Computer Science • 26(3) 2025 https://doi.org/10.7494/csci.2025.26.3.6563

Reza Hassanpour
Kasim Oztoprak
Niels Netten
Tony Busker
Mortaza S. Bargh
Sunil Choenni
Beyza Kizildag
Leyla Sena Kilinc

DEVELOPING EXPLAINABLE MACHINE-LEARNING MODEL USING AUGMENTED CONCEPT ACTIVATION VECTOR

Abstract *Machine-learning models use high-dimensional feature spaces to map their inputs to the corresponding class labels; however, these features often do not have a one-to-one correspondence with the physical concepts that are understandable by humans. This hinders the ability to provide meaningful explanations for the decisions that are made by these models. We propose a method for measuring the correlation between high-level concepts and the decisions that are made by machine-learning models. Our method can isolate the impact of a given high-level concept and accurately measure it quantitatively. Additionally, this study aims to determine the prevalence of frequent patterns in machine-learning models that often occur in imbalanced data sets. We successfully applied the proposed method to fundus images and managed to quantitatively measure the impacts of the radiomic patterns on the model's decisions.*

Keywords explainable AI, machine learning, radiomics

Citation Computer Science 26(3) 2025: 1–17

1. Introduction

Machine-learning models have achieved remarkable accuracy in decision-making; however, their complexity makes it difficult for humans to comprehend their inference processes. In most cases, machine-learning models are evaluated based on the correlations between their inputs and outputs. While these black-box models may seem sufficiently acceptable, there are scenarios where transparency in the decision-making process is essential for certifying the model's validity and clarifying the potential risks that are associated with the suggested decisions.

An important application area where such transparency plays a crucial role is medicine [4]. In this field, both practitioners and patients are keen to consider the safety of any decisions that are made by machine-learning models. Consequently, many researchers consider transparency to be an important factor in making the models trustworthy. The added transparency helps explain the factors that influence the selections of outputs and its relationships with the contextual parameters.

Therefore, explainability should be addressed in the context of the model's application. Subsequently, the transparent model can be improved and verified; in some cases, it can reveal hidden characteristics of the data. An important challenge in developing explainable models is the lack of explicit correlations between those low-level features that are extracted by deep-learning models and human-understandable concepts (such as the radiomics that are used in medical applications). This issue implies that even explaining the inference process of a model may not be easily comprehensible to humans.

The second challenge is the importance that is attached to specific features by domain experts – even if these features may be less frequent or relatively rare. In this article, we address both of these problems. We propose a novel method for establishing a correlation between human-understandable concepts and low-level model parameters. Additionally, we consider the impact of imbalanced data on machine-learning models. We used fundus images of patients with retinopathy complications to validate our proposed model.

This article is structured as follows. In Section 2, we review the related literature, while Section 3 details the proposed model. In Section 4, we present the experimental results of our research, and in Section 5, we draw our conclusions and discuss our results.

2. Related works

Deep-learning techniques have demonstrated remarkable efficacy across a range of medical diagnostic tasks – even surpassing human experts in some instances [22]. Nonetheless, the opacity that is inherent to these algorithms has limited their practical clinical adoption [8]. Recent investigations into explainability strive to reveal the primary drivers behind a model's decisions. The impetus behind these inquiries

has arisen from the realization that elucidation plays a pivotal role in critical medical treatments like cancer care. In such applications, the importance of predictions extends beyond their precision; their comprehensibility also holds significant weight. Yet, concurrently achieving both attributes poses a formidable challenge, as there exists a delicate balance between interpretability and accuracy [15].

Explainable AI (XAI) in medical-image processing focuses on the development of techniques that enhance the interpretability and transparency of the artificial-intelligence models that are used in medical imaging [1]. This is crucial for gaining trust from medical practitioners, ensuring patient safety, and providing insights into the decision-making processes of AI systems.

In the realm of the machine-learning literature, explainability techniques are classified into two main categories: model-based explanations [20], [16], and post-hoc [19], [21] explanations [9]. When delving into a trained model in order to gain a deeper understanding of the acquired correlations, this approach is termed a post-hoc explanation. A crucial differentiation between post-hoc and model-based explanations lies in their methodologies; the former involves training a neural network and then seeking to explicate the operations of the resulting black-box network, while the latter mandates that the model itself is inherently interpretable.

Researchers have considered explainability from different perspectives. Zhou et al. proposed an architecture to provide more-interpretable results. They claimed that their model was based on networks with attention mechanisms or gradient-based visualization methods that could highlight regions of interest in medical images, thus making any predictions more understandable [24]. In a similar approach, Simonyan et al. proposed a heatmap or saliency map model. Their model used techniques that highlighted the regions of input images that were most influential in making predictions, thus aiding in understanding why an AI model arrived at a particular decision [17]. Techniques like LIME (local interpretable model-agnostic explanations) and SHAP (SHapley Additive exPlanations) provided post-hoc explanations for predictions from any machine-learning model, thus enhancing interpretability in the medical-imaging domain [14].

LIME [18] and SHAP [11] are two popular techniques for explaining the predictions of machine-learning models. Both techniques are model-agnostic, meaning that they can explain any type of model and locally interpret individual predictions. LIME approximates a model locally around a specific prediction by training a simple interpretable model. This approximation is based on perturbing input data and observing any changes in the predictions to create an interpretable linear model around a specific instance. As a result, LIME provides weights of features in a local linear model, making it possible to interpret the impact of each feature on any decision of the model. In addition, SHAP outputs Shapley values, which represent each feature's contribution to a prediction in a fair way.

On the other hand, some researchers have focused on making the model consider specific features. Attention mechanisms in deep-learning models enable the model

to focus on particular parts of an image, thus contributing to a more transparent decision-making process. These mechanisms have been used to improve performance and interpretability in medical-image analysis [12].

Radiomics involves extracting a large number of quantitative features from medical images. Explainable-modeling techniques that are applied to radiomics help identify which features are most influential in making particular diagnoses or predictions [6]. These techniques are also used to reduce the dimensionality of the feature space, thus emphasizing the most salient features [3]. Explainable radiomics involves the application of interpretable techniques to the process of extracting quantitative features from medical images such as CT scans, MRI images, and X-rays inorder to gain insights into the factors that contribute to diagnoses or predictions.

In their foundational paper, iKumar et al. introduced the concept of radiomics and discussed the process of extracting large numbers of quantitative features from medical images. They highlighted the challenges in radiomics, including feature selection and validation [6]. Lambin et al. discussed the potential of radiomics in extracting more information from medical images using advanced feature-analysis techniques and emphasized the need for robust and reproducible radiomic features [7].

Aerts et al. demonstrated how radiomics could be used to decode tumor phenotypes by analyzing noninvasive medical images, thus showcasing the potential of radiomics for providing insights into tumor characteristics [2]. In their research, Parmar et al. applied machine learning to radiomic features that were extracted from head and neck cancer images to develop prognostic biomarkers, thus demonstrating the potential of radiomics in predictive modeling [13]. Nie et al. focused on breast MRIs and demonstrated how quantitative analyses of lesion morphologies and texture features could aid in diagnostic predictions, thus highlighting the potential of radiomics in characterizing breast lesions [10]. Zhang et al. introduced the IBEX software platform, which was designed to facilitate collaborative work in radiomics; this emphasized the importance of standardized tools for radiomic feature extraction and analysis [23].

The testing concept activation vector (TCAV) method that was proposed by Kim et al. was an interpretability technique that was designed to assess the influence of high-level concepts on the predictions that are made by machine-learning models [5]. TCAV works by comparing the model's responses when the concept of interest is present versus when it is absent. By manipulating the concept while keeping the other factors constant, TCAV calculates a concept's impact on the model's predictions by using statistical tests. If the model's predictions are found to be sensitive to the concept, this suggests that the model has learned to associate the concept with a prediction, thereby providing insights into how the model uses the concept to make decisions. In essence, TCAV provides a way for understanding and measuring the extent to which specific high-level concepts are used by a machine-learning model to arrive at its predictions. This method has applications in various domains, including

medical-image analysis (where it can help ensure that models are making decisions based on clinically relevant factors).

Our proposed method also uses high-level concepts to determine to what extent any decisions that are made by the model are affected by those high-level patterns. However, our method distinguishes itself from the TCAV method by augmenting input data with concept patterns and restricting the impacts of external factors. Additionally, our method uses a single network for both training-classification and explainability verification. In contrast to both of the SHAP and LIME techniques, the input to the our proposed model consists of raw data rather than extracted features; this is processed by a deep neural network. The proposed model therefore distinguishes itself by establishing a correspondence between the model's internal features and the high-level radiomic patterns that are used by practitioners.

3. Augmented Concept Activation Vector

We propose a new method for evaluating the impact of human-understandable visual patterns on the classification decisions of machine-learning models. Our proposed method follows the same idea as TCAV in measuring the contribution of a concept (high-level visual pattern) to the decisions of machine-learning models. However, it distinguishes itself from TCAV in a few ways. First, our proposed method (called the augmented concept activation vector [ACAV] method) utilizes the context of its input data to minimize the effects of external factors.

While TCAV considers two sets of data points (P_C and N) to represent the positive and negative samples of a given concept, ACAV defines data point sets T_P and T_N, where $d^n \subset R^n$ is the input data point set, and $T_P \subset d^n$ and $T_N \subset d^n$ are subsets of the original data set. T_P includes the data points with the given concept pattern and are classified accordingly, while data points in T_N are samples without having a concept pattern that are classified correctly.

In addition, $A_C : \{a_C | a_C = a \bigoplus C, a \in T_N, Cisaconceptpattern\}$ (a subset of T_N data points) is defined to represent the data points with an augmented concept. The main advantage of augmenting the concept pattern to a data point is preserving the context of a sample and avoiding external factors. ACAV therefore targets application areas where the input data points share a strong context (such as medical images).

Second, ACAV can handle imbalanced data sets where certain concept patterns occur less frequently but are highly valued by domain experts for classification purposes (e.g., rare symptoms that are used for more-reliable diagnoses in medical applications). Finally, ACAV employs a single machine-learning model, thus providing greater flexibility in designing and implementing the method (as shown in Figure 1).

We propose training a machine-learning model to classify data points as belonging to one the possible classes. Our assumption is that the classification process considers the existence and prevalence of specific concept patterns. The activation vector of layer l maps an input vector of dimension n to a vector of dimension $m \ll n$ as f_l :

$R^n \rightarrow R^m$. The direction of the activation vector alters with the value of the input data; however, the differences are limited when the input points belong to the same class (and are classified correctly). Equation (1) defines the activation of neuron i of layer l:

$$a_i^{(l)} = f\left(\sum_{j=1}^{n^{(l-1)}} w_{ij}^{(l)} a_j^{(l-1)} + b_i^{(l)}\right). \tag{1}$$

The activation vector of layer l is defined as is shown in Equation (2):

$$\mathbf{a}^{(l)} = f\left(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}\right). \tag{2}$$

First, we train a model to classify the labeled input data. Subsequently, we consider a sample data point d_i that lacks Concept Pattern C, which is classified as S_i by the model. The activation vector at layer l when this data point is fed into the model is given as $V_l = f(d_i)$.
Next, we augment the data point by including the Concept C pattern as $d_{ia} = d_i \bigoplus C$. We give the augmented data point to the model and find the activation vector at layer l as $V_{la} = f(d_i)$. The experiment is repeated with n data samples; the average rate of the deviation of the activation vector is found by using Equation (3):

$$\Delta V = \frac{1}{n}\sum_i |V_{la} - V_l|. \tag{3}$$

To evaluate the impact of a given concept on the decision-making process of the model, we define two metrics. The first metric measures the relative deviation of the activation vectors when the inputs are augmented with a concept pattern. The activation vectors of the fully connected feedforward networks serve as the foundations for the classification decisions that are made in the output layer. These vectors represent the combined processing of those features that were extracted in the earlier layers of the neural network (such as the convolutional layers). Consequently, any changes in the input data are reflected in the activation vectors. Our goal is to measure any deviations in the activation vectors when the changes in the input data were exclusively due to the presence of high-level concepts. This metric is the ratio of the number of data points are assigned to a different class after being augmented to the number of data points that preserve their initial class labels.[CHECK THE PREVIOUS SENTENCE] This metric is an indication of the contribution of the visual pattern to the decision that is made by the model. The second metric takes into account the abundance of the visual concepts when more than one concept is present in the data points. If a concept appears more frequently in the data set, it becomes the dominant factor in determining the class attributes. Therefore, imbalance data can deteriorate the accuracy of the model. We measure the data imbalance using an entropy-based metric (as shown in Equation 4):

$$H = -\sum_{i=1}^{C} p_i \log(p_i), \tag{4}$$

where H is the entropy of the high-level pattern distribution, p_i is the proportion of the high-level patterns in class i, and C is the total number of high-level patterns. Our second metric aims to measure how effective the distribution of the different visual concepts is on the decisions of the model. Figure 1 depicts a schematic diagram of our proposed method.

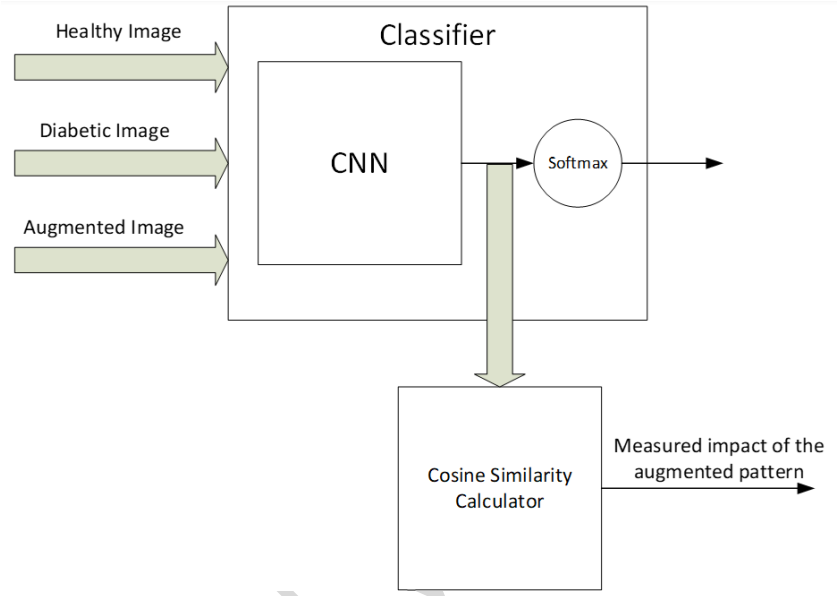

Figure 1. Schematic diagram of proposed method

4. Experimental results

To experimentally validate the proposed method, we conducted three sets of experiments. The data set that was used in our experiments was the publicly available RFMiD data set [1] data set. This data set included 3200 fundus images that were labeled with one of the five diabetic classes, where Class 0 included healthy cases, and Classes 1 through 4 represented diabetic classes with different severities (with 4 representing the most acute cases). The images that corresponded to the less-severe cases of diabetics incorporated fewer occurrences of radiomic patterns and was limited to only some of these patterns; therefore, we considered only two classes of healthy and diabetic cases in our experiments. We considered the images of the healthy cases to be the training samples of the first class and the fourth and fifth categories as the samples that belonged to the second class.

[1] RFMiD data set (Retinal Fundus Multi-Disease Image Data Set). Center of Excellence in Signal and Image Processing, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India

A subset of 50 healthy fundus images were augmented by adding a single cotton-wool pattern at locations that were close to the blood vessels. A sample image of this subset is illustrated in Figure 2, where the original and augmented images are shown.

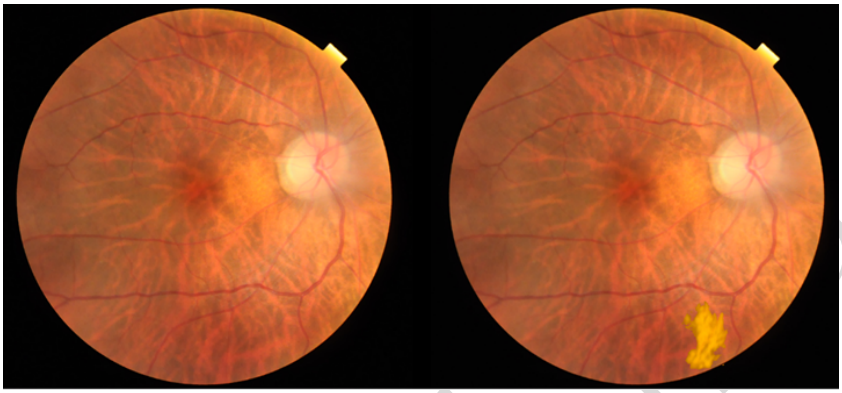

Figure 2. Healthy images augmented with cotton-wool radiomic patterns

The second experiment includes more than one augmented pattern per image in a subset of 50 healthy images. Figure 3 illustrates a sample from this subset.

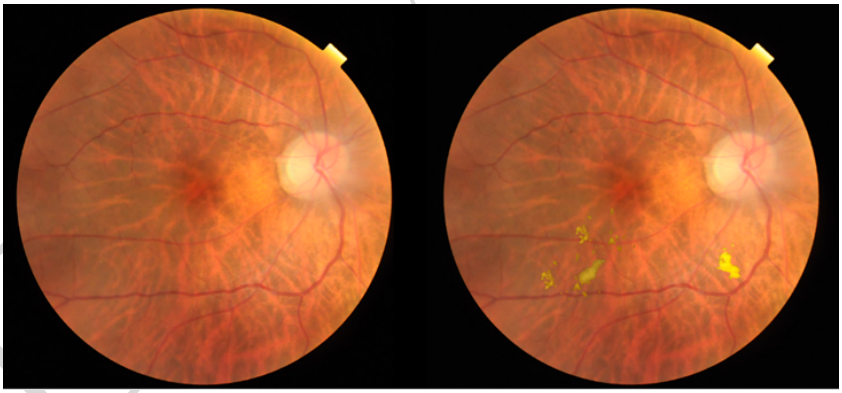

Figure 3. Healthy images augmented with cotton-wool and fatty dot radiomic patterns

The third experiment aimed to measure the impact of a less-frequent symptom such as cotton-wool patterns when the image included multiple high-level features. This experiment was to determine the behavior of the model in the presence of imbalanced data. For this experiment, we augmented those images had bleeding symptoms with a fatty dot pattern. As was mentioned, Categories 2 and 3 were not considered as parts of the experiments; however, some of these images were used to augment them with less-frequent radiomic patterns, as they generally included only more-frequent

bleeding patterns. A third subset with 50 images was used for the third experiment; Figure 4 illustrates a sample augmented image that was used for the third experiment.

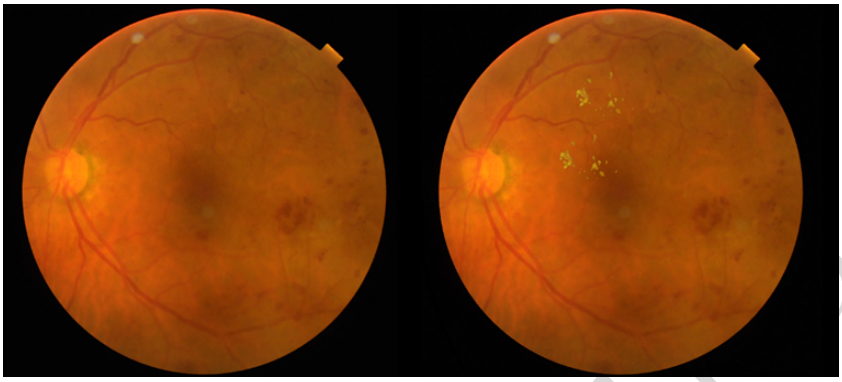

Figure 4. Images with bleeding pattern augmented with fatty dot radiomic pattern

We used a CNN with three convolution layers and three layers at the fully connected MLP, where the last layer included five neurons that featured the soft-max activation function. Figure 5 shows a schematic structure of the model with the parameters that were used.

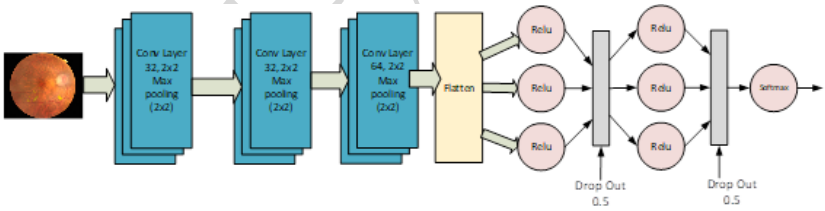

Figure 5. Schematic diagram of classifying model

The model was trained using two classes of training data; namely, healthy patients, and diabetics. Using the test data, we next evaluated the performance of the network. While classifying the healthy images, we stored the values of the layer before the last one as a vector of a 64×1 dimension if an image was correctly classified. At the end, we found the average of the stored vectors. In a similar way, we calculated the output of the layer before the last one for the diabetic images that were classified correctly. These two vectors indicated the direction of the activation vector for the two different classes. As is shown in the diagram of Figure 5, the last layer included a single neuron with the softmax activation function. To avoid those cases of which the model was not very confident, we labeled the softmax neuron output values that were greater than 0.6 as "healthy" and those values that were less than 0.4 as "unhealthy." Hence, an error margin of 0.2 was considered.

Subsequently, we provided the augmented images as the input and measured the activation vector value of the layer before the last one. This vector was compared with the average activation vectors of the healthy and diabetic inputs by using cosine similarity. As the original images before augmenting the radiomic patterns were available, we could find the deviation of the activation vector from its original direction. Table 1 summarizes the results as explained.

Table 1
Changes in directions of activation vector after augmenting healthy inputs

Augmented Pattern Type	Avg. Norm Vector Original Image	Avg. Norm Vector Augmented Image	Avg. Absolute Deviation
Fatty dots (single pattern)	0.864	0.835	0.03
Fatty dots (multi patterns)	0.864	0.813	0.06
Cotton wool (single pattern)	0.864	0.831	0.03
Cotton wool (multi patterns)	0.864	0.797	0.07
Bleeding (single pattern)	0.864	0.808	0.06
Bleeding (multi patterns)	0.864	0.759	0.10

Table 2
Changes in directions of activation vector after augmenting images with bleeding patterns

Augmented Pattern Type	Avg. Norm Vector Original Image	Avg. Norm Vector Augmented Image	Avg. Absolute Deviation
Fatty dots (single pattern)	0.843	0.792	0.05
Fatty dots (multi patterns)	0.843	0.755	0.09
Cotton wool (single pattern)	0.843	0.803	0.04
Cotton wool (multi patterns)	0.843	0.746	0.10

In addition, a similar procedure was repeated using images with bleeding radiomic patterns (generally from Category 2, which were excluded from our experiments) where other less-frequent radiomic patterns were augmented (Figure 4). The results of this experiment are summarized in Table 2. The experimental results revealed that increasing the radiomic patterns caused the activation vector to deviate in the direction of the diabetes class. In our experiments, the average deviation of the activation vector was small due to the fact that we augmented the input images with a small number of radiomic patterns; in the real images of the diabetic cases, these patterns occurred more frequently.

To further verify the effectiveness of the proposed method in determining the contribution of each radiomic pattern to the classifier's decision, we repeated the experiments with additional radiomic patterns that were augmented to the healthy images. We conducted experiments for three different cases where the augmented patterns were fatty dots and cotton-wool spots, bleedings, and a combination of both. Figure 6 depicts samples from these three cases.

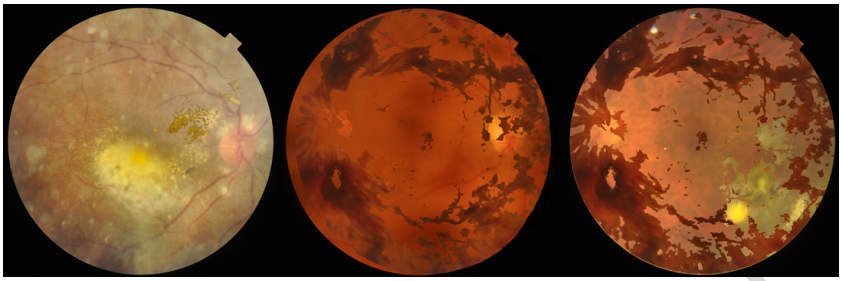

Figure 6. Sample images that were used in second experiment

In addition to measuring the deviation of the concept-activation vector in Layer n-1, we decided to include the deviations in Layer n-2 in this experiment to provide further insights into the behavior of the model in response to the presence of radiomic patterns in the input images. The results of this experiment are summarized in Table 3.

Table 3
Measured angles of concept activation vectors

Concept	Reference Vector	Angle (degrees)
Fatty dots/cotton (Layer n-1)	Healthy	41
Fatty dots/cotton (Layer n-2)	Diabetes	26
Fatty dots/cotton (Layer n-1)	Healthy	37
Fatty dots/cotton (Layer n-2)	Diabetes	11
Bleeding (Layer n-1)	Healthy	42
Bleeding (Layer n-2)	Diabetes	6
Bleeding (Layer n-1)	Healthy	44
Bleeding (Layer n-2)	Diabetes	6
Combined pattern (Layer n-1)	Healthy	60
Combined pattern (Layer n-2)	Diabetes	5
Combined pattern (Layer n-1)	Healthy	56
Combined pattern (Layer n-2)	Diabetes	4

Although we observed deviations in the concept-activation vector in both Layers n-1 and n-2 for Case 1 of the experiment, the average angle with diabetes cases was relatively high. This result was aligned with the rate of the changes in classification results (which was 80%). The deviation in the concept-activation vector in Case 2 was more consistent, as the angle of the diabetes cases became smaller. These results aligned with the fact that bleeding was generally the first symptom, and fatty dots or cotton-wool spots rarely occurred without the bleeding pattern in an image. In our experiments, all of the images that were augmented with both fatty dots and cotton-wool and bleeding patterns were classified as diabetes cases.

In our second set of experiments, we used a brain-tumor data set from Kaggle[2]. The data set consisted of 253 brain MRI images that were categorized into two groups: those with brain tumors, and those without. Each image was provided in the JPEG format at varying resolutions. Figure 7 presents a selection of the sample images from the data set.

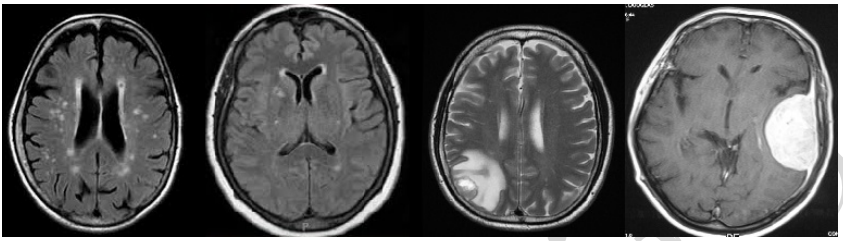

Figure 7. Sample images from brain MRI data set

The experiment was conducted by selecting an appropriate high-level concept. In this experiment, we considered the sizes of the brain tumors as a high-level concept. We augmented a subset of healthy images with brain tumors that were segmented from the other images. The segmented brain tumors were scaled into three different categories (large, average, and small). The purpose of the experiment was to determine the impact of the tumor size on the classification decision of the model. Figure 8 presents a healthy image after scaling that was augmented with a segmented tumor; the three images were samples of the test cases.

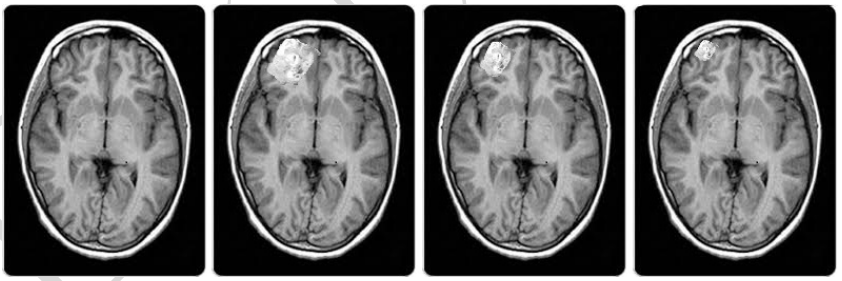

Figure 8. Sample augmented images with different pattern scales from brain MRI data set

A CNN model with the same structure as in the first experiment was used for the second experiment. The resolutions of the input images were adjusted, as these mostly had much lower resolutions than the fundus images. The amounts of the deviations at the activation vectors were measured; these are presented in Table 4.

[2]https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection

Table 4
Measured deviations of concept-activation vectors

Concept	Reference Vector	Angle (degrees)
Small Brain Tumor	Healthy	14
Small Brain Tumor	Diagnosed	31
Medium Brain Tumor	Healthy	29
Medium Brain Tumor	Diagnosed	9
Large Brain Tumor	Healthy	47
Large Brain Tumor	Diagnosed	3

The experimental results indicated that the trained model was sensitive to the sizes of the tumors. On average, the deviation from the diagnosis with a tumor-activation vector when a healthy image was augmented with a large scale tumor was insignificant, and all of the test images were classified as being diagnosed with tumors. On the other hand, only a small subset of images that were augmented with small-scale tumors were classified as being diagnosed with tumors. In this group, the average deviation of the activation vector from the healthy class was 14 degrees. The insignificant deviation that was observed from the healthy activation vector for the small tumors in our experiments can be attributed to the training data set, which predominantly contained images with large tumors. Consequently, the model learned to treat large tumor sizes as a consistent feature. The results of the second experiment demonstrated the applicability of the proposed method across different data sets; however, it is crucial to ensure the presence of the selected high-level concepts during the model's training phase.

Furthermore, our experiments showed that the proposed method had the potential of evaluating the significance of each pattern in isolation. In many medical examinations, the number of determining symptoms is large; it is of great significance to be able to determine the level of importance that the model considers for each of them. In addition, the proposed model facilitated specific considerations about high-level patterns; for instance, the impact of the brightness or darkness of radiomic patterns as well as their textures, sizes, and locations could be effectively verified using the proposed model.

In our experiments, we considered using convolutional neural networks; however, the proposed method can be used with any neural network (including recurrent networks). From a computational perspective, the proposed method has the advantage of training and using a single model. Despite the fact that both the ACAV and TCAV solutions are model-agnostic, however, TCAV requires the training of a second model; this requires not only extra processing but also a separate data set of a high-level concept.

Our experimental results were validated by a domain expert, who confirmed that the findings of our proposed method aligned with the expectations and observations from practitioners in hospitals.[6]

5. Conclusion

In this study, we extended our previous work on measuring the impacts of high-level patterns on the decisions of ML models by isolating these patterns and eliminating the influences of external factors during experiments. We proposed a method that can incorporate differences in the shapes, colors, intensities, locations, etc. of high-level patterns and effectively verify their impacts on the classification decisions that are made by the model. The proposed method not only provides insights into the decision-making process of ML models but also opens the way to incorporating domain-expert knowledge into classifiers for more-effective and -accurate results. Although the conducted experiments were limited to augmenting a small number of high-level radiomic patterns, we demonstrated that the proposed model could successfully isolate the impacts of these patterns independent of the surrounding data. This work can be extended by verifying the impacts of selecting different locations for augmenting radiomics. Additionally, the sizes and intensities/contrasts of radiomic patterns on the model's decisions can be investigated further.

6. Acknowledgment

We are sincerely grateful for the invaluable contributions of Dr. Dilay Ozek, Ophthalmologist at Ankara City Hospital, throughout the course of our research. Dr. Ozek's expertise and insights have been instrumental in shaping the direction and depth of our study. Their generous sharing of information, clinical experience, and thoughtful guidance have significantly improved the quality and relevance of our work.

References

[1] Abbasi-Asl R., Yu B.: Structural compression of convolutional neural networks, *arXiv preprint arXiv:170507356*, 2017.

[2] Aerts H.J., Velazquez E.R., Leijenaar R.T., Parmar C., Grossmann P., Carvalho S., Bussink J., Monshouwer R., Haibe-Kains B., Rietveld D., *et al.*: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, *Nature communications*, vol. 5(1), p. 4006, 2014. doi: 10.1038/ncomms5006.

[3] Hassanpour R., Netten N., Busker T., Shoae Bargh M., Choenni S.: Adaptive Feature Selection Using an Autoencoder and Classifier: Applied to a Radiomics Case. In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pp. 1256–1259, 2023. doi: 10.1145/3555776.3577861.

[4] Karaca E.E., Işık F.D., Hassanpour R., Oztoprak K., Evren Kemer Ö.: Machine learning based endothelial cell image analysis of patients undergoing descemet membrane endothelial keratoplasty surgery, *Biomedical Engineering/Biomedizinische Technik*, (0), 2024. doi: 10.1515/bmt-2023-0126.

[5] Kim B., Wattenberg M., Gilmer J., Cai C., Wexler J., Viegas F., *et al.*: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: *International conference on machine learning*, pp. 2668–2677, PMLR, 2018.

[6] Kumar V., Gu Y., Basu S., Berglund A., Eschrich S.A., Schabath M.B., Forster K., Aerts H.J., Dekker A., Fenstermacher D., *et al.*: Radiomics: the process and the challenges, *Magnetic resonance imaging*, vol. 30(9), pp. 1234–1248, 2012. doi: 10.1016/j.mri.2012.06.010.

[7] Lambin P., Rios-Velazquez E., Leijenaar R., Carvalho S., Van Stiphout R.G., Granton P., Zegers C.M., Gillies R., Boellard R., Dekker A., *et al.*: Radiomics: extracting more information from medical images using advanced feature analysis, *European journal of cancer*, vol. 48(4), pp. 441–446, 2012. doi: 10.1016/j.ejca.2011.11.036.

[8] Marcinkevičs R., Vogt J.E.: Interpretable and explainable machine learning: a methods-centric overview with concrete examples, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13(3), p. e1493, 2023. doi: 10.1002/widm.1493.

[9] Mitu M., Hasan S.M., Efat A.H., Taraque M.F., Jannat N., Oishe M.: An explainable machine learning framework for multiple medical datasets classification. In: *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, pp. 1–6, IEEE, 2023. doi: 10.1109/ncim59001.2023.10212821.

[10] Nie K., Chen J.H., Hon J.Y., Chu Y., Nalcioglu O., Su M.Y.: Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI, *Academic radiology*, vol. 15(12), pp. 1513–1525, 2008. doi: 10.1016/j.acra.2008.06.005.

[11] Nohara Y., Matsumoto K., Soejima H., Nakashima N.: Explanation of machine learning models using shapley additive explanation and application for real data in hospital, *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106584, 2022. doi: 10.1016/j.cmpb.2021.106584.

[12] Oktay O., Schlemper J., Folgoc L.L., Lee M., Heinrich M., Misawa K., Mori K., McDonagh S., Hammerla N.Y., Kainz B., *et al.*: Attention u-net: Learning where to look for the pancreas, *arXiv preprint arXiv:180403999*, 2018.

[13] Parmar C., Grossmann P., Rietveld D., Rietbergen M.M., Lambin P., Aerts H.J.: Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer, *Frontiers in oncology*, vol. 5, p. 272, 2015. doi: 10.3389/fonc.2015.00272.

[14] Ribeiro M.T., Singh S., Guestrin C.: " Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

[15] Rizopoulos D.: Book review: Max Kuhn and Kjell Johnson. Applied Predictive Modeling. New York, Springer, *Biometrics*, vol. 74(1), pp. 383–383, 2018.

[16] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. doi: 10.1109/iccv.2017.74.

[17] Simonyan K., Vedaldi A., Zisserman A.: Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint arXiv:13126034*, 2013.

[18] Palatnik de Sousa I., Maria Bernardes Rebuzzi Vellasco M., Costa da Silva E.: Local interpretable model-agnostic explanations for classification of lymph node metastases, *Sensors*, vol. 19(13), p. 2969, 2019. doi: 10.3390/s19132969.

[19] Springenberg J.T., Dosovitskiy A., Brox T., Riedmiller M.: Striving for simplicity: The all convolutional net, *arXiv preprint arXiv:14126806*, 2014.

[20] Tsang M., Cheng D., Liu Y.: Detecting statistical interactions from neural network weights, *arXiv preprint arXiv:170504977*, 2017.

[21] Tu Z., Gao S., Zhou K., Chen X., Fu H., Gu Z., Cheng J., Yu Z., Liu J.: SUNet: A lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1378–1382, IEEE, 2020. doi: 10.1109/isbi45749.2020.9098673.

[22] Zeiler M.D., Fergus R.: Visualizing and understanding convolutional networks. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833, Springer, 2014. doi: 10.1007/978-3-319-10590-1_53.

[23] Zhang L., Fried D.V., Fave X.J., Hunter L.A., Yang J., Court L.E.: IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics, *Medical physics*, vol. 42(3), pp. 1341–1353, 2015.

[24] Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A.: Learning deep features for discriminative localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016. doi: 10.1109/cvpr.2016.319.

Affiliations

Reza Hassanpour
Computer Science Department, Groningen University, The Netherlands,
r.zare.hassanpour@rug.nl

Kasim Oztoprak
Computer Engineering Department, Konya Food and Agriculture University, Turkey,
kasim.oztoprak@gidatarim.edu.tr

Niels Netten
Research Center Creating 010, Rotterdam University, The Netherlands, c.p.m.netten@hr.nl

Tony Busker
Research Center Creating 010, Rotterdam University, The Netherlands, a.l.j.busker@hr.nl

Mortaza S. Bargh
Research and Data Center, Ministry of Justice and Security, The Netherlands, m.shoae.bargh@wodc.n

Sunil Choenni
Research Center Creating 010, Rotterdam University, The Netherlands, and Research and Data Center, Ministry of Justice and Security, The Netherlands, r.choennie@hr.nl

Beyza Kizildag
Computer Engineering Department, Konya Food and Agriculture University, Turkey, beyzaberen01@gmail.com

Leyla Sena Kilinc
Computer Engineering Department, Konya Food and Agriculture University, Turkey, leylasenakilinc@gmail.com

Received: 09.09.2024
Revised: 15.10.2024
Accepted: 06.04.2025