INTERPRETABLE DEEP LEARNING METHOD FOR MEDICAL IMAGE DIAGNOSIS

Manziuk E.[1], Barmak O.[1], Krak I.[2,3], Petliak N.[1], Jin Zh.[4], Radiuk P.[1]

1. Dept. of Computer Science,
Khmelnytskyi National University 11, Instytuts'ka str., Khmelnytskyi, 29016, Ukraine
eduard.em.km@gmail.com, alexander.barmak@gmail.com,
npetlyak@khmnu.edu.ua, radiukp@khmnu.edu.ua

2. Dept. of Theoretical Cybernetics
Taras Shevchenko National University of Kyiv, 60 Volodymyrska str., Kyiv, 01033, Ukraine
3 Glushkov Cybernetics Institute, 40 Glushkov ave., Kyiv, 03680, Ukraine
yuri.krak@gmail.com

4 Key Laboratory of Disaster Prevention and Structural Safety of Ministry of Education,
Guangxi Key Laboratory of Disaster Prevention and Engineering Safety,
Guangxi University, Nanning 530004, China
sdkjdxjz@163.com

Incorporating artificial intelligence into the medical field holds immense potential, but it also raises significant challenges that must be addressed to ensure patient safety and ethical practices. While AI can enhance efficiency and support decision-making processes, its application in healthcare demands utmost caution and rigorous safeguards [1].

One of the primary concerns is the risk of erroneous decisions, which could have severe consequences for patients' well-being. Even highly accurate AI systems may occasionally generate incorrect recommendations, and there is a need for robust error detection mechanisms tailored to the medical domain [2]. Additionally, the interpretability of AI models is crucial, as healthcare professionals and patients alike should be able to comprehend the rationale behind AI-assisted diagnoses and treatment plans.

To foster trust and transparency, incentives may be required to encourage the development of interpretable AI models specifically designed for medical applications [3, 4]. These models, although potentially less accurate than their black-box counterparts, could offer valuable insights into the decision-making process, enabling more informed and collaborative decision-making between AI systems and human experts.

The integration of AI in healthcare raises questions of accountability and liability. Clear regulations must be established to delineate responsibilities in cases where AI systems contribute to adverse outcomes. Ethical considerations, such as preserving patient privacy and maintaining strict confidentiality standards, are also paramount when handling sensitive medical data.

AI systems in the medical field must be adaptive and capable of continuous learning, as medical knowledge and treatment protocols evolve over time. Regularly updating and refining these systems to align with the latest research and best practices is essential to ensure their long-term relevance and effectiveness. While the potential benefits of AI in healthcare are undeniable, addressing these challenges through robust oversight, ethical frameworks, and ongoing research and development is crucial to unlocking the full potential of this transformative technology while prioritizing patient safety and well-being. The introduction of AI in this area contributes to a radical transformation of the information space for its practical application. The synergistic effect of such integration blurs the line between theoretical research in the field of AI and its practical implementation in medicine through the development of specialized medical AI systems. The objective realities of practical use necessitate the cre- ation of AI solutions that take into account ethical aspects, comply with legal standards and build trust in these technologies.

We considered one basic model using a convolutional neural network. The results of this model were taken as the reference, according to which the decisions of the neural network were compared, and the results of its work were interpreted. The scheme of classifier formation ground on the interpreted model is shown in Figure 1.

The interpretability is determined by the DRN network, then intdecisions. Interpretability is a function of features and is determined on a certain set of features intdecisions, so we define interpretability by features.

One of the important aspects of the proposed method is that the DRN neural network interprets the decisions within the set of features F . However, the inter- pretation is based on the connections determined by the DRN neural network. This means that the correctness of the decision interpretation depends on how accurately the DRN network recognizes the relationships between the features. Hence, it can determine how correctly the decisions will be interpreted. Let's denote by int a certain conditional interpretability of a decision.

The evaluation of the DRN and VGG-16 models on the dataset revealed notable differences in their performance metrics. While the DRN model offered interpretable decisions, a desirable trait in medical applications, its precision, recall, and accuracy rates were significantly lower compared to the VGG-16 model. Specifically, the DRN model achieved a minimum precision rate of 0.65 and a minimum recall rate of 0.68 across all classes, resulting in an overall accuracy rate of 0.76. In contrast, the VGG-16 model demonstrated superior performance, surpassing the DRN model in these key

metrics. To leverage the strengths of both models, a hybrid approach was proposed. In cases where the decisions of the two models align, the DRN's interpretable decision can be utilized, benefiting from its ability to provide explanations for its predictions. However, when the decisions diverge, the VGG-16 model's decision is considered the reference, given its higher accuracy. To reconcile the discrepancies between the models' decisions, the feature set used by the DRN is modified, and the network is retrained to align its predictions with those of the VGG-16 model for the same input images.

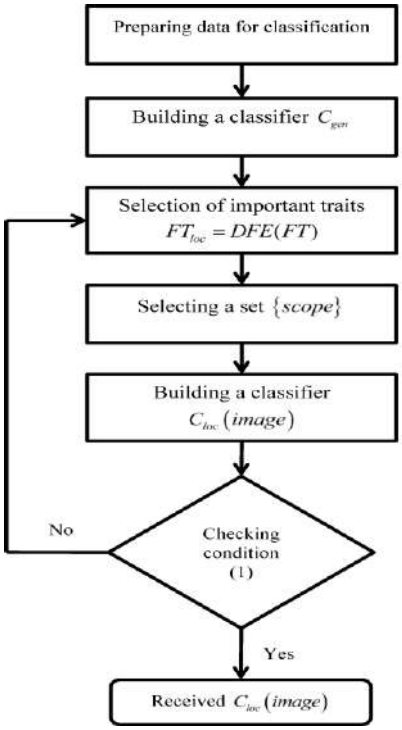

Figure 1. Scheme of classifier formation based on the interpreted classifier

The ROC curves for both models, as illustrated in Figure 2, provide a visual representation of their performance trade-offs between sensitivity and specificity. While the DRN's interpretability is valuable in the medical domain, its lower accuracy compared to the VGG-16 model necessitates a careful balancing act. The proposed hybrid approach aims to leverage the strengths of both models, prioritizing accuracy while incorporating interpretability when possible, ultimately enhancing the reliability and trustworthiness of AI-assisted medical decision-making.

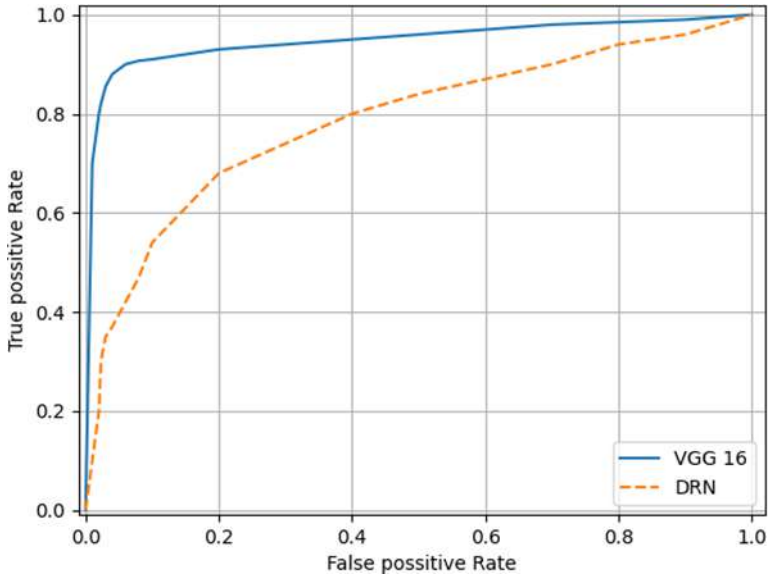

Figure 2. ROC curve of classification of neural networks VGG-16 and DRN

The use of DRN on the dataset gave the minimum precision and recall rates by class of 0.65 and 0.68, respectively. The accuracy rate is 0.76. Compared to the results obtained using VGG-16, these indicators are significantly lower. The use of DRN allows us to obtain an interpreted decision. However, this network gives correct decisions in fewer cases

compared to VGG-16, so it can be used properly in cases where the decisions of neural networks coincide. If the decisions do not coincide, the VGG-16 decision is taken as the reference decision. In this case, the set of features is changed and the DRN network is rebuilt to obtain a decision that corresponds to the VGG-16 decision for the same image.

The results show that the DRN manages to extract interpreted rules from the set of features used by the VGG-16 model for brain image classification. However, it is crucial to note that the precision, recall, and accuracy of the DRN are lower than those obtained from VGG-16. This indicates that DRN is able to extract the interpreted rules, but in some cases its decisions may be less accurate than VGG-16.

The integration of the DRN and VGG-16 models presents a promising approach to enhance the accuracy and interpretability of brain disease diagnosis. By leveraging the strengths of both models, this strategy aims to strike a balance between reliable classification and transparency in decision-making. In cases where the decisions of the DRN and VGG-16 models align, the DRN's interpretable nature can be leveraged to provide insights into the key features that influenced the classification outcome. This interpretability is particularly valuable in the medical domain, as it enables healthcare professionals to analyze and understand the rationale behind the model's predictions. Conversely, when the decisions of the two models diverge, the VGG-16 model's higher accuracy takes precedence, ensuring that the final diagnosis is based on the most reliable classification output. This approach acknowledges the VGG-16 model's superior performance, as evidenced by its higher Area Under the Curve (AUC) value of 0.948, compared to the DRN's AUC of 0.786. By combining the strengths of both models, this hybrid approach offers additional confidence in the diagnostic process. While the VGG-16 model serves as the primary classifier, the DRN's ability to highlight and interpret the salient features contributing to the classification decision enhances the transparency and explainability of the results.

This synergistic approach not only maintains the accuracy of the VGG-16 model but also provides medical professionals with deeper insights into the decision-making process. The interpretation of the VGG-16 model's outcomes is facilitated by the DRN's analysis of the key features, enabling healthcare professionals to better understand the factors that influenced the classification of brain images. Ultimately, the integration of the VGG-16 and DRN models offers a balanced solution, leveraging the former's superior accuracy while benefiting from the latter's interpretability, promoting trust and enabling more informed decision-making in the context of brain disease diagnosis.

REFERENCES

1. Barmak, O., Krak, I., Manziuk, E.: Diversity as the basis for effective clustering- based classification. In: CEUR Workshop Proceedings. vol. 2711, pp. 53–67. CEUR (2020), http://ceur-ws.org/Vol-2711/paper5.pdf
2. Bedue´, P., Fritzsche, A.: Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. Journal of Enterprise Information Management 35(2), 530–549 (2022)
3. Von Eschenbach, W.J.: Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology 34(4), 1607–1622 (2021)
4. Dhar, T., Dey, N., Borra, S., Sherratt, R.S.: Challenges of deep learning in medical image analysis— –improving explainability and trust. IEEE Transactions on Tech-nology and Society 4(1), 68–75 (2023)