sciendo

JAISCR, 2024, Vol. 14, No. 2, pp. 121 – 137
10.2478/jaiscr-2024-0007

A NOVEL EXPLAINABLE AI MODEL FOR MEDICAL DATA ANALYSIS

Nataliya Shakhovska[1,2,*], Andrii Shebeko[1], Yarema Prykarpatskyy[2,3]

[1]*Department of Artificial Intelligence Systems, Lviv Polytechnic National University, 5 Kniazia Romana St., Lviv, Ukraine*

[2]*Department of Applied Mathematics, University of Agriculture in Krakow, 21 Mickiewicza al., 31-120 Krakow, Poland*

[3]*Institute of Mathematics of NAS of Ukraine, 3, Tereschenkivska st., 01024 Kiyv-4, Ukraine*

[*]*E-mail: nataliya.b.shakhovska@lpnu.ua*

Submitted: 5th November 2023; Accepted: 26th January 2024

Abstract

This research focuses on the development of an explainable artificial intelligence (Explainable AI or XAI) system aimed at the analysis of medical data. Medical imaging and related datasets present inherent complexities due to their high-dimensional nature and the intricate biological patterns they represent. These complexities necessitate sophisticated computational models to decode and interpret, often leading to the employment of deep neural networks. However, while these models have achieved remarkable accuracy, their ”black-box” nature raises legitimate concerns regarding their interpretability and reliability in the clinical context.

To address this challenge, we can consider the following approaches: traditional statistical methods, a singular complex neural network, or an ensemble of simpler neural networks. Traditional statistical methods, though transparent, often lack the nuanced sensitivity required for the intricate patterns within medical images. On the other hand, a singular complex neural network, while powerful, can sometimes be too generalized, making specific interpretations challenging. Hence, our proposed strategy employs a hybrid system, combining multiple neural networks with distinct architectures, each tailored to address specific facets of the medical data interpretation challenges.

The key components of this proposed technology include a module for anomaly detection within medical images, a module for categorizing detected anomalies into specific medical conditions and a module for generating user-friendly, clinically-relevant interpretations.

Keywords: Explainable AI, deep neural networks, medical data analysis, interpretability, clinical applications

1 Introduction

In today’s rapidly evolving medical landscape, data-driven decision-making stands at the forefront of enhancing patient care and health outcomes. A significant portion of this data emerges from complex medical images and records, the inter-

pretation of which necessitates advanced computational technologies. Historically, clinicians and healthcare professionals have relied on traditional diagnostic techniques and their expertise to decode these medical images. However, the sheer volume and complexity of current medical data demand automated, yet highly accurate, analysis methods. This rising need for swift and precise data extraction has paved the way for the integration of Artificial Intelligence (AI) in medical data analysis [1, 2].

The application of AI in the medical realm, particularly deep learning techniques, has revolutionized diagnostics, yielding accuracy rates that often surpass human capabilities. Indeed, these sophisticated models can efficiently process vast datasets, identifying patterns and anomalies that might be overlooked by the human eye. Yet, their intricate inner workings, often termed the "black-box" nature, render their decisions somewhat enigmatic to clinicians and patients alike. Thus, there exists a critical need to demystify this black box, making AI not just powerful, but also transparent and interpretable.

Explainable AI (XAI) emerges as a promising frontier in addressing this challenge. XAI aims to produce models that are both high-performing and interpretable, allowing for transparency in decision-making processes. Applications of this technology span various sectors, including predictive patient outcomes, identification of disease patterns from imaging data, and personalized treatment recommendations. Despite the vast potential of XAI in medical data analysis, numerous nuances await refinement. For instance, while XAI techniques might elucidate the decisions of an AI model, the challenge remains in conveying this information in a clinically relevant and understandable manner. Statistical research reveals that while clinicians may understand the broad strokes of AI decisions, they might struggle with the granular intricacies and the 'why' behind certain predictions [3].

Refining these nuances bears profound implications not just for individual patients, but also for the broader healthcare sector and policy-making bodies. Enhanced transparency could empower patients with informed decision-making, bolster trust in AI-driven medical processes, and streamline regulatory procedures concerning AI in healthcare.

The evolving healthcare ecosystem necessitates cutting-edge technological solutions to optimize patient care, reduce overhead costs, and bolster the overall efficacy of medical procedures. With the advent of portable diagnostic devices and the increasing reliance on telemedicine, there's a pressing need for real-time, accurate, and, importantly, explainable data interpretation methods. Automating and refining these processes through XAI could indeed be the linchpin in realizing the full potential of AI in medicine.

The paper aimed to develop new explainable model based on ensemble of artificial neural networks and logistic regression for MRI imaged analysis and anomaly highlighting.

The paper contribution is given as follows:

- The novel XAI approach for MRI backbone analysis id proposed. The pipeline combines separated convolution neural network for each three slices, logistic regression for results aggregation, SHapley Additive exPlanations (SHAP) and Class Activation Mapping (CAM) for anomaly detection and visualization. It allows to increase the model performance comparing to VGG-11, VGG-16, Alexnet, Resnet and Efficientnet for separated slices in MRI and to add interpretability to the results.

- To improve the efficiency of SHAP, tree-base ensemble together with sampling and aggregating of Shapley values is used. It allows the speed improvement of feature selection for cloud and local environment at least $1,3$ times.

- Application of SHAP and CAM made it possible to conduct a detailed assessment of both the model itself and individual variables that influence its decision. This not only increased the transparency of the model, but also improved its interpretability and understanding for medical professionals. In addition, an OpenCV-based program was developed for image visualization and analysis and CAM. The program allows interaction with data at the level of individual patients and planes, which greatly facilitates the analysis process and can contribute to further clinical application.

2 State of the art

In the wake of digitization, the healthcare sector has been inundated with vast amounts of data. From patient records to diagnostic imagery, this data presents an unprecedented opportunity for improved patient care. However, leveraging such colossal datasets for meaningful insights necessitates sophisticated analytical tools. Artificial Intelligence (AI), particularly its subset machine learning, has emerged as a promising candidate to address this challenge. Yet, as machine learning models delve deeper into complexity, a new challenge emerges: interpretability. The need for explainability in AI, especially when applied to sensitive areas like healthcare, becomes paramount. This article aims to delve into the nuances of Explainable AI (XAI) in the context of medical data analysis.

Medical decisions have profound implications on patient outcomes, and thus any technology assisting these decisions must be transparent in its functioning. Traditional machine learning models, though effective, often operate as black boxes, rendering their decision-making processes opaque. As healthcare professionals increasingly rely on AI for diagnosis, treatment planning, and patient monitoring, understanding the ”why” behind AI predictions is crucial. This is where Explainable AI, with its focus on making AI's workings understandable to humans.

While AI has advanced by leaps and bounds in recent years, much of its decision-making remains enigmatic. In medical data analysis, where decisions can have life-altering implications, there is no room for uncertainty. Therefore, explaining AI predictions becomes not just desirable, but imperative.

Local Interpretable Model-agnostic Explanations (LIME): LIME is a widely adopted tool that allows users to comprehend the inner workings of any machine learning model. It operates by generating perturbed versions of the input data and analyzing the model's predictions based on these perturbations. Through this, LIME discerns which features are most influential in driving a specific prediction [4]. In the context of medical imagery, for instance, it could highlight areas in an X-ray or MRI that are pivotal in diagnosing a condition.

SHapley Additive exPlanations (SHAP): its foundation comes from game theory, SHAP endeavors to provide each feature with an importance score that indicates its contribution to a particular prediction. The essence of SHAP is its calculation method: it determines the average contribution of a feature over all potential combinations of features. This ensures that the contribution of every feature is fairly distributed based on its actual impact on the outcome, offering a comprehensive view of feature significance [5]. For clinical data, this could elucidate the impact of different medical tests or patient attributes in a diagnosis.

Class Activation Mapping (CAM): Before delving into Grad-CAM, it's crucial to understand its predecessor, CAM. CAM techniques are designed for convolutional neural networks (CNNs) and they highlight the regions of an input image that play a significant role in model decisions. This method leverages the global average pooling layer's weights in CNN to produce a coarse localization map highlighting important regions [6]. Gradient-weighted Class Activation Mapping (Grad-CAM): Used primarily with deep learning models, Grad-CAM visualizes the regions of an input image that were important for neural network decisions. Building on the foundation laid by CAM, Grad-CAM extends its applicability to a broader range of model architectures [7], not just CNNs. Primarily utilized with deep learning models, Grad-CAM creates a visual heatmap over input images, indicating which regions were pivotal for the neural network's decision. This approach is particularly beneficial as it doesn't require any modification to the original model and offers insights into why certain regions of an image were influential in the prediction process. In medical applications, it can spotlight areas in an image that led to certain pathological predictions.

Despite the availability of these tools, challenges persist. Medical data, by its nature, is multifaceted and often interlinked. While a tool might identify a symptom as a predictor for a disease, the underlying causality can be convoluted. Moreover, different models might give varying explanations for the same prediction, leading to confusion. Lastly, while these tools bridge the gap between AI models and human understanding, they do require

a basic comprehension of data science, which can be a hurdle for some medical professionals.

A 2021 study by Jordan Fuhrman's team [8] highlighted the role of explainable AI in COVID-19 imaging. Amidst controversies surrounding AI-driven evaluations, the review focuses on building trust using explainable and interpretable AI. Importantly, the study emphasized Shapley values' ability to identify key pixels, especially lower lobe features, in COVID-19 diagnosis.

Falk Schwendicke's 2020 research, 'Artificial Intelligence in Dentistry', emphasizes the importance of trust and generalizability in dental AI. Highlighting potential transformations in care quality and efficiency, the paper also mentions challenges like data limitations and ethics. The team differentiates between rule-based software and evolving AI models, pointing to the potential of deep learning [12].

Thanh Vo's group in 2022 discussed concerns about AI's "black-box" nature in predicting drug-drug interactions [13]. The study champions Explainable AI (XAI) for its transparency and safety, particularly in critical tasks like drug interactions, while also outlining XAI's limitations and future prospects.

Kipp Johnson in paper [12] serves as a clinician's guide to AI in cardiology. It touches upon predictive modeling, algorithms, and the promise of deep and unsupervised learning in enhancing patient outcomes.

Jana Lipková's 2022 research [15] from Brigham and Women's Hospital centers on AI for multimodal data in oncology. The paper highlights the need for data fusion to improve diagnostic models and AI's potential to identify novel patterns. Challenges and solutions for clinical adoption are also discussed.

These real-world examples underline the transformative potential of Explainable AI in healthcare. By bridging the trust gap and offering actionable insights, it augments the capabilities of healthcare professionals. However, with great potential comes great responsibility. The integration of these technologies must be approached with caution, ensuring that they complement, not replace, the human touch in medical care.

We explored the concept of Explainable AI, its importance in medical data analysis, real-world applications, and anticipated future developments. As the field of AI in healthcare continues to grow, the demand for transparency and understandability will invariably rise, further underlining the significance of making AI models explainable. The journey ahead is promising, but it mandates a collaborative effort between technologists and healthcare professionals to realize the full potential of this union.

The burgeoning interest in Explainable AI suggests a promising trajectory in medical AI research. As the healthcare sector gradually transits towards personalized medicine, the requirement for interpretable and transparent AI models will surge [1]. With genomic data playing a pivotal role in treatment plans, Explainable AI can unravel complex genetic patterns, offering insights that could lead to groundbreaking therapeutic strategies.

Modern healthcare is unimaginable without EHR systems. The integration of Explainable AI with EHR can revolutionize diagnostic procedures and treatment strategies, offering personalized care pathways based on comprehensive data analysis [7]. However, this amalgamation poses challenges, particularly regarding data privacy and ensuring the integrity of personal health information.

Explainable AI holds the potential to reshape the landscape of medical data analysis, offering clarity and fostering trust. While the prospects are enticing, careful and ethical integration is the key. The combination of human expertise with AI's computational prowess can elevate healthcare, ensuring optimal patient outcomes. The challenge lies in striking the perfect balance, ensuring that technology augments human capabilities without overshadowing the essence of personal care.

In [10], Grad-CAM (Gradient-weighted Class Activation Mapping) is analysed. Grad-CAM is a technique that visualizes the regions of an image that were most relevant to a deep learning model's decision. It has been applied to medical images to provide explanations for model predictions.

In healthcare applications, an Explainable AI system promises significant advancements, including predictive diagnostics, patient care personalization, and accelerating medical research [10]. In pa-

per [12], the extensive comparison of existing models of XAI in medical images analysis is given.

In [18], VGG-16, VGG-11, AlexNet and RestNet neural networks are used for backbones segmentation. Unfortunately, explainability is missed in the results. However, we will use the results to compare with proposed approach. Heatmaps and VGG-16 are used in [19] for image analysis and interpretability.

To improve the accuracy of AI model, the ensembling technic is used. Particularly in [20] various classification methods combined with the rough set theory. It allows improving of model's performing. The same stacking method is used in [21] together with hyperparameters tuning. Papers [23, 24] demonstrate the combination of fuzzy-logic and AI to increase the explainability of the model. That is why the paper aims to improve the explainability and performance of AI model based on ensembling of CNN with local interpretable model.

3 Materials and Methods

3.1 General Pipeline

From the sources reviewed, we can discern key blocks essential for the implementation of our algorithm to interpret medical data through Explainable AI (XAI). To enhance flexibility and streamline the refinement process, each block is designed to be independent. Thus, if the need arises, individual blocks can be improved or replaced for a more efficient or novel approach for specific sub-tasks. For ease of use and adaptability, the algorithm is encapsulated within a server with multiple access points. This setup allows for potential deployment on a standalone workstation, sending medical data for processing, retrieving processed results either immediately or upon request, and storing all data for future use and analysis. Consequently, we arrive at the following architecture:

1. **Model Training and Classification**

   1.1. Training plane-specific Deep CNNs. Three individual deep CNN models are trained, each tailored to a specific plane of medical imagery: axial, sagittal, and coronal. The nuances and distinct features of each plane are captured by their respective CNN models, ensuring high accuracy in the initial interpretation of the data.

   1.2. Consolidated Classification via Logistic Regression. Once each of the three CNN models processes the image slices and outputs a probability score (axial, sagittal, and coronal), these scores are integrated and fed into a Logistic Regression model. This classifier, specifically tuned with the 'lbfgs' solver, computes the final probability of a specific medical condition, in this instance, an ACL tear.

2. **Feature Visualization through SHAPs and CAMs**

   2.1. Generating CAMs, which visually highlight the areas in the medical image that were instrumental for the CNN models to arrive at their conclusions. This offers a heat-map styled overlay on the original image, indicating regions of importance.

   2.2. Generating SHAPs, which values provide a more granular understanding of feature importance. For each pixel or region in the medical image, SHAP values decipher the contribution of that specific feature to the model's final prediction. Like CAMs, SHAP values can be visualized as heat maps, spotlighting the key determinants in the diagnosis.

3. **Integrated Visualization Framework**. The visualization tool allows users to seamlessly switch between the original medical image, its CAM overlay, and its SHAP overlay. This layered approach ensures users can compare and contrast the regions highlighted by both CAMs and SHAP values against the pristine image, providing a comprehensive understanding of the model's decision-making process.

In summary, this three-tiered approach, from in-depth model training to sophisticated visualization techniques, sets the foundation for a highly interpretable AI system in medical data analysis. Such a system not only diagnoses but also clearly conveys the 'why' behind its conclusions, bridging the trust gap between AI and medical professionals.

In addition to ensembling, the improvement of Shapley value calculation is proposed. Shapley values provide a fair way to distribute the model's

prediction among its features and used in SHAP model. The Shapley value of a feature is its average marginal contribution across all possible combinations of features [22]. The exact calculation of the Shapley is computationally expensive for large feature sets, especially for medical data. That is why tree-based models is used. It means that Shapley values can be approximated more efficiently using the model's structure. Moreover, decision tree is good in parallel realization, that is why the time complexity can be reduced too. The TreeExplainer is a technique within the SHAP library specifically designed for explaining individual predictions of tree-based models. It uses a fast algorithm to approximate Shapley values for tree ensembles. In addition to that sampling and aggregating Shapley values is used too for improved speed.

3.2 Dataset model and classifier

The MRNet dataset consists of 1,370 knee MRI exams performed at Stanford University Medical Center [16]. The dataset contains 1,104 (80.6%) abnormal exams, with 319 (23.3%) ACL tears and 508 (37.1%) meniscal tears; labels were obtained through manual extraction from clinical reports.

Each record in the dataset has an ID, and the paths to the MRI scans are constructed using this ID. For labeling purposes, one-hot encoding is used to represent two classes, making it suitable for binary classification. If no transformations are specified for the dataset, the MRI scan, which is initially in grayscale, is stacked to mimic a three-channel image (similar to an RGB image). This ensures compatibility with models like AlexNet that expect three-channel inputs. Additionally, the dataset accommodates weighting, which can be beneficial for handling class imbalances.

The primary model used here is referred to as the MRNet, which is a convolutional neural network (CNN). This model leverages a pretrained version of the AlexNet architecture. The AlexNet has historically been a popular choice for image recognition tasks and has been pretrained on the ImageNet dataset. In the MRNet model, after passing through the pretrained layers, the extracted features undergo adaptive average pooling to reduce the spatial dimensions. This is followed by a classifier, which is a linear layer designed to produce two outputs. These outputs can be seen as a binary classification (like diagnosing a medical condition as positive or negative).

The MRNet is designed to handle MRI images that may come in batches or individually. Additionally, the model is flexible enough to process images with varying numbers of slices, accommodating the unique needs of medical imaging.

After extracting features from the MRI images using the MRNet model, these features are then utilized to make a diagnosis using a classifier. The primary classifier in this framework is a Logistic Regression model, which is a statistical method used for modeling binary dependent variables. In this context, it's used to diagnose the medical condition based on the features extracted from different planes of the MRI scans. The logistic regression model is trained using features extracted from the three planes: axial, coronal, and sagittal. Once trained, this classifier can then predict the probability of a positive diagnosis, which is evaluated using the AUC-ROC score—a common metric for binary classification tasks. This score provides insight into the classifier's ability to distinguish between the positive and negative classes.

In conclusion, the entire system is a blend of deep learning and classical machine learning. MRI scans are passed through a pretrained CNN to extract meaningful features, which are then used by a logistic regression model to make the final diagnosis. The integration of these components is designed to leverage the strengths of both deep learning (feature extraction) and classical machine learning (interpretable decision-making).

AlexNet is a deep convolutional neural network (CNN) that significantly impacted the field of machine learning when it was introduced. Developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, it won the 2012 ImageNet Large Scale Visual Recognition Challenge, a pivotal competition in computer vision [17].

The architecture comprises five convolutional layers followed by three fully connected layers and can differentiate between 1,000 distinct object categories. The innovative aspects of AlexNet include the use of ReLU (Rectified Linear Unit) activation functions which increased training speed, and dropout layers that helped in reducing overfitting. Additionally, it was one of the first models to use

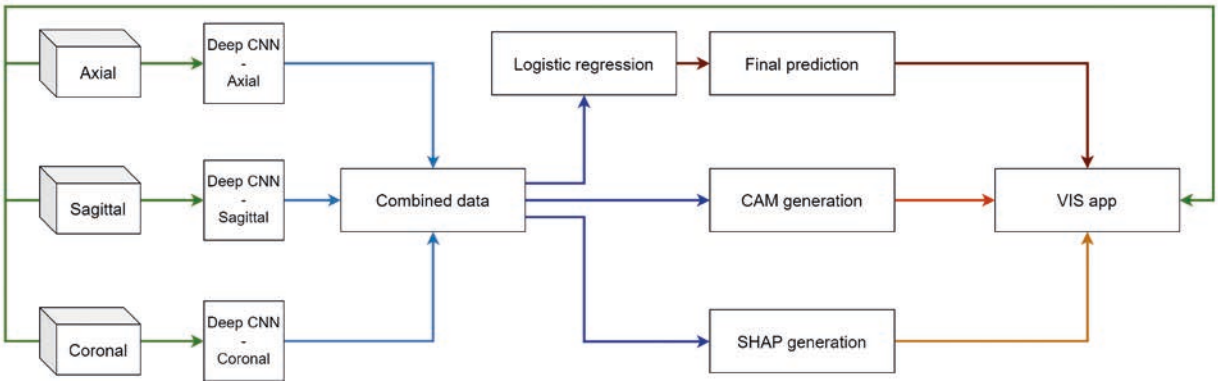

Figure 1. General pipeline

GPU training to speed up the computation.

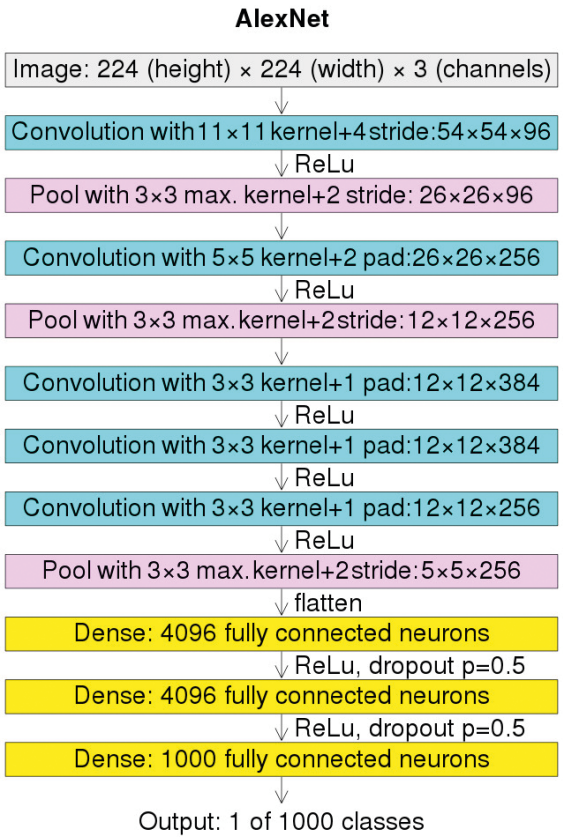

Figure 2. General pipeline

One of the primary reasons to use a pre-trained model like AlexNet in MRNet (Magnetic Resonance Neural network) or other projects is the advantage of transfer learning. AlexNet, trained on a vast dataset like ImageNet, has already learned essential features from millions of images. By leveraging these learned features, MRNet can bootstrap its training process and potentially reach a higher accuracy faster than training from scratch. Training a deep neural network from scratch requires significant computational resources and time. Utilizing a pre-trained version of AlexNet allows researchers and developers to bypass the resource-intensive initial training phase.

3.3 Core concepts

There are several ways of elucidating the internal mechanisms of complex machine learning models to make them interpretable for users. Primarily, these explanations can be divided into two main categories: Local Explanations and Global Explanations.

Local Explanations focus on explaining a subset of inputs [10]. A popular example of a local explanation is a per-decision or single-decision explanation, which elucidates the reasoning behind the algorithm's output or decision for a specific input instance. LIME is a frequently employed local explanation algorithm. LIME aims to understand a decision by investigating points in close proximity to the decision. It then constructs an interpretable model representing this local decision and leverages that model to offer explanations for each feature. In the context of images, LIME dissects each image into superpixels and then, through a random search mechanism, determines the impact of specific superpixels by replacing them with a uniform color, typically black. Another notable local explanation algorithm is SHAP (SHapley Additive exPlanations). SHAP offers insights into the importance of each feature for a given input in regression problems.

Global Explanations provide overarching, post-hoc insights about the entire algorithm [10]. These

explanations frequently involve creating a holistic model that can capture the broader behaviors of an algorithm or system. Partial Dependence Plots (PDPs) serve as an example of a global explanation method. These plots highlight the marginal change in the predicted response with variations in a particular feature. PDPs are instrumental in determining the nature of the relationship between a feature and the output, be it linear or more intricate. For deep neural networks, Testing with Concept Activation Vectors (TCAV) is an established global explanation approach. TCAV aims to make neural networks more interpretable by presenting their state as a linear combination of human-understandable concepts termed as Concept Activation Vectors (CAVs). This method has been utilized to decipher image classification algorithms, revealing insights into how certain concepts, like colors, influence the decision-making processes of these classifiers.

In our endeavor to make the intricate operations of our model comprehensible and transparent, we resorted to local explanation methods, specifically SHAP and CAM. We utilized SHAP because of its robustness and versatility. SHAP doesn't just explain the output of the model; it also provides a deep insight into how each feature influences the output. SHAP values, grounded in cooperative game theory, offer consistent and fairly distributed explanations. On the other hand, CAM (Class Activation Mapping) offers a visual perspective, making it particularly fitting for image data. CAM visually highlights the regions in the image that were pivotal for the model's decision. By using CAM, we were able to visually validate and ensure that our model was focusing on the correct regions in the image and not being misled by irrelevant or noisy parts.

3.4 SHAP

In the realm of machine learning, understanding model predictions is essential, especially for deep learning models, which are often perceived as "black boxes" [5]. One effective approach to unravel this complexity is by employing SHAP (SHapley Additive exPlanations), which offers a mechanism to interpret the output of machine learning models.

For our medical imaging analysis using MRNet, there are 3 possible SHAP approaches:

- **Explainer**: This fundamental class in SHAP caters to various data types and offers basic interpretations suitable for simpler models.
- **GradientExplainer**: Marrying the principles of SHAP, SmoothGrad, and Integrated Gradients, this method is tailored for intricate neural networks, such as those utilized in medical image diagnostics.
- **DeepExplainer**: Fusing SHAP with a variant of DeepLIFT, this method assists in comprehending deep learning models, illuminating how each imaging feature impacts the model's diagnostic predictions.

We particularly employed the GradientExplainer with the for several reasons. First of all MRNet is a deep learning model, the GradientExplainer is apt as it combines multiple gradient-based methods, providing detailed insights into how the model processes MRI images. The 'background tensor' serves as a reference or baseline for the model. By comparing an MRI image against this baseline, the explainer can determine how much each feature (or pixel) in the image contributes to the final prediction. In the context of MRNet, this could be distinguishing pathological regions from healthy tissue. Using GradientExplainer provides physicians and radiologists with an intuitive understanding of the model's diagnostic rationale. By visualizing the SHAP values, clinicians can gauge which regions in the MRI influenced the model's predictions the most, ensuring a higher level of trust and transparency.

3.5 CAM

In the world of medical imaging, the visualization of specific regions in MRI scans that influence the prediction of a machine learning model is crucial. A way to achieve this is by using Class Activation Mapping (CAM), which visually highlights those important areas in the image [6, 7].

For our MRI analysis, we've created a function which generates these visual maps for given patient cases and imaging planes. This function first sets up the necessary folders for storing the MRI slices and their corresponding CAM visualizations. It then retrieves the last but one layer's weights of the MRNet

model. These weights are essential for generating the CAM visualizations.

The MRI images are processed through the MRNet model, producing feature maps which are stored and applied to image as heatmap. For each MRI slice, we visualize the activated regions by combining the original MRI slice and a heatmap produced. The heatmap provides a color-coded representation of the importance of different regions in the MRI slice for the model's prediction. The resultant images are saved in the target folder, providing radiologists with a visual guide to understand which parts of the MRI were pivotal for the model's decision.

CAM extraction function takes in the feature maps, weights, and the target class index to generate the CAM visualizations. It computes the CAM for each MRI slice by combining the weight and feature map. This CAM is then resized and adjusted to produce a heatmap. The series of produced heatmaps are returned and used in the main function.

Lastly, we have a hook registered on a specific layer of the MRNet's pretrained model. This hook is designed to capture the feature maps every time data passes through that layer. By leveraging CAM, our approach provides a transparent way for doctors and radiologists to trust the model's predictions, ensuring that AI-driven decisions in medical imaging are both interpretable and actionable.

4 Results

4.1 Experiment organization and result

Overall, the combination of MRNet, SHAP and CAM has created a strong toolkit for the analysis and interpretation of medical images, providing a high level of accuracy and reliability, as well as opening avenues for further research and improvement Table 4.1.

The model and dataset for this work were implemented in the Python programming language using the PyTorch framework. Specifically, an MRNet class was created that inherits from the base nn.Module class in PyTorch. This model uses a pretrained AlexNet architecture as its basis for identifying salient features from medical images. One of the key features of the model is the use of adaptive averaging pooling, which unifies the dimensions of the extracted features before their classification.

For the dataset, the MRDataset class was created, which also inherits from the data.Dataset class in PyTorch. This class is used for loading and preprocessing medical images. It also maintains weights for each class, which is important for neural networks in tasks where the classes are unbalanced. Arguments for model training were specified via Python's built-in argparse module. The user can select the task (for example, ACL diagnosis), the anatomical plane of the image (sagittal, coronal, axial), as well as several other parameters for training the model.

The results of training the MRNet model for the classification of medical images are presented in three tables: Table 4.1 for the sagittal plane, Table 3 for the coronal plane and Table 4 for the axial plane. Each table presents model quality indicators before and after optimization: number of training epochs, AUC (Area Under the Curve) on training and validation data.

Table 1. Training parameters

Parameter	Value
Group size	1
Number of epochs	60 + (Early stopping)
Optimizer	Adam
Gamma	0.5
Initial training speed	1e-5
Dataset size	1 * n * 256 * 256
Augmentation	True

According to the results of training the model on the sagittal slice (Table 4.1), it can be seen that the final model showed an improvement in the AUC indicator on the validation data set from 0.85 to 0.88. Training AUC also increased from 0.80 to 0.83.

Table 2. Sagittal slice training and testing analysis

Training process	Initial value	Final value
Number of epochs	5	9
Val AUC	0.85	0.88
Train AUC	0.80	0.83
Train Loss		
Val Loss		

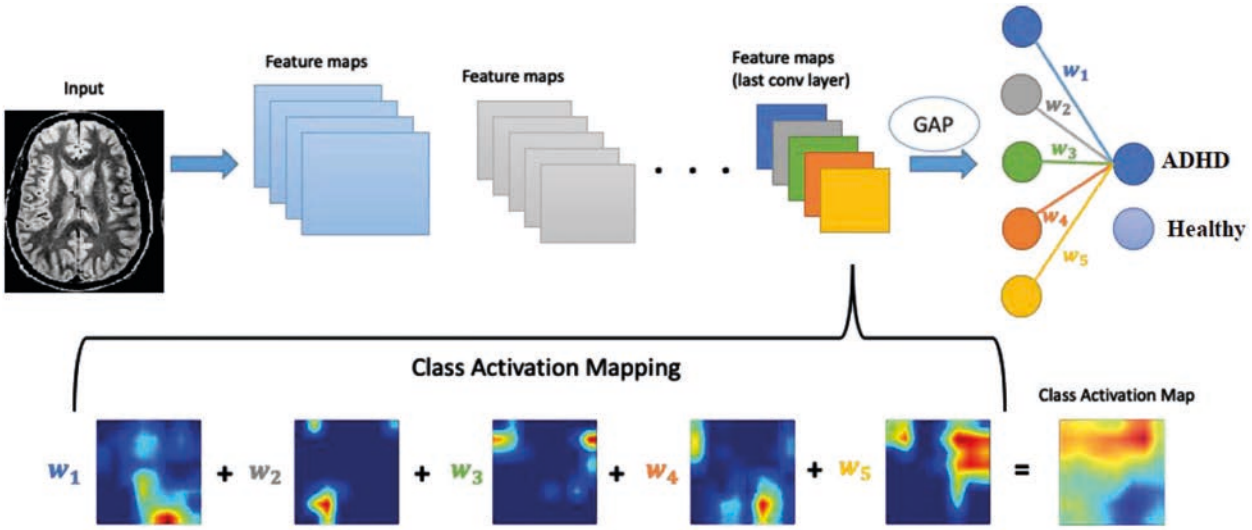

Figure 3. CAM extracting example

Relative to the coronal slice (Table 3), the final model showed a significant improvement in AUC on the validation data set – from 0.79 to 0.89. Training AUC increased from 0.81 to 0.87.

Table 3. Coronal slice training and testing analysis

Training process	Initial value	Final value
Number of epochs	17	10
Val AUC	0.79	0.89
Train AUC	0.81	0.87
Train Loss		
Val Loss		

For the axial slice (Table 4), the model performed impressively, with AUC on the validation dataset increasing from 0.82 to 0.93. Training AUC increased from 0.83 to 0.95.

Table 4. Axial slice training and testing analysis

Training process	Initial value	Final value
Number of epochs	11	22
Val AUC	0.82	0.93
Train AUC	0.83	0.95
Train Loss		
Val Loss		

It should also be noted that the training was performed on a dataset that has 1130 training examples with a total size of 6.51 GB, and on a test set consisting of 120 examples and 707 MB.

The overall improvement of the MRNet model can be assessed as significant, especially considering the AUC indicators on the validation data (Figure 4-Figure 5). During the initial phase of training on the Google Colab platform using the T4 GPU, good results were already achieved. However, after switching to local training using the more powerful 2070s GPU, the performance of the model improved significantly.

Figures 8-9 show a summary of pre-training and final training for visual evaluation. The tensorboard library was used to view the results.

These results highlight the effectiveness and reliability of the presented MRNet model for the medical image classification task. Particularly encouraging is the high level of AUC on the validation data, which is an indicator of the excellent overall performance of the model. The comparison with existing approaches for several diseases are given in Table 5 (best ROC-AUC is highlighted in bold).

Based on the table above, the proposed approach is not dominated by other pretrained networks.

4.2 Visualization

In the realm of machine learning, gaining insights into the intricacies of how models make decisions is pivotal. As these models become more sophisticated, the ability to interpret and understand them is crucial for a myriad of reasons including trust, model debugging, and regulatory com-

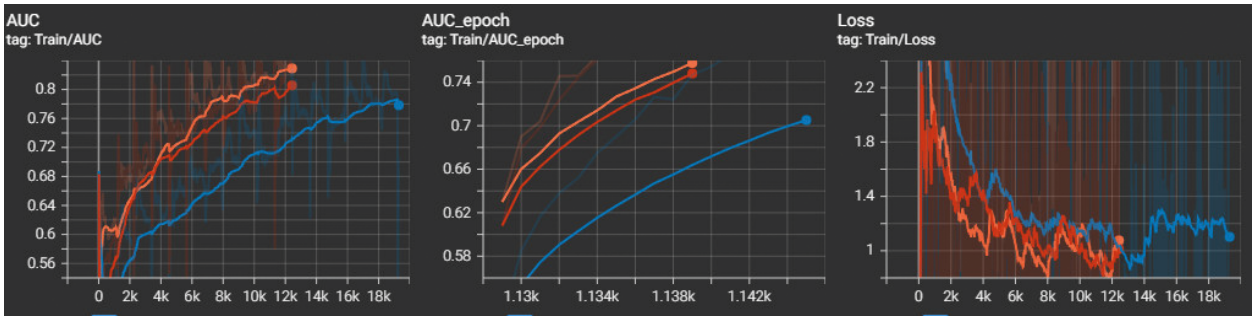

Figure 4. Results of pre-training (training set).

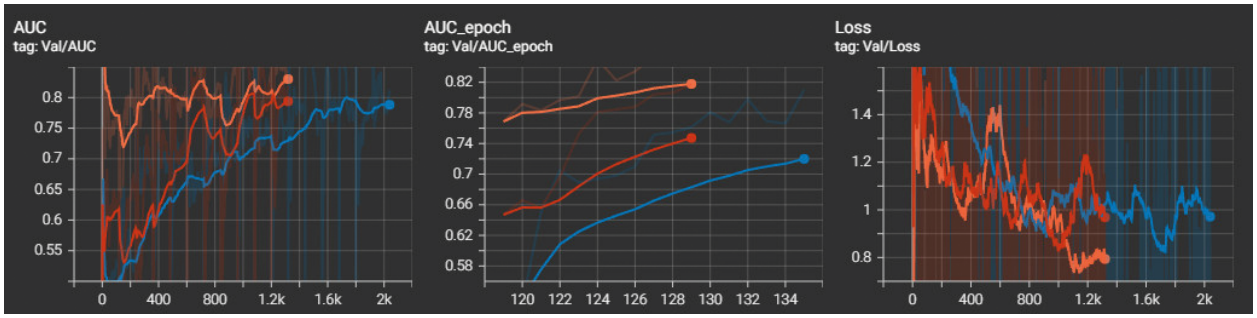

Figure 5. Results of pre-training (validation set).

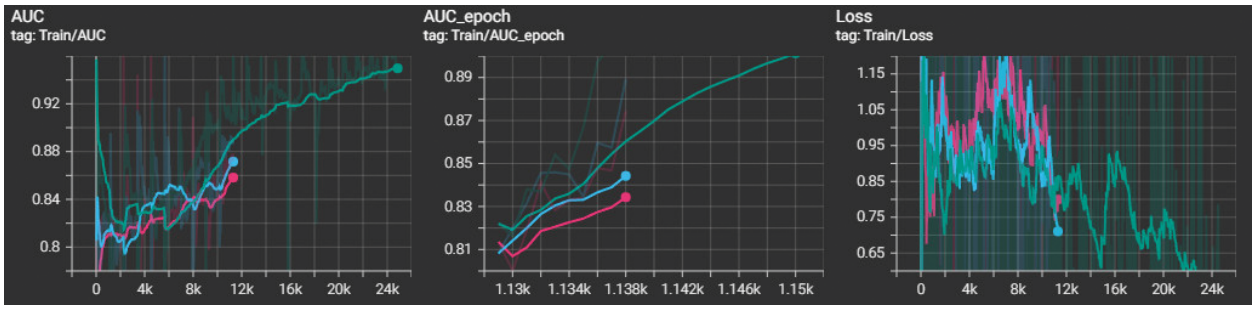

Figure 6. Results of final training (training set).

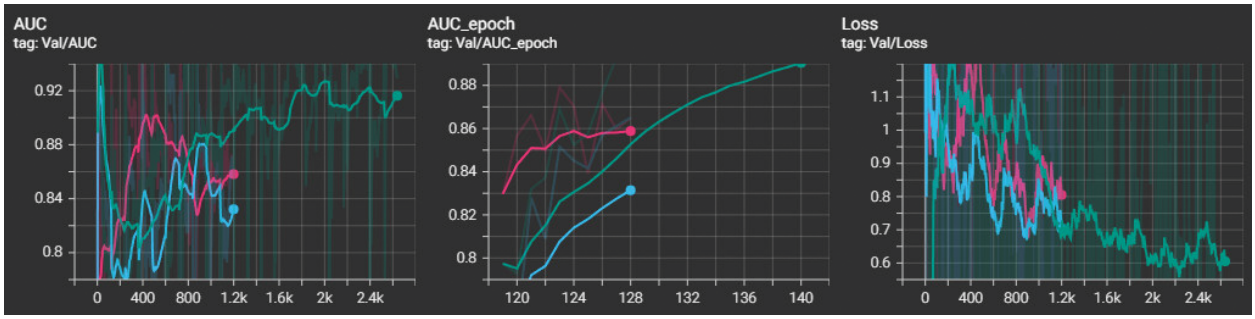

Figure 7. Results of final training (validation set).

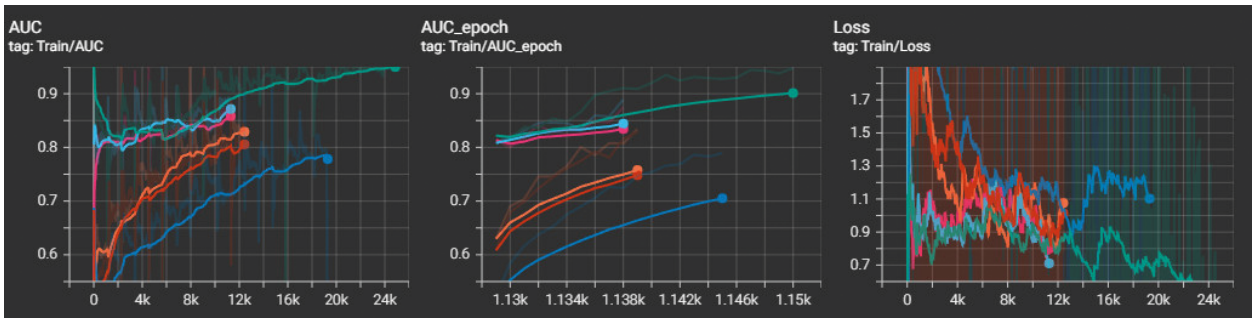

Figure 8. The training charts (training set).

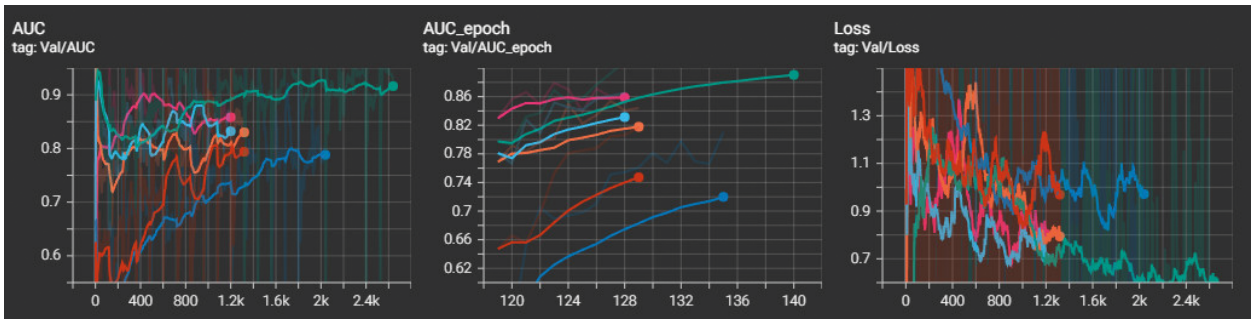

Figure 9. The training charts (validation set).

Table 5. The comparison with existing approaches

Backbone	**Slice**	General abnormality	ACL	Meniscus tear
Alexnet	axial	0.845	0.738	**0.813**
	coronal	0.788	0.616	0.633
	sagittal	0.935	0.834	0.707
VGG11	axial	0.892	0.758	0.811
	coronal	0.732	0.926	0.762
	sagittal	0.917	**0.903**	0.753
VGG16	axial	**0.919**	**0.829**	0.759
	coronal	**0.840**	**0.950**	**0.808**
	sagittal	0.909	0.883	**0.801**
Resnet	axial	0.848	0.763	0.618
	coronal	0.477	0.462	0.659
	sagittal	0.912	0.700	0.632
Efficientnet	axial	0.845	0.586	0.642
	coronal	0.667	0.593	0.535
	sagittal	0.723	0.631	0.591
Proposed approach	axial	0.852	**0.886**	**0.819**
	coronal	0.798	**0.956**	0.721
	sagittal	**0.936**	**0.901**	**0.801**

pliance. This is where the practice of visualizing machine learning model interpretations comes into play. Specifically, using three prominent sets: CAM (Class Activation Mapping), Original, and SHAP (SHapley Additive exPlanations) offers a comprehensive view into the model's decision-making process. Visualization makes complex data more tangible. When we can visually see how a model is thinking, we can better understand its behavior. This can lead to the identification of areas of improvement, model refinement, and even discovery of novel insights.

We decided to show 3 sets at once: CAM, SHAP outputs and original data offers a baseline, ensuring that any interpretations or visualizations are grounded in the true data. It's the reference point against which other visualizations can be compared.

Leveraging these three sets isn't just about depth of insight, but also ease of navigation and optimal performance. Having three distinct visualizations provides a structured approach to model interpretation. Users can start with the high-level view from CAM, ground themselves with the Original set, and then delve deep with SHAP. This structured approach avoids overwhelming the user while ensuring that all levels of detail are accessible.

Additionally, the combination of these tools has been optimized for performance. By avoiding redundant computations and using efficient algorithms, the visualization process remains swift, ensuring that users can get insights in real-time.

In addition to visualization, time complexity is analyzed too. Neural network training was performed both in the Google Colab cloud environment using the Nvidia Tesla T4 graphics accelerator and on a local machine with the Nvidia GeForce 2070s accelerator. The Shapley values without tree optimization generation on local machine was taking around $8-10$ mins and $6-8$ mins in Google Colab, which has now dropped to around 7 mins and 5 mins respectively using Tree optimizer and sampling.

Discussions

The obtained results are closely correlated with the set of tasks of increasing the accuracy and interpretability of AI models. The use of SHAP and CAM to visually interpret the MRNet model not only turned out to be expected, but also led to significant improvements in understanding the internal dynamics of the model. This confirms the effectiveness of the implementation of SHAP and CAM in medical applications of neural networks, in particular, in diagnostic tasks using MRNet. Thanks to this, medical professionals can more accurately and confidently use the models for diagnostic purposes.

Application of methods of Variable Importance Analysis (SHAP) and Class Activation Visualization (CAM) made it possible to conduct a detailed assessment of both the model itself and individual variables that influence its decision. This not only increased the transparency of the model, but also improved its interpretability and understanding for medical professionals.

In addition, an OpenCV-based program was developed for image visualization and analysis and CAM. The program allows interaction with data at the level of individual patients and planes, which greatly facilitates the analysis process and can contribute to further clinical application.

We also note the positive dynamics of model quality indicators. For example, the AUC value on the validation data improved for all three planes, which indicates the high ability of the model to correctly classify medical images.

In summary, SHAP and CAM contribute to the interpretability and understanding of AI models, fostering transparency, trust, and collaboration. These qualities are essential for improving the interoperability of AI systems across different domains and facilitating their integration into diverse applications. However, some improvements of SHAP and CAM together with ensemble based on CNN can be realized:

1. Computing efficiency: Calculating SHAP values can be computationally expensive, especially for complex models or large datasets. The improvement is realized in developing more efficient algorithms or approximations for calculating SHAP values.

2. Automated feature engineering: SHAP requires the definition of a background dataset for calculating Shapley values. That is why research on

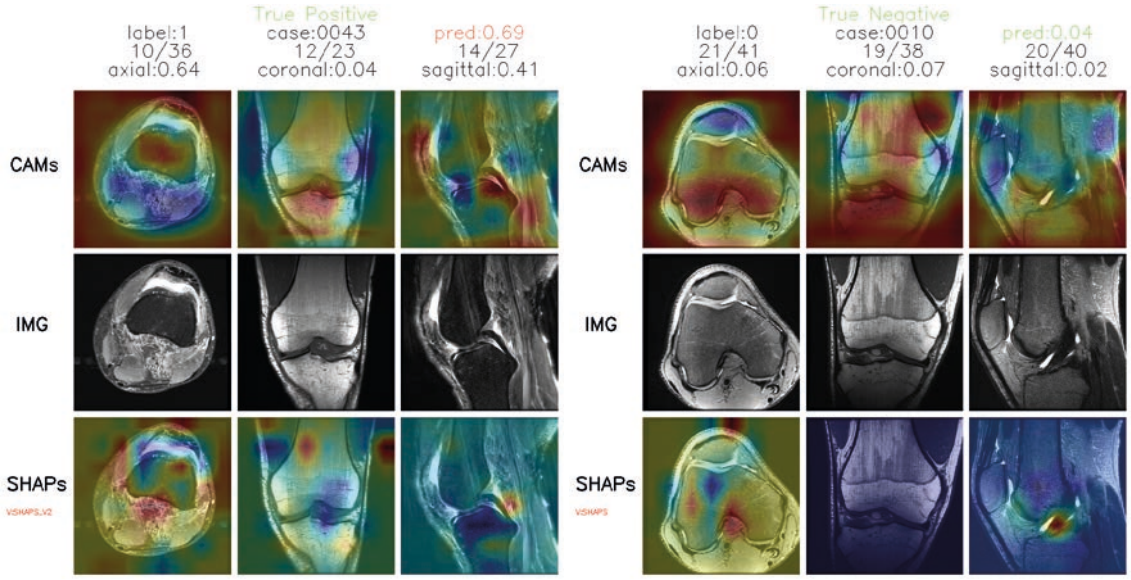

Figure 10. TP and TN predictions with SHAP and CAM visualizations

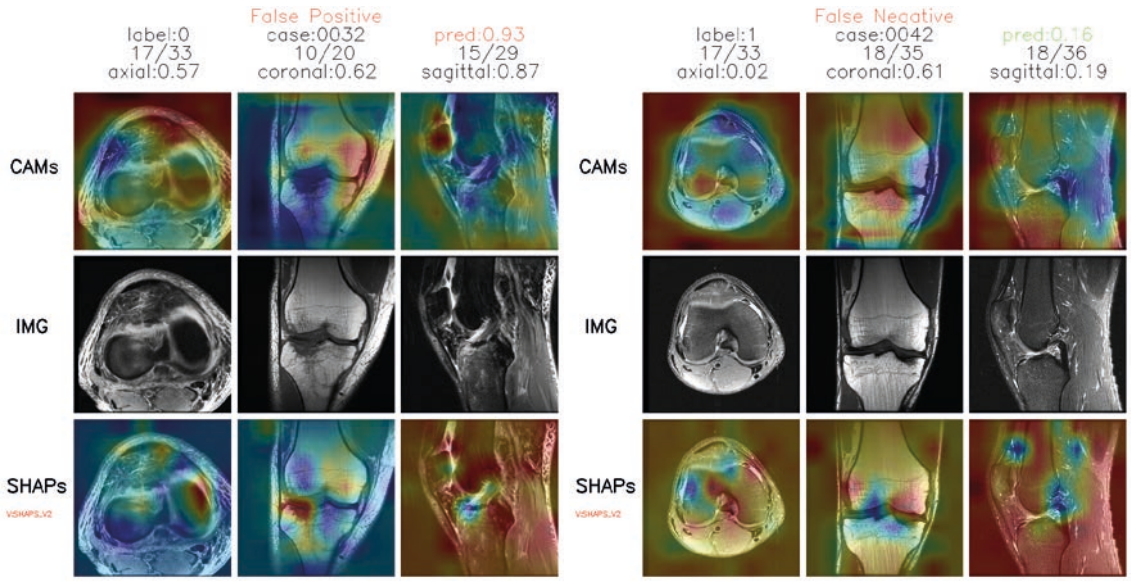

Figure 11. FP and FN predictions with SHAP and CAM visualizations

methods for automated or adaptive background dataset selection could simplify the usage of SHAP.

3. Integration with more architectures: CAM is commonly used with CNNs, that is extending it to other architectures, particularly to ensembling as in our case require additional adaptation.

4. Low performance improvement of stacking: Integrating diverse model architectures, including combinations of deep learning, tree-based models, and linear models, can enhance the ensemble's ability to capture a broader range of features and patterns.

Conclusion

The paper proposed the structured framework for Explainable AI, specifically tailored for medical data analysis. A pivotal component of our approach involves utilizing SHAP (SHapley Additive exPlanations) and CAM (Class Activation Mapping) methodologies, renowned for their efficacy in providing interpretability and visualization of neural network decisions. The modular nature of the algorithm is designed to accommodate future advancements and integrations.

One of the most significant attributes of this framework is its adaptability. Its architecture is inherently modular, enabling swift enhancements and allowing the system to incorporate various medical data types. The introduction of new data types is streamlined, thanks to this modular foundation. Our choice of using local explanation methods like SHAP and CAM was driven by the need for detailed, instance-specific insights. While SHAP provided a granular understanding of feature importance, CAM offered visual validation, ensuring our model's decisions were both transparent and interpretable.

At its core, our solution is designed for local applications. However, its structure envisages a potential split into a client-server model in future iterations. This separation would entail the visualization components residing on the client-side, while the neural network predictions, SHAP, and CAM computations would be processed server-side. Such a bifurcation would optimize computational loads and provide a scalable model for extensive datasets.

In healthcare applications, this Explainable AI system promises significant advancements, including predictive diagnostics, patient care personalization, and accelerating medical research [10]. The transparent nature of SHAP and CAM methodologies further reinforces trust and understanding in AI-driven decisions among medical professionals. The potential evolution into a client-server model heralds a new era of efficiency, expanding the horizons for democratized, interpretable AI in healthcare.

References

[1] Lane T. (2018). A short history of robotic surgery. Annals of the Royal College of Surgeons of England, 100(6_sup), 5–7. https://doi.org/10.1308/rcsann.supp1.5

[2] Liu P.-R., Lu L., Zhang J.-Y., Huo T.-T., Liu S.-X., & Ye Z.-W. (2021). Application of Artificial Intelligence in Medicine: An Overview. Current Medical Science, 41(6), 1105–1115. https://doi.org/10.1007/s11596-021-2474-3

[3] Zhang Y., Weng Y., & Lund J. (2022). Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. Diagnostics (Basel, Switzerland), 12(2), 237. https://doi.org/10.3390/diagnostics12020237

[4] Ribeiro M. T., Singh S., & Guestrin C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier (arXiv:1602.04938). arXiv. http://arxiv.org/abs/1602.04938

[5] Lundberg S. M., & Lee S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c4-3dfd28b67767-Abstract.html

[6] Camalan S., Mahmood H., Binol H., Araújo A. L. D. Santos-Silva, A. R. Vargas, P. A. Lopes, M. A. Khurram, S. A. & Gurcan, M. N. (2021). Convolutional Neural Network-Based Clinical Predictors of Oral Dysplasia: Class Activation Map Analysis of Deep Learning Results. Cancers, 13(6), 1291. https://doi.org/10.3390/cancers13061291

[7] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., & Batra D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision, 128(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[8] Fuhrman J. D., Gorre N., Hu Q., Li H., El Naqa I., & Giger, M. L. (2022). A review of explainable and interpretable AI with applications in COVID-19 imaging. Medical Physics, 49(1), 1–14. https://doi.org/10.1002/mp.15359

[9] Vinogradova K., Dibrov A., & Myers G. (2020, April). Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 10, pp. 13943-13944)

[10] Phillips P. J., Hahn C. A., Fontana P. C., Yates A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). Four principles of explainable artificial intelligence (NIST IR 8312; c. NIST IR 8312). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.IR.8312

[11] Shakhovska N., & Pukach P. (2022). Comparative Analysis of Backbone Networks for Deep Knee MRI Classification Models. Big Data and Cognitive Computing, 6(3), 69. https://doi.org/10.3390/bdcc6030069

[12] Johnson K. W., Torres Soto J., Glicksberg B. S., Shameer K., Miotto, R., Ali M., Ashley E., & Dudley J. T. (2018). Artificial Intelligence in Cardiology. Journal of the American College of Cardiology, 71(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

[13] Lipkova J., Chen R. J., Chen B., Lu M. Y., Barbieri M., Shao, D., Vaidya A. J., Chen C., Zhuang, L., Williamson D. F. K., Shaban M., Chen, T. Y., & Mahmood F. (2022). Artificial intelligence for multimodal data integration in oncology. Cancer Cell, 40(10), 1095–1110. https://doi.org/10.1016/j.ccell.2022.09.012

[14] Schwendicke F., Samek W., & Krois J. (2020). Artificial Intelligence in Dentistry: Chances and Challenges. Journal of Dental Research, 99(7), 769–774. https://doi.org/10.1177/0022034520915714

[15] Vo T. H., Nguyen N. T. K., Kha Q. H., & Le N. Q. K. (2022). On the road to explainable AI in drug-drug interactions prediction: A systematic review. Computational and Structural Biotechnology Journal, 20, 2112–2123. https://doi.org/10.1016/j.csbj.2022.04.021

[16] Štajduhar I., Mamula M., Miletić D., & Ünal G. (2017). Semi-automated detection of anterior cruciate ligament injury from MRI. Computer Methods and Programs in Biomedicine, 140, 151–164. https://doi.org/10.1016/j.cmpb.2016.12.006

[17] Krizhevsky A., Sutskever I., & Hinton G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c-8436e924a68c45b-Abstract.html

[18] Zhang R., Du L., Xiao, Q., & Liu J. (2020, May). Comparison of backbones for semantic segmentation network. In Journal of Physics: Conference Series (Vol. 1544, No. 1, p. 012196). IOP Publishing.

[19] Woldan P., Duda P., Cader A., & Laktionov I. (2023). A new approach to image-based recommender systems with the application of heatmaps maps. Journal of Artificial Intelligence and Soft Computing Research, 13(2), 63-72.

[20] Nowicki R. K., Seliga R., Żelasko D., & Hayashi Y. (2021). Performance analysis of rough set–based hybrid classification systems in the case of missing values. Journal of Artificial Intelligence and Soft Computing Research, 11(4), 307-318.

[21] Baradaran Rezaei, H., Amjadian, A., Sebt, M. V., Askari, R., & Gharaei, A. (2023). An ensemble method of the machine learning to prognosticate the gastric cancer. Annals of Operations Research, 328(1), 151-192.

[22] Dong H., Sun J., & Sun X. (2021). A multi-objective multi-label feature selection algorithm based on shapley value. Entropy, 23(8), 1094.

[23] Starczewski Janusz T., Przybyszewski Krzysztof, Byrski Aleksander, Szmidt Eulalia & Napoli Christian. (2022). A Novel Approach to Type-Reduction and Design of Interval Type-2 Fuzzy Logic Systems" Journal of Artificial Intelligence and Soft Computing Research, 12(3), 197-206.

[24] Laktionov I., Diachenko G., Rutkowska D. & Kisiel-Dorohinicki,M.(2023).An Explainable AI Approach to Agrotechnical Monitoring and Crop Diseases Prediction in Dnipro Region of Ukraine. Journal of Artificial Intelligence and Soft Computing Research,13(4) 247-272.

Nataliya Shakhovska, Doctor of science, Lviv Polytechnic National University, Ukraine, the head of artificial intelligence department, Lviv Polytechnic National University, Ukraine.
https://orcid.org/0000-0002-6875-8534

Yarema Prykarpatskyy, doctor habilitowany, professor at Uniwersytet Rolniczy im. Hugona Kołłątaja w Krakowie, Krakow, Poland
https://orcid.org/0000-0002-3033-7419

Andrii Shebeko, Master student at artificial intelligence department, Lviv Polytechnic National University, Ukraine.
https://orcid.org/0000-0002-0212-8855