electronics

Article

Towards Adversarial Attacks for Clinical Document Classification

Nina Fatehi [1], Qutaiba Alasad [2] and Mohammed Alawad [1,*]

[1] Electrical and Computer Engineering Department, Wayne State University, Detroit, MI 48202, USA
[2] Department of Petroleum Processing Engineering, Tikrit University, Al Qadisiyah P.O. Box 42, Iraq
* Correspondence: alawad@wayne.edu

Abstract: Regardless of revolutionizing improvements in various domains thanks to recent advancements in the field of Deep Learning (DL), recent studies have demonstrated that DL networks are susceptible to adversarial attacks. Such attacks are crucial in sensitive environments to make critical and life-changing decisions, such as health decision-making. Research efforts on using textual adversaries to attack DL for natural language processing (NLP) have received increasing attention in recent years. Among the available textual adversarial studies, Electronic Health Records (EHR) have gained the least attention. This paper investigates the effectiveness of adversarial attacks on clinical document classification and proposes a defense mechanism to develop a robust convolutional neural network (CNN) model and counteract these attacks. Specifically, we apply various black-box attacks based on concatenation and editing adversaries on unstructured clinical text. Then, we propose a defense technique based on feature selection and filtering to improve the robustness of the models. Experimental results show that a small perturbation to the unstructured text in clinical documents causes a significant drop in performance. Performing the proposed defense mechanism under the same adversarial attacks, on the other hand, avoids such a drop in performance. Therefore, it enhances the robustness of the CNN model for clinical document classification.

Keywords: adversarial attacks; document classification; CNN; NLP

Citation: Fatehi, N.; Alasad, Q.; Alawad, M. Towards Adversarial Attacks for Clinical Document Classification. *Electronics* **2023**, *12*, 129. https://doi.org/10.3390/electronics12010129

Academic Editors: Taiyong Li, Wu Deng and Jiang Wu

Received: 15 November 2022
Revised: 21 December 2022
Accepted: 22 December 2022
Published: 28 December 2022

1. Introduction

Although DL models for NLP have achieved remarkable success in various domains, such as text classification [1], sentiment analysis [2] and Named Entity Recognition (NER) [3], recent studies have demonstrated that DL models are susceptible to adversarial attacks, small perturbations and named adversarial examples (AEs), crafted to fool the DL model to make false predictions [4]. Such attacks are crucial in sensitive environments like healthcare where such vulnerabilities can directly threaten human life. Similar to other domains, DL in healthcare has obtained diagnostic parity with human physicians on various health information tasks such as pathology [5] and radiology [6]. The issue of AEs has emerged as a pervasive challenge in even state-of-the-art learning systems for health and has raised concerns about the practical deployment of DL models in such a domain. However, in comparison to non-clinical NLP tasks, adversarial attacks on Electronic Health Records (EHR) and tasks such as clinical document classification have gained the least attention.

Various approaches based on concatenation [7] or editing [8] perturbations have been proposed to attack NLP models. Attacking these models by manipulating characters in a word to generate AEs seems unnatural for some applications due to grammatical disfluency. Also, generating AEs is challenging in the text compared to images, due to the discrete space of input data as well as the fact that generating perturbations which can fool the DL model and at the same time be unperceivable for humans is not easy in text [4]. However, these approaches apply very well to the target application of this paper, i.e., pathology

report classification based on the cancer type. The unstructured text in pathology reports is ungrammatical, fragmented, and marred with typos and abbreviations. Also, the document text is usually long and results from the concatenation of several fields, such as microscopic description, diagnosis, and summary. Whenever they are combined, the human cannot easily differentiate between the beginning and end of each field. Moreover, the text in pathology reports exhibits linguistic variability across pathologists even when describing the same cancer characteristics [9,10].

Perturbations can occur at all stages of the DL pipeline from data collection to model training to the post-processing stage. In this paper, we will focus on two aspects of it. The first part will be the robustness during the training time. The case when the training set is unvetted, e.g., when the training set has arbitrarily chosen an outlier that the model is biased towards. The second aspect is the robustness during the test time. The case when the adversary is trying to fool the model.

The AEs that we will use in this paper are the class label names. These words are known to the attacker without accessing the target DL model. Also, due to the unbalanced nature of this dataset, the model is biased to majority classes or to specific keywords, which are mostly the class label names, that appear in their corresponding samples. Then, we propose a novel defense method against adversarial attacks. Specifically, we select and filter specific features during the training phase. Two criteria are followed when determining these features: (1) the DL model has to be biased to them, and (2) filtering them does not impact the overall model accuracy. We focus on the CNN model to carry out the adversarial evaluation on the clinical document classification, i.e., classifying cancer pathology reports based on their associated cancer type. This model performs equally or better than state-of-the-art natural language models, i.e., BERT [11]. This is mainly because in clinical text classification tasks on documents in which only very few words contribute toward a specific label, most of these subtle word relationships may not be necessary or even relevant to the task at hand [12].

The main contributions of the paper include:

- We compare the effectiveness of different black-box adversarial attacks on the robustness of the CNN model for document classification on long clinical texts.
- We evaluate the effectiveness of using class label names as AEs by either concatenating these examples to the unstructured text or editing them whenever they appear in the text.
- We propose a novel defense technique based on feature selection and filtering to enhance the robustness of the CNN model.
- We evaluate the robustness of the proposed approach on clinical document classification.

The rest of the paper is organized as follows: related works are briefly outlined in Section 2. Sections 3 and 4 present the method and experimental setup, respectively. In Section 5, the results are discussed. Finally, we conclude our paper in Section 6.

2. Related Works

Numerous methods have been proposed in the area of computer vision and NLP for adversarial attacks [4,13,14]. Since our case study focuses on adversarial attacks and defense for clinical document classification, we mainly review state-of-the-art approaches in the NLP domain. Zhang et al. present a comprehensive survey of the latest progress and existing adversarial attacks in various NLP tasks and textual DL models [4]. They categorize adversarial attacks on textual DL as follows:

- Model knowledge determines if the adversary has access to the model information (white-box attack) or if the model is unknown and inaccessible to the adversary (black-box attack).
- Target type determines the aim of the adversary. If the attack can alter the output prediction to a specific class, it is called a targeted attack, whereas an untargeted attack tries to fool the DL model into making any incorrect prediction.

- Semantic granularity refers to the level to which the perturbations are applied. In other words, AEs are generated by perturbing sentences (sentence-level), words (word-level) or characters (character-level).

The work investigated in this paper relates to the adversarial attack on document classification tasks in the healthcare domain and focuses on the targeted/untargeted black-box attack using word/character-level perturbations. We choose black-box attacks as they are more natural than white-box attacks.

In the following subsections, we first present the popular attack strategies with respect to the three above-mentioned categories. Then, we discuss the adversarial defense techniques.

2.1. Adversarial Attack Strategies

Adversarial attacks have been widely investigated for various NLP tasks, including NER [15,16], semantic textual similarity [16], and text classification [17]. These attacks generate perturbations by modifying characters within a word [18], adding or removing words [16], replacing words with semantically similar and grammatically correct synonyms using a word embedding optimized for synonyms replacement [17], or by synonym substitution using WordNet where the replaced word has the same Part of Speech (POS) as the original one [19]. The drawback of character-level methods is that generated AEs can easily be perceived by human and impact the readability of the text [20], and the drawback of AEs generated in word-level is their dependency on the original sentences [20].

Clinical text comes with its unique challenges, where the findings in the non-clinical text might not be applied to clinical text. For instance, in non-clinical text, character-level or word-level perturbations that change the syntax can be easily detected and defended against by the spelling or syntax check. However, this does not apply to clinical text, which often contains incomplete sentences, typographical errors, and inconsistent formatting. Thus, domain-specific strategies for adversarial attacks and defense are required [21].

There are relatively few works that have examined the area of clinical NLP. Mondal et al. propose BBAEG, which is a black-box attack on biomedical text classification tasks using both character-level and word-level perturbations [22]. BBAEG is benchmarked on a simple binary classification and the text is relatively short and clean when compared to real-world clinical text. There are also some works that investigate adversarial attack on EHR including [23,24]. However, they are different from this paper's work as they use the temporal property of EHR to generate AEs and none of them investigates the adversarial attack on unstructured clinical text.

2.2. Adversarial Defense Strategies

As explained in the previous section, detection-based approaches, such as spelling check, have been used as a defense strategy. Gao et al. use python's spelling check to detect the adversarial perturbations in character level; however, this detection method can be performed only on character-level AEs [18]. Another approach in detection to evaluate the model's robustness under adversarial attacks is discriminator training. Xu et al. train another discriminator using a portion of original samples plus AEs to discriminate AEs from original examples [25]. Adversarial training has also been used to enhance the robustness of DL model [4,26]. In adversarial training, adversarial perturbations are involved in the training process [27]. The authors of [15–17] utilize an augmentation strategy to evaluate or enhance the robustness of DL models. In this approach, the model is trained on the augmented dataset that includes original samples plus AEs. The drawback of adversarial training, which makes it an ineffective defense strategy against some adversarial attacks, is overfitting [27]. If the model is biased to the AEs, as in the case of our paper, augmentation will make the bias issue worse.

3. Method

In this section, we first formalize the adversarial attack in a textual CNN context and then describe two methods, namely concatenation adversaries and edit adversaries to generate AEs.

3.1. Problem Formulation

Let us assume that a dataset consists of N documents $X = \{X_1, X_2, ..., X_N\}$ and a corresponding set of N labels $y = \{Y_1, Y_2, ..., Y_N\}$. On such a dataset, $F : X \rightarrow y$ is the CNN model which maps input space X to the output space y. Adversarial attack and adversarial example can be formalized as follows:

$$X_{adv} = X + \Delta$$

$$F(X_{adv}) \neq y \quad (Untargeted \quad attack)$$

$$F(X_{adv}) = y\prime, \quad y \neq y\prime \quad (Targeted \quad attack)$$

3.2. Concatenation Adversaries

Given an input document X_1 of n words $X_1 = \{w_1, w_2, ..., w_n\}$, in concatenation adversaries, there is a list of selected perturbation words that are supposed to be added (one at a time) to different locations of documents. In this paper, we consider three locations: "random", "end", and "beginning".

- Adding perturbation words at random locations: In this attack, we attempt to add a various number of a specific perturbation word in random locations of input documents. If w_{adv} denotes the added word, the adversarial input would be as $X = \{w_1, w_2, w_{adv}, \ldots, w_{adv}, w_{n-4}, w_{adv}, w_n\}$. The location of each w_{adv} is determined randomly.
- Adding perturbation words in the beginning: In this attack, the aim is to append a various number of a specific perturbation word in the beginning of each input document. In this way, the adversarial input would be as $X = \{w_{adv}, w_{adv}, w_{adv}, \ldots, w_1, w_2, \ldots, w_n\}$.
- Adding perturbation words at the end: This attack is carried out to add a various number of a specific perturbation word at the end of each document. The adversarial inputs would be as $X = \{w_1, w_2, \ldots, w_n, w_{adv}, w_{adv}, w_{adv}\}$.

3.3. Edit Adversaries

Instead of adding perturbation words to the input document text, edit adversaries manipulate specific words in the input document text. In this paper, we apply two edit adversaries forms: synthetic perturbation, which is an untargeted attack; and replacing strategy, which in contrast is a targeted attack [28].

- Synthetic perturbation: In this attack, AEs are generated by perturbing characters including swapping all two neighboring characters ($breast \rightarrow rbaets$), randomly deleting one character ($breast \rightarrow breat$), randomly changing orders of all characters ($breast \rightarrow reastb$) and randomly changing orders of characters except the first and last ones ($breast \rightarrow beasrt$) of a specific list words. Each of these perturbations are performed on all selected words at the same time.
- Replacing strategy: This attack is a targeted attack in which the selected words in all input documents are replaced with a specific word that leads to the targeted prediction (for instance $F(X_{tadv}) = Y_d$ and w_{adv} is the perturbation word that makes the prediction Y_d instead of Y_t).

3.4. Defense Strategy

For the defense mechanism, we propose a novel method called feature selection and filtering, in which features are selected and filtered from input documents during the model training. These features are selected based on two criteria: (1) the CNN model has to be

biased to them, and (2) filtering them does not impact the overall model accuracy. In this paper, we select the class label names as the target features. Other techniques can also be used to determine which features should be selected, such as model interpretability tools, attention weights, scoring functions, etc.

3.5. Evaluation Metrics

The focus of this study is to evaluate the performance of the CNN model for document classification against adversarial examples. The following common performance metrics for classification tasks are used for model evaluation:

F1 Score: The overall accuracy is calculated using the standard micro- and macro- F1 scores as follows:

$$Micro\ F1 = 2(\frac{Precision * Recall}{Precision + Recall})$$

$$Macro\ F1 = \frac{1}{|c|}\Sigma_{c_i}^{c} Micro\ F1(c_i)$$

where $|C|$ is the total number of classes and c_i represents the number of samples belonging to class i.

Accuracy per class: To evaluate the vulnerability of the model per class, we use the accuracy per class metric, which is the percentage of correctly predicted classes after an attack to the number of all samples of the class.

$$Accuracy = \frac{TP_i}{c_i}$$

Number of Perturbed Words: For the attack itself, we include a metric to measure the amount of required perturbations to fool the CNN model. We call this metric "number of perturbed words". In this way, we can determine the minimum number of perturbation words, in concatenation adversaries, that leads to a significant degradation in accuracy.

4. Experimental Setup

4.1. Data

In this paper, we benchmark the proposed adversarial attack and defense on a clinical dataset, specifically The Cancer Genome Atlas Program pathology reports dataset (TCGA) (https://www.cancer.gov/tcga, accessed on 1 October 2021).

The original TCGA dataset consists of 6365 cancer pathology reports; five of which are excluded because they are unlabeled. Therefore, the final dataset consists of 6360 documents. Each document is assigned a ground truth label for the site of the cancer, the body organ where the cancer is detected. In the TCGA dataset, there is a total of 25 classes for the site label. Figure A1 in Appendix A shows the histograms of the number of occurrences per class. Standard text cleaning, such as lowercasing and tokenization, is applied to the unstructured text in the documents. Then, a word vector of size 300 is chosen. The maximum length of 1500 is chosen to limit the length of documents in pathology reports. In this way, reports containing more than 1500 tokens are truncated and those with less than 1500 tokens are zero-padded. Also, we choose 80%/20% data splitting strategy.

4.2. Target Model

In this paper, we use a CNN network as the DL model. ADAM adaptive optimization is used to train the network weights. For all the experiments, the embedding layer is followed by three parallel 1-D convolutional layers. The number of filters in each convolution layer is 100, and the kernel sizes are 3, 4, and 5. ReLU is employed as the activation function and a dropout of 50% is applied to the global max pooling at the output layer. Finally, a fully connected softmax layer is used for the classification task. These parameters are optimized following previous studies [29,30]. We use NVIDIA V100 GPU for all the experiments.

4.3. Adversarial Attack

In this subsection we go through the details of each adversarial attack. For these attacks, we use the dataset class label names, which are the cancer types, as the selected perturbation words to perform concatenation and edit adversaries. The reasons for selecting the label names as the AEs are as follows:

1. Using these names is considered a black-box attack as the adversary does not need to know the CNN model details.
2. The presence of a specific class label name frequently in the unstructured text of pathology reports biases the CNN model to a specific prediction [29,30].
3. As we will see later, filtering these names during the model training does not impact the overall model accuracy.

Therefore, we note that the practical dataset is the one whose class label names exist in the document text.

Since there are 25 different labels' class names in the dataset, we select three of them as AEs to report the result in this paper. Specifically, we select one of the majority classes (breast), one of the minority classes (leukemia/lymphoma- in short lymphoma) and one of the moderate classes (sarcoma). From that selection, we can see how the classes with different distributions can impact the performance of the CNN model under adversarial attacks. In this way, the impact of class distribution on the CNN model's performance can be evaluated as well.

4.3.1. Concatenation Adversaries

In this attack, we investigate the impact of adding selected class names (breast, leukemia/lymphoma, and sarcoma) to the input documents as perturbation words. Three different concatenation adversaries are used:

- Concat-Random: For each selected perturbation word, 1, 2, 3, 5, 10, 20 or 40 words are randomly added to all input documents. For instance, Concat-Random-Breast-1 means randomly adding one "breast" perturbation word to the documents.
- Concat-Begin: For each selected perturbation word, 1, 2, 3, 5, 10, 20 or 40 words are added at the beginning of all input documents. For instance, Concat-Begin-Breast-1 denotes appending one "breast" perturbation word at the beginning of all documents.
- Concat-End: For each selected perturbation word, 1, 2, 3, 5, 10, 20 or 40 words are appended at the end of all input documents. For instance, Concat-End-Breast-1 means adding one "breast" perturbation word at the end of the documents.

4.3.2. Edit Adversaries

We apply the following edit adversaries to the text dataset:

- Edit-Synthetic: In this attack, we perturb the letters of all the selected words (breast, leukemia/lymphoma, and sarcoma) whenever they appear in the document text. Different approaches are applied to edit the targeted tokens, such as swapping all two neighboring characters (Swap), randomly deleting one character (Delete), randomly changing orders of all character (Fully Random), or randomly changing orders of characters except the first and last ones (Middle Random).
- Edit-Replacing: In this attack, all class label names (the 25 different labels) are replaced with one of the target words (breast, leukemia/lymphoma, or sarcoma) whenever they appear in the unstructured text. For instance, Edit-Replacing-Breast means all class label names that appear in the input document text are replaced with the word "breast" as the perturbed word.

4.4. Defense

In defense, all class label names are filtered from the input documents during the new model training. Then, we attack the model using the same AEs as before to investigate

the word-level and character-level adversarial training impacts on enhancing the CNN model's robustness.

5. Results

In this section, we present the results related to each experiment.

5.1. Concatenation Adversaries

Figure 1 illustrates the impact of increasing the number of perturbed words on the overall accuracy. We can see, as expected, that the drop in accuracy increases when adding more perturbation words to the document text.

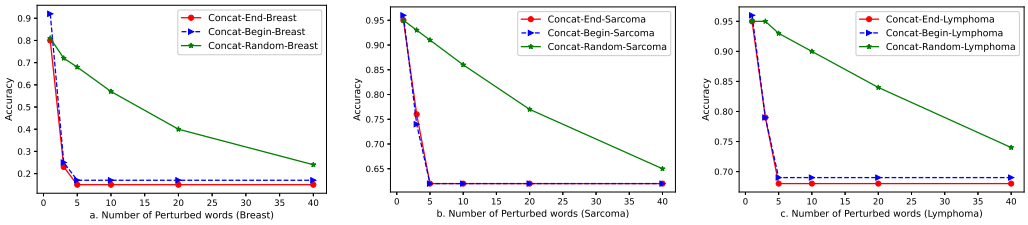

Figure 1. Impact of increasing number of words in Concatenation adversaries; for (**a**). breast, (**b**). sarcoma and (**c**). lymphoma.

The figure also shows that Concat-Random's accuracy degrades slowly with an increasing number of perturbation words; however, in Concat-Begin and Concat-End, there is a sharp drop in accuracy by adding only 3 perturbation words and this decrease continues until adding 5 words. Adding more than 5 words does not change the accuracy. This indicates that if the perturbation words are adjacent in the input text, they have higher impact on the model predictions.

Another observation is the different impact of the selected perturbed words (breast, sarcoma and lymphoma) on the overall model accuracy. From the accuracy values for each class, we see that accuracy drop in breast as a majority class is significant, as adding 3 words causes accuracy to become less than 30%. However, in lymphoma and sarcoma as minority and moderate classes, accuracy drops to 79% and 74%, respectively.

In Table 1, a comparison between different concatenation adversaries is provided. In this table, we consider 3 perturbed words. Compared with the baseline model, we can see that adding only 3 words can reduce the accuracy significantly, which is an indication of the effectiveness of the attack. From the results of Table 1, we came to conclude that in an imbalanced dataset and under an adversarial attack, majority classes contribute at least 3 times more than the minority classes. This conclusion is drawn from the fact that the CNN model is biased towards the majority classes in an imbalanced dataset; therefore, minority classes contribute less to the overall accuracy than majority classes.

Table 1. Comparison between different concatenation adversaries attack strategies.

Model	Micro F1			Macro F1		
	Beginning	**End**	**Random**	**Beginning**	**End**	**Random**
Baseline		0.9623			0.9400	
Concat-Breast-3	0.2461	0.2335	0.7193	0.2337	0.2029	0.7501
Concat-Sarcoma-3	0.7429	0.7594	0.9261	0.6666	0.6818	0.8794
Concat-lymphoma-3	0.7862	0.7932	0.9465	0.7262	0.7367	0.9028

To gain more insight on the impact of concatenation adversaries, we investigate the accuracy per class. Figure 2 illustrates the accuracy of each class when the perturbed word is "breast" for Concat-End attack. The figures for the other two perturbation words and the

other concatenation strategies are included in Appendix B. The interesting observation is that adding the perturbed word contributes in an accuracy drop of all classes except the "breast" class. In other words, the adversarial attack was able to fool the CNN model to a target attack and give 100% accuracy for the perturbed class word.

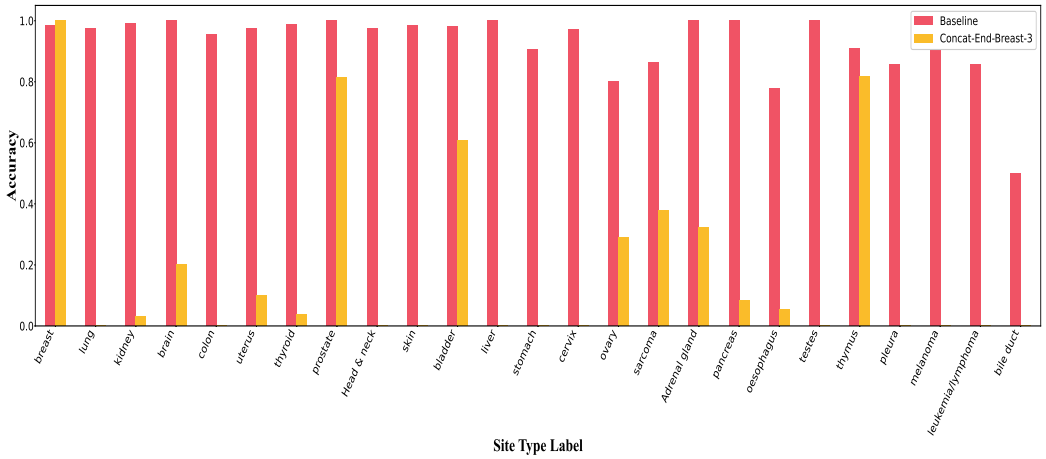

Figure 2. Accuracy per class in Concat-End for breast.

With further analysis, we also realize that adding the perturbed word causes an increase in number of false predictions such that the CNN model is most likely to classify the documents of other classes as the class equal to the perturbed word. Table 2 shows the number of documents classified as the perturbed word after an adversarial attack.

While analysing the two-term word class names, such as "leukemia/lymphoma", "bile duct" and "head and neck", we noticed that such classes seem to have one term neutral which does not cause any changes in the accuracy; however, the other term follows almost the same pattern as the other single-term word class names in the dataset. To find the reason, we looked into the input dataset to see the occurrence of each word in the whole dataset (Table A1 in Appendix B). We found that the term that occurred more often is likely to impact the performance more under adversarial attacks.

Table 2. Number of documents classified as the perturbed word before and after adversarial attack.

	Number of Documents
Baseline-breast	134 out of 1272
Concat-Random-Breast-1	359
Concat-Random-Breast-10	671
Concat-Random-Breast-20	878
Baseline-sarcoma	31 out of 1272
Concat-Random-sarcoma-1	61
Concat-Random-sarcoma-10	196
Concat-Random-sarcoma-20	312
Baseline-lymphoma	6 out of 1272
Concat-Random-lymphoma-1	22
Concat-Random-sarcoma-10	90
Concat-Random-lymphoma-20	179

5.2. Edit Adversaries

Table 3 depicts the comparison of accuracy on different edit adversaries attacks. As we can see from the results, compared to the baseline model, all edit adversaries attack strategies degrade the accuracy. We also see that all character-level perturbations cause the same

amount of drop in accuracy (4% in micro F1 and 6% in macro F1). The reason is that, only class names have been targeted in this set of experiments and no matter how they are edited, the CNN model interprets them all as unknown words; therefore, they all contribute in the same amount of accuracy drop. This also confirms that there are keywords other than the class names that are critical to the class prediction. On the contrary, Edit-Replacing strategies result in a significant decrease in accuracy (12% in micro F1 and 17% in macro F1) and (58% in micro F1 and 44% in macro F1) when all 25 class names in the text are replaced with "lymphoma" and "breast" perturbation words, respectively. It shows that although the CNN model is biased towards all class names, majority classes seem to have a more significant impact than the minority. Figure 3 shows accuracy per class under Edit-Synthetic adversarial attack. From the figure, we see that minority classes are impacted more than majority classes. Figures of accuracy per class in Edit-Replacing attacks for breast, sarcoma and lymphoma are included in Appendix B.

Table 3. Comparison between different edit adversaries attack strategy.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Swap	0.9230	0.8815
Delete	0.9230	0.8815
Fully Random	0.9230	0.8815
Middle Random	0.9230	0.8815
Edit-Replacing-Breast	0.3774	0.4209
Edit-Replacing-Sarcoma	0.7987	0.7366
Edit-Replacing-Lymphoma	0.8373	0.7648

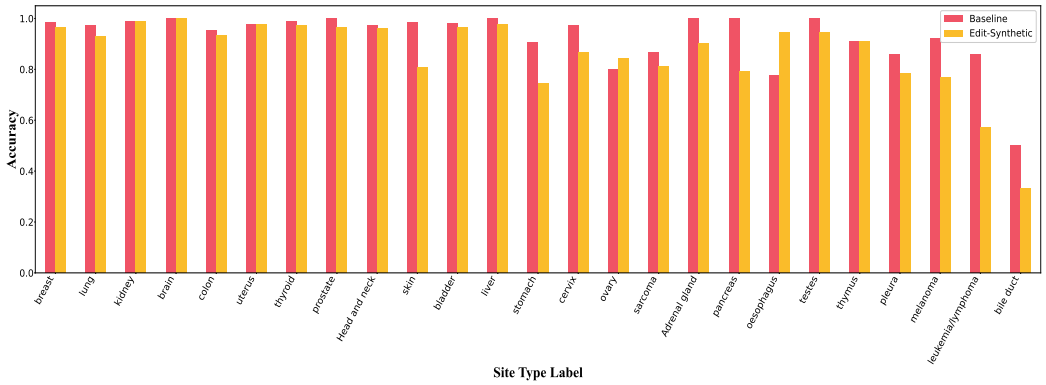

Figure 3. Accuracy per class in Edit-Synthetic.

5.3. Defense

Tables 4 and 5 demonstrate the performance results of the CNN model after filtering the class names from the text during the training, as well as the model performance under adversarial attacks using the concatenation and edit adversaries. From the result, we can easily see that the defense strategy was able to successfully defend against adversarial attacks with little to no degradation of the performance of the baseline CNN model under the same adversarial attack. From the macro-F1 score, we see that after performing the defense strategy, the accuracy of minority classes increases while the accuracy of majority classes remains unchanged; so, we came to conclude that the defense strategy is able to enhance the CNN model's robustness not only by immunizing the model against adversarial attack but also by tackling the class imbalance problem as well.

Table 4. Comparison between different concatenation adversaries attack strategies while defense strategy is imposed.

Model	Micro F1			Macro F1		
	Beginning	**End**	**Random**	**Beginning**	**End**	**Random**
Baseline		0.9544			0.9240	
Concat-Breast-3	0.9544	0.9544	0.9544	0.9240	0.9240	0.9243
Concat-Sarcoma-3	0.9544	0.9544	0.9544	0.9240	0.9243	0.9243
Concat-lymphoma-3	0.9544	0.9544	0.9544	0.9240	0.9240	0.9243

Table 5. Overall micro/macro F1 by performing defense.

	Micro F1	**Macro F1**
Baseline	0.9544	0.9240
Swap	0.9583	0.9369
Delete	0.9583	0.9369
Fully Random	0.9583	0.9369
Middle Random	0.9583	0.9369
Edit-Replacing-Breast	0.9583	0.9369
Edit-Replacing-Sarcoma	0.9583	0.9369
Edit-Replacing-Lymphoma	0.9583	0.9369

6. Conclusions

In this paper, we investigate the problem of adversarial attacks on unstructured clinical datasets. Our work demonstrates the vulnerability of the CNN model in clinical document classification tasks, specifically cancer pathology reports. We apply various black-box attacks based on concatenation and edit adversaries; then, using the proposed defense technique, we are able to enhance the robustness of the CNN model under adversarial attacks. Experimental results show that adding a few perturbation words as AEs to the input data will drastically decrease the model accuracy. We also indicate that by filtering the class names in the input data, the CNN model will be robust to such adversarial attacks. Furthermore, this defense technique is able to mitigate the bias of the CNN model towards the majority classes in the imbalanced clinical dataset.

Author Contributions: Conceptualization, M.A. and Q.A.; methodology, M.A. and N.F.; software, M.A. and N.F.; validation, M.A., N.F. and Q.A.; formal analysis, M.A.; investigation, M.A. and N.F.; resources, M.A.; writing, review, and editing, M.A., N.F., Q.A.; visualization, M.A. and N.F.; supervision, M.A.; project administration, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga, accessed on 1 October 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. TCGA Dataset

Figure A1 shows the histograms of the number of occurrences per class for the cancer site.

Figure A1. Classes Distribution in TCGA Dataset for Site.

Appendix B. Adversarial Attack

Table A1 shows the frequency of each term of two-term Label's classes word in the whole dataset.

Table A1. Two-term word Labels' occurrence in whole dataset.

	Occurrence
duct	1542
bile	1012
gland	2589
adrenal	1786
lymphoma	90
leukemia	3
neck	2817
head	356

Appendix B.1. Concatenation Adversaries

The overall micro- and macro- F1 scores for various number of perturbed words in Concat-End-Breast, Concat-End-sarcoma and Concat-End-lymphoma adversarial attacks are depicted in Tables A2–A4.

Table A2. Overall micro/macro F1 in Concat-End-Breast adversarial attack for various number of perturbed word.

	Micro F1	Macro F1
Baseline	0.9623	0.9400
Concat-End-Breast-1	0.8003	0.8156
Concat-End-Breast-3	0.2335	0.2029
Concat-End-Breast-5	0.1486	0.0915
Concat-End-Breast-20	0.1486	0.0915

Table A3. Overall micro/macro F1 in Concat-End-sarcoma adversarial attack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-End-sarcoma-1	0.9520	0.9172
Concat-End-sarcoma-3	0.7594	0.6818
Concat-End-sarcoma-5	0.6156	0.5506
Concat-End-sarcoma-20	0.6156	0.5506

Table A4. Overall micro/macro F1 in Concat-End-lymphoma adversarial attack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-End-lymphoma-1	0.9520	0.9091
Concat-End-lymphoma-3	0.7932	0.7367
Concat-End-lymphoma-5	0.6824	0.6203
Concat-End-lymphoma-20	0.6824	0.6203

The overall micro- and macro- F1 scores for various number of perturbed words in Concat-Begin-Breast, Concat-Begin-sarcoma and Concat-Begin-lymphoma adversarial attacks are depicted in Tables A5–A7.

Table A5. Overall micro/macro F1 in Concat-Begin-Breast adversarial attack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-Begin-Breast-1	0.9198	0.9157
Concat-Begin-Breast-3	0.2461	0.2337
Concat-Begin-Breast-5	0.1682	0.1332
Concat-Begin-Breast-20	0.1682	0.1332

Table A6. Overall micro/macro F1 in Concat-Begin-sarcoma adversarial attack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-Begin-sarcoma-1	0.9615	0.9157
Concat-Begin-sarcoma-3	0.7429	0.6666
Concat-Begin-sarcoma-5	0.6211	0.5684
Concat-Begin-sarcoma-20	0.6211	0.5684

Table A7. Overall micro/macro F1 in Concat-Begin-lymphoma adversarial attack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-Begin-lymphoma-1	0.9638	0.9289
Concat-Begin-lymphoma-3	0.7862	0.7262
Concat-Begin-lymphoma-5	0.6863	0.6209
Concat-Begin-lymphoma-20	0.6863	0.6209

The overall micro- and macro- F1 scores for various number of perturbed words in Concat-Random-Breast, Concat-Random-lymphoma and Concat-Random-sarcoma adversarial attacks are depicted in Tables A8–A10.

Table A8. Overall micro/macro F1 in Concat-Random-Breast adversarial atttack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-Random-Breast-1	0.8066	0.8240
Concat-Random-Breast-10	0.5660	0.6006
Concat-Random-Breast-20	0.4049	0.3992

Table A9. Overal micro/macro F1 in Concat-Random-lymphoma adversarial atttack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-Random-lymphoma-1	0. 9520	0.9105
Concat-Random-lymphoma-10	0.9033	0.8567
Concat-Random-lymphoma-20	0.8381	0.7924

Table A10. Overall micro/macro F1 in Concat-Random-sarcoma adversarial attack for various number of perturbed word.

	Micro F1	**Macro F1**
Baseline	0.9623	0.9400
Concat-Random-sarcoma-1	0. 4049	0.3992
Concat-Random-sarcoma-10	0.8585	0.8051
Concat-Random-sarcoma-20	0.7720	0.7148

Figures A2–A9 illustrates the accuracy per class for each perturbed word (breast, sarcoma and lymphoma) in concatenation adversaries.

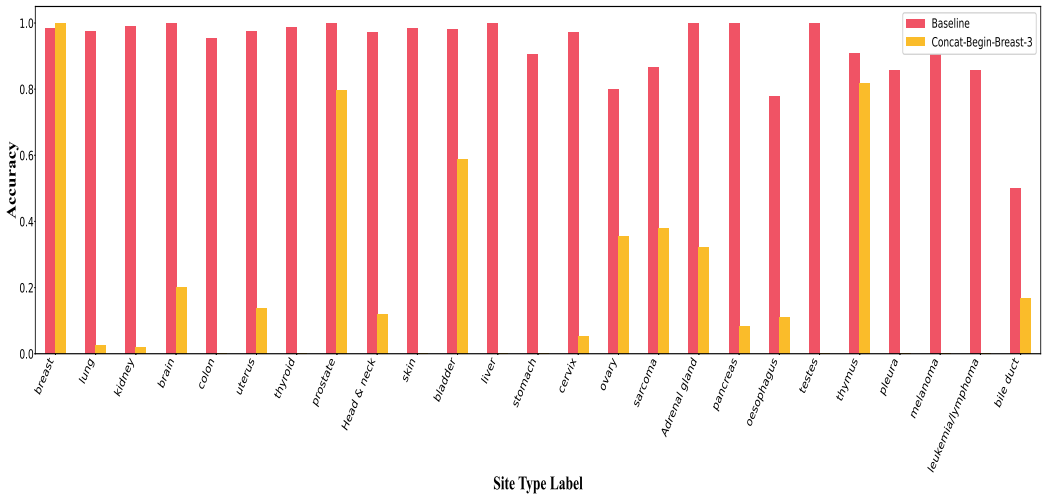

Figure A2. Accuracy per class in Concat-Begin for breast.

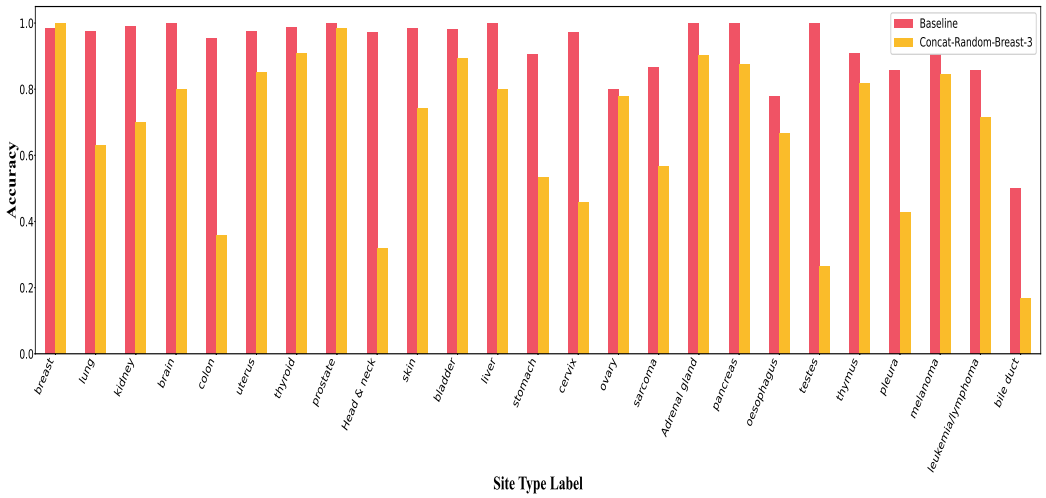

Figure A3. Accuracy per class in Concat-Random for breast.

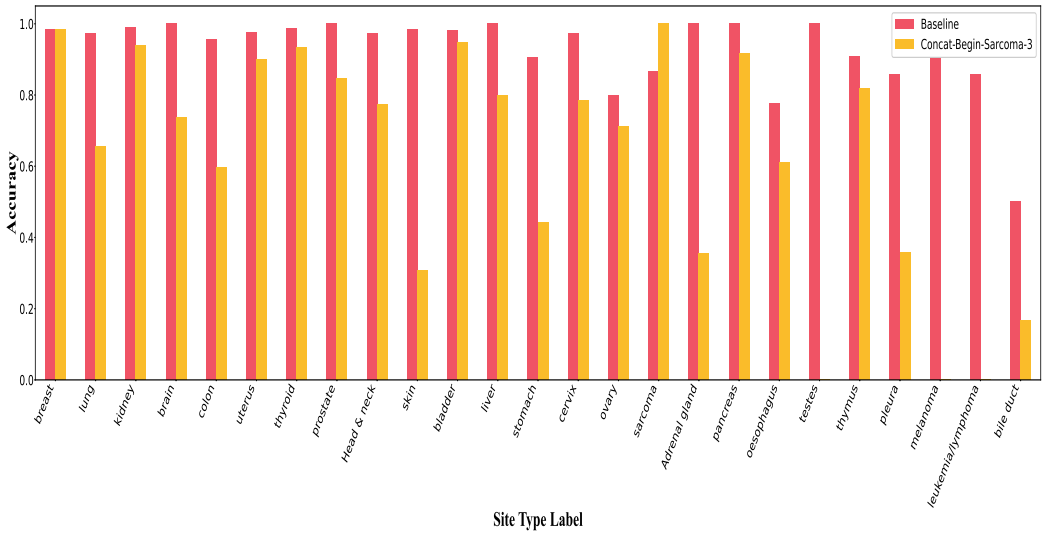

Figure A4. Accuracy per class in Concat-Begin for sarcoma.

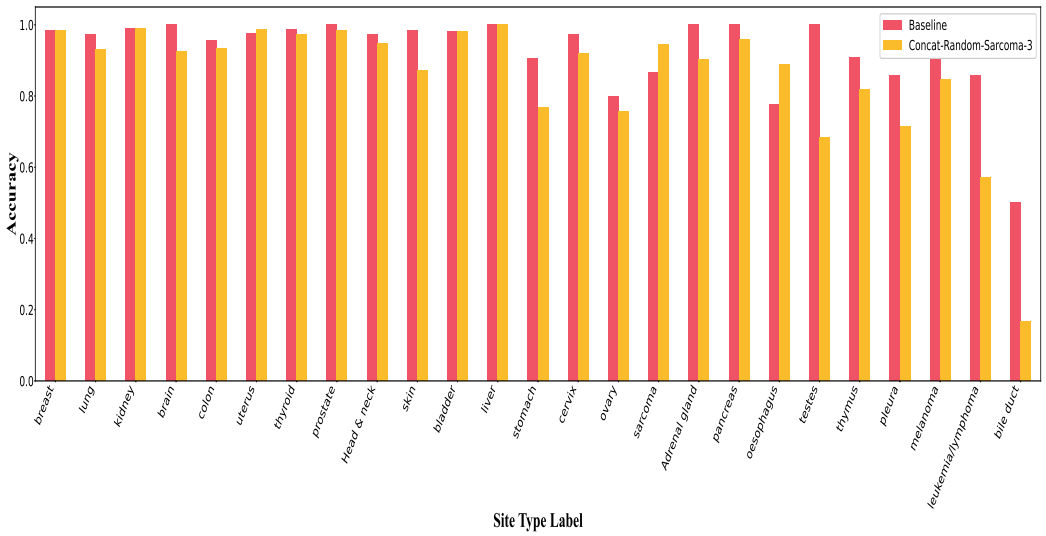

Figure A5. Accuracy per class in Concat-Random for sarcoma.

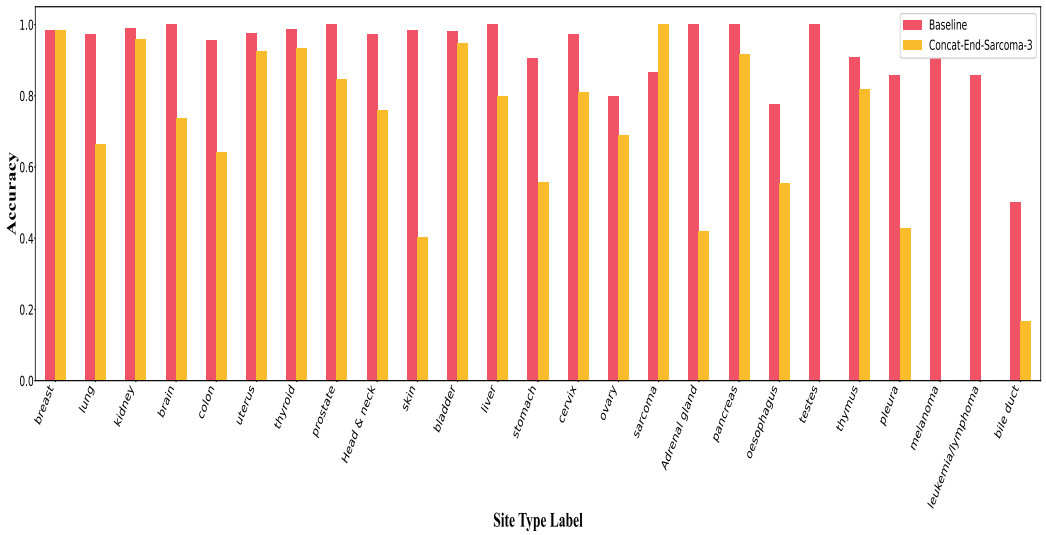

Figure A6. Accuracy per class in Concat-End for sarcoma.

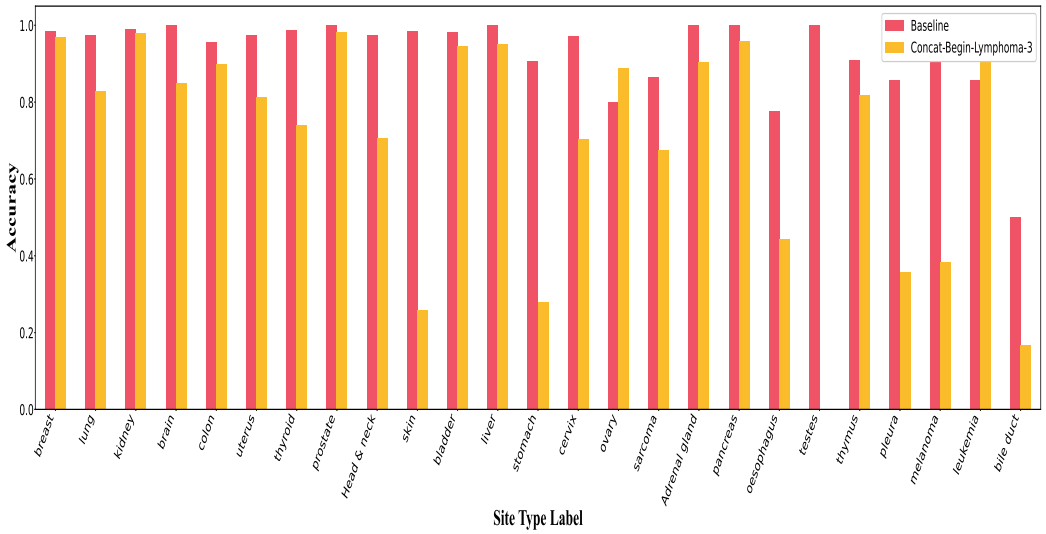

Figure A7. Accuracy per class in Concat-Begin for lymphoma.

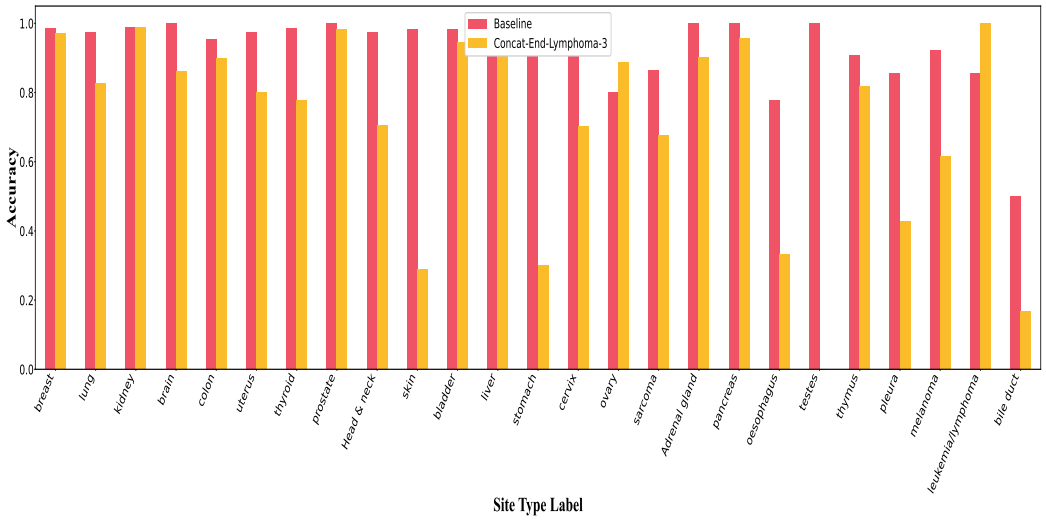

Figure A8. Accuracy per class in Concat-End for lymphoma.

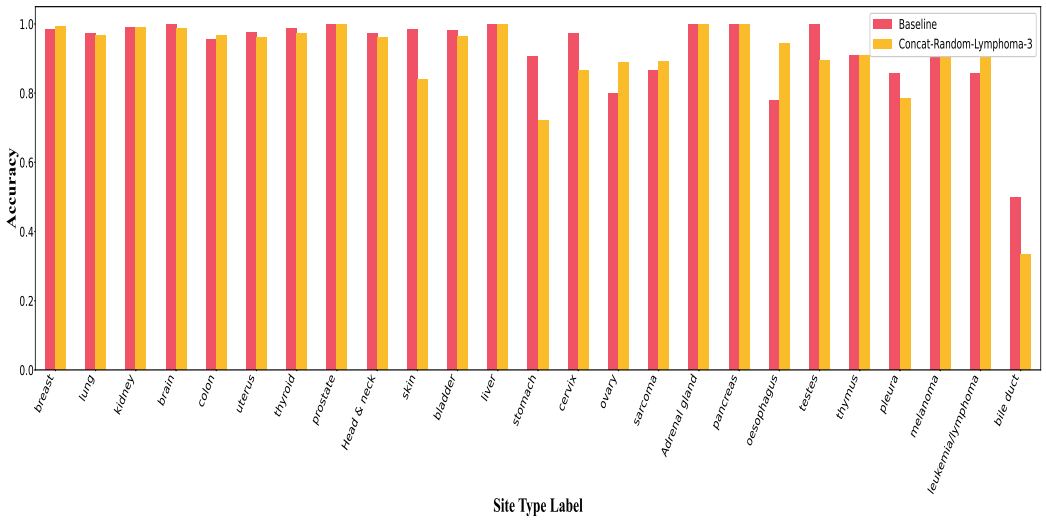

Figure A9. Accuracy per class in Concat-Random for lymphoma.

Appendix B.2. Edit Adversaries

Figures A10–A12 show accuracy per class in Edit-Replacing-Breast, Edit-Replacing-Sarcoma and Edit-Replacing-Lymphoma attacks, respectively.

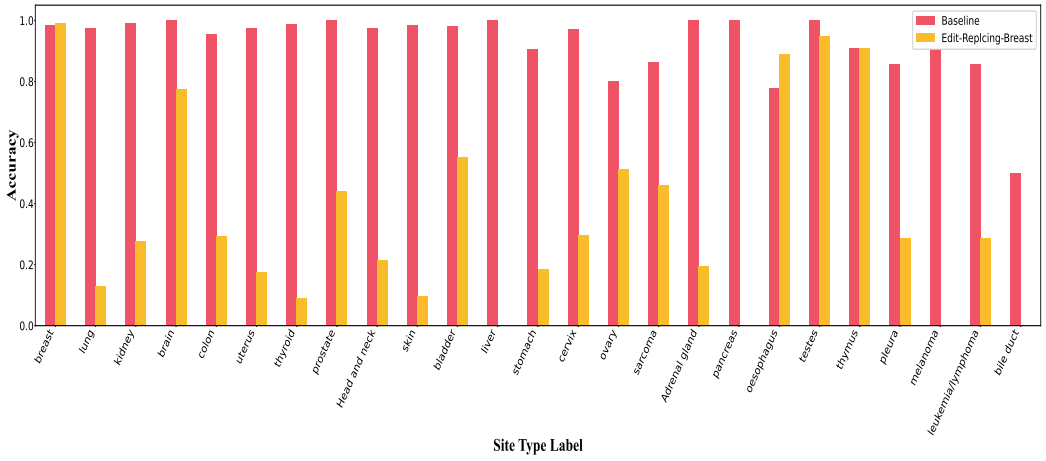

Figure A10. Accuracy per class in Edit-Replacing-breast.

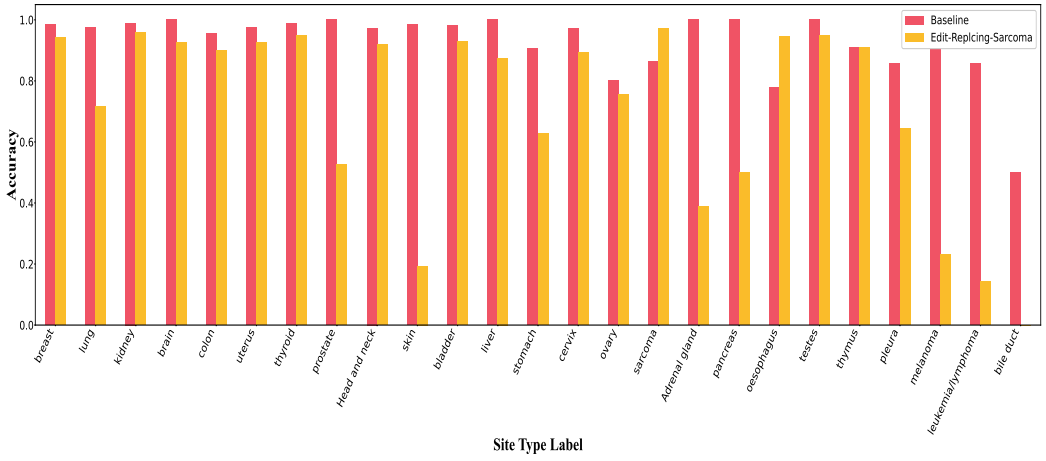

Figure A11. Accuracy per class in Edit-Replacing-Sarcoma.

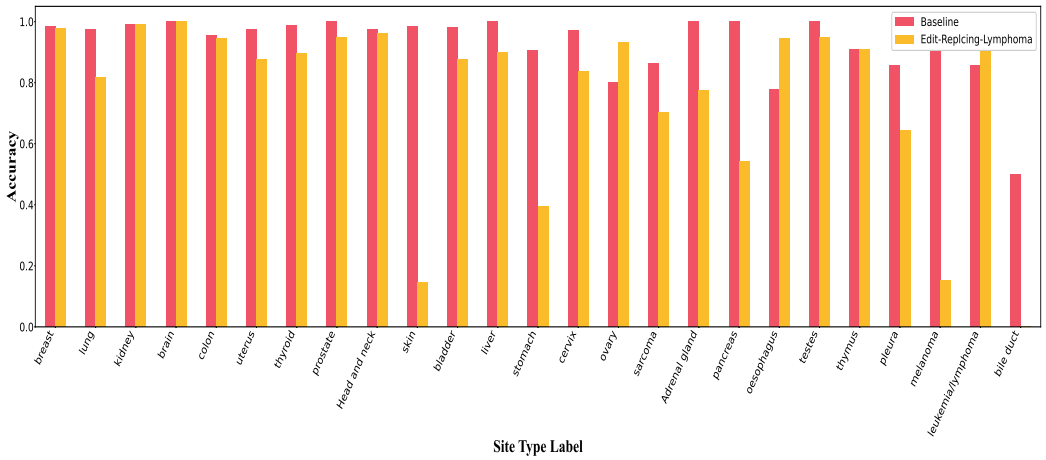

Figure A12. Accuracy per class in Edit-Replacing-lymphoma.

Appendix C. Defense

In this section, we provide figures and tables that are related to the defense under different adversarial attacks. Figures A13–A15 illustrate accuracy per class under concatenation and edit adversaries attacks when defense strategy is applied.

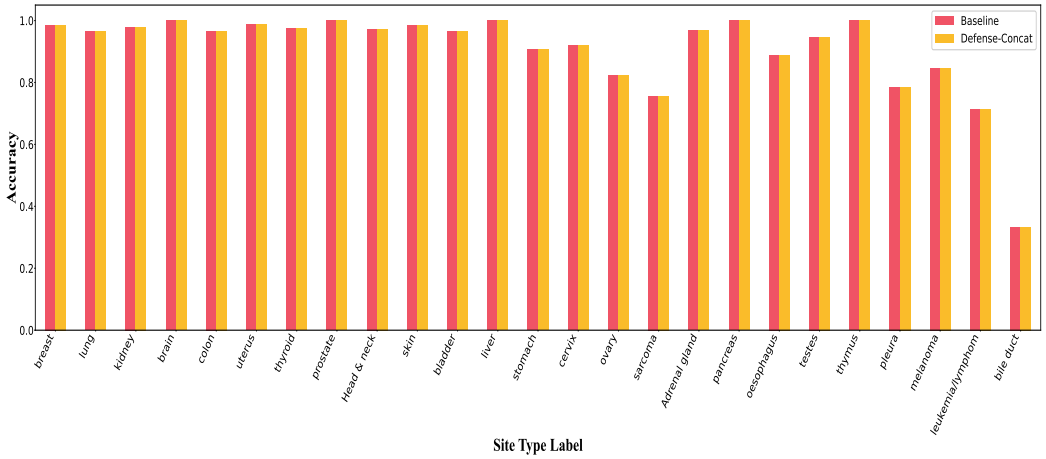

Figure A13. Accuracy per class in Defense-Concat.

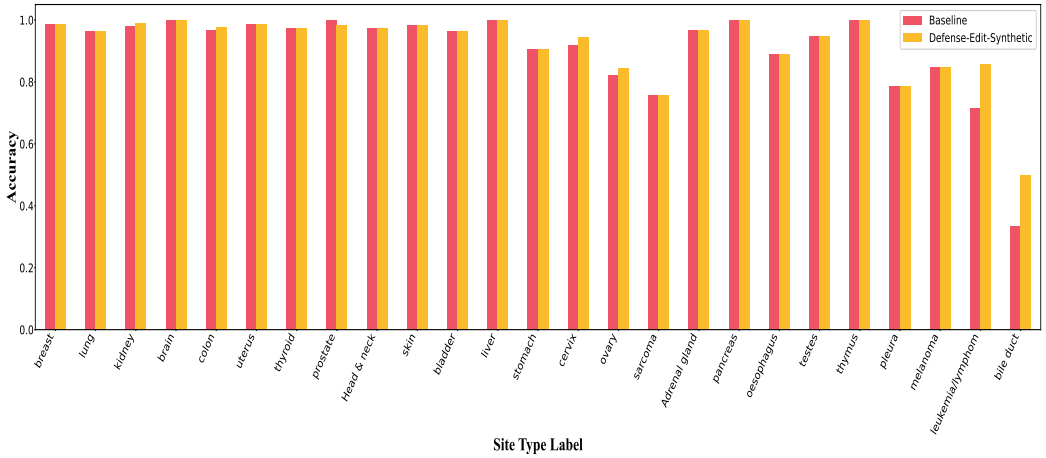

Figure A14. Accuracy per class in Defense-Edit-Synthetic.

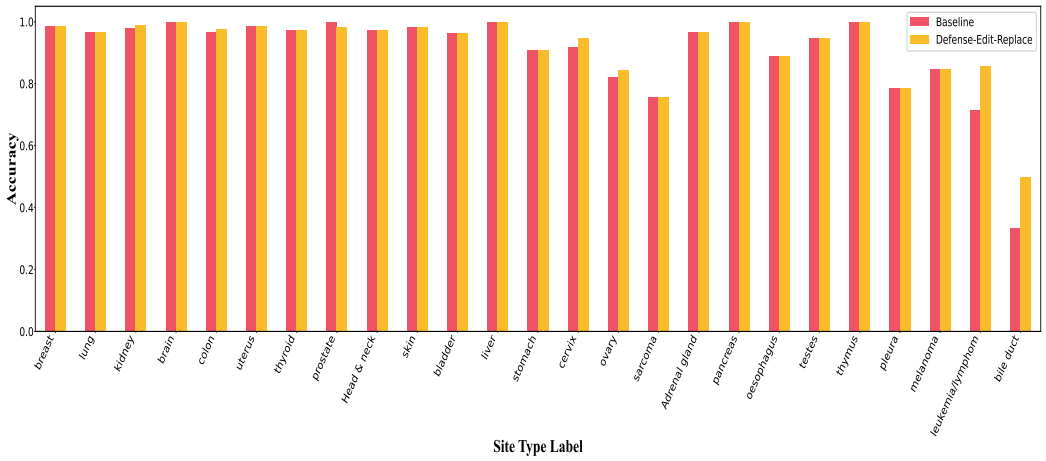

Figure A15. Accuracy per class in Defense-Edit-Replace.

Table A11 lists the results of overall micro/macro F1 by performing defense on Edit-Replacing for all classes names. From the result, we can easily see that defense strategy enhance the robustness of the CNN model.

Table A11. Overall micro/macro F1 by performing defense.

	Micro F1	**Macro F1**
Baseline	0.9544	0.9240
Edit-Replacing-Breast	0.9583	0.9369
Edit-Replacing-Lung	0.9583	0.9369
Edit-Replacing-Kidney	0.9583	0.9369
Edit-Replacing-Brain	0.9583	0.9369
Edit-Replacing-colon	0.9583	0.9369
Edit-Replacing-uterus	0.9583	0.9369
Edit-Replacing-thyroid	0.9583	0.9369
Edit-Replacing-prostate	0.9583	0.9369
Edit-Replacing-head and neck	0.9583	0.9369
Edit-Replacing-skin	0.9583	0.9369
Edit-Replacing-bladder	0.9583	0.9369
Edit-Replacing-liver	0.9583	0.9369
Edit-Replacing-stomach	0.9583	0.9369
Edit-Replacing-cervix	0.9583	0.9369
Edit-Replacing-ovary	0.9583	0.9369
Edit-Replacing-sarcoma	0.9583	0.9369
Edit-Replacing-adrenal gland	0.9583	0.9369
Edit-Replacing-pancreas	0.9583	0.9369
Edit-Replacing-oesophagus	0.9583	0.9369
Edit-Replacing-testes	0.9583	0.9369
Edit-Replacing-thymus	0.9583	0.9369
Edit-Replacing-melanoma	0.9583	0.9369
Edit-Replacing-leukemia/lymphoma	0.9583	0.9369
Edit-Replacing-bile duct	0.9583	0.9369

References

1. Köksal, Ö.; Akgül, Ö. A Comparative Text Classification Study with Deep Learning-Based Algorithms. In Proceedings of the 2022 9th International Conference on Electrical and Electronics Engineering (ICEEE), Alanya, Turkey, 29–31 March 2022; IEEE: New York, NY, USA, 2022; pp. 387–391.
2. Varghese, M.; Anoop, V. Deep Learning-Based Sentiment Analysis on COVID-19 News Videos. In Proceedings of the International Conference on Information Technology and Applications, Lisbon, Portugal, 20–22 October 2022; Spinger: Berlin/Heidelberg, Germany, 2022; pp. 229–238.

3. Affi, M.; Latiri, C. BE-BLC: BERT-ELMO-Based deep neural network architecture for English named entity recognition task. *Procedia Comput. Sci.* **2021**, *192*, 168–181. [CrossRef]
4. Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol. (TIST)* **2020**, *11*, 1–41. [CrossRef]
5. Alawad, M.; Yoon, H.J.; Tourassi, G.D. Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. In Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, USA, 4–7 March 2018; pp. 218–221. [CrossRef]
6. Olthof, A.W.; van Ooijen, P.M.A.; Cornelissen, L.J. Deep Learning-Based Natural Language Processing in Radiology: The Impact of Report Complexity, Disease Prevalence, Dataset Size, and Algorithm Type on Model Performance. *J. Med. Syst.* **2021**, *45*. [CrossRef] [PubMed]
7. Wang, Y.; Bansal, M. Robust machine comprehension models via adversarial training. *arXiv* **2018**, arXiv:1804.06473.
8. Suya, F.; Chi, J.; Evans, D.; Tian, Y. Hybrid batch attacks: Finding black-box adversarial examples with limited queries. In Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), Boston, MA, USA, 12–14 August 2020; pp. 1327–1344.
9. Yala, A.; Barzilay, R.; Salama, L.; Griffin, M.; Sollender, G.; Bardia, A.; Lehman, C.; Buckley, J.M.; Coopey, S.B.; Polubriaginof, F.; et al. Using Machine Learning to Parse Breast Pathology Reports. *bioRxiv* **2016**. [CrossRef] [PubMed]
10. Buckley, J.M.; Coopey, S.B.; Sharko, J.; Polubriaginof, F.C.G.; Drohan, B.; Belli, A.K.; Kim, E.M.H.; Garber, J.E.; Smith, B.L.; Gadd, M.A.; et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J. Pathol. Inform.* **2012**, *3*, 23. [CrossRef] [PubMed]
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
12. Gao, S.; Alawad, M.; Young, M.T.; Gounley, J.; Schaefferkoetter, N.; Yoon, H.J.; Wu, X.C.; Durbin, E.B.; Doherty, J.; Stroup, A.; et al. Limitations of Transformers on Clinical Text Classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3596–3607. [CrossRef] [PubMed]
13. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. Adversarial Attacks and Defences: A Survey, 2018. *arXiv* **2018**, arXiv:1810.00069.
14. Long, T.; Gao, Q.; Xu, L.; Zhou, Z. A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions. *Comput. Secur.* **2022**, *121*, 102847. [CrossRef]
15. Simoncini, W.; Spanakis, G. SeqAttack: On adversarial attacks for named entity recognition. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 308–318.
16. Araujo, V.; Carvallo, A.; Aspillaga, C.; Parra, D. On adversarial examples for biomedical nlp tasks. *arXiv* **2020**, arXiv:2004.11157.
17. Jin, D.; Jin, Z.; Zhou, J.T.; Szolovits, P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8018–8025.
18. Gao, J.; Lanchantin, J.; Soffa, M.L.; Qi, Y. Black-box generation of adversarial text sequences to evade deep learning classifiers. In Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24–24 May 2018; IEEE: New York, NY, USA, 2018; pp. 50–56.
19. Yuan, L.; Zheng, X.; Zhou, Y.; Hsieh, C.J.; Chang, K.W. On the Transferability of Adversarial Attacksagainst Neural Text Classifier. *arXiv* **2020**, arXiv:2011.08558.
20. Pei, W.; Yue, C. Generating Content-Preserving and Semantics-Flipping Adversarial Text. In Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May–3 June 2022; pp. 975–989.
21. Finlayson, S.G.; Kohane, I.S.; Beam, A.L. Adversarial Attacks Against Medical Deep Learning Systems. *CoRR* **2018**, *abs/1804.05296*. Available online: http://xxx.lanl.gov/abs/1804.05296 (accessed on 1 December 2022).
22. Mondal, I. BBAEG: Towards BERT-based biomedical adversarial example generation for text classification. *arXiv* **2021**, arXiv:2104.01782.
23. Zhang, R.; Zhang, W.; Liu, N.; Wang, J. Susceptible Temporal Patterns Discovery for Electronic Health Records via Adversarial Attack. In Proceedings of the International Conference on Database Systems for Advanced Applications, Taipei, Taiwan, 11–14 April; Springer: Berlin/Heidelberg, Germany, 2021; pp. 429–444.
24. Sun, M.; Tang, F.; Yi, J.; Wang, F.; Zhou, J. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 793–801.
25. Xu, H.; Ma, Y.; Liu, H.C.; Deb, D.; Liu, H.; Tang, J.L.; Jain, A.K. Adversarial attacks and defenses in images, graphs and text: A review. *Int. J. Autom. Comput.* **2020**, *17*, 151–178. [CrossRef]
26. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
27. Wang, W.; Park, Y.; Lee, T.; Molloy, I.; Tang, P.; Xiong, L. Utilizing Multimodal Feature Consistency to Detect Adversarial Examples on Clinical Summaries. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, Online, 19 November 2020; pp. 259–268.
28. Belinkov, Y.; Bisk, Y. Synthetic and natural noise both break neural machine translation. *arXiv* **2017**, arXiv:1711.02173.

29. Alawad, M.; Gao, S.; Qiu, J.; Schaefferkoetter, N.; Hinkle, J.D.; Yoon, H.J.; Christian, J.B.; Wu, X.C.; Durbin, E.B.; Jeong, J.C.; et al. Deep transfer learning across cancer registries for information extraction from pathology reports. In Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Chicago, IL, USA, 19–22 May 2019; IEEE: New York, NY, USA, 2019; pp. 1–4. [CrossRef]
30. Gao, S.; Alawad, M.; Schaefferkoetter, N.; Penberthy, L.; Wu, X.C.; Durbin, E.B.; Coyle, L.; Ramanathan, A.; Tourassi, G. Using case-level context to classify cancer pathology reports. *PLoS ONE* **2020**, *15*, e0232840. [CrossRef] [PubMed]