MedOptNet: Meta-Learning Framework for Few-shot Medical Image Classification

Liangfu Lu, Xudong Cui, Zhiyuan Tan*, Yulei Wu

Abstract—In the medical research domain, limited data and high annotation costs have made efficient classification under few-shot conditions a popular research area. This paper proposes a meta-learning framework, termed MedOptNet, for few-shot medical image classification. The framework enables the use of various high-performance convex optimization models as classifiers, such as multi-class kernel support vector machines, ridge regression, and other models. End-to-end training is then implemented using dual problems and differentiation in the paper. Additionally, various regularization techniques are employed to enhance the model's generalization capabilities. Experiments on the BreakHis, ISIC2018, and Pap smear medical few-shot datasets demonstrate that the MedOptNet framework outperforms benchmark models. Moreover, the model training time is also compared to prove its effectiveness in the paper, and an ablation study is conducted to validate the effectiveness of each module.

Index Terms—few-shot; meta learning; convex optimization; medical image classification

1 Introduction

In recent years, deep learning has demonstrated remarkable prowess and boundless potential in fields such as computer vision [1], [2], [3], [4], [5], [6] and natural language processing [7], [8]. In certain instances, models, including ResNets [9], have even outperformed humans on specific datasets. This has spurred significant interest in applying computer vision to the medical field, with researchers exploring areas such as breast cancer immunohistochemical image generation [10], lesion image segmentation [11], and X-ray projectomic reconstruction [12]. Despite some progress, leveraging medical imaging for paramedical technology remains a formidable challenge.

The primary obstacle in applying computer vision to the medical field stems from deep learning's reliance on large-scale, well-annotated datasets [1], [6], [13], [14], [15], [16]. Acquiring a substantial number of labeled images is both expensive and time-consuming, which constrains the model's applicability. Although there are numerous open datasets available in the computer vision domain, acquiring sufficient training data in the medical field is considerably difficult [17] as they need to be annotated by medical experts for their usefulness. Therefore, by addressing the challenge of few-shot learning, which aims to achieve efficient classification tasks using few images, it will help integrate machine learning into doctors' workflows and significantly alleviate the burden on medical personnel.

Meta-learning, also known as "learning to learn," is a machine learning paradigm that seeks to emulate the human cognitive process, which can quickly acquire new knowledge without the need for large amounts of data [18]. The ultimate goal of meta-learning is to develop models that can generalize to previously unseen tasks by learning at the task level, rather than the sample level. It is one of the potentially effective ways to deal with the data scarcity problem in real-life few-shot learning scenarios [19]. There are two primary phases involved in meta-learning: the meta-learning phase and the adaptation phase(Meta-test phase).

1) Meta-learning phase: In this stage, the model is exposed to a diverse set of tasks and learns to recognize patterns and relationships across these tasks. The primary objective is to acquire a broad understanding of the problem space and develop a strong foundation that can be adapted to new tasks. This process typically involves training the model on a large number of tasks, each with its own dataset and objective.
2) Adaptation phase: Once the model has been trained on various tasks, it is then fine-tuned on a new task with a limited amount of training data. The model leverages its prior knowledge, acquired during the meta-learning phase, to quickly adapt to the novel task. The adaptation phase is crucial for achieving high performance on the target task, as it allows the model to utilize its prior knowledge and make accurate predictions with only a few training samples.

By following these two phases, meta-learning models can effectively tackle few-shot learning problems, which are characterized by their limited availability of training data. This approach enables the development of models that can rapidly adapt to new tasks, much like human cognitive learning, making them more versatile and robust in real-world applications.

Although, Few-shot learning has emerged as a prominent research area in computer vision and cultivated several outstanding models [20], [21], [22]. It is worth highlighting that most of these models are evaluated on conventional

- *Liangfu Lu, Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China*
- *Xudong Cui, School of Mathematics, Tianjin University, Tianjin 300350, China.*
- *Yulei Wu, Department of Computer Science, Faculty of Environment, Science and Economy, University of Exeter, UK.*
- *Zhiyuan Tan, School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Edinburgh EH10 5DT, Scotland, UK. E-mail:z.tan@napier.ac.uk*

datasets , such as Mini-ImageNet [23]. It is questionable whether such datasets truly capture the complexities and nuances of practical applications. For instance, the medical domain often grapples with ethical and privacy concerns when it comes to data collection, and even when available, datasets can have an uneven sample distribution across different classes. These challenges hinder the development and evaluation of few-shot learning models that can generalize well to real-world medical cases. Therefore, it is imperative to explore new datasets that can represent real-world scenarios and develop models that can learn from them. Additionally, further research is required to devise novel methods and theoretical foundations that can address the challenges of few-shot learning under diverse practical settings. Training and evaluating few-shot learning models that can effectively generalize to new medical cases becomes challenging due to the difficulty in obtaining appropriate datasets and ethical concerns. Moreover, real-world data is usually more complex and variable than data in controlled laboratory settings, thereby presenting additional challenges for few-shot learning methods. To increase the applicability of few-shot learning to practical problems, it is crucial to explore novel datasets that represent real-world scenarios and build models capable of effectively learning from them. Additionally, further research is required to develop innovative methods and theoretical frameworks that can address the challenges in few-shot learning under diverse practical settings.

To address the unique characteristics of medical datasets, we propose a meta-learning framework called MedOptNet, which is based on convex optimization. In contrast to classical meta-learning methods that utilize nearest-neighbor classifiers or one-layer neural networks, our trained convex optimization model sees greater generalization even for few-shot classification. Besides, convex optimization models are promising for medical image processing. In this paper, our meta-learning approach is experimented with several convex optimization models, including multi-class kernel support vector machines, ridge regression. In terms of model training , convex optimization models essentially solve quadratic programming problems, using the optnet approach that enables end-to-end training with an efficient quadratic programming solver [24]. Different convex optimization models for different medical datasets to achieve optimal performance.

In summary, the contributions of our approach are as follows:

- To address the issue of few-shot medical classification, we propose a meta-learning framework called MedOptNet. This framework allows for the utilization of various high-performance convex optimization models as classifiers to conduct non-linear classification. These models can undergo end-to-end training using quadratic programming solvers. Our method aims to enhance the performance of classifiers in situations where only a limited number of labeled samples are available.
- In order to tackle the imbalanced distribution of medical data, several regularization techniques are employed, including image augmentation, to improve the model's generalization capability. The effectiveness of each module is demonstrated through ablation experiments.
- We employ various convex optimization models, such as multi-class kernel SVM and ridge regression, as classifiers for a range of medical datasets. By comparing the performance of our model to classical meta-learning methods (including MAML, Reptile, ProtoNet, MatchingNet, and RelationNet), we demonstrate that our model achieves optimal performance and effectively tackles the challenges presented by the BreakHis, ISIC2018, and Pap smear datasets.
- We compared the runtime of the models, and the computational cost of our model only slightly increased, demonstrating the practical value of our model.

The rest of the paper is structured as follows: In Section 2, we provide an overview of the related work. In Section 3, we introduce our proposed model, named MedOptNet, and provide details about its architecture. Section 4 describes the experimental setup, including the implementation details, and presents the analysis of the classification results. Lastly, in Section 5, we conclude our study.

2 Related work

Few-shot learning methods can be broadly classified into three categories: measure-based [25], [26], [27], [28], [29], [30], [31], [32], model-based [33], [34], and optimization-based methods [26], [35]. Measure-based methods employ the nearest-neighbor idea to classify samples. Model-based methods are designed to establish a mapping from input space to feature space, which enables quick adaptation to new parameters in few-shot scenarios. Optimization-based methods formulate the adaptation process as an optimization problem and use better optimization strategies. Despite the variety of methods, many few-shot learning models adopt a two-stage meta-learning framework comprising meta-training and meta-test stages. Typically, these models consist of two components: an embedded model for feature extraction and a classifier. In the case of extremely limited training data, traditional convolutional neural networks utilize four-layer frameworks or ResNets-12. Classifiers typically use simple nearest-neighbor algorithms [28] or one-layer neural networks [26]. Support vector machines are often used as classifiers in few-shot learning because they can effectively utilize the high-dimensional embeddings of samples and introduce parameter regularization to prevent over-fitting.

Few-shot classification involves the use of a limited amount of training data to classify samples from unseen classes. As previously mentioned, few-shot learning methods can be broadly classified into three categories, and our focus is mainly on optimization methods, which are most relevant to our work. Model-Agnostic Meta Learning (MAML) [35] is a task-independent algorithm for meta-learning that trains model parameters to enable fast learning of new tasks with minimal gradient updates. Distance-based models aim to learn supervised or unsupervised classifiers with strong generalization ability. In traditional machine

learning, the base class is used for training, but due to the imbalance between the base and new classes, overfitting can occur. To address this, Koch et al. [25] proposed Siamese Neural Networks for One-shot learning. In this approach, the model extracts features in parallel and compares them to obtain similarity, allowing for effective classification with a small amount of data. Vinyals et al. [26] proposed a matching network based on memory and attention as an extension of prototype networks, while some researchers use graph neural networks for information transmission [36], [37].

Several researchers have explored the application of meta-learning in medical image processing. Prabhu et al. [38] proposed Prototypical Clustering Networks, which are based on the prototype network and enable fast generalization. They applied this method to dermatology datasets. Li et al. [39] introduced a meta-learning method for difficulty perception that was trained on a dataset of common diseases and then applied to rare diseases. Singh et al. [40] applied Reptile and Prototypical Networks to medical image datasets and compared the confidence of models between transfer learning and meta-learning. He et al.. [41] presented a meta-learning-based approach called MetaMed, which can adapt to rare disease classes with only a few available images and less computational resources. Xie et al. [37] explored a novel method based on metric-based meta-learning for Cross-Domain Few-Shot (CDFS) problems in the classification of welding defects. Crammer and Singer f [42] described an algorithm implementation of multi-class kernel functions, which can transform multi-class problems into quadratic programming problems. Bertinetto et al. [35] proposed an iterative solver based on ridge regression and logical regression, viewing machine learning as a part of deep learning. Amos et al. [24] utilized a primal-dual interior point algorithm to solve quadratic programming problems, which can be deployed on a GPU. Agrawal et al. [43] described three general convex optimization models, including maximum a posteriori models, utility maximization models, and agent models.

While existing few-shot models have shown strong performances, their application to medical datsets has primarily relied on classic MAML or prototype network models. But it is not optimized for medical datasets and has poor performance. In this work, we propose a novel approach using a convex optimization model as the classifier. This approach offers the benefits of a mathematically elegant and interpretable model that can be efficiently solved using a primal-dual interior point algorithm, while also leveraging the computational power of modern GPUs to enable rapid training.

3 Framework

In this section, we first introduce the definition of the few-shot classification task, then introduce the MedOptNet model in detail, mainly detailing the training of support vector machines with multi-class kernel functions, and then introduce various other convex optimization models, such as SiMSVM and ridge regression.

3.1 Problem definition

The goal of few-shot classification models is to train a classification model using a limited number of labeled samples and then use the trained model to classify unlabeled samples. However, with only a few labeled samples, it is far from being enough to train a neural network. Thus, this work adopts meta-learning, an effective approach for few-shot learning. Each classification task $\mathcal{T}$ includes a support (training) set $\mathcal{S}$ and a query (test) set $\mathcal{Q}$. In the meta-test phase, if the support set $\mathcal{S}$ contains N classes, each class has K training samples; the task is called an N-way K-shot problem.

In this case, meta-learning is instantiated in the following way: during the meta-learning phase, the model is trained on multiple similar tasks to learn how to rapidly adapt to new tasks. In each task, the model is trained using the support set $\mathcal{S}$ and calculates the meta-loss based on the query set $\mathcal{Q}$. By optimizing the meta-loss, the model is able to fine-tune itself for each new task. In the meta-test phase, the trained meta-learner adapts to the new task using a small number of labeled samples (i.e., the support set $\mathcal{S}$). This enables the model to perform effective classification on unlabeled samples with limited labeled data.

3.2 MedOptNet

A few-shot classification model typically consists of two parts: a convolutional neural network and a classifier. Fig. 1 illustrates the proposed method. During the meta-training phase, images are passed through the convolutional neural network f_θ to obtain feature vectors, which are then fed into the classifier. Suppose the classifier has received feature vectors from the support set of 3 classes, the query set images must belong to one of these 3 classes. In this work, a convex optimization model is used as the classifier. The convex optimization model classifies the images in the query set. Then, the predicted results are compared with the true labels to calculate the loss value. The network parameters in the backbone are updated using backpropagation, completing the training of one episode. This method employs a convex optimization model to learn the classifier and calculates the generalization error of a set of labeled training samples during the meta-training phase.

In terms of convolutional neural networks, the most popular ResNet-12 is used as the feature extractor. The focus of the following content is on the forward and backward propagation of the convex optimization model. First, we discuss the forward propagation of the convex optimization model, which involves solving the convex optimization model and using the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$ to predict the label $y \in \mathcal{Y} = 1, ..., k$, where $x \in \mathcal{X}$. Specifically, we consider the model $\phi : f_\theta(\mathcal{X}) \to \mathcal{Y}$, which predicts the output y by solving a convex optimization problem that depends on the input $f_\theta(x)$. Unlike most multiclass classification methods, the multiclass nonlinear SVM used in the MedOptNet model [42] requires solving only one

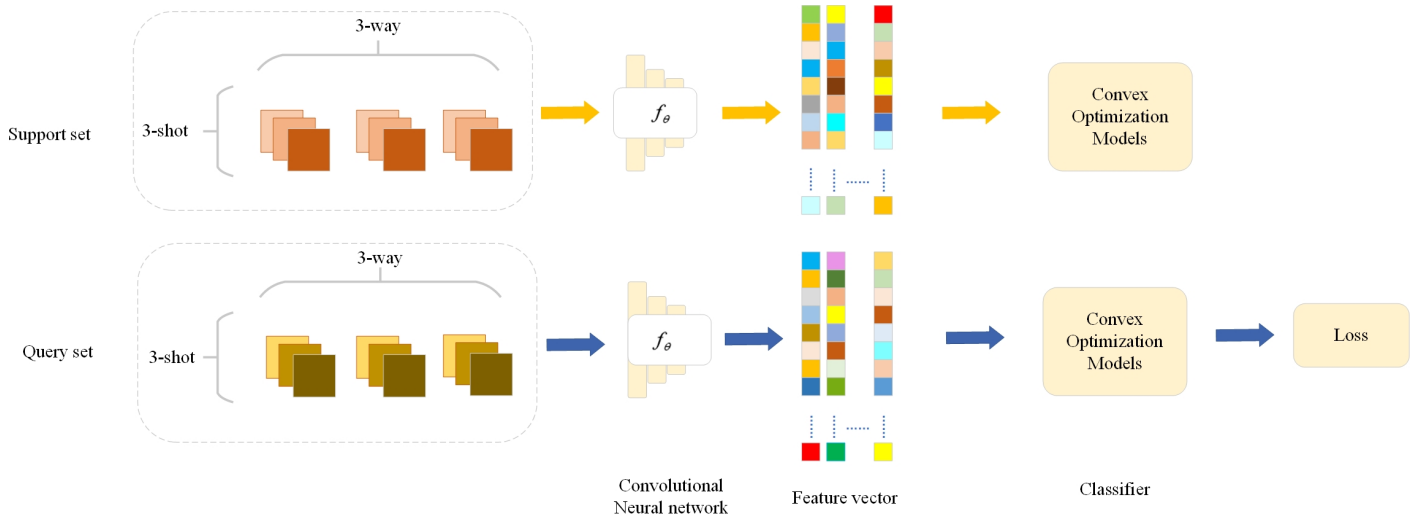


Fig. 1: The overall framework of the MedOptNet.

optimization problem (1).

$$\min_{W,\xi} \frac{1}{2}\beta \sum_{r=1}^{k} \|w_r\|_2^2 + \gamma \sum_{i=1}^{m} \xi_i \tag{1}$$

subject to:

$$w_{y_i} \cdot f_\theta(x_i) + \delta_{y_i,r} - w_r \cdot f_\theta(x_i) \geq 1 - \xi_i \quad \forall i, r.$$

W is a matrix of size $k \times n$ and w_r is the rth row of W. We define the l_2-norm of w_r as $\|w_r\|_2^2 = \sum_j w_{r,j}^2$. Given a regularization constant $\beta > 0$, the inequality constraint for $r = y_i$ becomes $\xi_i \geq 0$, where ξ is the slack variable defined as $\xi = \gamma \sum_{i=1}^{m} \xi_i$, and $\delta_{.,.}$ denotes the Kronecker delta function.

Eventually, the classifier takes the form:

$$H(f_\theta(x)) = \arg\max_{r=1}^{k} \{w_r \cdot f_\theta(x)\}. \tag{2}$$

To solve the optimization problem, we introduce a dual set of variables, one for each constraint, and derive the Lagrangian of the optimization problem.

$$\begin{aligned} L(W, \xi, \eta) =& \frac{1}{2}\beta \sum_{r} \|w_r\|_2^2 + \gamma \sum_{i=1}^{m} \xi_i \\ &+ \sum_{i,r} \eta_{i,r}[w_{y_i} \cdot f_\theta(x_i) + \delta_{y_i,r} + 1 - \xi_i] \end{aligned} \tag{3}$$

subject to :

$$\forall i, r \quad \eta_{i,r} \geq 0.$$

To determine the minimum for the primal variables W, ξ and the maximum for the dual variables η, we need to find a saddle point of the Lagrangian. In order to obtain the minimum over the primal variables, we need to take the partial derivatives of the Lagrangian with respect to W and ξ, and set them equal to zero. Similarly, to obtain the maximum over the dual variables, we need to take the partial derivative of the Lagrangian with respect to η and set it equal to zero. This gives us a system of equations, which can be solved using the Karush-Kuhn-Tucker (KKT) conditions [44]. The KKT conditions provide a set of necessary conditions for a point to be a saddle point of the Lagrangian, and they ensure that the primal and dual solutions are optimal. By satisfying the KKT conditions, we can obtain the optimal solutions for the primal and dual variables of the optimization problem.

$$\frac{\partial}{\partial \xi_i} L = \gamma - \sum_{r} \eta_{i,r} = 0 \quad \Rightarrow \quad \sum_{r} \eta_{i,r} = \gamma, \tag{4}$$

Similarly, for w_r we require,

$$\begin{aligned} \frac{\partial}{\partial w_r} &= \sum_{i} \eta_{i,r} f_\theta(x_i) - \sum_{i, y_i = r} \underbrace{\left(\sum_{q} \eta_{i,q} \right)}_{=\gamma} f_\theta(x_i) + \beta w_r \\ &= \sum_{i} \eta_{i,r} f_\theta(x_i) - \gamma \sum_{i} \delta_{y_i r} f_\theta(x_i) + \beta w_r = 0, \end{aligned} \tag{5}$$

which results in the following form

$$w_r = \beta^{-1} \left[\sum_{i} \left(\gamma \delta_{y_i,r} - \eta_{i,r} \right) f_\theta(x_i) \right]. \tag{6}$$

Next, we can obtain the objective function of the dual program by substituting Equations (5) and (6) into Equation (3) to develop the Lagrangian using only the dual variables. The resulting expression is as follows:

$$\begin{aligned} &\frac{\gamma - 1}{2\beta} \sum_{i,j} (f_\theta(x_i) \cdot f_\theta(x_j)) [\sum_{r} (\delta_{y_i,r} - \eta_{i,r})(\delta_{y_j,r} - \eta_{j,r})] \\ &- \sum_{i,r} \eta_{i,r} \delta_{y_i,r}. \end{aligned} \tag{7}$$

We can use the notation $\overline{1}_i$ to represent the vector whose components are all zero except for the i-th component which is equal to one, and $\overline{1}$ to represent the vector whose components are all one. With this notation, we can rewrite the dual program in the following vector form:

$$\begin{aligned} \max_{\eta} &- \frac{1}{2}\beta^{-1} \sum_{i,j} (f_\theta(x_i) \cdot f_\theta(x_j)) \left[(\overline{1}_{y_i} - \eta_i) \cdot (\overline{1}_{y_j} - \eta_j) \right] \\ &- \sum_{i} \eta_i \cdot \overline{1}_{y_i} \end{aligned} \tag{8}$$

$$\text{subject to :} \qquad \forall i : \eta_i \geq 0 \quad \text{and} \quad \eta_i \cdot \overline{1} = 1.$$

It can be shown that $\mathcal{Q}(\eta)$ is a concave function of η. Since the set of constraints is convex, there exists a unique maximum value of $\mathcal{Q}(\eta)$. To simplify the problem, we introduce a change of variables. Specifically, we define $\tau i = \overline{1} y_i - \eta i$, which represents the difference between the point distribution $\overline{1} y_i$ that concentrates on the correct label and the distribution η_i obtained by solving the optimization problem. With this change of variables, the expression for w_r given in Equation (6) can be written as follows:

$$w_r = \beta^{-1} \sum_i \tau i, r f_\theta(x_i). \tag{9}$$

Using Lagrange duality, the dual form of Equation (1) is obtained as shown in Equation (10):

$$\begin{aligned} \max_{\tau} & -\frac{1}{2} \sum_{i,j} \left(f_\theta(x_i) \cdot f_\theta(x_j)\right) (\tau_i \cdot \tau_j) + \beta \sum_i \tau_i \cdot \overline{1}_{y_i} \\ & \text{subject to} : \forall i \quad \tau_i \leq \overline{1}_{y_i} \quad \text{and} \quad \tau_i \cdot \overline{1} = 0. \end{aligned} \tag{10}$$

Finally, in relation to the variable τ, the classifier $H(f_\theta(x))$ is as follows:

$$\begin{aligned} H(f_\theta(x)) &= \arg\max_{r=1}^{k} \{w_r \cdot f_\theta(x)\} \\ &= \arg\max_{r=1}^{k} \left\{ \sum_i \tau_{i,r} \left(f_\theta(x_i) \cdot f_\theta(x)\right) \right\}. \end{aligned} \tag{11}$$

The dual form and the final classifier only rely on the inner product to calculate $(f_\theta(x_i) \cdot f_\theta(x))$, so we can introduce the kernel equation $K(\cdot, \cdot)$ inner product operation in high-dimensional space to replace $(f_\theta(x_i) \cdot f_\theta(x))$. The dual problem with kernel equation is as follows:

$$\begin{aligned} \max_{\tau} & -\frac{1}{2} \sum_{i,j} K\left(f_\theta(x_i), f_\theta(x_j)\right) (\tau_i \cdot \tau_j) + \beta \sum_i \tau_i \cdot \overline{1}_{y_i} \\ & \text{subject to: } \forall i \ \tau_i \leq \overline{1}_{y_i} \quad \text{and} \quad \tau_i \cdot \overline{1} = 1. \end{aligned} \tag{12}$$

Classifier $H(f_\theta(x))$ has the form:

$$H(f_\theta(x)) = \arg\max_{r=1} \left\{ \sum_i \tau_{i,r} K\left(f_\theta(x), f_\theta(x_i)\right) \right\}. \tag{13}$$

The Linear Kernel is actually a linearly separable SVM, and its expression is:

$$K(x, z) = x \bullet z, \tag{14}$$

In the above formula, "•"represents the inner product of vectors. However, for nonlinear classification problems, kernel support vector machines perform better. In the following, we introduce several commonly used kernel functions.

Polynomial Kernel,

$$K(x, z) = (\gamma x \bullet z + r)^d, \tag{15}$$

Among them, γ, r, d all need their own parameter definition.

Gaussian Kernel

$$K(x, z) = \exp\left(-\gamma \|x - z\|^2\right), \tag{16}$$

Note that γ is a positive parameter that must be defined by researchers, and its value should be adjusted accordingly.

Sigmoid Kernel

$$K(x, z) = \tanh(\gamma x \bullet z + r), \tag{17}$$

Among them, γ and r all need to be defined by themselves.

To ensure end-to-end trainability of our system, it is crucial that the solution of the SVM solver is differentiable with respect to its input. The SVM objective is convex and has a unique global minimum, which enables us to use the implicit function theorem on the Karush-Kuhn-Tucker (KKT) optimality conditions and obtain the necessary gradients. To provide a comprehensive understanding, we present the derivation of the theorem for convex optimization problems.

$$\begin{aligned} L(W, \xi, \eta) =& \frac{1}{2} \beta \sum_r \|w_r\|_2^2 + \gamma \sum_{i=1}^{m} \xi_i \\ & + \sum_{i,r} \eta_{i,r} [w_{y_i} \cdot f_\theta(x_i) + \delta_{y_i,r} + 1 - \xi_i], \\ p(W, \xi) =& 1 - \xi_i - \{W_{y_i} f_\theta(x_i) + \delta_{y_i,r} - W_r f_\theta(x_i)\}. \end{aligned} \tag{18}$$

The vector $f_\theta(x_i)$ If and only if there are $\tilde{\eta}$ that satify, the KKT conditions:

$$\begin{aligned} & p(W, \xi) \preceq 0, \\ & \tilde{\eta}_i \geq 0, \quad i = 1, \ldots, m \\ & \tilde{\eta}_i p_i(W, \xi) = 0, \quad i = 1, \ldots, m \\ & \nabla_W L(W, \tilde{\eta}, \xi) = 0. \end{aligned} \tag{19}$$

We are going to reduce the KKT equations to an algebraic equation and apply the implicit function theorem. We first let $z = (W, \eta)$ for notational convenience and then define the function:

$$g(z, \xi) = \begin{bmatrix} \nabla_W L(W, \eta, \xi) \\ \operatorname{diag}(\eta) p(W, \xi) \\ 0 \end{bmatrix}, \tag{20}$$

where $\operatorname{diag}(\cdot)$ transforms a vector into a diagonal matrix. We define the (partial) Jacobian matrix as follows:

$$\mathrm{D}_z g(\tilde{z}, \xi) = \begin{bmatrix} \mathrm{D}_W \nabla_W L(\tilde{W}, \tilde{\eta}, \xi) & \mathrm{D}_W p(\tilde{W}, \xi)^T \\ \operatorname{diag}(\tilde{\eta}) \mathrm{D}_W p(\tilde{W}, \xi) & \operatorname{diag}(p(\tilde{W}, \xi)). \end{bmatrix} \tag{21}$$

If $g(\tilde{z}, \xi) = 0$, $\mathrm{D}_W g(\tilde{z}, \xi)$ is non-singular, then the solution mapping has a single-valued localization s around $\tilde{W}, \tilde{\eta}$ that is continuously differentiable in a neighborhood Q of ξ with Jacobian satisfying:

$$\mathrm{D}_\xi s(\xi) = -\mathrm{D}_z g(\tilde{W}, \tilde{\eta}, \xi)^{-1} \mathrm{D}_\xi g(\tilde{W}, \tilde{\eta}, \xi), \text{ for every } \xi \in Q. \tag{22}$$

The loss $\mathcal{L}(W, \theta)$ we use to optimize the feature extractor and the classifier is as follows:

$$\mathcal{L}(W, \theta) = \sum_{(\boldsymbol{x}, y) \in \mathcal{D}^{\text{test}}} \left[-\boldsymbol{w}_y \cdot f_\theta(\boldsymbol{x}) + \log \sum_k \exp\left(\boldsymbol{w}_k \cdot f_\theta(\boldsymbol{x})\right) \right]. \tag{23}$$

3.3 Other convex optimization models

The MedOptNet framework serves as a classification tool, as previously analyzed. Researchers can select from various convex optimization models based on the characteristics of their datasets.

By utilizing the MedOptNet framework, researchers can leverage the flexibility of convex optimization models to achieve optimal results for their specific datasets. The following section will provide further details on the specific

Algorithm 1 MedOptNet: few-shot medical image classification using Convex Optimization Models.

Input: $\mathcal{D}$: dataset; α: learning rates; N:number of training epochs; K:number of support examples per training class; β, γ: regularization parameter; Q:number of query examples per test class;
Output: Accuracy;
Randomly initialize θ, W
for the number of training iterations **do**
  Sample a batch of tasks$\mathcal{T} \sim p(\mathcal{T})$;
  Sample a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1:N\times K}$ a query set$\mathcal{Q} = \{(x_i, y_i)\}_{i=N\times K+1:N\times(K+Q)}$ from $p(\mathcal{T})$;
  Get the embedding of samples;
  Solve the convex optimization model and get the prediction of the query set;
  Envalue cross entropy loss $\mathcal{L}(W, \theta) by Equation(23)$;
  $\theta, W \longleftarrow \theta, W - \bigtriangledown \alpha \mathcal{L}(W, \theta)$;
end for
for the number of testing iterations **do**
  Sample a batch of tasks$\mathcal{T} \sim p(\mathcal{T})$;
  Sample a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1:N\times K}$ a query set$\mathcal{Q} = \{(x_i, y_i)\}_{i=N\times K+1:N\times(K+Q)}$ from $p(\mathcal{T})$;
  Get the embedding of samples;
  Solve the convex optimization model and get the prediction of the query set;
end for

models available through the framework.

Weston and Watkins multi-class SVM The Support Vector Machine (SVM) algorithm is commonly used for binary classification problems. However, for multi-class pattern recognition problems, traditional methods often rely on combining multiple binary classification decision functions using a voting scheme. To address this issue, Weston and Watkins [45] proposed two extensions to the SVM algorithm that enable k-class pattern recognition problems to be solved in a single step, without the need for multiple binary classifiers.

Weston and Watkins proposed k-class SVM algorithm offers a straightforward approach to multi-class pattern recognition problems, as their methods do not require the use of a voting scheme. The implications of their work are significant for the development of SVM-based approaches in pattern recognition. Their proposed methods should be considered for solving k-class problems in a single step, as neither of their methods requires the use of multiple binary classifiers.

The classifier form of their proposed method is as follows:

$$H(f_\theta) = \arg\max\{w_i \bullet f_\theta(x_i) + b_i\}, i = 1, \ldots, n. \quad (24)$$

Their method generalizes the binary SVM optimization problem to minimize and optimize k classes in a single step, eliminating the need to combine multiple binary classification rules. The first method is a direct generalization of the binary classification SVM approach, and in the special case of $k = 2$, it yields identical support vectors and hyperplane. The second method involves solving a linear program, rather than a quadratic one. Overall, the proposed methods offer a more efficient and effective solution for k-class pattern recognition problems.

The optimization problem for their proposed method is given by:

$$\begin{aligned} &\min_{W,\xi} \frac{1}{2}\sum_{r=1}^{k}\|w_r\|_2^2 + \gamma\sum_{i=1}^{m}\sum_{r\neq y_i}\xi_i^r \\ &\text{subjec to:} \\ &(w_{y_i}\cdot f_\theta(x_i)) + b_{y_i} \geq (w_r \cdot f_\theta(x_i)) + b_r + 2 - \xi_i^r \\ &\xi_i^r \geq 0, \quad i = 1, \ldots, \ell \quad r \in \{1, \ldots, k\}\backslash y_i. \end{aligned} \quad (25)$$

This optimization problem minimizes and optimizes k classes in a single step, where $\|w_r\|_2^2$ represents the regularization term for the r-th class and ξ_i^r is the slack variable for the i-th instance and r-th class. The objective function balances the trade-off between maximizing the margin between classes and minimizing the classification error. The constraints ensure that the correct class is assigned to each instance while allowing for a margin of error.

SimMSVM He et al. [46] proposed a simplified multi-class SVM algorithm that directly solves a multi-class classification problem. By introducing a relaxed classification error bound to modify Crammer and Singer's multi-class SVM, the proposed SimMSVM reduces the size of the dual variables from $l \times k$ to l, where l and k are the sizes of the training data and number of classes, respectively. Moreover, we prove that the dual formulation of the proposed SimMSVM is exactly the same as that of Crammer and Singer's approach, with an additional constraint.

The optimization problem for the proposed SimMSVM is formulated as follows:

$$\begin{aligned} &\min_{W,\xi} \frac{1}{2}\sum_{r=1}^{k}\|w_r\|2^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{subject to: } \boldsymbol{w}_{yi}^T f_\theta(x_i) - \frac{1}{k-1}\sum_{m\neq y_i}\boldsymbol{w}_r^T f\theta(x_i) \geq 1 - \xi_i, \\ &\xi_i \geq 0, \quad i = 1, \ldots, l. \end{aligned} \quad (26)$$

The dual form is as follows:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}\in\mathcal{R}^l} \quad \tfrac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{G}\boldsymbol{\alpha} - \boldsymbol{e}^T\boldsymbol{\alpha}, \\ &\text{subject to:} \quad \boldsymbol{0} \geq \boldsymbol{\alpha} \geq C\boldsymbol{e}. \end{aligned} \quad (27)$$

The Hession G is an $l \times l$ matrix with its entries

$$G_{i,j} = \begin{cases} \frac{k}{k-1}K_{i,j}, & \text{if } y_i = y_j, \\ \frac{-k}{(k-1)^2}K_{i,j}, & \text{if } y_i \neq y_j, \end{cases} \quad (28)$$

$K_{i,j}$ is the kernel value.

Ridge regression model Bertinetto et al. [35] proposed a novel approach to use ridge regression for classification tasks, which traditionally has been used for regression problems. By introducing a novel interpretation of the regression model, they showed that ridge regression can be used effectively for classification tasks. In contrast, Lee et al. [33] conducted experiments to compare the performance of ridge regression and linear SVMs for classification tasks. They found that ridge regression falls short of linear SVMs

in terms of performance. However, the quadratic programming formulation of ridge regression can still be leveraged by implementing it within the existing framework.

The model for ridge regression is expressed as follows:

$$\boldsymbol{w}_k\left(\boldsymbol{\alpha}^k\right)=\sum_n \alpha_n^k f_\theta\left(\boldsymbol{x}_n\right) \quad \forall k$$

$$\max_{\{\boldsymbol{\alpha}^k\}}\left[-\frac{1}{2}\sum_k\left\|\boldsymbol{w}_k\left(\boldsymbol{\alpha}^k\right)\right\|_2^2-\frac{\lambda}{2}\sum_k\left\|\boldsymbol{\alpha}^k\right\|_2^2+\sum_n \alpha_n^{y_n}\right] \quad (29)$$

$$\text{subject to:} \quad \alpha_n^{y_n}\leq C, \quad \alpha_n^k\leq 0 \quad \forall k\neq y_n$$

$$\sum_k \alpha_n^k=0 \quad \forall n$$

4 Experimental Results

4.1 Data description and segmentation

BreakHis dataset [17]. The BreakHis dataset of breast cancer contains 9,109 microscopic images of breast tumor tissue, with magnifications of 40, 100, 200, and 400. The dataset includes 8 classes, from which we selected 5 categories for meta-training: Duck Al Carcinoma (903 images), Fibroadenoma (260 images), Mucinous Carcinoma (222 images), Lobular Carcinoma (170 images), and Papillary Carcinoma (150 images), one-tenth of the data is used as a validation set. In the meta-test stage, we chose three categories as meta-test categories: Phyllostachys Tumors (142 images), Tubular Adenoma (121 images), and Adenosis (113 images). All images are in the RGB format and stored in PNG format.

ISIC 2018 skin lesion dataset [47]. The ISIC 2018 Skin Damage datset contains 10,015 dermatoscopic images in seven categories. For our meta-learning experiments, we selected four categories as meta-training categories: Melanocytic Nevus (6,705 images), Melanoma (1,133 images), Benign Keratosis (1,099 images), and Basal Cell Carcinoma (514 images), one-tenth of the data is used as a validation set. In the meta-test stage, we chose three categories, namely Actinic Keratosis (327 images), Vascular Lesion (142 images), and Dermatofibroma (115 images). All images are in RGB format and stored as JPEG files.

Pap smear dataset [48]. The Pap-smear database comprises two versions created by the Herlev University Hospital. The images were prepared and analyzed by hospital staff using the CHAMP commercial software package (Dimac). For our meta-learning experiments, we selected four categories as meta-training categories: Severe Dysplastic, Moderate Dysplastic, Light Dysplastic, and Carcinoma in Situ, one-tenth of the data is used as a validation set. These categories contain 196, 146, 182, and 150 pictures, respectively. In the meta-test stage, we used three categories, namely Normal Superficial, Normal Intermediate, and Normal Columnar, with 74, 70, and 98 pictures, respectively. All images are RGB and stored in JPEG format.

4.2 Implementation Details

In this paper, we present a model that addresses the challenge of processing a limited number of samples by employing two distinct network architectures: a four-layer Convolutional Neural Network (ConvNet) and ResNet-12.The ConvNet is structured with four Conv-BN-ReLU layers, featuring convolutional kernels of varying sizes: 64, 96, 128, and 256. This design allows for the extraction of diverse features from the input data. ResNet-12, on the other hand, is a deep residual network that consists of multiple residual blocks, which significantly improves the flow of information and gradients through the network by utilizing skip connections. Fig.2 provides a detailed illustration of the network structures for both the four-layer ConvNet with specified convolutional kernel sizes and ResNet-12, showcasing the architectural differences and highlighting the unique characteristics of each approach.

Meta-training. During the meta-training phase, we employ

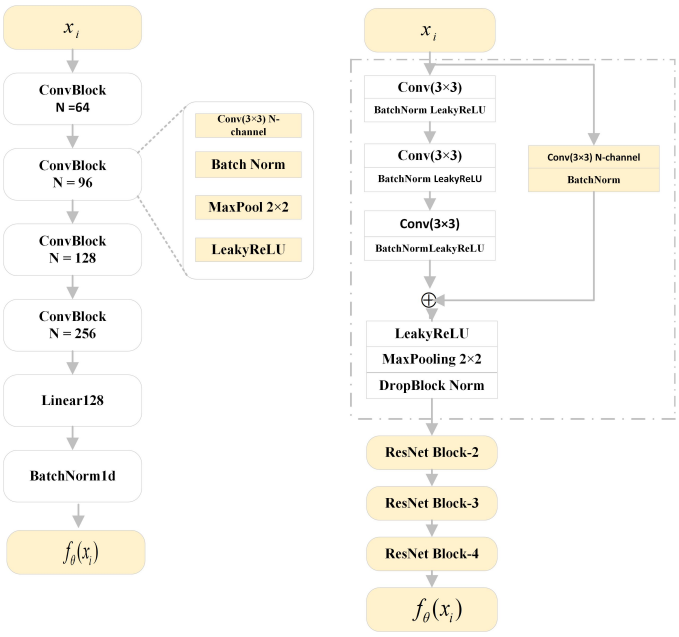
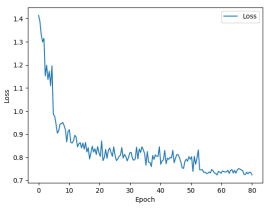
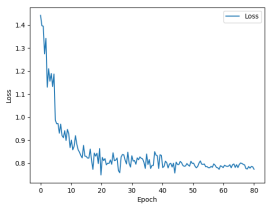


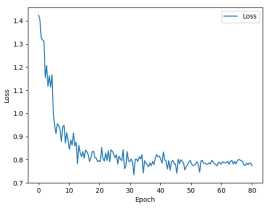
Fig. 2: Convolutional neural network structures.



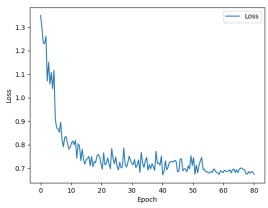
(a) The validation loss on the BreakHis 40X dataset.



(b) The validation loss on the BreakHis 100X dataset.



(c) The validation loss on the ISIC2018 dataset.



(d) The validation loss on the Pap smear dataset.

Fig. 3: The validation loss on the dataset.

TABLE 1: The 3-way, 1-shot and 3-shot Classification results on BreakHis datset. Average few-shot classification accuracies with 95% confidence intervals on BreakHis meta-test splits. "64-96-128-256" denotes a 4-layer convolutional network with 64, 96, 128 and 256 filters in each layer. "RR" stands for ridge regression model.

Model	Backbone	BreakHis 40X 3-way		BreakHis 100X 3-way	
		1-shot	3-shot	1-shot	3-shot
MAML	64-96-128-256	49.43±0.24%	52.10±0.32%	50.42±0.46%	59.63±0.66%
Reptile	64-96-128-256	56.20±1.84%	64.11±0.27%	60.53±0.31%	66.86± 0.74%
Matching Networks	64-96-128-256	58.56±0.44%	67.21±0.42%	59.34±0.67%	67.34±0.46%
Prototypical Networks	64-96-128-256	62.42±0.33%	71.23±0.36%	62.31±0.89%	70.89±0.24%
Relation Networks	64-96-128-256	57.34±0.29%	63.32±0.31%	61.77±0.13%	67.32±0.70%
MedOptNet-KernelSVM	64-96-128-256	68.62±0.62%	72.23±0.45%	66.70±0.32%	71.47±0.47%
MAML	ResNet-12	52.73±0.51%	59.16±0.72%	51.72±0.15%	58.91± 0.31%
Reptile	ResNet-12	56.20±1.84%	64.11±0.27%	52.13±0.57%	58.62±0.66%
Matching Networks	ResNet-12	63.46±0.62%	75.45±0.33%	59.61±0.55%	63.49±0.41%
Prototypical Networks	ResNet-12	61.68±0.30%	70.20±0.66%	61.41±0.88%	72.92±0.14%
Relation Networks	ResNet-12	63.41±0.82%	67.32±0.70%	63.48±0.23%	70.46±0.70%
MedOptNet-RR	ResNet-12	68.72±0.21%	70.23±0.45%	66.40±0.32%	78.41±0.22%
MedOptNet-SVM-WW	ResNet-12	76.45±0.53%	80.12±0.68%	68.34±0.32%	78.12±0.71%
MedOptNet-SimMSVM	ResNet-12	**78.32±0.10%**	81.12±0.24%	69.11±0.56%	78.32±0.34%
MedOptNet-KernelSVM	ResNet-12	77.41±0.67%	**81.23±0.75%**	**70.40±0.64%**	**79.37±0.43%**

TABLE 2: The 3-way, 1-shot and 3-shot Classification results on ISIC2018 and Pap smear. Average few-shot classification accuracies with 95% confidence intervals on ISIC2018 and Pap smera meta-test splits. "RR" stands for ridge regression model.

Model	Backbone	ISIC2018 3-way		Pap smear 3-way	
		1-shot	3-shot	1-shot	3-shot
MAML	ResNet-12	51.37±0.41%	58.12±0.75%	63.74±0.43%	69.21± 0.34%
Reptile	ResNet-12	53.20±0.75%	57.21±0.26%	61.23±0.57%	70.34±0.36%
Matching Networks	ResNet-12	53.46±0.33%	60.45±0.71%	61.91±0.45%	69.31±0.41%
Prototypical Networks	ResNet-12	61.58±0.38%	71.20±0.32%	65.41±0.88%	79.92±0.14%
Relation Networks	ResNet-12	63.41±0.82%	67.32±0.70%	73.48±0.23%	82.46±0.70%
MedOptNet-RR	ResNet-12	72.52±0.21%	79.23±0.45%	66.40±0.32%	78.41±0.22%
MedOptNet-SVM-WW	ResNet-12	72.13±0.34%	79.32±0.38%	**80.55**±0.75%	84.17±0.78%
MedOptNet-SimMSVM	ResNet-12	74.88±0.32%	81.12±0.24%	69.11±0.56%	78.32±0.34%
MedOptNet-KernelSVM	ResNet-12	**74.32±0.72%**	**83.37±0.49%**	79.87±0.84%	**84.32±0.43%**

three data augmentation methods: RandomCrop, ColorJitter for changing image attributes, and RandomHorizontalFlip. For the kernel function support vector machine, we use the Sigmoid kernel function. In the meta-learning stage, the support set adopts a 3-way 15-shot configuration, while the query set uses 3-way 15-shot configuration. During the meta-testing phase, we evaluate our model using 3-way 1-shot and 3-way 3-shot on the support set, and 3-way 15-shot on the query set. We set the regularization parameter of SVM to 0.1. For the prototype network, we measure the distance between feature vectors using the square Euclidean distance. Furthermore, Fig.3 displays the loss curve of MedOptNet on the validation set. By illustrating the relationship between the number of training iterations (epochs) and the corresponding loss values, the loss plot aids in evaluating the performance of the machine learning model. By analyzing the loss plot, informed decisions can be made to adjust hyperparameters (such as learning rate, batch size, and regularization strength) to improve the model's performance.

Experimental results and analysis. According to Table 1, the models can be divided into two categories: baseline methods (including MAML, Reptile, Matching Networks, Prototypical Networks, and Relation Networks) and MedOptNet methods (including MedOptNet-RR, MedOptNet-SVM-WW, MedOptNet-SimMSVM, and MedOptNet-KernelSVM). Among the baseline methods, Prototypical Networks and Relation Networks perform bet-

TABLE 3: Compare different convolutional neural networks and classifiers on the Breakhis dataset. "64-96-128-256" denotes a 4-layer convolutional network with 64, 96, 128 and 256 filters in each layer. "RR" stands for ridge regression model.

Model	Backbone	BreakHis 40X 3-way		BreakHis 100X 3-way	
		1-shot	3-shot	1-shot	3-shot
Matching Networks	64-96-128-256	0.030(s)	0.036(s)	0.030(s)	0.037(s)
Prototypical Networks	64-96-128-256	0.016(s)	0.020(s)	0.017(s)	0.020(s)
Relation Networks	64-96-128-256	0.035(s)	0.042(s)	0.033(s)	0.044(s)
MedOptNet-KernelSVM	64-96-128-256	0.047(s)	0.051(s)	0.047(s)	0.055(s)
Matching Networks	ResNet-12	0.103(s)	0.116(s)	0.110(s)	0.116(s)
Prototypical Networks	ResNet-12	0.99(s)	0.102(s)	0.097(s)	0.099(s)
Relation Networks	ResNet-12	0.101(s)	0.110(s)	0.102(s)	0.113(s)
MedOptNet-RR	ResNet-12	0.103(s)	0.117(s)	0.103(s)	0.120(s)
MedOptNet-SVM-WW	ResNet-12	0.094(s)	0.102(s)	0.096(s)	0.105(s)
MedOptNet-SimMSVM	ResNet-12	0.095(s)	0.104(s)	0.097(s)	0.105(s)
MedOptNet-KernelSVM	ResNet-12	0.104(s)	0.110(s)	0.107(s)	0.114(s)

TABLE 4: Compare different convolutional neural networks and classifiers on the ISIC2018 and Pap smear datasets. "RR" stands for ridge regression model.

Model	Backbone	ISIC2018 3-way		Pap smear 3-way	
		1-shot	3-shot	1-shot	3-shot
Matching Networks	ResNet-12	0.120(s)	0.132(s)	0.102(s)	0.106(s)
Prototypical Networks	ResNet-12	0.098(s)	0.102(s)	0.102(s)	0.103(s)
Relation Networks	ResNet-12	0.119(s)	0.125(s)	0.121(s)	0.133(s)
MedOptNet-RR	ResNet-12	0.127(s)	0.133(s)	0.123(s)	0.132(s)
MedOptNet-SVM-WW	ResNet-12	0.0951(s)	0.103(s)	0.097(s)	0.104(s)
MedOptNet-SimMSVM	ResNet-12	0.101(s)	0.112(s)	0.104(s)	0.114(s)
MedOptNet-KernelSVM	ResNet-12	0.118(s)	0.131(s)	0.128(s)	0.145(s)

ter in most cases. These methods make use of the distance information between samples, enabling the model to better learn discriminative features between classes. Among the baseline methods, MAML and Reptile perform relatively poorly. This indicates that these two methods are not suitable for medical small-sample classification tasks, especially at higher magnifications (such as 100X). In all experimental settings, the MedOptNet methods generally outperform the baseline methods. Due to the more suitable optimization algorithm and model structure design for medical small-sample classification tasks, MedOptNet methods have stronger performance in medical small-sample classification tasks. Among the MedOptNet methods, the MedOptNet-KernelSVM model performs best in most cases. It is the reason that there are more samples for each category in the 3-shot setting, which requires more computation cost. This is because it introduces a nonlinear mapping in the feature space, allowing the model to capture more complex classification boundaries. For both baseline methods and MedOptNet methods, models using ResNet-12 as the backbone perform better overall than models using 64-96-128-256 as the backbone. This is because ResNet-12 has stronger representation learning ability, enabling the model to capture more image features. The experimental results on the BreakHis dataset show that the MedOptNet methods have superior performance in few-shot classification tasks, especially when using ResNet-12 as the backbone. Among the baseline methods, Prototypical Networks and Relation Networks perform relatively well. In addition, the model performance at 100X magnification is generally better than that at 40X magnification, indicating that higher magnifications may help improve model performance on these tasks.

According to Table 2, the 3-way, 1-shot, and 3-shot classification results of various models on the ISIC201

and Pap smear datasets can be seen. All models use ResNet-12 as the backbone. The models can be divided into two categories: baseline methods (including MAML, Reptile, Matching Networks, Prototypical Networks, and Relation Networks) and MedOptNet methods (including MedOptNet-RR, MedOptNet-SVM-WW, MedOptNet-SimMSVM, and MedOptNet-KernelSVM). On the ISIC2018 3-way classification task, all MedOptNet methods outperform the baseline methods in 1-shot and 3-shot tasks. Among them, MedOptNet-KernelSVM achieved an accuracy of 74.32$\pm$0.72% in the 1-shot task, and the highest accuracy of 83.37$\pm$0.49% in the 3-shot task. On the Pap smear 3-way classification task, MedOptNet-SVM-WW achieved the highest accuracy of 80.55$\pm$0.75% in the 1-shot task, and MedOptNet-KernelSVM achieved the highest accuracy of 84.32$\pm$0.43% in the 3-shot task. This indicates that MedOptNet methods are also superior to the baseline methods in this task. Among the baseline methods, Prototypical Networks and Relation Networks perform relatively well in the 1-shot and 3-shot tasks of both datasets. This indicates that these two methods are relatively more competitive in these two tasks. It can be concluded that the performance of MedOptNet methods in the 3-way 1-shot and 3-shot classification tasks on the ISIC2018 and Pap smear datasets is superior to that of baseline methods, with MedOptNet-KernelSVM performing the best.

According to Table 3, the single-episode runtime of different convolutional neural networks and classifiers on the Breakhis dataset can be analyzed. For the 64-96-128-256 convolutional network structure, the Prototypical Networks have the shortest runtime, while the MedOptNet-KernelSVM has the longest runtime. This indicates that Prototypical Networks have an advantage in computational efficiency, while MedOptNet-KernelSVM requires more computational resources. The runtime of other models falls between these two. For the ResNet-12 convolutional network structure, the 1-shot runtime of Prototypical Networks is the shortest, while the 3-shot runtime of MedOptNet-RR is the longest. This implies that the computational efficiency of different models varies on different datasets. For the same model, the 3-shot runtime is generally slightly longer than the 1-shot runtime. It is the reason that there are more samples for each category in the 3-shot setting, which requires more computation cost. When using ResNet-12 as the backbone, the runtime of all models is longer than that with the 64-96-128-256 structure. This is due to ResNet-12 being a more complex network structure, requiring more computational resources.

According to Table 4, the single-episode runtime of different models on the ISIC2018 and Pap smear datasets can be analyzed. For the ResNet-12 structure, Prototypical Networks have the shortest runtime on the ISIC2018 dataset, while MedOptNet-SVM-WW has the shortest runtime on the Pap smear dataset. This indicates that Prototypical Networks and MedOptNet-SVM-WW have advantages in computational efficiency. On the ISIC2018 dataset, the 3-shot runtime of Relation Networks and MedOptNet-KernelSVM is relatively longer, which may imply that these models require more computational resources when processing this dataset. For the same model, the 3-shot runtime is generally slightly longer than the 1-shot runtime. This is because there are more samples for each category in the 3-shot setting, which requires more computation.

In medical few-shot classification tasks, the MedOptNet method overall outperforms the baseline methods, especially when using ResNet-12. Among them, MedOptNet-KernelSVM performs the best, thanks to the design of optimization algorithms and model structure, as well as the introduction of non-linear mapping in the feature space to capture more complex classification boundaries. For baseline methods, Prototypical Networks and Relation Networks perform well in most cases, as they utilize distance information between samples, allowing the model to better learn discriminative features between classes. However, the performance of MAML and Reptile methods on medical few-shot classification tasks is relatively poor, suggesting that these two methods may not be well-suited for such tasks.Using the ResNet-12 model overall performs better than models using simple convolutional network structures, as ResNet-12 has stronger representation learning capabilities, allowing the model to capture more image features. In terms of computational efficiency, Prototypical Networks and MedOptNet-SVM-WW have certain advantages on different datasets. Meanwhile, when each category has more samples, the model requires more computation, and therefore the runtime will be relatively longer.The MedOptNet method has strong performance on medical few-shot classification tasks, and among the baseline methods, Prototypical Networks and Relation Networks perform well. Higher magnification levels may help improve the model's performance on these tasks. In practical applications, suitable methods and models can be selected according to task requirements and computational resource constraints.

Influence of various regularization methods. As shown in Table 5, to demonstrate the impact of the various regularization methods on classification accuracy a 3-way, 1-shot and a 3-way, 3-shot meta-tests were conducted on the on the ISIC2018 datset for the classification of ablation images. In addition, we compared the effects of linear SVM and kernel SVM, and found that kernel SVM outperforms the simple basic classifier in general, which proved the effectiveness of the model.

TABLE 5: **Ablation study.** Various regularization techniques improves the accuracy of 3-way experiments on ISIC2018 datset. "Data Aug" stand for data augmentation.

Data Aug	Weight Decay	Kernel function	1-shot	3-shot
Yes	No	No	71.34	80.34
Yes	Yes	No	73.21	81.54
Yes	Yes	Yes	74.32	83.37

5 Conclusions

In this paper, we propose a convex optimization solver-based meta-learning method (MedOptNet) applied to medical datasets, using multiclassification and functional support vector machines as classifiers, with better generalization capability than nearest-neighbor classifiers, with a slight increase in computational cost. We have used several convex optimization models as classifiers and performed

experiments on several medical datasets, and MedOptNet achieves optimal performance. We also introduce various regularization methods to prevent overfitting of the model.Potential future research directions for few-shot medical image classification based on convex optimization models are as follows. Improving the model robustness and generalization: Although the proposed model demonstrated good performance in few-shot situations, it still has limitations in complex and variable medical scenarios. Therefore, future research can explore methods to improve the robustness and generalization of the model by incorporating specific factors relevant to the data, such as specialized fine-tuning or incremental learning. Strengthening the model's interpretability and data visualization: Given the importance of interpretability and data visualization in medical imaging applications, future work could investigate using interpretable models for feature extraction or combining image visualization to better understand the model's classification process and improve the model's credibility in practical applications. Expanding to other medical imaging analysis domains: While the proposed method focused on medical image classification, further research could explore expanding the method to other medical imaging analysis domains such as medical image segmentation, registration, and so on, to improve the practicality and efficiency of medical imaging analysis. Overall, the few-shot medical image classification based on convex optimization model displays potential for future research and application. Efforts in multiple areas could lead to further improvements to the model's performance and usability.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[4] X. Zhou, W. Liang, I. Kevin, K. Wang, and L. T. Yang, "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 171–178, 2020.

[5] W. Liang, Y. Hu, X. Zhou, Y. Pan, I. Kevin, and K. Wang, "Variational few-shot learning for microservice-oriented intrusion detection in distributed industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5087–5095, 2021.

[6] X. Zhou, X. Xu, W. Liang, Z. Zeng, S. Shimizu, L. T. Yang, and Q. Jin, "Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1377–1386, 2021.

[7] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[10] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin, "Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1815–1824, 2022.

[11] O. Maier, B. H. Menze, J. Von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, *et al.*, "Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri," *Medical image analysis*, vol. 35, pp. 250–269, 2017.

[12] T. Nguyen, M. Narwani, M. Larson, Y. Li, S. Xie, H. Pfister, D. Wei, N. Shavit, L. Mi, A. Pacureanu, *et al.*, "The xpress challenge: Xray projectomic reconstruction–extracting segmentation with skeletons," *arXiv preprint arXiv:2302.03819*, 2023.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[15] X. Zhou, X. Xu, W. Liang, Z. Zeng, and Z. Yan, "Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart iot," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12588–12596, 2021.

[16] X. Zhou, W. Liang, S. Shimizu, J. Ma, and Q. Jin, "Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5790–5798, 2020.

[17] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *Ieee transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[18] T. Munkhdalai and H. Yu, "Meta networks," in *International Conference on Machine Learning*, pp. 2554–2563, PMLR, 2017.

[19] M. Abdullah Jamal, G.-J. Qi, and M. Shah, "Task-agnostic meta-learning for few-shot learning," *arXiv e-prints*, pp. arXiv–1805, 2018.

[20] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[21] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.

[22] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11719–11727, 2019.

[23] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," *Advances in neural information processing systems*, vol. 31, 2018.

[24] B. Amos and J. Z. Kolter, "Optnet: Differentiable optimization as a layer in neural networks," in *International Conference on Machine Learning*, pp. 136–145, PMLR, 2017.

[25] G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, Lille, 2015.

[26] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[27] E. Triantafillou, R. Zemel, and R. Urtasun, "Few-shot learning through an information retrieval lens," *arXiv preprint arXiv:1707.02610*, 2017.

[28] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[29] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[30] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018.

[31] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, and J. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8642–8649, 2019.

[32] B. Zhang, X. Li, Y. Ye, Z. Huang, and L. Zhang, "Prototype completion with primitive knowledge for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3754–3762, 2021.

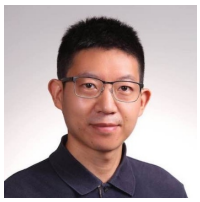
[33] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10657–10665, 2019.
[34] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.
[35] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017.
[36] T. Wei, J. Hou, and R. Feng, "Fuzzy graph neural network for few-shot learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2020.
[37] H. Yao, C. Zhang, Y. Wei, M. Jiang, S. Wang, J. Huang, N. Chawla, and Z. Li, "Graph few-shot learning via knowledge transfer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 6656–6663, 2020.
[38] V. Prabhu, A. Kannan, M. Ravuri, M. Chaplain, D. Sontag, and X. Amatriain, "Few-shot learning for dermatological disease diagnosis," in *Machine Learning for Healthcare Conference*, pp. 532–552, PMLR, 2019.
[39] X. Li, L. Yu, Y. Jin, C.-W. Fu, L. Xing, and P.-A. Heng, "Difficulty-aware meta-learning for rare disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 357–366, Springer, 2020.
[40] R. Singh, V. Bharti, V. Purohit, A. Kumar, A. K. Singh, and S. K. Singh, "Metamed: Few-shot medical image classification using gradient-based meta-learning," *Pattern Recognition*, vol. 120, p. 108111, 2021.
[41] X. Han, J. Wang, S. Ying, J. Shi, and D. Shen, "Ml-dsvm+: A meta-learning based deep svm+ for computer-aided diagnosis," *Pattern Recognition*, vol. 134, p. 109076, 2023.
[42] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of machine learning research*, vol. 2, no. Dec, pp. 265–292, 2001.
[43] A. Agrawal, S. Barratt, and S. Boyd, "Learning convex optimization models," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 8, pp. 1355–1364, 2021.
[44] G. Gordon and R. Tibshirani, "Karush-kuhn-tucker conditions," *Optimization*, vol. 10, no. 725/36, p. 725, 2012.
[45] J. Weston and C. Watkins, "Multi-class support vector machines," tech. rep., Citeseer, 1998.
[46] X. He, Z. Wang, C. Jin, Y. Zheng, and X. Xue, "A simplified multi-class support vector machine with reduced dual optimization," *Pattern Recognition Letters*, vol. 33, no. 1, pp. 71–82, 2012.
[47] J. Zou, X. Ma, C. Zhong, and Y. Zhang, "Dermoscopic image analysis for isic challenge 2018," *arXiv preprint arXiv:1807.08948*, 2018.
[48] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," *Nature inspired Smart Information Systems (NiSIS 2005)*, pp. 1–9, 2005.

Liangfu Lu received the B.S. and M.S. degrees in computational mathematics from Ludong University and Nanjing University of Aeronautics and Astronautics, China in 2001 and 2004, respectively, and his Ph.D. in computer science from Tianjin University, China in 2008. He worked as a visiting scholar in the University of California, Los Angeles (UCLA), USA, from Dec. 2016 to July. 2017. He is currently an associate professor in Academy of Medical Engineering and Translational Medicine, Tianjin University. His research interests include tensor analysis, optimization algorithms in AI and image processing etc. He has published over 50 papers in top journals and conference proceedings and principally investigated some research projects Natural Science Foundation of China.

Xudong Cui received the B.S.degree from TianJin Normal University, TianJin China in 2020. He is currently pursuing the M.S. degree in the School of Mathematice, TianJin University, TianJin, China. His current research interests include meta learning and few-shot learning.

Zhiyuan Tan is an Associate Professor with the School of Computing, Engineering and the Built Environment, Edinburgh Napier University, UK. He received his Ph.D. degree from the University of Technology Sydney, Australia, in 2014, and was a Postdoctoral Researcher with the University of Twente, NL between 2014 and 2016. His research focus on Cybersecurity, Machine Learning, Cognitive Computation. He is an Associate Editor of IEEE Transactions on Reliability and the Journal of Ambient Intelligence and Humanized Computing, as well as an Academic Editor of Security and Communication Networks. He is a Senior Member of the IEEE and a Member of the ACM.



Yulei Wu is a Senior Lecturer with the Department of Computer Science, Faculty of Environment, Science and Economy, University of Exeter, UK. He received the B.Sc. degree in Computer Science and the Ph.D. degree in Computing and Mathematics from the University of Bradford, UK, in 2006 and 2010, respectively. His main research interests include network digital twins, AI-based networks, connected systems, and edge intelligence. He is an Associate Editor of IEEE Transactions on Network and Service Management and IEEE Transactions on Network Science and Engineering, as well as an Editorial Board Member of Computer Networks and Future Generation Computer Systems. He is a Senior Member of the IEEE and the ACM.