University of Reading

Challenges of deep learning in medical image analysis – improving explainability and trust

Article

Accepted Version

Dhar, T., Dey, N., Borra, S. and Sherratt, R. S. ORCID: https://orcid.org/0000-0001-7899-4445 (2023) Challenges of deep learning in medical image analysis – improving explainability and trust. IEEE Transactions on Technology and Society, 4 (1). pp. 68-75. ISSN 2637-6415 doi: 10.1109/TTS.2023.3234203 Available at https://centaur.reading.ac.uk/109789/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1109/TTS.2023.3234203

Publisher: IEEE

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

University of Reading

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Full Text Version

Challenges of Deep Learning in Medical Image Analysis – Improving Explainability and Trust

IEEE Transactions on Technology and Society

DOI: 10.1109/TTS.2023.3234203
Accepted: 1st Jan 2023.
Published on-line: 5th Jan 2023.

Authors:

Tribikram Dhar,
Department of Electrical Engineering, Jadavpur University, Kolkata-712235, India. (Email: dhartribikram@gmail.com)

Nilanjan Dey, Senior Member, IEEE,
Department of Computer Science Engineering, Techno International New Town, Kolkata-700156, India. (Email: nilanjan.dey@tint.edu.in)

Surekha Borra, Senior Member, IEEE,
Department of Electronics and Communication, K. S. Institute of Technology, Bengaluru-560109, India. (Email: borrasurekha@gmail.com)

R. Simon Sherratt, Fellow, IEEE,
Department of Biomedical Engineering, University of Reading, United Kingdom. (Email: r.s.sherratt@reading.ac.uk).

Abstract—Deep learning has revolutionized the detection of diseases and is helping the healthcare sector break barriers in terms of accuracy and robustness to achieve efficient and robust computer-aided diagnostic systems. The application of deep learning techniques empowers automated AI-based utilities requiring minimal human supervision to perform any task related to medical diagnosis of fractures, tumors, and internal hemorrhage; preoperative planning; intra-operative guidance, etc. But deep learning faces some major threats to the flourishing healthcare domain. This paper traverses the major challenges that the deep learning community of researchers and engineers faces, particularly in medical image diagnosis, like the unavailability of balanced annotated medical image data, adversarial attacks faced by deep neural networks and architectures due to noisy medical image data, a lack of trustability among users and patients, and ethical and privacy issues related to medical data. This study explores the possibilities of AI autonomy in healthcare by overcoming the concerns about trust that society has in autonomous intelligent systems.

Index Terms—Adversarial attacks, Computer-aided diagnostic systems, Convolutional neural network, Data augmentation, Deep learning, medical image analysis.

I. Introduction

Artificial intelligence (AI) has become one of the most ground-breaking fields that has revolutionized healthcare [1], remote sensing [2], robotics [3], autonomous driving [4], and other interdisciplinary domains over the past decade. Over time, machine learning (ML) algorithms have been used by researchers and engineers to solve many biomedical tasks as well as simple data processing tasks [5, 6]. ML accomplishes a job by learning and utilizing previously obtained task-specific data. While machine learning algorithms excel when dealing with simple linear data, they do less well when handling complicated medical data [7]. Over time, deep learning, a subtype of AI has surpassed machine learning in the diagnosis of medical images [8], thanks to its improved accuracy, capacity for deep feature learning, and flexibility in tackling challenging diagnostic issues. However, the use of deep learning has been met with some difficulties, including the lack of annotated medical image data, lossy medical image data with attenuations and artefacts, user mistrust of deep learning-based tools like computer-aided diagnostic (CAD) systems, and ethical and privacy concerns regarding the data. Researchers and engineers are very concerned about these issues, given that they are hindering the advancement of deep learning in the medical field.

Image artefacts and noise are considered adversarial attacks as they result in a biased and wrong prediction. While motion artefacts are common in magnetic resonance imaging (MRI) scans [9], computed tomography (CT) scans frequently exhibit ring artefacts, beam-hardening, scatter artefacts, and metal artefacts [10]. Further, a blend of speckle, Poisson, and Gaussian noise is the most prevalent form of noise observed in radiological medical images like X-rays [11]. Deep learning practitioners in the medical domain must denoise image data manually to generate a clean dataset for training and testing. However, this procedure necessitates more time and space, which drives up the price of creating the CADs.

While the availability of annotated medical image data is low, large datasets are required to train and validate robust deep learning architectures. The image labelling is a tedious task that requires supervision by medical professionals, making it expensive and time-consuming. This crunch in data generates a big challenge for training deep neural networks. Further, any imbalance in data for a particular class in the dataset could create a bias in the deep neural network, which could lead to overfitted prediction and diagnosis from the model. Modern deep learning techniques use representation learning algorithms to generate high volumes of medical images from a small medical image dataset [12, 13]. Although the artificial datasets generated after augmentation produce good results when training deep neural networks, a lack of trust towards these methods of data generation prevails in the community.

Since deep learning models are typically “black box” algorithms, the public has a critical perspective on them. There are significant risks associated when a black box machine is employed as a diagnostic tool in healthcare facilities. Even though deep networks are frequently utilized by doctors and other medical experts, not everyone accepts the risk, which makes it difficult to trust these AI approaches. To increase user confidence and trust in CADs, it is imperative that the explainability of deep networks be thoroughly researched and made available to users.

Privacy and ethical issues are great concerns that prevent the deep learning community from sharing medical image data in the open-source domain. It is estimated that by 2027, cloud computing in health care will be worth 89 billion US dollars [14], and to ensure the users' privacy and prevent counterfeiting, certain measures must be taken unless it is deemed to be open source. Medical data is protected by various international regulations in different nations. For instance, to prohibit the disclosure of sensitive patient health information without the knowledge or consent of the patient the United, States' Health Insurance Portability

and Accountability Act (HIPAA) of 1996 created national standards. Since patients' personally identifiable information is now legally protected, healthcare professionals are now compelled to secure, restrict the use of, and distribute it. Therefore, maintaining the secrecy of patient identity is one of the main responsibilities of deep learning engineers, which is challenging. The United Nations member states adopted the “Sustainable Development Goals” in 2015 with an aim to achieve 17 goals by 2030, with the third goal to "*Ensure healthy lives and promote well-being for all at all ages*." Healthcare and diagnosis in rural areas in some underdeveloped countries face a major barrier due to a lack of infrastructure. A disruption in healthcare of 92% global member states of the UN was recorded during the pandemic [15]. It created havoc in the healthcare community, and the major pitfalls of manual, contact-based diagnosis were noticed by society. Proposals for contact-less diagnosis were achieved during the latter half of the pandemic using CADs. As a result, to achieve health supremacy, society must look up to and trust AI-based diagnostic systems. The shift from manual healthcare diagnosis to efficient CAD-based diagnostic systems could facilitate a major growth in the healthcare system by reducing diagnosis time, preventing erroneous diagnoses, and thus reducing mortality rates. The acceptance of AI based systems will help society accelerate in the domain of healthcare and therefore improve the quality of human life.

The primary objective of our research was to explore and identify the challenges, explainability, trustability, and prospects of deploying deep learning algorithms to healthcare and medical image analysis. Specifically, in this paper, we consider:

1) Adversarial attacks in deep networks due to attenuation.
2) Unavailability of data and imbalanced data.
3) Trust issues and explainability.
4) Privacy and legal issues of medical data.

The rest of the paper is organized as follows. Section II explains the role of explainable AI (XAI) in the CAD systems related to healthcare. Section III provides the challenges of deep learning in healthcare. Section IV discusses the ethical principles and Section V concludes the paper.

II. Deep Learning in Healthcare

Healthcare organizations and facilities are gradually becoming more inclined toward the methods in which artificial intelligence can be used to diagnose diseases and treat patients. In the past several years, deep learning has revolutionized the ability of machines to analyze and handle data at a very high speed and with a precision that has never been achieved before. It is a hierarchical method that utilizes complex and deep structures and efficiently learns non-linear data with high accuracy and precision. Deep learning has been investigated in biological image processing, the diagnosis of illnesses, and the design of surgical systems for intraoperative and preoperative support, and the findings are encouraging. According to a survey [16] conducted in the US, people trust and are aware of AI in healthcare. A range of patient-facing healthcare technologies that allow individuals to communicate with clinicians and access their own medical information are highly acquainted with and used by more than half of patients (58%), according to the survey. The fact that 52% of people do trust AI for their medical requirements was also underlined. This shows that there is a pressing need to develop high-performance AI that solves a range of issues.

For artificial intelligence to grow in the healthcare industry, it is essential to comprehend medical data, discern techniques for processing it, and use the resulting CADs to provide precise and trustworthy findings.

A proper understanding of medical data is required to effectively use data and resources in the healthcare business and to deliver trustworthy outcomes. Fig. 1 represents a typical CAD system and its functionality of diagnosing medical image data or real time medical image time series data in surgical environment.

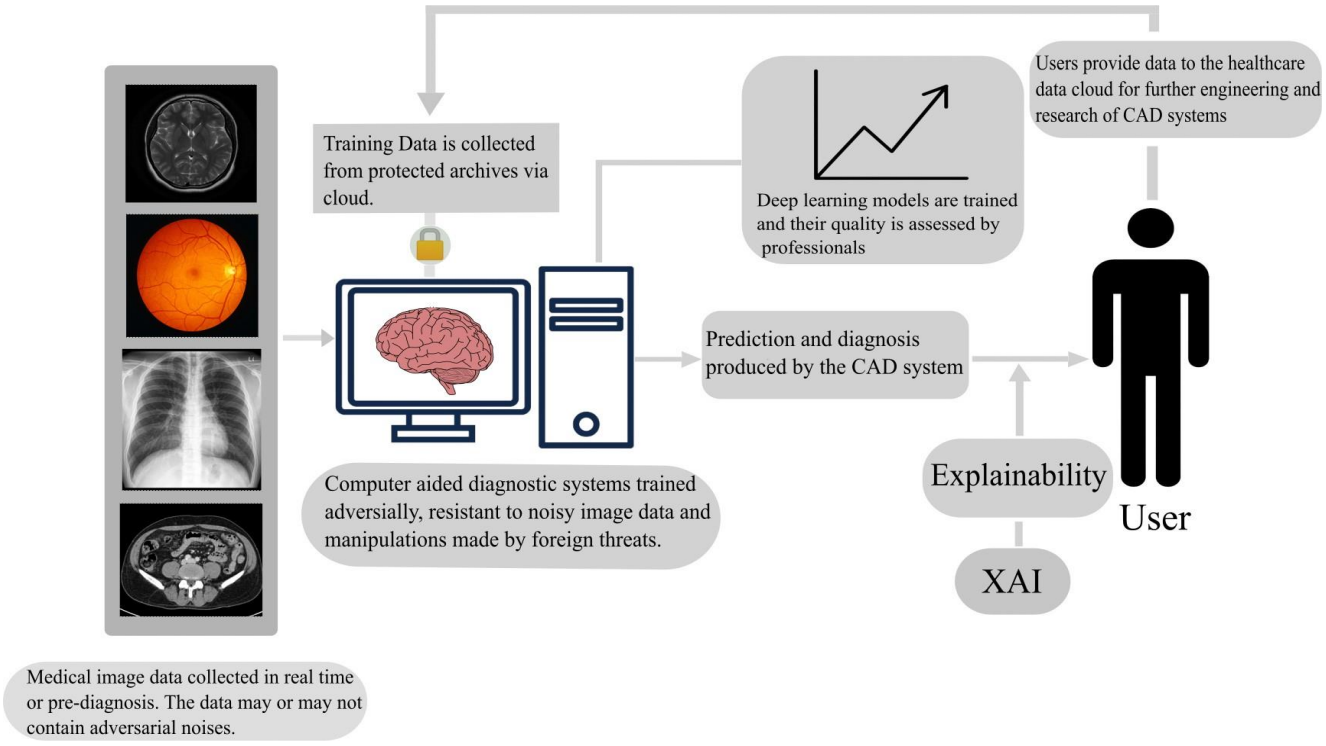

Fig. 1. A general overview of a computer aided diagnostic system and how it tackles deep learning challenges.

Once the deep learning model incorporated into the CAD is trained and perfected by deep learning engineers, the system takes the medical image as an input and generates the diagnosis prediction for the user. The CAD system is trained so that it is immune to adversarial attacks and attenuations. In the case of a very corrupted medical image, it could provide the user with an inconclusive result for the diagnosis. Inclusion of an XAI module in the CAD, as shown in Fig. 1, gives the user both a visual and a numerical explanation, upholding the user's faith and trust in the outcomes of the CADs. The models' internal workings and predictions are made openly available to be adequately interpretable by human analysts to develop trust in model predictions and find potential defects or limitations in the model. To support the expansion of CAD research, users contribute their own data to the healthcare facility's data storage cloud. The user has the option of reporting an incorrect diagnosis to the development team for evaluation. Further, to avoid misdiagnosis and to gain users' trust, CADs must be gradually retrained repeatedly based on real time data and thoroughly field tested before they are deployed into the medical industry. The performance should be assessed under the guidance of medical practitioners, whose expertise and supervision could be used to validate these AI-based CADs, which would help them improve.

III. Challenges of Deep Learning in Healthcare

A. Adversarial Attacks in Deep Networks due to Attenuation

Deep neural networks, when subjected to noise, could produce misclassifications and erroneous results [17]. Medical image data is not devoid of noise and attenuations; it contains a wide variety of attenuations and motion artefacts. Multiplicative noise is often present in biomedical imaging, such as MRI, CT, ultrasound, and positron emission tomography (PET) imaging. The denoising varies the spatial and temporal distribution, and the medical image contrast so a substantial amount of detail is lost. If the AI systems are not well protected from malicious attacks, it might bring mistrust in AI among the common people. If any malicious third party tries to introduce any form of adversarial attack, the CADs must be prepared forehand to deal with them. Failure of being resistant to these attacks will make the users of the technology lose faith in the CADs. Linear algorithms fail in removing multiplicative noise which is signal dependent and is described mostly by complex Rayleigh and Gamma models. Further, since both noises and edges of medical images contain high frequencies, linear filters often do not provide sufficient performance when denoising.

The removal of the artefacts poses a massive challenge to deep learning researchers, engineers, and biomedical professionals in the healthcare decision support system. In the early days of ML usage in healthcare industry, the image processing community engaged in developing pre-processing methods to root out these artefacts and attenuations in medical data. Liao *et al.* [18] presented a bilateral filter (CBF) that when applied to a smoother version of the image (context), suppresses the propagation of noise (PoN) and performs blind image denoising effectively. Bhonsle *et al.* [19] used bilateral filtering of MRI and X-ray images perturbed with additive Gaussian noise of different variances, and achieved better results than linear filtering methods like the Wiener filter, mean filter, etc. There has been a considerable amount of research exploring the use of deep learning for medical image denoising. Gondara [20] showed that denoising autoencoders using convolutional layer blocks can be used as effective denoising tools for medical imaging. The author showed that a small data volume containing about 300 samples can be sufficient to train the autoencoder and produce efficient denoising performance. Jifara *et al.* [21] proposed a residual learning approach along with batch normalization to learn the noise directly from noisy images. In recent years, adversarial training has become more effective in dealing with lossy data, and it does not manipulate the lossy data to produce results. A wide range of adversarial training methods using generative adversarial networks (GANs) [22] were explored by Yi *et al.* [23]. Although a fair amount of research has been conducted in the deep learning community related to adversarial training, very few studies have been implemented in the medical industry. The basic intuition behind adversarial training is that data is augmented by adversarial samples, which contain the general features of the noisy image data. The training data is enforced with accurate ground truth, and the network is trained so that in the event of an attack with noisy data, the deep network can be flexible to the attack and produce accurate results thereby improving user's trust [24].

B. Unavailability and Imbalanced Data

An ideal medical image dataset should have the metadata, the identifier, and an adequate volume of images, be correctly annotated with ground truth by a medical professional and be reinforced with a proper license for distribution in the deep learning research community. Annotations and the data that is generated using imaging modalities are part of metadata. Medical images are indeed time-consuming and expensive to collect. There remains a scarcity of a large volume of authentic annotated medical image data, which prevents the growth of deep learning in medical image diagnosis. To compensate for the scarcity of medical image datasets, the generation of synthetic datasets has begun. This procedure is known as "image data augmentation," which includes the generation of images from the original dataset by manipulation while keeping the important features of the image data preserved. The deep learning model considers the new image to be alien data but still increases the effective volume of the image dataset. There has been a substantial amount of research over the years on data augmentation techniques. Fig. 2 explores various state-of-the-art

image augmentation techniques, apart from the classical image manipulation methods that are widely used with deep learning methods. Basic image manipulation and augmentation techniques include image translation and flipping, jittering of image pixel intensity values, random cropping of images, random rotation, image shearing using affine transformation, adding Gaussian blur and noise, and color space transformation.

Overfitting is a condition where the AI model associates itself closely with the training data and may not perform well when subjected to unseen data. It is usually experienced when the training data is small and not clean enough, i.e., when it contains noise.

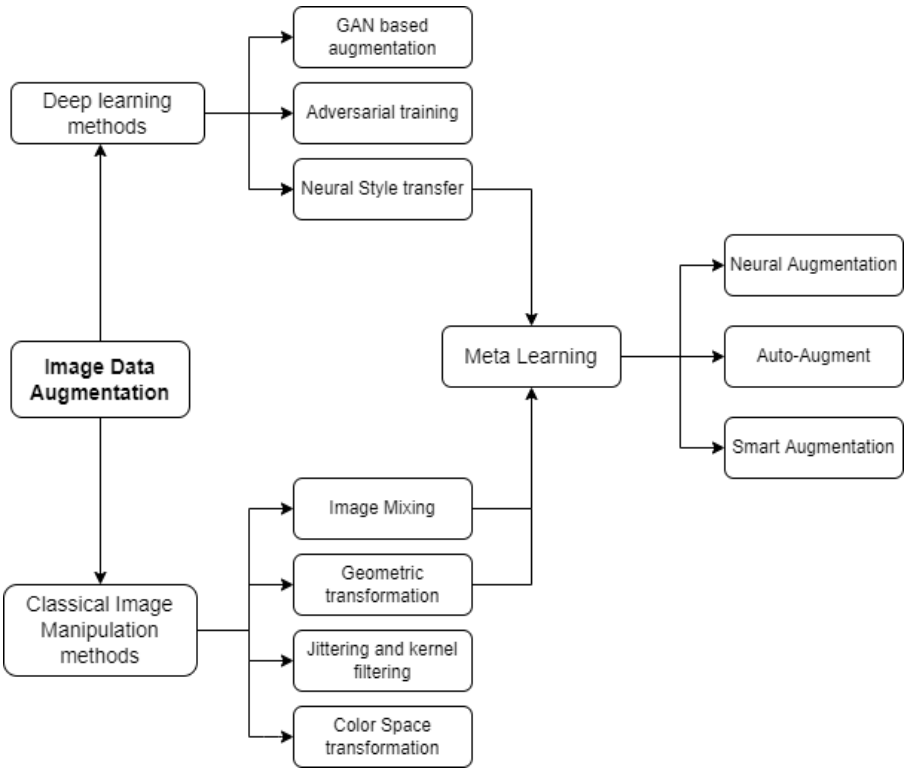

Fig. 2. Different Image Augmentation approaches.

Image manipulation methods do not generate new images for that feature space, and since there is structural similarity with the original data, the deep learning model is overfitted for using the augmented data. To resolve this challenge, deep learning methods using autoencoders and generative adversarial networks (GANs) were implemented, which learned the features of the data for every different class of disease and generated synthetic images for the respective classes. A GAN contains two neural networks: a generator and a discriminator. While synthetic data is generated by the generator from random noise, the discriminator predicts if the data is real or synthetic. Both networks get better through training by trying to outperform each other in a min-max game. In a typical GAN-based augmentation, the generator is trained to generate synthetic medical image data by training it on a real medical image dataset. The image dataset, which is generated by the generator, is used to train or validate deep networks for medical diagnosis.

Pix2pix [25], Pix2pixHD [26], SPADE [27], are some translations based GANs that translate images using masks that improve biological and medical coherence. While GANs like DCGAN [28] are flexible, the PGAN is good at generating high-resolution images. Data augmentation [29, 30] and patch extractions are often used with noise-based generation models when dealing with small datasets.

Sundaram and Hulkund [30] used GAN generated chest X-ray images to train and classify lung lesions, fractures, and pleural effusion categories. Bissoto *et al.* [31] used GAN-based augmentation to augment skin lesion medical images and studied the different synthetic data generated from different GANs. Their study revealed that for skin lesion image augmentation, StyleGAN 2 [32] outperformed other contemporary architectures in generating synthetic image data. Frid-Adar *et al.* [33] implemented the generation of augmented CT images for liver lesion classification. A convolutional neural network (CNN) trained on augmented data from the GAN outperformed the CNN trained on a real dataset, thus showing the improvement of the deep network by using augmentation.

Conditional GANs are used by Kyuchkov *et al.* [34] on CT images to improve the reproducibility and detection probability of pulmonary nodules. For data augmentation, the model used an open-source CT-GAN network, Wasserstein loss, and adaptive instance normalization. Minne *et al.* [35] conducted experiments for a comparative analysis of the classical and GAN-based augmentation techniques to augment 3D MRI data to diagnose and treat Alzheimer's disease and dementia. Zhang *et al.* [36] used a domain adaptation technique to overcome the unavailability of medical image data by segmenting medical images using meta-learning [37]. Singh *et al.* [38] proposed a "MetaMed" approach aiming to reduce computation and improve accuracy with small datasets. While there is a 2–5% improvement in generalization capability, 70% accuracy was obtained using the hybrid augmentation and regularization models when tested on three datasets: ISIC 2018, BreakHis, and Pap smear. While these are some recent encouraging findings made by deep learning researchers to tackle the challenge of data scarcity and variance, integrating patients' trust information and trust propagation [39] with GANs while generating new samples improves the performance of recommender systems and patient' trust.

C. Trust Issues and Explainability

Deep learning training algorithms are mathematically comprehensible, but their architectures and the complex mathematics behind them are more of a "*black box"* model. Thus, the prediction of deadly diseases, medical diagnosis for treatment, or surgical guidance using these AI methods raises ethical concerns for people. The public generally refuse to trust these CAD systems and their evaluations of the given data. As a result, the deep learning community in healthcare needs highly qualified individuals to create these deep learning-based diagnostic tools.

Accuracy, precision, recall, and the F1 score are the standard metrics to evaluate a model's performance in terms of classification

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{TN}+\text{FN}) \quad (1)$$

$$\text{Specificity} = \text{TP}/(\text{TP}+\text{FP}) \quad (2)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) \quad (3)$$

$$\text{F1 score} = (2*\text{precision}*\text{recall})/(\text{precision}+\text{recall}) \quad (4)$$

Where TP is true positive, FP is false positive, TN is true negative, and FN is false negative.

It is important for the model to have strong F1, precision, and recall in addition to good accuracy. While sensitivity corresponds to the accurate identification of the diseased, specificity corresponds to the correct identification of people without disease. Since accuracy, precision, and other metrics usually estimate how well the model performed on the test set, they could be manipulated and changed to reflect different values. AI engineers could readily fabricate the data used to represent the model's performance. As a result, users never have trust in the performance indicators of the model. This lack of trust led to the demand for XAI, which enables consumers to understand the predictions and diagnoses made by deep learning algorithms.

The wide breach of trust between the society and the black box deep learning models can be mended by implementing explainable deep learning in CADs. Unlike machine learning algorithms, deep learning handles visual information very explicitly, and visual explainability is more informative and captivating to the users than numerical quantifiers. Thus, for society to accept AI, visual explainability must be actively implemented in CAD systems.

Research on the explainability of deep learning in medical imaging has been extensively conducted. Magesh *et al.* [40] studied the explainability of Parkinson's disease prediction by CNNs on DaTscan (dopamine transporter scan) data using the Local Interpretable-Model Agnostic Explainer (LIME) method and produced encouraging results. In deep learning, to explain image classification tasks using CNN, guided backpropagation, gradient-based class activation maps (Grad-CAMs) [41], class activation maps (CAMs), and deconvolution techniques are used. These techniques generally fall into the category of "visual explanation" of deep learning methods. Cohen *et al.* [42] demonstrated the use of Grad-CAMs, CAMs, and heatmaps to localize infected regions in lungs of a COVID-19 patient to predict the opacity and extent of the disease and to produce an interpretable output. Successful experiments have been conducted by Gu *et al.* [43] using a comprehensive attention convolutional neural network and have achieved remarkable results in skin lesion segmentation and fetal MRI segmentation based on their explainability and disease localization. Fig. 3 represents a plot of performance versus explainability for the various machine learning models, including deep learning models and Bayesian models.

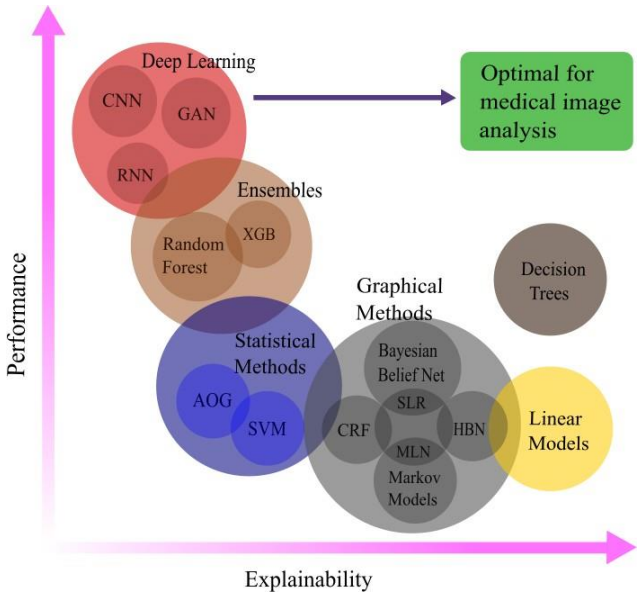

Fig. 3. Performance vs explainability plot for various machine learning, statistical learning, Bayesian learning and deep learning methods.

1) Model agnostic and model-specific explainability: Model agnostic XAI methods, by definition, are model or architecture independent and provide a high level of flexibility. Instead of developing a different evaluation method for a different model, it has been found that model agnostic methods produce praiseworthy results for different types of AI architectures and algorithms, from decision trees, support vector machines (SVM), nearest neighbor classifiers, logistic regression classifiers to some black-box methods like CNN. A reliable model-agnostic method that is approved by clinical experts and scientists can be used globally as a standard method. Once a model is standardized, it can be easily explained to the common people while comparing with other reliable methods to ensure the trustworthiness. For example, the users would find a linearly explainable algorithm or a visual representation more promising than statistical values that they

might not be aware of. Unlike model-agnostic methods, model-specific methods do not explain a large spectrum of deep learning models, although their explanations are much more specific than their model-agnostic counterparts. For example, with GradCAM tool, using a method that is specific to visual data-based architectures like CNN, gradient maps can be generated. The pixels in the image responsible for a particular prediction can be described, which in turn is logically understandable by a human being. The deep learning research community has made significant efforts to bring model explainability tools to society. Tools such as LIME, SHAP (SHapley Additive exPlanations), WIT (Whatif Tool), AI Explainability 360 toolset and Skater, are widely used to produce reports on a model's operation and attempt to explain the algorithm.

2) Post hoc explainability: Post hoc explainability is a technique for analyzing a trained deep network to gain understanding of the learned mapping. The model-based and model-specific explanations are different from the post hoc explanation methods. The former train a deep network and then attempt to explain the pattern of predictions given by the black box deep network, whereas the latter try to make the deep network explainable by various modifications.

D. Privacy and Legal Issues of Medical Data

Medical data is one of the most valuable entities in the healthcare system; it contains all the valuable health-related information of everyone who has ever been registered in a public or private medical facility for treatment or diagnosis. In recent years, medical data has been stored digitally, and the sharing of this data is protected by various laws. The Patient Data Protection Act, or Patientendaten-Schutz-Gesetz (PDSG), was passed by the German parliament (The Bundestag) in 2020 for the complete digitalization of the German healthcare system. Several database providers, organizations, and websites—including Harvard Dataverse, U.K. Biobank, Kaggle, and IEEE Dataport—have recently made medical data more accessible. They exchange licensed data that registered data scientists, machine learning, and deep learning programs can use. However, these medical databases can be compromised by external malicious software on the system, and antivirus programs and individuals demanding high cybersecurity measures like strong password policies, firewalls, and penetration testing. If malicious programs can steal the private health information, the patient might be subjected to harassment, cyberbullying, paranoia, or mental pain. Thus, the decision to disclose medical information rests with the data's owner, or the patient, who frequently expresses reluctance due to the data's confidentiality and integrity. Only a small percentage of patients volunteer to engage in various data gathering questionnaires and actively provide their medical information for study. For deep learning researchers and engineers, getting access to confidential medical data presents a significant challenge. To remedy such challenges, researchers could use techniques involving pseudonymized data or differential privacy rather than identified data [44]. Privacy audits can ensure patients trust, appropriate use, and security standards should guard against unauthorized use.

IV. Ethical Principles, Opportunities and Future Research Directions

The world of deep learning in medical image diagnosis faces the challenges discussed in the previous section, has created a divide between society and the deep learning community. The misunderstanding and suspicion that society has towards intelligent autonomous systems in healthcare and diagnosis hinders the autonomy of AI-based healthcare systems.

Transparency, explainability, fairness, non-maleficence, accountability, or privacy are some of the ethical principles that may be used in the design of deep learning systems to alleviate patient concerns about data security, confidentiality, and integrity. Actions are required to address the ethical standards of AI at various organizational levels, and it should be a continuous process. The deployment of ethical AI involves several governance procedures that are not only technical, and there are many stakeholders involved. Risk

assessment, competence and knowledge building, stakeholder communication, cross-functional collaboration, data governance, IT governance, MLOps, and AI design are emerging practices.

Competence and knowledge development, refers to a set of practices used to promote the abilities, expertise, and awareness necessary to implement ethical AI. Stakeholder communication refers to the set of communication practices organizations use to inform about their ethical AI practices, algorithms, or data. While the terms "AI design and development" and "data and AI governance" both refer to a set of operational decisions and practices that organizations use to address ethical concerns regarding the deployment, development, and use of AI systems, respectively, the former refers to the practical methods and practices that make up MLOps.

Building responsible AI systems being prime importance, the deep learning engineers need to be careful while designing task-specific deep learning systems. The models need to be carefully assessed for various data and tasks through numerous performance tests and explainability assessments, while keeping the privacy of the data secure and avoiding any algorithmic bias caused by noisy data or foreign attacks. Further, robust software should be incorporated into the CADs, and they should undergo vigorous penetration testing [45].

The following are a few best practices for responsible AI systems to overcome the limitations of conventional AI in healthcare:

a. Share information with the greater community, including research, tools, databases, and other resources.
b. Create trustworthy and efficient user-centered AI systems while considering machine learning-specific issues.
c. Keep an eye out for both long-term and short-term problems, estimate how the system will operate overall, and predict how users will react to future improvements.
d. Update systems regularly, considering user input and real-time performance while maintaining a good balance between simple and optimal solutions.
e. Conduct iterative integration and isolation tests on subsystems while considering a range of gold standard datasets, users, and use cases.
f. Recognize the constraints imposed by the dataset and model while communicating the depth and breadth of the training to the users.
g. Use a variety of indicators to understand the trade-offs between different errors and experiences.

Tools that produce reports on a model's operation and attempt to explain it are known as "explainable AI frameworks." SHAP (SHapley Additive exPlanations) is a model-independent approach that uses the shapley values from game theory to describe models. It illustrates how various features impact output or what role they play in the model's conclusion. LIME (Local Interpretable Model-agnostic Explanations) generates a list of explanations that show how each attribute affects a data sample's ability to predict the future. The Google-developed Whatif Tool (WIT) aids users in comprehending the operation of machine learning (ML) trained models, test performance regarding hypothetical scenarios, various data attributes, subsets of input data, and various machine learning fairness measures. IBM created the open-source AI Explainability 360 toolset to enhance the explainability of datasets and the interpretability of ML techniques. Skater is a framework that aids in the creation of interpretable machine learning systems, which are typically implemented in real use cases.

Potential opportunities of XAI in healthcare include but are not limited to maintenance and evolution in health care technologies, patient-specific health care, context-aware systems, robotic-assisted surgical procedure planning, medical image segmentation, targeted drug delivery, disease diagnosis and forecasting, treatment planning, basic and translational biomedical research, and survival analysis.

Although the study encompasses the challenges qualitatively, it does not consider the quantitative aspects of the demographics of different countries on continents where CADs could be implemented in healthcare. Many rural regions of the world do not have the facilities to implement such CADs, which causes a major restraint in achieving the supremacy of AI in healthcare. Future studies may examine the interactions between organizations involved in AI governance. The intersections of IT governance, data governance, and AI governance might be another area of future research.

While the explainability of AI and adversarial training were previously unexplored topics in medical imaging, significant research in these areas has produced groundbreaking results. CADs implemented with responsible AI systems, secure data protection protocols, and explainable tools could help healthcare flourish in the most rural areas of the globe, where medical staff is scarce. As new regulations and AI legislation take effect, it will be interesting to see how different organizations continue to assess the risks and adopt the AI standards, certifications, and audits that are anticipated to occur in the future.

V. Conclusion

Although deep learning faces barriers in its growth in healthcare, the future holds big prospects for AI in healthcare, from classifying diseases to probing AI supervised robots to perform surgical tasks. The growth of deep learning in medical image diagnosis can be achieved only by restoring the faith of humans in AI through awareness. While it is very important to explain to the user how systems deliver the output, how the model is developed, and what the impact of the model is, the deep learning-based designs should also be adaptively fine-tuned, considering the quality degradations with respect to time. Apart from resolving AI explainability issues, the deep learning-based designs must be made transparent without compromising the user's privacy and security and without any bias with respect to data collection, labelling, treatment, and model operations. Further, the technical, ethical, and societal factors of deep learning-based designs for medical image analysis can be optimized for better decision making and usability by combining responsible and XAI, which involves the complementing, co-creation, and coexistence of AI and humans.

REFERENCES

[1] A. Panesar, *Machine learning and AI for healthcare*. Springer, 2019.
[2] J. E. Estes, C. Sailer, and L. R. Tinney, "Applications of artificial intelligence techniques to remote sensing," *The Professional Geographer*, vol. 38, no. 2, pp. 133–141, May 1986, doi: 10.1111/j.0033-0124.1986.00133. x.
[3] K. Rajan and A. Saffiotti, "Towards a science of integrated AI and robotics," *Artificial Intelligence*, pp. 1–9, Jun. 2017, doi: 10.1016/j.artint.2017.03.003.
[4] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robotics*, vol. 37, no. 3, pp. 362–386, Apr. 2020, doi: 10.1002/rob.21918.
[5] C. Park, C. C. Took, and J.-K. Seong, "Machine learning in biomedical engineering," *Biomedical Engineering Letters*, vol. 8, no. 1, pp. 1–3, Feb. 2018, doi: 10.1007/s13534-018-0058-3.
[6] K. R. Foster, R. Koprowski, and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research-commentary," *Biomedical Engineering Online*, vol. 13, no. 1, pp. 1–9, Jul. 2014, doi: 10.1186%2F1475-925X-13-94.
[7] G. Olague, S. Cagnoni, and E. Lutton, "Introduction to the special

issue on evolutionary computer vision and image understanding," *Pattern Recognition Letters*, vol. 27, no. 11, pp. 1161–1163, Aug. 2006, doi: 10.1016/j.patrec.2005.07.013.

[8] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Apr. 2021, doi: 10.1007/s12525-021-00475-2.

[9] J. D. Blumenthal, A. Zijdenbos, E. Molloy, and J. N. Giedd, "Motion artifact in magnetic resonance imaging: implications for automated analysis," *Neuroimage*, vol. 16, no. 1, pp. 89–92, May 2002, doi: 10.1006/nimg.2002.1076.

[10] F. E. Boas and D. Fleischmann, "CT artifacts: causes and reduction techniques," *Imaging in Medicine*, vol. 4, no. 2, pp. 229–240, Apr. 2012, doi: 10.2217/iim.12.13.

[11] E. Manson, V. Ampoh, E. Fiagbedzi, J. Amuasi, J. Flether, and C. Schandorf, "Image noise in radiography and tomography: causes, effects and reduction techniques," *Current Trends in Clinical and Medical Imaging*, vol. 3, no. 4, pp. 86–91, Oct. 2019, doi: 10.19080/CTCMI.2019.02.555620.

[12] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. ISBI*, Washington, DC, USA, 2018, pp. 289–293, doi: 10.1109/ISBI.2018.8363576.

[13] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 12, pp. 2720–2730, Dec. 2018, doi: 10.1109/TBME.2018.2814538.

[14] A. Mehra, "Healthcare cloud computing market worth $89.4 billion by 2027", 2022. [Online]. Available: https://www.marketsandmarkets.com/PressReleases/cloud-computing-healthcare.asp

[15] United Nations, "Ensure heathy life and promote well-being for all ages," Department of Economic and Social Affairs, Sustainable development. [Online]. Available: https://sdgs.un.org/goals/goal3

[16] Lisa Hedges, "Software advice healthcare tech survey," 2021. [Online]. Available: https://www.softwareadvice.com/resources/healthcare-technology-trends

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013, doi: 10.48550/arXiv.1312.6199.

[18] Z. Liao, S. Hu, Z. Yu, and D. Sun, "Medical image blind denoising using context bilateral filter," in *Proc. MIACA*, Guangzhou, China, 2010, pp. 12–17, doi: 10.1109/MIACA.2010.5528280.

[19] D. Bhonsle, V. Chandra, and G. Sinha, "Medical image denoising using bilateral filter," *Int. J. Image, Graphics and Signal Processing*, vol. 4, no. 6, p. 36, Jul. 2012, doi: 10.5815/ijigsp.2012.06.06.

[20] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *Proc. ICDMW*, Barcelona, Spain. 2016, pp. 241–246, doi: 10.1109/ICDMW.2016.0041.

[21] W. Jifara, F. Jiang, S. Rho, M. Cheng, and S. Liu, "Medical image denoising using convolutional neural network: a residual learning approach," *J. Supercomputing*, vol. 75, no. 2, pp. 704–718, Feb. 2019, doi: 10.1007/s11227-017-2080-0.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 63, no. 11, pp. 139-144, Nov. 2020, doi: 10.1145/3422622.

[23] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, Dec. 2019, doi: 10.1016/j.media.2019.101552.

[24] Chen, Honglong, Shuai Wang, Nan Jiang, Zhe Li, Na Yan, and Leyi Shi. "Trust-aware generative adversarial network with recurrent neural network for recommender systems." *International Journal of Intelligent Systems*, vol.36, no. 2, pp.778-795, 2021, doi: 10.1002/int.22320

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 1125–1134, doi: 10.1109/CVPR.2017.632.

[26] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. CVPR*, Salt Lake City, UT, USA, 2018, pp. 8798–8807, doi: 10.1109/CVPR.2018.00917.

[27] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 2337–2346, doi: 10.1109/CVPR.2019.00244.

[28] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015, doi: 10.48550/arXiv.1511.06434.

[29] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, “Differential data augmentation techniques for medical imaging classification tasks,” in *Proc. AMIA Annual Symp.*, 2017, pp. 979–984. [Online]. Available: http://europepmc.org/article/MED/29854165.

[30] S. Sundaram and N. Hulkund, “Gan-based data augmentation for chest x-ray classification,” *arXiv preprint arXiv:2107.02970*, 2021, doi: 10.48550/arXiv.2107.02970.

[31] A. Bissoto, E. Valle, and S. Avila, “Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review,” in *Proc. CVPRW*, Nashville, TN, USA, 2021, pp. 1847–1856. doi: 10.1109/CVPRW53098.2021.00204.

[32] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *Proc. CVPR*, Seattle, WA, USA, 2020, pp. 8110–8119, doi: 10.1109/CVPR42600.2020.00813.

[33] 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, Dec. 2018, doi: 10.1016/j.neucom.2018.09.013.

[34] M. Kryuchkov, N. Khanzhina, I. Osmakov, and P. Ulyanov, “CT images GAN-based augmentation with AdaIN for lung nodules detection,” in *Proc. Int. Conf. Machine Vision*, Rome, Italy, 2021, vol. 11605, pp. 628–635, doi: 10.1117/12.2587940

[35] P. Minne, A. Fernandez-Quilez, D. Aarsland, D. Ferreira, E. Westman, A. W. Lemstra, M. T. Kate, A. Padovani, I. Rektorova, L. Bonanni, F. M. Nobili, M. G. Kramberger, J.-P. Taylor, J. Hort, J. Snædal, F. Blanc, A. Antonini, and K. Oppedal, “A study on 3D classical versus GAN-based augmentation for MRI brain image to predict the diagnosis of dementia with Lewy bodies and Alzheimer's disease in a European multi-center study,” *Medical Imaging 2022: Computer-Aided Diagnosis*, vol. 12033. pp. 624–632, Apr. 2022, doi: 10.1117/12.2611339.

[36] P. Zhang, J. Li, Y. Wang, and J. Pan, “Domain adaptation for medical image segmentation: a meta-learning method,” *J. Imaging*, vol. 7, no. 2, p. 31, Feb. 2021, doi: 10.3390/jimaging7020031.

[37] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, “Meta-learning in neural networks: A survey,” *arXiv preprint arXiv:2004.05439*, 2020, doi: 10.48550/arXiv.2004.05439.

[38] R. Singh, V. Bharti, V. Purohit, A. Kumar, A. K. Singh, and S. K. Singh, “Metamed: Few-shot medical image classification using gradient-based meta-learning,” *Pattern Recognition*, vol. 120, p. 108111, Dec. 2021, doi: 10.1016/j.patcog.2021.108111.

[39] Ahmadian, Sajad, Milad Ahmadian, and Mahdi Jalili. "A deep learning-based trust-and tag-aware recommender system." *Neurocomputing*, vol. 488, pp. 557-571, June 2022, doi: 10.1016/j.neucom.2021.11.064.

[40] P. R. Magesh, R. D. Myloth, and R. J. Tom, “An explainable machine learning model for early detection of Parkinson’s disease using lime on datscan imagery,” *Computers in Biology and Medicine*, vol. 126, pp. 104041, Nov. 2020, doi: 10.1016/j.compbiomed.2020.104041.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proc. ICCV*, Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[42] J. P. Cohen, L. Dao, K. Roth, P. Morrison, Y. Bengio, A. F. Abbasi, B. Shen, H. K. Mahsa, M. Ghassemi, H. Li, and T.Q. Duong “Predicting covid-19 pneumonia severity on chest x-ray with deep learning,” *Cureus*, vol. 12, no. 7, Jul. 2020, doi: 10.7759/cureus.9448.

[43] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, “Ca-net: Comprehensive attention convolutional neural networks for explainable medical image

segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021, doi: 10.1109/TMI.2020.3035253.
[44] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI—Two Frameworks for Ethical Design Practice," *IEEE Trans. Tech. and Soc.*, vol. 1, no. 1, pp. 34–47, Mar. 2020, doi: 10.1109/TTS.2020.2974991.
[45] A. K. Ghosh and G. McGraw, "An approach for certifying security in software components," in Proceedings of 21st NIST-NCSC National Information Systems Security Conference, 1998, pp. 42–48, doi: 10.1007/11563228_27

Tribikram Dhar has recently completed his B.S. in Electrical Engineering at Jadavpur University. His research interests are deep learning in biomedical image analysis, algorithm optimization and explainable AI.

Nilanjan Dey (Senior Member, IEEE) received his PhD from Jadavpur University in 2015. He is currently an Associate Professor in the Department of Computer Science and Engineering, Techno International New Town, Kolkata, India. He is also a visiting fellow of the University of Reading, UK and holds the position of Adjunct Professor at Ton Duc Thang University, Vietnam. Previously, he held an honorary position of Visiting Scientist at Global Biomedical Technologies Inc., CA, USA (2012–2015). Dr. Dey is the Editor-in-Chief of the International Journal of Ambient Computing and Intelligence. He is the Series Co-Editor of Springer Tracts in Nature-Inspired Computing, Data-Intensive Research Series, Co-Editor of Advances in Ubiquitous Sensing Applications for Healthcare. He is an associate editor of IET Image Processing and editorial board member of Complex & Intelligent Systems, Springer. He is the Indian Ambassador of the International Federation for Information Processing – Young ICT Group.

Surekha Borra (Senior Member, IEEE) received her PhD in 2015 from Jawaharlal Nehru Technological University, Hyderabad, India. She started her academic career as Assistant Professor in 2004 and served in various engineering colleges for 17 years. Currently, she is Professor in the Department of Electronics and Communication Engineering, K. S. Institute of Technology, Bengaluru, India. Dr Borra's research interests include Image and Video Analytics, Information Security and Signal Processing. She has received Woman Achiever's Award from The Institution of Engineers (India) for her prominent research and innovative contributions, and several research grants from the Government of Karnataka, India.

R. Simon Sherratt (Fellow, IEEE) received the B.Eng. from Sheffield City Polytechnic in 1992, both the M.Sc. in 1993, and Ph.D. in 1996 from The University of Salford. In 1996, he was appointed as a Lecturer in Electronic Engineering with the University of Reading, where he is currently a Professor of Biosensors. His research area is wearable devices, mainly for healthcare and emotion detection. Eur Ing Professor Sherratt was awarded the 1st place IEEE Chester Sall Award in 2004, 2nd place in 2014, 3rd place in 2015 and 3rd place in 2016 for best papers in the IEEE Transactions on Consumer Electronics. He is currently Chair of the IEEE Masaru Ibuka Consumer Technology Award.