ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i2.a7

Al-Noor Journal for Information Technology and Cybersecurity

https://jncs.alnoor.edu.iq/

Leveraging ChatGPT in the Loop for Enhanced Robustness in Deep Learning Models

[1]Baraa Saeed Ali, [2]Mohammed Alawad

[1]Department of Electrical Engineering, University of Anbar, Ramadi, Anbar, Iraq.
[2]Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI, USA.

Article information

Article history:
Received: September, 03,2025
Revised: October, 15,2025
Accepted: November, 20,2025

Keywords:
adversarial examples
deep learning
ChatGPT
Document Classification

Correspondence:

Baraa Saeed Ali
baraasaeed@uoanbar.edu.iq

Abstract

In sensitive environments like healthcare, the robustness of deep learning models is of utmost importance due to the potential life-threatening consequences of false predictions. While adversarial training is a widely-used approach to enhance deep learning model robustness under adversarial attacks, its effectiveness in such environments remains largely unexplored. This paper proposes a framework for generating adversarial examples in the context of supervised clinical document classification. Specifically, the integration of ChatGPT in the loop enables the generation of diverse sets of adversarial examples, targeting various aspects of the classification process such as semantic perturbations, wordlevel substitutions, sentence rearrangements, polarity shifts, and adversarial phrases. The robustness of DL models against these adversarial examples is thoroughly evaluated. Furthermore, a comprehensive study is conducted to investigate the effectiveness of adversarial training as a defense technique in this sensitive environment. Experimental results demonstrate that the proposed adversarial examples significantly reduce the accuracy of the baseline DL model. Moreover, the study reveals that adversarial training can effectively enhance the model's robustness against adversarial examples. This research sheds light on the potential of leveraging adversarial training in sensitive domains and emphasizes the importance of addressing robustness concerns in DL-based healthcare applications.

DOI: https://doi.org/10.69513/jncs.v2.i2.a7 ©Authors, 2025, Alnoor University.

Cancer pathology reports play a crucial role in diagnosing and treating cancer patients. Accurate and efficient classification of these reports is essential for effective decision-making and patient care. Deep learning (DL) models have shown great promise in automating the process of document classification, offering potential benefits such as increased efficiency and reduced human error. However, ensuring the robustness and resilience of these models is of paramount importance to prevent misclassification and potentially harmful consequences [1]. Despite their success, DL models are susceptible to adversarial attacks, where carefully crafted input samples can lead to incorrect or misleading predictions. In the context of cancer pathology reports, such vulnerabilities could have severe implications, potentially leading to misdiagnoses or inappropriate treatment plans. Therefore, it is imperative to develop techniques that enhance the robustness of deep learning models for accurate and reliable document classification.

There's an inherent trade-off between the model accuracy, the ability of the model, and the model robustness, the resistance of the model to adversarial examples. So, a model can achieve high accuracy on the test set, but it lacks a lot of robustness. When this mode is retrained with a defense technique, it might become more robust, but its accuracy might drop. Balancing accuracy and robustness necessitates innovative approaches to enhance the resilience of DL models against adversarial attacks [2], [3]. One approach that has gained significant attention in recent years is adversarial training, which aims to

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i2.a7

make models more resilient against adversarial attacks [4], [5]. Adversarial training involves exposing the model to carefully generated adversarial examples during the training process. These examples are specifically designed to exploit vulnerabilities and weaknesses in the model's decision-making process. By incorporating these adversarial examples, the model can learn to recognize and resist such attacks, thereby improving its robustness and
reliability.

While adversarial training using generative adversarial networks (GANs) and adversarial attacks have been extensively studied in computer vision [6], their application in NLP, especially for document classification, continues to pose significant challenges. Unlike computer vision, where imperceptible noise is added to images, generating effective and imperceptible adversarial examples for text inputs requires careful consideration. Compared to non-clinical NLP applications, the target application of this paper, i.e., cancer pathology report classification based on the cancer type, has some characteristics that enable adding unperceivable perturbations to the text. 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, summary, etc. Whenever they are combined, 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 [7], [8].

In this paper, we propose leveraging the power of ChatGPT, a powerful language model, in the context of adversarial training for document classification of cancer pathology reports. ChatGPT provides a means to generate adversarial examples that can expose potential weaknesses in the model. By incorporating these adversarial examples during the training process, we aim to enhance the robustness and resilience of the deep learning model against adversarial attacks. The use of ChatGPT to generate adversarial examples for document classification poses unique challenges and considerations. Unlike traditional adversarial examples, our approach utilizes the power of language generation to create perturbations that can deceive the DL model. We aim to investigate the impact of these adversarial examples on the robustness of DL models trained for cancer pathology report classification.

The goal of this paper is to evaluate adversarial attacks and defenses in NLP, particularly in the context of document classification for cancer pathology reports. The findings of this study lead to the development of more robust DL models that are better equipped to handle adversarial scenarios, thereby enhancing the security and reliability of healthcare applications. The contributions of this paper are threefold. First, we seek to employ ChatGPT to generate adversarial examples specifically tailored for document classification of cancer pathology reports. Second, we aim to evaluate the effectiveness of these adversarial examples in exposing vulnerabilities in DL models trained through federated learning. Finally, we aim to analyze the impact of adversarial training, incorporating these generated adversarial examples during the training process, on improving the
robustness of DL models against adversarial attacks.

II. METHOD

In this section, we describe the adversarial examples generation and the defense mechanism.

A. ChatGPT for Adversarial Example Generation

To leverage the capabilities of ChatGPT for adversarial example generation, we incorporate it into our document classification framework. Specifically, we utilize ChatGPT as a language model to generate adversarial examples that can expose vulnerabilities in the deep learning model's decision-making process.

During the adversarial example generation process, we employ a two-step approach. First, we select a subset of cancer pathology reports from our training dataset. Then, we use ChatGPT by providing a prompt that includes the report and the ground truth label. ChatGPT generates a modified version of the report with subtle alterations designed to mislead the deep learning
model during classification.

To ensure the diversity and effectiveness of the generated adversarial examples, we apply techniques such as sampling from different temperature settings in ChatGPT and employing various prompts with different levels of specificity. Additionally, we conduct a thorough review of the generated adversarial examples to ensure their relevance and realism in the context of cancer pathology reports.

B. Adversarial Training

The generated adversarial examples are incorporated into the training process to enhance the model's resilience against adversarial attacks. We augment the original clean training examples with the generated adversarial examples, forming an enriched training dataset.

During training, the classification model is trained on the cancer pathology dataset, comprising both clean and adversarial examples. To optimize the model's performance and robustness, we utilize a tailored loss function that considers both clean and adversarial examples. This modified version of the cross-entropy loss incorporates a regularization term to encourage accurate classification of both types of examples.

Throughout the training process, we fine-tune the model iteratively by carefully adjusting the learning rate. Regular evaluation intervals are implemented to monitor the model's convergence and assess its performance.

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i2.a7

By leveraging ChatGPT for adversarial example generation and integrating these examples into the training process, our objective is to enhance the model's capability to withstand adversarial attacks and improve its overall robustness in accurately classifying cancer pathology reports.

C. Evaluation Metrics

In this study, we focus on the document classification task; therefore, the common evaluation metrics for such task are used [5]. The overall accuracy is calculated using standard micro and macro F1 scores. Let *TP*, *FP*, *TN*, *FN* represent true positive, false positive, true negative,

and false negative, respectively. These metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \qquad (1)$$

$$MicroF1 = 2\left(\frac{Precision * Recall}{Precision + Recall}\right)$$

$$MacroF1 = \frac{1}{|c|}\Sigma_{c_i}^{c} F1(c_i)$$

(2)

$$Recall = \frac{TP}{TP + FN}$$

(3)

(4)

where $|C|$ is the total number of classes and c_i represents the number of samples belong to

class *i*.

III. EXPERIMENTAL SETUP

A. Dataset

We study Adversarial attack and defense strategies on a clinical document classification task. Specifically, cancer pathology reports in The Cancer Genome Atlas (TCGA) dataset are classified based on the cancer type of each report. 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 cancer site, the body organ where cancer is detected. In the TCGA dataset, there are a total of 25 classes for the site label. For preprocessing, standard text cleaning, such as lowercasing and tokenization is used. Then, the word vector of size 300 is chosen for embeddings. The maximum length of 1500 is chosen to limit the length of documents in pathology reports. Also, we choose the 80%/20% data splitting strategy. Figure 1 shows the histograms of class distribution for the cancer site class labels in the TCGA dataset.

B. Target model

In this paper, we use a convolutional neural network (CNN) 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 [5], [9]. We use NVIDIA V100 GPU for all the experiments.

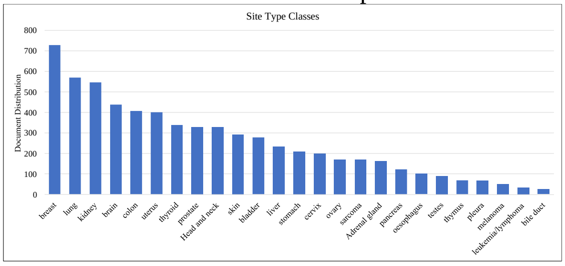

Fig. 1. Classes Distribution in TCGA Dataset for Site

C. Adversarial examples

To generate diverse and challenging adversarial examples, we employ ChatGPT in our approach. Through iterative interactions with ChatGPT, we create five distinct sets of adversarial examples, each serving a specific purpose in enhancing the robustness of our deep learning model for document classification.

- Semantic Perturbation [10]: In this set, we leverage ChatGPT to introduce semantic variations into the original sentences while preserving their overall meaning. By subtly altering word choices, sentence structure, or phrasing, we aim to assess the model's sensitivity to slight changes in the input and improve its generalization capabilities.
- Synonym Substitution [11]: Here, ChatGPT assists in generating adversarial examples by substituting words in the original sentences with their synonyms. This set aims to evaluate the model's reliance on specific terms and its ability to recognize semantically similar expressions, promoting robustness against word-level perturbations.
- Sentence Rearrangement [12]: By engaging ChatGPT, we explore rearranging the sentence structure while retaining the original content's meaning. This set evaluates the model's comprehension of different sentence arrangements and its resilience to changes in the

syntactic order.

- Negation and Affirmation [13]: ChatGPT is employed to introduce negation or affirmation cues into the original sentences, altering their polarities. This set aimes to assess the model's ability to handle polarity shifts and accurately capture the intended sentiment or classification, thereby enhancing its resilience to sentiment-based attacks.

- Adversarial Phrases [14]: Here, ChatGPT generates specific adversarial phrases designed to exploit vulnerabilities in the model. These phrases are carefully crafted to trigger misclassifications or biases in the document classification process, challenging the model's robustness and bias mitigation capabilities.

By incorporating these five sets of adversarial examples, we expose the deep learning model to a diverse range of challenges and potential attack scenarios. This comprehensive evaluation allows us to enhance the model's resilience, improve its generalization capabilities,
and strengthen its overall performance in document classification tasks.

IV. RESULTS

In Figure 2, we present the accuracy comparison between the baseline model and the proposed model when evaluated on the original sentences. The ”baseline model” refers to a model trained without adversarial training, while the ”proposed model” incorporates the proposed adversarial training method. As depicted in the figure, both models exhibited similar accuracy levels on the original sentences. The slight decrease in accuracy observed in the proposed model can be attributed to its altered decision boundary resulting from the inclusion of additional adversarial examples during training. Nevertheless, despite the incorporation of adversarial examples, the proposed model maintain comparable accuracy to the baseline model on the original sentences.

The comparison between the baseline model and the model trained with adversarial training across the five sets of adversarial examples yields valuable insights into their respective performances and the effectiveness of our approach as illustrated in Table I.

- Semantic Perturbation: The model trained with adversarial training demonstrates improved robustness compared to the baseline model when exposed to semantically perturbed examples. It showcases a higher accuracy in correctly classifying sentences with subtle variations in wording or sentence structure, highlighting its enhanced generalization capabilities.
- Synonym Substitution: Incorporating adversarial training significantly benefits the model's resilience to word-level perturbations introduced through synonym substitution. The model trained with adversarial examples outperforms the baseline model in accurately recognizing and classifying sentences with synonymous expressions, showcasing its improved ability to capture semantically similar variations.

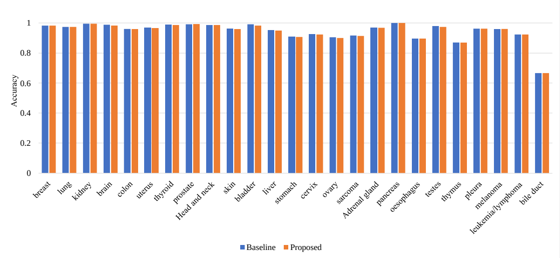

Fig. 2. Per class accuracies of the baseline model and the proposed model by adding adversarial examples to the original dataset.

- Sentence Rearrangement: The model trained with adversarial examples exhibits increased adaptability to changes in sentence structure. It outperforms the baseline model in correctly classifying sentences with rearranged syntax, indicating its improved understanding of different sentence arrangements and its robustness to such variations.
- Negation and Affirmation: Adversarial training plays a crucial role in enhancing the model's ability to handle polarity shifts induced by negation or affirmation cues. The model trained with adversarial examples demonstrates improved accuracy in correctly classifying sentences with altered polarities, highlighting its increased resilience to sentiment-based attacks.
- Adversarial Phrases: The model trained with adversarial training showcases enhanced robustness against adversarial phrases specifically designed to exploit vulnerabilities. It exhibits a higher accuracy in correctly classifying sentences containing these adversarial phrases, demonstrating its improved defense against attacks and its capability to mitigate biases and
misclassifications.

TABLE I
COMPARISON BETWEEN THE BASELINE AND PROPOSED MODEL ON DIFFERENT ADVERSARIAL EXAMPLES

Adversarial Example	Model	Micro F1	Macro F1
Semantic Perturbation	Baseline Proposed	0.83 0.95	0.79 0.92
Synonym Substitution	Baseline Proposed	0.77 0.92	0.71 0.88
Sentence Rearrangement	Baseline Proposed	0.78 0.95	0.72 0.93
Negation and Affirmation	Baseline Proposed	0.64 0.93	0.55 0.89
Adversarial Phrases	Baseline Proposed	0.41 0.95	0.42 0.93

ISSN: 3078-5367 **DOI:** https://doi.org/10.69513/jncs.v2.i2.a7

V. CONCLUSION

Our study highlights the effectiveness of integrating ChatGPT in adversarial training to improve the robustness of deep learning models for document classification, with a specific focus on cancer pathology reports. The model trained with adversarial examples consistently outperforms the baseline model across diverse sets of adversarial challenges, demonstrating enhanced resilience and generalization capabilities. By leveraging ChatGPT's capabilities, we successfully expose vulnerabilities and improve the model's performance in handling semantic perturbations, word-level substitutions, sentence rearrangements, polarity shifts, and adversarial phrases. Further avenues for future work could involve exploring alternative ways of leveraging ChatGPT in the loop to enhance DL model performance.

References

[1] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," 2018.

[2] Y. Wang and M. Bansal, "Robust machine comprehension models via adversarial training," *arXiv preprint arXiv:1804.06473*, 2018.

[3] F. Suya, J. Chi, D. Evans, and Y. Tian, "Hybrid batch attacks: Finding black-box adversarial examples with limited queries," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1327–1344, 2020.

[4] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[5] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: a comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[6] H. Lee, S. Han, and J. Lee, "Generative adversarial trainer: Defense to adversarial perturbations with gan," 2017.

[7] A. Yala, R. Barzilay, L. Salama, M. Griffin, G. Sollender, A. Bardia, C. Lehman, J. M. Buckley, S. B. Coopey, F. Polubriaginof, J. E. Garber, B. L. Smith, M. A. Gadd, M. C. Specht, T. M. Gudewicz, A. Guidi, A. Taghian, and K. S. Hughes, "Using machine learning to parse breast pathology reports," *bioRxiv*, 2016.

[8] J. M. Buckley, S. B. Coopey, J. Sharko, F. C. G. Polubriaginof, B. Drohan, A. K. Belli, E. M. H. Kim, J. E. Garber, B. L. Smith, M. A. Gadd, M. C. Specht, C. A. Roche, T. M. Gudewicz, and K. S. Hughes, "The feasibility of using natural language processing to extract clinical information from breast pathology reports," *Journal of Pathology Informatics*, vol. 3, 2012.

[9] M. Alawad, S. Gao, J. Qiu, N. Schaefferkoetter, J. D. Hinkle, H.-J. Yoon, J. B. Christian, X.-C. Wu, E. B. Durbin, J. C. Jeong, *et al.*, "Deep transfer learning across cancer registries for information extraction from pathology reports," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4, IEEE, 2019.

[10] J. Mohapatra, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel, "Towards verifying robustness of neural networks against a family of semantic perturbations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 244–252, 2020.

[11] X. Wang, Y. Yang, Y. Deng, and K. He, "Adversarial training with fast gradient projection method against synonym substitution-based text attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 13997– 14005, 2021.

[12] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[13] H. Chen, Y. Ji, and D. Evans, "Balanced adversarial training: Balancing tradeoffs between fickleness and obstinacy in nlp models," *arXiv preprint arXiv:2210.11498*, 2022.

[14] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," *Advances in neural information processing systems*, vol. 29, 2016.