IB-M: A Flexible Framework to Align an Interpretable Model and a Black-box Model

Mengze Li
Zhejiang University
3150104805@zju.edu.cn

Kun Kuang*
Zhejiang University
kunkuang@zju.edu.cn

Qiang Zhu
Zhejiang University
zhuq@zju.edu.cn

Xiaohong Chen
Beijing Tong Ren Hospital
Capital Medical University
trchxh@163.com

Qing Guo
Beijing Tong Ren Hospital
Capital Medical University
guoqingxt@163.com

Fei Wu
Zhejiang University
wufei@zju.edu.cn

***Abstract*—Both interpretation and accuracy are very important for a predictive model in real applications, but most of previous works, no matter interpretable models or black-box models, cannot simultaneously achieve both of them, resulting in a trade-off between model interpretation and model accuracy. To break this trade-off, in this paper, we propose a flexible framework, named IB-M, to align an Interpretable model and a Black-box Model for simultaneously optimizing model interpretation and model accuracy. Generally, we think most of samples that are well-clustered or away from the true decision boundary can be easily interpreted by an interpretable model. Removing those samples can help to learn a more accurate black-box model by focusing on the left samples around the true decision boundary. Inspired by this, we propose a data re-weighting based framework to align an interpretable model and a black-box model, letting them focus on the samples what they are good at, hence, achieving both interpretation and accuracy. We implement our IB-M framework for a real medical problem of ultrasound thyroid nodule diagnosis. Extensive experiments demonstrate that our proposed framework and algorithm can achieve a more interpretable and more accurate diagnosis than a single interpretable model and a single black-box model.**

***Index Terms*—Interpretable model, Black-box model, Thyroid nodules**

I. Introduction

Owing to the big data and computing power, many machine learning methods, especially deep learning methods, have been proposed and shown to be successful in many real applications. For example in medical fields, [1], [2] based on the neural network can significantly improve the performance and even surpass the ability of corresponding experts. However, the increase in model performance always comes as a cost of increasing model complexity and opacity. As a result, most of those "great performance" models are used in a black-box way, resulting in a big gap between model knowledge and human understanding. The lack of interpretability of deep learning algorithms limits their applications in real scenarios, especially those requiring human understanding. Hence, it is highly demanding to develop a kind of statistical model with great performance on both accuracy and interpretation.

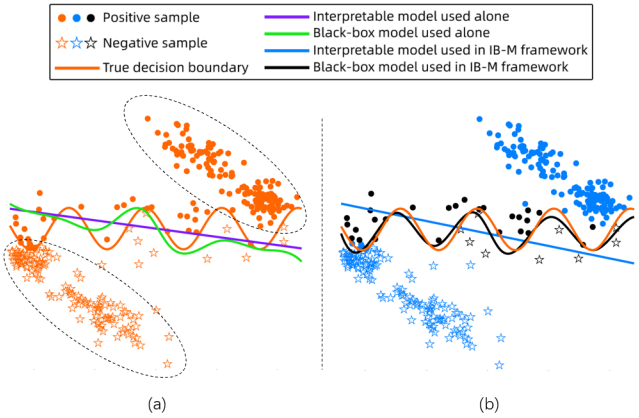

Fig. 1: **Comparison between a single model and an aligned model**. (a) Performance of a single interpretable (linear) model or a single black-box (non-linear) model, where the interpretable model cannot explain the samples (outside the dashed circle) that around the decision boundary and the black-box model would be affected by the samples (in the dashed circle) that well-clustered or away from the decision boundary; (b) Performance of an aligned model between an interpretable model and a black-box model, where by data separation/re-weighting, the interpretable model and the black-box model mainly focus on the samples that they are good at, respectively. The interpretable model focuses on blue samples, while the black-box model focuses on the black samples.

Recently, many interpretable models have been proposed to bring an understanding of black-box models. These methods can be roughly categorized into two branches: (i) "posthoc" interpretable models [3]–[5], and (ii) "intrinsic" interpretable models [6]–[8]. The "posthoc" interpretable models focus on

* Corresponding author

The authors would like to thank Prof. Tong Wang for her helpful comments and discussions. This work was supported in part by National Natural Science Foundation of China (No. 62006207 and No. 62037001), National Key Research and Development Program of China (No. 2018AAA0101900 and No. 2020YFC0832500), the Fundamental Research Funds for the Central Universities.

increasing human understanding of models by interpreting the predictive results of models. For example, [3] proposed to highlight the important features via heatmap of gradients for interpreting the predictive result of each sample; [5] explained the predictions of any classifier by learning a linear but interpretable model locally around the prediction. While the "intrinsic" interpretable models attempt to explain the inference process/logic of models to achieve human understanding and trust. For example, [6] designed a ProtoPnet based on human inference logic and visualized the prototypes of each class to interpret the inference logic of ProtoPnet; [7] proposed to replace the black-box model by a simple (interpretable) model on the data-space where the simpler model can be almost as accurate as the black-box model. Most of these interpretable methods, both "posthoc" and "intrinsic", can achieve a good interpretation of deep models in real applications, but their performance on predictive accuracy would be always worse than the black-box model. The trade-off between model accuracy and interpretation puts the practitioners in a dilemma of choosing between high accuracy and model interpretability.

Is a trade-off necessary between model accuracy and interpretability? No, it is not always necessary. In this paper, we focus on how to break that dilemma by aligning the interpretable model and the black-box model for simultaneously optimizing the accuracy and interpretability. The main idea is demonstrated in Fig. 1. Fig. 1(a) demonstrates a setting where the true decision boundary (orange line) is very complex but most of the samples can be easily classified by a interpretable classifier (e.g., linear model). If one only uses a linear model for training, the majority of the samples (data point in the dashed circle) that well-clustered or away from the true decision boundary can be interpreted by the learned linear model, but it would sacrifice the accuracy on the samples (data point outside the dashed circle) that around the true decision boundary. On the other hand, if one employs a black-box model (e.g., non-linear classifier) for training, the accuracy might be guaranteed, but lacking the interpretability. Moreover, the learned decision boundary of the black-box (green line) would be affected by those samples (data point in the dashed circle) that are well-clustered and near to the true decision boundary, leading the loss of model accuracy. To address this problem, cooperation between an interpretable model and a black-box model is needed as shown in Fig. 1(b), where the interpretable model only focuses on the samples that are well-clustered or away from the true decision boundary for optimizing model interpretation, then the black-box model concentrates on the left samples that are around the decision boundary but hardly explained by the interpretable model for optimizing model accuracy. Hence, our key idea to align an interpretable model and a black-box model is to let each model perform its own functions on the samples it is good at.

To simultaneously optimize the accuracy and interpretability, we propose a flexible framework, named IB-M, to align an Interpretable model and a Black-box Model. Our IB-M framework consists of three modules: an interpretable model, a black-box model, and a data re-weighting and model selection module. The interpretable model and black-box model in our framework can be *any* model by the practitioners. Given the interpretable model and the black-box model, the data re-weighting and model selection module is to re-weight/separate the whole data into two subsets, one set for training the interpretable model and the other for training the black-box model. The criteria for data re-weighting is mainly based on whether the prediction of the interpretable model on a sample can be as accurate as the black-box model. With the data re-weighting and model selection module, our IB-M framework can let both interpretable and black-box models perform their own functions on the samples they are good at, respectively, hence achieve both interpretation and accuracy.

In this paper, we apply our IB-M framework into the real medical problem of ultrasound thyroid nodule diagnosis. Thyroid nodule disease has become one of the hot research questions in the medical field with its extremely high prevalence [10]. A number of researchers have attempted to use deep learning to classify nodules based on their main diagnostic basis, ultrasound images, but the interpretation and accuracy of existing studies are not satisfactory. We design a specific model to solve Thyroid Nodule Classification using the IB-M framework, named TNC-IB-M. Considering medical interpretability, we combine medical diagnosis logic of thyroid nodules with the ProtoPnet model [6] to design an interpretable model suitable for the problem. Cooperating it with a powerful black-box model under the framework of IB-M, the TNC-IB-M obtains a very high accuracy rate.

The main contributions of this paper can be summarized as follows:

- We study the symbiotic relationship between model accuracy and interpretability, while previous methods make a trade-off between them.
- We propose a flexible IB-M framework to align an interpretable model and a black-box model for simultaneously optimizing model accuracy and interpretability.
- We apply our IB-M framework in the practical medical problem, and propose a TNC-IB-M algorithm for ultrasound thyroid nodule diagnosis.
- Extensive experiments demonstrate that our framework and algorithm can achieve better performance on both accuracy and interpretability than a single interpretable model or a single black-box model.

II. Related Work

A. The Interpretability of Deep Learning

In recent years, researchers have attempted to achieve the interpretability of deep learning models in a variety of ways, and have made some breakthroughs. Some researchers have tried to use the "posthoc" approach to interpret the trained deep learning model [3]–[5], [11], [12]. [3] used salient visualization to present the model's attention to different image regions as a heat map. [11] used an inverse convolution approach to explain the intermediate layers of a neural network, further clarifying the semantic information of each layer. [12] used

the method of maximum activation to find the important high-level features of the model. The "posthoc" method can enhance the interpretability of the model to some extent, but it cannot explain the real inference process of the model.

Recognizing the shortcoming of the "posthoc" approach, some researchers have attempted to improve the interpretability of the neural network by the "intrinsic" approach [6]–[8], [13], which is to design the model, part of which is interpretable. [6] simulated the way that the human analyzes things and proposed a prototype matching approach. [13] further developed this approach by combining it with segmentation to visualize the model's division of semantic regions. These methods make the inference process of the model reality transparent and are easier to gain the trust of users, but it comes at the expense of model accuracy. In order to improve the accuracy of the interpretable model, some researchers try to use the joint classification framework of the interpretable model and the black-box model. [7] used the interpretable model to approach the black-box model's classification plane, so that the black-box model can be partially replaced. [14] extended this idea by using multiple linear classifiers competing to achieve multi-class classification. Although such methods can improve the accuracy of the model to some extent, their upper limit of accuracy is limited by the black-box model. Thus, the trade-off between the interpretability and accuracy still exists. In this paper, we break this trade-off by designing a new type of joint classification model framework, which can achieve a more reasonable division of labor and cooperation between the interpretable model and the black-box model.

B. Application of Deep Learning in Ultrasound Diagnosis of Thyroid Nodules

In recent years some researchers have attempted to apply deep learning to the ultrasound diagnosis of thyroid nodules. [15] applied the VGG-16 model to this problem and initially achieved the classification of thyroid nodules. [16] employed a deeper convolution neural network to further improve the accuracy of thyroid nodule classification by relying on the model's stronger feature extraction capability. Such methods that rely on model depth to improve classification accuracy suffer from accuracy bottlenecks and poor medical interpretability. [17] used a basic multi-task model to solve the classification problem of thyroid nodules. The accuracy and interpretability were improved compared with previous algorithms. [18] improved the interpretability and classification accuracy by designing a multi-task model using multi-semantic attention. While these studies have improved the interpretability of the models to some extent, the inference process is still not transparent enough to show, which prevents these models from gaining sufficient trust.

III. IB-M Framework and Application

In this section, we introduce how our IB-M framework aligns an interpretable model and a black-box model for simultaneously optimizing both accuracy and interpretation. Then, we introduce a specific implementation of each part of the framework on a real medical application of the thyroid nodule classification task.

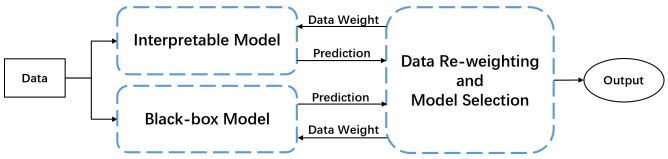

Fig. 2: **IB-M framework.** It consists of three main components: an interpretable model, a black-box model, and a data re-weighting and model selection module. Given any interpretable model and black-box model, the data re-weighting and model selection module learns two weights of samples by analyzing the interpretable model output, one weight for highlighting the samples that the interpretable model is good at for interpretation, and the other for the black-box model for accuracy.

A. IB-M Framework

To break the dilemma between model accuracy and interpretation in traditional methods, in this paper, we propose an IB-M framework to align an interpretable model and a black-box model for simultaneously optimizing the accuracy and interpretability. Fig. 2 shows our IB-M framework, which consists of the following three parts:

- An Interpretable Model (IM). The IM is designed for interpreting a part of samples (data points) that can be accurately predicted by IM, ensuring the interpretation and accuracy on that part of samples. The IM can be *any* interpretable model, such as linear regression model, decision tree, and ProtoPNet, as long as it is thought to be interpretable by the practitioners.
- A Black-box Model (BM). The BM is designed for collaborating with the IM to ensure the predictive accuracy of those samples that IM cannot guarantee its interpretation and accuracy. The BM can also be *any* black-box model, such as CNN and RNN, as long as it is thought to be accurate by the practitioners.
- The Data Re-weighting and Model Selection module (DR&MS). It is assigned to coordinate the division of labor between the predefined interpretable model and the black-box model by sample (data points) re-weighting during the model training process. Considering the prediction from IM, we measure the uncertainty of each sample about whether the prediction of IM can be as accurate as BM. By sample weighting with uncertainty, we let IM focus on the samples with high certainty for ensuring their interpretation and accuracy, and let BM focus on other samples that cannot be well-interpreted by IM for ensuring the accuracy. During the model inference process, DR&MS is designed for model selection to select IM or BM for final prediction based on the uncertainty of each sample during the inference process.

Among three modules, DR&MS is the core component to align the interpretable model and black-box model in our IB-M framework. It's responsible for calculating the weight of samples during the training phase, and selecting models during the inference phase.

In the training phase, since the classification plane of well-trained IM is far from the samples, which is well-clustered or away from the true decision plane, IM prediction accuracy of these samples is high. DR&MS generates the uncertainty u of samples using the distance d between samples and the classification plane of IM:

$$u = log(\frac{1+d}{d+\epsilon}) \tag{1}$$

Setting the threshold δ and comparing u with it, the weight of sample i to train IM and BM is calculated out:

$$w_{IM}^i = \begin{cases} r_{IM}(u < \delta) \\ 1(u \geqslant \delta) \end{cases} \tag{2}$$

$$w_{BM}^i = \begin{cases} r_{BM}(u \geqslant \delta) \\ 1(u < \delta) \end{cases} \tag{3}$$

where $r_{IM} > 1$ and $r_{BM} > 1$.

Then, with sample weights w_{IM}, the loss function for retraining IM can be represented as:

$$L_{IM}^{re-weight} = \sum_{i=1}^{n} w_{IM}^i L_{IM}^i, \tag{4}$$

where L_{IM}^i refers to the loss on sample i predicted by IM. If IM is a linear model, the $L_{IM}^i = (y_i - x_i\beta)^2$, where x_i and y_i are the features and outcome of sample i, β is linear regression coefficient.

Similarity, the loss function for retaining BM with sample weight w_{BM} can be represented as:

$$L_{BM}^{re-weight} = \sum_{i=1}^{n} w_{BM}^i L_{BM}^i, \tag{5}$$

where L_{BM}^i is the loss of sample i calculated by BM.

Combining Eq.(4) and Eq.(5), the total loss function L_{IB-M} during the training phase can be obtained:

$$L_{IB-M} = L_{IM}^{re-weight} + L_{BM}^{re-weight} \tag{6}$$

During the inference phase, if $w_{IM}^i > w_{BM}^i$, DR&MS outputs the prediction of IM for the sample i; otherwise, the prediction of BM is output.

B. Specific Realization of Thyroid Nodules Classification Task

In the process of diagnosing benign and malignant thyroid nodules, doctors focus on observing the properties of nodules such as echogenic foci, composition, margin, and echogenicity as the key basis for diagnosis. At present, many studies apply the convolution neural network model on the ultrasound image dataset of thyroid nodules and claim that they reach a high accuracy [15]–[18], but most of the studies do not use thyroid attribute information. Based on the medical diagnosis logic, we propose a specific implementation for IB-M to solve Thyroid Nodule Classification, named TNC-IB-M. In order to make full use of attribute information, we adopt the multi-tasking approach. In detail, we regard the classification of benign and malignant thyroid nodules as the main task and attribute classification as the subtask. In addition, considering that both the interpretable model and the black-box model need to include feature extractors, in order to reduce the computational complexity, we use a shared CNN instead of two feature extractors to process the input images.

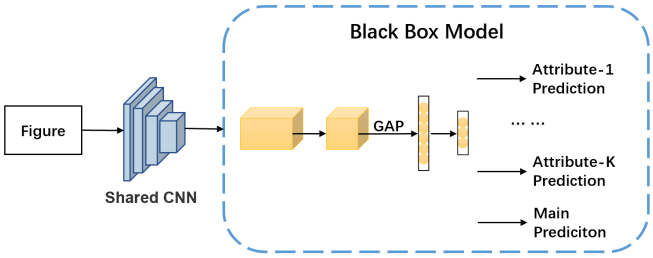

Fig. 3: **Black-box model structure.** The black-box model classifies the image features, which is extracted by the shared CNN, through multi-layer stacked convolution layers and fully connected layers.

Black-box Model in TNC-IB-M. We design a black-box model named TNC-IB, whose structure is shown in Fig. 3. Suppose there are 1 main task label and K attribute labels in the image classification task. A set of feature maps F is extracted by shared CNN from the image. After the feature maps F are input into the TNC-BM, a highly nonlinear model composed of multi convolution layers CNN and multi fully connected layers FC completes the classification task, then generates multi-task classification prediction $P_{TNC-BM} = \{p_b^0, p_b^1, \ldots, p_b^K\}$:

$$V_{TNC-BM} = GAP(CNN(F)) \tag{7}$$

$$P_{TNC-BM} = softmax(FC(V_{TNC-BM})) \tag{8}$$

The calculation formula of the loss function L_{TNC-BM} is:

$$\begin{aligned} L_{TNC-BM} = &- \sum_{c=1}^{C_0} log(p_b^0(c)) q^0(c) \\ &- \alpha * \sum_{k=1}^{K} \sum_{c=1}^{C_k} log(p_b^k(c)) q^k(c) \end{aligned} \tag{9}$$

α is the weight of the loss function of the attribute classification task. C_k is the number of task k categories. For the task k, if the target is c, $q^k(c) = 1$; otherwise, $q^k(c) = 0$.

Interpretable Model in TNC-IB-M. We combine multi-tasking and ProtoPnet model [6] to specifically realize module IM and generate a new model named TNC-IM. The specific

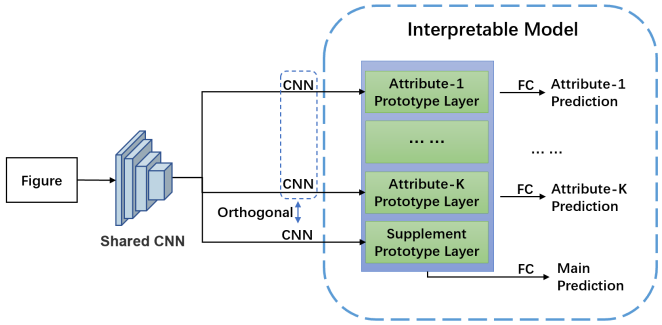

Fig. 4: **Interpretable model structure.** Shared CNN and a set of independent CNNs extract feature maps from the image. Then, the interpretable model is responsible for analyzing feature maps using built-in prototype layers, and generating multi-task prediction using the fully connected layers FC.

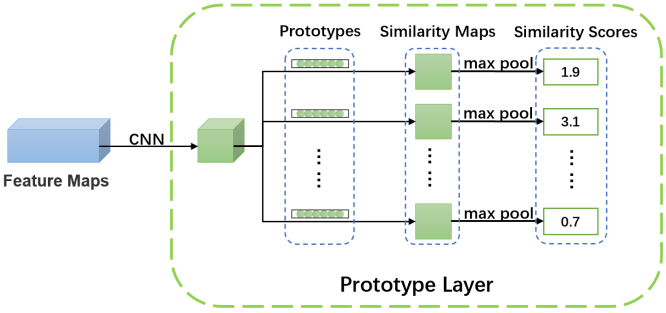

Fig. 5: **Prototype layer model structure.** The feature maps extracted by CNN are input into the prototype layer. The prototype layer uses built-in prototypes to match the feature maps, then generates similarity maps and similarity scores.

structure of TNC-IM is shown in Fig. 4. Before inputting a set of feature maps F extracted by the shared CNN to TNC-IM, K+1 independent CNNs are used to extract the feature maps required for each classification task from F. Subsequently, the feature maps are analyzed in parallel by the internal multi prototype layers of TNC-IM, and transformed into interpretable feature vectors V_{TNC-IM}. After dealt by the fully connected layers FC, the predictions of 1 main task and K attribute classification tasks $P_{TNC-IM} = \{p_i^0, p_i^1, \ldots, p_i^K\}$ are generated. The process can be formally described as:

$$V_{TNC-IM} = Prototype\ Layer(CNN(F)) \tag{10}$$

$$P_{TNC-IM} = softmax(FC(V_{TNC-IM})) \tag{11}$$

The loss of task k to penalize misclassification is:

$$L_{TNC-IM}^{CrsEnt-k} = -\sum_{c=1}^{C_k} log(p_i^k(c))q^k(c) \tag{12}$$

Considering the independence between attributes and the master-slave relationship between the main task and the attribute classification tasks, each attribute classification task uses their own prototype layer independently, named attribute-k prototype layer, to analyze the feature maps of each attribute. The classification of the main task relies on the comprehensive analysis of attribute characteristics generated by attribute prototype layers. Since the known attributes often cannot fully describe the main task, we add a supplement prototype layer to analyze feature maps that contain supplementary information. To ensure the orthogonality of information, the one-way orthogonal constraint is added between the CNN convolution kernel parameter W_s used to extract supplement information and the CNN convolution kernel parameter $W_{Attribute-k}$ used to extract attribute information:

$$L_{TNC-IM}^{Ortho} = \sum(|W_s W_{Attribute-k}^T - I|) \tag{13}$$

I represents the identity matrix. The loss function term can adjust the W_s parameter during the back propagation process, but cannot modify the $W_{Attribute-k}$ parameter.

The prototype layer model structure refers to the ProtoPnet model structure [6], as shown in Fig. 5. The feature map F_P generated by CNN is matched with the built-in prototypes Pr:

$$F_D = match(F_P, Pr) \tag{14}$$

For the "match" operation, F_P is divided into many patches of size 1*1. Calculating L^2 distance between the patches and Pr, we get the distance maps F_D. After that, we invert F_D into the similarity maps F_A:

$$F_A = log(\frac{1 + F_D}{\epsilon + F_D}) \tag{15}$$

After the max pooling operation, the global maximum similarity scores s between the input image and the built-in prototypes are obtained:

$$s = max\ pool(F_A) \tag{16}$$

The attribute-k prototype layer is responsible for analyzing and transforming features that contain both the attribute k information and the main task information. Each of its built-in prototypes contains information, which is about a certain category of attribute k and a certain category of the main task. Thus, the number of prototypes $\#Pr_k$ should be determined by the number of attribute k categories C_k and the number of main task categories C_0:

$$\#Pr_k = C_k * C_0 \tag{17}$$

The supplement prototype layer is only responsible for analyzing and transforming features containing the main task information. Each built-in prototype of it contains information which is about a certain category of the main task. Thus, the number of prototypes $\#Pr_s$ is determined by the number of main task categories:

$$\#Pr_s = C_0 \tag{18}$$

To train the prototypes, for task k, we choose one whose similarity score is the largest, from the prototypes Pr_c which contains the information of target category c, and optimize how similar it is to the feature; another one whose similarity score is the largest, is chosen from prototypes not containing the information of target category, and we punish its similarity to the feature:

$$L_{TNC-IM}^{Proto-k} = \min_{pr \in Pr_c} \min_{z \in patches(F_P)} (z - pr)^2 - \min_{pr \notin Pr_c} \min_{z \in patches(F_P)} (z - pr)^2 \tag{19}$$

Combining the formulas Eq.(12), Eq.(13) and Eq.(19), we can get the loss of TNC-IM:

$$L_{TNC-IM} = L_{TNC-IM}^{Ortho} + L_{TNC-IM}^{CrsEnt-0} + L_{TNC-IM}^{Proto-0} + \sum_{k=1}^{K} (L_{TNC-IM}^{CrsEnt-k} + L_{TNC-IM}^{Proto-k}) \tag{20}$$

Data Re-weighting and Model Selection. Use the prediction P_{TNC-IM} output by TMC-IM and the prediction P_{TNC-BM} output by TNC-BM as the input of the data re-weighting and model selection module, which is named TNC-DR&MS. TNC-DR&MS calculates the difference between maximum main task prediction probability of TNC-IM and $\frac{1}{C_0}$ to evaluate the distance d between samples and the main task classification plane of TNC-IM:

$$d = max(p_i^0) - \frac{1}{C_0} \tag{21}$$

After getting the distance d, according to Eq.(21), Eq.(1), Eq.(2), Eq.(3), the weightes of samples w_{TNC-IM} and w_{TNC-BM}, which are used to train TNC-IM and TNC-BM can be calculated by setting threshold δ. Substituting the loss of TNC-IM L_{TNC-IM} and the weight of samples w_{TNC-IM} into Eq.(4), the loss of TNC-IM after re-weighting $L_{TNC-IM}^{re-weight}$ is calculated:

$$L_{TNC-IM}^{re-weight} = \sum_{i=1}^{n} w_{TNC-IM}^i L_{TNC-IM}^i \tag{22}$$

Substituting the loss of TNC-BM L_{TNC-BM} and the weight of samples w_{TNC-BM} into Eq.(5), the loss of TNC-BM after re-weighting $L_{TNC-BM}^{re-weight}$ is calculated:

$$L_{TNC-BM}^{re-weight} = \sum_{i=1}^{n} w_{TNC-BM}^i L_{TNC-BM}^i \tag{23}$$

With the loss $L_{TNC-IM}^{re-weight}$ and $L_{TNC-BM}^{re-weight}$, the loss of TNC-IB-M $L_{TNC-IB-M}$ is:

$$L_{TNC-IB-M} = L_{TNC-IM}^{re-weight} + L_{TNC-BM}^{re-weight} \tag{24}$$

During the inference phase, the inference process of TNC-IB-M refers to IB-M.

TABLE I: Accuracy comparison among TNC-IB-M, TNC-IM, and TNC-BM.

Methods	**Accuracy**	
	Trained in IB-M	***Trained alone***
TNC-IM	79.47%	78.69%
TNC-BM	82.01%	80.35%
TNC-IB-M	**83.63%**	

IV. Experiment

In this section, we check the performance, including interpretation and accuracy, of our proposed IB-M framework and TNC-IB-M algorithm on the real medical application of ultrasound thyroid nodule diagnosis.

A. Dataset Description

The ultrasound image dataset of thyroid nodules dataset used in this paper is collected from 1,790 patients with a collection of 2,285 ultrasound images, including 1,055 images of malignant nodules and 1,230 images of benign nodules. The labeling of the images is done by thyroid professional doctors, as follows: (1) Compare pathological anatomy which is the gold standard, and label thyroid nodules benign and malignant. (2) Label the four key thyroid nodule attributes of echogenic foci, composition, margin, and echogenicity, which are considered to be the most important by the medical community for thyroid nodule diagnosis. Due to the extremely difficult medical data collection, in our cognitive category, this dataset is the largest ultrasound image dataset of thyroid nodules with finely labeled attribute information.

B. Implementation Details

We use DenseNet-201 [19] as the backbone and set the size of feature maps output by shared CNN to 128*7*7. In the loss function, we set the attribute loss weight coefficient α to 0.1, the parameter r_{TNC-IM} used in calculating w_{TNC-IM} to 1.3, and the parameter r_{TNC-BM} used in calculating w_{TNC-BM} to 2.Before the training phase, the input images are resized to 224*224 and normalized. During training, the batch size is set to 8, the learning rate is set to 0.03 and the total training epochs is set to 150.

C. Comparison with Single IM and Single BM

First, we conduct extensive experiments to demonstrate the effectiveness of our proposed IB-M framework and the TNC-IB-M algorithm on both model accuracy and interpretation.

To validate the advantage of our method on the accuracy, we compare our TNC-IB-M algorithm with a single interpretable model (including trained in our IB-M framework and trained alone) and a single black-box model (including trained in our IB-M framework and trained alone). We report the result in Table I. From the results, we have the following observations and analysis: (1) With aligning the TNC-IM and TNC-BM and letting each of them focus on the samples that they are good at respectively, our TNC-IB-M algorithm achieves the

TABLE II: Comparison with the state-of-art algorithms on the ultrasound image dataset of thyroid nodules.

Methods	Accuracy	F1-score	k value
Ko et al., 2019 [15]	78.47%	0.7853	0.5722
Song et al., 2019 [16]	78.78%	0.7771	0.5749
Li et al., 2019 [21]	79.70%	0.7894	0.5953
Buda et al., 2019 [17]	78.99%	0.7757	0.5781
Li et al., 2019 [18]	80.09%	0.8009	0.5981
TNC-IB-M	**83.63%**	**0.8272**	**0.6720**

best performance on predictive accuracy. (2) The performance of TNC-BM trained in our IB-M framework is better than it trained alone. The main reason is that by the DR&MS module, our IB-M framework can help to reduce the side effect of the samples that can be well-interpreted and predicted by TNC-IM, and let TNC-BM focus on those samples around true decision boundary. (3) With the help of DR&MS, the TNC-IM trained in the IB-M framework can also focus on the samples that it's good at, hence achieving a better performance than it trained alone.

D. Comparison with the State-of-Art Algorithms

We also compare our proposed method with *five* current most advanced algorithms for the problem of thyroid nodules classification as shown in Table II. Among them, [15], [16], [21] train a single-task learning model with only benign and malignant thyroid nodule labels. [17], [18] train a multi-task learning model to jointly predict thyroid nodule benign and malignant labels and attribute labels. [17] proposes a shared CNN followed by fully connected layers to achieve multi-tasking; and [18] uses multiple semantic branches comprehensive classification to achieve multi-tasks learning. In experiments, we use 10-fold cross-validation to ensure the reliability of results.

We report the results in Table II, and from the result, we have followed observations. Our proposed TNC-IB-M algorithm achieves the best performance on the task of ultrasound thyroid nodules classification, where the accuracy of our algorithm is 83.63%, achieving an improvement of 3.54% than baselines. The results on the other two measurements, F1-score and k value, also demonstrate the effectiveness of our proposed algorithm on the task of ultrasound thyroid nodule diagnosis.

V. Conclusion

In this paper, we focus on how to simultaneously optimize the model accuracy and interpretation. We argue that the trade-off between model accuracy and interpretation is not necessary for some scenarios. To break this trade-off, we propose an IB-M framework to align an interpretable model and a black-box model for simultaneously optimize the model accuracy and interpretation. Based on the IB-M framework, moreover, we design a TNC-IB-M algorithm for a real medical problem of ultrasound thyroid nodule diagnosis. Experimental results verify the practical usefulness of our proposed framework and method on model accuracy and model interpretation.

References

[1] Ardila, Diego, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." Nature medicine 25.6 (2019): 954-961.

[2] Liang, Huiying, et al. "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence." Nature medicine 25.3 (2019): 433-438.

[3] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[4] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems. 2017.

[5] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "' Why should I trust you?' Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

[6] Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." Advances in neural information processing systems. 2019.

[7] Wang, Tong, and Qihang Lin. "Hybrid predictive model: When an interpretable model collaborates with a black-box model." arXiv preprint arXiv:1905.04241 (2019).

[8] Zhengqing Fang, Kun Kuang, Yuxiao Lin, Fei Wu and Yu-Feng Yao. "Concept-based Explanation for Fine-grained Images and Its Application in Infectious Keratitis Classification." Proceedings of the 28th ACM international conference on Multimedia. 2020.

[9] Ghorbani, Amirata, et al. "Towards automatic concept-based explanations." Advances in Neural Information Processing Systems. 2019.

[10] Haugen, Bryan R., et al. "2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer." Thyroid 26.1 (2016): 1-133.

[11] Erhan, Dumitru, et al. "Visualizing higher-layer features of a deep network." University of Montreal 1341.3 (2009): 1.

[12] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

[13] Huang, Zixuan, and Yin Li. "Interpretable and Accurate Fine-grained Recognition via Region Grouping." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[14] Rafique, Hassan, et al. "Transparency Promotion with Model-Agnostic Linear Competitors."

[15] Ko, Su Yeon, et al. "Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound." Head Neck 41.4 (2019): 885-891.

[16] Song, Junho, et al. "Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules." Medicine 98.15 (2019).

[17] Buda, Mateusz, et al. "Management of thyroid nodules seen on US images: deep learning may match performance of radiologists." Radiology 292.3 (2019): 695-701.

[18] Li, Shuai, et al. "Fine-Grained Thyroid Nodule Classification via Multi-Semantic Attention Network." 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019.

[19] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[20] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[21] Li, Xiangchun, et al. "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study." The Lancet Oncology 20.2 (2019): 193-201.