AN ANALYSIS OF THE INTERPRETABILITY OF MACHINE LEARNING IN MEDICAL DIAGNOSIS

S. Shanmugapriya,L. Gnanaprasanambikai*

ABSTRACT

Machine learning has been really good at medical tasks, sometimes even better than doctors. But there's a big problem: these deep learning models are like locked boxes. They're hard to understand because they don't show how they make decisions. This makes it tough to use them in real medical situations because we need to trust and understand how they work.

To solve this problem, many studies have tried to make deep learning more understandable. In our paper, we review these efforts and what they've found. We look at the ways people have tried to make deep learning in medicine easier to understand, what they've used it for, how they've measured its success, and what data they've used. We also talk about the challenges and what researchers should focus on next.

Keywords: Machine learning, Methods of interpretation, Applications, Disease detection

I. INTRODUCTION

In recent times, A lot of fields now use machine learning as a strategy, like understanding language and working with images.[1] It often does much better than older ways of teaching computers. In medicine, we now have powerful computer programs that help doctors diagnose diseases using machine learning. These programs are really good at finding diseases and making diagnoses quickly.

But here's the problem: machine learning is like a locked box. It's hard for people to see how it decides things. Because of this, these advanced programs aren't as useful in real medical situations as they could be.

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
* Corresponding Author

Interpretability in machine learning models is a crucial topic gaining more attention, especially in the medical field. It helps make these complex models more transparent and trustworthy. While some interpretability methods exist, there's a lack of comprehensive review papers in medicine.

This paper aims to fill that gap by examining various interpretability techniques, their use in disease detection based on recent research, discussing challenges, and suggesting future directions. Ultimately, the goal is to encourage further exploration of interpretability in medicine and the development of clinical CAD systems.[2] Key contributions of this paper include providing an overview of current research, summarizing applications, and inspiring further advancements in this area.

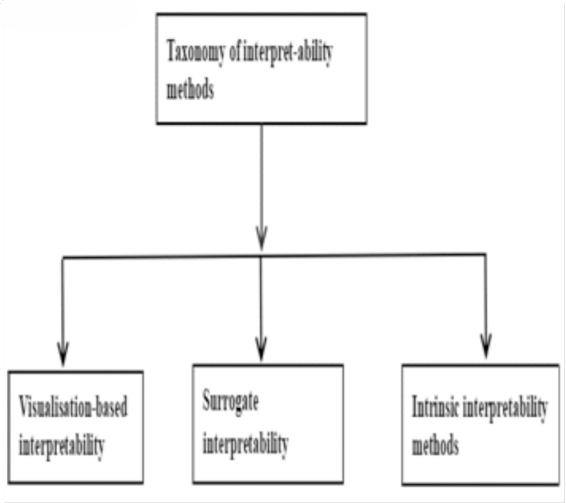

Fig.1 Techniques for interpreting data: a taxonomy

II. INTERPRETABILITY METHODOLOGY TAXONOMY

There isn't a universally agreed-upon definition for interpretability, but make the final outcomes easier to understand; however, numerous studies have suggested various approaches. [3-4] In general terms, there are two kinds of interpretability methods: "ante-hoc interpretability" and "post-hoc interpretability.

This section explains categorize and describe various regularly used interpretability methods in the medical profession. It's important to note that classification approaches aren't set in stone; they can overlap or create different categories based on the specific features of interpretability methods. You can see a visual representation of this taxonomy in Figure 1.

A. Visualization-based interpretability methods

Visualization-based interpretability methods in machine learning are techniques that use graphical or visual representations to help users understand how a model works and why it makes specific predictions. These methods leverage visualization to make complex model behaviors more accessible. Visualization-based interpretability methods provide intuitive and transparent insights into machine learning models, making them more accessible to domain experts and stakeholders who may not be familiar with the technical intricacies of the models themselves.

1. Methods of back-propagation interpretability

Methods of back-propagation interpretability in machine learning are techniques that utilize information from the back-propagation process of neural networks to learn about the behavior and thought processes of the model. Back-propagation is a fundamental algorithm used for training neural networks, and it can also be leveraged for interpretability purposes.

Back-propagation interpretability methods harness the principles of neural network training to explain how these models arrive at their predictions.[5] They are particularly valuable for gaining insights into deep learning models, especially in complex applications like linguistic processing and computer vision.

2. Methods of CAM-based interpretability

CAM-based(Class Activation Map-based) interpretability methods in machine learning are techniques that use Class Activation Maps to visualize and understand how a neural network, particularly convolutional neural networks (CNNs), focuses on different regions of an input image when making predictions. These methods help identify which parts of an image are important for a particular class or category prediction. These methods are especially valuable for image-related tasks as they offer insights into where a model pays attention within an image, making it easier to comprehend the decision-making process [6-8] of CNNs in image classification, object detection, and related applications.

3. Methods of interpretability based on perturbations

Methods of interpretability based on perturbationsare techniques used in machine learning to understand and interpret the predictions made by complex models. With these techniques, the input data is subjected to slight adjustments or perturbations, and the output of the model is then evaluated in response to these adjustments.The response of the model to these perturbations can be studied, we can gain insights into its decision-making process and identify important features or patterns that influence its predictions. These methods are valuable tools for gaining insights into complex machine learning techniques, especially deep learning techniques, which can be challenging to interpret. They provide a means to assess feature importance, model behavior, and robustness, helping users build trust in AI systems and make informed decisions based on model predictions [11].

B. Methods of surrogate interpretability

Techniques substitute interpretability, which are intended to make complex "black-box" models more justifiable via preparing interpretable models like shallow choice trees or straight models to imitate the way of behaving of the mind-boggling models. They explicitly notice two famous strategies, one of which is Neighborhood Interpretable Model-skeptic Clarification (LIME).

The main idea behind [9-10] is to take inputs used by the complex model, analyze them, and then use these insights to build simpler networks that come close to the black-box imitation predictions.This guess interaction permits LIME to give interpretability to profound learning models by looking at these improved, interpretable organizations.

The following equation can be used to explain the LIME approach of interpretability:

$(x) = \text{argmin}_{6} \in GL(f, g, \prod x) + \Omega(g)$

In simpler terms, this equation represents the process of finding an interpretable model (g) that closely approximates the way of behaving of the perplexing model (f) within a given locality $\prod$ x while also considering a regularization term Ω (g). The objective is to make an additional justifiable and locally reliable model to make sense of the expectations of the first perplexing model.

In simpler terms, LIME seeks to find a simple and interpretable model (represented by 'g') that, when applied to the analyzed inputs ($\prod$ x), closely approximates the predictions of the complex model (f). The goal is to minimize The variations among the predictions of the straightforward model and the complicated model while also taking into account a regularization term Ω (g) to ensure the simplicity of the interpretable model.

Overall, the paragraph explains that surrogate interpretability methods like LIME aim to make complex technique more interpretable by approximating their behavior with simpler models, which can help in understanding the decisions made by machine learning models.

1. LIME

LIME, an abbreviation for "Nearby Interpretable Model-rationalist Clarifications," is a system utilized inside the domain of machine learning to elucidate the rationale behind the decisions made by intricate machine learning models. It proves especially valuable when dealing with models that resist easy comprehension, for example, profound brain organizations or ensembletechnique,as well as goal is to gain insights into the reasons behind specific predictions for individual instances or a set of instances [12,13]. Here is a breakdown of the key steps involved in the LIME process:

- Select Instances
- Perturbation
- Model Fitting
- Local Interpretation Model
- Feature Importance
- Explain ability

2. Knowledge condensation

Knowledge condensation is a technique used to transfer the knowledge learned by a preplex or "teacher" model to anequivalent or "student" model. It is often employed when you have a large, accurate, and computationally expensive model and you want to create a smaller, faster, and more efficient model that approximates the behavior of the teacher model.

Here's how knowledge distillation works:

- Teacher Model
- Student Model
- Soft Targets
- Loss Function

3. Intrinsic interpretability methods

Intrinsic interpretability in machine learning refers to the inherent transparency and explain ability of a model's decision-making process without the need for additional post-hoc techniques or tools. In other words, an intrinsically interpretable machine learning model is designed in a way that makes it inherently easy to understand and interpret how it arrives at its predictions or classifications. [14,15].

Here are some key aspects of intrinsic interpretability in machine learning:

- Simple Model Architectures&
- Additive Models&Rule-Based Models
- Sparse Models
- Local Interpretability&Visualisation
- Model Documentation
- Monotonicity Constraints
- Domain-Specific Constraints

III. APPLICATIONS OF INTERPRETABILITY TECHNIQUES IN DISEASE DIAGNOSIS

Interpretability techniques in machine learning and artificial intelligence play a crucial role in disease diagnosis by providing insights into how models arrive at their predictions

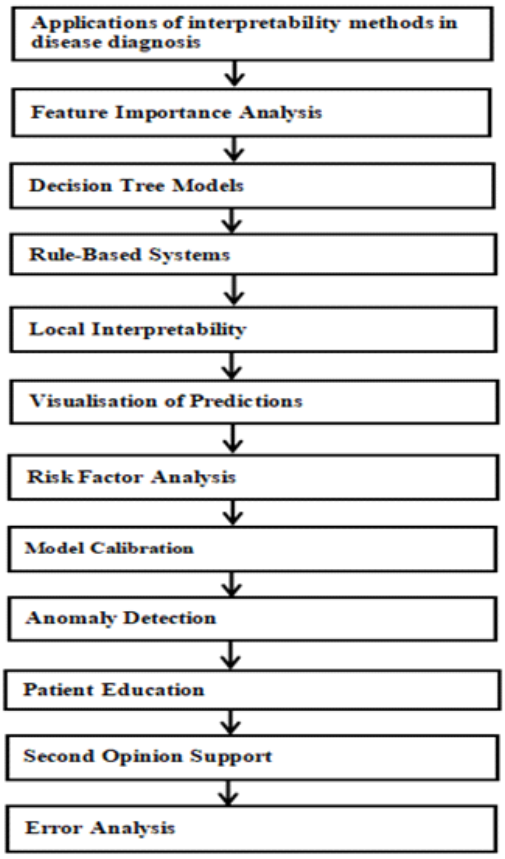

Fig.2 Interpretability approaches used in disease diagnosis

Interpretability strategies in illness diagnosis:

1. Eye diseases
2. Lung diseases
3. Brain diseases
4 .Other diseases

A. Applications in eye diseases

Machine learning has made significant advancements in the field of ophthalmology and eye diseases. It has been applied in various ways to assist with diagnosis, treatment, and management of eye conditions. Here are some key applications of machine learning in eye diseases:

Ophthalmology Machine Learning:Machine learning has transformed the diagnosis and treatment of numerous eye disorders, making it a valuable tool in the area of ophthalmology.

Diabetic Retinopathy Detection: Diabeticeye disease, a primary cause of blindness in diabetic patients, can be identified early using machine learning algorithms that analyze retinal images, enabling timely intervention.

Advancements in Glaucoma Diagnosis: Glaucoma, a progressive eye disease leading to irreversible vision loss, benefits from machine learning models that analyze optic nerve images and visual field tests to detect early signs and monitor disease progression.

Early ARMD (Age-Related Macular Degeneration) detection: Early ARMD (Age-Related Macular Degeneration) detection is widely used by older adults. Machine learning aids in the early identification of AMD by analyzing retinal images, spotting drusen, and other characteristic features.

Streamlined Cataract Diagnosis: The automation of cataract detection is achieved through machine learning, which analyses lens images to facilitate early diagnosis and timely surgical intervention.

Diagnosing Various Retinal Diseases: Machine learning models are instrumental in diagnosing a range of retinal diseases, including retinal detachment, retinitis pigmentosa, and macular holes, by scrutinizing retinal images for abnormalities.

Enhancing(OCTA) Optical Coherence Tomography Angiography:OCTA filters, a harmless imaging innovation, benefit from machine learning algorithms that aid in diagnosing and monitoring retinal diseases.

Optimizing Visual Field Testing: Automated perimetry, used to assess the visual field in glaucoma patients, is enhanced by machine learning, which delivers more accurate and efficient results.

Revolutionizing Drug Discovery: Machine learning is employed in identifying potential drug candidates for treating eye diseases by analyzing molecular and genetic data, predicting the efficacy of novel compounds.

Telemedicine and Remote Monitoring Solutions: Machine learning enables telemedicine applications for remotely monitoring eye diseases. Patients can capture retinal images at home, and healthcare providers receive analyzed data and feedback.

Tailored Treatment Plans: Personalized treatment plans for patients with eye diseases are developed with the assistance of machine learning. Calculations investigate patient information and clinical records, suggesting the best treatment choices

Exploring Prosthetic Vision: The exploration of prosthetic vision devices using machine learning, which harnesses cameras and computer vision algorithms to provide visual information to individuals with vision impairments.

Predicting Patient Outcomes: Machine learning assists in predicting the progression of eye diseases and forecasting outcomes for individual patients, aiding in treatment decisions and patient counselling.

B. Applications in lung diseases

Machine learning has also found numerous applications in the field of lung diseases, [16-18] ranging from early detection and diagnosis to treatment optimization and research. Here are some applications in lung diseases:

Pulmonary Disease Diagnosis: Machine learning models can break down clinical imaging information, for example, chest X-beams and CT checks, to aid in the diagnosis of certain pulmonary diseases,pneumonia, lung cancer, and tuberculosis, to name a few.

Cancer Detection and Diagnosis:Machine learning calculations can be prepared to perceive examples and peculiarities in clinical imaging, possibly further developing endurance rates.

Drug Discovery: Machine learning is used in drug discovery for lung diseases, helping researchers identify potential drug candidates and predict their effectiveness through molecular and genetic data analysis.

Predictive Modelling for COPD: Chronic Obstructive Pulmonary Disease (COPD) progression and exacerbation can be predicted using machine learning algorithms. This allows for timely interventions and management strategies.
Asthma Management: Machine learning can assist in the management of asthma by predicting asthma attacks based on environmental data, patient history, and physiological measurements, enabling patients to take preventive measures.

Radiomics:Radiomics includes the extraction of quantitative highlights from clinical pictures.Machine learning can analyze these radiomic features to provide insights into lung diseases' characteristics and behavior.

Lung Function Assessment: Machine learning models can predict lung function parameters, such as forced expiratory volume (FEV1), based on patient data, helping in the diagnosis and monitoring of respiratory conditions.

Telemedicine for Remote Monitoring: Machine learning enables remote monitoring of lung disease patients. Smart devices and wearables can collect data on lung function, symptoms, and medication adherence, which is then analyzed by machine learning algorithms.

Treatment Optimization: Machine learning can assist healthcare providers in optimizing treatment plans for lung disease patients by analyzing patient data, treatment responses, and potential side effects.

Patient Outcome Prediction: Machine learning models can predict patient outcomes and disease progression for conditions like idiopathic pulmonary fibrosis (IPF), aiding in treatment decisions and patient counselling.

Air Quality Monitoring: Machine learning can be used to predict air quality and pollution levels, which are crucial for individuals with respiratory conditions like asthma and COPD.

Clinical Decision Support: Machine learning support decision support tools for healthcare providers, assisting in selecting the most appropriate tests, treatments, and interventions for patients with lung diseases.

These applications demonstrate how machine learning is contributing to the early detection, diagnosis, treatment, and management of various lung affection, eventually enhancing patient outcomes and progressing related research.

C. Applications in Brain diseases

The study of neuroscience and medical diagnosis greatly benefits from machine learning, treatment, and understanding of brain diseases. Here are some applications of machine learning in brain diseases:

Brain Tumor Detection:Brain tumors can be identified and categorized using medical imaging data from MRI and CT scans analyzed by machine learning models. This aids in early diagnosis and treatment planning. [19,20]

Alzheimer's Disease Diagnosis: Machine learning can aid the earlier analysis of dementia or Alzheimer's sickness by dissecting cerebrum imaging,genetic data, and cognitive assessments, helping with timely intervention.

Parkinson's Disease Diagnosis: Machine learning algorithms can analyze voice and movement data to assist in the early analysis of Parkinson's infection and screen its movement.

Stroke Prediction and Prevention: Machine learning models can predict a person's risk of having a stroke based on various risk factors, including medical history, lifestyle, and genetics, allowing for preventive measures.

Seizure Prediction: Machine learning can be used to develop predictive models for epilepsy patients to forecast seizure occurrences, enabling timely medication or intervention.

Neuroimaging Biomarkers: Machine learning can identify neuroimaging biomarkers associated with specific brain diseases, aiding in disease characterization and understanding.

Brain-Computer Interfaces (BCIs): Machine learning is used in BCIs to decode brain signals and assist individuals with neurological disorders in controlling external devices or prosthetics.

Treatment Planning: Machine learning can optimize treatment plans for brain diseases, such as determining the most effective treatment options for brain tumors or suggesting personalized medication regimens for neurological conditions.

Mental Health Assessment: Machine learning models can assist in the assessment of mental health conditions by analyzing speech, text, and behavioral data, helping clinicians make more accurate diagnoses and treatment recommendations.

Neurodegenerative Disease Monitoring: Machine learning can continuously monitor and analyze patient data to track the progression of neurodegenerative diseases, providing valuable insights for healthcare providers.

Neurological Rehabilitation: Machine learning-driven rehabilitation programs can provide personalized exercises and interventions for individuals recovering from brain injuries or strokes.

Genomic Analysis:Machine learning can be used to evaluate genomic data in order to find genetic risk factors for brain illnesses and contribute to personalized treatment techniques.

Drug Discovery: Machine learning assists in drug discovery for brain diseases by predicting potential drug candidates and their efficacy in treating neurological conditions.

Neuroscience Research: Machine learning is used in analysing large-scale neuroscience data, including brain connectivity networks and neural activity patterns, to gain insights into brain function and disorders.

Telemedicine for Remote Consultations: Machine learning-powered medicine platforms allow patients with neurological conditions to consult with specialists remotely, improving access to care.

These applications demonstrate the versatility of machine learning in addressing various brain diseases, from diagnosis and treatment to research and patient support, ultimately advancing our understanding of the brain and improving the lives of individuals affected by these conditions.

D. Aapplicability to other diseases

Besides the aforementioned interpretable techniques, various transparent and easily understandable approaches have been introduced for diagnosing prevalent medical conditions. In this context, we will primarily emphasize interpretable applications concerning three frequently occurring ailments: heart diseases, dermatological conditions, and breast disorders.

1. Heart disease diagnosis

Iin the heart undeniablyamong the body's vital organs, is in charge of supplying all tissues with blood. When heart disease develops, it poses a significant threat to human health. Currently, various research studies have been dedicated to detecting heart diseases by scrutinizing alterations in electrocardiogram (ECG) patterns [21,22].

Additionally, Puyol-Antón et al. [29] fostered a structure for picture-based characterization, integrating clinical information into the model through an optional classifier. This addition enhances classification results interpretability and empowers clinicians to get a handle on the thinking behind the model's choices.While experimental results indicated performance similar to the baseline VAE, the critical novelty lies in the technique interpretability.

In addition to it, Agha Mohammadi et al. [30] reported a useful categorization technique for predicting cardiac attacks. This method combines the capabilities of evolutionary algorithms, neural networks, and fuzzy logic, and is supplemented by the use of logical diagrams to give interpretability to conclusive expectations. The experimental outcomes affirm the satisfactory performance of this proposed algorithm.

Additionally, a novel architecture named SAU-Net [31] was introduced for cardiac MRI image segmentation, prioritizing interpretability and model robustness. Notably, it achieves dual attention decoder module to provide multi-level interpretability. Unlike other post-hoc interpretable methods, SAU-Net can outfit different degrees of interpretability and multi-objective saliency maps can be used during the forward pass without adding extra computational overhead.

Skin diseases diagnosis

A Skin cancershown a consistent increase in its incidence, becoming a growing concern year after year [23]. Consequently, numerous interpretable applications have arisen to support the compelling analysis of skin infections. Barata et al. [24] spearheaded the improvement of a CAD (Computer-Aided Diagnosis) framework for skin sore finding.

Similarly, a consideration-driven U-Net engineering [25] was introduced for the programmed discovery of skin injuries,as input, multi-scale photos are used. Each lesion property is regarded as its own network. addressing class imbalance issues. Remarkably, it achieves optimal skin lesion detection results with reduced parameters and computational requirements.

Additionally, a lightweight deep learning architecture with an attention mechanism [26] was designed to differentiate among hhistopathological photos of 11 different forms of skin disorders. To provide visual explanations for deep model predictions, it leverages the Class Activation Mapping (CAM) approach. This approach not only identifies the disease but also highlights the areas that are impacted.

Gu et al. [27] introduced the CA-Net engineering for skin sore division, which utilizes a far-reaching consideration-based brain organization. Dissimilar to other interpretable methodologies that just dissect spatial data, this strategy consolidates spatial, channel, and scale consideration, giving a thorough clarification of expectation discoveries. Importantly, it computes attention coefficients directly without the need for further calculations.

Moreover, Stieler et al. [28] family classifier that merges the LIME methodology with the ABCD-principle, a characteristic procedure used by dermatologists.This approach distinguishes between melanocytes and non-melanocytes. While it allows for the identification of significant features, the challenge lies in translating this information into a user-friendly explanation.

These innovative approaches aim to facilitate accurate skin disease diagnosis while providing transparent and understandable insights for medical practitioners and patients alike.

IV. DATASETS AND MMEASUREMENT METRICS

In this part, our consideration goes to the assessment of normal assessment models and datasets inside the clinical field. These resources have been gathered from the practical applications of interpretable methods discussed in Section 3. Although these measurements and datasets are crucial for evaluating how accurately diseases are diagnosed, it's essential to note that there exists a dearth of standardized evaluation metrics for interpretability in this context. As a result, much of the assessment of interpretability relies on human judgement. In spite of this, the consideration of these evaluation rules is as yet significant while surveying the interpretability execution of machine learningtechnique in the clinical field, with the ultimate objective of giving our readers useful references.

Measurement metrics

Evaluation metrics is a fundamental method for estimating the presentation of the models, where the disarray lattice can work out a few measurements (for example exactness, accuracy, and so forth.) by ascertaining the quantity of genuine positive (GP), bogus positive (BP), genuine negative (GN) and bogus negative (BN). They can be indicated as:(33)

Accuracy=(GP+GN)/(GP+BP+BN+GN)

Exactness=GP/(GP+BP),

Recall(Sensitivity)=GP/(GP+BN),

Specificity=GN/(GN+BP)

F1=2* (exactness*recall)/(exactness+recall)

Moreover, we offer two famous measurements, Dice coefficient and Jaccard record (IoU), in view of cross-over estimations, to assess the model's exhibition. We will guess that A and B are two given sets. They can then be addressed as:

$$Dice,(A,B)=(2|A \cap B|)/(|A|+|B|)$$

$$Jaccard,(A,B)=(|A \cap B|)/(\ |A \cup B|)$$

V. CONCLUSION

Interpretability has garnered significant attention, particularly within the medical sector, as it serves as an efficient means to demystify the "black-box" nature of machine learning technique and establish entrust with users. Throughout this document, we have conducted a comprehensive review of various interpretability techniques extensively applied in the medical domain. By categorizing disease locations and outlining diverse applications with interpretability, we have revealed insight into both the current difficulties and potential exploration headings.With the introduction of this paper, we desire to provide perusers with a careful handle of the condition of interpretability in the field of medication,using the ultimate goal of facilitating its swift integration into clinical practice.

REFERENCES

1. Chrysostomou, G., Aletras, N.: Improving the faithfulness of attention-based explanations with task-specifc information for text classifcation (2021). at preprint arxiv:2105.02657

2. Sun, M., Huang, Z., Guo, C.: Automatic diagnosis of alzhei mer's disease and mild cognitive impairment based on cnn+ svm networks with end-to-end training. In: 2021 13th International Conference on Advanced Computational

Intelligence (ICACI), pp. 279–285 (2021) https://doi.org/10.1109/ICACI52617.2021.9435894. IEEE

3. Da Cruz, H.F., Pfahringer, B., Martensen, T., Schneider, F., Meyer, A., Böttinger, E., Schapranow, M.-P.: Using interpretability approaches to update black-box clinical prediction models: an external validation study in nephrology. Artif. Intell. Med. 111, 101982 (2021). https://doi.org/10.1016/j.artmed.2020.101982

4.Pedapati, T., Balakrishnan, A., Shanmugam, K., Dhurandhar, A.: Learning global transparent models consistent with local contras tive explanations. Adv. Neural. Inf. Process. Syst. 33, 3592–3602 (2020)

5.Qin, Z., Yu, F., Liu, C., Chen, X.: How convolutional neural network see the world-a survey of convolutional neural network visualization methods (2018). at print.

6. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Pro ceedings of the IEEE Conference on Computer Vision and Pat tern Recognition, pp. 2921–2929 (2016)

7. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 24–25 (2020)

8. Lee, J.R., Kim, S., Park, I., Eo, T., Hwang, D.: Relevance-cam: Your model already knows where to look. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14944–14953 (2021)

9.Pintelas, E., Livieris, I.E., Pintelas, P.: A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms 13(1), 17 (2020). https://doi.org/10. 3390/a13010017

10. Mi, J.-X., Li, A.-D., Zhou, L.-F.: Review study of interpretation methods for future interpretable machine learning. IEEE Access 8, 191969–191985 (2020). https://doi.org/10.1109/ACCESS.2020.3032756

11. Joshi, A., Mishra, G., Sivaswamy, J.: Explainable disease classifcation via weakly-supervised segmentation. In: Interpretable and Annotation-Efcient Learning for Medical Image Computing, pp. 54–62 (2020). https://doi.org/10.1007/978-3-030-61166-8_6

12. Ras, G., Xie, N., van Gerven, M., Doran, D.: Explainable deep learning: a feld guide for the uninitiated. J. Artif. Intell. Res. 73, 329–397 (2022). https://doi.org/10.1613/jair.1.13200

13. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Gradient-based attribution methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 169–191. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_9

14.Morafah, R., Karami, M., Guo, R., Raglin, A., Liu, H.: Causal interpretability for machine learning-problems, methods and evaluation. ACM SIGKDD Explor. Newslett. 22(1), 18–33 (2020). https://doi.org/10.1145/3400051.3400058

15.Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R.: Explainable artifcial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai.Inf. Fus. 58, 82–115 (2020). https://doi.org/10.1016/j.infus. 2019.12.012

16. Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., Yu, Y.: Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10632–10641 (2019)

17. Guan, Q., Huang, Y.: Multi-label chest x-ray image classifcation via category-wise residual attention learning.

Pattern Recogn. Lett. 130, 259–266 (2020). https://doi.org/10.1016/j.patrec.2018. 10.027

18. Huang, Z., Fu, D.: Diagnose chest pathology in x-ray images by learning multi-attention convolutional neural network. In: 2019 IEEE 8th Joint International Information Technology and Artifcial Intelligence Conference (ITAIC), pp. 294–299 (2019). https://doi.org/10.1109/ITAIC.2019.8785431. IEEE

19. Tang, Z., Chuang, K.V., DeCarli, C., Jin, L.-W., Beckett, L., Keiser, M.J., Dugger, B.N.: Interpretable classifcation of Alzheimer's disease pathologies with a convolutional neural network pipeline. Nat. Commun. 10(1), 1–14 (2019)

20. Nigri, E., Ziviani, N., Cappabianco, F., Antunes, A., Veloso, A.: Explainable deep cnns for mri-based diagnosis of alzheimer's disease. In: 2020 International Joint Conference on Neural Net works (IJCNN), pp. 1–8 (2020) IEEE

21.Van Steenkiste, T., Deschrijver, D., Dhaene, T.: Interpretable ecgbeat embedding using disentangled variational auto-encoders. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), pp. 373–378 (2019) https://doi.org/10.1109/CBMS.2019.00081. IEEE

22. Mousavi, S., Afghah, F., Acharya, U.R.: Han-ecg: an interpretable atrial fbrillation detection model using hierarchical attention networks. Comput. Biol. Med. 127, 104057 (2020). https://doi.org/10.1016/j.compbiomed.2020.104057

23. Jones, O.T., Ranmuthu, C.K., Hall, P.N., Funston, G., Walter, F.M.: Recognising skin cancer in primary care. Adv. Ther. 37(1), 603–616 (2020)

24. Barata, C., Celebi, M.E., Marques, J.S.: Explainable skin lesion diagnosis using taxonomies. Pattern Recogn. 110, 107413 (2021). https://doi.org/10.1016/j.patcog.2020.107413

25. Nguyen, D.M.H., Ezema, A., Nunnari, F., Sonntag, D.: A visually explainable learning system for skin lesion detection using multiscale input with attention u-net. In: German Conference on Artifcial Intelligence (KünstlicheIntelligenz), pp. 313–319 (2020) https://doi.org/10.1007/978-3-030-58285-2_28 Springer

26. Jiang, S., Li, H., Jin, Z.: A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. IEEE J. Biomed. Health Inform. 25(5), 1483–1494 (2021). https://doi.org/10.1109/JBHI.2021.3052044

27. Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging 40(2), 699–711 (2020). https://doi.org/10.1109/TMI.2020.3035253

28. Stieler, F., Rabe, F., Bauer, B.: Towards domain-specifc explain able ai: model interpretation of a skin image classifer using a human approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1802–1809 (2021)2355 A survey on the interpretability of deep learning in medical diagnosis 1 3

29. Puyol-Antón, E., Chen, C., Clough, J.R., Ruijsink, B., Sidhu, B.S., Gould, J., Porter, B., Elliott, M., Mehta, V., Rueckert, D.: Interpret able deep models for cardiac resynchronisation therapy response prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 284–293 (2020) https://doi.org/10.1007/978-3-030-59710-8_28.Springer

30. Aghamohammadi, M., Madan, M., Hong, J.K., Watson, I.: Predicting heart attack through explainable artifcial intelligence. In: International Conference on Computational Science, pp. 633–645 (2019) https://doi.org/10.1007/978-3-030-22741-8_45. Springer

31.Sun, J., Darbehani, F., Zaidi, M., Wang, B.: Saunet: Shape attentive u-net for interpretable medical image segmentation.

In: International Conference on Medical Image Computing and 2352 Q. Teng et al. 1 3 Computer-Assisted Intervention, pp. 797–806 (2020) https://doi.org/10.1007/978-3-030-59719-1_77.Springer

31. Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.-M., Tengg-Kobligk, H.V., Summers, R.M., Wiest, R.: On the interpretability of artifcial intelligence in radiology: challenges and opportunities. Radiology 2(3), 190043 (2020). https://doi.org/10.1148/ryai.2020190043

32. Qin, Z., Yu, F., Liu, C., Chen, X.: How convolutional neural network see the world-a survey of convolutional neural network visualization methods (2018). at print.

33. Morafah, R., Karami, M., Guo, R., Raglin, A., Liu, H.: Causal interpretability for machine learning-problems, methods and evaluation. ACM SIGKDD Explor. Newslett. 22(1), 18–33 (2020). https://doi.org/10.1145/3400051.3400058