JOURNAL
COMPUTER SCIENCES INSTITUTE

JCSI 38 (2026) 51-58
Received: 21 September 2025
Accepted: 3 December 2025

Comparative analysis of interpretable artificial intelligence methods

Aleksandra Kuszewska*, Małgorzata Charytanowicz

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract

The aim of this article is to analyze and compare methods for explaining the results of artificial intelligence methods. Three methods were analyzed: Grad-CAM, SHAP, and LIME, evaluated in terms of their effectiveness on different data types. The analysis used five datasets: Iris, Wine Quality, Brain Tumor Dataset, PHCD, and WheatGrain. Two datasets are tabular, two are image, and one is mixed. SHAP and LIME were applied to tabular datasets, while all three methods were used for image data. Grad-CAM proved the fastest and most effective in locating key regions, while SHAP was slower but more accurate in pixel attribution, and LIME achieved the lowest precision. For tabular data, SHAP provided more accurate and consistent explanations than LIME, especially for high-dimensional datasets.

Keywords: Explainable Artificial Intelligence (XAI); Grad-CAM; SHAP; LIME, deep learning interpretability; visual explanations

*Corresponding author

Email address: s95466@pollub.edu.pl (A. Kuszewska)

1. Introduction

Artificial intelligence is currently one of the fastest-growing areas of technology. When used properly, it can streamline and support work in the most important fields, such as medicine, education, finance, industry, data analysis, and many others [1]. This allows specialists to focus on more complex challenges and the development of new ideas, both in the context of their fields and in terms of innovative ways of using artificial intelligence in that field. The development of machine learning, particularly deep neural networks, has led to a significant increase in the effectiveness of predictive models. These models not only allow for the correct processing of natural language, but also for detailed analysis of tabular and image data [2]. Despite these tremendous achievements, many professionals who were not previously familiar with machine learning may have difficulty interpreting and analyzing the results of artificial intelligence. A detailed understanding of all the decision-making processes of language models is crucial, especially in such important fields as medicine, where decisions can have a direct impact on human life.

Deep learning-based models, such as convolutional neural networks (CNNs), are often referred to as „black boxes” [3] because it is difficult to understand why they made specific decisions. The inability to fully understand them raises mistrust in the use of this type of assistance by both doctors and, indirectly, patients. A survey conducted by Stack Overflow Developer Survey 2024 shows that only 2.7% of respondents have complete confidence in the results of language models [4]. Even despite the very high effectiveness of the models, errors can always occur. To increase the level of trust in language models, a number of Explainable Artificial Intelligence (XAI) methods have been developed in recent years to improve the comprehensibility of decisions made [5].

XAI methods such as Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive explanations (SHAP), Gradient-weighted Class Activation Mapping (Grad-CAM) enable the analysis of the impact of input features on model. These explainability techniques can be applied to various types of machine learning models, including regression and classification models [6]. In this work, we focus specifically on the problem of classification. This work makes a clear contribution to the field of explainable artificial intelligence (XAI). It compares three XAI techniques: Grad-CAM, SHAP, and LIME and to evaluate the effectiveness of these techniques in two contexts – both for image and tabular data. The comparison will cover both quantitative aspects, such as computation time and explanation accuracy, and qualitative aspects, including the comprehensibility of the results. The results provide practical insights into the strengths and limitations of each method. They support informed selection of XAI techniques across different application contexts.

2. Characteristics of the data

The research material consisted of image collections and tabular data sets. Below is a detailed description of the collections used in the analysis.

2.1. Image data sets

This subsection describes three image data sets.

PHCD (Polish Handwritten Characters Database) [7] is a dataset of images containing handwritten characters. The collection contains 89 different characters, including lowercase and uppercase letters of the Latin alphabet, diacritical marks of the Polish language, numbers, and selected special characters. Each character is written in at least 6000 different ways, which guarantees a high diversity of data. Sample images of characters are shown below in Figure 1.

Figure 1: Sample images from PHCD Dataset (author's own visualization based on the original dataset [7]).

WheatGrain Dataset [8] is a collection of both image and tabular data. The image section contains 288 X-ray images of wheat grains. The dataset includes three classes: Canadian (108 samples), Kama (76 samples), and Rosa (108 samples). Figure 2 shows a sample of this collection.

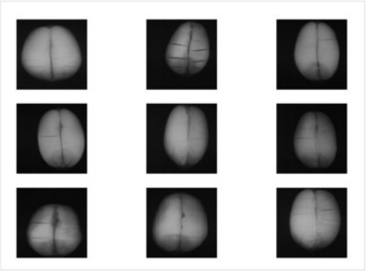

Figure 2: Sample images from the WheatGrain Dataset (author's own visualization based on the original dataset [8]).

Brain Tumor Dataset [9] is a collection of brain images taken using magnetic resonance imaging for binary classification (tumor vs. healthy brain). It contains both scans showing tumorous changes and images of healthy brains. The collection is characterized by high complexity and diversity of data. The images also show noise typical for medical images. Figure 3 shows a fragment of the Brain Tumor Dataset.

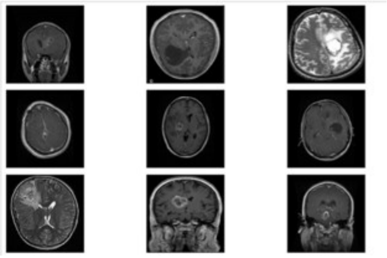

Figure 3: MRI sample images from the Brain Tumor Dataset (author's own visualization based on the original dataset [9]).

2.2. Tabular data sets

This subsection describes three sets of tabular data.

Iris Dataset [10] is a collection of 150 flower samples of three different species: Iris setosa, Iris versicolor, and Iris virginica. Each species is described by four numerical characteristics: sepals length, sepals width, petals length, and petals width.

Wine Quality Dataset [11] is a collection of physicochemical properties of Portuguese wines. A subset of only red wine is analysed, which contains 1,599 samples. Each sample is described by eleven independent variables: volatile acidity, citric acid, sulphur dioxide concentration, alcohol level, residual sugar, density, pH, sulphates, chlorides, sulphur dioxide. The dependent variable is wine quality, rated on a scale of 0 to 10 by social experts.

WheatGrain Dataset [12] is the one discussed earlier in the context of images. It also includes a tabular data section that describes the morphological characteristics of wheat grains. The dataset contains data such as: id, unique sample identifier, grain surface area, grain circumstance, grain compactness, grain length, grain width, asymmetry coefficient, groove length on the surface, germ area, germ length.

The issue of determining the geometric characteristics of wheat grain varieties, on the basis of which it is possible to discriminate between them, is discussed in articles [13].

3. Methods

This chapter describes explainable artificial intelligence methods, classifiers and metrics used in the analysis.

3.1. Explainable AI methods

This chapter describes the methods used to explain the results, i.e., Grad-CAM, SHAP, and LIME.

Grad-CAM is a technique used to explain results in deep neural network models. Its purpose is to indicate the relevant parts of an image that most influenced the model's decision. The Grad-CAM method uses gradients to obtain weights for the filter channels. Next, an activation map is created and ReLU is applied to retain only positive influences on the class. [14].

The **SHAP** method [15] s an explanatory method based on Shapley's game theory, which offers global and local explanations. The idea behind game theory is to determine the contribution of a single player to the outcome of the game. SHAP assigns values to each input feature that show its contribution to the model's final decision.

The **LIME** method [16] is an explanatory technique that allows us to understand which input features have the greatest impact on the model's decision. LIME divides images into superpixels, and creates multiple variants of the image around each input point, then a temporary simple model is created. This model is used to explain the main model's prediction for the changed area. The result indicates which fragments caused the model's decision to change.

3.2. Classification models

The following classifiers were used in the study:

Deep neural networks (DNN) **–** these are neural networks with multiple layers that process data. Neural networks learn input data representations through nonlinear activation functions and backpropagation [17]. Convolutional neural networks (CNNs) were used for both tabular and image data.

Random Forest is a machine learning algorithm [18]. It is based on multiple decision trees. Each tree is built on the basis of random data samples and a random subset of features. Each decision tree generates its own prediction, and the final classification decision is made based on the decision of the majority of trees. In regression, on the other hand, the average of all trees' predictions is taken. Random Forest is characterized by high accuracy, resistance to noise and missing data, and good generalization.

DenseNet121 (Densely Connected Convolutional Network) is an architecture distinguished by dense connections between layers. It consists of 121 layers. Each layer receives feature maps from all previous layers as inputs and passes its own to subsequent layers. This dense connectivity enables better gradient flow, preventing gradient vanishing. Small gradients (close to 0) cause significant slowdowns in network learning. DenseNet achieves strong performance with fewer parameters. [19].

MobileNetV2 is an efficient and fast deep convolutional network [20]. The architecture uses inverted residual blocks. It works by keeping the channel space narrow at the block input, then expanding it to allow for more complex features to be processed and then narrowing the number of channels again. The block's input data is added to its output through residual connections. This allows for better gradient flow during learning.

3.3. Metrics

This subsection presents quantitative metrics used to evaluate the performance of explainable artificial intelligence (XAI) methods. The following metrics were applied:

Computation Time – represents the time (in seconds) required to generate an explanation.

Fidelity – measures how well the explanation reflects the model's decision. Values closer to 1 indicate higher agreement between the explanation and the model output. [21]

Comprehensiveness – evaluates how well the explanation captures the most influential features. [22]

Sufficiency– indicates whether the selected features alone are sufficient to obtain a similar prediction. [23]

Stability – assesses robustness to perturbations and noise. Higher values denote greater stability. [24]

Robustness – measures consistency under data transformations such as rotation or scaling.[25]

4. Classification process

To ensure a fair comparison, key issues were taken into account, such as: the same learning model, identical input data processing, use of the same number of background samples, consistent scale, and data normalization. Classifiers were selected based on data characteristics: Random Forest for small tabular datasets (Iris), CNNs for complex pattern recognition (Wine Quality, WheatGrain tabular, PHCD), and pre-trained models (DenseNet121, MobileNetV2) for image classification to utilize weights trained on ImageNet, reducing training time and improving performance on limited datasets.

4.1. Image data sets

Brain Tumor Dataset

- **Classification model**: DenseNet121 (CNN)
- **Objective**: Binary classification of brain MRI images (tumor vs. healthy)
- **Description**:
  - Deep networks with dense connections between layers
  - Weights pre-trained on ImageNet
  - Added layers: `Flatten, Dropout, BatchNormalization, Softmax`
  - Loss function: `binary crossentropy`
  - Optimizer: `Adam, learning rate` = 0.0001
  - Data augmentation
- **Results**:
  - Accuracy: 99.97%
  - Loss: 0.0056

PHCD

- **Classification model**: CNN (Convolutional Neural Network), library Keras
- **Objective:** Recognition of characters from 89 classes of handwritten characters, i.e.: Digits, Letters, Special characters
- **Description**:
  - Two convolutional blocks
  - First block: two `Conv2D` convolutional layers with 32 filters (5x5), `padding, ReLU, MaxPooling2D` (32x32 to 16x16), `Dropout`
  - Second block: two convolutional layers with 64 filters, `ReLU`, `MaxPooling2D` (16x16 to 8x8), `Dropout`
  - Fully connected layer: `Flatten, Dense + ReLU, Dropout, Dense` (89 neurons) + `softmax`
  - Difficulties in recognizing similar characters, e.g., "0" and "O"
- **Results:**
  - Model accuracy: 82.16%
  - Loss: 28.76%

WheatGrain Dataset (image data)

- **Classification model**: MobileNetV2 (CNN)
- Objective: Classification of wheat grain varieties (Canadian, Kama, Rosa)
- **Description**:
  - Weights pre-trained on ImageNet
  - Real-time data normalization and augmentation
  - Layers: `GlobalAveragePooling2D, Dense + ReLU, Dropout`
  - `Callbacks: EarlyStopping, ReduceLROnPlateau, ModelCheckpoint`
- **Results**:
  - Accuracy: 86%
  - Loss: 36%

4.2. Tabular data sets

Iris Dataset

- **Classification model**: Random Forest
- **Objective**: Classification of iris flower species (setosa, versicolor, virginica)

- **Description**: One hundred decision trees
- **Results**:
  - Accuracy: 100%
  - Loss: 1%

Wine Quality Dataset

- **Classification model**: CNN (`Conv1D`)
- **Objective**: Wine quality classification based on physicochemical properties
- **Description**:
  - Two convolutional layers + `ReLU`
  - `L2 regularization`
  - Layers: `Flatten, Dense` with normalization, `Softmax`
- **Results**:
  - Accuracy: 86.05%
  - Loss: 5.78%

WheatGrain Dataset (tabular data)

- **Classification model**: CNN (`Conv1D`)
- **Objective**: Classification of wheat grain varieties (Canadian, Kama, Rosa)
- **Description**:
  - Two convolutional blocks with two `Conv1D` layers
  - Activation functions: `LeakyReLU`
  - `BatchNormalization, MaxPooling1D, Dropout`
  - Additionally: loss function + `SGD` with `learning rate` and `momentum`
  - Stage one of training: 10-fold cross-validation after 100 epochs
  - Stage two of training: final model on the full dataset for 100 epochs
- **Results**:
  - Accuracy: 96%
  - Loss: 9.62%

Classification accuracy ranged from exceptional results for Iris (100%) and Brain Tumor (99.97%) to more challenging performance for PHCD (82.16%) due to visual similarity between classes such as digit "0" and letter "O", and WheatGrain images (86%) due to subtle morphological differences between grain varieties. These baseline performance differences provide crucial context for evaluating XAI method effectiveness in subsequent analyses.

5. Results

5.1. Image data sets

This chapter describes the results of explaining the results of explaining the Grad-CAM, SHAP, and LIME methods for image data.

Interpretation of visualization colors for all methods indicates that red/yellow areas indicate high activation (important regions), blue/black areas indicate low activation. Color intensity corresponds to the strength of feature importance

PHCD

Three methods for explaining CNN model results for the PHCD dataset were compared. Three characters were compared: the letter “i”, the letter “ś” and the dollar sign “$”. One character was selected at random, “i”, and the characters “ś” and “$” were selected due to their visual similarity. The selection was dictated by the desire to check which elements of the image are important during model classification.

On the left side of Figure 4 **Grad-CAM** focuses on the key elements of the letter “i”. This is indicated by the intense light blue turning into yellow in the central area of the character, which is shown in the image on the right titled “Grad-CAM Heatmap.” It can be seen that the lower part of the letter proved to be the most important during the analysis, as this area was marked in yellow and red. The heatmap almost completely overlaps with the original image, which indicates the effectiveness of the Grad-CAM method. The second character analysed was the letter “ś”. On the heatmap, the enhanced areas accurately indicated the shape of the letter “ś”. The third character examined was the “$” sign. It can be seen that the model correctly distinguished and classified the characters ‘ś’ and “$”. The Grad-CAM method enhanced the line passing through the character “S,” indicating the dollar sign, while during the classification of the letter “ś,” the line above the letter was enhanced, allowing both characters to be distinguished. The analysis time was 0.158 seconds.

The **SHAP** value map shows the precise distribution of the influence of individual pixels on the model prediction. The SHAP method allows for a mathematically justified assignment of the impact of each feature, which makes it exceptionally accurate at the pixel level. The SHAP analysis time was 24.248 seconds. The results are presented in Figure 4.

The figure 4 also shows the result of the **LIME** method. The segmentation is clear, and the boundaries between regions are sharp. The red and yellow areas coincide with the key shape of letters, which confirms their importance in the classification process. The analysis time in this case was 15.692 seconds.

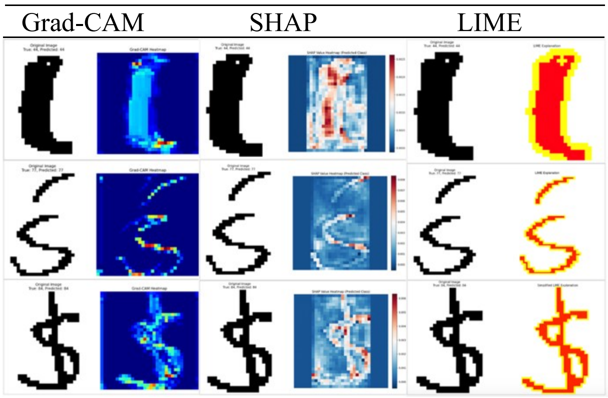

Figure 4: PHCD Dataset – original images with Grad-CAM, SHAP, and LIME explanations.

Table 1 presents metrics for image data for the PHCD collection.

Table 1: Metrics for the PHCD dataset

Metrics	Grad-CAM	SHAP	LIME
Computation time [s]	0.158	24.248	15.692
Fidelity	0.682	0.725	0.417
Comprehensiveness	0.719	0.325	0.603
Sufficiency	0.880	0.356	0.446
Stability	0.783	0.887	0.526
Robustness	0.432	0.657	0.600

The Grad-CAM method proved to be the fastest, while the SHAP method was the slowest. In terms of fidelity, SHAP achieved the highest score, Grad-CAM a moderate score, and LIME the lowest score. Grad-CAM performed well in comprehensiveness and faithfulness. LIME excels in explanation transformation, and SHAP is the most effective here.

Brain Tumor Dataset

In the case of the **Grad-CAM** method, high activation concentration was observed in areas corresponding to pathological changes. Grad-CAM effectively localized tumor masses in various sections of the brain, especially in the upper, posterior, and subcortical areas. This method proves to be intuitive for specialists, allowing for precise identification of relevant regions of the image. The analysis time was 0.468 seconds. The results are shown in Figure 5.

Analysis of **SHAP** results also showed high effectiveness in detecting the diseased area. However, the marked areas are only slightly highlighted, which indicates lower sensitivity to pathological changes. The results are presented in Figure 5.

The areas marked by the **LIME** method in Figure 5 only partially overlapped with the tumor areas. The effectiveness of this method proved to be the lowest in comparison with the previous interpretability methods.

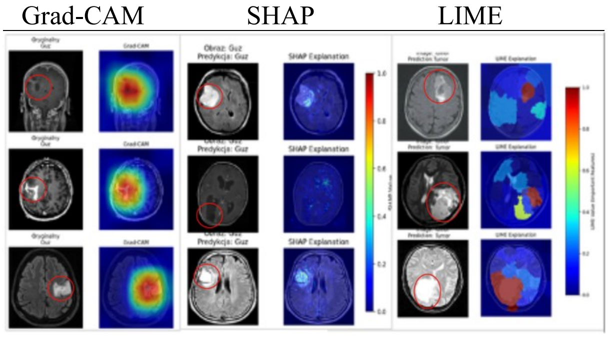

Figure 5: Brain Tumor Dataset – original images with Grad-CAM, SHAP, and LIME explanations.

Table 2: Metrics for the Brain Tumor Dataset

Metrics	Grad-CAM	SHAP	LIME
Computation time [s]	0.468	28.293	43.670
Fidelity	0.950	0.804	0.781
Comprehensiveness	0.307	0.189	0.147
Sufficiency	0.919	0.909	0.842
Stability	0.881	0.675	0.732
Robustness	0.716	0.453	0.467

The analysis of metrics (Table 2) showed that the **Grad-CAM** method achieved the best results among all three methods tested. The **LIME** method proved to be better than the **SHAP** method only in the case of noise and image transformation.

WheatGrain Dataset

This subsection shows a comparison of explanatory methods in the context of wheat image classification.

Figure 6 shows the original sample and the heat map. In the Canadian sample (top row), the Grad-CAM method indicated enhancements in the left part of the grain, which may indicate the importance of asymmetry during classification. In the example of the Kama grain (middle row), the enhancement occurred in the central part of the image. In contrast, in the Rosa sample (bottom row), the enhancements were much less accurate, spreading beyond the grain area. The classification confidence in this case was very low, at only 54.63%.

The **SHAP** method indicated enhancements in the contour area for the Canadian and Kama varieties. The method highlighted irregular reinforcements on the left side of the grain for the Canadian variety and reinforcements on the right side for the Kama variety. The SHAP method highlighted the grain groove in the Rosa variety sample.

The **LIME** method identified the most relevant area in the central part of the wheat grain (grain groove).

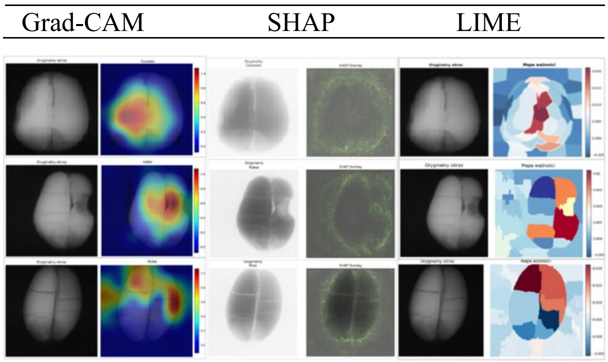

Figure 6: WheatGrain Dataset – original images with Grad-CAM, SHAP, and LIME explanations.

Table 3 presents metrics for image data for the WheatGrain Dataset.

Table 3: Comparison of LIME and SHAP metrics

Metrics	Grad-CAM	SHAP	LIME
Computation time [s]	0.113	13.067	5.76
Fidelity	0.717	0.898	0.318
Comprehensiveness	0.598	0.689	0.147
Sufficiency	0.749	0.678	0.542
Stability	0.899	0.780	0.732
Robustness	0.760	0.360	0.167

An analysis of six metrics showed that the Grad-CAM method achieved the fastest time. The SHAP method proved to be the slowest and most accurately maps significant areas of the image. LIME is characterized by significantly lower precision compared to the other methods. The SHAP and Grad-CAM methods effectively indicate significant pixels, and both methods achieved high metric values in terms of the sufficiency of key pixels. Grad-CAM achieved the highest resistance to noise stability and transformations. None of the methods dominates in all categories, which means that each has its strengths and weaknesses.

5.2. Tabular data

This chapter describes the results of explaining the SHAP and LIME methods for tabular data.

For **SHAP** and **LIME** methods, bar charts show feature contributions to model predictions. Positive values (green bars extending right) indicate features that increase the prediction probability, while negative values (red bars extending left) decrease it. Bar height reflects the strength of feature contribution to the explanation.Iris Dataset

This subsection shows a comparison of explanatory methods in the context of Iris data classification.

Figure 7 shows a comparison of the impact of features on the model's decision using the LIME and SHAP methods. Both methods consistently indicate that “petal width” and “petal length” are the most important features. Both LIME and SHAP assign them the highest weights.

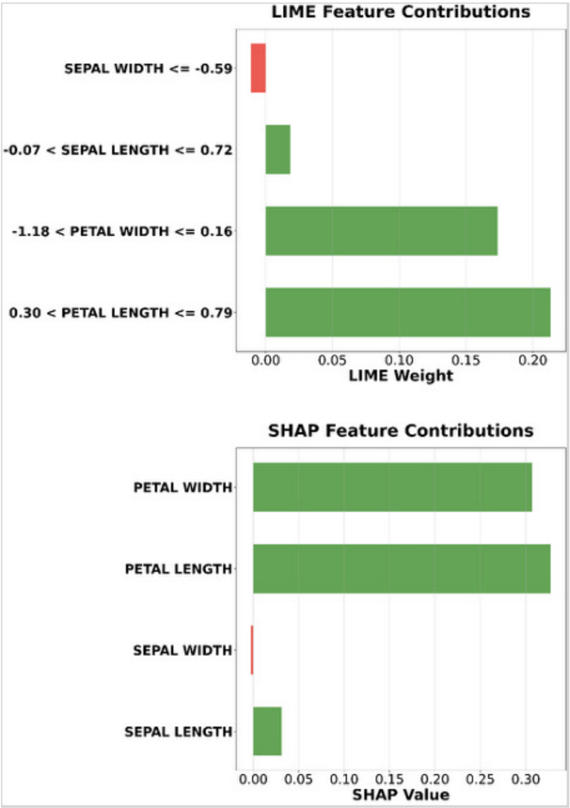

Figure 7: SHAP and LIME plots for the Iris Dataset.

Table 4 presents a comparison of metrics evaluating the LIME and SHAP methods. In terms of computation time, SHAP proves to be significantly slower than LIME, which can be important when working with large datasets. In terms of fidelity, LIME performs better, meaning it is more faithful to predictions in the vicinity of the analyzed example. However, both methods achieve chemical accuracy. In comprehensiveness, which assesses whether explanations cover all important features on the model's decision, SHAP achieves a better result. In terms of stability, LIME shows greater repeatability. Resistance to noise indicates how interpretations remain stable despite data perturbations, and in this case, SHAP is more resistant.

Table 4: Comparison of LIME and SHAP metrics

Metrics	SHAP	LIME
Computation time [s]	0.150	0.023
Fidelity	0.881	0.961
Comprehensiveness	0.930	0.930
Sufficiency	0.880	0.790
Stability	0.938	0.980
Robustness	0.999	0.990

The choice between these methods depends on the specific needs of the user, as both methods have their advantages and disadvantages but achieve similar results.

Wine Quality Dataset

This subsection shows a comparison of explanatory methods in the context of wine data classification.

The graph shows a local picture of the importance of individual features in the SHAP and LIME method. The results are shown in Figure 8.

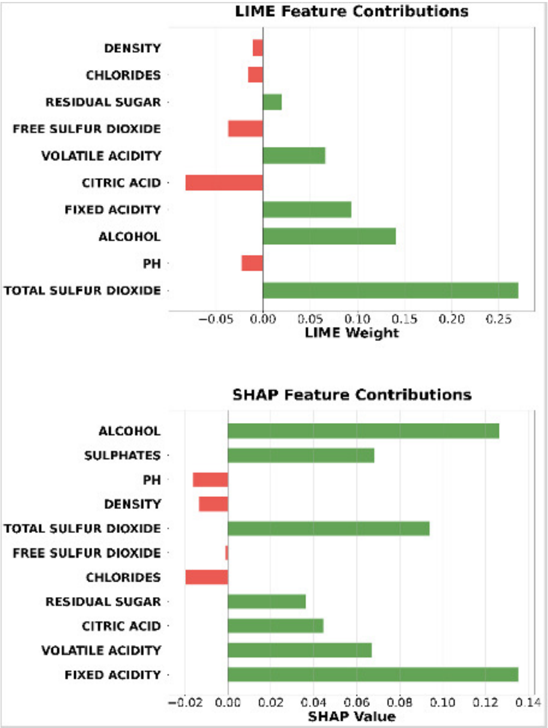

Figure 8: SHAP and LIME plots for the Wine Quality Dataset

Table 5 presents metrics for tabular data for the Iris Dataset.

Table 5: Comparison of LIME and SHAP metrics

Metrics	SHAP	LIME
Computation time [s]	3.005	0.240
Fidelity	0.521	0.350
Comprehensiveness	0.511	0.287
Sufficiency	0.985	0.698
Stability	0.919	0.922
Robustness	0.923	0.924

The time achieved by the LIME method is significantly shorter than that achieved by the SHAP method. The method calculating Shapley values achieved higher results such as explanation fidelity, the importance of key pixels, and the sufficiency of relevant pixels during model prediction. The LIME method proved to be slightly more stable and resistant to noise and transformations.

5.3. WheatGrain Dataset

Analysis of the X-ray image of the Canadian10_11 wheat sample showed that grain circumference, asymmetry coefficient, and groove length had the greatest positive impact on classification. The results are shown in Figure 9.

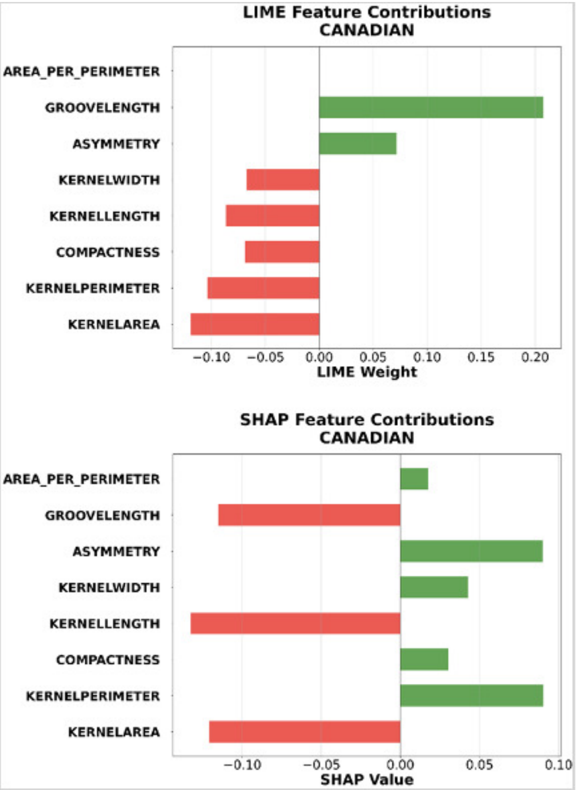

Figure 9: SHAP and LIME plots for the WheatGrain Dataset

Table 6 presents metrics for tabular data for the WheatGrain Dataset.

Table 6: Comparison of LIME and SHAP metrics

Metrics	SHAP	LIME
Computation time [s]	0.720	0.138
Fidelity	0.671	0.557
Comprehensiveness	0.589	0.347
Sufficiency	0.905	0.797
Stability	0.923	0.942
Robustness	0.958	0.918

When analyzing metric values, the SHAP method is slower. This may be due to the specifications of this method, as calculating Shapley values takes more time than creating a simple local model, as LIME does. The SHAP method is slower than LIME when analyzing metric values. This may be due to the specifications of this method, as calculating Shapley values takes more time than creating a simple local model, as LIME does.

7. Summary and conclusions

A comparative analysis of interpretable artificial intelligence methods has shown that there is no universal method that surpasses the others in all aspects. The most important factor in the analysis is the most accurate selection of a model tailored to the analysis of a specific type of data (tabular, image). The Grad-CAM method proved to be the fastest of all methods when analyzing images. In the case of brain classification, it achieved the highest accuracy. It can therefore be concluded that it is a good solution for medical diagnostics, as time and high location precision are key. The SHAP method achieved the longest computation time. However, the method proved to be very stable during noise and resistant to transformations. When analyzing the WheatGrain dataset and characters from the PHCD dataset, the SHAP method achieved the highest fidelity among all three methods. A major advantage of this method is the ability to analyze multiple samples simultaneously on a global scale. In some cases, e.g., in an industrial context, SHAP analysis can provide general patterns of the influence of individual features. In contrast, LIME and Grad-CAM show only a single example for one sample, which can be misleading if that sample is not representative. The LIME method proved fastest for tabular data but demonstrated limitations in image analysis through oversimplified explanations and lowest fidelity scores, though its conversion of continuous values to intervals provides more intuitive explanations for non-technical users.

A particularly interesting finding emerges from the WheatGrain dataset, which uniquely combines both image and tabular data for the same samples. The analysis reveals partial convergence between image-based and tabular-based explanations. Image analysis methods primarily focused on grain contours, central groove, and asymmetrical features visible in X-ray images. Meanwhile, tabular data analysis consistently identified grain circumference, asymmetry coefficient, and groove length as the most important features. This convergence suggests that both approaches capture similar underlying morphological characteristics, with image-based methods detecting spatial patterns that correspond to the numerical features calculated from the same geometric properties. Future research should focus on developing hybrid approaches that combine strengths of multiple XAI methods, particularly for multisource datasets like WheatGrain, presents a promising research direction for improving explanation quality.

Literature

[1] Y. N. Harari, Nexus. Krótka historia informacji – od epoki kamienia do sztucznej inteligencji, Wydawnictwo Literackie, Kraków, 2024.

[2] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521(7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[3] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, S. Turini, Meaningful Explanations of Black Box AI Decision Systems, Proceedings of the 33rd AAAI Conference on Artificial Intelligence (2019) 9780–9784, https://doi.org/10.1609/aaai.v33i01.33019780.

[4] Survey results on AI usage by developers, Stack Overflow, Developer Survey 2024 – AI, https://survey.stackoverflow.co/2024/ai#developer-tools-ai-acc, [22.09.2025].

[5] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[6] S. Nazim, M. M. Alam, S. S. Rizvi, J. C. Mustapha, S. S. Hussain, M. M. Suud, Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM, PLoS ONE 20(5) (2025) e0318542, https://doi.org/10.1371/journal.pone.0318542.

[7] Polish Handwritten Characters Database (PHCD), https://cs.pollub.pl/phcd/, [22.09.2025].

[8] WheatGrain Images, https://zenodo.org/records/15172506, [22.09.2025].

[9] Brain Tumor Dataest, https://www.kaggle.com/datasets/preetviradiya/brian-tumor-dataset, [22.09.2025].

[10] Iris Dataset, https://www.kaggle.com/datasets/himanshunakrani/iris-dataset, [22.09.2025].

[11] Wine Quality Dataset, https://www.kaggle.com/datasets/yasserh/wine-quality-dataset, [22.09.2025].

[12] WheatGrain Features, https://zenodo.org/records/15172506, [22.09.2025].

[13] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, Discrimination of Wheat Grain Varieties Using X-ray Images, In: E. Pietka, P. Badura, J. Kawa, W. Więcławek (eds), Information Technologies in Biomedicine. Advances in Intelligent Systems and Computing, Springer 471 (2016) 39–50, https://doi.org/10.1007/978-3-319-39796-2_4.

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

[15] V. Vimbi, N. Shaffi, M. Mahmud, Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection, Brain Informatics 11 (2024) 10, https://doi.org/10.1186/s40708-024-00222-1.

[16] A. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, G. Menegaz, K. Lekadir, A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME, Advanced Intelligent Systems 7(1) (2024) 2400304, https://doi.org/10.1002/aisy.202400304.

[17] A. Kuanar, A. Akbar, P. Sujata, D. Kar, Deep Neural Network (DNN) Modelling for Prediction of the Mode of Delivery, European Journal of Obstetrics & Gynecology and Reproductive Biology 297 (2024) 241–248, https://doi.org/10.1016/j.ejogrb.2024.04.012.

[18] S. J. Rigatti, Random Forest, Journal of the American Medical Informatics Association 24(6) (2017) 1015–1024, https://doi.org/10.1093/jamia/ocx133.

[19] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2261–2269, https://doi.org/10.1109/CVPR.2017.243.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.

[21] M. Pawlicki, A. Pawlicka, F. Uccello, S. Szelest, S. D'Antonio, R. Kozik, M. Choraś, Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination, Neurocomputing 602 (2024) 128282, https://doi.org/10.1016/j.neucom.2024.128282.

[22] W. J. Yeo, W. van der Heever, R. Mao, E. Cambria, R. Satapathy, G. Mengaldo, A comprehensive review on financial explainable AI, Artificial Intelligence Review 58 (2025) 189, https://doi.org/10.1007/s10462-024-11077-7.

[23] D. S. Watson, L. Gultchin, A. Taly, L. Floridi, Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice, Minds and Machines 32(3) (2022) 475–498, https://doi.org/10.1007/s11023-022-09598-7.

[24] M. Saarela, L. Geogieva, Robustness, Stability, and Fidelity of Explanations for a Deep Skin Cancer Classification Model, Applied Sciences 12(19) (2022) 9545, https://doi.org/10.3390/app12199545.

[25] L. Coroamă, A. Groza, Evaluation Metrics in Explainable Artificial Intelligence (XAI), Proceedings of the International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability (ARTIIS) (2022) 401–413, https://doi.org/10.1007/978-3-031-20319-0_30.