosen as a starting point to create the forgery dataset. It was originally created for scanned receipts OCR and information extraction (SROIE) as an ICDAR 2019 competition and contains 1,000 scanned receipt images along with their transcriptions.

One characteristic of the scanned receipts of this dataset is that some have been modified, either digitally or manually, with different types of annotations. These annotations are not considered as forgeries. Even though the documents have been modified, they are still authentic, as they have not undergone any forgery. These annotations suit our case study, as most of them are context-specific notes found in real document applications. For instance, some annotations are consistent with notes left on the receipts, such as “staff outing” to

[4] http://madm.dfki.de/downloads-ds-scandist
[5] http://madm.dfki.de/downloads-ds-distorted-textline
[6] http://madm.dfki.de/downloads-ds-doctor-bills

describe the nature of the event (Figure 3), numbers that can describe a mission or a case number (or any contextual numerical information) (Figures 1, 2, and 5), names (Figure 6) or markings to highlight key information on the document, such as the price in Figure 4. Many of such annotations might even come from the collection process of the dataset and are difficult to interpret without contextual queues (names, numbers, etc.).

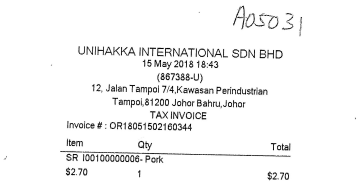

Fig. 1. Numerical insertion.

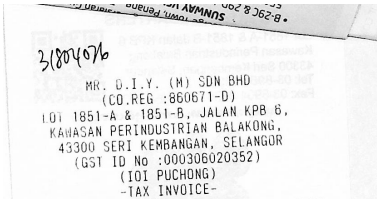

Fig. 2. Numerical insertion.

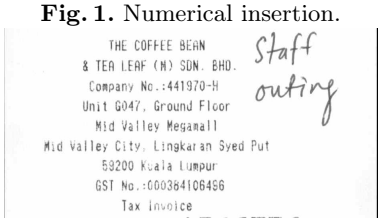

Fig. 3. Note on the receipt.

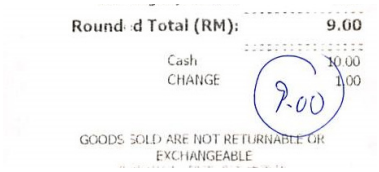

Fig. 4. Highlight of the total amount.

Fig. 5. Numerical insert.

Fig. 6. Name insert.

These modifications in authentic documents address the challenging issue in fraud detection to distinguish between a fraudulent modification and a non-malicious modification. A fraudulent modification is characterized not only by the ill intent of the perpetrator, but also by the fact that it changes crucial

structural or meaningful features of the document that can be used to distort the meaning the original document supports.

In order to evaluate the impact of such modifications, we manually annotated the authentic receipts according to the type of modifications they have undergone. We consider a digital annotation as a particular recurrent case of a sequence of numbers or names added digitally in several receipt headers, as shown in Figures 5 and 6, respectively. We consider that there has been a manual annotation (Figures 1, 2, 3 and 4) if there is a handwritten note of any type on the receipt (words, checkmarks, highlighted or underlined areas, etc.), as well as stamps. Table 1 shows the results of our manual annotation of the SROIE documents. We noted that the digital annotations are reported in the transcriptions, whereas manual annotations are not.

3.2 Forgery Campaign

In order to provide forgeries as realistically as possible, we organized several forgery workshops similar to the ones carried out by pseudo-realistic forged datasets [2, 18]. The 19 participants were volunteers, mainly from a computer science background, even if we attempted to enlarge the scope of our project to different levels of competence and expertise in digital documents and image editing software. The goal was not to create a dataset of expert forgeries, but to have a realistic representation of different skills and time commitments. Participants were not provided with any specific guidelines on what tools and techniques to use, in order for them to use whatever they were most comfortable with. Five different software were used: `preview` (15 documents), `paint` (70), `paint3d` (10), `GIMP` (65), and `kolourpaint` (3).

Image and Text Modification Participants were provided with examples and scenarios to get started, such as the reimbursement of travel expenses (to earn extra money, to hide unauthorized products), proof of purchase for insurance, and proof of purchase for warranty (e.g., date too old). They were asked to modify the image as well as its text file (transcription).

Forgery Annotation Next, participants were asked to annotate the forgeries they just had performed using the VGG Image Annotator[7]. These annotations are provided with the dataset in JSON format, as shown in the following example:

```
{'filename': 'X51005230616.png', 'size': 835401, 'regions':
[{'shape_attributes': {'name': 'rect', 'x': 27, 'y': 875,
'width': 29, 'height': 43},
'region_attributes': {'Modified area': {'IMI': True},
'Entity type': 'Product', 'Original area': 'no'}},
{'shape_attributes': {'name': 'rect', 'x': 458, 'y': 883,
```

[7] https://www.robots.ox.ac.uk/∼vgg/software/via/

'width': 35, 'height': 37},
'region_attributes': {'Modified area': {'IMI': True},
'Entity type': 'Product', 'Original area': 'no'}}],
'file_attributes': {'Software used': 'paint', 'Comment': ''}}

The process consisted, first, in the annotation of the areas they had modified using rectangular region shapes, and second, in the description of the forgery type according to the categorization proposed in [5] (copy-paste from within the same document, copy-paste from another document, information suppression and imitation). Furthermore, we include an extra forgery type, PIX, for all "freehand" modifications [9]. We, thus, proposed the following forgery types for tampering:

- **CPI**: Copy and paste inside the document, i.e., copy a section of the image (a character, a whole word, a sequence of words, etc.) and paste it into the same image;
- **CPO**: Copy and paste outside the document, that is to say, copy a section of the image (a character, a whole word, a sequence of words, etc.) and paste it into another document;
- **IMI**: Text box imitating the font, using a text insertion tool to replace or add a text;
- **CUT**: Delete one or more characters, without replacing them;
- **PIX**: Pixel modification, for all modifications made "freehand" with a brush type tool to introduce a modification (for example, transforming a character into another by adding a line);
- ***Other***: Use of filters or other things (to be specified in comments).

If the same area has undergone two types of changes, more than one modification type can be selected. For example, an internal copy-paste can be followed by some retouching with a brush-type tool in order to change the colour. That would result in a combination of types CPI and PIX.

Modified Entity Annotation Participants were also asked to identify the entity type they had tempered with for every modified area from the following list:

- **Company**: Information related to the company and its contact details (address, phone, name);
- **Product**: Information related to a product (name, price, removal or addition of a product);
- **Total/Payment**: Total price, the payment method or the amount paid;
- **Metadata**: Date, time.

The annotators were also asked, in the particular case of copy-paste forgeries in the same document (CPI), to locate the original area that has been copied. This annotation was performed sequentially: after annotating the modified area, they were asked to annotate the area they had used to modify it.

These three steps that the participants were asked to follow (image and text modification, forgery annotation, and modified entities annotation) made it possible to get pseudo-realistic forgeries and annotations on their spatial location and semantic content.

3.3 Post-processing

All annotations provided for the forged receipts are manual. As manual annotations can be error-prone, we manually corrected all the annotations to ensure that the modified areas were correctly annotated. Annotating the original area of copy-paste forgeries posed the most problems for the participants, as they often forgot the area they had copied (especially when they had modified several characters of different items in a single receipt). Some problems also arose in specific scenarios that were harder to annotate: the case of switching two characters out, as the original area no longer exists in the forged document, and the case of a character copied in several places, as the instructions did not specify how to annotate in this situation. As different annotators treated those cases differently, we normalized the annotations. Only one original annotation was kept per area, even when the area was used in more than one copy-paste, and even when the original area was itself modified. However, there is nothing we could do for forgotten original areas, or clearly erroneous ones (different characters). We removed the erroneous annotations in order to keep only annotations we were sure of. Therefore, not all the original areas have been annotated: 200 original areas/356 CPI areas.

The most common errors we encountered were missing the original area, mislabelling between the original and the modified area, mislabelling of the entity types, missing software-used annotation and missing transcript updates. Corrections were made manually by comparing the annotations of the forged documents to the originals to ensure the consistency of the labels, which was a very time-consuming process. The goal was to have usable annotations, in order to provide a dataset that could be used not only for a classification task (between the forged and authentic classes) but also for a forgery localization challenge. In order to correct missing software-used clauses, we emitted the hypothesis that every participant that did not specify a software for every modified receipt used the same one as in its other forgeries.

3.4 Dataset Description

The resulting dataset contains 988 PNG images with their corresponding transcriptions in text format[8]. We propose a data split into train, validation and test sets in order to allow comparison between different methods. The data split is described in Table 1, with counts of the forgeries committed during the forgery campaign (Section 3.2) as well as the annotations present in the authentic receipts (Section 3.1).

[8] The data is available for download at http://l3i-share.univ-lr.fr/2023Finditagain/findit2.zip.

Table 1. Data splits.

	Train	**Validation**	**Test**	**Total**
Number of receipts	577	193	218	988
Number of forged receipts	94	34	35	163
% of forged receipts	16	18	16	16
Number of digitally annotated receipts	34	9	11	54
% of digitally annotated receipts	6	5	5	5
Number of manually annotated receipts	305	86	109	500
% of manually annotated receipts	53	45	50	50

In total, 455 different areas were modified, across 163 receipt documents. Table 2 details how many modifications have been conducted by type: one area can have been affected by more than one type of modification. With regard to the entities, most modifications targeted the total or payment information. Also, we can observe in the table that the most used forgery technique is CPI.

Table 2. Modified areas description.

Modification type	Counts	Entity type	Counts
CPI	353	Total/payment	234
IMI	36	Product	95
CUT	36	Metadata	82
PIX	33	Company	26
CPO	10	Other	18

4 Experiments - Baselines

We describe below four baselines on the proposed dataset that can be used for the comparison of new research work on this dataset. We present two text-based methods and two image-based methods.

4.1 Text Classification

We tested two methods of text classification, logistic regression and ChatGPT, which are described below.

Bag-of-words (BoW) & Logistic Regression First, we chose this statistical language model used to analyse the text and documents based on word count since it generally serves as a foundation model and can be used as a benchmark to evaluate results and gain a first insight regarding the difficulty of the task. We consider the most commonly utilized model for a simple and straightforward baseline: logistic regression (LR).

ChatGPT The recent model created by OpenAI[9], proposed in November 2022, has gained immense attention in both academic and industrial communities, being shortly adopted by all types of users, not only due to its impressive ability to engage in conversations, but also to its capacity of responding to follow-up questions, paraphrasing, correcting false statements, and declining inappropriate requests [8]. Specifically, the technology behind ChatGPT is a Transformer-based architecture trained through reinforcement learning for human feedback on a vast corpus of web-crawled text, books, and code. We, thus, were curious to compare the responses of an expert human and ChatGPT to the same question [3]. We followed a straightforward zero-shot approach to retrieve responses from ChatGPT via the official web interface[10] between January 17th and 19th, 2023.

We, thus, decide on the following prompt:

```
Extract the locations (LOC), products (PROD) and prices (PRI)
from the following receipt and tell me if it's fraudulent:{receipt}
```

Based on this prompt, for each document, we utilize ChatGPT to generate answers to these questions by replacing **{receipt}** with the unmodified text of each receipt. Since ChatGPT is currently freely available only through its preview website, we manually input the questions into the input box, and get the answers, with the aid of some automation testing tools. The answers provided by ChatGPT can be influenced by the chatting history, so we refresh the thread for each question (each document). ChatGPT can generate slightly different answers given the same question in different threads, which is perhaps due to the random sampling in the decoding process. However, we found that the differences can be very small, thereby we only collect one answer for most questions, and we propose an evaluation with several configurations.

4.2 Image Classification

We tested two methods of image classification: SVM and JPEG compression artefact detection.

Pixels & SVM Classification To provide an initial baseline based on image information, we chose the support vector machine (SVM), which is more suited to the size of our dataset than convolutional neural networks. The idea was also to evaluate a simple and straightforward baseline on our dataset, not a specific forgery detection approach. We resized the images to 250×250 and normalized them. We, then, trained an SVM with a linear kernel with default hyperparameters.

JPEG Compression Artefact Detection Based on the fact that the images of the SROIE dataset are JPEG images, we make the hypothesis that in case

[9] https://openai.com/blog/chatgpt/
[10] https://chat.openai.com

of fraud on the images they would be saved using a simple (Ctrl+S) in JPEG format. Thus, the modified areas would undergo a simple JPEG compression (because it would be original content) while the rest of the image would be subject to additional JPEG compression (double or triple compression).

We used the bounding boxes of the SROIE files as well as the modified bounding boxes to split each image into overlapping crops (128×128), the areas containing a modification are thus labelled as fraudulent. In order to make the fraudulent image crops more realistic (in the JPEG context), from the PNG file, we compressed the fraudulent crop using the same quantization matrix as its non-fraudulent pair. Otherwise, the generated JPEG images would have been too easy to detect, because they would have all had exactly the same quantization matrix.

Finally, we selected all the crops containing a tampered area and gave them the label "tampered", and we equally and randomly selected crops from original images and from images that have been tampered with, but whose crops do not contain any. This results in three sub-datasets, balanced between the tampered and non-tampered classes.

Table 3. Data splits for JPEG double compression artefact detection.

	Train	**Validation**	**Test**	**Total**
Number of 128×128 crops	7,747	2,583	2,582	12,912

We used a convolutional neural network (CNN) model, which was previously proposed for the detection of document manipulations in JPEG documents [19]. The proposed method utilizes a one-hot encoding of the DCT (Discrete Cosines Transformed) coefficients of JPEG images to compute co-occurrence matrices. The authors declined this network through two approaches: OH-JPEG and OH-JPEG+PQL. Both use a one-hot encoding of the JPEG coefficients for each image (OH-JPEG), the second approach uses a Parity-Quantization-Layer (OH-JPEG+PQL) which consists in inserting parity information (provided by the quantization matrix of the JPEG file) thus allowing the network to detect possible discrepancies in the images.

After training the network for 200 epochs on the designated database, the metrics reached a plateau, upon which they were recorded. The training was performed using the AdaMax optimizer (a variant of Adam [11] based on the infinity norm) and the multistep learning rate scheduler (a scheduling technique that decays the learning rate of each parameter group by gamma once the number of epochs reaches one of the milestones), with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}.

4.3 Evaluation

For all baselines, we utilize the standard metrics: precision (P), recall (R), and F1-score. For ChatGPT, as the answers in the free-form text do not correspond to binary classification results, we align them following two configurations:

- **Strict**: Only the answers that expressed precise doubts or notable elements about the receipt or its legitimacy were labelled as "forged". Only for seven receipts, the answer did explicitly state that something was "worth noting", "suspicious" or seemed or appeared "fraudulent".
- **Relaxed**: We also considered that a receipt was labelled as forged if the answer did not lean towards authentic or suspicious. We thus considered that if the answer did not refer to any appearance of authenticity, then the receipt was suspicious, and was therefore labelled as forged.

The two configurations we chose to evaluate the ChatGPT results are intended to give an account of the confidence expressed in the answers. Let us consider two different answers:

> "It is not possible for me to determine whether the receipt is fraudulent or not, as I do not have enough information and context."

> "It is not possible for me to determine if the receipt is fraudulent or not, as I don't have enough information about the context or the business. However, it appears to be a legitimate receipt based on the format and information provided."

In the strict configuration, we consider that the receipt that prompted the first answer, as it does not remark on anything suspicious, is just as authentic as the receipt that produced the second answer. However, in the relaxed evaluation, the first receipt is considered forged. These two evaluation set-ups come from a qualitative analysis of the results.

4.4 Results

This section presents the results for the above-presented baseline methods. Table 4 reports the results of the classification task. As the dataset is very imbalanced, we report only the results of the "Forged" class. As the JPEG compression artefact detection approach, was tested on a balanced version of the dataset (see Table 3), only the precision results are reported. Indeed, as there is approximately the same number of tampered and non-tampered documents, the precision is also approximately equal to the recall and F1-score.

The best precision results are obtained for the JPEG compression artefact detection method. While the text and image classification methods we have tested yield better precision results than the ChatGPT approach, they score significantly lower in terms of recall. In a forgery detection task, one would prioritize a high recall, as it is preferred to have approaches that are more sensitive towards

identifying the "Forged" class. In that respect, the image classification approach and the very low recall show how insufficient it is. The text classification approach, even if it performed slightly better, remains equally insufficient. Only four forged receipts were correctly labelled by the text classifier. For these reasons, we will only analyse the results of the ChatGPT and the JPEG compression artefact detection methods.

Table 4. Results of the tested approaches.

Method	Precision	Recall	F1-score
Text classification (BoW + LR)	40.00	11.43	17.78
Image classification (SVM)	30.00	8.57	13.33
ChatGPT (strict)	14.69	**88.57**	25.20
ChatGPT (relaxed)	18.33	62.86	**28.39**
OH-JPEG	**79.41**	-	-
OH-JPEG+PQL	78.39	-	-

ChatGPT Analysis ChatGPT performed better overall than the text and image classification approaches. However, it is worth noting that ChatGPT, in a strict evaluation configuration, predicted only seven receipts as forged, which explains the high recall. However, the first three baselines proposed yield low results, as we can observe from the F1-score. For 113 receipts (out of the 218 test receipts), the answers underline the task's difficulty without leaning towards an authentic or forged label, such as

> "It is not possible to determine if the receipt is fraudulent based on the information provided."

> "It is not clear from the receipt provided whether it is fraudulent."

> "I'm sorry, as a language model, I am unable to determine if the receipt is fraudulent or not, as I don't have access to the context such as the store's standard price list, so I can't compare the prices of the products."

The rest of the answers (for 105 receipts) do express a certain decision on whether the receipt is fraudulent, with varying degrees of certainty: "It doesn't appear to be fraudulent." or "It is not a fraudulent receipt." for instance. These answers can be accompanied by justifications, either related to the task or the receipt and its contents. Only seven answers explicitly declare that the receipt could be fraudulent. In two of them, the answers state that the company name ("TRIPLE SIX POINT ENTERPRISE 666") and the discount offered to make it suspicious. Even if those are entities that could be modified and do deserve to be checked, both of these receipts are legitimate, and so is the company name. ChatGPT correctly found discrepancies between prices and amounts in two receipts:

> "This receipt may be fraudulent, as the quantity and price of WHOLEMEAL seem to be incorrect. Ten units at a unit price of 2.78 RM is 27.8 RM, but it appears to be 327.8 RM in the receipt, which is a significant discrepancy. I'd recommend you to verify the receipt with the vendor and the government tax authority."

> "It appears that the receipt is fraudulent, as the total and cash values do not match with the calculation of the product's total prices. It would be best to double-check with the seller or authorities for proper investigation."

However, in another case, the difference between the total amount of the products (5RM) and the cash paid (50RM) was reported as suspicious, even if the change matched the values and the receipt was authentic. In two instances, some issues were noted with the dates of the receipts. For one of them, two different dates were present in the receipt and one of them was indeed modified: the inconsistency was therefore apparent and duly noted by ChatGPT. However, the other receipt contained only one date that was modified (2018 changed to 2014). Surprisingly, this date was reported suspicious, even if there is no apparent inconsistency. The justification given was that "This receipt appears to be fraudulent as the date is 03/01/2014 and the knowledge cut-off is 2021."

JPEG Compression Artefact Detection Analysis This method outperforms the other baselines in terms of precision. Several key observations can be made from the results:

- The network is capable of detecting the presence of original content, or more specifically, detecting correlation breaks between blocks, even when the original images have undergone multiple compressions;
- The performance of the network is limited, which can be attributed to several factors.

One of the main challenges in implementing these approaches is the low entropy nature of document images, making it difficult to extract meaningful statistics in the JPEG domain. Furthermore, the document images used in this study are mostly blank, making it challenging for a CNN to accurately determine the authenticity of the image. This can be also due to the size of the manipulated regions which is relatively small, and the fact that spatial and DCT-domain semantics are relatively consistent, given that most manipulations consist of internal copy-paste operations. These results leave great windows for improvement, as the JPEG artefact detection approaches show their limits here. Indeed, these approaches ignore the semantics of the image and are vulnerable to some basic image processing such resizing, binarization, etc.

5 Conclusions and Perspectives

This paper presents the freely available receipt dataset for document forgery detection, containing both images and transcriptions of 988 receipts. It also

provides semantic annotations on the modified areas, as well as details on the forgery techniques used and their bounding boxes. Thus, the dataset can be used for classification and localization tasks. We also experimented with straightforward methods for a classification task, using either the textual content or the image. These experiments are very limited and aim to provide examples of what can be done, as well as underline the difficulty of the task at hand.

We believe that this dataset can be an interesting resource for the document forgery detection community. The experiments presented can be considered as a starting point to compare with other methods, in particular multimodal approaches, which we believe to be very promising in this field, but also specific forgery detection approaches, such as copy-move detection [20] and further JPEG compression artefact detection. Indeed, the method that yielded the best results was the only one from the forensic document field, and it still leaves room for improvement. The focus of this dataset is its semantic and technical consistency: by undergoing a forgery campaign where participants were free to use the techniques they felt most comfortable with, images acquired by different means, some even with digital or manual annotations, it establishes a challenging task to test forgery detection methods within a realistic context.

Limitations

Since we considered a ChatGPT baseline in a zero-shot manner, we are aware that ChatGPT is lacking context for predicting the existence of fraud. Following previous research that explored GPT-3 for different tasks [7], including plagiarism detection [6], public health applications [10, 12] and financial predictions, it was interesting to explore it as it can act as a fraud detector. Our intention was to study its ability. ChatGPT is not as powerful as GPT-3, but it is better suited for chatbot applications. Moreover, for now, it is freely available, which is not the case for GPT-3, thus, this posed another limitation to experimenting with GPT-3. While we found that ChatGPT is able to simulate an answer that seemed realistic, most of them were invented and thus, invalid.

Ethics Statement

With regard to the chosen baselines, ChatGPT, while it can generate plausible-sounding text, the content does not need to be true, and, in our case, many answers were not. Being grounded in real language, these models inevitably inherit human biases, which are amplified and cause harm in sensitive domains such as healthcare, if not properly addressed. As previously demonstrated through the use of the Implicit Association Test (IAT), such Internet-trained models as ChatGPT and GPT-3 tend to reflect the level of bias present on the web [14]. While the impact of this model is not direct as being associated with gender biases or the usage in healthcare considering forgery detection, we still draw attention to the fact that an undetected fraud for the wrong reasons could impact drastically the confidence of the systems.

Acknowledgements

We would like to thank the participants for their contribution to the creation of the dataset. This work was supported by the French defence innovation agency (AID), the VERINDOC project funded by the Nouvelle-Aquitaine Region and the LabCom IDEAS (ANR-18-LCV3-0008) funded by the French national research agency (ANR).

References

1. Artaud, C., Doucet, A., Ogier, J.M., d'Andecy, V.P.: Receipt dataset for fraud detection. In: First International Workshop on Computational Document Forensics (2017)
2. Artaud, C., Sidère, N., Doucet, A., Ogier, J.M., Yooz, V.P.D.: Find it! Fraud detection contest report. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 13–18 (2018)
3. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al.: A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861 (2021)
4. Behera, T.K., Panigrahi, S.: Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In: 2015 Second International Conference on Advances in Computing and Communication Engineering (2015)
5. Cruz, F., Sidère, N., Coustaty, M., Poulain d'Andecy, V., Ogier, J.M.: Categorization of document image tampering techniques and how to identify them. In: International Conference on Pattern Recognition. pp. 117–124. Springer (2019)
6. Dehouche, N.: Plagiarism in the age of massive generative pre-trained transformers (gpt-3). Ethics in Science and Environmental Politics **21**, 17–23 (2021)
7. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines **30**, 681–694 (2020)
8. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y.: How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597 (2023)
9. James, H., Gupta, O., Raviv, D.: Ocr graph features for manipulation detection in documents (2020)
10. Jungwirth, D., Haluza, D.: Feasibility study on utilization of the artificial intelligence gpt-3 in public health. Preprints (2023)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Korngiebel, D.M., Mooney, S.D.: Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. NPJ Digital Medicine **4**(1), 93 (2021)
13. Kowshalya, G., Nandhini, M.: Predicting fraudulent claims in automobile insurance. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (2018)
14. Lucy, L., Bamman, D.: Gender and representation bias in gpt-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding. pp. 48–55 (2021)

15. Mishra, A., Ghorpade, C.: Credit card fraud detection on the skewed data using various classification and ensemble techniques. In: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (2018)
16. Rabah, C.B., Coatrieux, G., Abdelfattah, R.: The supatlantique scanned documents database for digital image forensics purposes. In: 2020 IEEE International Conference on Image Processing (ICIP) (2020)
17. Rizki, A.A., Surjandari, I., Wayasti, R.A.: Data mining application to detect financial fraud in indonesia's public companies. In: 2017 3rd International Conference on Science in Information Technology (ICSITech) (2017)
18. Sidere, N., Cruz, F., Coustaty, M., Ogier, J.M.: A dataset for forgery detection and spotting in document images. In: 2017 Seventh International Conference on Emerging Security Technologies (EST) (2017)
19. Taburet, T., Rouis, K., Coustaty, M., Gomez-Krämer, P., Sidère, N., Kébairi, S., d'Andecy, V.P.: Document forgery detection using double JPEG compression. In: 2022 ICPR Workshop on Artificial Intelligence for Multimedia Forensics and Disinformation Detection (AI4MFDD) (2022)
20. Teerakanok, S., Uehara, T.: Copy-move forgery detection: A state-of-the-art technical review and analysis. IEEE Access **7**, 40550–40568 (2019). https://doi.org/10.1109/ACCESS.2019.2907316
21. Tornés, B.M., Boros, E., Doucet, A., Gomez-Krämer, P., Ogier, J.M., d'Andecy, V.P.: Knowledge-based techniques for document fraud detection: A comprehensive study. In: Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019, La Rochelle, France, April 7–13, 2019, Revised Selected Papers, Part I. pp. 17–33. Springer (2023)
22. Vidros, S., Kolias, C., Kambourakis, G., Akoglu, L.: Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. Future Internet **9**(1) (2017)