International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deepfakes In Healthcare: Reviewing the Transformation Potential and its Challenges

Shashank Agarwal*[1], Sumeer Peta[2], Sriram Panyam [3]

Submitted:12/03/2024 **Revised**: 27/04/2024 **Accepted**: 04/05/2024

Abstract: Artificial Intelligence has become an integral aspect of the medical sector. The healthcare industry has benefited greatly from deep learning models that are based on machine learning, particularly from their propensity to handle massive volumes of data constantly. "Deepfake" is one of the most exciting Artificial Intelligence (AI) technological advancements in deep learning. The technique known as "deepfake" uses deep learning models to generate synthetic images, sound, or videos. Conventional medical procedures have limits concerning patient engagement, access, as well as training effectiveness. Deepfake technology has the potential to transform healthcare delivery, but it also poses ethical and practical concerns. The review article goes into the intricate narrative of artificial intelligence deepfakes in the healthcare sector, thoroughly reviewing and summarizing both the promising avenues they promise and the inherent problems involved with their application. It also addresses two potentially lucrative approaches to the generation of deepfakes i.e., variational autoencoders or VAEs and Generative Adversarial Networks or GANs. Collaboration among healthcare providers, technological professionals, as well as lawmakers is vital for crafting moral frameworks that guarantee responsible development and leverage the beneficial assets of deepfakes.

Keywords: *Artificial intelligence, deepfakes, generative adversarial networks healthcare, medical imaging, variational autoencoders*

1. Introduction

The state of generative algorithms for deep learning has advanced to the point where it is difficult to differentiate between authentic and fabricated data. Deep fakes, an intriguing technology, are digital frauds that create or alter audible or visual content to deceive audiences using sophisticated deep-learning techniques [1]. Much curiosity has been generated by this revolutionary technology regarding its possible uses in the healthcare industry, especially in the area of patient education.

According to Qureshi and Khan [2], supporters see a time when deepfakes would tailor education to each student's own learning preferences, resulting in a more profound comprehension of intricate medical concepts. There is thus a compelling reason to investigate how deepfakes can improve patient understanding as well as engagement in the healthcare setting. For example, a patient having a difficult diagnosis receives personalized education from a virtual doctor whose look, mannerisms, as well as patterns of speech are suited to their particular requirements and cultural environment. This artificial interaction showcases deepfakes' transformational potential in patient awareness and communication. However, like with any strong technology, ethical concerns and possible repercussions cannot be overlooked.

[1] *Independent Researcher, Chicago, IL, USA*
ORCID ID: 0009-0003-7679-6690
[2] *Independent Researcher, Boston, MA, USA*
ORCID ID: 0009-0001-5886-0498
[3] *CTO – DagKnows, San Francisco, CA, USA*
ORCID ID: 0009-0006-0025-9110
** Corresponding Author Email: shashanka757@gmail.com*

Malicious actors may use deepfakes to distribute inaccurate information, impersonate medical practitioners, and deceive susceptible patients, offering considerable emotional, financial, and medical threats. In addition, ethical considerations concerning informed consent, security of data, and possible psychological harms must be carefully considered [3]. Moreover, technological constraints in producing and identifying deepfakes offer practical hurdles for wider application [2].

Existing studies have explored the potential benefits as well as concerns regarding deepfakes in healthcare [2][4][5]. The present review article goes into the intricate narrative of artificial intelligence deepfakes in the healthcare sector, thoroughly reviewing and summarizing both the promising avenues they promise and the inherent problems involved with their application. Evaluating the potential positive and negative aspects of this disruptive technology can lead to appropriate development and incorporation within the medical community, eventually serving the best options for patients and improving care quality.

2. Generation of Deepfakes

Variational autoencoders or VAEs and Generative Adversarial Networks or GANs serve as the two potentially lucrative approaches to the generation of deepfake. Both devices are intended and used for the same usage, picture, and video creation, but the methods utilized are very different. Much research has been conducted to demonstrate the GAN and VAE's capabilities of learning crucial 306 D, which are discussed in the following section.

2.1. GAN in the Healthcare Industry

Since the year 2014, GAN has been employed as a deep learning model, while since 2016, this tool has been utilized for medical image interpretation [6]. The several GAN types for healthcare deepfake generation, include Original

Generative Adversarial Networks, Pulse2pulse GAN, Temporal GAN, StyleGAN, and WaveGAN. To produce, manipulate, and assess medical information, GAN makes use of its computation abilities as well as the availability of vast medical records. Image copying, image cropping, image inverting, and alignment have been achievable with standard data augmentation, but producing a totally new visual is not, what GAN can accomplish. Han et al. [7] employed one-dimensional noises as input to construct an MRI scan of the brain with the Generative Adversarial Networks model. This technique produced more realistic and higher resolution artificial/fake photographs of brain MRI and accomplished image denoising, although the training of the network was achieved on a dataset consisting of only 528 images. Table 1 shows some of the widely used data sets and the associated deepfake outputs in the healthcare sector [8]. Zhao et al. [9] developed a method for generating fake 128x128 MRI images of the brain. Frid-Adar et al. [10] achieved augmentation of data by employing GAN to improve Convolutional neural network (type of GAN) performance in hepatic lesion classification. Deepfake prostate tumors in MRI scans are generated using GAN [11]. Figure 1 depicts actual and deepfake-based artificial images taken and synthesized [8].

Table 1. Datasets and the Deepfake Outputs in the Healthcare Sector

Dataset	Anatomical Region	Deepfake Product
SPACE	Foetal profile	Ultrasound image
DRIVE	Blood vessels, eyes	Image of fundus or retina
SRC	Lungs	Chest image
ADNI	Nervous system	CT scan
HRF	Neuro	Neuronal image
NeuB1	Neuro	Neuronal image
STARE	Neuro	Neuronal image

2.2. VAE in the Medical Domain

Variational autoencoders are made up of an encoder encoding the incoming data input along with a decoder that reconstructs it. To accomplish the learning method, the encoder and decoder comprise distinct convolutional neural networks or CNNs [12]. The VAE technology is utilized for generating new synthetic images, whilst in medicine, it is typically employed for categorization and disease detection. To deal with limited and uneven medical imaging data, VAE is used in conjunction with StyleGAN for creating deepfake cell images for white blood cell classification. This approach displays a light for clinicians to make an effortless and precise diagnosis [13]. VAE is employed to improve the act of risk evaluation in treating patients of hepatocellular cancer by revealing detailed representations of normal tissue characteristics for future stereotactic body radiotherapy. In order to identify anomalies in chest imaging or radiography, a hybrid model combining VAE-GAN is suggested. This model does not need lesion labelling or anomalous photographs for network training. According to the research, GAN is employed within healthcare deepfake more frequently as compared to VAE [8].

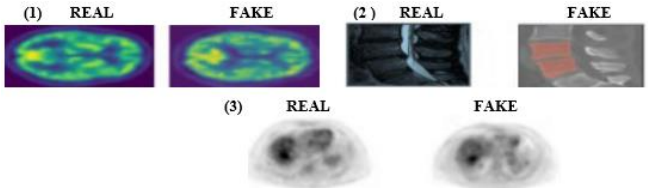

Fig.1. Images Showing (1) Real and Fake PET Scans of a Normal Patient (2) Real Input MRI and Fake CT scan (3) Real and Fake PET Scans

3. Benefits of Deepfakes in Healthcare

3.1. Customized Patient Education and Awareness

Deepfakes provide fascinating and relevant explanations for complicated medical ideas. Tobon et al. [14] found that they are successful at simulating real-life scenarios and virtual patients. This tailored approach may assist patients in comprehending and adhering to regimens more effectively. However, ethical concerns concerning data protection and possible manipulation necessitate careful examination.

3.2. Helping in Mental Disorders

Deepfake avatars have recently been investigated for their ability to provide secure settings for training social interactions and controlling phobias in individuals with anxiety conditions [15]. The success rate of virtual treatment sessions is being studied. However, uncertainties regarding possible psychological effects and the necessity for robust controls must be resolved.

3.3. Online Consultations/Telemedicine

Deepfakes possess the ability to anonymously capture patient data while maintaining authentic interactions with medical experts during virtual consultancies. These virtual meetings and appointments can help the patients get answers to their queries related to disease, dose, treatment, etc. Nevertheless, ethical concerns around informed consent as well as the possibility of abuse require meticulous investigation [16].

3.4. Medical Training

Distinct advantages are offered by deepfake patients' avatars over conventional techniques of clinical training. These advantages include improved scalability, individualization, better communication, better skill development, etc. [17]. Despite this, producing extremely realistic and complex

deepfakes yet again remains technically challenging.

3.5. Medical Image Synthesis

Deepfakes are utilized primarily in medical imaging processes to re-create and denoise imaging data, providing patients as well as healthcare professionals with a visual representation of the procedures. Deepfake technology was used to generate simulated patient data, saving the expenses of treating real patients [18]. Deepfake avatars would generate virtual patients based on data currently available to improve the safety and confidentiality of information among healthcare providers. Similar to this, multiple-domain image production was achievable in deepfake techniques that create MRI using CT [19].

4. Risks of Deepfakes in Healthcare

4.1. Manipulation in Medical Images

Although it is a fact that deepfakes are increasingly being utilized in medical settings, they are also used improperly. Medical photographs can be manipulated to add or remove indicators of health challenges to engage in insurance scams, sabotage ongoing investigations, execute terrorist acts, seek vengeance against others, or even commit murders. Attackers having control over medical imaging information modify it, for instance infusing or erasing lung cancer from a 3D scan, resulting in a misdiagnosis [8]. It is difficult to distinguish from such deepfake-generated photos, and creating fraudulent material has gotten considerably easier. Accidental victims of such crimes may suffer serious consequences throughout their entire lives.

4.2. Incorrect Data and Healthcare Frauds

Deepfakes cause a danger to the transmission of misleading data and continue the spread of medical crimes. Criminal actors may:

• Establish phony testimonials from famous people for skeptical therapies, making use of the public's faith [2].

• Act as medical experts to advertise bogus remedies or get personal data [2].

• Develop deepfake videos portraying people who experience adverse effects from authentic remedies, which increase vaccine skepticism or mistrust toward healthcare systems [3].

4.3. Moral Aspects

For the sake of accuracy and comprehension, procedures for informed consent must be thoroughly revised when patients engage with deepfakes or seek treatments from AI-powered simulators [3]. Creating and storing deepfakes demands confidential information, posing concerns regarding intrusions as well as possible breaches [3]. Although legal structures such as the General Data Protection Regulation (GDPR) provide confidentiality instructions, particular requirements for deepfakes keep growing.

4.4. Exploitation of Susceptible Patients

Deepfakes that convey customized false information can exploit individuals who are vulnerable and already have concerns and limited medical knowledge. Systems like the "Montreal Declaration for Responsible AI Development" highlight the value of avoiding these impacts. Deepfakes demand massive technological resources as well as training data, which limits their overall adoption, especially in resource-limited contexts. Furthermore, accurately recognizing deepfakes continues to be an obstacle, which highlights doubts regarding the accuracy of connections and data [1]. For execution to be secure and credible, specific limitations regarding this aspect need to be addressed.

4.5. Influence on Discrimination in Healthcare

Differences in medical care can be exacerbated by inadequate usage of deepfake-based approaches. A lack of accessibility to such possibly helpful innovations to individuals living in remote regions or with struggling financial means may increase gaps in quality of medical care [20]. Much consideration needs to be put towards improving availability and assuring an equitable allocation of deepfakes in the medical field.

5. Detection of Deepfakes in Healthcare

Two types of methods have been suggested for detecting tampering in medical images. Figure 2 depicts two types of detection techniques: active and passive. Active detection approaches, such as signatures, digital watermarking, etc., require embedding a code of authentication into a picture using specific software or hardware prior to distribution. However, active detection has two major issues: (a) adding additional data following the image was shot, and (b) the impact of watermarks on image quality [21].

Passive detection approaches, such as copy-move and picture splicing, compare the image's frequency domain attributes or statistical information to detect alterations in both local and overall characteristics. Copy-move conceals the region of interest by repeating an uninteresting region over the desired region. Unlike copy-move, picture splicing duplicates a targeted region from an external image input. The main advantages of this type of detection technique are: (a) No prior information is required for validating the image; (b) it eliminates the chances of visual degradation caused by integrated watermark information [22].

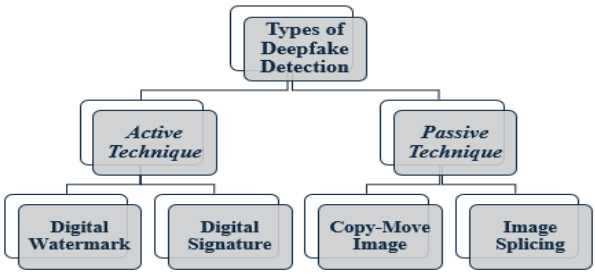

Fig. 2. Types of Deepfake Detection Methods Used in Healthcare

The problems associated with GANs, along with their two primary uses in clinical imaging are discussed in the study by Yi et al. [23]. Images with generative features served as the focal point of the initial application. The second usage, on the other hand, focused on the discerning component; when unusual or fraudulent images are displayed, discriminator D may be employed as a detector. In order to identify abnormal photographs, this study used GAN to perform tasks on image generation, segmentation, reconstruction, categorization, identification, and registration.

Another study by Zhang et al. [24] investigated detecting images generated by GAN and found that there existed no viable solution. They conducted the experiment with two techniques, principal component analysis or PCA alongside the support vector machine or SVM. These were tested to categorize GAN-produced images via a two-phase cascade framework that detects counterfeits on a mere one percent of the original visual. Doctors were perplexed by the classification, since they were unable to distinguish between actual and fabricated photographs, demonstrating the importance of this study. The finalized proposed model had 99.80% accuracy in classifying CycleGAN-altered medical photos and actual images.

Ghadi et al. [25] used a variety of CNNs to examine how effectively they could recognize GAN-produced deep fake visuals. Because deep fakes use machine learning to produce high-quality imitations, distinguishing between synthetic and real photos becomes much more difficult. This study investigated several CNN-based systems for classifying deep fake images, such as DenseNet121, VGG16, VGGFace, etc. and analyzed them by applying five metrics i.e. F1-scoring, the area under receiver operating characteristic) curve, precision, recall, and accuracy.

Siddharth et al.'s paper [26] presented a deep fake identification algorithm of medical images that allowed for differentiation between tampered and untampered photos. It is built on three machine learning techniques including five deep learning models. The learning strategies employed in the investigation were built using CT-GAN datasets. The experiment's findings demonstrated the improved classification of tumor injection images using deep learning and localization in the targeted region. For dispersed medical deepfake imaging the DenseNet framework achieved the highest 80% accuracy score.

The potential problems associated with the manipulation of data and regeneration of images in healthcare settings were addressed by Suk et al. [27]. The research dataset consisted of four lesion images utilized for manipulating data of the fundus (i.e. normal, retinopathy caused by diabetes, glaucoma, as well as macular degeneration). The whole architecture was built on a Sparse convolutional neural network for operating the manipulation detection mechanism, which used Cycle GAN plus U-Net and achieved a 91% detection rate.

Reichman et al. [28] presented ConnectionNet (a framework based on deep learning) that allows for automatic detection of manipulated images. The presented ConnectionNet operated on tiny, altered areas in images, produces promising findings, and can be used as a foundation for future medical imaging research. LuNoTim CT scan represents a new dataset. It comprises a large number of CT scans that have been manipulated using three techniques: copy-move fraud, and deep and classical inpainting. The proposed framework achieved an 85% accuracy in detecting deep fakes.

Gite et al. [29] centered their research on tuberculosis (TB), a Mycobacterial lung infection. X-ray scans are the main tool used by physicians for the diagnosis of tuberculosis. In order to determine which deep learning algorithms worked best, semNet (which is a semantic segmentation framework), U-Net++, U-Net, and Fully Convolution Network or FCN were compared. Following lung segmentation analysis and comparison, the models were assessed, and U-Net++ outperformed U-Net in terms of accuracy. U-Net++ attained an accuracy that exceeded 98% upon completion of the study, whereas FCN, U-Net, and SegNet attained 78%, 95%, and 84% respectively.

6. Potential Suggestions to Combat Negative Aspects

6.1. Promote Clearly Defined Frameworks

• Involve lawmakers by encouraging them to back policies that deal with informed consent, confidentiality of data, and the possible abuse of deepfakes in medical care.

• Taking part in campaign groups by raising voices in favour of organizations who fight for ethical deepfakes as well as accountable AI development.

6.2. Promote Development and Research

• Support experimental projects by participating in studies on deepfake identification, and ethical AI creation, and lowering risks by contributing or volunteering with organizations.

• Support open-source teamwork by promoting clarity, and good conduct and by striving for the advancement of deepfake technology.

6.3. Encourage Open Discussion

• Have conversations with medical professionals by discussing the possibilities of deepfakes in their line of work and inquire about the appropriate execution options.

• Inform your community by organizing seminars that increase consciousness regarding deepfakes in medical care and promote educated discussion.

• Enable patients by encouraging them to take part in medical discussions about deepfakes.

6.4. Hold Stakeholders Responsible

• Ensure transparency by requesting particular clarifications from system developers, medical professionals, and legislators about the usage of deepfakes.

• Promote ethical businesses by choosing physicians and technological companies who encourage ethical standards and privacy of personal information.

• Address harmful practices by reporting any improper or immoral usage of deepfakes in medical settings to appropriate authorities.

Appendixes, if needed, appear before the acknowledgment.

7. Conclusion

The key benefits of deepfakes in the healthcare industry are outlined in the paper, along with their inherent drawbacks. Additionally, it provides an overview of the most widely used techniques, such as GAN and VAE, for the generation of deepfake medical images, demonstrating that the GAN model has become the most frequently used model currently in use. Various methods have also been highlighted for detecting tampering in medical images and it is concluded

that passive detection types are more beneficial than active type. Even while deepfakes present exciting opportunities for the healthcare industry, managing the risks involved is essential to their responsible development as well as application. It is possible to pave the path for utilizing the positive attributes of technological advancement while making sure it supports patients effectively and ethically by recognizing the possibility of false information, ethical issues, technical constraints, and inequities.

References

[1] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," Journal of Business Research, vol. 154, p. 113368, 2023.

[2] J. Qureshi and S. Khan, Artificial Intelligence (AI) Deepfakes in Healthcare Systems: A Double-Edged Sword? Balancing Opportunities and Navigating Risks, 2024.

[3] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Machine Intelligence, vol. 1, no. 11, pp. 501–507, 2019.

[4] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 14, no. 2, p. e1520, 2024.

[5] O. Cabrera, The World of Deepfakes.

[6] X. Li, Y. Jiang, J. J. Rodriguez-Andina, H. Luo, S. Yin, and O. Kaynak, "When medical images meet generative adversarial network: recent development and research opportunities," Discover Artificial Intelligence, vol. 1, pp. 1–20, 2021.

[7] C. Han, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, and H. Hayashi, "Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection," IEEE Access, vol. 7, pp. 156966–156977, 2019.

[8] D. Lakshmi and D. J. Hemanth, An Overview of Deepfake Methods in Medical Image Processing for Health Care Applications, 2024.

[9] J. Zhao, D. Li, Z. Kassam, J. Howey, J. Chong, B. Chen, and S. Li, "Tripartite-GAN: Synthesizing liver contrast-enhanced MRI to improve tumor detection," Medical Image Analysis, vol. 63, p. 101667, 2020.

[10] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing, vol. 321, pp. 321–331, 2018.

[11] K. Falahkheirkhah, S. Tiwari, K. Yeh, S. Gupta, L. Herrera-Hernandez, M. R. McCarthy, and R. Bhargava, "Deepfake histologic images for enhancing digital pathology," Laboratory Investigation, vol. 103, no. 1, p. 100006, 2023

[12] S. Zavrak and M. Iskefiyeli, "Anomaly-based intrusion detection from network flow features using variational autoencoder," IEEE Access, vol. 8, pp. 108346–108358, 2020.

[13] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: a review," Journal of Imaging, vol. 9, no. 4, p. 81, 2023.

[14] D. P. Tobon, M. S. Hossain, G. Muhammad, J. Bilbao, and A. E. Saddik, "Deep learning in multimedia healthcare applications: a review," Multimedia Systems, vol. 28, no. 4, pp. 1465–1479, 2022.

[15] A. Werntz, S. Amado, M. Jasman, A. Ervin, and J. E. Rhodes, "Providing human support for the use of digital mental health interventions: systematic meta-review," Journal of Medical Internet Research, vol. 25, p. e42864, 2023.

[16] M. Langarizadeh, F. Moghbeli, and A. Aliabadi, "Application of ethics for providing telemedicine services and information technology," Medical Archives, vol. 71, no. 5, pp. 351, 2017.

[17] S. Szymanowicz, V. Estellers, T. Baltrušaitis, and M. Johnson, "Photo-Realistic 360° Head Avatars in the Wild," presented at the European Conference on Computer Vision, 2022, pp. 660–667.

[18] S. Neethirajan, "Is Seeing Still Believing? Leveraging Deepfake Technology for Livestock Farming," Frontiers in Veterinary Science, vol. 8, Nov. 23, 2021. doi: 10.3389/fvets.2021.740253.

[19] R. Oulbacha and S. Kadoury, "MRI to CT synthesis of the lumbar spine from a pseudo-3D cycle GAN," presented at the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1784–1787.

[20] World Health Organization, World Health Report 2023: Health Inequalities: The Root of the Problems and the Path to Solutions, 2023.

[21] C. S. Prakash, H. Om, and S. Maheshkar, "Authentication of medical images using passive approach," IET Image Processing, vol. 13, no. 13, pp. 2420–2427, 2019. https://doi.org/10.1049/iet-ipr.2018.6035

[22] A. Alsaheel, R. Alhassoun, R. Alrashed, N. Almatrafi, N. Almallouhi, and S. Albahli, "Deep Fakes in Healthcare: How Deep Learning Can Help to Detect Forgeries," Computers, Materials & Continua, vol. 76, no. 2, 2023.

[23] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," Medical Image Analysis, vol. 58, p. 101552, 2019.

[24] J. Zhang, X. Huang, Y. Liu, Y. Han, and Z. Xiang, "GAN-based medical image small region forgery detection via a two-stage cascade framework," PLOS ONE, vol. 19, no. 1, p. e0290303, 2024.

[25] I. Ghadi, S. Akhter, T. A. Alsuhibany, A. A. Shloul, and A. Jalal, "Comparative analysis of deep fake image detection method using convolutional neural network," *Computational Intelligence and*

Neuroscience, vol. 88, no. 6, pp. 910–931, 2021.

[26] H. J. Suk, J. Song, and J. H. Han, "A study on the development of deep fake-based deep learning algorithm for the detection of medical data manipulation," Webology, vol. 19, no. 1, pp. 4396–4409, 2022.

[27] L. Reichman, O. Jing, A. Akin, and T. Yingli, "Medical image tampering detection: A new dataset and baseline," Pattern Recognition, vol. 68, no. 4, pp. 266–277, 2021.

[28] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," Medical Image Analysis, vol. 58, p. 101552, 2019.

[29] S. Agarwal, "Machine Learning Based Personalized Treatment Plans for Chronic Conditions," *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, 2024, pp. 1127-1132.

[30] A. Gite, A. Mishra, and K. Kotecha, "Enhanced lung image segmentation using deep learning," Neural Computing and Applications, vol. 8, no. 3, pp. 1–15, 2022.