Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 3, 2024

Research on Multimodal Generative Adversarial Networks in the Framework of Deep Learning

Ruilin Xu[1, a], Yutian Yang[2, b], Hongjie Qiu[3, c], Xiaoyi Liu[4, d], Jingbo Zhang[5, e]

[1] The University of Chicago, USA
[2] University of California, Davis, USA
[3] University of Washington, USA,
[4] Arizona State University, USA,
[5] Independent Researcher, USA

[a]harveytsui915@gmail.com; [b]yytyang@ucdavis.edu; [c] hongjieq@uw.edu; [d] xliu472@asu.edu; [e]jingbozhangsummer@gmail.com

Abstract: This project aims to align facial and vocal characteristics within a closely related common space through the construction of multi-modal generative adversarial networks (GANs). The project proposes a multi-modal approach grounded in visual perception, utilizing the Graph Cut algorithm to align feature components with the image features of each corresponding local context, thereby achieving adaptability in multi-modal information. To enhance the speed and accuracy of the modeling process, a regional attention strategy is integrated. Experimental results demonstrate that the proposed algorithm enhances the accuracy of image recognition tasks.

Keywords: Image Recognition; Cross-modal; Generate Adversarial Network; Triplet Loss.

1. Introduction

In recent years, scholars have extensively explored the integration of feature and decision layers [1]. Previous studies have predominantly utilized conventional methods of phonetic feature transfer. For example, scholars have conducted a classical correlation analysis between phonetic and auditory features [2]. Other researchers have employed discriminative association analysis to merge sound and sound features, yet such transformation methods risk the loss of information in images [3]. In the realm of decision-maker fusion, algorithms have been applied to obtain face recognition scores, which then inform voice recognition scores. Some scholars introduced the Bayesian algorithm to amalgamate facial and voice features, maintaining data integrity [4]. In response to these challenges, this paper introduces an image-based data fusion algorithm. Deep learning technologies, which have been widely employed in recent years, have shown promise in many areas, such as medical imaging and predictive analysis [5][6][7]. Researchers have successfully utilized CNNs to integrate facial and voice features. The multi-modal short-term memory network has proven capable of extracting information across multiple channels and prioritizing it to enhance speech recognition accuracy [8]. Nonetheless, these methods have not fully exploited the strong correlation possible between channels. There have been suggestions to use CNNs for feature extraction from images and LSTMs for classification [9]. This approach aims to minimize the gap between sound modes in a common space through adversarial means while increasing the distance between different categories within the same scenario via triplet loss, ultimately advancing classification precision [10].

2. Generate adversarial networks across modes

2.1. Generate adversarial networks

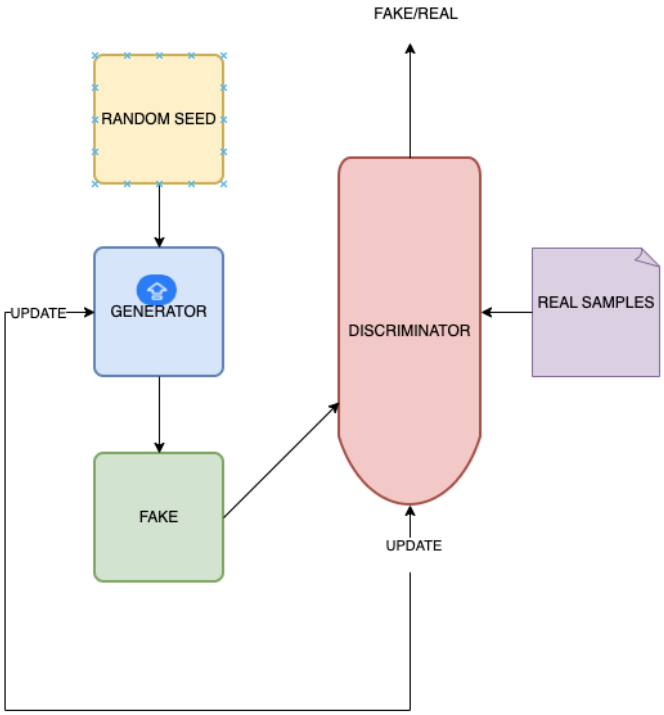

Fig.1 Generating adversarial network architecture

Since generative adversarial networks were introduced in 2014, image recognition, multi-perspective learning, and background elimination have become hot spot in machine vision research [11]. The network generally consists of a generator and separate two components, using a set of codecs to extract an initial signal and reconstruct it into a generated sample [12]. Then, the generated sampling value is used as the negative sampling of the discriminator, and together with

the actual data, it is used to train the discriminator's recognition ability, so as to improve the accuracy of the discriminator and make it able to accurately distinguish the true and false information. Corrected the error score, so that the generated false data as much as possible to confuse identification. This is repeated until the network converges to a plateau, and finally the input and output can produce a distribution close to the actual data. The Generative Adversarial Network architecture is shown in Figure 1.

2.2. Generate adversarial networks across modes

Most of the existing research work is based on a single channel network to establish the transformation between models, so its adaptability is not strong[13]. For the acquisition of multimodal information, CMGAN has been proposed by researchers [14]. In other words, an adversarial training network (CMGAN) is used to establish the sharing characteristics of multi-modal information, and then establish the correlation relationship between various types of heterogeneous information. Figure 2 shows the basic structure for generating adversarial networks across patterns (Image cited in Medical Image Analysis, Volume 52, February 2019, Pages 174-184).

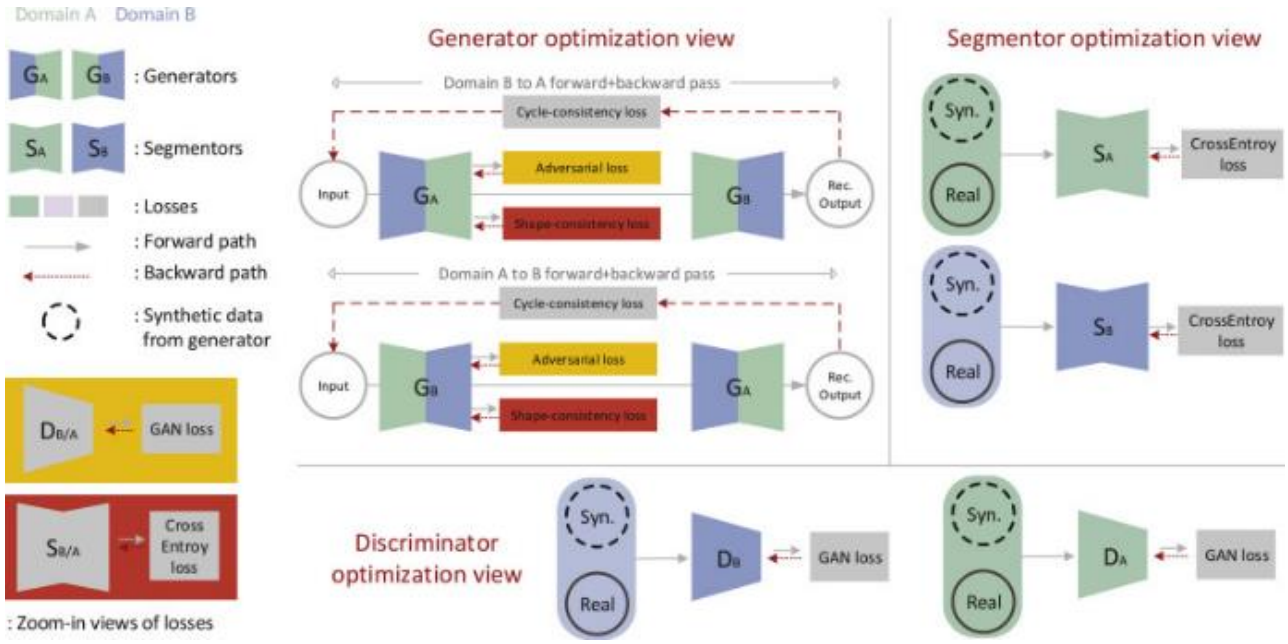

Fig. 2 Generate adversarial networks across modes

The network builds a generator for each image and text, and at the encoder end of the generator, two fully connected layers are added to each mode, s_p^t and s_p^i the last layer parameters are shared to extract common features [15]. Among them, the original features h_p^t and h_p^i samples extracted from the generator are used as real data, and the reconstructed outputs of the decoder in the generator are r_p^t and r_p^i respectively. In the discriminator section, the network is an inter-modal discriminator D_{Ci}, D_{Ct} and an intra-modal discriminator D_I, D_T. Therefore, the objective function of the network also becomes as shown in equation (1)

$$\max_{D_I,\ D_T,\ D_{Ci},\ D_{Ct}} \mathcal{L}_{\mathrm{GAN}_1}(G_I, G_T, D_I, D_T) = \mathcal{L}_{\mathrm{GAN}_2}(G_I, G_T, D_{Ci}, D_{Ct}) \quad (1)$$

Where, $\mathcal{L}_{\mathrm{GAN}_1}$ and $\mathcal{L}_{\mathrm{GAN}_2}$ are GAN objective functions within modes and GAN objective functions between modes in multimodal networks, respectively, and the specific formulas are

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}_1} = {} & E_{i\sim p_i}[D_I(i) - D_I(G_{\mathrm{Idec}}(i))] \\ & + E_{t\sim p_t}[D_T(t) - D_T(G_{\mathrm{Tdec}}(t))] \\ \mathcal{L}_{\mathrm{GAN}_2} = {} & E_{i,\ t\sim p_{i,\ t}}\Big[D_{Ci}(G_{\mathrm{Tenc}}(i)) - \tfrac{1}{2}D_{Ci}(G_{\mathrm{Tenc}}(t)) \\ & - \tfrac{1}{2}D_{Ci}(G_{\mathrm{Tenc}}(\hat{\imath})) + D_{Ct}(G_{\mathrm{Tenc}}(t)) \\ & - \tfrac{1}{2}D_{Ct}(G_{\mathrm{Tenc}}(i)) - \tfrac{1}{2}D_{Ct}(G_{\mathrm{Tenc}}(\hat{t}))\end{aligned} \quad (2)$$

Where, i and $t$are the original picture samples and text samples respectively, $\hat{\imath}$ ten and $\hat{t}$ ten represent the mismatched samples belonging to different categories, G_{Idec} and G_{Tdec} are reconstruction features, G_{Tenc} and G_{Tenc} generator features. Because the cosine distance and voting mode between multiple sampling points are adopted, the method is more suitable for information extraction of multiple modes, but it cannot give two different types of information intuitively, so this paper applies the improved algorithm to speech person recognition.

3. Experiment and result analysis

3.1. Network Settings

The project selected a set of open-source video libraries, the facial expressions and voice information in the video are from the TV series "The Big Bang Theory", and the five main characters in the drama Sheldon, Leonard, Howard, Raj, Penny, and so on. In this paper, a deep convolutional neural network (CNN), represented by a 53x49 deep Convolutional neural network based on a time window, is proposed to process the face information of the same person in the time window in parallel to obtain 25x49 MFCC features. In addition, the mismatching samples of face and voice are added under the same ID by combining the fixed samples of 3-tuple with the positive samples and the negative samples based on generating the samples with the same ID of face and voice. The Linked Data approach is used to harmonize different data formats in the dataset, which is essential for

academic research[16]. This structured technique allows researchers to link and cross-check information, promoting interoperability across multiple datasets. This feature is particularly beneficial in machine learning and artificial intelligence, where reliable data is vital for training models and achieving precise outcomes. The final experiment will use 25,000 matched and 25,000 unmatched samples respectively, and each training set will be divided into 40,000 training and 10,000 test samples.

3.2. Parameter analysis

Of these, 30,000 samples in the first phase were used to train the adversarial network, and 40,000 samples were used to train a marker and a measurement learning network. In the learning process of GAN, the discriminator is updated each time, and the generating function is updated 5 times each time. In addition, the BN layer is added after each layer of the network. In the experiment, through the forward transmission of the sample, the weighted average of all the connected levels of the two modes at the final level is carried out. When the distance between the two modes exceeds a certain critical value, it means that the face and voice of the sample are the same. When it is determined to be the same individual, the output value of its identification network is used as the estimated identification value for this sample. In terms of model selection, this paper adopted a large number of tests to determine the parameters of the model, and obtained the ROC curve of pattern matching and identification (Figure 3).

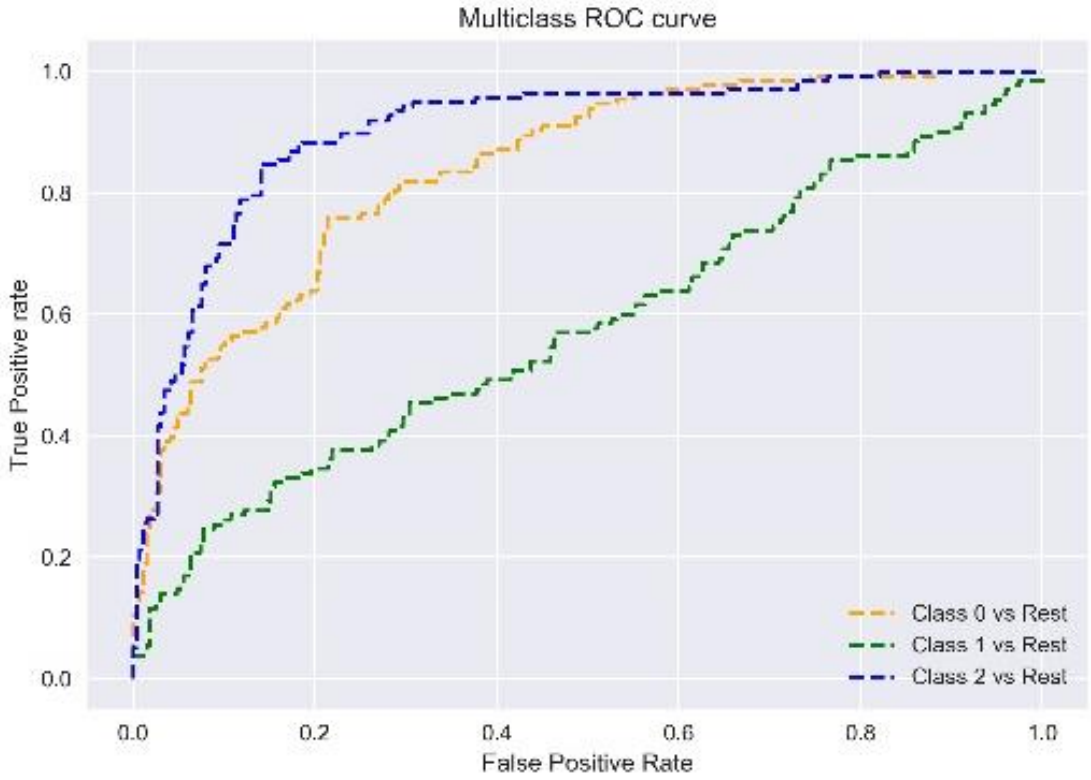

Fig. 3 ROC of different margin values

Here, the pseudo-ratio refers to the proportion of the sample judged to be a match to the sample that is not a match, unlike the general ROC curve, where the true ratio represents the proportion of the ID accurately determined in the matching sample. It can be seen from Figure 3 that when the margin is 0.2, the AUC region of the corresponding ROC curve is the largest, so the margin is finally set to 0.2. In terms of threshold selection, this paper compares the identification effect of multiple indices with multiple thresholds in the test set. Through the analysis of the experimental results, it is found that the method has the best effect when the threshold value is set to 0.5, so 0.5 is chosen as the experimental threshold value in this paper. This paper also finds that after training with other margin values, the total performance reaches a maximum value around 0.5.

3.3. Module necessity analysis and feature selection

3.3.1. Necessity analysis of the public layer

By comparing the fusion effect of the two modes in different modes, the ROC curves of different modes are obtained, which proves the necessity of constructing the common space of acoust-shadow modes in multi-modes. As shown in Figure 4, after removing the common layer, the authentication and decision performance of the network is greatly reduced. This is mainly due to the fact that without a shared layer, the characteristics of the two patterns are very different, so the distance limit of the 3-tuple cannot be used to build connections between patterns.

3.3.2. Network necessity analysis of feature matching

In view of the fact that the forward adversarial neural network can establish the common space position between the two models, this project intends to study the classification method based on the three classification methods by combining the feature selection based on the common level and the feature-based classification and discrimination based on the attribute. The results of the test for selecting the common layer characteristics are shown in Figure 5. The matching accuracy shown in FIG. 5 refers to the number of samples that have been correctly judged for conforming samples and the percentage in the whole test sample. The ID identification accuracy is also the real ratio mentioned above. Under the condition of full threshold, the experimental effect

of non-feature matching decision network is better than that of feature matching decision network, but its accuracy is above 50%. The experimental results show that when the feature matching network is not used, the two samples to be matched and the two samples to be matched are indistinguishable in cosine distance, and they are confused with each other.

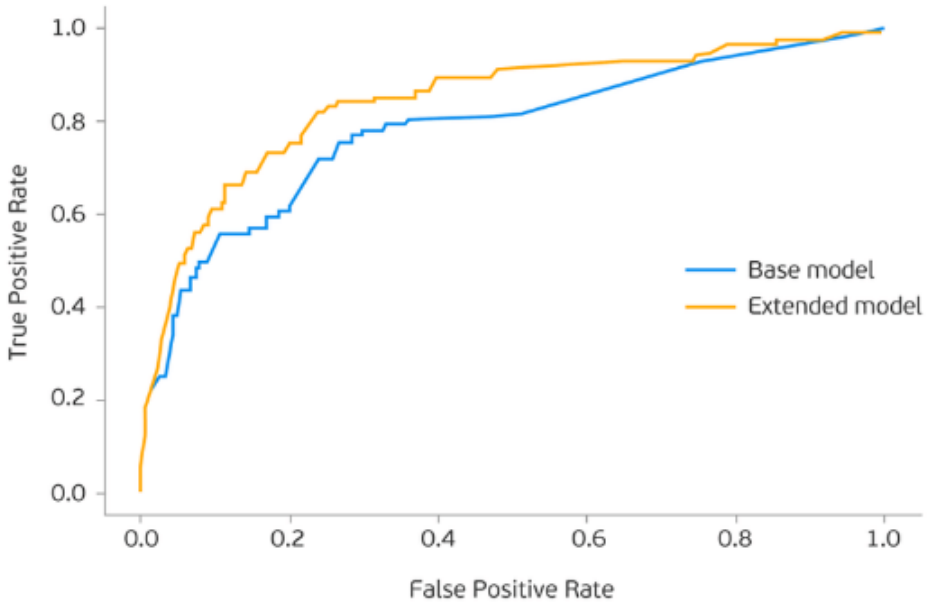

Fig. 4 Comparison of ROC curves with or without a common layer

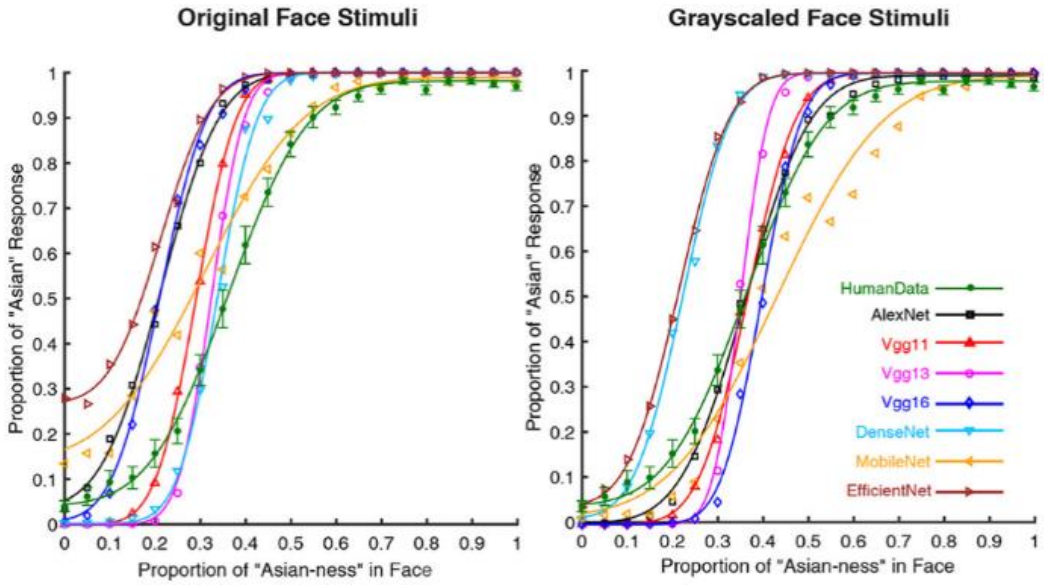

Fig. 5 Comparison of the recognition results of the network with or without feature matching

The results show that the cosine distance of the eigenvectors tends to be 1 regardless of whether the two modes are matched or not, without the feature matching decision. Therefore, this experiment proves that the feature-matching decision network can reduce the spacing of each channel in the mismatched sample, so as to significantly improve the discriminant ability of each channel. When the serial feature is used, the recognition effect is better than that of common facial features because the information on speech patterns is added to the feature to improve the discrimination ability. Among them, joint learning based on a multi-modal adversarial network makes the similarity between the two modes closer and integrates the characteristics of speech into learning [17]. The method proposed in this paper can significantly improve the classification recognition ability, that is, the adversarial network is used to uniformly distribute the features of the two models, and the distance between the classes and the classes is shortened by the ternary loss method, so as to improve the classification discrimination ability [18]. Compared with predecessors, the accuracy of this method is obviously improved.

4. Conclusion

The research delineated in this paper encapsulates a novel approach to multimodal style transfer, which innovatively employs the Graph Cut algorithm alongside a specialized style loss function that inherently possesses multimodal characteristics. This blend of techniques offers a significant advance in optimizing network parameters for enhanced image recognition accuracy. The incorporation of a "mixed

region" attention mechanism stands out as a pivotal enhancement within the proposed algorithm. This mechanism ensures a nuanced and comprehensive reflection of the object's characteristics, fostering a more precise alignment of facial and vocal features. Such precision underpins the broader objective of this work: to reduce the dissonance between facial and vocal data within a shared representational space, thereby enhancing the cohesion of the multimodal information synthesis. This work lays a robust foundation for the next generation of multimodal recognition systems, promising enhancements in both the user experience and the technological capabilities of these systems.

References

[1] Liu, Z., Yang, Y., Pan, Z., Sharma, A., Hasan, A., Ding, C., ... & Geng, T. (2023, July). Ising-cf: A pathbreaking collaborative filtering method through efficient ising machine learning. In 2023 60th ACM/IEEE Design Automation Conference (DAC) (pp. 1-6). IEEE.

[2] Wang, X. S., Turner, J. D., & Mann, B. P. (2021). Constrained attractor selection using deep reinforcement learning. Journal of Vibration and Control, 27(5-6), 502-514.

[3] Xin Chen , Yuxiang Hu, Ting Xu, Haowei Yang, Tong Wu. (2024). Advancements in AI for Oncology: Developing an Enhanced YOLOv5-based Cancer Cell Detection System. International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12(2),75-80, doi:10.55524/ijircst.2024.12.2.13.

[4] Yao, J., Wu, T., & Zhang, X. (2023). Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn. arXiv preprint arXiv:2308.08333.

[5] Yan, X., Wang, W., Xiao, M., Li, Y., & Gao, M. (2024). Survival prediction across diverse cancer types using neural networks. doi:10.48550/ARXIV.2404.08713

[6] Yulu Gong , Haoxin Zhang, Ruilin Xu, Zhou Yu, Jingbo Zhang. (2024). Innovative Deep Learning Methods for Precancerous Lesion Detection. International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12(2),81-86, doi:10.55524/ijircst.2024.12.2.14.

[7] Yan, C., Qiu, Y., Zhu, Y. (2021). Predict Oil Production with LSTM Neural Network. In: Liu, Q., Liu, X., Li, L., Zhou, H., Zhao, HH. (eds) Proceedings of the 9th International Conference on Computer Engineering and Networks . Advances in Intelligent Systems and Computing, vol 1143. Springer, Singapore. https://doi.org/10.1007/978-981-15-3753-0_34.

[8] Hu, Z., Li, J., Pan, Z., Zhou, S., Yang, L., Ding, C., ... & Jiang, W. (2022, October). On the design of quantum graph convolutional neural network in the nisq-era and beyond. In 2022 IEEE 40th International Conference on Computer Design (ICCD) (pp. 290-297). IEEE.

[9] Dai, W., Tao, J., Yan, X., Feng, Z., & Chen, J. (2023, November). Addressing Unintended Bias in Toxicity Detection: An LSTM and Attention-Based Approach. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 375-379). IEEE.

[10] Wang, X. S., & Mann, B. P. (2020). Attractor Selection in Nonlinear Energy Harvesting Using Deep Reinforcement Learning. arXiv preprint arXiv:2010.01255.

[11] Li, S., Kou, P., Ma, M., Yang, H., Huang, S., & Yang, Z. (2024). Application of Semi-supervised Learning in Image Classification: Research on Fusion of Labeled and Unlabeled Data. IEEE Access.

[12] Xiao, M., Li, Y., Yan, X., Gao, M., & Wang, W. (2024). Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. doi:10.48550/ARXIV.2404.08279

[13] Liu, Y., Yang, H., & Wu, C. (2023). Unveiling patterns: A study on semi-supervised classification of strip surface defects. IEEE Access, 11, 119933-119946.

[14] Abdulatif, S., Cao, R., & Yang, B. (2022). CMGAN: Conformer-based metric-GAN for monaural speech enhancement. arXiv preprint arXiv:2209.11112.

[15] Guo, A., Hao, Y., Wu, C., Haghi, P., Pan, Z., Si, M., ... & Geng, T. (2023, June). Software-hardware co-design of heterogeneous SmartNIC system for recommendation models inference and training. In Proceedings of the 37th International Conference on Supercomputing (pp. 336-347).

[16] Li, Y., Yan, X., Xiao, M., Wang, W., & Zhang, F. (2024). Investigation of Creating Accessibility Linked Data Based on Publicly Available Accessibility Datasets. In Proceedings of the 2023 13th International Conference on Communication and Network Security (pp. 77–81). Association for Computing Machinery.

[17] Zi, Y., Wang, Q., Gao, Z., Cheng, X., & Mei, T. (2024). Research on the Application of Deep Learning in Medical Image Segmentation and 3D Reconstruction. Academic Journal of Science and Technology, 10(2), 8-12.

[18] Foody, G. M., & Arora, M. K. (1997). An evaluation of some factors affecting the accuracy of classification by an artificial neural network. International Journal of Remote Sensing, 18(4), 799-810.