Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2021.027385
Article

Tech Science Press

An Explanatory Strategy for Reducing the Risk of Privacy Leaks

Mingting Liu[1], Xiaozhang Liu[1,*], Anli Yan[1], Xiulai Li[1,2], Gengquan Xie[1] and Xin Tang[3]

[1]Hainan University, Haikou, 570228, China
[2]Hainan Hairui Zhong Chuang Technol Co., Ltd., Haikou, 570228, China
[3]School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore
[*]Corresponding Author: Xiaozhang Liu. Email: lxzh@hainanu.edu.cn

Received: 16 January 2022; Accepted: 25 February 2022

Abstract: As machine learning moves into high-risk and sensitive applications such as medical care, autonomous driving, and financial planning, how to interpret the predictions of the black-box model becomes the key to whether people can trust machine learning decisions. Interpretability relies on providing users with additional information or explanations to improve model transparency and help users understand model decisions. However, these information inevitably leads to the dataset or model into the risk of privacy leaks. We propose a strategy to reduce model privacy leakage for instance interpretability techniques. The following is the specific operation process. Firstly, the user inputs data into the model, and the model calculates the prediction confidence of the data provided by the user and gives the prediction results. Meanwhile, the model obtains the prediction confidence of the interpretation data set. Finally, the data with the smallest Euclidean distance between the confidence of the interpretation set and the prediction data as the explainable data. Experimental results show that The Euclidean distance between the confidence of interpretation data and the confidence of prediction data provided by this method is very small, which shows that the model's prediction of interpreted data is very similar to the model's prediction of user data. Finally, we demonstrate the accuracy of the explanatory data. We measure the matching degree between the real label and the predicted label of the interpreted data and the applicability to the network model. The results show that the interpretation method has high accuracy and wide applicability.

Keywords: Machine learning; model data privacy risks; machine learning explanatory strategies

1 Introduction

Machine learning has gradually been applied to face recognition systems, autonomous driving technology, medical analysis, criminal justice and other realistic sensitive tasks [1–3]. However, due to the inherent complexity of the black box model, there has always been a crisis of confidence. As a result, machine learning cannot be fully applied in real life, and its development is restricted by various aspects. Solving the problem of trust between artificial intelligence models and human beings has become a key factor in the development and extension of artificial intelligence. To solve this problem, one has to provide some additional information to explain the model so that one can understand the decisions made by the model [4–8].

Machine learning model complexity is correlated with accuracy. In general, the structure of the model is simple, and people can easily understand the model decision, but its poor fitting ability leads to low accuracy. The more complex the structure of the model, the more accurate the model. However, due to the complexity of the model, its interpretability becomes more difficult. In order to improve the interpretability of machine learning, people have carried out multi-angle and multi-level research on the interpretability of models. For example, model-based interpretation provides users with some additional information, such as model information, training set information, gradient information, and so on, which is related to model decision-making to a certain extent. Unfortunately, providing users with interpretable information poses a threat to data privacy security while improving model transparency. For machine learning, Martin et al. combined machine learning methods with advanced privacy protection mechanisms [9], and used additional protection mechanisms to protect privacy. In order to protect the data privacy security of machine learning interpretability, more consideration will be given to designing an interpretive approach that does not compromise the data privacy of the model.

Take the example-based explanations proposed by Koh Liang et al., for example, the method provides the data points that have the greatest influence on the decision as explanatory information [10], while the training data has privacy information, but the model should not disclose the model data privacy. This method is not related to the model structure, but intuitively discloses the information of the model data set, which makes the model have data privacy risk. Shokri et al. have shown that this interpretation can be easily used for data privacy attacks [11]. Model interpretability is the key to build a bridge of trust between AI and people. However, if these explanations lead to model privacy disclosure, the development of model interpretability will be limited and the development of artificial intelligence will also be hindered. However, the research on artificial intelligence can explain privacy security is in a preliminary exploration stage and has not yet formed a research system. How to balance the sensitive relationship between model transparency and privacy security is an urgent problem to be solved. To solve this problem, we propose a new interpretation strategy, which reduces the risk of privacy disclosure while providing explanatory information.

In this paper, inspired by the example-based explanations, we study a data-based explanation. Unlike the example-based explanations, we reduce the transparency of the model while reducing the privacy risk of the explainable strategy. Our interpretation strategy focuses on the user's own query data. The main contributions of this paper can be summarized as follows:

- We propose a new data-based interpretation method, which provides explanatory information for the model to help users understand model decisions and reduce data privacy disclosure at the same time.
- We evaluate the effectiveness and correctness of this strategy according to the similarity degree of model prediction confidence and the label matching degree between interpretation data and user data as evaluation criteria.
- We test a variety of classification network models to prove the effectiveness of this interpretation strategy and to verify the universality of the model.

2 Related Work

In recent years, the research on the interpretability of artificial intelligence has attracted extensive attention from academia and enterprises [7]. In view of the development of interpretable artificial intelligence, multi-angle and multi-level interpretation methods have emerged at present [12–15]. For example, rule-based interpretation [16,17], activation maximization interpretation technology [18–20], gradient-based interpretation method [21–24], agnostic model interpretation [25] and other interpretation methods keep emerging.

Koh et al. [10] assumed that some point in the training set was y, and obtained model Ω by using training set $A(y \in A)$ and model Ω' by using training set $A'(y \notin A')$, and compared the difference of prediction effect between model Ω and model Ω' on point x to judge the contribution degree of training point y to decision making of point x. Koh et al.'s [10] explanation method of black box model is the first to explain the model based on data, which provides a new research perspective for the explicable model. However, the research progress of this perspective has been slow in recent years. In addition, there are privacy risks associated with this approach, as confirmed by Shokri et al. [11]. They designed a membership inference attack and a data reconstruction attack against the explanation model. The attacker accessed the model through continuous query, continuously obtained the data of the training set, repeated query with the feedback explanation information, constructed the graph structure, and finally realized the data set reconstruction. Although example-based explanations provide a new research perspective for interpretable models, it inevitably brings model data privacy risks. Shokri et al. [11] first focused on the privacy risks of the interpretability of the black box model in their study. This potential threat limits the development of interpretability. The researchers' ignorance of the privacy risks of interpreting information for the model provided an advantage for our study.

At present, machine learning is gradually applied to many sensitive fields, and people need to rely on the interpretable choice trust model of machine learning because of the particularity of the scene. Although research on interpretable machine learning has blossomed, as far as we know, No one studying interpretability has focused on model privacy risks, and after Koh, very few people have made models interpretable from a data point of view.

In our research, we continued to use the perspective of example-based explanations to seek the explanatory method of the model from the data. In addition, we also consider the model data privacy risks on this basis. Through this study, we propose a new machine learning explanation approach for black-box models.

3 Methodology

Our study aims to ensure model transparency while providing explanatory information for the model to balance the sensitive relationship between model transparency and privacy risk. In order to achieve the research objectives of this study, this study was carried out in three stages (see Fig. 1). First, we use transfer learning to train ResNet152 network model to obtain the best performance model. Secondly, we design an explanation algorithm based on our proposed explanatory strategy, and apply the model obtained in the first part to the algorithm. User's data and explanatory data are input into the model, the model outputs decision information, and the algorithm filters interpretation information according to the output calculation results. Finally, in the evaluation phase, we evaluate the model performance as well as the confidence similarity, tag matching and fitness of the interpretation information provided based on the user interpretable algorithm.

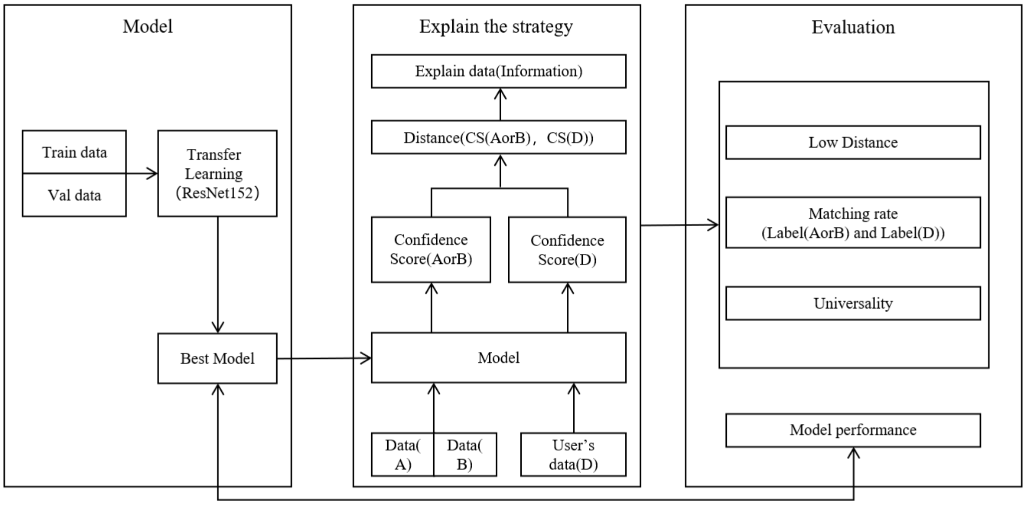


Figure 1: An overview of interpretable methods based on user data

3.1 Machine Learning Models

Example-based explanations return the data point with the largest contribution to the user training set by calculating the contribution degree of the training set to the model. This method is the first to provide explanation for machine learning models from a data set perspective, independent of the internal structure of the model. Our study provides interpretability for the model based on data sets. Before the research, a high-precision classification model was needed, so I chose to use Pytorch framework for transfer learning and training to compare the training of multiple models. In the model selection, we choose Resnet152 training model with better classification effect, in order to obtain high precision and low loss classification network. Due to the verification results required by the experiment, we choose the model training method with less cost.

Our goal is to obtain well-classified models. According to the current research results, we choose the ResNet model with the best effect for training. ResNet was proposed by Kaiming He et al. in order to reduce the training difficulty of deep neural network learning framework [26]. The complexity of neural network determines the accuracy of the model, but with the increase of network layers of deep neural network, network degradation will occur. In order to optimize the training of the network, the ResNet does not let the network directly fit the original mapping, but fits the residual mapping. More specifically, ResNet are the addition of fast connections to the forward network that allow raw data to skip layers. The proposed ResNet alleviates the training difficulty of the network. With the same network, the complexity and the number of parameters will be reduced when the residual network is added. The loss function is also called the objective function. The loss function we use is nn.CrossentRoPyLoss () in Pytorch as the loss function. Reasons for using this loss function This loss function is very effective when doing classification tasks. The loss function combines softmax-log-nllLoss. The specific process is as follows:

1) Softmax will constrain the value to the interval [0,1].

2) Take logarithms of the results after Softmax to ensure the monotonicity of the function.

3) NLLLoss corresponds the logarithmic result to the label and calculates the mean value by removing the sign.

3.2 Interpretable Algorithm Based on User Query Data

Unlike the example-based explanation, to reduce the data privacy risk of the model, the strategy proposed in this paper explanatory data from different sources. Due to different data sources, the search rules (influence functions [10]) for explaining information are no longer applicable. In the strategy proposed by us, the contribution degree is not sought, but the data that the search model makes similar decisions on user data are calculated.

3.2.1 Source of Explanatory Information

There are two parts to the strategy of explaining the source of information: 1) We need to set a default explanation dataset to search for explanation data when the user first queries and the amount of data is insufficient. 2) When the number of historical query data of the current user saved by the model is enough, the explanatory information is searched from the historical query data set of the user. In this paper, the rules for finding explanatory information in the two datasets are consistent.

3.2.2 Explanatory Information Generation Rules

Given two explanatory datasets A and B. In which, $A=\{x_0, x_1,\dots, x_t\}$ (t is constant) is the default explanation dataset; User history dataset B is an empty dataset under the initial conditions. We collectively refer to A and B as explanatory datasets, and explanatory data is also generated from A and B.

The user's current query dataset is $D=\{y_1, y_2,\dots,\mathrm{y}_m\}, m\in N$. The query model is Ω. After the explanation dataset A or B and user's dataset D are input into the model, the model outputs the confidence scores of classification decision and category decision (see formula (1)):

$$CScore_{AorB} = \begin{cases} \alpha_0, & 0 \\ \alpha_1, & 1 \\ \ldots, & \ldots \\ \alpha_n, & \text{n} \end{cases}, \quad （0,\ldots,n \text{ classes}） \tag{1}$$

where α_i means that in this decision, the model has the confidence of α_i to classify the target data into categories.

The essence of this method is to find the data point in the explanation dataset where the model decision confidence score is most similar to the user's current query data decision. Euclidean distance was chosen to calculate the similarity (see formula (2)):

$$Distance(CSore_{AorB}, CScore_D) = \sqrt{\sum_{i=0}^{n}(\partial_i - \beta_i)^2} \tag{2}$$

where x represents the list of category confidence scores obtained in the model for data decisions. This formula represents the Euclidean distance between the classification decision score of the interpreted data and the classification decision score of the user's input data, and we use the distance to indicate the degree of similarity between the two

Based on the interpretability method of user data, the target model is obtained through transfer learning and training, and the user queries the data into the model to obtain various confidence scores. Similarly, various confidence scores of datasets A or B are obtained, Euclidean distances of the two confidence scores are calculated, and explanation data is selected from A or B according to the minimum distance and returned to the user. The query data in this query is stored in dataset B for the next query to provide explanation information. The data pipeline is shown in Fig. 2.

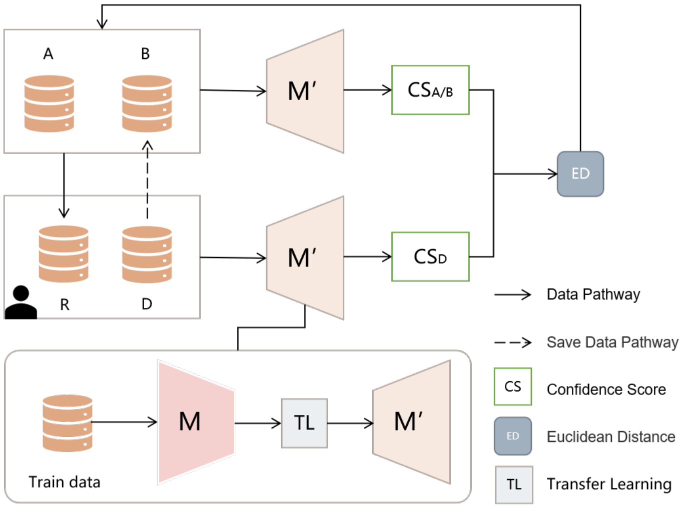


Figure 2: The data pipeline of the method

Algorithm 1: Based on user query data interpretability strategy

Initial conditions: Classification model Ω

Initial interpretation dataset A

Input: User dataset D

The user needs to explain the number of data: k

Output: Explanation information B

```
for $y_i$ in $D$ do
    $CScore_{yi} = \Omega(y_i)$
    if Num( $B$ ) < k
        $CScore_A = \Omega(A)$
        // Calculate the Euclidean distance
        $Sim = Distance(CScore_A, CScore_D)$
        Save $D$ into $B$
    else
        $CScore_B = \Omega(B)$
        $Sim = Distance(CScore_B, CScore_D)$
        Save $D$ into $B$
    $R = min_k(Sim)$
Return $R$
```

Here, Num(B) represents the number of data of each category in data set B . $CScore_y = \Omega(y)$ represents the confidence scores of each class after data y is input into the model. Sim represents the confidence score distance between the current forecast picture and each data point in the interpretation set, and min_k represents the k data point closest to the confidence score of the current forecast picture. After each image is predicted, k data points with the smallest Euclidean distance will be saved into interpretation information set R according to user requirements. We can see the explanation information in R . The query data of the same user will be stored in dataset B with the data and labels after each query.

In this paper, in order to more intuitively observe the explanatory information of our strategy selection, we choose the dichotomy model.

4 Experimental Design and Implementation

4.1 Experimental Dataset

In the selection of data sets, we tend to use the small data sets commonly used in classification models—cat and dog datasets and MNIST dataset. The dataset, which was made public on Kaggle in 2013, includes training sets and test sets. There are 25,000 pictures of cats and 12,500 pictures of dogs in the training set. Test set cat and dog mix 12,500 data. The MNIST dataset contains a total of 60,000 images and labels in the training set and 10,000 images and labels in the test set.

In this paper, in order to save time and cost, we select part of the data of the training set as the training set and test set of the transfer learning training model. We selected 2,806 images from the training set, including 1,403 for each dog and cat, 532 for the test set, 266 for dogs and cats. Set the test set of the original data set to the original data set for our experimental tests.

In order to compare the experimental results, we also train the handwritten digital classification model for MNIST data sets, and apply this model to our interpretation method. We selected about 1/5 of the data in the training set for transfer learning to train the model, and about 1/10 of the data in the test set as the default interpretation set.

4.2 Evaluation Indicators

4.2.1 Model Performance

In providing interpretability for machine learning models, one of the key issues that needs to be addressed is the dependence of model performance on interpretive performance. This is important because whether it's post-hoc explanation or ante-hoc explanation, users will see it as a system, and the explanatory we provide is essentially tied to the model, and users will make trust judgments about the model based on the explanatory information provided. That is, when we use our model interpretability, we first require good performance from the model. We help us select the classification model with the best classification effect according to the simplest accuracy and loss value of model performance evaluation as performance indicators. We test multiple classification networks through transfer learning and select the model with the best performance.

4.2.2 Confidence Score Similarity

As is known to all, the machine learning classification model makes decisions about the classification results of predicted data according to the confidence scores of each category predicted by the model. In other words, the model predicts the probability of the data being classified into each category. If we can find a piece of data that is very similar to how likely it is to be classified into categories, we can provide a sample explanation to the user. Therefore, in this study, we need to look for data points that are very similar to the forecast data. The confidence score similarity is obtained as the evaluation standard. The higher the confidence score similarity is, the higher the model similarity is to the decision process of interpreting data and predicting data.

4.2.3 The Matching Rate of Explanatory Data Label and Query Data Label

After obtaining the confidence score similarity between the explanatory data and the predicted data, we need to evaluate whether the explanatory information selected based on this similarity is deceptive. In order to prove the correctness of the explanatory information provided by this method, we calculate the consistency between the real labels of the explanatory data and the predicted labels of the predicted data. This data reflects whether the explanatory information provided by the model is consistent with the category of the predicted data. Specifically, if users predict dogs, the model will provide a picture of a dog as the explanatory data, which will increase users' trust in the model. Therefore, this data is used to evaluate interpretation methods.

4.2.4 Explain the Generality of the Method

This interpretation method is not directly related to the internal structure of the model, but closely related to the results of the model decision. We assume that this method can be applied to other classification models. This indicator is used to evaluate whether the interpretation method works well in other classification networks. In this experiment, the image dichotomy model is used as the experimental model, and we try to prove the universality of this method to other models, not to a particular model. We replace the machine learning model of other network structures to obtain the comparison of confidence score similarity and the change of tag matching rate to prove the universality of this method.

5 Experiments

5.1 Model Performance

In the existing classification model network, we use transfer learning to test the popular classification network. The specific performance is shown in Table 1. The optimal network training classification model ResNet152 is selected by accuracy and loss rate and is used in other aspects of this paper. Accuracy: 0.9850, Loss value: 0.0424. Through the test experiment of cat and dog datasets, we directly selected ResNet152 network model for training of MNIST dataset. The accuracy is 0.9896, and the loss value is 0.0408.

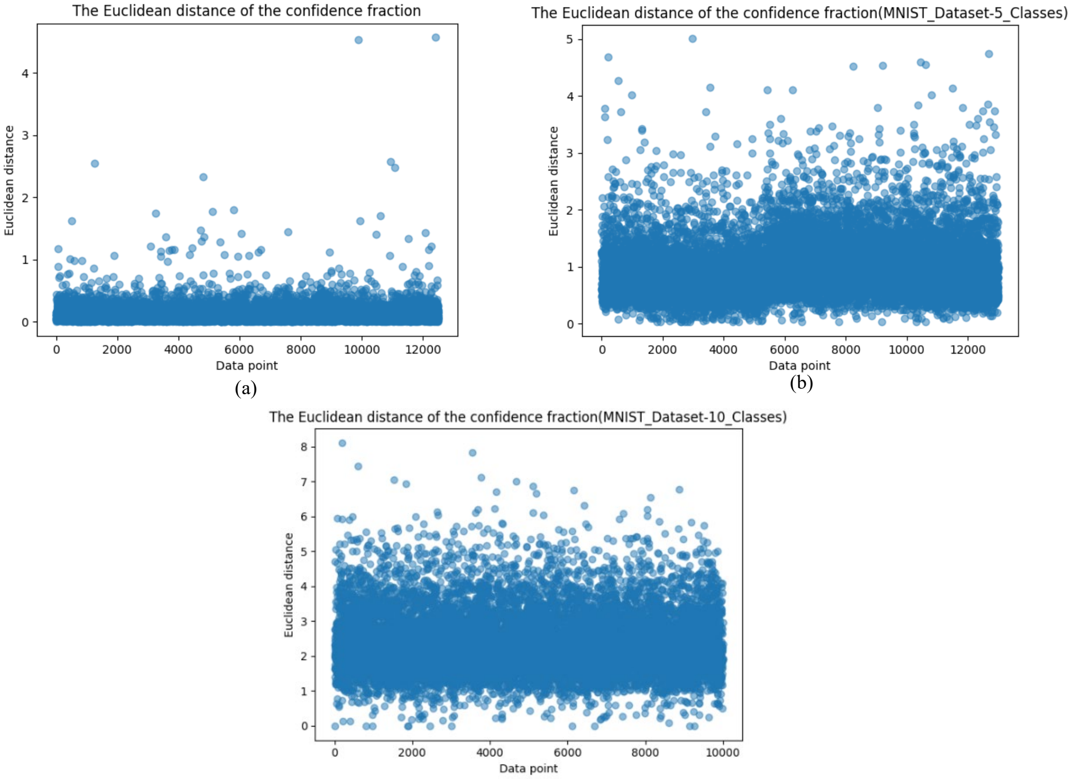
Table 1: Model performance evaluation

Dataset	Model	Accuracy	Loss
Cat and dog	AlexNet	0.9643	0.1045
	VGG19	0.9737	0.0439
	ResNet101	0.9831	0.0426

	DenseNet161	0.9793	0.0399
	ResNet152	0.9850	0.0424
	InceptionV3	0.9756	0.0941
MNIST	ResNet152	0.9896	0.0408

5.2 Confidence Score Similarity

The algorithm needs to calculate the Euclidean distance between the category confidence score of the explained data and the predicted confidence score of the predicted data according to the model, and the smallest distance is the highest similarity. According to the algorithm, we tested 12,500 data on cat and dog datasets and obtained the highest similarity between the confidence scores of the predicted data and the interpreted data. The results show (Fig. 3) that the Euclidean distance between the predicted confidence score of almost 99% of the interpreted data and the predicted confidence score of the query data is between [0,1]. However, the explanatory and predictive data of different models are not all very similar, which proves that the explanatory provided is related to the decision logic of the model itself. As a comparison, we tested 10,000 images of the five classification model and the ten classification model on the MNIST dataset, and the results showed that although the five classification distance was between [0,2] and the ten classification distance was between [0.5,5] with the increase of classification complexity, the distance gradually increased. But the results are still good.



(a)

(b)

(c)

Figure 3: (a) Bi-classification model of cat and dog classification (b) MNIST five-classification model and (c) MNIST ten-classification model. The Euclidean distance between prediction data and interpretation data classification confidence score

5.3 Label Matching Rate

Fig. 4 is an example of the explainable method proposed in this study. Among 12,500 test data, we calculated the matching degree between the interpretation data label with the highest similarity and the prediction label of the test data, as well as the matching degree between the real label and the prediction label of the K interpretation data with the highest similarity respectively (see Table 2). It is worth noting that despite multiple interpretation data, the match is still substantial. Moreover, since the calculation of confidence depends on the prediction of the model, factors leading to these differences include the error rate of the model itself.

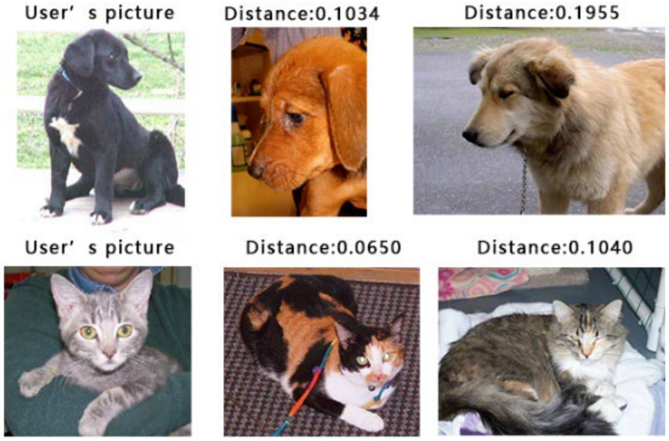


Figure 4: A comparison of the two examples. The left column is the picture currently predicted by the user, and the right two columns are the explanatory picture provided by this interpretation method and the Euclidean distance with the confidence score of the corresponding predicted picture

Table 2: Top k data matching tables

k	Label matching rate
1	96.62%
2	95.82%
3	93.31%
4	89.50%

5.4 Explain the Generality of the Method

We refer to different models for testing. Table 3 shows in detail the changes of the optimal similarity and tag matching rate obtained for different models. We test and calculate the mean, variance and standard deviation of the optimal similarity, and the results show that the similarity of the confidence scores of different models is very superior, and the difference of the model will make the similarity change very little. In addition, compared with Table 1, we can also observe that the influence of the superiority of the model itself on the tag matching degree is related.

To prove that our method is applicable to more models, we not only test different network models, but also experiment with classification complexity. After selecting the best ResNet152 network, we compare the binary model, the five-class model and the ten-class model respectively. It can be seen from Table 4 that despite the higher model complexity, our method can still achieve a high matching degree of more than 89%. However, with the increase of classification complexity, the Euclidean distance of its confidence score gradually increases, and the corresponding tag matching degree also gradually decreases, but the results are still in line with expectations.

Table 3: The optimal interpretation data similarity and label matching are obtained by different models

Model	The Euclidean distance of the confidence fraction			Label matching rate (k = 2)
	Mean	Var	Std	
Binary model	0.2490	0.6897	0.8305	89.00%
ResNet101	0.1289	0.0255	0.1597	95.18%
DenseNet161	0.1463	0.0615	0.2480	93.84%
VGG19	0.2275	0.4253	0.6522	94.42%
ResNet152	0.1573	0.0792	0.2814	95.82%
InceptionV3	0.1870	0.0407	0.2016	94.14%

Table 4: The optimal interpretation data similarity and label matching are obtained by different classification complexity models

Model (ResNet152)	The Euclidean distance of the confidence fraction			Label matching rate (k=2)
	Mean	Var	Std	
Binary classification model	0.1573	0.0792	0.2814	95.82%
Five classification model	0.9806	0.2691	0.5187	93.45%
Ten classification model	2.4477	0.9329	0.9658	89.22%

6 Conclusion

We propose an interpretable strategy that based on data that does not exposes model data privacy for the black box model. The method first trains a classification model with good performance, and then makes a predicted confidence score for the data according to the model. Finally, the explanation information is filtered according to the confidence score similarity as the rule of explanation information generation. We also conducted additional tests on the MNIST data set by training the five-classification model and the ten-classification model for comparative tests on the MNIST data. Experimental results show that this method has good performance. The mean, variance and standard deviation of the optimal similarity were 0.1573, 0.0792 and 0.2814, respectively, and the matching degree between the prediction data label and the optimal interpretation information label was 96.62%. In addition, we also test the confidence fraction similarity and tag matching degree of other network models. The optimal similarity of explanatory data provided by this method is excellent regardless of model performance. In addition, although classification complexity affects the matching degree of interpretation information to a certain extent, the effect is small. It is undeniable that tag matching degree depends on model performance and classification complexity. All in all, the explanatory information provided by this study for the black box model can not only help users understand the classification of the model, but more importantly, the explanatory information provided by this method does not involve model data privacy, thus avoiding the disclosure of model data privacy by explanatory information. In the future, we will optimize this method to reduce the impact of classification complexity on its results as much as possible. In addition, we will pay more attention to the defects of preference in the research of machine learning interpretable methods to achieve a more comprehensive interpretable information extraction method.

Funding Statement: This work is supported by the National Natural Science Foundation of China (Grant No. 61966011), Hainan University Education and Teaching Reform Research Project (Grant No. HDJWJG01), Key Research and Development Program of Hainan Province (Grant No. ZDYF2020033), Young Talents' Science and Technology Innovation Project of Hainan Association for Science and Technology (Grant No. QCXM202007), Hainan Provincial Natural Science Foundation of China (Grant No. 621RC612), Hainan Provincial Natural Science Foundation of China (Grant No. 2019RC107).

Conflicts of Interest: We declare that we have no conflicts of interest to report regarding the present study.

References

[1] T. Ching, D. S. Himmelstein and B. K. Beaulieu-Jones, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, 2018.

[2] R. Elshawi, M. H. Al-Mallah and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–32, 2019.

[3] C. Xiao, E. Choi and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[4] A. Bussone, S. Stumpf and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *2015 Int. Conf. on Healthcare Informatics*, IEEE, pp. 160–169, 2015.

[5] A. Datta, S. Sen and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE Sym. on Security and Privacy (SP)*, IEEE, pp. 598–617, 2016.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[7] C. Xiao, E. Choi and J. Sun., "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[8] S. M. Lundberg, G. G. Erion and S. I. Lee, "Consistent individualized feature attribution for tree ensembles," arXiv preprint arXiv:1802.03888, 2018.

[9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov *et al.*, "Deep learning with differential privacy," in *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security*, pp. 308–318, 2016.

[10] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Int. Conf. on Machine Learning*, PMLR, pp. 1885–1894, 2017.

[11] R. Shokri, M. Strobel and Y. Zick, "On the privacy risks of model explanations," in *Proc. of the 2021 AAAI/ACM Conf. on AI, Ethics, and Society*, pp. 231–241, 2021.

[12] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems*, pp. 1–18, 2018.

[13] A. Holzinger, C. Biemann, C. S. Pattichis and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" arXiv preprint arXiv:1712.09923, 2017.

[14] Q. S. Zhang and S. C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.

[15] S. M. Lundberg, G. G. Erion and S. I. Lee, "Consistent individualized feature attribution for tree ensembles," arXiv preprint arXiv:1802.03888, 2018.

[16] O. Bastani, C. Kim and H. Bastani, "Interpreting blackbox models via model extraction," arXiv preprint arXiv:1705.08504, 2017.

[17] A. Bondarenko, T. Zmanovska and A. Borisov, "Decompositional rules extraction methods from neural networks," in *Proc. of the 16th Int. Conf. on Soft Computing MENDEL'10*, Brno, Czech Republic, pp. 256–262, 2010.

[18] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A.Torralba, "Learning deep features for discriminative localization," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

[19] H. Wang, Z. Wang, M. Du, F. Yang, Z. J. Zhang *et al.*, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 618–626, 2017.

[21] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional neural networks,” in *European Conf. on Computer Vision*, Springer International Publishing, pp. 818–833, 2013.

[22] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, “Striving for simplicity: The all convolutional net,” arXiv preprint arXiv:1412.6806, 2014.

[23] M. Sundararajan, A. Taly and Q. Yan, “Gradients of counterfactuals,” arXiv preprint arXiv:1611.02639, 2016.

[24] D. Smilkov, N. Thorat, B. Kim, F. Viégas and M. Wattenberg, “Smoothgrad: Removing noise by adding noise,” arXiv preprint arXiv:1706.03825, 2017.

[25] W. Guo, D. Mu, J. Xu, J. Xu, P. Su *et al.*, “Lemna: Explaining deep learning based security applications,” in *Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security*, pp. 364–379,2018.

[26] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.