2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) | 979-8-3503-8354-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC2PCT60090.2024.10486235

Adversarial Attack and Defense Mechanisms in Medical Imaging: A Comprehensive Review

Surekha M
Computer Science and Engineering
Sharda University
Greater Noida, Uttar Pradesh, India.
0000-0002-7338-8267
2021357873.surekha@dr.sharda.ac.in

Anil Kumar Sagar
Computer Science and Engineering
Sharda University
Greater Noida, Uttar Pradesh, India
0000-0002-5991-2835
anil.sagar@sharda.ac.in

Vineeta Khemchandani
Computer Science and Engineering
Galgotias University
Greater Noida, Uttar Pradesh, India
0000-0002-7934-8493
vineetakh05@gmail.com

Abstract— Medical imaging is essential in modern healthcare, allowing for accurate diagnoses and treatment planning. The practice of machine learning algorithms into clinical imaging uses has significantly enhanced diagnostic abilities. In contrast, as technology has forward-looking, new problems have emerged, main among them existence adversarial assaults. This complete research study investigates the background information of adversarial attacks and its rising defense measures on medical image model. Adversarial assaults involve intelligent alterations of input data to change machine learning algorithms, possibly important to misdiagnosis or inaccurate clinical diagnosis. This paper thoroughly analyses many adversarial assault practices precisely tailored for medical imaging based on division, classification and rebuilding of clinical image with different defensive tactic. Also, the research study provides the better path for collect rising counter measures on adversarial attacks posed by threats. These protective measures contain different types of proposals. For example, adversarial practices, input pre-processing and algorithm understandable strategies.

Keywords—Adversarial attack, adversarial defense, medical image analysis, taxonomy, evaluation.

I. INTRODUCTION

Usage of DNNs model in medical image processing is a transformative period in healthcare and Bringing notable prospects. With implementation of Artificial Intelligence (AI), these progressive models have confirmed to be special skilled at analysing difficult medical pictures [10], providing researchers and doctors with an innovative approach. Though, an important challenge that these deep learning systems look their susceptibility to adversarial attacks, sagging the way for a detailed investigation of adversarial ML approaches. Despite the fact deep learning has shown abundant potential in a lot of areas, DNNs experience an important issue for the reason that of their vulnerability to adversarial examples [26]. Particularly in safety-critical applications like clinical image analysis [2], [3] these opposing examples can introduce slight, nearly hidden modifications to valid examples, positively discounting manual verification.

The presence of adversarial samples pretence an important hurdle to the well-organized implementation of DNNs in clinical diagnosis systems, which might outcome in misdiagnosis, fake insurance claims, and a drop in belief for AI-powered medical technology. Computer-aided diagnosis models are susceptible to adversarial assaults in a variety of situations, containing semi-white-box, black-box, white-box, and [2], [3], and [4], [5] as confirmed by latest studies. These aggressive inputs, which are designed with smart disturbances, have the possible to complicate AI algorithms, leading to improper forecasts and risking the accuracy of diagnostics.

The implementation of AI driven answers in dangerous healthcare setups is severely vulnerable by these weaknesses, demanding the development of strong counter techniques to decrease the threats.

This complete study means to examine the state of adversarial machine learning theories and their consequences for DNN based medical image processing.

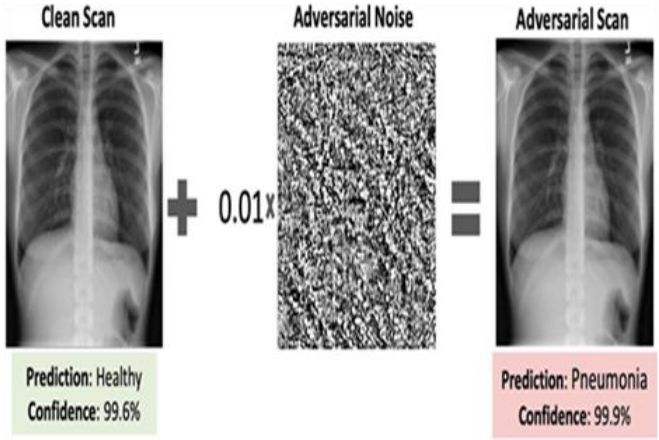

Fig. 1. Adversarial Attack (FGSM) on a medical image

II. BACKGROUND

Deep learning has been widely applied in a variety of disciplines in recent years. With the healthcare division, where it plays an essential role in medication improvement, medical decision making and innovative medical approaches. Medical imaging, encompassing modalities like X-rays, ultrasound, CT scans, PET scans, and MRI, is fundamental for computerized clinical outcomes [6], [26].

A. Medical Adversarial Attack

Medical adversarial assaults involve the purposeful changes of medical imaging data in order to fool machine learning algorithms used in medical image analysis. These assaults aim to exploit vulnerabilities in the algorithms and compromise the accuracy and untrustworthiness of diagnostic or analytical results. Adversarial attacks in medical picturing can have serious intimations, as they can cause wrong diagnoses, treatment ideas, or research findings.

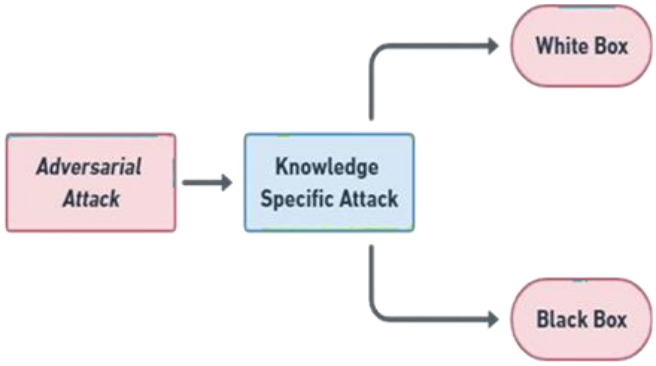

Fig. 2. Taxonomy of Adversarial Attacks

1) White-Box Attacks

"White box attacks" is a type of attack that is leading in adversarial machine learning. These attacks are illustrious by the assaulters' thorough empathetic of the parameters, exercise set, and architecture distinctive to the machine learning algorithm that is being assaulted. This deep information gives assaulters the capability to create difficult, tailored adversarial assaults that attempt to take benefit of weaknesses and make available incorrect outputs or misclassified data. Paschali et al. [7] were the first to widely examine the white-box attack against a range of medical picturing models of many medical tasks, including skin lesion classification and whole brain division, to the best of our knowledge.

2) Black-Box Attacks

An adversarial attack known as a "black-box attack" occurs when the attacker has little no access to the underlying workings of the machine learning model that is being attacked. The attacker in a black-box attack scenario is not aware of the parameters, training data, or architecture of the model. The attacker's goal, despite these drawbacks, is to produce adversarial instances that can trick the model into generating false predictions. Medical imaging models, which are frequently confidential and protected for patient privacy, attract the attention of attackers who want to tamper with diagnostic results without knowledge of the model's internal workings of the model or its training data.

In medical imaging, black-box attack usually involves iterative relation with the algorithms, producing adversarial examples using query-based strategies. If effective, these assaults might seriously make vulnerable patient health by Cooperating the accuracy of diagnosis and treatment plans.

Corresponding to the study [30], even when two models are trained with distinct datasets, adversarial methods designed to cause misclassification in one model can easily mislead another machine. The authors used this concept [9] to take on black-box assaults on a Deep Neural Network so as to train an substitute model on artificial inputs wrong labelled by the target DNN. The replacement model was then used to create adversarial examples capable of undermining the target DNN.

Paschali and Conjeti [10] examined the fine balance of general applicability and heftiness, emphasising the vital features in refining both wings. [11] Ilyas, Engstrom, made considerable hand-outs by examining black-box adversarial attacks, put on bandit based strategies, and exploiting past information to target machine learning algorithms, which is especially significant when internal details are restricted.

Byra et al. [12] in this study attentive on adversarial assaults on deep learning algorithms engaged in fatty liver disease classification. Ultrasound image rebuilding methods alterations dedicated by their research, offering information on the influence of these alterations on algorithm achievements. Chen et al.[14] publicized an in deepness examination on the adversarial bounciness of CNN in the framework of MR based (lumbar disc)shape renovation, providing valuable visions for uses in medical picture processing and reconstruction. Lastly, P. R.BMS, Anusree [13] conduct an investigation of black box attack on medical image classification models, examining susceptibility and considering real life robustness consequences.

B. Medical adversarial Defenses

In General, it denotes to the techniques and methods used to protect medical imaging systems particularly deep learning algorithms against aggressive assaults. In an adversarial attack, input data is deliberately influenced to trick a machine learning model and give wrong classifications or predictions. The main aim of medical defensive technique is to safeguard the accuracy and integrity of medical imaging. By reducing the effect of different attacks, these defense techniques hope to develop the general security and dependability of medical imaging models.

C. Adversarial Training

Adversarial exercise is a machine learning tactic that progresses model bounciness counter to adversarial attacks. Adversarial attacks involve making slight, cautiously prepared alterations to input data to trick a algorithm, resulting in inaccurate predictions. Adversarial training tries to rise the resilience of machine learning models against such attacks by showing them to adversarial examples throughout the training process.

A significant portion of medical adversarial defense work has focused on the use of adversarial training to improve the resilience of diagnostic systems [15]. Many research have adapted classic adversarial training approaches built for natural images to handle the problems given by medical image classification tasks. Vatian et al. [16] investigated the

context of adverse instances in medical imaging, evaluating several techniques to protect against these potentially malevolent examples. Furthermore, these defense techniques aim to improve and alter pre-existing adversarial training methods, which were originally developed for standard pictures, for the specific area of medical image analysis.

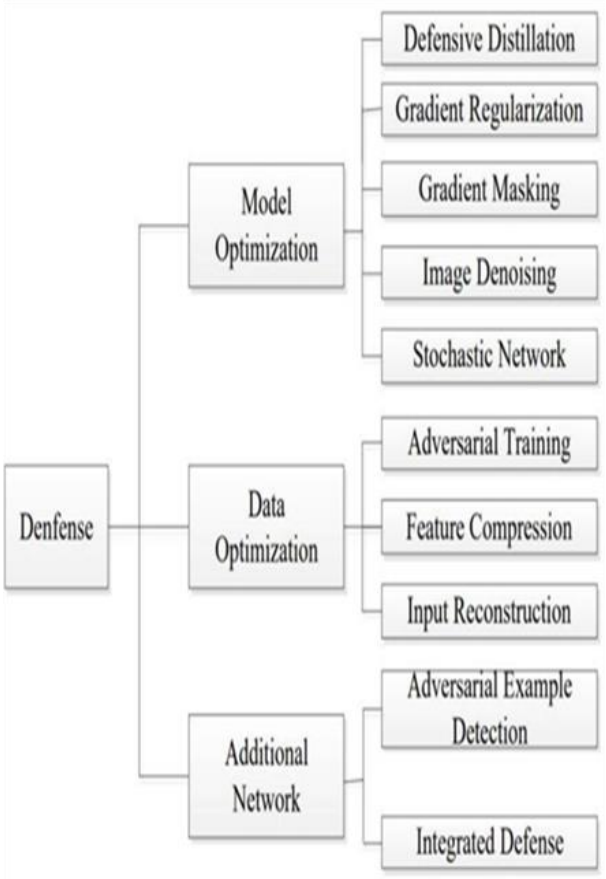

Fig. 3. Taxonomy of Adversarial Defenses [21]

D. Image-level Pre-Processing

1) Image-level preprocessing refers to the approaches

Used to change or improve an entire image before it's introduced as input into the machine learning model. This pre-processing phase is critical for increasing the quality of input data and permitting improved model performance. Adversarial pictures are generally made up of an initial clean image and an accompanying adversarial disturbance. Deep Neural Networks (DNNs) perform well on clean pictures but are subject to adversarial instances. Therefore, denoising the adversarial example and eliminating the perturbation can improve subsequent network diagnostics. Importantly, picture-level preprocessing for denoising does not necessitate retraining or change of medical models, assuring simplicity and safety in biomedical image analysis. Numerous image-level preprocessing algorithms have been proposed in this area to safeguard computer-aided diagnostic systems against adversarial cases [17].

TABLE I. RELATED WORK

TITLE	MERIT	LOSS
[31] Dong, J., (2023) Adversarial attack and defense for medical image analysis: Methods and application	This survey also incorporates a unified theoretical framework with a comprehensive analysis of different types of attack and defense methods in the context of medical images.	There is no Architecture to fulfill required challenging gap.
[22] Comprehensive Review by Muoka, G. W., Yi,[2023]	Detailed study on all recent attacks and defenses	Not use Unet architecture to segment the attacked image.
[32] S. Kaviani et al.,[2022] "Adversarial attacks and defenses on AI in medical imaging informatics	Proposed adversarial attack methods to medical imaging DNNs and defense techniques against these attacks.	Attack and defense of Medical Image Analysis discussed only about classification and segmentations.
[21] Auto encoder Yadav, A., Upadhyay [2022]	Defend both Black box & White box attack.	Not work well with medical images.
[20] A Variational Autoencoder to Purify Adversarial Examples [2019]	Takes only0.114 second to Process.	Only filter Limited examples.
[19] Jang, H. and Yoon [2019]	Reduced the Level of White Box attacks from 100% to 13%	Unable to defend every kind of attacks.
[18] Su, J. and Vargas [2017]	Less Adversarial information Needed.	For high quality images, many pixel perturbations are necessary.

2) Feature Enhancement

Feature enhancement is a method or technique that enhances the visibility, clarity, or prominence of certain characteristics in an image or dataset. Feature enhancement methods are frequently used in the context of medical imaging or computer-aided diagnosis to emphasize key structures, anomalies, or features that may benefit healthcare practitioners in making correct diagnoses. Feature enhancement methods are critical in boosting the visibility of anatomical structures or pathological situations, hence assisting radiologists and healthcare professionals in their analysis. The qualities of the healthcare images and the diagnostic criteria influence the selection of a certain feature augmentation approach.

To increase the resilience of medical classification models, a number of solutions have been proposed. To minimize the size of feature maps, pooling layers are commonly utilized in neural network architectures. Max-pooling layers in medical classification networks were replaced with average- pooling layers in a change presented by Taghanaki et al.[22], resulting in a considerable increase in robustness against adversarial samples across various network designs.

Autoencoders (AE) are being used in computer-aided diagnostic application, namely for feature-level denoising. This method acts independently of image-level pre-processing and includes feature invariance advice to decrease model sensitivity to adversaries. In accordance with, Han et al.[24] introduced dual-batch normalization in adversarial training, resulting in a significant gain in diagnostic model robustness without affecting clean accuracy.

III. EXPERIMENTATION

This section provides an overview of our practical analysis of adversarial attacks and defenses learning in the context of medical image reconstruction and classification. We begin by introducing the datasets used and outlining the data pre-processing techniques applied.

A. Datasets and Models

TABLE II. SUMMARY OF MODELS USED IN STUDY

S.NO.	Model	Description
1.	MobileNetV2	CNN architecture designed for mobile a edge devices with resource constraints.
2.	ResNet50	Architecture known for its use of residual blocks. It has 50 layers.

TABLE III. SUMMARY OF DATASETS USED IN STUDY

S.NO.	Datasets	Description
1.	ChestX-ray14	Collection of over 100k+ frontal-view chest X-rays linked to 14 different thoracic conditions.
2.	Messidor Dataset	1,200 color numerical pictures of the eye fundus are used to identify retinopathy.
3.	Imagenet	It consists of a vast number of images across a 1000 of categories around Half million

A
"
A
in
F
a
a

The proposed defensive model has the following characteristics:

- Base Feature extraction using model ResNet50 or MobilenetV2.
- Implement Auto-encoder [18] for Dimensionality reduction and image reconstruction.
- Use randomization for robustness.
- Integrated U-Net [28] in architecture for Segmentation of image.
- Defend against vulnerable Attacks, Using two coupled models.
- Using Grad-CAM [25], highlight relevant regions.

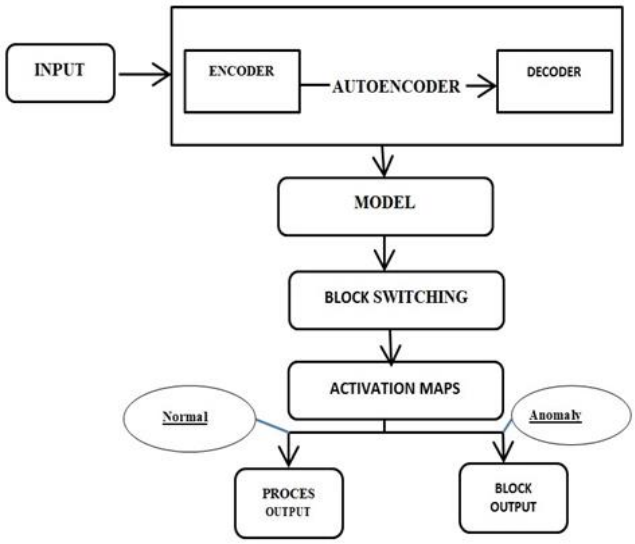

Fig. 4. Auto Encoder-Block Switching

V. CONCLUSION

In this comprehensive review, we have meticulously explored the intricate landscape of knowledge specific attacks and their defense mechanisms within the area of medical imaging and their application. By leveraging powerful architectures like ResNet-50, MobileNetV2, Auto Encoder-Block Switching, Grad-Cam and U-Net, we delved into both the strategies of attack and defense, emphasizing the prevailing "norm threat model." Our rigorous evaluation encompassed the application of potent adversarial attack techniques, showcasing the robustness of our proposed defense strategies across various perturbation ranges. The study underscores the paramount importance of addressing adversarial vulnerabilities in medical imaging models given their pivotal role in critical aspects of healthcare, such as diagnosis and treatment planning. Our findings emphasize the necessity for holistic defense mechanisms, seamlessly integrating image-level preprocessing, feature enhancement, and robust model architectures. This multifaceted approach serves as a robust defense against adversarial intrusions.

Concluded both qualitative and quantifiable evaluates, this work pays valuable visions that cover out there a simple understanding of susceptibilities. We present a directional plan for future research actions, inspiring the improvement of more safe, reliable, and clinically appreciated deep learning systems personalized specially for medical picture examination.

As the field advances, the constant enhancement and novelty of defense strategies become authoritative. This review gives strong foundation, motivating researchers to pioneer revolutionary results that not only strengthen in contradiction of adversarial threats but also increase the overall security and reliability of medical image exploration

systems. This collective determination holds the assurance of advancing healthcare technology, confirming the uprightness exactness of diagnoses in the ever evolving section of medical picturing.

REFERENCES

[1] Z. Wang et al., "Deep learning for image superresolution: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3365-3387, 2021.

[2] K. Muhammad et al., "Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, pp. 507-522, 2020.

[3] P. Bountakas et al., "Defense strategies for Adversarial Machine Learning: A survey," Comput. Sci. Rev., vol. 49, p. 100573, 2023.

[4] S. G. Finlayson et al., "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287-1289, 2019.

[5] K. D. Apostolidis and G. A. Papakostas, "Digital watermarking as an adversarial attack on medical image analysis with deep learning," J. Imaging, vol. 8, no. 6, p. 155, 2022.

[6] T. V. Maliamanis et al. (2022). How Resilient Are Deep Learning Models in Medical Image Analysis? The Case of the Moment- Based Adversarial Attack (Mb-AdA). Biomedicines, 10(10), 2545.

[7] M. Paschali et al., "Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples" in Med. Image Comput. Comput. Assist. Interv. MICCAI, Proc. Part I: 21st International Conference, Granada, Spain, September 16-20, 2018, vol. 2018. Springer International Publishing, 2018, pp. 493-501.

[8] M. Zhang et al., "Fooling examples: Another intriguing property of neural networks," Sensors (Basel), vol. 23, no. 14, p. 6378, 2023.

[9] N. Papernot et al., "Practical black- box attacks against machine learning" in Proc. 2017 ACM on Asia Conference on Computer and Communications Security, 2017, Apr., pp. 506-519.

[10] M. Paschali et al., "Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples" in Med. Image Comput. Comput. Assist. Interv. MICCAI, Proc. Part I: 21st International Conference, Granada, Spain, September 16-20, 2018, vol. 2018. Springer International Publishing, 2018, pp. 493-501.

[11] A. Ilyas et al., 2018, Prior convictions: Black-box adversarial attacks with bandits and priors. arXiv preprint arXiv:1807.07978.

[12] M. Byra et al., "Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method" in IEEE International Ultrasonics Symposium (IUS), vol. 2020. IEEE, 2020, Sept., pp. 1-4.

[13] BMS et al., "Analysis of the effect of black box adversarial attacks on medical image classification models" in Third International Conference on Intelligent Computing Instrumentation and Control. Technologies (ICICICT), vol. 2022. IEEE, 2022, Aug., pp. 528-531.

[14] C. Chen et al., "Realistic adversarial data augmentation for MR image segmentation" in Med. Image Comput. Comput. Assist. Interv. MICCAI, Proc. Part I 23: 23rd International Conference, Lima, Peru, October 4-8, 2020, vol. 2020. Springer International Publishing, 2020, pp. 667-677.

[15] L. Hu et al., "Adversarial training for prostate cancer classification using magnetic resonance imaging," Quant. Imaging Med. Surg., vol. 12, no. 6, p. 3276-3287, 2022.

[16] A. Vatian et al., "Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images" in 24th Conference of Open Innovations Association (FRUCT), vol. 2019. IEEE, 2019, Apr., pp. 472-478.

[17] K. Kansal et al., "Defending against adversarial attacks on Covid-19 classifier: A denoiser- based approach," Heliyon, vol. 8, no. 10, e11209, 2022.

[18] J. Su et al., "One pixel attack for fooling deep neural networks," IEEE Trans. Evol. Comput., vol. 23, no. 5, pp. 828-841, 2019.

[19] U. Hwang et al., "Puvae: A variational autoencoder to purify adversarial examples," IEEE Access, vol. 7, pp. 126582-126593, 2019.

[20] A. Yadav et al., 2022, An integrated Auto Encoder-Block switching defense approach to prevent adversarial attacks. arXiv preprint arXiv:2203.10930.

[21] G. W. Muoka, ... et al., "A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense" Mathematics, vol. 11, no. 20, p. 4272, 2023.

[22] S. A. Taghanaki et al., "A kernelized manifold mapping to diminish the effect of adversarial perturbations" in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11340-11349.

[23] T. Han et al., "Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization," Nat. Commun., vol. 12, no. 1, p. 4315, 2021.

[24] R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization" in Proc. IEEE International Conference on Computer Vision, 2017, pp. 618-626.

[25] Y. Li and S. Liu, "The threat of adversarial attack on a COVID-19 CT image-based deep learning system," Bioengineering (Basel), vol. 10, no. 2, p. 194, 2023.

[26] X. Wang et al., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly- supervised classification and localization of common thorax diseases" in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3462-3471.

[27] S. Minaee et al., "Image segmentation using deep learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 7, pp. 3523-3542, 2022.

[28] E. Decencière et al., "Feedback on a publicly distributed image database: The Messidor database," Image Anal. Stereol., vol. 33, no. 3, pp. 231-234, 2014.

[29] C. Szegedy et al., 2013, Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[30] "Dong, Junhao & Chen, Junxi & Xie, Xiaohua & Lai", Jianhuang et al., 2023, Adversarial Attack and Defense for Medical Image Analysis: Methods and Applications.

[31] S. Kaviani et al., "Adversarial attacks and defenses on AI in medical imaging informatics: A survey," Expert Syst. Appl., vol. 198, p. 116815, 2022.