E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia Science & Technology Asia
Vol.30 No.4 October - December 2025 Page: [254-269]
Original research article

A Model for Smart Detection: Modified Explainable Machine Learning for Interpreting Detection Results

Alifia Revan Prananda[1,*], Eka Legya Frannita[2]

[1]*Department of Information Technology, Universitas Tidar, Magelang 56116, Indonesia*
[2]*Department of Leather Product Processing Technology, Politeknik ATK Yogyakarta, Yogyakarta 55188, Indonesia*

Received 9 May 2025; Received in revised form 9 September 2025
Accepted 7 October 2025; Available online 17 December 2025

ABSTRACT

Medical imaging analysis using artificial intelligence has become a powerful system for assisting the doctor in diagnosing some diseases. Most of CAD performed excellent performance with average accuracy of more than 80%. Regardless of the excellent performance of artificial intelligence in the CAD, implementation of artificial intelligence in medical cases is still causing controversy. It happened due to the black-box principle of AI. Actually, both machine learning and deep learning worked in the black-box direction in which it was difficult to recognize how the model performed and how it analyzed the data. Hence, it became controversial since there remained some big questions about "how can the doctor trust the AI result?" Regarding this problem, a continued solution was needed. In this study we proposed a modified SHAP for explaining the artificial intelligence result. The modification itself is conducted by performing correlation in the perturbation process of SHAP. Our proposed solution was performed into two different datasets to evaluate the significance and the reliability of the proposed solution. According to both the visual analysis and statistical test, we conclude that the proposed solution gave a more rational explanation compared to the original SHAP.

Keywords: Artificial intelligence; Explainable method; SHAP method; Statistical evaluation; Thyroid cancer

***Corresponding author:** revan@untidar.ac.id **doi:** 10.14456/scitechasia.2025.83

1. Introduction

The maturity of research work in the implementation of artificial intelligence has been extending tremendous opportunities for innovation in numerous applications and in several disciplines. One of the prominent applications was the employment of artificial intelligence for medical sector [1]. The development of contemporary assistant technologies based artificial intelligence has inspired the development of new innovation in the medical clinical process [2], such as a well-understanding of genome principle [3], smart coaching for medical cases [4], intelligent scan reader [5], intelligent cancer detection and diagnosis [6, 7][8], intelligent mutations identification [9], and predicting the mortality [10]. Implementation of AI in medical cases ushered in an intelligent era by re-engineering and re-imagining the clinical and study abilities [11]. Hence, it has recently earned popularity across various medical domains for aiding in the recognizing the abnormality based on pathology samples or findings medical examination [12]. It also offered some promising methods for assessing subtle diseases in neuroimaging data. Machine learning and deep learning, as the most popular approach, had ability to detect and recognize the pattern automatically that may otherwise be challenging to distinguish through conventional procedures. AI has earned influential popularity in medical cases due to its ability in producing accurate detection, which similar to or sometimes more acceptable than, the medical personnels for determining several abnormalities (12), such as breast cancer [13–15], lung cancer [16], skin cancer [17], thyroid cancer [18–22], malaria [7, 23] and retinal diseases [24–26].

In radiological imaging cases, artificial intelligence has been perpetually evolving and used in plenteous applications. One of the most implementations was on thyroid cancer cases. Thyroid is an essential organ placed in the lower front of the human neck [27]. It is shaped like a butterfly and has function for managing human body's metabolism by secreting hormone regularly [27, 28]. It also helps to manage all components in human blood like calcium. Other functions are that it controls the blood pressure, helps to manage temperature of human body, and regulates the heart rate [27]. Thyroid cancer is like an abnormal condition occurred when the thyroid gland is disrupted so that it produces unstable amount of the hormones [27]. The epidemiological report has been analyzed that the abnormality in human thyroid gland is estimated occurred in 19%-68% of population in which 5%-15% of them define as malignant cases [29]. Some countries also have been reported that the grown up of thyroid cancer has reached up to 4.5% per year in which it increased faster than other cancers [30].,Nowadays, the medical imaging analysis using artificial intelligence becomes powerful system for assisting the doctor in diagnosing some diseases including thyroid cancer [31]. One of well-known intelligent medical imaging analysis system -called computer aided diagnosis (CAD)- has been widely developed including in the thyroid cases [22, 32–35]. Almost CAD performed excellent performance with average accuracy of more than 80%. According to those results, we can conclude that artificial intelligence has great impact for helping the medical problems.

Achieving high performance in solving the medical cases using artificial intelligence may not be enough to be used in daily tasks. Regardless of the aforementioned research obtained excellent performance, implementation of artificial intelli-

gence in medical cases is still causing controversy. It happened due to the black-box principle of AI. Actually, we know that both machine learning and deep learning worked in the black-box direction in which we did not recognize how the model performed and how it analyzed our data. Hence, it became controversy since it remained a big question about "how can the doctor trust to the AI result?" [36]. Regarding the aforementioned issue, a continued solution was needed. An explainable method came by offering some abilities including explaining what happens inside the AI model. It was extremely useful for the doctor since it can assist the doctor in believing the AI result. Several research communities have been done in employing explainable artificial intelligence for medical cases. Seethi et.al (37) proposed an explainable artificial intelligence for diagnosing Covid-19 by employing MALDI-ToF mass spectrometry. In this research work, random forest classifier followed by shapely additive explanations was performed for interpreting 152 human gargle samples. This method successfully achieved accuracy of more than 90% and they also successfully provided trusty interpretation about the classification result. Deepanshi et. al [38] proposed explainable artificial intelligence for Covid-19 using NG-IoT models. This research work was done by employing several deep learning architectures such as ResNet-50, Inception V3, Densenet-121 and DCNN followed by explained method using Grad CAM++. According to their experimental results and medical professional evaluation, deep learning architecture followed with Grad CAM++ provided acceptable interpretation. Caroprese et. al [39] proposed an explainable artificial intelligence using argumentation approach. This study successfully obtained acceptable justification of artificial intelligence result. Martino et. al [40] proposed an explainable artificial intelligence method for predicting malnutrition risk. The explanation was created by employing shapely additive explanation. This research concluded that by applying shapely additive explanation method, they can produce clinically relevant prediction. Niranjan et. al [41] proposed fused classification and segmentation based XAI for diagnosing Covid-19. This proposed solution was performed in chest CT scans dataset. According to the experimental result, they concluded that an XAI method can visualize the region of interest localization perfectly in which it can assist the next process easily. Cao et. al [42] proposed an interpretable artificial intelligence for assisting the medical personnels in creating the decision in radiosurgery of brain metastases. This proposed method was employed in CT scan dataset collected from 152 patients. After conducting evaluations with the physician, they concluded that the proposed solution successfully obtained acceptable prediction.

Among the various explainable artificial intelligence (XAI) techniques, SHAP (SHapley Additive exPlanations) is widely regarded as one of the most robust methods. Theoretically, SHAP provides a framework for interpreting model predictions by computing the marginal contribution of each feature to the output. Thus, SHAP enables both global and local interpretability, offering insights into overall model behavior as well as individual prediction rationales. This dual capability makes SHAP particularly advantageous for uncovering complex feature–class relationships in high-dimensional datasets. Several recent studies successfully employed SHAP for explaining the AI results [43–48].

The aforementioned solution suc-

cessfully obtained significant methods, however in certain cases SHAP still remained the issues. Since SHAP only focus on the relationship between feature contribution with classes, it not considers the relationship between features. For example, in the classification process, the major problem in explainable method was that this method worked by only considering the relationship between each feature to the class. However, the relationship between feature to feature was not considered. Consequently, in some cases the explanation resulted from the SHAP method was not rational and logic. To solve that problem, we proposed modified explainable method by considering the relationship between feature to feature for generating more rational explanation. In this study, we focused on modifying shapely additive explanation (SHAP) method by adding correlation between features in the perturbation process. To address the major problem, we introduced the contribution including:

(a) (a) Modifying SHAP by considering correlation between features in developing the explainable model.

(b) Conducting experiments in different datasets to prove the significancy and the reliability of the proposed solution.

2. Materials and Methods

2.1 Brief description about dataset

This study used two datasets which are thyroid cancer dataset and breast cancer dataset. Both are tabular datasets. Although our research focuses on the thyroid nodule, we considered using breast cancer dataset since both thyroid and breast cancer had similar characteristics, hence the second dataset can be used to evaluate the significancy and the reliability of the proposed solution. Summary of both datasets can be seen in Table 1.

Table 1. Summary of datasets.

Component	Dataset 1	Dataset 2
Case and resource	Thyroid cases resulted from previous study [18]	Breast cancer Wisconsin Diagnostic dataset from UCI machine learning repository [49]
Type of data	Tabular data	Tabular data
Number of data	165 instances	569 instances
Number of features	8 features: convexity, circularity, tortuosity, rectangularity, width and height ratio, compactness, dispersy, and chain code difference	5 features: perimeter, radius, area, smoothness, and texture
Classes	2 classes: irregular/ malignant and smooth/benign	2 classes: benign and malignant

The first was thyroid instance dataset resulted from our previous study [18]. The dataset consisted of 165 instances comprising of 8 attributes and 2 classes (smooth and irregular). Smooth class indicated benign cancer, while irregular class indicated malignant cancer. Example of the dataset is depicted in Table 2.

Table 2. Example of thyroid dataset.

ID	Convexity	Circularity	…	Class
1.	0.6010	0.6989	· · ·	Smooth
2.	0.4813	0.6417	· · ·	Smooth
3.	0.7673	0.6329	· · ·	Smooth
4.	0.6833	0.6949	· · ·	Smooth
5.	0.6010	0.6989	· · ·	Smooth
⋮	⋮	⋮	⋮	⋮
165.	2.2955	0.7293	· · ·	Irregular

The second dataset was public breast cancer dataset called Breast Cancer Wisconsin Diagnostic [49] provided by UCI machine learning repository. The second dataset consisted of 569 instances having 5 features and 2 classes (benign and malignant). Example of this dataset can be seen in Table 3.

Table 3. Example of breast cancer dataset.

ID	Perimeter	Radius	…	Class
1.	17.99	10.38	· · ·	Benign
2.	20.57	17.77	· · ·	Benign
3.	19.69	21.25	· · ·	Benign
4.	11.42	20.38	· · ·	Benign
5.	20.29	14.34	· · ·	Benign
⋮	⋮	⋮	⋮	⋮
569	7.76	24.54	· · ·	Malignant

2.2 Experimental design

Development of the proposed solution began by collecting the data. In this study, we used tabular datasets of thyroid cancer and breast cancer cases. The datasets were then divided into two parts (training data and testing data) with proportion of 80:20 and were performed into two scenarios. In the first scenario, we trained and tested both datasets using SVM for classifying the data. After getting high classification accuracy, we took the testing data to be used in the next stage. The next stage was machine learning explanation process. This process was very important to support the classification result and to validate the prediction, hence the prediction became more transparent. The transparent result can be used to overcome the black box problem of artificial intelligence including machine learning approach. In this step, we performed original SHAP [11] for explaining the prediction result of applying SVM in classification process. After performing SHAP, we saved the visual result and SHAP value to be compared with our proposed solution in the second scenario of experiment.

In the second scenario, we also performed SVM to classify. After obtaining high accuracy, we explained the classification result using modified SHAP. The modification itself was conducted by performing correlation into perturbation process of SHAP. Results of conducting modified SHAP were then saved to be compared with the result of first scenario. In the next stage, we performed two types of analysis which are visual analysis and statistical analysis. The visual analysis was used to explain visually how the comparison results between scenario 1 and the proposed solution in scenario 2. While the statistical test was used to evaluate how the proposed solution give more impact to the prediction result and to show the significancy of the proposed solution. The detail of experimental design can be seen in Table 4 and Fig. 1.

Table 4. Components of the experiment.

Component	Scenario 1	Scenario 2
Dataset	Dataset 1 and dataset 2 (performing in parallel process)	Dataset 1 and dataset 2 (performing in parallel process)
Proportion of data	80:20 for training and testing	80:20 for training and testing
Classifier	SVM	SVM
Explainable method	Original SHAP	Modified SHAP
Result	Visual explanation and SHAP value	Visual explanation and SHAP value

2.3 Modified SHAP

SHAP (SHapley Additive exPlanations) is one of the famous explainable methods. This method explained the machine learning result by calculating the contribution of each feature to the prediction result. Contribution was calculated by measuring the important level of each feature to the prediction. After calculating important levels and determining contributions of each feature, the process was then continued by creating the coalition of each feature and calculating the marginal contribution of each feature in each coalition. Coalition was the combination between some features that may have contribution to the prediction result [50]. The result of this process was a SHAP value that indicated the important level of each feature to the prediction result. This process was mathematically for-

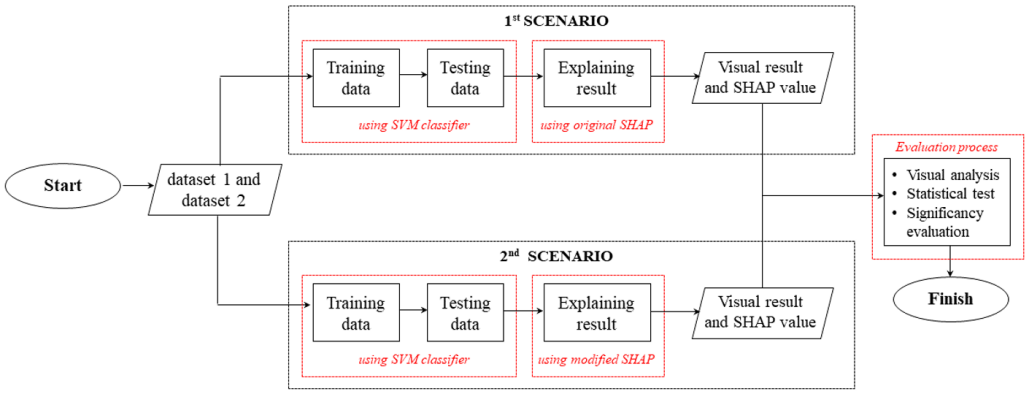

Fig. 1. Flowchart of the proposed solution.

mulated by Eq. (2.1) [51].

$$g(z') = \emptyset_) + \sum_{j=1}^{M} \emptyset_j z'_j, \qquad (2.1)$$

with g is explainable model, z' is coalition vector, M is maximum number of coalitions, $\emptyset_0$ is the first coalition that was appeared, $\emptyset_j$ is coalition of data j [51]. In this research work, the SHAP method itself was modified by considering correlation between features in permutation process of the measurement of SHAP value. The original permutation process of SHAP value is formulated in Eq. (2.2) [52]. Then, non-interventional conditional expectation was performed to simplify the equation as formulated in Eq. (2.3) [(2].

$$\emptyset_i = \frac{1}{M!}\sum_{R} E[f(X)|\mathrm{do}(X_{s_I^R \cup i} = x_{s_I^R \cup i}) - E[f(X)|\mathrm{do}(X_{s_I^R} = x_{s_I^R})], \qquad (2.2)$$

$$\emptyset_i = \frac{1}{M!}\sum_{R} E[f(Xx)|x_{s_I^R \cup i}] - E[f(x)|(X_{s_I^R})], \qquad (2.3)$$

with $f(x) = \beta x + b.\beta$ is row of vector and b is a scalar. Then, assumed that the input data x considered normal distribution with mean μ and covariance $\sum$. Then, we got expected value as follow:

$$E[x|x_S] = [Q_{\bar{S}} - R_S]\mu + [Q_S + R_S]x, \qquad (2.4)$$

or

$$E[x|x_{S_i^R \cup i}] = [Q_{S_i^{\bar{R}} \cup i} - R_{S_i^R \cup i}]\mu + [Q_{S_i^R \cup i} + R_{S_i^{\bar{R}} \cup i}]x. \qquad (2.5)$$

Then, the final Shapley formula is defined as follow:

$$\emptyset_i = \beta[\frac{1}{M!}\sum_{R}([Q_{S_i^{\bar{R}} \cup i} - R_{S_i^R \cup i}][Q_{\bar{S}_i^R \cup i} + R_{S_i^R \cup i}])]\mu + \beta[\frac{1}{M!}\sum_{R}([Q_{S_i^R \cup i} - R_{S_i^R \cup i}][Q_{S_i^R} + R_{S_i^R}])]x \qquad (2.6)$$

2.4 Evaluation

To evaluate performance of the proposed solution, we conducted two types of evaluation which were visual analysis and statistical analysis. Visual analysis was conducted by analyzing the visualization result of SHAP in which it should consider

the relationship of features to class and relationship between features so that the explanation can be more rational. The statistical analysis was used to prove the significancy of the proposed solution by comparing results in scenario 1 and scenario 2 in each dataset. In this study we analyzed the comparison of correlation coefficient between original dataset, SHAP value in the original SHAP and SHAP value resulted from the combination of SHAP and correlation function. In this evaluation process, the correlation coefficient of original dataset played a role as a ground truth. Hence, we evaluated the result of original SHAP and SHAP+correlation by comparing its value to the ground truth. If the correlation coefficient value was quite similar to the ground truth, then we concluded that it was correct.

3. Experimental Results

Our experiment was divided into two scenarios performed in two dataset parallelly. Results in both scenarios and both datasets are described as follows.

3.1 Results in dataset 1

Dataset 1 is thyroid cancer dataset consisting of 165 instances and two classes distributed in 8 features. At the beginning step, we trained and tested the data using several classifier methods. Table 5 illustrates the comparison result.

Table 5. Comparison result of classification.

Classifier Method	Accuracy
Support Vector Machine	97.37%
Multilayer Perceptron	95.61%
Naïve Bayes	95.62%
Decision Tree	94.74%
KNN	93.86%

According to Table 5, SVM achieves the highest accuracy. Hence, we choose SVM result and took the testing data for explaining the result using original SHAP (as described in Fig.1. scenario 1). In general, SHAP resulted in two types of output which were visual explanation and SHAP value. Visual explanation showed the justification of each feature, while the SHAP value illustrated how each feature can impact the prediction. An example visual result can be seen in Fig. 2.

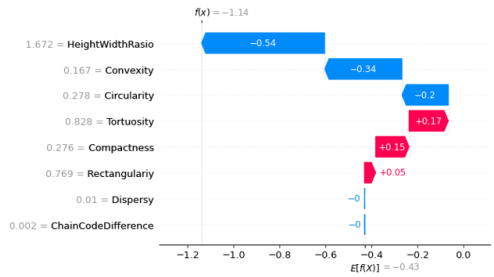

Fig. 2. Visual result of scenario 1 in dataset 1.

Fig. 2 shows an example of SHAP visualization result in the random testing set in dataset 1 using scenario 1. The ground truth of its data is smooth. In Fig. 2 we can see the list of features containing of eight features with each feature value. In the middle of image, we find two types of color which are blue and red. The blue one indicates negative class (in this case, the negative class is irregular class). Number in the middle of blue arrow indicates SHAP value of each feature. SHAP value is value that indicates the important level between each feature to the prediction result. Negative class always contains negative SHAP value. Negative here does not indicate the negative value but only for showing the class which means the negative symbol is used for indicating negative class (in this case the negative class is irregular class and positive class is smooth class). Conversely, the red color in Fig. 2 indicates the positive class or smooth class. Similar with the blue one, number in the middle of red arrow indicates SHAP value of each feature. Hence, we can

summarize that the higher the SHAP value in the blue arrow, the stronger the prediction indicates a negative class. Otherwise, the higher the SHAP value in the red arrow, the stronger the prediction indicates a positive class. The prediction itself is determined by summing up all feature values in both colors diminished by f(x) for normalizing the value. The resulted number is then called as base value or expected value symbolized as E[f(x)]. In Fig. 2, base value is 0.43 indicating the positive class or smooth class. According to visual results and base value, the explanation is relevant enough.

In the second scenario, we applied different SHAP by adding correlation in SHAP method. The visualization result in the random testing set in dataset 1 can be seen in Fig. 3.

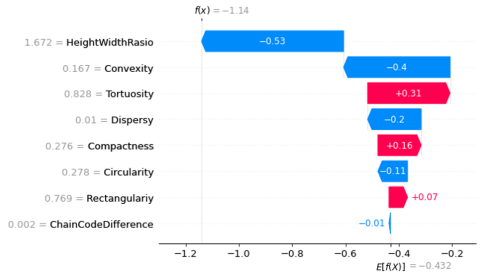

Fig. 3. Visual result of scenario 2 in dataset 1.

Visualization results in Figs. 2-3 seem different. In the Fig. 3, almost all features make contribution to the prediction result. Contrasted with Fig. 2, there are some features that do not make any contribution to the prediction result. However, the explanation of SHAP itself between Figs. 2-3 seems almost the same. It occurred since all features have similar characteristics which have strong correlation. Hence, if we added correlation, it would not have well impact.

In the next step, we proved our method in other datasets that consisted random features. Hence, we can evaluate the impact of adding correlation in the original SHAP. The result of this experiment is described in section below.

3.2 Results in dataset 2

In this experiment, we used different datasets that have more complex features compared with the first dataset. The second dataset had five features that were strongly different from each other. Hence, we tested our proposed solution to validate whether correlation can give well impact or not.

The five features in the second dataset were perimeter, radius, area, smoothness, and texture. Theoretically, the characteristics of those features were strongly different. For example, is that texture with radius. Texture was calculated by the histogram of data, while radius means distance between objects. Thus, both of them gave different characteristics in which correlation between those features may affect the prediction result. However, other features seem to have correlated with the others such as perimeter, area, and radius. Hence, we applied modified SHAP to prove it. Results of applying original SHAP and modified SHAP can be seen in Figs. 4-5.

Fig. 4 shows an example of SHAP visualization result in the random testing set in dataset 2 using scenario 1 (using original SHAP). The ground truth of its data is malignant. The blue one indicates negative class (in this case, the negative class is benign class). Similar with result in dataset 1, number in the middle of blue arrow indicates SHAP value of each feature.

SHAP value is value that indicates the important level between each feature to the prediction result. Negative class always contains negative SHAP value. Negative here does not indicate the negative value but only for showing the class which

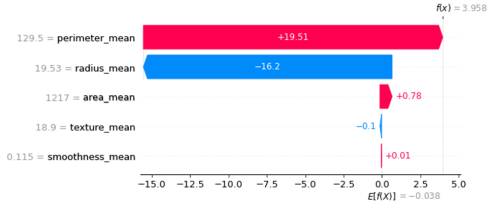

Fig. 4. Visual result of scenario 1 (original SHAP) in dataset 2.

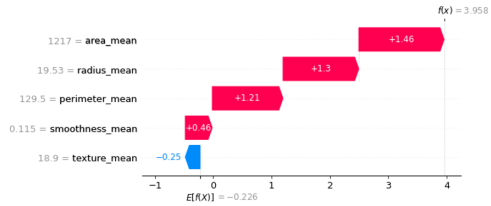

Fig. 5. Visual result of scenario 2 (modified SHAP) in dataset 2.

means the negative symbol is used for indicating negative class (in this case the negative class is benign class and positive class is malignant class). Conversely, the red color in Fig. 4 indicates the positive class or malignant class. According to Fig. 4 we can see that almost features indicate strongly malignant since almost SHAP values are positive. However, it seems any miss-interpretation in Fig. 4. Theoretically, perimeter and radius had strongly positive correlation in which value of each feature was strongly affected by each other. However, in Fig. 4 perimeter and radius had different interpretation to the prediction. Perimeter strongly correlated to malignant class, while radius strongly correlated to benign class. Hence, the explanation result seems not rational since those two features actually are strongly correlated, however in SHAP explanation those two features seem not correlated. It indicates that there is any error in SHAP explanation that should be handled.

In the experiment, we applied different SHAP by adding correlation in SHAP method. The visualization results in the random testing set in dataset 2 can be seen in Fig. 5.

Visualization result between Fig. 4 and Fig. 5 seems very different. Fig. 5 seems more rational because relation between feature appeared clearly. For example, in Fig. 5, area, radius and perimeter had strong relationship indicated by almost close SHAP value. To prove the result, we also conducted statistic test to validate whether correlation gave significant impact to the explanation or not. In this step, we calculate the coefficient correlation value between features of original dataset, SHAP value without correlation and SHAP value with correlation. Here is the result of each component.

Table 6. Coefficient correlation between features of original dataset.

	radius	texture	perimeter	area	smoothness
radius	1				
texture	0.324	1			
perimeter	0.998	0.330	1		
area	0.987	0.321	0.987	1	
smoothness	0.171	-0.023	0.207	0.177	1

Table 7. Coefficient correlation between features of SHAP Value resulted from original SHAP (without correlation).

	radius	texture	perimeter	area	smoothness
radius	1				
texture	-0.2748	1			
perimeter	-0.9974	0.2868	1		
area	-0.9905	0.2743	0.9869	1	
smoothness	-0.2573	0.0235	0.3025	0.2372	1

After we gained the coefficient correlation value of original dataset, SHAP value without correlation and SHAP value with correlation, we compared those results. In this step, we performed the coefficient cor-

Table 8. Coefficient correlation between feature of SHAP Value resulted from SHAP with correlation.

	radius	texture	perimeter	area	smoothness
radius	1				
texture	0.0841	1			
perimeter	0.9890	0.0940	1		
area	0.8671	0.0505	0.7848	1	
smoothness	0.2633	0.0330	0.3333	0.053	1

relation value of original dataset as the ground truth. Hence, if the coefficient correlation value of original SHAP or SHAP with correlation are closed to the coefficient correlation value of original dataset, it can be inferred that the explanation from original SHAP or SHAP with correlation is more rational. Here is the comparison between original dataset and original SHAP also between original dataset and SHAP with correlation.

Table 9 shows that the description of SHAP+ correlation was more similar to the original description that played a role as the ground truth compared to the original SHAP. Hence, we can conclude that after adding the correlation in the perturbation process, the explanation results were more rational. Our two scenarios in two different datasets had different results. According to the result we concluded that adding correlation in SHAP should consider the characteristics of each feature. If almost features had similar characteristics, adding correlation will not have significant impact. However, if we used correlation for more complex features, it will be very valuable.

3.3 Discussion

The resulted findings highlight the advantages of incorporating feature correlations into the SHAP method for explainable artificial intelligence (XAI). The proposed solution presented clear improvements in generating more rational explanations compared to the original SHAP, while its implementation also provided important insights into practical considerations and limitations.

One of the key findings is that explicitly accounting for correlations among features leads to explanations that are more logical, realistic, and consistent with the true structure of the data. Since features in real-world datasets rarely occur separately, ignoring their interdependence may result in misleading or oversimplified interpretations. In contrast, the correlation-adjusted approach enables a more reliable identification of the most influential features, thereby improving the interpretability and trustworthiness of model outputs. This has significant implications for critical applications, since transparency and accountability are essential.

Despite of the resulted finding, error can be appeared in certain cases. When correlations between features are weak or inconsistent, the method may provide an error analysis. Another consideration is computational efficiency. Embedding correlations analysis in SHAP architecture will increase complexity particularly in high-dimensional datasets, which risks limiting large-scale implementation. Hence, further research can focus on broader empirical testing, computational optimization, and the development of user-oriented visualization tools that enhance the accessibility of explainability.

In general, the proposed solution increases the rationality and reliability of SHAP-based explanations by employing correlation-based XAI methods, which potential to be implemented future research and practical adoption.

Table 9. Comparison of correlation coefficient result between original dataset, original SHAP and SHAP+correlation.

Correlation	Correlation coefficient value			Description		
	Original dataset	SHAP value		Original dataset	SHAP value	
		original method	SHAP+ correlation		original method	SHAP+ correlation
Texture x radius	0.324	-0.2748	0.0841	Low correlation (+)	Not correlated (-)	Not correlated (+)
Perimeter x radius	0.998	-0.9974	0.9890	Highly correlated (+)	Highly correlated (-)	Highly correlated (+)
Area x radius	0.987	-0.9905	0.8671	Highly correlated (+)	Highly correlated (-)	Highly correlated (+)
Smoothness x radius	0.171	-0.2573	0.2633	Low correlation (+)	Low correlation (-)	Low correlation (+)
Perimeter x texture	0.330	0.2868	0.0940	Low correlation (+)	Low correlation (+)	Not correlated (+)
Area x texture	0.321	0.2743	0.0505	Low correlation (+)	Low correlation (+)	Not correlated (+)
Smoothness x texture	-0.023	0.0235	0.0330	Not correlated (-)	Not correlated (+)	Not correlated (+)
Area x perimeter	0.987	0.9869	0.7848	Highly correlated (+)	Highly correlated (+)	Highly correlated (+)
Smoothness x perimeter	0.207	0.3025	0.3333	Low correlation (+)	Low correlation (+)	Low correlation (+)
Smoothness x area	0.177	0.2372	0.053	Low correlation (+)	Low correlation (+)	Not correlated (+)

4. Conclusion

This research work aimed to develop modified SHAP for explaining machine learning results in medical applications. The proposed solution was conducted by adding correlation in the perturbation process. The proposed solution was tested in two different datasets to prove and validate the impact of using correlation in perturbation of SHAP method. According to both the visual analysis and statistical test, we conclude that the proposed solution gave a more rational explanation compared to the original SHAP. In the other hand, we also concluded that our proposed solution was very useful for dataset that have complex features since if the dataset consisted of similar feature, adding correlation function would not give significant impact to our result. Since the proposed solution yielded more reasonable outcomes in the theoretical analysis, its implementation further emphasized the significance of incorporating feature correlations within the SHAP method. Correlation reflects the inherent interrelationships among features; thus, when explanatory models explicitly account for these dependencies, several advantages can be realized. First, the generated explanations reveal greater logical consistency and realism, as they more accurately capture the underlying data structure. Second, this approach provides a more precise identification of the most influential features, thereby improving the reliability of the interpretation and providing suitable support for data-driven decision-making.

References

[1] Calisto FM, Nunes N, Nascimento JC. Modeling adoption of intelligent agents in medical imaging. Int J Hum Comput Stud [Internet]. 2022;168:102922. Available from: https://www.sciencedirect.com/science/article/pii/S1071581922001422

[2] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med [Internet]. 2019;25(1):44–56. Available from: https://doi.org/10.1038/s41591-018-0300-7

[3] Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet [Internet]. 2018;50(8):1161–70. Available from: https://doi.org/10.1038/s41588-018-0167-z

[4] Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. J Med Internet Res. 2018 Sep;20(9):e11510.

[5] Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. npj Digit Med [Internet]. 2018;1(1):59. Available from: https://doi.org/10.1038/s41746-018-0065-x

[6] Nugroho HA, Frannita EL, Hutami AHT, Chondah L, Nugroho A, Fauzi RN, et al. Deep learning for analyzing thyroid nodule malignancy based on the composition characteristic of ultrasonography images. In: 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS). 2020. p. 77–82.

[7] Prananda AR, Nugroho HA, Ardiyanto I. Enumeration of Plasmodium parasites on thin blood smear digital microscopic images. In: 2019 5th International Conference on Science in Information Technology (ICSITech). 2019. p. 223–8.

[8] Nugroho HA, Frannita EL, Hutami AHT. Thyroid nodules stratification based on orientation characteristics using machine learning approach. In: 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE). 2020. p. 52–7.

[9] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med [Internet]. 2018;24(10):1559–67. Available from: https://doi.org/10.1038/s41591-018-0177-5

[10] Ahmad MA, Eckert C, McKelvey G, Zolfagar K, Zahid A, Teredesai A. Death versus data science: predicting end of life. Proc 30th Innov Appl Artif Intell Conf IAAI 2018. 2018;7719–26.

[11] Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. J Med Imaging Radiat Sci [Internet]. 2019;50(4):477–87. Available from: https://www.sciencedirect.com/science/article/pii/S1939865419305041

[12] Gleichgerrcht E, Munsell BC, Alhusaini S, Alvim MKM, Bargalló N, Bender B, et al. Artificial intelligence for classification of temporal lobe epilepsy with ROI-level MRI data: a worldwide ENIGMA-Epilepsy study. NeuroImage Clin [Internet]. 2021;31:102765. Available from: https://www.sciencedirect.com/science/article/pii/S2213158221002096

[13] Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW). 2018. p. 117–22.

[14] Peng WV, Mayorga RV, Hussein EMA. An automated confirmatory system for analysis of mammograms. Comput Methods Programs Biomed [Internet]. 2016;125:134–44. Available from:

https://www.sciencedirect.com/science/article/pii/S0169260715002461

[15] Koh J, Yoon Y, Kim S, Han K, Kim E-K. Deep learning for the detection of breast cancers on chest computed tomography. Clin Breast Cancer [Internet]. 2021; Available from: https://www.sciencedirect.com/science/article/pii/S1526820921001154

[16] Nurfauzi R, Nugroho HA, Ardiyanto I, Frannita EL. Autocorrection of lung boundary on 3D CT lung cancer images. J King Saud Univ - Comput Inf Sci [Internet]. 2021;33(5):518–27. Available from: https://www.sciencedirect.com/science/article/pii/S1319157818308632

[17] Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol [Internet]. 2018;29(8):1836–42. Available from: https://www.sciencedirect.com/science/article/pii/S0923753419341055

[18] Frannita EL, Nugroho HA, Nugroho A, Zulfanahri, Ardiyanto I. Thyroid nodule classification based on characteristic of margin using geometric and statistical features. In: 2018 2nd International Conference on Biomedical Engineering (IBIOMED). 2018. p. 54–9.

[19] Nugroho HA, Frannita EL. Thyroid cancer classification using transfer learning. In: 2021 International Conference on Computer Science and Engineering (IC2SE). 2021. p. 1–5.

[20] Ouahabi A, Taleb-Ahmed A. Deep learning for real-time semantic segmentation: application in ultrasound imaging. Pattern Recognit Lett [Internet]. 2021;144:27–34. Available from: https://www.sciencedirect.com/science/article/pii/S0167865521000234

[21] Ma J, Wu F, Jiang T, Zhao Q, Kong D. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int J Comput Assist Radiol Surg. 2017 Jul 1;12:1–16.

[22] Zhang S, Du H, Jin Z, Zhu Y, Zhang Y, Xie F, et al. A novel interpretable computer-aided diagnosis system of thyroid nodules on ultrasound based on clinical experience. IEEE Access. 2020;8:53223–31.

[23] Nugroho HA, Marsiano AFD, Xaphakdy K, Sihakhom P, Frannita EL, Nurfauzi R, et al. Multithresholding approach for segmenting Plasmodium parasites. In: 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE). 2019. p. 1–5.

[24] Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. Comput Med Imaging Graph [Internet]. 2019;74:61–71. Available from: https://www.sciencedirect.com/science/article/pii/S0895611118305573

[25] Deperlioglu O, Kose U, Gupta D, Khanna A, Giampaolo F, Fortino G. Explainable framework for glaucoma diagnosis by image processing and convolutional neural network synergy: Analysis with doctor evaluation. Futur Gener Comput Syst [Internet]. 2022;129:152–69. Available from: https://www.sciencedirect.com/science/article/pii/S0167739X21004556

[26] Vinícius dos Santos Ferreira M, Oseas de Carvalho Filho A, Dalília de Sousa A, Corrêa Silva A, Gattass M. Convolutional neural network and texture descriptor-based automatic detection and diagnosis of glaucoma. Expert Syst Appl [Internet]. 2018;110:250–63. Available from: https://www.sciencedirect.com/science/article/pii/S0957417418303567

[27] Mugasa H, Dua S, Koh JEW, Hagiwara Y, Lih OS, Madla C, et al. An adaptive feature extraction model for classification of thyroid lesions in ultrasound images. Pattern Recognit Lett [Internet]. 2020;131:463–73. Available from: https://www.sciencedirect.com/science/article/pii/S0167865520300490

[28] Chen J, You H, Li K. A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Comput Methods Programs Biomed [Internet]. 2020;185:105329. Available from: https://www.sciencedirect.com/science/article/pii/S0169260719308454

[29] Yang W, Dong Y, Du Q, Qiang Y, Wu K, Zhao J, et al. Integrate domain knowledge in training multi-task cascade deep learning model for benign–malignant thyroid nodule classification on ultrasound images. Eng Appl Artif Intell [Internet]. 2021;98:104064. Available from: http://www.sciencedirect.com/science/article/pii/S0952197620303304

[30] Tulsani A, Kumar P, Pathan S. Automated segmentation of optic disc and optic cup for glaucoma assessment using improved UNET++ architecture. Biocybern Biomed Eng [Internet]. 2021;41(2):819–32. Available from: https://www.sciencedirect.com/science/article/pii/S0208521621000656

[31] Khudhair Abbas S. Thyroid tissue segmentation and classification in ultrasound image of artificial intelligence. Mater Today Proc [Internet]. 2021; Available from: https://www.sciencedirect.com/science/article/pii/S221478532103279X

[32] Nugroho HA, Zulfanahri, Frannita E, Ardiyanto I, Choridah L. Computer aided diagnosis for thyroid cancer system based on internal and external characteristics. J King Saud Univ Comput Inf Sci. 2019 Jan 23;

[33] Kesarkar XA, Kulhalli KV. Thyroid nodule detection using artificial neural network. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). 2021. p. 11–5.

[34] Bayrak EA, Kırcı P, Ensari T. Comparison of machine learning methods for breast cancer diagnosis. In: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). 2019. p. 1–3.

[35] Bayrak EA, Kirci P. Fractal analysis of thyroid ultrasound image data evaluation. In: 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC). 2020. p. 1–4.

[36] Ribeiro MT, 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 [Internet]. New York (NY): Association for Computing Machinery; 2016. p. 1135–44.

[37] Seethi VDR, LaCasse Z, Chivte P, Bland J, Kadkol SS, Gaillard ER, et al. An explainable AI approach for diagnosis of COVID-19 using MALDI-ToF mass spectrometry. Expert Syst Appl [Internet]. 2024;236:121226. Available from: https://www.sciencedirect.com/science/article/pii/S0957417423017281

[38] Deepanshi, Budhiraja I, Garg D, Kumar N. Choquet integral based deep learning model for COVID-19 diagnosis using explainable AI for NG-IoT models. Comput Commun [Internet]. 2023;212:227–38. Available from: https://www.sciencedirect.com/science/article/pii/S0140366423003481

[39] Caroprese L, Vocaturo E, Zumpano E. Argumentation approaches for explainable AI in medical informatics. Intell Syst Appl [Internet]. 2022;16:200109. Available from:

https://www.sciencedirect.com/science/article/pii/S2667305322000473

[40] Di Martino F, Delmastro F, Dolciotti C. Explainable AI for malnutrition risk prediction from m-health and clinical data. Smart Health [Internet]. 2023;30:100429. Available from: https://www.sciencedirect.com/science/article/pii/S2352648323000570

[41] Niranjan K, Shankar Kumar S, Vedanth S, Chitrakala DS. An explainable AI driven decision support system for COVID-19 diagnosis using fused classification and segmentation. Procedia Comput Sci [Internet]. 2023;218:1915–25. Available from: https://www.sciencedirect.com/science/article/pii/S1877050923001680

[42] Cao Y, Kunaprayoon D, Ren L. Interpretable AI-assisted clinical decision making for dose prescription in radiosurgery of brain metastases. Radiother Oncol [Internet]. 2023;187:109842. Available from: https://www.sciencedirect.com/science/article/pii/S0167814023897365

[43] Kruk M. SHAP-NET, a network based on Shapley values as a new tool to improve the explainability of the XGBoost-SHAP model for the problem of water quality. Environ Model Softw [Internet]. 2025;188:106403. Available from: https://www.sciencedirect.com/science/article/pii/S1364815225000878

[44] Zhou H, Ma X, Guan H, Yang J, Wei B, Zhang Y, et al. Prediction of maize crude fat content based on improved conditional mutual information maximization and SHAP analysis. Food Chem [Internet]. 2025;493:146054. Available from: https://www.sciencedirect.com/science/article/pii/S0308814625033059

[45] Lee MJ, Choi SY. The impact of financial statement indicators on bank credit ratings: Insights from machine learning and SHAP techniques. Financ Res Lett [Internet]. 2025;85:107758. Available from: https://www.sciencedirect.com/science/article/pii/S1544612325010165

[46] Nourani V, Dehghan M, Baghanam AH, Kantoush SA. Dual purpose of Shapley additive explanation (SHAP) in model explanation and feature selection for artificial intelligence-based digital twin of wastewater treatment plant. J Water Process Eng [Internet]. 2025;75:107947. Available from: https://www.sciencedirect.com/science/article/pii/S2214714425010190

[47] Qurat Ul Ain S, Islam Rather KU. Integrated statistical modeling and machine learning techniques with SHAP for epidemiological data analysis. Ann Epidemiol [Internet]. 2025;108:85–91. Available from: https://www.sciencedirect.com/science/article/pii/S1047279725001334

[48] Zhao J, Wang F, Wen J, Dai H, Wu L, Guo Y, et al. Enhanced magnetite separation prediction by PLIMS multifactor coupling based on machine learning and SHAP interpretability analysis. Powder Technol [Internet]. 2026;467:121546. Available from: https://www.sciencedirect.com/science/article/pii/S0032591025009416

[49] UCI. Breast Cancer Wisconsin (Diagnostic) Data Set [Internet]. Diagnostic Wisconsin Breast Cancer Database. 1995. Available from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[50] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;2017-Decem(Section 2):4766–75.

[51] Molnar C. A guide for making black box models explainable [Internet]. Interpretable Machine Learning. 2021 [cited 2021 Mar 13]. Available from:

https://christophm.github.io/interpretable-ml-book/

[52] Lundberg S. Math behind LinearExplainer with correlation feature perturbation [Internet]. SHAP Documentation. 2018 [cited 2021 Jul 15]. Available from: https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/linear_models/Math%20behind%20LinearExplainer%20with%20correlation%20feature%20perturbation.html