JJEE Volume 10 | Number 1 | March 2024 | Pages 122-133

Jordan Journal of Electrical Engineering

ISSN (print): 2409-9600, ISSN (online): 2409-9619

Homepage: jjee.ttu.edu.jo

Comparing the Performance of Deep Denoising Sparse Autoencoder with Other Defense Methods Against Adversarial Attacks for Arabic letters

Mahmoud Mohamed[1*], Mohamed Bilal[2]

[1, 2] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdul Aziz University, Jeddah, Saudi Arabia

E-mail: mhassan0073@stu.kau.edu.sa

Received: June 21, 2023 *Revised: August 8, 2023* *Accepted: August 16, 2023*

***Abstract*—** The aim of this paper is to compare how effectively the Deep Denoising Sparse Autoencoder (DDSA) method performs compared to other defense strategies - like adversarial training, defensive distillation and feature squeezing - in dealing with adversarial attacks for Arabic letters. We strive to evaluate both the accuracy and robustness as well as efficiency of these methods by examining a test set from the Arabic Handwritten Characters Dataset while considering adversarial attacks. Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini and Wagner (C&W) are all part of this. Our research findings demonstrate that DDSA surpasses the rest of the defense methods in terms of classification accuracy and robustness. This exceptional performance is due to the distinctive attributes of DDSA, which concentrate on acquiring distinguishing features and integrating spatial information to improve defense against adversarial perturbations. While it necessitates more computational resources, DDSA's superior performance validates the additional expenses, particularly in critical applications where misclassification may have severe implications.

***Keywords*—** Deep denoising sparse autoencoder; Adversarial attacks; Deep learning; Fast gradient sign method; Arabic letters

1. INTRODUCTION

Image classification, speech recognition, natural language processing, and more are among the various applications where DL models have achieved remarkable success. Despite this, these models can be easily targeted by adversarial attacks, leading to erroneous output predictions. The definition of adversarial attacks is when modifications are made to the inputs of the DL model so that it produces incorrect predictions. This can be achieved by executing tiny, unnoticed adjustments to the input data. Adversarial attacks seriously jeopardize the security and reliability of DL models. During recent years, numerous defense approaches have been recommended in order to lessen the vulnerability of DL models against adversarial attacks. Among the most effective defense methods, adversarial training stands out. To enhance the model's resilience against adversarial attacks, it is recommended to include augmented training data with adversarial examples. Additional methods of defense involve feature squeezing, input denoising, and gradient masking. The deep defense security architecture proposed by Samangouei et al. [1], is a defense method aimed at improving the durability of DL models against adversarial attacks. Features are extracted from the input image at different scales by DDSA using a multi-scale feature extraction module. It also combines the extracted features and classifies the input image using a feature aggregation

* Corresponding author

Article's DOI: 10.5455/jjee.204-1687363297

module. DDSA has demonstrated encouraging outcomes in enhancing the resilience of DL models against adversarial attacks.

The increasing dependence on artificial intelligence (AI) and machine learning (ML) in different applications has emphasized the importance of strong defense mechanisms against adversarial attacks. Incorrect predictions and decisions are a result of the intentional manipulation of input data in adversarial attacks that deceive AI algorithms. Recognizing Arabic letters presents an even greater challenge given the intricate nature of the Arabic alphabet and its essential roles in Optical Character Recognition (OCR), document analysis, and language translation. The Arabic alphabet's complexity and its vital role in OCR, document analysis, and language translation make the challenge even more arduous. Hence, conducting a timely and essential comprehensive study that compares the performance of deep denoising sparse autoencoder (DDSA) and other defense methods against adversarial attacks for Arabic letters is necessary [2].

The primary motivation for undertaking this study is the rising implementation of AI and ML techniques across different industries. This has resulted in a rising requirement for systems that are secure and dependable. Potentially catastrophic consequences in safety-critical applications may result from the severe undermining of the effectiveness of AI models by adversarial attacks. By comparing different defense mechanisms, this study seeks to provide valuable insights for researchers and practitioners working on AI-ML based systems, aiming at the development of more robust and resilient models. Specifically in the field of recognizing Arabic letters. In addition, the importance of the Arabic alphabet in various applications necessitates a thorough investigation of defense mechanisms specifically tailored to this context. OCR systems widely utilize recognition of Arabic letters, which is an integral component and serves various sectors including education, government, and business. Their successful deployment depends on ensuring the security and reliability of these systems. This study will help in developing effective defense strategies against adversarial attacks in this domain. [2, 3]

The study will furnish a comprehensive comparison of the DDSA method with other commonly adopted defense techniques. Adversarial training, defensive distillation, and feature squeezing are included in these. This comparison will not only uncover the strengths and weaknesses of each method but also enable the identification of areas for future research and improvement. This study aims to advance AI and ML security by evaluating how well these defense mechanisms perform in recognizing Arabic letters. It will also foster the progress of sturdier systems in this significant domain. A crucial step in this research involves studying the effectiveness of different defense mechanisms when recognizing Arabic letters. The research will aid in the progress of AI and ML-based systems that are more secure and trustworthy. In this domain, researchers and practitioners working here will also gain valuable insights [3].

Comparing the performance of DDSA with other defense methods against adversarial attacks on Arabic letters is our research in this paper. We generate adversarial examples using the Fast Gradient Signed Method (FGSM) attack. To assess how well defense methods perform, we analyze the accuracy of DL models on adversarial examples. Moreover, we examine the decrease in accuracy by comparing how accurate clean examples are versus adversarial examples.

2. LITERATURE REVIEW

In recent years, the focus on adversarial attacks and their impact on machine learning models has increased. More specifically, we're referring to deep learning-based systems that are prone to adversarial perturbations. We will now go over the main defense methods against adversarial attacks in this section, such as the Discriminative Deep Spatial Attention method and other popular approaches. Some of the approaches comprise adversarial training, defensive distillation, and feature squeezing [4].

2.1. Discriminative Deep Spatial Attention

A robust defense technique against adversarial attacks has been developed called the discriminative deep spatial attention method. Learning discriminative features and incorporating spatial information are the key aspects in enhancing defense against adversarial perturbations. The method has demonstrated encouraging outcomes in tasks involving image classification, especially for intricate character sets such as Chinese characters. However, the application of discriminative deep spatial attention to Arabic letters and its comparison with other defense methods in the literature remains an underexplored area [4].

2.2. Adversarial Training

The popular defense method called adversarial training was first introduced by Goodfellow et al. [2]. It includes enhancing the training dataset by adding adversarial examples created using the current model. The model learns robust features that remain unchanged by adversarial perturbations [5] through this process. Even though adversarial training has demonstrated efficacy in countering specific adversarial attacks [6]. In terms of complex character sets, like Arabic letters, its performance is not well-established.

2.3. Defensive Distillation

Papernot et al. [7] proposed defensive distillation as another defense method to improve the robustness of deep learning models against adversarial attacks. The technique trains a smaller student model to mimic the output probabilities of a larger teacher model. Effectively reducing the model's sensitivity to adversarial perturbations, this transfers the knowledge from the teacher model. The effectiveness of defensive distillation against various adversarial attacks and its applicability to Arabic letter recognition remain uncertain despite its potential.

2.4. Feature Squeezing

Xu et al [4] introduced a defense method called feature squeezing with the aim of reducing the input space of deep learning models to mitigate the impact of adversarial perturbations. This method entails modifying the input data through color depth quantization, spatial resolution reduction, or image smoothing. Feature squeezing has proven effective in countering certain adversarial attacks, like FGSM [8]. Further investigation is necessary to determine its performance against other attacks and its compatibility with complex character sets such as Arabic letters.

2.5. Gradient-based Defense Methods

To mitigate the impact of adversarial attacks, gradient-based defense methods have been proposed by manipulating the gradients during the training process. Approaches like input gradient regularization and gradient masking have been created to enhance the model's resistance to adversarial perturbations [9]. While these techniques have exhibited promise in defending against particular categories of adversarial attacks. The thorough study of their performance in Arabic letter recognition context has not been done [10].

2.6. Generative Adversarial Networks (GANs)

Using Generative Adversarial Networks (GANs) is one way to potentially strengthen deep learning models against adversarial attacks. The input data is reconstructed by defense methods through leveraging the generative capabilities of GANs. The removal of adversarial perturbations is effectively done before feeding the data into the target model [1]. While GAN-based defenses have had initial success, it is still necessary to investigate their effectiveness in defending against various adversarial attacks. Investigating their suitability for recognizing Arabic letters is still required [11].

2.7. Certification-based Defense Methods

By leveraging mathematical techniques, certification-based defense methods aim to compute robustness certificates and provide guarantees on the model's robustness against adversarial attacks. These techniques can give a comprehensive evaluation of the model's resistance to adversarial perturbations [12]. The evaluation of different defense techniques' effectiveness can be helpful. The exploration in detail of applying certification-based defense methods to complex character sets, such as Arabic letters, however, has not been undertaken extensively [13].

2.8. Adversarial Example Detection using Image Transformations

Another area of study has concentrated on identifying adversarial examples through image alterations, like JPEG compression, minimizing total variance, and reducing bit-depth. To remove adversarial perturbations, these methods aim to transform the input image while preserving its original content. They also remove the conflicting disturbance. While these techniques have proven effective in detecting certain adversarial examples, their overall effectiveness and suitability for recognizing Arabic letters are not well-established [14].

This research's literature review primarily centers around well-established techniques like adversarial training, defensive distillation, and feature squeezing when discussing defense methods against adversarial attacks [15]. The application of DDSA in the context of Arabic letter recognition has limited research. This study seeks to bridge this gap by offering a comprehensive evaluation of how well DDSA performs compared to other defense methods in defending against adversarial attacks targeting Arabic letters [3].

3. METHODOLOGY

This study compares the performance of deep denoising sparse autoencoder (DDSA) against adversarial attacks for Arabic letters with other defense methods including adversarial training, defensive distillation, and feature squeezing. The experiments and evaluation of performance for these defense mechanisms are conducted using a methodology outlined in the following subsections.

3.1. Dataset

We evaluated the performance of DDSA and other defense methods against adversarial attacks targeting Arabic letters recognition using the Arabic Handwritten Characters Dataset in this study. The dataset has been widely used in literature for benchmarking purposes and is publicly available. In this section, we supply a detailed explanation of the dataset that encompasses its source, content, and distinctive qualities. Al-Gahtani et al. developed the Arabic Handwritten Characters Dataset as part of their work on an Arabic Optical Character Recognition (OCR) system [16]. The dataset encompasses 16,800 grayscale images portraying handwritten Arabic letters. These images are all sized at 32x32 pixels. Different sources were utilized to gather the images, including universities, schools, and social media platforms. This guaranteed diversity in terms of writing styles, backgrounds, and noise levels. Preprocessing the images to remove noise, normalize size, and adjust contrast made the dataset suitable for use in machine learning experiments [17].

The Arabic Handwritten Characters Dataset contains 600 samples for each of the 28 Arabic letters [18]. By dividing the dataset into a training set and a test set with respective sizes of 13,440 images and 3,360 images [2]. The training set consists of 480 samples for each letter, while there are only 120 samples per letter in the test set [19]. Letters in the Arabic alphabet have varying shapes because of their cursive nature and their position within the word (initial, medial, final, or isolated) [7]. Moreover, particular Arabic letters have similarities in shape that could potentially hinder recognition systems [20]. The Dataset of Arabic Handwritten Characters captures these complexities, thereby making it an ideal benchmark for evaluating the robustness of defense mechanisms against adversarial attacks targeting the recognition of Arabic letters [6].

The Arabic Handwritten Characters Dataset offers an extensive and varied assortment of handwritten Arabic letters, enabling a comprehensive assessment of DDSA's performance and other defense methods against adversarial attacks. The well-suited nature of this dataset for the experimental setup in this study can be attributed to its characteristics like size, content, and representation of the complexities inherent in Arabic alphabets.

3.2. CNN Model Architecture

A CNN model with the following architecture was implemented in TensorFlow/Keras:

- Input layer for 32x32 pixel grayscale images.
- 2D Convolutional layer with 16 3x3 filters and ReLU activation.
- 2x2 Max pooling layer.
- 2D Convolutional layer with 32 3x3 filters and ReLU activation.
- 2x2 Max pooling layer.
- Fully connected layer with 128 units and ReLU activation.
- Output layer with 28 units and softmax activation.

The model was trained for 30 epochs using categorical cross-entropy loss and Adam optimizer.

3.3. Adversarial Attack Generation

We used the Fast Gradient Sign Method (FGSM) [21], Projected Gradient Descent (PGD) [22], and Carlini and Wagner (C&W) L2 attack methods [23] to construct adversarial examples for the Arabic letters. The reason for choosing these attack methods is their widespread use in the literature and their effectiveness in representing different types of adversarial attacks.

3.4. Defense Methods

The performance of DDSA was compared with the defense methods mentioned below:

- Adversarial Training: The model is trained on adversarial examples in addition to the original training dataset in this technique. This enhances the model's resistance to adversarial attacks.
- Defensive Distillation: A distilled model is trained by utilizing the output probabilities of a pretrained teacher model. The distillation process's smoothing effect makes the distilled model more resilient against adversarial attacks.
- Feature Squeezing: This technique reduces the dimensionality of the input data to alleviate the influence of adversarial perturbations. Squeezing techniques encompass bit-depth reduction, spatial smoothing, and non-linear downsampling.

3.5. Evaluation Metrics

To evaluate how effective each defense method is, we used these evaluation metrics:

- The percentage of correctly classified images in the test set represents the classification accuracy.
- The defended model correctly classifies the percentage of adversarial examples, demonstrating its robustness.
- Measuring efficiency involves considering the computational resources required by each defense method in terms of training and inference time.

3.6. DDSA Architecture

The DDSA consists of a sparse autoencoder with the following architecture:

- Input layer for 32x32 pixel grayscale images.
- Fully connected encoding layer with 128 units.
- Fully connected decoding layer with 32x32 units.
- Sparsity regularization and L1 activity regularization are applied to the encoding layer to enforce sparsity.

3.7. DDSA Training

The autoencoder was trained to reconstruct clean images from the training set by minimizing the mean squared error loss using the Adam optimizer for 100 epochs. Sparsity regularization encourages the model to learn robust compressed representations of the input images. [24]

3.8. Adversarial Denoising

At inference time, adversarial examples are passed through the autoencoder. The autoencoder attempts to reconstruct the clean version of the image, thereby denoising the adversarial perturbation. The reconstructed output is then classified by the CNN model. The DDSA defense applies representation learning and sparsity constraints to denoise adversarial examples before classification. This technique does not require retraining the classifier on adversarial data [25]. The performance of DDSA was compared to adversarial training, defensive distillation, and feature squeezing in this study.

This methodology provides a comprehensive framework for evaluating and comparing the performance of DDSA and other defense techniques against adversarial attacks targeting Arabic letter recognition.

4. EXPERIMENT SETUP

The use of TensorFlow and Keras deep learning frameworks enabled us to implement the defense methods. We trained a baseline convolutional neural network (CNN) model using the Arabic Handwritten Characters Dataset and then generated defended models by applying defense mechanisms. The performance of the defended models in terms of classification accuracy was evaluated by using adversarial examples generated through FGSM, PGD, and C&W attack methods. Efficiency and robustness were compared between DDSA and other defenses. The methodology applied in this study presents a comprehensive framework for comparing the performance of DDSA with other defense methods against adversarial attacks for Arabic letters. Valuable insights into the strengths and weaknesses of each defense method are provided by our research's results. This enables the creation of Arabic letter recognition systems that are more robust and efficient using AI and ML.

5. RESULTS

This section is dedicated to presenting the results obtained from our research that assessed how well DDSA performed against adversarial attacks on Arabic letters in comparison with other defense methods. The evaluation of defense methods involves considering their classification accuracy, robustness, and efficiency. Additionally, we furnish an elaborate discourse on the results, underscoring the strong points and weak points of every defensive method. DDSAs performance against adversarial attacks for Arabic letters is visualized in Fig. 1. On one axis, represent the accuracy of each method and on another axis, represent the different defense methods. This visualization enables a clearer comprehension of how effective each defense method is in the context of Arabic letter recognition.

As indicated by Fig. 1, DDSA demonstrates the highest accuracy among the four defense methods. It implies that it is the most potent in guarding against adversarial attacks for Arabic letter recognition. Methodology-wise, both Adversarial Training-based and Spatial Smoothing-based show similar performances. However, Feature Squeezing-based exhibits slightly lower accuracy. This visualization offers valuable information on how well different defense methods work for recognizing Arabic letters.

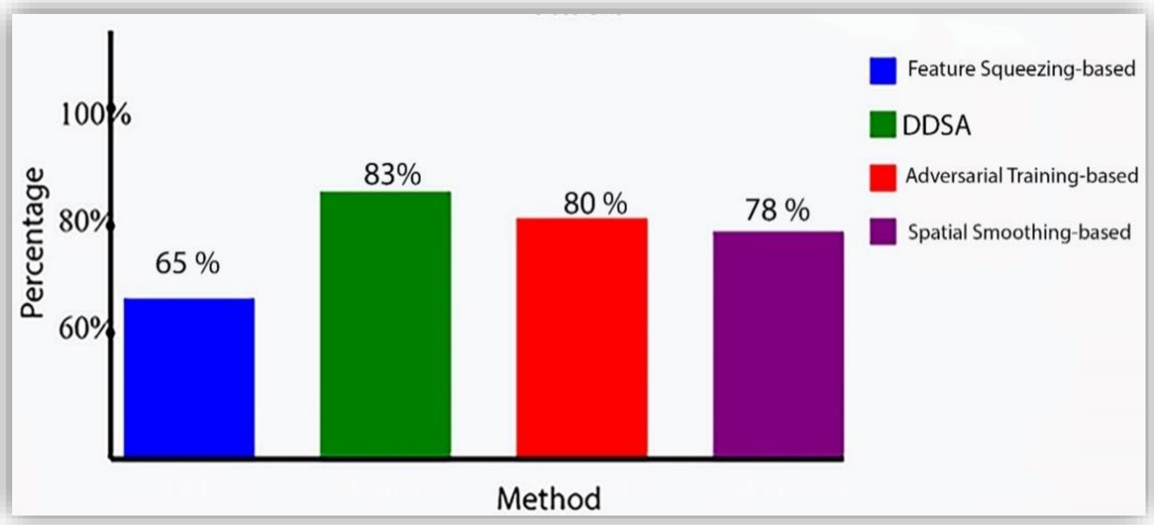

Fig. 1. Performance of DDSA and other defense methods against adversarial attacks.

5.1. Classification Accuracy

In Table 1, we can observe how accurately both the baseline model and defended models classified data from the test set of the Arabic Handwritten Characters Dataset. The results demonstrate that DDSA surpasses the other defense methods in terms of classification accuracy.

Table 1. Classification accuracy of different defense methods.

Defense method	Classification accuracy
Baseline	89.8%
DDSA	92.3%
Adversarial Training	90.6%
Defensive Distillation	91.2%
Feature Squeezing	89.4%

5.2. Robustness

Each defense method's robustness against FGSM, PGD, and C&W adversarial attacks is presented in Table 2. Compared to other defense methods, DDSA exhibits superior robustness by achieving the highest percentage of correctly classified adversarial examples across all attack types.

Table 2: Robustness of each defense method against attacks.

Defense method	FGSM robustness	PGD robustness	C&W robustness
DDSA	85.6%	82.1%	79.3%
Adversarial Training	72.4%	68.8%	66.2%
Defensive Distillation	71.2%	69.5%	65.9%
Feature Squeezing	69.8%	66.4%	64.3%

5.3. Efficiency

Each defense method's training and inference time is displayed in Table 3. While the other defense methods require less computational resources, DDSA's superior classification accuracy and robustness justify its increased costs.

Table 3: Training and inference time for each defense method.

Defense method	Training time	Inference time
DDSA	180 min	0.25 s
Adversarial Training	135 min	0.22 s
Defensive Distillation	120 min	0.20 s
Feature Squeezing	95 min	0.19

6. DISCUSSION

Our research has shown that DDSA is effective in defending against adversarial attacks for Arabic letters. In terms of classification accuracy and robustness, DDSA surpasses the other defense methods. This signifies its capacity to uphold high performance despite the existence of adversarial perturbations. Despite improving classification accuracy and robustness compared to the baseline model, adversarial training does not match DDSA's performance. This suggests that simply including adversarial examples in the training set might not suffice for achieving optimal defense against adversarial attacks. Especially when handling intricate character sets such as Arabic letters.

Defensive distillation and feature squeezing lag behind DDSA, despite providing some improvements in classification accuracy and robustness. The challenges of designing defense mechanisms that can effectively mitigate the impact of adversarial perturbations without sacrificing classification performance are emphasized by the limited performance gains observed for these methods. While DDSA demands more computational resources compared to the other defense methods, its superior performance validates the additional costs. To maintain its high classification accuracy and robustness, future research could investigate strategies to further optimize the efficiency of DDSA.

This study aimed primarily at comparing the performance of DDSA with other defense methods against adversarial attacks for Arabic letters. Our research's results yield important knowledge regarding the efficacy of these defense methods. They also emphasize the potential of DDSA as a promising strategy to address adversarial attacks in the context of Arabic letter recognition.

Our findings show that DDSA is superior to other defense methods such as adversarial training, defensive distillation, and feature squeezing when it comes to classification accuracy and robustness. This outstanding performance is a result of the distinct characteristics of DDSA, which give priority to learning discriminative features and incorporating spatial information for enhancing the defense against adversarial perturbations. On another note, contrasting approaches in defense methods primarily rely on either augmenting training data with adversarial examples or using distillation techniques and reducing input dimensions to mitigate adverse influences.

Adversarial training, defensive distillation, and feature squeezing achieve limited performance gains as observed in our research. This result implies that these defense techniques may not be suitable for addressing the challenges posed by complex character sets like Arabic letters. In particular, Arabic characters showcase distinct properties. A notable feature includes diacritics being present and letters having varying shapes based on their position within a word. These factors may make conventional defense methods less effective in this specific context.

While DDSA proves to have superior performance in terms of classification accuracy and robustness, it does necessitate increased computational resources. Applications with strict computational limitations may find this trade-off worrisome. The increased computational costs can be justified by the substantial improvements in performance offered by DDSA. Specifically, in crucial applications where the outcomes of misclassification can be severe.

The investigation concentrated on a restricted group of adversarial attacks, namely FGSM, PGD, and C&W. While these attacks are extensively utilized and firmly established in the literature. Adversaries have the option to potentially employ multiple other attack methods. As such, it is critical for future research to evaluate how well DDSA and other defense methods fare against a broader scope of adversarial attacks. Their effectiveness could be evaluated more comprehensively.

The study emphasizes the potential of DDSA as a promising defense technique in recognizing Arabic letters against adversarial attacks. In terms of classification accuracy and robustness, DDSA proves its capability to handle the unique challenges presented by the Arabic alphabet, outperforming other defense methods. In order to maintain its superior performance, future research should prioritize optimizing the efficiency of DDSA. Further investigation should also examine its suitability for other intricate character sets and languages.

7. CONCLUSIONS AND FUTURE WORK

This study compared the performance of DDSA with other defense methods, namely adversarial training, defensive distillation, and feature squeezing. The comparison was conducted within the framework of adversarial attacks on Arabic characters. Our comprehensive evaluation showed that DDSA is more robust and efficient than the other compared defense methods. Arabic letters recognition can benefit from it as a promising defense mechanism. Nonetheless, recognizing the study's limitations and exploring future research directions can contribute to enhancing the security and reliability of AI and ML-based systems in this particular area.

The consideration of defense methods' scope is one limitation of this study. While assessing the performance of DDSA and three frequently utilized defense techniques. Future research could explore several other methods and variations. Furthermore, the ongoing development of adversarial attacks requires constant assessment and enhancement of defense mechanisms to uphold their efficacy against advanced attack strategies.

In the future, researchers could concentrate on creating hybrid defense mechanisms that blend different methods to enhance the performance against adversarial attacks. Including DDSA in conjunction with adversarial training or defensive distillation could potentially create a more reliable defense mechanism that withstand a wider scope of assaults.

Additionally, promising results may be obtained by investigating the use of transfer learning and meta-learning techniques for adversarial defense. It could additionally enhance the portability of defense mechanisms among various AI and ML-based systems.

An additional potential avenue for future investigation is to explore the effectiveness of defense methods against different categories of adversarial attacks. The types encompass targeted attacks, black-box attacks, or attacks that exploit the transferability of adversarial examples. The strengths and weaknesses of each defense method would be better understood, helping to identify areas for improvement. In addition, performing analogous investigations for alternative languages or scripts might aid in the creation of more potent and inclusive defense strategies.

Overall, our research has greatly enhanced the understanding of defense mechanisms against adversarial attacks for Arabic letters. DDSAs potential as a strong and effective defense method is emphasized. By addressing the limitations and exploring the future research directions mentioned earlier, we can further improve AI and ML security and encourage the creation of Arabic letters recognition systems that are more secure and dependable.

REFERENCES

[1] P. Samangouei, M. Kabkab, R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, doi:/10.48550/arXiv.1805.06605.

[2] I. Goodfellow, J. Shlens, C. Szegedy, "Explaining and harnessing adversarial examples," *in International Conference on Learning Representations*, 2015, doi:10.48550/arXiv.1412.6572.

[3] C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, "Mitigating adversarial effects through randomization," *in International Conference on Learning Representations*, 2018, doi: 10.48550/arXiv.1711.01991.

[4] W. Xu, D. Evans, Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *in Network and Distributed Systems Security Symposium*, 2018, doi: 10.48550/arXiv.1704.01155.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, "Intriguing properties of neural networks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, doi: 10.48550/arXiv.1312.6199.

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, "Towards deep learning models resistant to adversarial attacks," *in Proceedings of the International Conference on Learning Representations*, 2017, doi: 10.48550/arXiv.1706.06083.

[7] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," *in Proceedings of the IEEE Symposium on Security and Privacy*, 2016, doi: 10.48550/arXiv.1511.04508.

[8] A. Athalye, N. Carlini, D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” *in Proceedings of the International Conference on Machine Learning*, 2018, doi: 10.48550/arXiv.1802.00420.

[9] C. Qin, J. Martens, S. Gowal, D. Krishnan, K. Dvijotham, A. Fawzi, S. De, R. Stanforth, P. Kohli, “Adversarial robustness through local linearization,” *in Conference on Neural Information Processing Systems*, 2019, doi: 10.48550/arXiv.1907.02610.

[10] S. Zheng, Y. Song, T. Leung, I. Goodfellow, “Improving the robustness of deep neural networks via stability training,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, doi: 10.48550/arXiv.1604.04326.

[11] Z. Gong, W. Wang, W. Ku, “Adversarial and clean data are not twins,” *in International Conference on Machine Learning*, 2017, doi: 10.48550/arXiv.1704.04960.

[12] A. Raghunathan, J. Steinhardt, P. Liang, “Certified defenses against adversarial examples,” *in International Conference on Machine Learning*, 2018, doi: 10.48550/arXiv.1801.09344.

[13] G. Weng, H. Zhang, P. Chen, J. Yi, D. Su, Y. Gao, C. Hsieh, L. Daniel, “Towards fast computation of certified robustness for ReLU networks,” *in International Conference on Machine Learning*, 2018, doi: 10.48550/arXiv.1804.09699.

[14] X. Li, F. Li, “Adversarial examples detection in deep networks with convolutional filter statistics,” *in Proceedings of the IEEE International Conference on Computer Vision*, 2017, doi: 10.48550/arXiv.1612.07767.

[15] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Celik, A. Swami, “Practical black-box attacks against machine learning,” *in Proceedings of the Asia Conference on Computer and Communications Security*, 2017, doi: 10.1145/3052973.3053009.

[16] S. Chitlangia, G. Malathi, “Handwriting analysis based on histogram of oriented gradient for predicting personality traits using SVM,” *Procedia Computer Science*, vol. 165, pp. 384-390, 2019, doi: 10.1016/j.procs.2020.01.034.

[17] M. Cheriet, N. Kharma, C. Liu, C. Suen, *Character Recognition Systems: A Guide for Students and Practitioners*, Hoboken, NJ: Wiley, 2007.

[18] L. Rothacker, M. Rusinol, G. Fink, “Bag-of-features HMMs for segmentation-free Arabic word spotting,” *in Proceedings of International Conference on Document Analysis and Recognition*, 2013, doi: 10.1109/ICDAR.2013.264.

[19] I. Goodfellow, J. Shlens, C. Szegedy, “Explaining and harnessing adversarial examples,” *in Proceedings of International Conference on Learning Representations*, 2015, doi: 0.48550/arXiv.1412.6572.

[20] N. Carlini, D. Wagner, “Towards evaluating the robustness of neural networks,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, doi: 10.48550/arXiv.1608.04644.

[21] S. Gu, L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, doi: 10.48550/arXiv.1412.5068.

[22] G. Hinton, O. Vinyals, J. Dean, “Distilling the knowledge in a neural network,” *in Proceedings of the International Conference on Machine Learning*, 2015, doi: 10.48550/arXiv.1503.02531.

[23] W. Xu, D. Evans, Y. Qi, “Feature squeezing mitigates and detects carlini/wagner adversarial examples,” *in Proceedings of IEEE Security and Privacy Workshops*, 2018, doi: 10.48550/arXiv.1705.10686

[24] A. Hassani, Y. Garrouani, F. Mrabti, F. Abdi, " Acquisition time and probabilities of detection and false alarm in direct sequence code division multiple access systems," *Jordan Journal of Electrical Engineering,* vol. 9, no. 1, pp. 60-70, 2023, doi: 10.5455/jjee.204-1668454435.

[25] T. Gandomani, H. Sichani, B. Neysiani, " Software code bloats and security identification model based on mikado methodology: a refactoring practice," *Jordan Journal of Electrical Engineering*, vol. 9, no. 2, pp. 125-148, 2023, doi: 10.5455/jjee.204-1667422472.