# Medical Image Tampering Detection

Priscila Moreira, Mahsa Mitcheff
pmoreira@nd.edu,mmitchef@nd.edu
University of Notre Dame

## ABSTRACT

The transmission of medical data over the Internet has increased with the more frequent use of the Telemedicine. At the same time, the health area has become a target for cyberattacks [1]. That vulnerability raises concerns about disclosing patient personal data and its consequences when tampering with medical images. Thus, it is desirable to design a solution that aims to solve the medical image forgery **localization** and **detection** problem. We adopt a well established neural network-based object detection called Single Shot Multibox Detector (SSD) [5] to solve this challenging issue. Using the LuNoTim dataset [7] which consists of images of tampered CT scan slices, the trained SSD model achieved a mean average precision (mAP) of 73.4% on the test set, which is promising and can offer some analyses for this type of image domain.

## 1 INTRODUCTION

With the rising usage of Telemedicine, the transmission of medical data over the Internet has expanded. At the same time, cyberattacks have turned their attention to the health sector [1]. That vulnerability raises concerns about disclosing patient personal data and its consequences when tampering with medical images. This project aims to solve the medical image forgery localization and detection problem. We adopt a well established neural network-based object detection called Single Shot Multibox Detector (SSD) [5] to solve this challenging issue.

In the training process, we adopted the LuNoTim dataset [7] which consists of images of tampered CT scan slices. we also have tested a ManTraNet [8], a tampering detection algorithm for natural image, on the images from LuNoTim dataset to verify if this model is able to find tampered regions in medical images. we could observe that it does not generalize for the new dataset of medical images. So the decision was starting training a SSD model on this data. My final SSD model achieves a mean average precision (AP) of 73.4% on the test set, which is promising and can offer some analyses for this type of image domain.

## 2 METHODOLOGY

### 2.1 Model Selection

SSD is a unified object detection framework that uses a single network [5]. The input image and ground truth data, which is information about an object's bounding box, are required for this object detector model. SSD begins the training phase by using the default bounding boxes, which vary in location and aspect ratio (to handle a variety of objects of different sizes), with the goal of minimizing the jacquard overlap to a threshold value and matching the default box to the ground truth box.

*Conference'17, July 2017, Washington, DC, USA*
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

### 2.2 Data Collection

For training the SSD, we decided to use data from the LuNoTim-CT dataset contains 7, 202 total tampered lung CT scans ($512 * 512$) with 356, 217 slices by different tampering methods: *copy-move*, *classical inpainting*, and *deep inpainting*, see Fig.1. For the last one, they used the CTGAN [6] for removing and adding nodules from/to the original CT scans in the original LIDC-IDRI dataset [2].

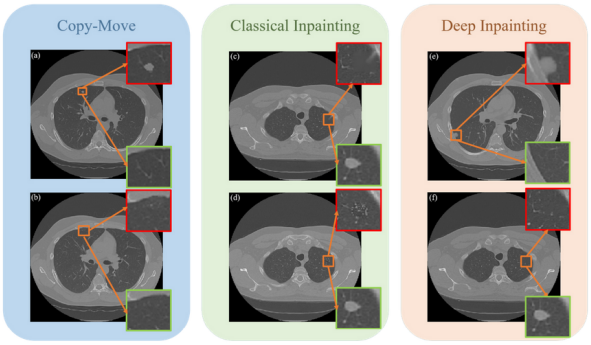



**Figure 1: LuNoTim Dataset. Samples of different tampering methods. Obtained from [7].**

There are two steps to prepare the dataset:

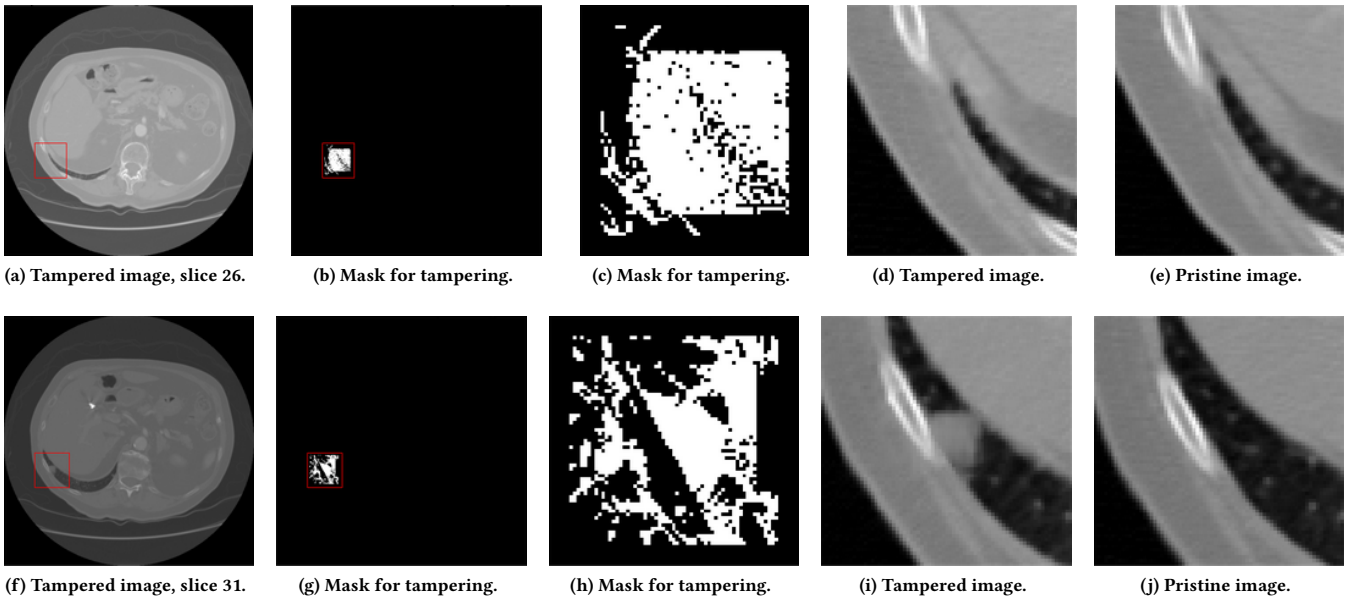
- Determining the quality of tampered images, where we curated the database, and
- Preparation of the curated dataset for the Pascal VOC format and then the conversion for tfrecords structures, which is required by the SSD TensorFlow implementation, which contains all information about bounding boxes of tampered regions and image information in byte format.

For the step 1, we created scripts to visualize the tampered images and its masks provided by LuNoTim, also, we had to integrated this dataset with the original source of CT Scan dataset LIDC-IDRI dataset [2], so we could understand the type and quality of tampered areas.

In the Fig.2 you can see the the data integration, Fig.2(d) presents a "bad" forgery that only noise was added, and Fig.2(e) presents a good forgery where we can see a nodule added. That "bad" forgery occurs because they were generated by deep inpainting using the CTGAN model using neighborhood texture and adding Gaussian noise.

More specifically, the data inside LuNotim dataset is subdivided into eight categories. Table 1 summarizes information about the number of samples for each subcategory.

We split the precessed dataset after the curation into train, validation, and test with 67%,13%, and 20%, from each type of tampering, respectively. See Table 2.



(a) Tampered image, slice 26. (b) Mask for tampering. (c) Mask for tampering. (d) Tampered image. (e) Pristine image.

(f) Tampered image, slice 31. (g) Mask for tampering. (h) Mask for tampering. (i) Tampered image. (j) Pristine image.

**Figure 2: Data integration and curation.**

**Table 1: Total No. tampered slices in each tampering category.**

| Tampering Category | Total No. tfrecords |
|---|---|
| Added inner tissue different slice (ITDS-add) | 031,666 |
| Added inner tissue same slice (ITSS-add) | 029,887 |
| Added outer tissue different slice (OTDS-add) | 032,074 |
| Added outer tissue same slice (OTSS-add) | 030,597 |
| Added ct gan inpainting (ctGAN-add) | 001,595 |
| Removed patchmatch guided inpainting (PGI-rem) | 013,678 |
| Removed simple inpainting (SI-rem) | 044,497 |
| Removed ct gan inpainting (ctGAN-rem) | 927,000 |

**Table 2: Dataset split.**

| | |
|---|---|
| Total No. Train | 099,148 |
| Total No. Validation | 024,787 |
| Total No. Test | 030,984 |
| Total No. Samples | 154,920 |

## 3 EXPERIMENTAL RESULTS

we used a TensorFlow implementation of the SSD which has as input feature maps from a VGG-16 backbone, which was trained on the pascal VOC07+12 dataset[3, 4]. The detection process has two main steps: 1) training the SSD model on the input image and 2) post-processing the output using common algorithms, *top-k filtering* and *Non-Maximum Suppression algorithm*. The SSD model was set to classify two types of classes: background (non tampered areas) and tampered areas.

When we ran the SSD with the initial values of the hyperparameters (weight decay=0.0005, optimizer=adam, learning rate=0.001, and batch size=32), we found that it produced a loss value of "Nan." To address this issue, we reduced the learning rate to avoid model underfitting. We changed the learning rate value in the range of 1e-2 to 1e-7 following as recommended in the literature. I also switched between two distinct optimizers: "ADAM" and "SGD". Changing the hyperparameters and using the "ADAM" optimizer helped with starting with a lower loss function value and decreasing the loss value faster.

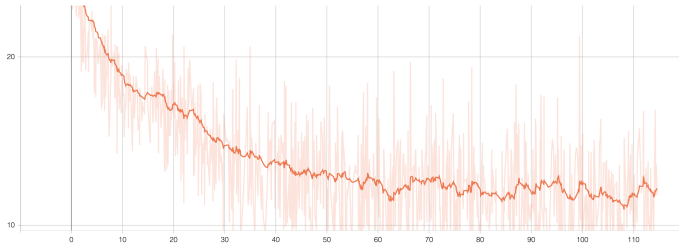



**Figure 3: Training loss curve for SSD**

Although at the beginning the loss value was very high (Fig. 3), it started to decrease along the iterations and the model learned to detect and localize some types of forgeries in medical images. The model as evaluated in the test set unseen in the training and presented an mean average precision around 73%.

Fig. 4 shows total and localization losses and Fig. 5 shows cross entropy positive and negative) losses, respectively.

At prediction time, the network generates scores for the presence of each class in each default box and produces adjustments to the box to better match the object shape. The trained model could correctly localize some of the tampered regions although sometimes it failed to correctly detect and localize the tampered areas. See

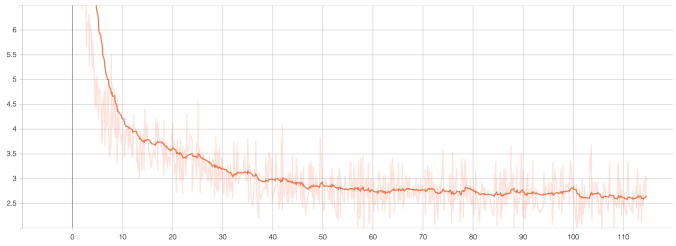
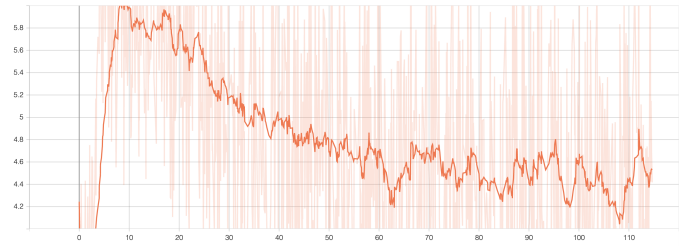
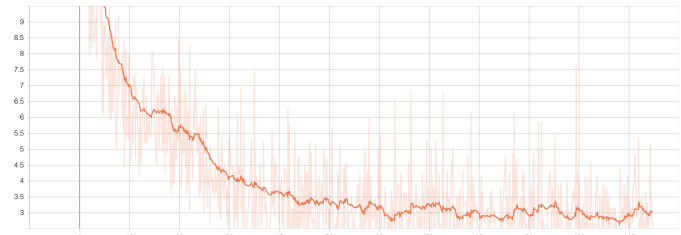


**Figure 4: localization loss**



(a)



(b)

**Figure 5: cross entropy loss(a) positive (b) negative**

some results by category of forgery in Fig. 6 and Fig. 7, where the tampered regions were correctly detected (cyan boxes), and some red rectangles indicating truly tampered areas that were not detected. First number on the left side shows the class number (1: means tampered) and the second number is the Confidence score, which is the probability that an anchor box contains an object and how accurate the bounding box is.

In some tampered areas that were not detected, sometimes size of the tampered area is very small or have a texture similar to the inner part of the lung.

To show the efficiency of SSD in detecting and localizing the tampered areas we ran ManTraNet model [8] on our dataset. Although ManTraNet could detect some of the obvious tampering, it had a lot of false positives. Remembering that the ManTraNet model was trained in natural images, we would like to investigate its performances when trained in medical images. ManTra-Net may fail when: a) a forged image was intentionally contaminated with highly correlated noise, and b) multiple regions were manipulated differently.The highest, and lowest F1 score was 0.46 (for "removed simple inpainting"), and 0.0 (for added ct gan inpainting), respectively. The Average overall F1 score was 0.08 which is very low remembering that this algorithm was designed to detect forgey in natural images

## 4 ANALYSES

It is important to notice that the tampered regions in our dataset ranged from 400 to 6000 in area, which is significantly smaller than the image area of 262,144. The largest tampered area on a CT slice was only 0.022 of the whole slice, making it difficult for SSD to locate its position. This problem was also observed in the original SSD model trained on natural images and they observed a worse performance on smaller objects than larger objects.

Another issue is that the texture of tampered areas is similar to the texture of the background (inner part of the lung), and the difference between the two classes is minimal, making it even more difficult for SSD to discern between the objects specially for inner tissue tampering. Based on the results, it seems that it is easy for SSD to locate obvious tampering like added outer tissues. But, maybe it cannot detect CT-GAN removed/added nodules because in ctGAN they tried to hide tampered areas by adding gaussian noise and combining the texture of pristine areas with the texture of tampered areas.

For the future research we can take several steps:

- Train SSD on individual tampering separately especially for CTGAN which had more false positive;
- Train a shallow CNN to classify the bound boxes returned by the SSD model;
- Make a model based on the original CT images and feed them to the new model trying to distinguish the tampered images. This approach is similar to NoisePrint and ManTraNet algorithms.

## 5 CONCLUSION

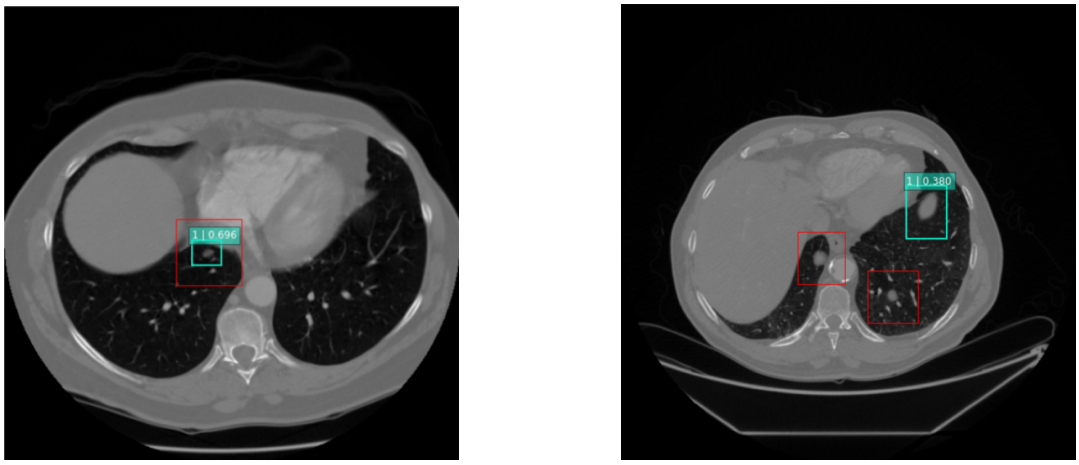
In this work, we proposed an object detection algorithm that learned from images of CT scan slices and performed tampering detection in multiple tampering categories. We trained an SSD with a VGG-16 backbone in the LuNoTim processed dataset. The final model achieved on the test set a mean average precision of 73.4% (mean averaged over different IOUs). Experimental results showed that the model could detect different types of manipulations, but it still misses some small tampered areas.

## 6 CONTRIBUTION

The contributions by different group members are shown in the Table 3.

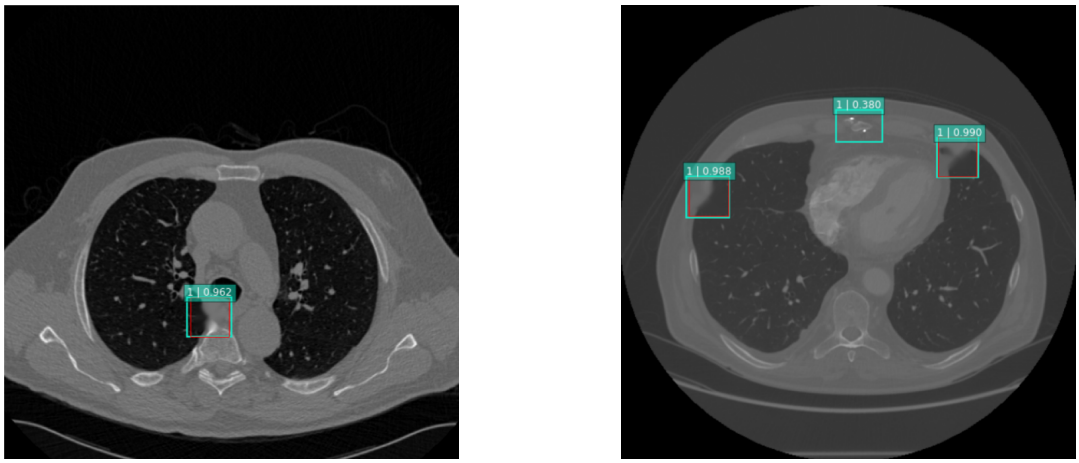
**Table 3: Contribution**

| | Priscila | Mahsa |
|---|---|---|
| Data preparation | 50% | 50% |
| SSD model training | 60% | 40% |
| ManTranet testing | 40% | 60% |
| Report writing | 50% | 50% |




(a) (b)

**Figure 6: Tampered images using CTGAN. Ground truth: red boxes. Predicted: cyan boxes.**




(a) (b)

**Figure 7: Classical Inpainting. Groundtruth: red boxes. Predicted: cyan boxes.**

## REFERENCES


[1] Salem T Argaw, Nefti-Eboni Bempong, Bruce Eshaya-Chauvin, and Antoine Flahault. 2019. The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review. *BMC medical informatics and decision making* 19, 1, 1–11.

[2] McLennan G. Bidaut L. McNitt-Gray M. F. Meyer C. R. Reeves A. P. Zhao B. Aberle D. R. Henschke C. I. Hoffman E. A. Kazerooni E. A. MacMahon H. Van Beek E. J. R. Yankelevitz D. Biancardi A. M. Bland P. H. Brown M. S. Engelmann R. M. Laderach G. E. Max D. Pais R. C. Qing D. P. Y. Roberts R. Y. Smith A. R. Starkey A. Batra P. Caligiuri P. Farooqi A. Gladish G. W. Jude C. M. Munden R. F. Petkovska I. Quint L. E. Schwartz L. H. Sundaram B. Dodd L. E. Fenimore C. Gur D. Petrick N. Freymann J. Kirby J. Hughes B. Casteele A. V. Gupte S. Sallam M. Heath M. D. Kuhn M. H. Dharaiya E. Burns R. Fryd D. S. Salganicoff M. Anand V. Shreter U. Vastagh S. Croft B. Y. Clarke L. P. Armato III, S. G. [n.d.]. (2015). Data From LIDC-IDRI [Data set]. The Cancer Imaging Archive. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. [n.d.]. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

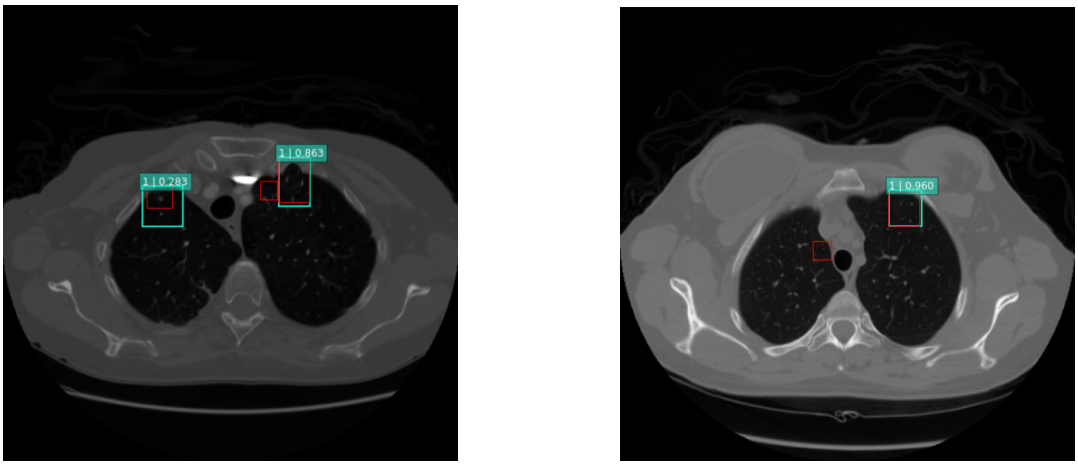
[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. [n.d.]. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[6] Yisroel Mirsky, Tom Mahler, Ilan Shelef, and Yuval Elovici. 2019. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. arXiv:1901.03597 [cs.CR]

[7] Benjamin Reichman, Longlong Jing, Oguz Akin, and Yingli Tian. 2021. Medical Image Tampering Detection: A New Dataset and Baseline. In *International Conference on Pattern Recognition*. Springer, 266–277.

[8] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9543–9552.




(a) (b)

**Figure 8: Copy-Move. Groundtruth: red boxes. Predicted: cyan boxes.**