SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I8P111

Volume 12 Issue 8, 112-119, August 2025


*Original Article*

# Susceptibility Analysis Using Adversarial Attacks: A Deep Learning Perspective on Contextual Clinical Text

Jaya A. Zalte[1,2], Harshal Shah[1]

[1]*Computer Science Engineering Department, Parul University, Vadodara, Gujarat, India.*
[2]*Computer Engineering Department, Shah and Anchor Kutchhi Engineering College, Mumbai, India.*

[1]*Corresponding Author : harshal.shah@paruluniversity.ac.in*

Received: 07 June 2025 Revised: 09 July 2025 Accepted: 08 August 2025 Published: 30 August 2025

***Abstract -*** *Deep learning models have demonstrated a strong performance in various classifications of varied amounts of data. As these models are prone to various attacks, even the smallest change can generate errors and lead to the classification of data. Adversarial attacks, which can significantly impact the model's performance, pose a threat to these models. In this work, the vulnerability of deep learning models in clinical contextual text classification using adversarial perturbations is demonstrated. By applying the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), evaluating the model robustness and data sensitivity, and were able to demonstrate the attacks with a decrease in accuracy drop of 23%. With white box attacks, trained a DistilBERT model and optimized the model accordingly to sustain the attacks. Our results demonstrate significant prediction shifts from minor input perturbations and suggest a new metric for calculating the susceptibility of the underlying text that generates a susceptible score. Further, the adversarially trained model can withstand FGSM and PGD attacks significantly.*

***Keywords -*** *Adversarial attacks, Deep learning, FGSM, PGD, Text classification.*

## 1. Introduction

Recent advances in Artificial Intelligence (AI) have led to the utilization of Deep Learning in sensitive and critical domains such as healthcare. Despite high predictive accuracy, these models are prone to adversarial attacks, where subtle input perturbations cause incorrect predictions. As the healthcare domain incorporates Artificial Intelligence systems, the records fetched from a database stored electronically can contain critical and sensitive data. As in the world of Natural language processing, each text record has a semantic and syntactical meaning. The context of the text is sometimes ignored. Hidden context behind the sentences can reveal a lot about a patient, it could be leading to various diseases and are susceptible in nature, like records indicating the influence of drugs, depression symptoms. Such susceptible data points need to be identified. This study aims to explore and address such vulnerabilities through a structured framework, where classification of textual data, depending on the susceptible data that can be vulnerable, is classified using the Binary Encoder Representations (BERT) model. Machine learning and deep learning algorithms have gained huge popularity for efficiently classifying and predicting on trained data [1]. Incorporating such models into real-world applications like disease prediction and medical imaging using MRI in medical systems has presented its role in healthcare as a prominent future for medical and healthcare domains. Identification of critical data points in such a model can lead us to susceptible data that can be insecure and prone to various attacks. It becomes difficult to detect if there are minute changes in input data, causing a major hazardous effect on the misuse of information, incorrect medical treatments, and classification of wrong data and hence wrong disease predictions. Such events can mislead the medical practitioner into giving away missed or incorrect information to patients. Relying on AI completely, the dependent trained model should not be vulnerable to attacks. One such attack is well defined by Goodfellow [2], which outlines the AI models' vulnerability to black box and white box attacks with various adversarial examples. Model vulnerabilities are often overlooked in critical systems. Since the healthcare domain includes various sub-domains, clinical text records are considered in this study to assess susceptible parameters. To develop a robust model that caters to textual data, which can be well-trained with adversarial attacks. White box attacks can be employed to target data to detect susceptible models. Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) white box adversarial examples are implemented. These attacks misclassify the word embeddings in clinical text with a variation in confidence drop as compared to the original confidence. Finally, to mitigate such attacks, a defensive adversarial trained neural network model is in place to sustain such attacks. Unlike prior work focused on image or general NLP tasks, the tasks were examined for adversarial susceptibility in clinical free-text classification, where data is

sensitive, context-heavy, and error-prone. This identifies the Gap between the shift of work from image-based data to textual data.

The key Contributions of this study are:

- Identification of Vulnerable Data: Utilizing deep learning models to detect data points susceptible to adversarial attacks, which are crucial in domains like healthcare, where data integrity is needed.
- Enhancement of Model Robustness: Implementing strategies such as adversarial training to fortify models against potential threats, ensuring reliability in real-world applications.

## 2. Related Work

In [3], the author performed attacks with the Fast Gradient Sign method, the Projected Gradient Descent method, and the Carlini and Wagner attack to change the input images and observed the model's accuracy change for the perturbation test. In the growing age of AI, healthcare systems consist of critical and sensitive data that can be altered by an attacker. Manipulate using Fast Gradient Sign Method (FGSM) for medical images. The smallest changes to MRI and X-ray images can lead to error-prone identification and misclassification of illness or produce erroneous treatments for patients; this could be a huge disaster if not prevented. In [4], the authors demonstrated that an FGSM adversarial attack misclassified the images, which caused the model's accuracy to drop to 11% compared to the original accuracy of 88%. When it comes to numerical data in healthcare, the breast cancer dataset was utilized by authors [5, 6], who demonstrated the use of FGSM, which reduced the accuracy of the model from 98% to 53%.

Author [7] proposed a defense mechanism for adversarial attacks using adversarial training and Gaussian data noise augmentation to recover the model's accuracy to 92%. FGSM and PGD attacks were performed on the ECG dataset, and demonstrated that white and black box attacks together can be applied to achieve a defense mechanism against the threats to the ECG signal. The attackers can generate the input perturbations easily just by adding adversarial examples with different generating methods [8, 9]. G. Chang et.al [10] used the black box technique to generate adversarial text with input shift with 0.2 epsilon values. J. Xu and Q. Du [11] proposed a white box technique with publicly available databases, which notably shows improved performance as compared to the black box technique. Distil Bidirectional Encoder Representations from Transformers (BERT) [12] was implemented to design a faster and smaller model with fewer computations to yield faster results of the trained model. W. Wang [13] compared the target value obtained from word-level perturbation with the output achieved. They were able to set the number of words to less than 5, not more than 5 words at a time, which could severely fail the defense model demonstrated. X. Han, et.al [14], provided a list of taxonomy, issues related to security and various types of attacks pertaining to testing the vulnerability of models. Authors M. Behjati [15] were able to provide the perturbations to the word input by bringing down the accuracy from 93% to 50%. Which is quite impressive, but they do not represent the defense model. Multilabel classification becomes again a challenging part, where the move from binary labels to multilabel and then having adversarial training on those texts [16]. M. Qaraei and R. Babbar were able to target multilabel but only one target at a time, which can be extended to more targeted labels. In [17], a synonym word was related to getting the same semantics and to unchanged the meaning of the sentences. The accuracy dropped from 95% to 86%, which is only a partial change that can be further improved by adding more perturbations to the model. Similar work was observed in [18-19], where authors have used a clinical document that consists of unstructured texts with a CNN model. X. Li, et.al [20] have used a CNN and an LSTM model to expose vulnerability while examining the test on text perturbations.

Finlayson et al. [21] showed that imperceptible pixel-level perturbations could cause deep learning models to misclassify dermatology images, potentially leading to incorrect skin cancer diagnoses. In another case, Paschali et al. [22] demonstrated that adversarial noise could substantially reduce segmentation accuracy in medical imaging systems, risking misidentification of tumor boundaries. From the literature studied, one can realize the importance of robustness of models in clinical settings to avoid major misinterpretations of results from machine learning or deep learning models.

Concern over the security and resilience of neural network models to hostile examples is on the rise. Many scholars have tackled this problem from various perspectives, putting forward methods and algorithms to produce negative examples and create efficient defenses. Analysis of the work that has stood out in this field in our examination of related works and score them based on how innovative and relevant our suggestion is. All the referred papers provide us with the need to identify a sustainable model framework that applies to text as well. Most of the work elaborates on the use of trained models for images. Very few articles discuss and demonstrate the use of contextual data for adversarial attacks. Most of the work discussed demonstrates the failure of deep learning models and their defensive mode. However, a model is proposed that identifies the susceptible text. A defensive model that can be very well trained for clinical texts that are sensitive and can be vulnerable is demonstrated.

## 3. Methodology

Data is pre-processed by cleaning the data, removing punctuation, and stop words to get plain text data. Each of the sentences is then tokenized and converted to vector format. Pre-trained Glove Embeddings are used to feed to the model. About 45437 unique tokens are extracted. A total of 400001

word vectors are generated. Each text is first tokenized using a tokenizer using Distil BERT. It is then fed to the Classifier for text classification of susceptible data points. Further, clean accuracy is calculated from the classified data. Then, FGSM and PGD attacks are implemented on the input data. Then the model is trained for adversarial training. A susceptible score is generated for both attacks. About 206926 sentences are used, out of which 80% are used for training and the rest for testing, with an accuracy of 90% approx. This set is used with the Distil BERT model with an accuracy of approximately 91%. The dataset was collected from an online platform (Kaggle website). The architecture of DistilBERT consists of 6 Transformer encoder layers, each containing a self-attention mechanism and a position-wise feed-forward network, and the model uses 12 attention heads. Despite being smaller, DistilBERT preserves the original BERT's key features. In this study, DistilBERT was used for clinical text classification. The model was initialized with pretrained weights with language patterns. Clinical text samples were first tokenized into units using DistilBERT's tokenizer, converted to token IDs, and truncated to a fixed sequence length. These token sequences were then fed into the DistilBERT encoder, which generated contextualized embeddings for each token. The input sequence was passed through a fully connected classification with a softmax activation to produce class probabilities. Fine-tuning was performed end-to-end, updating both the pretrained DistilBERT weights and the classification parameters.

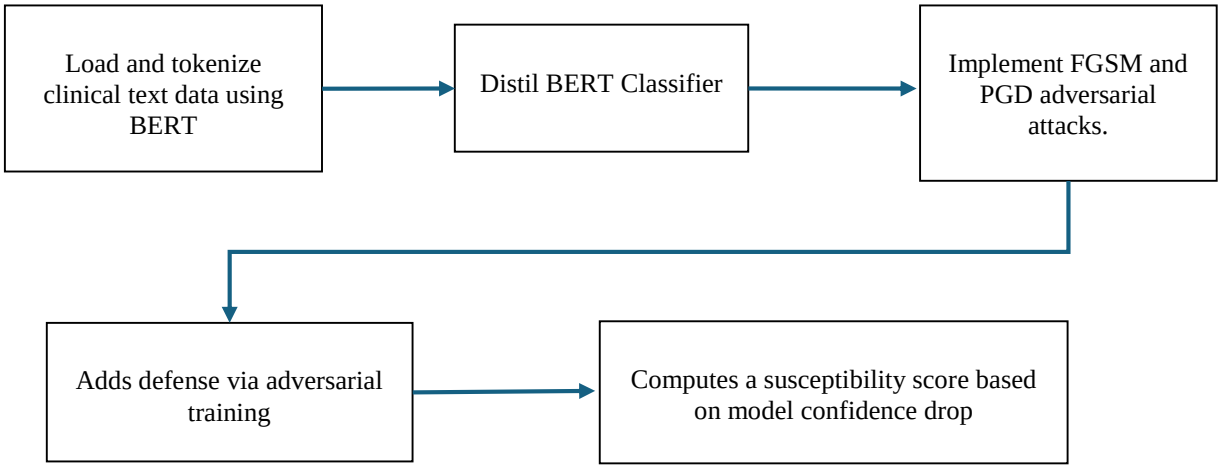


**Fig. 1 Methodology**

### *3.1. Algorithm*

- Data is preprocessed for text summaries. Each text is tokenized to get word embeddings.
- With BERT classification, texts are classified, and accuracy is calculated.
- Clean accuracy is generated from a trained model.
- With Adversarial attacks, various epsilon values are used to demonstrate the attacks on text embeddings.
- Two white box attacks, namely FGSM and PGD, are performed on the trained model to check the robustness of the model after the attack.
- The susceptibility score is calculated based on the confidence drop.
- The values generated determine the strength of the trained model, even with good accuracy.
- The next step is to create a robust model that sustains such attacks.
- The model is again trained with adversarial examples.
- Accuracy after adversarial training is generated.
- Susceptibility scores are generated for both attacks, and a comparative analysis is done.

### *3.2. Attack Method*

Fast Gradient Sign Method (FGSM) is a method to perturb input embeddings (text) or feature vectors (tabular). FGSM is a single-step, white-box adversarial attack introduced by Goodfellow et al. in 2015. It perturbs the input data in the direction that increases the model's loss the most, aiming to cause misclassification.

$$o = x + \varepsilon \times \text{sign}\left(\nabla_x J(\theta, x, y)\right) \quad (1)$$

o: Our output adversarial data
x: The original input embeddings
y: The ground-truth label of the input data
Ɛ: Small value is multiplied by the signed gradients to ensure the perturbations are small enough that the human eye cannot detect them, but large enough that they fool the neural network
θ: Our neural network model
J: The loss function

The gradient is calculated with respect to the inputs of the model in order to maximize the loss. Higher values of epsilon,

starting with 0.01 to 0.2, yield larger perturbations, while lower values are more subtle. FGSM is mostly used to demonstrate the vulnerability of deep learning models to adversarial attacks. Projected Gradient Descent (PGD) is an iterative, white-box adversarial attack considered one of the most potent first-order attacks. It extends FGSM by applying it multiple times with small step sizes, projecting the perturbed input back onto the valid data domain after each step. To minimize the loss function, this algorithm fine-tunes the model parameters. The equation is,

$$\theta_{t+1} = \theta_t - \alpha \times \nabla J(\theta_t) \quad (2)$$

Where $\theta_t$ represents the parameters used at iteration t,

$\alpha$ is the learning rate, and $\nabla J(\theta_t)$ is the gradient of the loss function.

The PGD method is a multi-step variant of the FGSM that generates adversary examples that are difficult to detect.

The PGD algorithm is as follows:

- Hyperparameters are epsilon, alpha, and the number of iterations.
- The input embeddings of the text and input IDs are considered. The number of iterations is set to 10.
- The gradient of the loss function for the input considered is calculated.
- Gradient with alpha parameters is tuned.

### *3.3. Metrics parameters used*

Table 1 describes the parameters used for the implementation of the model.

**Table 1. Parameters used for text classification**

| Category | Parameter | Value / Description |
|---|---|---|
| **Model** | Model architecture | DistilBERT (6-layer Transformer, 768 hidden units, 12 attention heads) |
| **Training** | Optimizer | AdamW |
| | Batch size | 32 |
| | Epochs | 10 |
| | Loss function | Cross-entropy loss |
| **Data Processing** | Max sequence length | 128 tokens |
| | Padding | Dynamic padding per batch |
| | Truncation | Applied to sequences >128 tokens |
| **Adversarial Setup** | Attack methods | FGSM, PGD |
| | FGSM epsilon values | [0.05, 0.1, 0.15, 0.2] |
| | PGD step size | 0.01 |
| | PGD iterations | 10 |
| | Adversarial training | FGSM-based augmentation with $\varepsilon = 0.1$ |
| **Environment** | Python version | 3.11.13 |
| | PyTorch version | 2.7.0 |
| | Transformers version | 4.36.2 |
| | GPU | NVIDIA Tesla T4 (16 GB) |

Accuracy: It measures the fraction of text classified by the model from the test set.

Accuracy = Number of output predictions/Total number of predictions

Susceptibility score is given as:

susceptibility=clean_confidence - perturbed_confidence

Clean confidence is generated by training models without adversarial training.

Perturbed confidence is given by the dropped confidence after a change in input text embeddings.

FGSM attack function:

perturbed_embeddings = embeddings + epsilon × grad_sign

Here, embeddings are calculated by adding the smallest input perturbations with epsilon values ranging from 0.01 to 0.2. For each value, the FGSM values are observed.

PGD attack function:

delta = (delta + alpha × grad_sign) × (Ɛ, epsilon)

where delta are the input embeddings, the gradient sign can be negative or positive, alpha 0.05, and epsilon varies from 0.01 to 2.0. This function is fed the number of iterations given as input by users. Figure 2 explains the detailed flow chart for adversarial attacks.

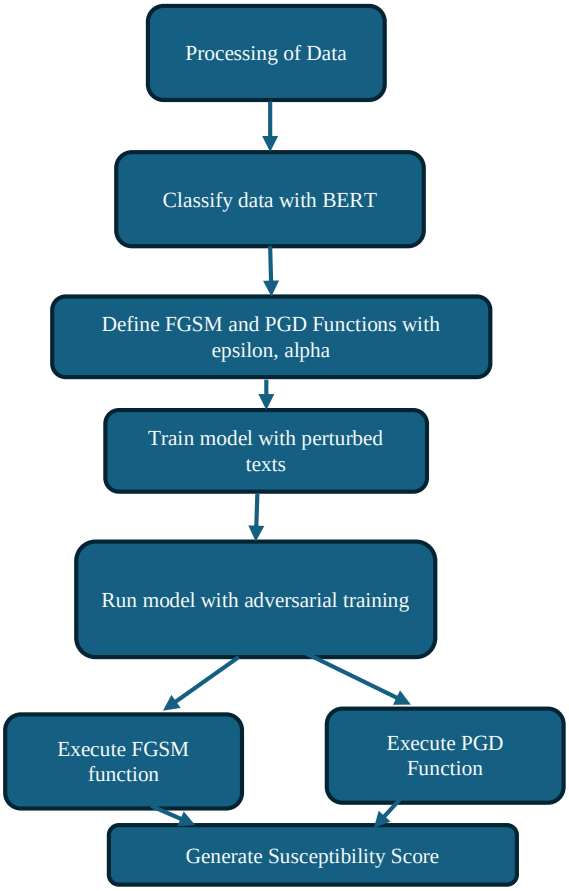


**Fig. 2 Processing steps for adversarial attack**

## 4. Results and Discussions

In this section, a comparison of adversarial attacks is made to the input data word embeddings with a demonstration of different experiments. With adversarial examples, each attack is performed independently to assess the vulnerability of the trained model. The input data consists of text records and is labelled as 0 as non-sensitive and 1 as sensitive. The BERT model is already trained to classify text with attention to contextual meaning, tagging the input text as sensitive or nonsensitive. The vulnerable input data points are generated before giving them to adversarial example attacks. Here, the susceptibility data and the model are demonstrated. In Figure 3, input text is used; there is a significant confidence drop with every row of text given as input to the attack functions. Each label is further misclassified with a change of adversarial predictions. Making it more vulnerable and generating wrong predictions.

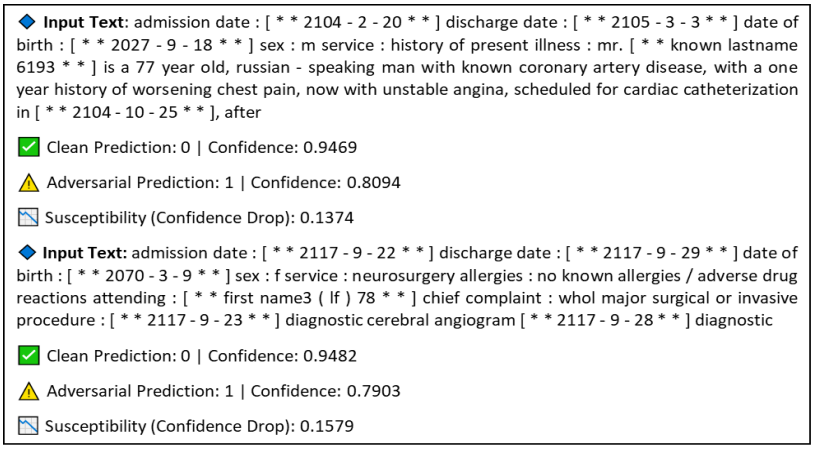


**Fig. 3 Sample text data after each input text, displaying confidence drop and susceptibility score**

In Table 2, the comparative results for both attacks pertaining to different epsilon values are shown. The accuracy generated on the clean model is approximately. 91%.

The table describes the adversarial accuracy for each epsilon value on FGSM and PGD attacks. From the observed values, one can see that, for the FGSM attack, the values remain the same after adversarial training. Meanwhile, PGD accuracy drops to 0.7181 from 0.9195 and sustains the attack for the successive epsilon values, maintaining its strength against PGD attack. Accuracy on clean test data generated is 0.9195

**Table 2. Comparison of the accuracy score for both attacks**

| Epsilon | Adversarial Accuracy- FGSM | Adversarial Accuracy-PGD |
|---|---|---|
| 0.00 | 0.9195 | 0.9195 |
| 0.01 | 0.9195 | 0.7181 |
| 0.05 | 0.9195 | 0.7181 |
| 0.10 | 0.9195 | 0.7181 |
| 0.15 | 0.9195 | 0.7181 |
| 0.20 | 0.9195 | 0.7181 |

Figure 4 shows that the accuracy remains unchanged with a susceptible value of 0.017 for the FGSM attack, with input perturbations. Models' accuracy remains unchanged because it can withstand a single-step white box FGSM attack. Figure 5 shows the robustness of the trained model, where with an epsilon of 0.01, accuracy remains the same, whereas it falls to 0.71 after 0.5 epsilon value and withstands further vulnerability, achieving the same constant accuracy till 0.2 epsilon value. From Table 3. As shown, the input data with clean accuracy is 91%.

After performing the attacks, FGSM and PGD, the susceptible scores are 0.651 and 0.2332, respectively. This indicates that, with FGSM, 6% drop in accuracy was generated. The 23% drop in the original accuracy with PGD attacks shows a significant drop in attacks where the model showed us vulnerability.

Figure 6 shows the overall susceptibility score before training and after training with FGSM and PGD attacks. The graph shows that it is highly sensitive to PGD attack, which is a stronger attack than FGSM, which is a single-step attack. An attack after training the model with an adversarial attack can withstand the FGSM attack with a 0.0017 score with input perturbations. As opposed to the PGD attack, it can sustain an attack with a 16% accuracy drop, and it further does not decrease the score.

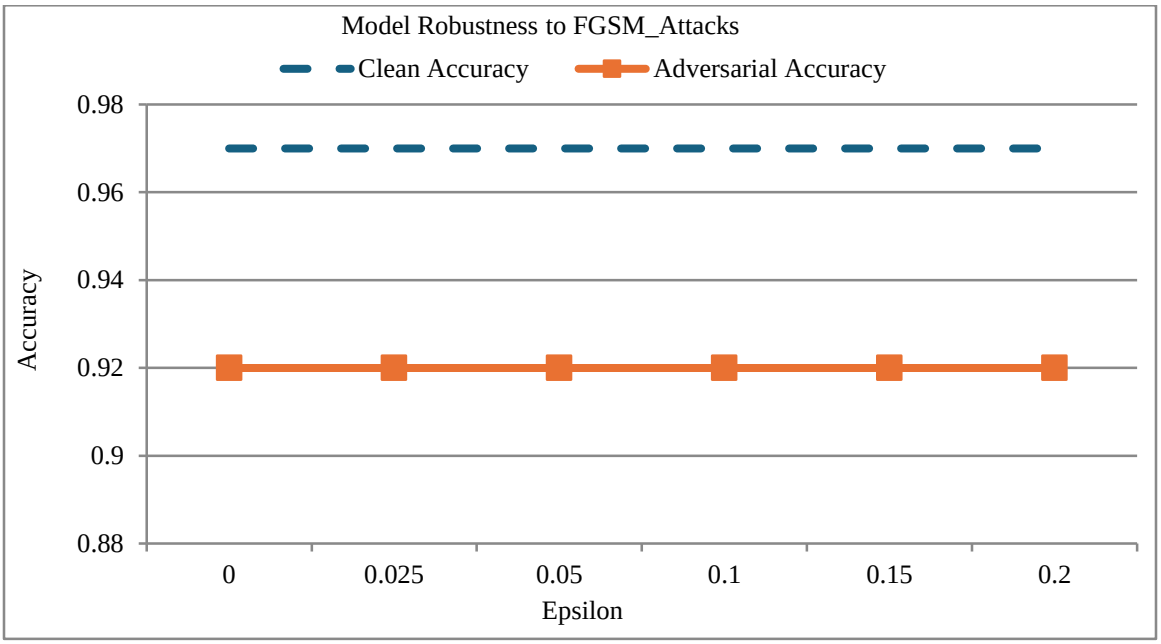


**Fig. 4 Model robustness to FGSM attacks with accuracy and different epsilon values**

**Table 3. Comparison of susceptibility scores for both attacks**

| Data | Before Attack | Susceptibility score with attack | Susceptibility score after adversarial training |
|---|---|---|---|
| Clinical Text with word embeddings | Clean Accuracy-91% | FGSM Susceptibility Score: 0.0651<br>PGD Susceptibility Score: 0.2332 | FGSM Susceptibility Score: 0.0017<br>PGD Susceptibility Score: 0.1681 |

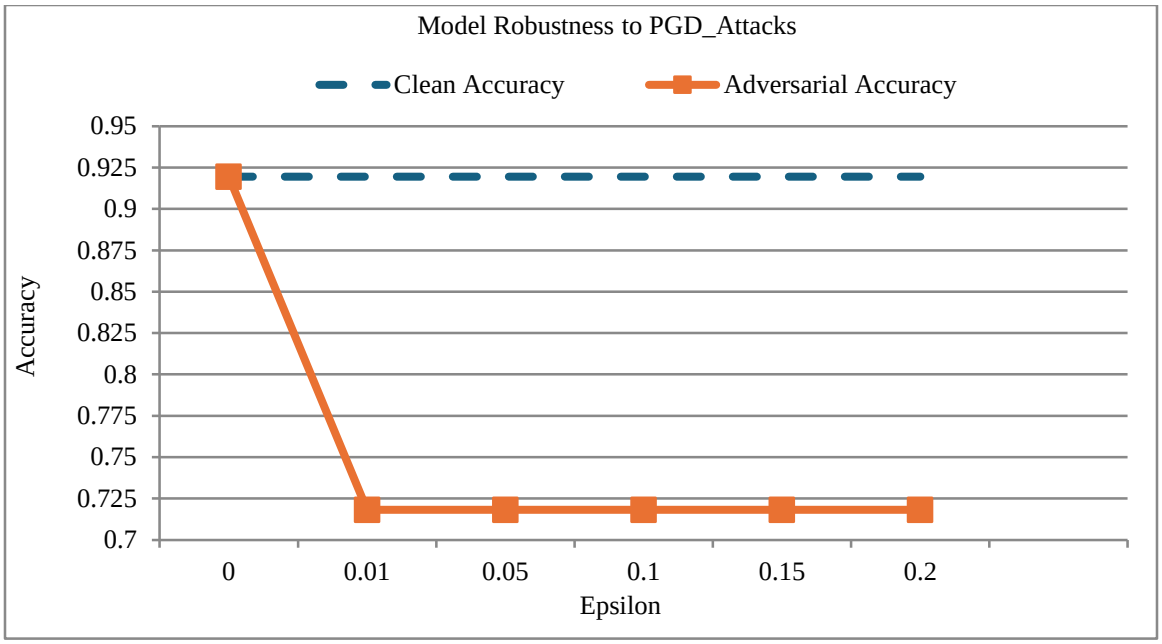


**Fig. 5 Model robustness to PGD attacks with accuracy and different epsilon values**

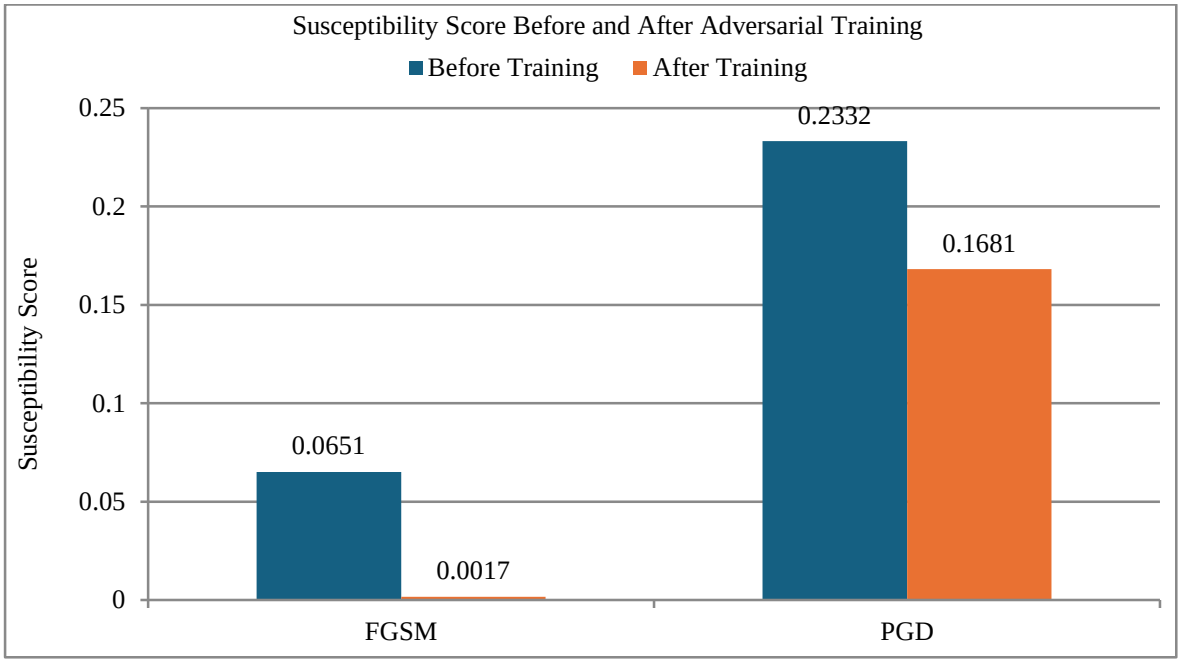


**Fig. 6 Susceptibility score before and after training with adversarial examples with FGSM and PGD attacks**

## 5. Conclusion

The model was evaluated for robustness against adversarial perturbations using FGSM and PGD attacks by computing susceptibility scores, defined as the average confidence drop between clean and adversarial inputs. Prior to adversarial training, the model exhibited susceptibility scores of 0.0651 and 0.2332 for FGSM and PGD attacks, respectively, indicating moderate vulnerability. After incorporating FGSM-based adversarial training, the susceptibility scores significantly decreased to 0.0017 for FGSM and 0.1681 for PGD attacks. This demonstrates the effectiveness of adversarial training in reducing model vulnerability, especially against single-step FGSM attacks.

While PGD susceptibility was also reduced, the model remains partially vulnerable to stronger iterative perturbations, suggesting that future work could explore PGD-based adversarial training for enhanced defense. Although FGSM-based adversarial training significantly improved robustness against single-step perturbations, the model

remained partially vulnerable to stronger iterative attacks such as PGD. This suggests that the defense mechanism is not fully generalizable across different attack strategies. This study explores only textual data; further, it could be merged with multi-level records with text and image data.

## Acknowledgments

J.Zalte contributed to this research with data preparation, literature work, analysis, and demonstration, and H.Shah contributed equally to this work with more supervision and guidance.

## References

[1] Christian Janiesch, Patrick Zschech, and Kai Heinrich, "Machine Learning and Deep Learning," *Electronic Markets*, vol. 31, no. 3, pp. 685-695, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ian Goodfellow et al., "Generative Adversarial Networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139-144, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] William Villegas-Ch, Angel Jaramillo-Alcázar, and Sergio Luján-Mora, "Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW," *Big Data and Cognitive Computing*, vol. 8, no. 1, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Sanjaykrishnan Ravikumar et al., "Securing AI of Healthcare: A Selective Review on Identifying and Preventing Adversarial Attacks," *2024 IEEE Opportunity Research Scholars Symposium (ORSS)*, USA, pp. 75-78, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Sarfraz Brohi, and Qurat-ul-ain Mastoi, "From Accuracy to Vulnerability: Quantifying the Impact of Adversarial Perturbations on Healthcare AI Models," *Big Data and Cognitive Computing*, vol. 9, no. 5, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Etidal Alruwaili, and Tarek Moulahi, "Prevention of Data Poisonous Threats on Machine Learning Models in e-Health," *ACM Transactions on Computing for Healthcare*, pp. 1-20, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sheikh Burhan Ul Haque et al., "Threats to Medical Diagnosis Systems: Analyzing Targeted Adversarial Attacks in Deep Learning-Based COVID-19 Diagnosis," *Soft Computing*, vol. 29, no. 3, pp. 1879-1896, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Lu Sun, Mingtian Tan, and Zhe Zhou, "A Survey of Practical Adversarial Example Attacks," *Cybersecurity*, vol. 1, no. 1, pp. 1-9, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Han Xu et al., "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151-178, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Guoqin Chang et al., "TextGuise: Adaptive Adversarial Example Attacks on Text Classification Model," *Neurocomputing*, vol. 529, pp. 190-203, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Jincheng Xu, and Qingfeng Du, "TextTricker: Loss-Based and Gradient-Based Adversarial Attacks on Text Classification Models," *Engineering Applications of Artificial Intelligence*, vol. 92, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Nicolas Papernot et al., "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," *2016 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA, pp. 582-597, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[13] Wenqi Wang et al., "TextFirewall: Omni-Defending against Adversarial Texts in Sentiment Classification," *IEEE Access*, vol. 9, pp. 27467-27475, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Xu Han et al., "Text Adversarial Attacks and Defenses: Issues, Taxonomy, and Perspectives," *Security and Communication Networks*, vol. 2022, no. 1, pp. 1-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Melika Behjati et al., "Universal Adversarial Attacks on Text Classifiers," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, pp. 7345-7349, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mohammadreza Qaraei, and Rohit Babbar, "Adversarial Examples for Extreme Multilabel Text Classification," *Machine Learning*, vol. 111, no. 12, pp. 4539-4563, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Hyun Kwon, and Sanghyun Lee, "Detecting Textual Adversarial Examples through Text Modification on Text Classification Systems," *Applied Intelligence*, vol. 53, no. 16, pp. 19161-19185, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Nina Fatehi, Qutaiba Alasad, and Mohammed Alawad, "Towards Adversarial Attacks for Clinical Document Classification," *Electronics*, vol. 12, no. 1, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Kexin Zhao et al., "Intriguing Properties of Universal Adversarial Triggers for Text Classification," *2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Wenzhou, China, pp. 480-487, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Xinzhe Li et al., "Exploring the Vulnerability of Natural Language Processing Models via Universal Adversarial Texts," *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pp. 138-148, 2021. [Google Scholar] [Publisher Link]

[21] Samuel G. Finlayson et al., "Adversarial Attacks on Medical Machine Learning," *Science*, vol. 363, no. 6433, pp. 1287-1289, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Magdalini Paschali et al., "Generalizability vs. Robustness: Investigating Medical Imaging Networks using Adversarial Examples," *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018*, Granada, Spain, pp. 493-501, 2018. [CrossRef] [Google Scholar] [Publisher Link]