Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19316-3

MTUNet + +: explainable few-shot medical image classification with generative adversarial network

Ankit Kumar Titoriya[1] · Maheshwari Prasad Singh[1] · Amit Kumar Singh[1]

Received: 9 March 2024 / Revised: 15 April 2024 / Accepted: 27 April 2024

Abstract

Medical imaging, a cornerstone of disease diagnosis and treatment planning, faces the hurdles of subjective interpretation and reliance on specialized expertise. Deep learning algorithms show improvements in automating medical image analysis, reducing radiologists' burden, and potentially enhancing patient outcomes. However, these algorithms require substantial quantities of high-quality labelled data for effective training and refinement. This paper proposes an innovative approach that harnesses few-shot learning (FSL) and generative adversarial networks (GANs) to overcome conventional methods' limitations in medical image classification. FSL, capable of learning from limited labelled examples, holds promise for scenarios where labelled data is scarce. However, the lack of interpretability in existing FSL models impedes their clinical adoption. To tackle this, this paper proposes a explainable FSL network, "MTUNet + +," which integrates an attention mechanism to emphasize relevant regions in medical images. Furthermore, integrating a generative adversarial network, enhances the performance of MTUNet + + by generating synthetic medical images. Systematically eliminating misleading synthetic images improves the reliability and accuracy of medical image classification. Empirical evaluation on benchmark datasets underscores the effectiveness of the approach, achieving 85.19% and 69.28% accuracy for the HAM10000 and Kvasir datasets, respectively. This paper contributes to advancing AI-driven solutions in clinical practice, facilitating enhanced patient care and streamlined workflows within real-world healthcare settings.

Keywords Few-shot Learning · Medical Imaging · Generative Adversarial Networks · HAM10000 · Kvasir · Explainable Artificial Intelligence

✉ Ankit Kumar Titoriya
ankitt.ph21.cs@nitp.ac.in

Maheshwari Prasad Singh
mps@nitp.ac.in

Amit Kumar Singh
amit.singh@nitp.ac.in

[1] Department of Computer Science and Engineering, National Institute of Technology, Patna, India

Published online: 08 May 2024

1 Introduction

The integration of artificial intelligence (AI) into healthcare has emerged as a critical imperative in addressing the increasing complexity and challenges within the medical field [1]. The rapid expansion of medical data and the increasing demand for accurate and efficient diagnostic methods position AI as a pivotal force in transforming healthcare delivery, especially in medical imaging. Medical imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), and X-rays play fundamental roles in disease diagnosis, treatment planning, and patient progress monitoring. However, the interpretation of medical images often requires specialized expertise and is prone to subjectivity, which may lead to potential diagnostic errors and delays in treatment. Deep learning has emerged as a powerful tool for automating the analysis of medical images and improving the accuracy and efficiency of diagnoses. However, its effectiveness depends on a large amount of labelled data for training. The lack of labelled data in medical fields presents a significant challenge, hindering the complete utilization of machine learning algorithms for diagnostic tasks. Therefore, deep learning may not always be the most reliable choice for medical image analysis and diagnosis. Deep learning may show less than optimal accuracy in classifying rare diseases. Its performance may differ among various datasets due to difficulties in generalizing within deep learning models.

In recent years, few-shot learning (FSL) has emerged as a promising approach to medical image classification, particularly in scenarios where labelled data is scarce or unavailable. FSL aims to train machine learning models to recognize patterns from a small number of labelled examples, or support images, and generalize to classify new, unseen images, or query images. This paradigm closely mimics the human ability to learn new concepts from a limited number of examples, making it well-suited for medical image analysis, where labelled data is often limited and expensive to obtain. By leveraging FSL, researchers and healthcare professionals can potentially improve the efficiency and accuracy of diagnosing medical conditions from imaging data. Despite the potential of FSL in medical image classification, existing models often lack interpretability and transparency, hindering their adoption in clinical practice. Deep learning models, which are commonly used in FSL approaches, are often perceived as black boxes, making it difficult for healthcare professionals to understand the rationale behind their decisions. Explainable artificial intelligence (XAI) aims to address this challenge by providing transparent insights into the decision-making process of AI models, enabling clinicians to validate and trust the decisions made by these models. XAI methods like feature visualization, saliency maps, and attention mechanisms can assist healthcare professionals in identifying the specific regions of an image that are impacting the model's decision-making process. This not only increases confidence in the model's predictions but also allows clinicians to provide better patient care by integrating AI technology into their workflow [2].

The integration of generative adversarial networks (GANs) into medical image analysis holds significant promise for enhancing the quality and diversity of training data. GANs, a class of AI algorithms that generate synthetic data by learning from real data distributions, can produce realistic medical images that closely resemble real patient data. Augmenting the limited labelled datasets available to FSL with synthetic data generated by GANs may improve the generalization performance and robustness of FSL models to changes in imaging conditions. Furthermore, the integration of GANs may also aid in addressing the issue of data scarcity in medical imaging by generating large amounts of realistic data for training [3]. This may help improve the accuracy and reliability of FSL models for detecting and diagnosing medical conditions. While GANs can generate realistic data, there may still be limitations in the ability of synthetic data to fully capture the complexity and variability of real patient data. Additionally, there is a risk of introducing bias or inaccuracies into the FSL models if the synthetic data does not accurately reflect the true distribution of medical images. To tackle this, this paper introduces an approach that uses two convolutional neural networks (CNNs) to remove misleading synthetic images from the training dataset before training the FSL models. GANs is a versatile tool that can be applied in a wide range of domains beyond image generation. For example, it can be employed in recommendation systems [38, 39] to generate synthetic embeddings from content features, thereby improving recommendation accuracy and addressing challenges in recommending cold items. By incorporating GANs into the process of training FSL models, researchers can ensure that the synthetic data used is both accurate and representative of the true distribution of medical images, ultimately improving the performance and reliability of the FSL models in medical image analysis tasks.

In summary, the integration of AI into healthcare, particularly in medical image analysis, offers transformative opportunities to improve patient care, enhance diagnostic accuracy, and streamline clinical workflows. Advanced computational techniques such as FSL and GANs may be leveraged to develop innovative solutions to address the complex challenges faced by healthcare professionals in diagnosing and treating diseases. This paper introduces a novel approach to explainable few-shot medical image classification using generative adversarial networks (GANs), aiming to enhance the trustworthiness and adoption of AI-driven solutions in clinical practice. This paper demonstrates the effectiveness and interpretability of the proposed approach through empirical evaluation and validation. This facilitates its practical implementation in real-world healthcare environments. The key contributions of this paper are outlined as follows:

- This paper proposes "MTUNet+ +," a novel explainable few-shot medical image classification network exhibiting superior accuracy compared to current state-of-the-art methods. The proposed network integrates an attention mechanism, emphasizing relevant regions in medical images for classification. Notably, the attention mechanism incorporates learnable positional embedding and gated recurrent unit with a skip connection, diverging from traditional approaches employing fixed positional embedding and gated recurrent unit architectures.
- This paper proposes leveraging a generative adversarial network to enhance MTUNet+ +'s performance. It integrates FASTGAN [4] to generate additional medical images from available training data. Employing two CNNs, the approach identifies synthetic images prone to misclassification, removing them from the dataset. This systematic elimination aims to improve the accuracy and reliability of the classification task, showcasing the potential of FSL techniques in medical image classification.

2 Related work

In this section, the literature on few-shot learning is initially addressed. Following this, the literature on explainable artificial intelligence is examined. Finally, the literature on generative adversarial networks is discussed.

2.1 Few-shot learning

FSL, a subfield of machine learning, addresses the challenge of training models with limited labelled data. Initial methods for FSL primarily utilized meta-learning strategies, training models that quickly adapted to new tasks utilizing limited data. Vinyals et al. [6] introduce matching networks, which utilize attention mechanisms to compare support set examples with query examples. Ravi et al. [7] introduce the concept of learning to learn, which involves utilizing a recurrent neural network to acquire shared initialization for multiple tasks. Snell et al. [8] introduce prototypical networks, which are designed to learn class prototypes to improve classification efficiency. Finn et al. [9] present Model-Agnostic Meta-Learning (MAML), which learns an initialization that enables fast adaptation to new tasks. Nichol et al. [10] streamline the meta-learning process with Reptile by employing a minimal number of inner loop updates during training. Relational reasoning models, such as Relation Network (Sung et al. [11]), explicitly represent relationships between pairs of instances to enhance classification. Gidaris et al. [12] present Dynamic Few-Shot Visual Learning, which adjusts feature embeddings dynamically to enhance classification performance. Cross-modal FSL is now a significant area of research. Xing et al. [13] introduce Cross-Modal Few-Shot Learning, a method that enhances classification by sharing knowledge across different modalities. Hu et al. [14] propose transductive meta-learning, which involves using unlabelled data in meta-training to improve generalization. Zhou et al. [15] introduce Few-Shot Deep Learning via Information Bottleneck, leveraging information bottleneck theory to improve few-shot classification performance. Wang et al. [16] introduce meta-prototypical networks, which combine meta-learning with prototypical networks to enhance few-shot learning. Gauch et al. [17] introduce a method called SubGD, which discovers low-dimensional parameter subspaces in stochastic gradient descent updates. Wang et al. [26] suggest a few-shot classification method that uses a two-stream neural network to pull out features and adds circle loss to the loss function. This leads to higher accuracy rates on standard datasets. Wang et al. [37] investigate the efficacy of nearest-neighbour baselines without meta-learning in few-shot learning, revealing competitive accuracies with basic feature transformations. Mix-MAML is a hybrid optimization meta-learning method that Jia et al. [27] have created to improve MAML's limited generalization performance by adding techniques for data enhancement, initialization attenuation, and resolution enhancement. Zhang et al. [28] propose RaPSPNet, a network tailored for few-shot fine-grained image classification, enhanced feature distinction, and similarity evaluation. While the existing literature has made remarkable progress in advancing few-shot learning techniques, there is still a significant research gap in terms of scalability, real-world applicability, generalization across diverse domains, as well as interoperability and integration.

2.2 Explainable artificial intelligence

XAI has garnered substantial attention in recent years as a pivotal element in constructing trustworthy and interpretable machine learning models. Initial XAI methods concentrated on model-agnostic techniques that explain predictions of machine learning models without needing access to their internal mechanisms. Ribeiro et al. [18] introduce LIME (Local Interpretable Model-Agnostic Explanations) as a method that explains the predictions of any classifier in a human-interpretable way. It generates local surrogate models to approximate the behavior of the black-box model near a specific prediction. Lundberg et al. [19] propose

SHAP (SHapley Additive exPlanations), which leverages cooperative game theory to assign feature importance scores to individual input features. Recently, there has been an increasing interest in incorporating XAI methods into the training phase to improve model transparency and interpretability. This trend has resulted in the creation of interpretable neural network structures like attention mechanisms and self-attention mechanisms. Vaswani et al. [20] present the Transformer architecture, which uses self-attention mechanisms to capture distant relationships in sequential data and offers interpretability by using attention weights. Wang et al. [21] introduce Score-CAM, a novel visual explanation method for CNNs, surpassing previous approaches with gradient-independent weight determination for improved interpretability. Selvaraju et al. [22] propose Grad-CAM. It creates visual explanations for decisions in CNN-based models by highlighting important image regions for different tasks without changing the architecture. Chattopadhay et al. [23] introduce Grad-CAM++, an enhanced method building on Grad-CAM, aiming to improve visual explanations of CNN predictions significantly. Li et al. [24] suggest SCOUTER, a slot attention-based classifier that offers classification by adding explanations directly to final confidence scores, which makes them easier to understand. Sun et al. [25] introduce a new way to train models for cross-domain few-shot classification tasks that uses explanation scores to highlight important features in real time. This makes the models more generalized across a wider range of datasets. Wang et al. [5] propose a novel FSL approach for image classification, utilizing a visual representation from the backbone model and patterns generated by a self-attention based explainable module to enhance transparency in knowledge transfer. While recent advancements in XAI have demonstrated promising results in enhancing the interpretability of machine learning models and improving model generalization across various domains, there are still several critical gaps and considerations that require attention. Future research should focus on developing robust, versatile, and user-centric XAI methods that are computationally efficient, maintain high performance, and consider XAI's broader ethical and societal implications.

2.3 Generative adversarial networks

GANs have emerged as a powerful tool in medical imaging, particularly for image generation, segmentation, and enhancement tasks. The application of GANs in medical imaging gained momentum with the pioneering work of Goodfellow et al. [29]. They introduce the concept of adversarial training. In the context of medical imaging, GANs have been employed for various tasks such as image generation, image-to-image translation, and anomaly detection. They employ Parzen window-based log-likelihood estimates to figure out the performance of the GAN network on MNIST and TFD datasets. Nie et al. [30] propose a data-driven approach, using adversarial training and an image gradient difference loss, improving accuracy in predicting Computed tomography (CT) images from magnetic resonance imaging (MRI). Iqbal et al. [31] propose MI-GAN, which synthesizes retinal images and segmented masks for STARE and DRIVE datasets. Beers et al. [32] suggest a GAN network that generates synthetic medical images, including fundus photographs of retinopathy of prematurity and multi-modal MRI images of glioma. Ren et al. [33] suggest using GANs to create real medical image stimuli. They have created tumor-like stimuli with specific shapes and authentic textures in a controlled manner. Liu et al. [4] introduce a lightweight GAN structure for few-shot image synthesis, achieving high quality at 1024 × 1024 resolution with minimal computing resources. They use the Frechet Inception Distance metric to measure the overall semantic realism of synthesized images. Joseph et al. [34] propose a GAN-based data augmentation model to generate synthetic mammograms, addressing class imbalances in the MIAS dataset and leading to improved breast cancer classification performance. They use CNN classifiers, including binary and ternary classifiers, to evaluate the performance of their GAN network. While GANs have shown promise in creating realistic medical images, there may be some concerns regarding the use of synthetic data for medical diagnosis and treatment decisions. Additionally, the reliability and accuracy of GAN-generated images compared to real patient data may still be a concern in clinical settings.

In conclusion, recent advancements in FSL and XAI have shown promise for enhancing model adaptation and interpretability across diverse domains. Additionally, GANs offer valuable tools for medical imaging tasks, although concerns remain regarding their reliability and accuracy in clinical settings. Therefore, this paper proposes an explainable FSL network tailored for medical image classification, integrating it with GAN. This integrated approach aims to address the limitations of current models by providing both accuracy and interpretability in medical image classification.

3 Method

In this section, the problem definition is initially discussed, along with the proposed few-shot medical image classification network. Furthermore, the generation of medical images through a generative adversarial network is discussed.

3.1 Problem definition

In the context of classification tasks, the training set D represents a collection of paired instances (x_i, y_i), where x_i denotes the input data and y_i signifies the corresponding output label. The fundamental objective is to derive a functional mapping $f : X \rightarrow Y$, where X denotes the input space and Y signifies the output space, thereby enabling the model to accurately assign class labels to new inputs. In few-shot learning (FSL), on the other hand, the training dataset D is very limited, with only a few examples of each class. This is different from the usual assumption of a lot of examples in supervised learning methods. A typical approach in FSL is characterized by an M-way N-shot strategy, where M denotes the number of classes and N signifies the number of samples per class. This strategy works in an episodic manner. The algorithm parses the whole dataset into various episodes. A single episode consists of two distinct sets: a support set and a query set. The support set D_s encompasses a set of labelled instances, formulated as $D_s = \left\{\left(x_{mn}, y_{mn}\right) | m = 1, 2, 3, \dots M, n = 1, 2, 3, \dots N\right\}$ while the query set D_q contains unlabelled instances belonging to classes present in the support set but distinct from those in the support set. The basic purpose of a Few-Shot Learning (FSL) algorithm is to acquire knowledge from the support set to make predictions for instances in the query set.

3.2 Few-shot medical image classification network

Figure 1 illustrates the proposed MTUNet+ +network, which is designed for few-shot image classification. It is the enhanced version of [5]. During each episode, feature maps $F_{map} = f_{\phi}(x) \in \mathbb{R}^{a \times h \times w}$ are extracted from each image x in the support set D_s and the query image using the CNN backbone f_{ϕ}, where ϕ represents the set of learnable parameters.

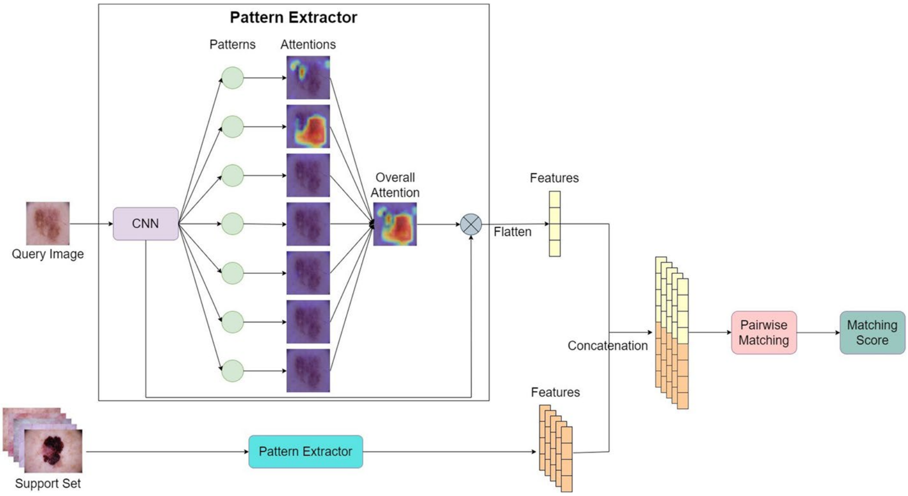

Fig. 1 The structure of MTUNet++. A single query image undergoes processing by the pattern extractor to derive distinctive patterns, which are then aggregated into an overall attention mechanism. Then, features of each query image are combined with features of every support image by concatenation, which enables final classification through pairwise matching

These feature maps are then fed into the pattern extractor module f_{pe}, which generates attention $Att = f_{pe}(F_{map}) \in \mathbb{R}^{u \times v}$ over F_{map}. The pairwise matching module utilizes a Multi-Layer Perceptron (MLP) to determine a score indicating the probability of the query image x_q belonging to one of the M classes in D_s. The Pattern extractor module is essential in the learning process for FSL tasks. It is designed with a specific focus on developing a transferable attention mechanism that can effectively recognize and analyse common patterns that are present in the dataset.

Figure 2 illustrates the structure of the proposed pattern extractor module. It is the enhanced version of [24]. First, the input feature map F_{map} goes through a 1×1 convolutional layer, which is then followed by a Rectified Linear Unit (ReLU) activation function. This process is intended to reduce the dimensionality of F_{map} from a to b. Afterwards, the spatial

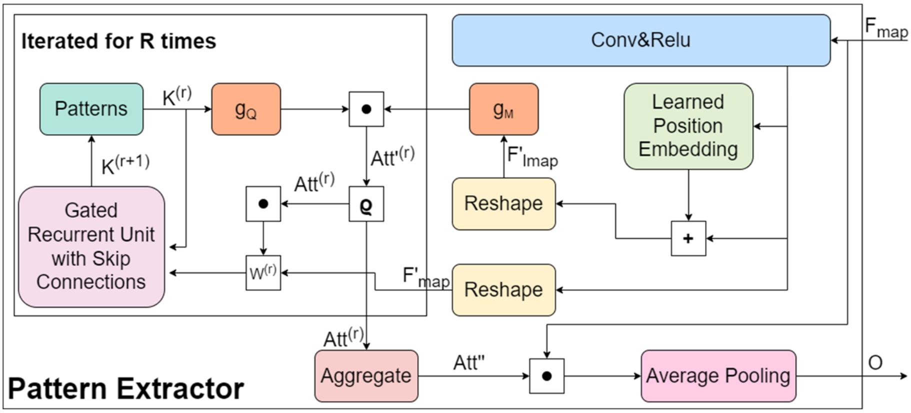

Fig. 2 The structure of proposed pattern extractor

dimension of the reduced features is flattened to construct $F_{map}' \in \mathbb{R}^{b\times v}$, where $v = h \times w$. To maintain spatial information, a learnable positional embedding P_l is integrated with the features, represented as $F_{map}' = F_{map}' + P_l$. By utilizing a self-attention mechanism, attention is allocated over F_{map} concerning spatial dimensions. This is done by calculating the dot-product similarity between a set of u patterns and F_{map}' post nonlinear transformations. This process is iterated R times, where patterns are updated using a Gated Recurrent Unit with Skip Connections (GRU_{sc}) to improve attention. Let $K^{(r)} \in \mathbb{R}^{u\times b}$ represent the patterns in the r-th iteration, where $r = 1, 2, \dots, R$ and $K^{(1)} = K$ denotes the learnable parameters. Nonlinear transformations for $K^{(r)}$ and F_{map}' are denoted by

$$g_q\left(K^{(r)}\right) \in \mathbb{R}^{u\times b}, g_M\left(F_{map}'\right) \in \mathbb{R}^{b\times v} \dots\dots \tag{1}$$

The attention is applied by utilizing a normalization function ϱ, as follows:

$$Att'^{(r)} = g_q\left(K^{(r)}\right) g_M\left(F_{map}'\right) \dots\dots \tag{2}$$

$$Att^{(r)} = \varrho\left(Att'^{(r)}\right) \in \mathbb{R}^{u\times v} \dots\dots \tag{3}$$

The patterns are updated through

$$W^{(r)} = Att^{(r)} F_{map}'^{R} \dots\dots \tag{4}$$

$$K^{(r+1)} = GRU_{sc}\left(W^{(r)}, K^{(r)}\right) \dots\dots \tag{5}$$

Let *Softmax* represent a softmax function and let σ represent a sigmoid function. MTU-Net+ +adjusts the attention map through

$$Att^{(r)} = \varrho\left(Att'^{(r)}\right) = \sigma\left(Att'^{(r)}\right) \odot Softmax\left(Att'^{(r)}\right) \dots\dots \tag{6}$$

This function effectively diminishes weak attention across various patterns sharing the same spatial location, utilizing $\odot$ Hadamard product. This mechanism pushes the network to find very specific and different patterns with less duplication, which improves the precise allocation of attention. The attention map predominantly focuses on individual patterns, typically excluding their peripheral regions. The input feature map F_{map} is then defined by the comprehensive attention Att'' that corresponds to the extracted patterns.

$$Att'' = \frac{1}{u} Att^{(r)} 1_u \dots\dots \tag{7}$$

The vector 1_u represents a row vector wherein all u elements are aggregated to yield 1. The attention Att'' is reshaped from v into a spatial structure identical to that of F_{map}. Subsequently, features corresponding to the comprehensive attention are extracted and subjected to average pooling across spatial dimensions, denoted as:

$$O = \frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} Att''_{ij} F_{map_{ij}} \dots\dots \tag{8}$$

The process of performing a Few-Shot Learning (FSL) classification task typically entails establishing a similarity between a query image and one of the support images provided. The proposed network utilizes a learnable distance framework that includes a MLP.

This MLP takes the feature vectors of the query and support images as input and outputs a similarity score. Let O_q and O_{mn} denote the feature extracted by applying the pattern extractor to the query image $x_q \in D_Q$ and support images $x_{mn} \in D_s$. Here the subscripts $m = 1, 2, 3 \ldots\ldots M$ and $n = 1, 2, 3 \ldots\ldots N$ denote the n^{th} image in the M-way N-shot episodic paradigm. In scenarios where $N > 1$, an average over the N images is computed the generate the representative feature $\overline{O_m}$; otherwise $\overline{O_m} = O_{m1}$. To compute the similarity score *score* between O_q and $\overline{O_m}$, a MLP f_θ with learnable parameters θ is employed:

$$score\left(O_q, \overline{O_m}\right) = \sigma\left(f_\theta\left(\left[O_q, \overline{O_m}\right]\right)\right) \ldots\ldots \quad (9)$$

Here [., .] denotes concatenation. The similarity score is then used to measure the similarity between the query image and each support image. By comparing the scores, the network can determine which support image is most similar to the query image and classify it accordingly. The query image x_q is assigned to class m^*, with the highest similarity scores over m classes.

$$m^* = \underset{m}{\mathrm{argmax}}\, score\left(O_q, \overline{O_m}\right) \ldots\ldots \quad (10)$$

This learnable distance framework enhances the performance of Few-Shot Learning by effectively capturing the underlying similarities and differences between images.

3.3 Generation of medical images through generative adversarial network

This study employs FASTGAN [4] for generating synthetic images in the training dataset. FASTGAN employs a single convolution layer per resolution with restricted channels, yielding a smaller and quicker-to-train model. Moreover, it integrates the Skip-Layer Excitation (SLE) module, enhancing gradient flow by modifying skip-connections for efficient gradient signal propagation across resolutions. SLE utilizes channel-wise multiplications and extends skip-connections between distant resolutions to improve gradient flow without notable computational overhead. While resembling the Squeeze-and-Excitation module, SLE operates between distant feature-maps, aiding gradient flow and channel-wise feature recalibration essential for disentangling content and style attributes in generated images. FASTGAN also incorporates a self-supervised discriminator trained with small decoders, enhancing image feature extraction via auto-encoding. This approach improves the discriminator's capacity to extract comprehensive representations from inputs, thereby enhancing model robustness and synthesis quality. This approach maintains a pure GAN framework, using auto-encoding solely for discriminator regularization. Additionally, the method employs the hinge version of the adversarial loss for iterative training of the discriminator and generator. Overall, these methodological advancements contribute to more efficient and effective GAN training, which has implications for various image synthesis tasks.

Figure 3 illustrates the integration of the GAN with the few-shot classifier network. This paper initially undertakes the generation of synthetic medical images by using original images. Considering that some synthetic medical images may not be completely realistic or accurately categorized, the possibility of inaccuracies necessitates manual examination by domain experts for correction. To circumvent the need for expert intervention, the study employs two CNNs, namely EfficientNetV2 and ShuffleNetV2, trained on the original medical images for the accurate classification of synthetic images. Consequently, only the

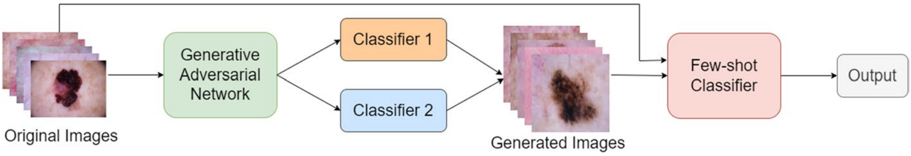

Fig. 3 Illustration of the integration of the GAN with the few-shot classifier network

synthetic images correctly classified by these CNNs are utilized. These mutually classified images, along with the original medical images, form the training set for the Few-Shot Learning (FSL) classifier. This approach ensures the selection of high-quality synthetic images for inclusion in the training dataset, thereby enhancing the performance and reliability of the FSL classifier. By leveraging the capabilities of pretrained CNNs, the study establishes a mechanism to automate the classification process. It mitigates the need for manual intervention and facilitates the seamless integration of synthetic images into the training pipeline.

4 Experiments

4.1 Dataset

This paper employs two distinct datasets to evaluate the effectiveness of the proposed method in medical image classification. The first dataset comprises dermoscopic images of skin cancer, while the second dataset encompasses gastrointestinal endoscopy images for gastrointestinal disorders. The HAM10000 dataset, abbreviated as "Human Against Machine with 10,000 Training Images" [35], is a comprehensive collection of dermoscopic images of pigmented skin lesions, consisting of 10,015 images. The dataset encompasses dermatoscopic images sourced from diverse populations and acquired through various modalities, resulting in a comprehensive collection suitable for academic machine learning purposes. Over a span of 20 years, the Department of Dermatology at the Medical University of Vienna, Austria, and Cliff Rosendahl's skin cancer practice in Queensland, Australia, collected the 10,015 dermatoscopic images in the HAM10000 dataset. This dataset covers a diverse range of benign and malignant lesions, including 'Actinic Keratoses', 'Basal cell carcinoma', 'Benign Keratosis', 'Dermatofibroma', 'Melanocytic Nevi', 'Melanoma', and 'Vascular Skin Lesions'. Furthermore, histopathology confirms over 50% of the lesions in the HAM10000 dataset. Expert consensus, follow-up examination, or in-vivo confocal microscopy established the ground truth for the remaining cases. The dataset is a valuable resource for developing and evaluating algorithms for dermatology and skin cancer detection. The Kvasir dataset [36] comprises gastrointestinal endoscopy images, facilitating research in medical image classification for gastrointestinal disorders. Vestre Viken Health Trust (VV) in Norway, consisting of four hospitals providing healthcare to 470,000 people, collected the data using endoscopic equipment. The dataset includes eight classes: 'Normal Z-line', 'Pylorus', 'Cecum', 'Esophagitis', 'Polyps', 'Ulcerative Colitis', 'Dyed and Lifted Polyps', and 'Dyed Resection Margins'. Medical experts from VV and the Cancer Registry of Norway (CRN) manually annotated the images. Some of the included image classes feature a green picture-in-picture illustration of the endoscope's position and configuration inside the bowel, facilitated by an electromagnetic imaging system (ScopeGuide, Olympus Europe), which may aid in interpreting the image. The dataset offers a diverse set of images representing various gastrointestinal conditions. This makes it an appropriate dataset for training and evaluating medical image classification models, specifically for gastrointestinal diseases.

4.2 Experimental Setup

The backbone CNN is initially trained using a task-based method on the utilized medical dataset. Subsequently, the attention module is trained independently, and finally, the few-shot classifier is trained to optimize its performance for medical image classification tasks. This sequential training approach is adopted to ensure that each component of the network is effectively trained to perform its specific task, thereby enhancing the overall performance of the model. The proposed network (MTUNet++) is assessed through experimentation involving 2000 episodes of 2-way classification. The first step in this evaluation process is to randomly pick two classes from the dataset. Next, support and query images from these 2 classes are randomly picked, with the number of support images ($N = 5$ or 10) and the fifteen query images per class. The average accuracy is calculated for ($15 \times 2 = 30$) query images from all 2000 episodes. This paper adopts ResNet-18 as the backbone for MTUNet+ +, utilized in Few-Shot Learning (FSL) tasks. Changes are made to the ResNet-18 architecture. The first two downsampling layers are left out, and the kernel size of the first 7×7 convolutional layer is changed to 3×3. After the Rectified Linear Unit (ReLU) activation, the feature maps used in the model are taken from the hidden vector of the last convolutional layer. For ResNet-18, the respective numbers of feature maps amount to 512. In the pre-training phase of the pattern extractor module, the hyperparameter hidden dimension for GRU_{sc} is configured to 256, with 3 update iterations. The number of patterns is set to seven. Both the g_q and g_M consist of three fully connected layers with ReLU nonlinearities. All parameters within the backbone (ResNet-18) remain fixed. The initial learning rate for training commences at 10^{-4} and undergoes a reduction by a factor of 10 at the 40^{th} epoch, with a total training duration spanning 150 epochs. To train the proposed network, MTUNet+ +, the CNNs and pattern extractor parameters are fine-tuned using a slow learning rate of 10^{-5} over 20 iterations. Each epoch involves the sampling of 500 episodes for 2-way tasks. The remaining trainable components of the model commence training with an initial learning rate of 10^{-4}, divided at the 10th epoch by 10. The model selection process involves saving the model demonstrating the best performance after evaluation on 2,000 episodes sampled from the validation set. The AdaBelief algorithm facilitates optimization as part of the model's implementation using PyTorch. Input images undergo resizing to dimensions of 80×80, alongside data augmentation techniques such as random flipping and affine transformations. Experiments were run on a GPU workstation with one NVIDIA A4000 GPU (16 GB of GDDR6 memory) and a Xenon Gold 6226R CPU and 64 GB of RAM.

4.3 Results and Discussion

For image classification, the HAM10000 and Kvasir datasets are used for training and testing the proposed architecture. The HAM10000 dataset comprises 7 classes of skin lesions, while Kvasir dataset contains 8 classes of gastrointestinal tract diseases. The distribution of both datasets is depicted in Figs. 4 and 5. Additionally, 1000 synthetic medical images are added to each class to evaluate the proposed model's performance

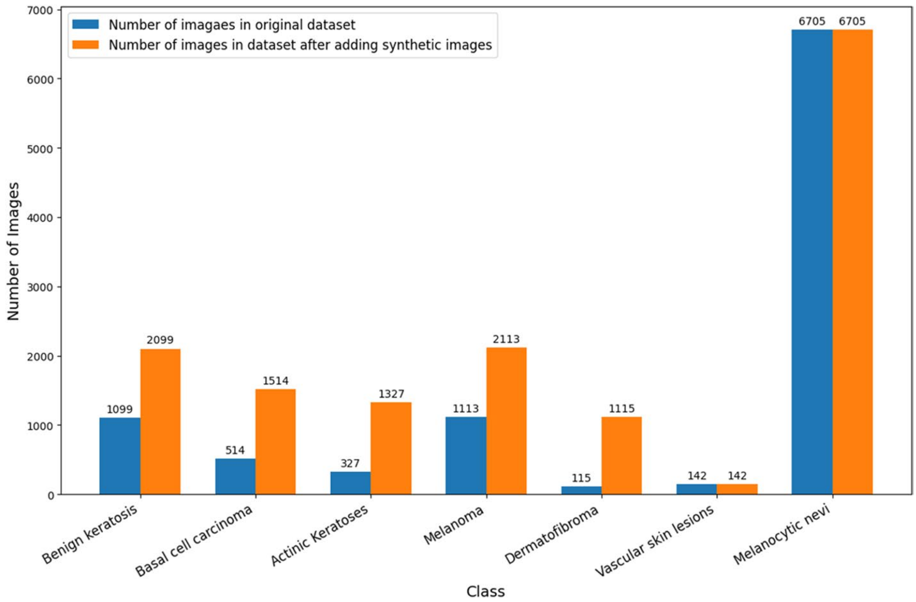

Fig. 4 Data distribution plot of HAM10000 dataset

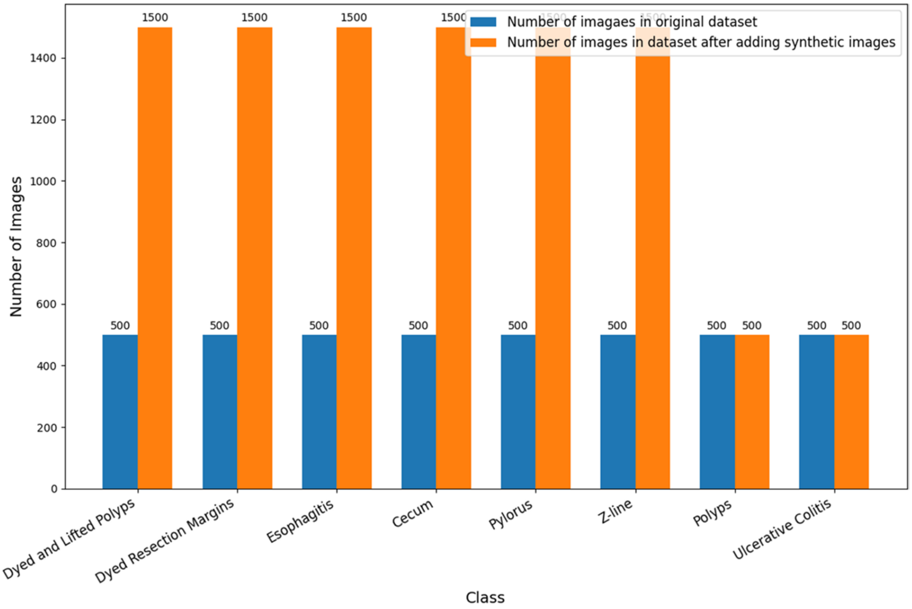

Fig. 5 Data distribution plot of Kvasir dataset

with GAN. In the HAM10000 dataset, ’Benign keratosis’, ’Basal cell carcinoma’, and ’Actinic Keratoses’ constitute the training set, while ’Melanoma’ and ’Dermatofibroma’ are used for validation. ’Vascular skin lesions’ and ’Melanocytic nevi’ are designated for the test dataset. In the Kvasir dataset, 'Dyed and Lifted Polyps', 'Dyed Resection Margins', and 'Esophagitis' are utilized for training, with 'Cecum', 'Pylorus', and 'Z-line' reserved for validation. 'Polyps' and 'Ulcerative Colitis' are allocated for the test dataset. Thus, only two-way classification is possible for both the datasets.

Figure 6 presents the depiction of original images alongside generated synthetic images belonging to classes from both the training and validation sets of the HAM10000 dataset. Similarly, Fig. 7 illustrates original images and generated synthetic images from classes within the training and validation sets of the Kvasir dataset. The synthetic images are generated using a FASTGAN model to increase the diversity of the dataset and improve the performance of the classification model. By incorporating synthetic images, the model can better generalize to new, unseen data and improve overall accuracy.

GANs are widely utilized in generating synthetic images that mimic the characteristics of the original images. The GAN framework typically employs a discriminator network to distinguish between these synthetic images and the original images. However, due to the inherent limitations in the discriminator's discriminative capabilities, it is not always feasible for the GANs to accurately classify all synthetic images. Given the critical nature of medical imaging, manual verification by domain experts is often indispensable to ensuring the quality and accuracy of the synthetic images. Incorporating these synthetic images directly into the training dataset can introduce additional noise, compromising the model's accuracy and potentially rendering it less useful in medical applications. This constraint presents a significant challenge for the integration of GANs with limited medical data.

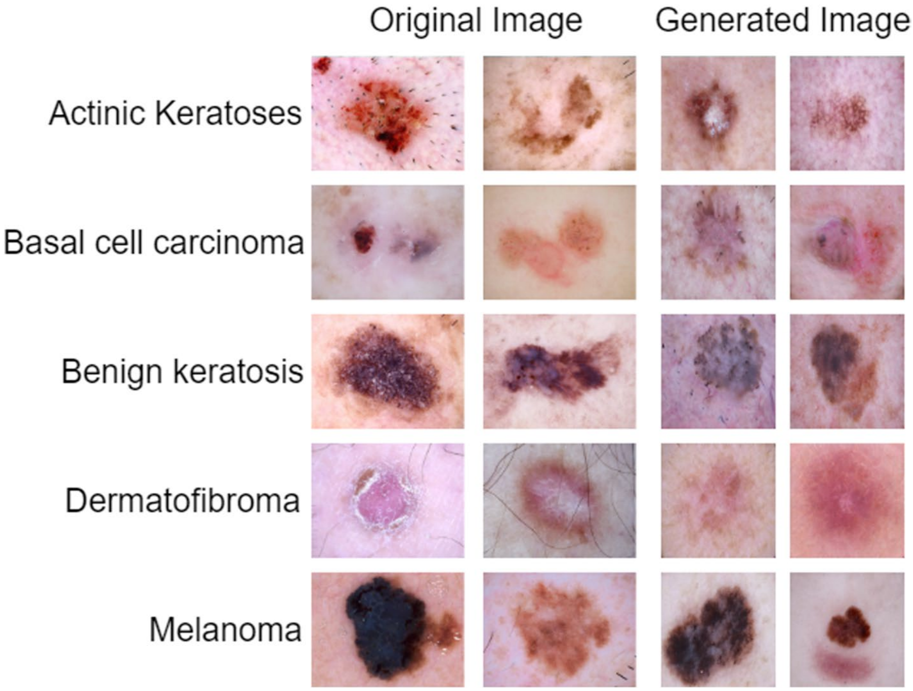

Fig. 6 Illustration of original images and generated synthetic images of the HAM10000 dataset

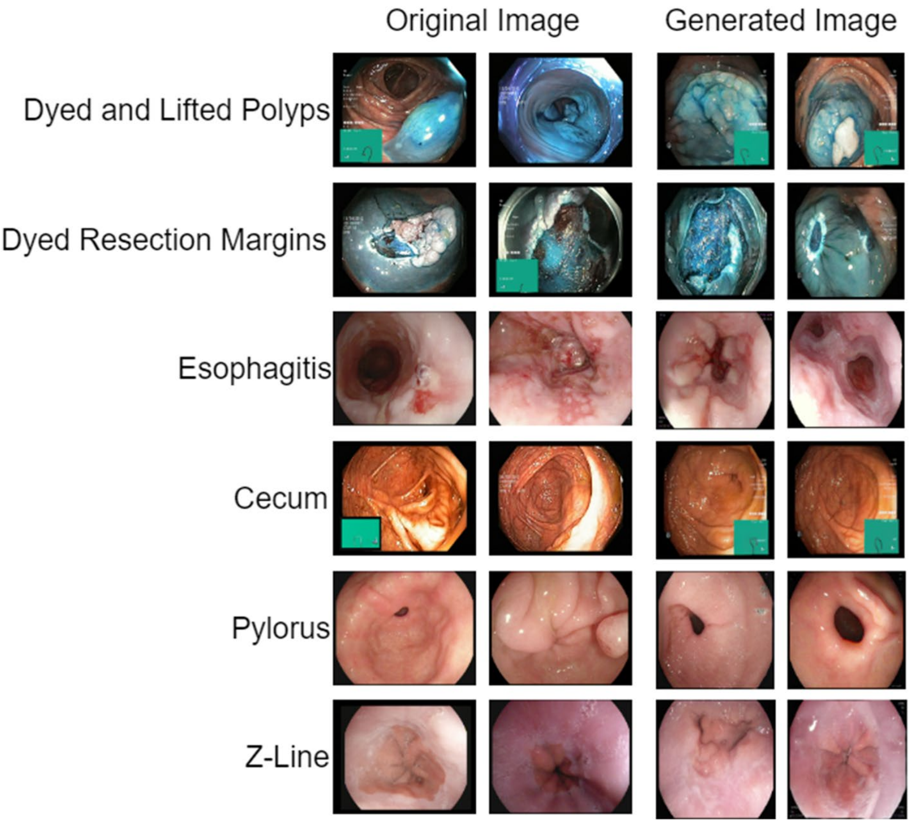

Fig. 7 Illustration of original images and generated synthetic images of the Kvasir dataset

Although a trained expert has the potential to handle this issue, the medical community frequently lacks such experts. Additionally, relying solely on expert verification can be time-consuming and may not be scalable for large datasets. To address this challenge, this paper proposes a novel approach that leverages two CNNs trained on the original dataset to classify the synthetic images. This paper utilizes EfficientNetV2 and ShuffleNetV2, pre-trained on the original medical images, for the accurate classification of synthetic images. The newly created dataset only includes the synthetic images that both CNNs mutually classify as correct. This selective approach ensures that the dataset remains noise-free and includes only correctly generated synthetic images, thereby mitigating the limitations associated with the integration of GANs in medical applications.

The proposed network is compared with state-of-the art methods in terms of accuracy. The proposed network achieves superior performance and outperforms all other methods. For model testing, the results are based on the model that did the best on the validation dataset. This was done by picking 2000 (1-shot, 5-shot and 10-shot) tasks at random from the test dataset across both datasets. During testing, the model assigns query images to one of the support image classes by extracting regions from each query and support image, deriving features with the pattern extractor, and subsequently matching these features with the pattern matching module. Table 1 and Table 2

Table 1 Average accuracy of 2000 episodes of 2-way tasks on the HAM10000 dataset test set

Approach	2w-1 s-15q	2w-5 s-15q	2w-10 s-15q
Simpleshot [37]	51.14	64.21	66.18
MTUNet [5]	75.08	65.77	78.22
MTUNet+ +	76.29	68.64	80.9
MTUNet+ +with GAN	**80.71**	**84.35**	**85.19**

Table 2 Average accuracy of 2000 episodes of 2-way tasks on the Kvasir dataset test set

Approach	2w-1 s-15q	2w-5 s-15q	2w-10 s-15q
Simpleshot [37]	51.14	58.04	64.21
MTUNet [5]	51.62	51.86	53.46
MTUNet+ +	53.27	54.03	55.01
MTUNet+ +with GAN	**62.51**	**63.49**	**69.28**

present the obtained results for the HAM10000 and Kvasir datasets, respectively. The results show that MTUNet+ +with GAN achieved higher accuracy in both datasets compared to other existing methods. The pattern extractor and matching module in MTUNet+ +along with GAN make feature extraction and matching more reliable. This results in better performance in few-shot learning tasks on the HAM10000 and Kvasir datasets.

Figure 8 provides a visual representation of individual patterns alongside the average features extracted from a sampled task within the HAM10000 dataset. This figure displays two images from the support set and two images from the query set. Notably, the illustration comprises 1–7 images, each portraying the pattern extracted from corresponding slots. The proposed network utilizes a pattern extractor with seven slots. Furthermore, the term 'Overall' encapsulates the average pattern visualization, ranging from 1 to 7 patterns, which is integral to the few-shot classification process. Similarly, Fig. 9 showcases the visualization of each pattern and the average features derived from a sampled task within the Kvasir dataset.

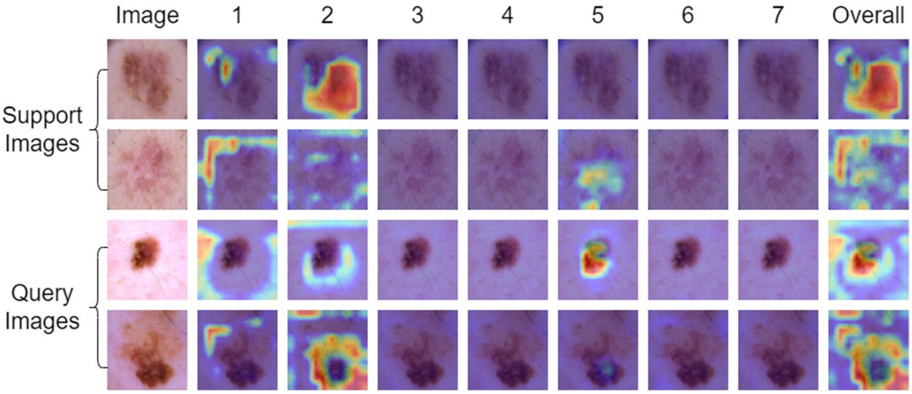

Fig. 8 Visualization of each pattern and the average features from a sampled task in the HAM10000 dataset

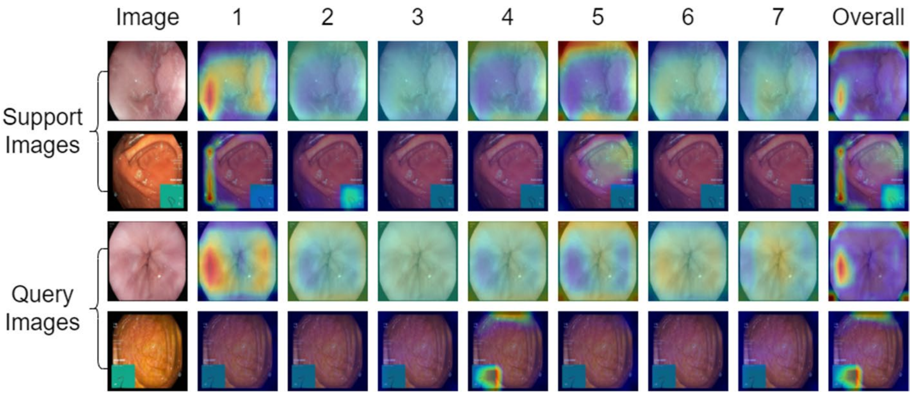

Fig. 9 Visualization of each pattern and the average features from a sampled task in the Kvasir dataset

4.4 Ablation study

Spatial relationships between pixels or regions are pivotal for accurate modeling in computer vision tasks such as image classification and object detection. Position embedding plays a crucial role in capturing spatial context in visual data by encoding positional information. This paper employs learned position embedding as opposed to fixed position embedding to enhance the performance of the Few-Shot Learning (FSL) network in image classification. Fixed-position embedding involves assigning predetermined positional vectors to each location or region within the image grid. These vectors remain static during training and are not adjusted based on the input data. While fixed position embedding is computationally efficient and straightforward to implement, it may not fully capture the complex spatial relationships present in the data. Furthermore, the fixed positional embeddings eliminate the risk of overfitting. However, fixed positional embeddings may not capture specific patterns and structures in the data as effectively as learned positional embeddings, potentially resulting in information loss as they do not adapt to the specific task or dataset. In contrast, the learned position embedding approach treats position embedding vectors as trainable parameters. These vectors are adjusted during training to enable the model to dynamically learn spatial relationships from the input data. This adaptability allows the model to capture intricate spatial contexts more effectively, potentially enhancing performance in tasks where spatial relationships are crucial. Furthermore, in learned positional embedding, the model can adapt the positional embeddings to the specific patterns and structures in the data it is trained on. However, there are also disadvantages to learned positional embeddings, including the risk of overfitting, especially when the model is trained on a small dataset. Furthermore, learning positional embeddings requires additional computational resources and may increase training time. In medical images, where the spatial arrangement of features can be critical for accurate diagnosis, this adaptability of learned embeddings proves beneficial. Table 3 displays the average accuracy of 2000 two-way tasks on the HAM10000 dataset test set, with both fixed and learned position embeddings. The proposed network attains an accuracy of 68.64% when utilizing learned position embeddings, surpassing the performance achieved with fixed position embedding. To train the learned position embeddings, the number of epochs has been

Table 3 Mean accuracy of 2000 episodes of 2-way 5-shot 15-query tasks on the HAM10000 dataset test set with various position embeddings

Position embedding	2w-5 s-15q
Fixed position embedding	66.48
Learned position embedding	**68.64**

increased from 60 to 150. The attention mechanism in the proposed FSL network can effectively leverage the learned positional embeddings to focus on relevant regions or features within the medical images. The attention mechanism can prioritize the most informative spatial contexts, which is critical for accurate medical image classification. The learned positional embeddings facilitate the attention mechanism's adaptability, enhancing the model's ability to identify subtle patterns and structures in the medical images, thereby improving classification accuracy.

Adding skip connections to Gated Recurrent Unit (GRU) networks improves the flow of information and makes it easier for gradients to propagate through the network. These skip connections can assist in learning long-term dependencies by establishing a direct path for gradient flow during backpropagation. This feature aids in mitigating the vanishing gradient problem, which is a prevalent challenge in training deep recurrent neural networks (RNNs). By retaining information from prior time steps and enabling it to bypass certain GRU layers, skip connections allow the model to capture both short-term and long-term dependencies more effectively. However, adding skip connections introduces additional complexity and increases the number of parameters in the model. To effectively manage this increased complexity and ensure optimal utilization of skip connections, selecting an appropriate hidden dimension is essential. The hidden dimension determines the network's capacity to capture and represent the underlying patterns in the data while also accommodating the information flow facilitated by the skip connections. Insufficient hidden dimensions may limit the network's ability to exploit the benefits of skip connections, leading to suboptimal performance or even degradation in performance due to overfitting or underfitting. Therefore, choosing the right hidden dimension is crucial in GRU with skip connections to maintain a balance between model capacity and computational efficiency. Table 4 presents the mean accuracy results obtained from 2000 episodes consisting of 2-way 5-shot tasks with 15 queries each, conducted on the HAM10000 dataset's test set. These experiments were conducted using GRU with skip connections with varying numbers of hidden dimensions. The proposed network achieves its highest accuracy when configured with a hidden dimension value set at 256. Table 5 further illustrates the effectiveness of GRU with skip connections by presenting the mean accuracy of 2000 episodes of

Table 4 Mean accuracy of 2000 episodes of 2-way 5-shot 15-query tasks on the HAM10000 dataset test set with varying numbers of hidden dimensions in Gated Recurrent Unit with skip connections

Hidden Dimension	2w-5 s-15q
64	64.49
128	66.83
256	**68.64**
512	67.25

Table 5 Mean accuracy of 2000 episodes of 2-way 5-shot 15-query tasks on the HAM10000 dataset test set with GRU and GRU with skip connections

Network	2w-5 s-15q
GRU	66.18
GRU with skip connections	**68.64**

2-way 5-shot 15-query tasks on the HAM10000 dataset test set with a hidden dimension set at 256. The proposed network achieves the highest accuracy when GRU with skip connections is employed.

4.5 Limitations

A significant limitation of the proposed network lies in its inability to effectively classify histopathological images when integrated with GANs. While MTUNet+ +demonstrates advancements in classification compared to MTUNet, the inclusion of synthetic images generated by FASTGAN diminishes the performance of the proposed network. Specifically, FASTGAN fails to produce realistic histopathological images, posing a significant challenge to achieving optimal classification results. Furthermore, histopathological images require a substantial input size within the network to preserve the intricate details inherent in histopathological images. Consequently, employing a smaller input size compromises the retention of crucial information within the images. So, more research needs to be done to make it easier for GANs to make realistic histopathological images, which would then help classification networks like MTUNet+ +work better. Advanced GAN architectures, specially designed for complex medical images, may be beneficial in addressing this issue. Additionally, exploring ways to effectively handle the large input size required for histopathological images without losing important details will be crucial for advancing the accuracy and efficiency of such classification systems in the future.

Numerous images in the Kvasir dataset feature the scope configuration for the cecal position, which is a common occurrence across various images. In the proposed network, the pattern extractor identifies this common pattern within images of the same class and allocates attention to this green component, thereby impeding the network's performance. Figure 9 illustrates this phenomenon. The prevalent green element in images across diverse classes introduces noise into the classification process, thereby hindering accurate categorization. To address this issue, future research may focus on implementing advanced data augmentation techniques, such as Cutout, Mixup, or AutoAugment, to diversify the dataset and reduce the impact of common patterns, such as the green component in the Kvasir dataset. By introducing variations in the images, the network will be less likely to rely solely on these common patterns for classification. Additionally, experimentation with modifications to the attention mechanism to focus on other relevant features instead of distracting common patterns may enhance the overall efficacy and accuracy of the network. Another technique of self-attention mechanisms or adaptive attention models could be explored to improve the network's ability to capture and prioritize relevant features in the images.

5 Conclusion

This paper proposed a novel few-shot classification network with explainable artificial intelligence within the context of medical image classification. The network demonstrated better performance compared to existing methods, especially when using GAN-generated images. It achieved an accuracy of 85.19% for the HAM10000 dataset and 69.28% for the Kvasir dataset in 2000 episodes (2-way, 10-shot, 15-query). The network's increased precision in the 1-shot scenario underscores its potential for practical medical uses, especially in situations where there is a scarcity of labelled data. This paper has shown how using synthetic images created by GAN can enhance dataset diversity and improve the performance of classification models. The network's spatial context comprehension has been improved by adopting learned position embedding instead of fixed embedding, resulting in enhanced classification capabilities. GRU with skip connections has facilitated improved information flow and gradient propagation, particularly with carefully selected hidden dimensions. However, challenges persist, particularly regarding the accurate classification of images containing common patterns, such as the green component observed in the Kvasir dataset. Further investigation is needed to address these challenges. Future research could focus on using advanced data augmentation techniques and improving attention mechanisms to reduce their effects and enhance network performance. Furthermore, visualizing patterns in the network's decision-making process has offered valuable insights into its operations, improving transparency and interpretability. In medical settings, these features are extremely valuable because they are essential for building trust and encouraging acceptance among medical professionals due to their explainability. By addressing the challenges outlined and continuing to innovate in this field, researchers can advance the reliability and usability of medical image classification models, ultimately enhancing patient care and outcomes.

Author Contributions Ankit Kumar Titoriya contributed to conceptualization, methodology, and writing; Prof. Maheshwari Prasad Singh and Dr. Amit Kumar Singh contributed to validation, editing, and supervision. All authors read and approved the final manuscript.

Funding Ankit Kumar Titoriya receives a Ph.D. scholarship from the National Institute of Technology Patna.

Data Availability Not applicable.

Declarations

Competing Interests The authors declare no competing interests regarding this manuscript.

Ethics Approval Not applicable.

Consent to Participate No requirement for informed consent for this study.

Consent to Publish Not applicable.

References

1. Boeken, T, Feydy, J, Lecler, A, Soyer, P, Feydy, A, Barat, M, Duron, L (2023) Artificial intelligence in diagnostic and interventional radiology: where are we now?. Diagnostic and Interventional Imaging, 104. https://doi.org/10.1016/j.diii.2022.11.004
2. Jin W, Li X, Fatehi M, Hamarneh G (2023) Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical Image Analysis 84:102684. https://doi.org/10.1016/j.media.2022.102684

3. Zhou T, Li Q, Lu H, Cheng Q, Zhang X (2023) GAN review: models and medical image fusion applications. Inf Fusion 91. https://doi.org/10.1016/j.inffus.2022.10.017
4. Liu B, Zhu Y, Song K, Elgammal A (2020) Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In: International Conference on Learning Representations
5. Wang B, Li L, Verma M, Nakashima Y, Kawasaki R, Nagahara H (2023) Match them up: visually explainable few-shot image classification. Appl Intell 53(9):10956–10977. https://doi.org/10.1007/s10489-022-04072-4
6. Vinyals O, Blundell C, Lillicrap T, Wierstra D (2016) Matching networks for one shot learning. Adv Neural Inf Process Syst 29
7. Ravi S, Larochelle H (2016) Optimization as a model for few-shot learning. In: International conference on learning representations
8. Snell J, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning. Adv Neural Inf Process Syst 30
9. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning, pp 1126–1135
10. Nichol A, Schulman J (2018) Reptile: a scalable metalearning algorithm. arXiv preprint arXiv:1803.02999
11. Sung F, Yang Y, Zhang L, Xiang T, Torr PH, Hospedales TM (2018) Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1199–1208
12. Gidaris S, Komodakis N (2018) Dynamic few-shot visual learning without forgetting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4367–4375
13. Xing C, Rostamzadeh N, Oreshkin BN, Pinheiro PO (2019) Adaptive cross-modal few-shot learning. Adv Neural Inf Process Syst 32
14. Hu SX, Moreno PG, Xiao Y, Shen X, Obozinski G, Lawrence ND, Damianou A (2020) Empirical bayes transductive meta-learning with synthetic gradients. arXiv preprint arXiv:2004.12696. https://doi.org/10.48550/arXiv.2004.12696
15. Zhou L, Liu Y, Zhang P, Bai X, Gu L, Zhou J, Hancock E (2023) Information bottleneck and selective noise supervision for zero-shot learning. Mach Learn 112. https://doi.org/10.1007/s10994-022-06196-7
16. Wang RQ, Zhang XY, Liu CL (2021) Meta-prototypical learning for domain-agnostic few-shot recognition. IEEE Trans Neural Netw Learn Syst 33(11):6990–6996. https://doi.org/10.1109/TNNLS.2021.3083650
17. Gauch M, Beck M, Adler T, Kotsur D, Fiel S, Eghbal-zadeh H, Lehner S (2022) Few-shot learning by dimensionality reduction in gradient space. In: Conference on Lifelong Learning Agents, pp 1043–1064
18. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144. https://doi.org/10.1145/2939672.2939778
19. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
21. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Hu X (2020) Score-CAM: score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 24–25
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
23. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV), pp 839–847. https://doi.org/10.1109/WACV.2018.00097
24. Li L, Wang B, Verma M, Nakashima Y, Kawasaki R, Nagahara H (2021) Scouter: slot attention-based classifier for explainable image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1046–1055
25. Sun J, Lapuschkin S, Samek W, Zhao Y, Cheung NM, Binder A (2020) Explain and improve: cross-domain few-shot-learning using explanations. arXiv preprint arXiv:2007.08790, 1. https://doi.org/10.48550/arXiv.2007.08790
26. Wang J, Song B, Wang D, Qin H (2022) Two-stream network with phase map for few-shot classification. Neurocomputing 472. https://doi.org/10.1016/j.neucom.2021.11.074

27. Jia J, Feng X, Yu H (2024) Few-shot classification via efficient meta-learning with hybrid optimization. Eng Appl Artif Intell 127. https://doi.org/10.1016/j.engappai.2023.107296
28. Zhang W, Zhao Y, Gao Y, Sun C (2024) Re-abstraction and perturbing support pair network for few-shot fine-grained image classification. Pattern Recog 148. https://doi.org/10.1016/j.patcog.2023.110158
29. Goodfellow, I, Pouget-Abadie, J, Mirza, M, Xu, B, Warde-Farley, D, Ozair, S, Bengio, Y (2014) Generative adversarial nets. Advances in neural information processing systems, 27
30. Nie, D, Trullo, R, Lian, J, Petitjean, C, Ruan, S, Wang, Q, Shen, D (2017) Medical image synthesis with context-aware generative adversarial networks. In Medical Image Computing and Computer Assisted Intervention− MICCAI 2017, (pp. 417–425). https://doi.org/10.1007/978-3-319-66179-7_48
31. Iqbal T, Ali H (2018) Generative adversarial network for medical images (MI-GAN). J Med Syst 42. https://doi.org/10.1007/s10916-018-1072-9
32. Beers A, Brown J, Chang K, Campbell JP, Ostmo S, Chiang MF, Kalpathy-Cramer J (2018) High-resolution medical image synthesis using progressively grown generative adversarial networks. arXiv preprint arXiv:1805.03144. https://doi.org/10.48550/arXiv.1805.03144
33. Ren, Z, Stella, XY, Whitney, D (2021) Controllable medical image generation via generative adversarial networks. In IS&T International Symposium on Electronic Imaging (Vol. 33). https://doi.org/10.2352/ISSN.2470-1173.2021.11.HVEI-112
34. Joseph, AJ, Dwivedi, P, Joseph, J, Francis, S, Pournami, PN, Jayaraj, PB, Sankaran, P (2024) Prior-guided generative adversarial network for mammogram synthesis. Biomed Signal Process Control, 87. https://doi.org/10.1016/j.bspc.2023.105456
35. Tschandl, P, Rosendahl, C, Kittler, H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 5. https://doi.org/10.1038/sdata.2018.161
36. Pogorelov, K, Randel, KR, Griwodz, C, Eskeland, SL, de Lange, T, Johansen, D, Halvorsen, P (2017) Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In Proceedings of the 8th ACM on Multimedia Systems Conference (pp. 164–169)
37. Wang Y, Chao WL, Weinberger KQ, Van Der Maaten L (2019) Simpleshot: revisiting nearest-neighbor classification for few-shot learning. arXiv preprint arXiv:1911.04623. https://doi.org/10.48550/arXiv.1911.04623
38. Huang F, Wang Z, Huang X, Qian Y, Li Z, Chen H (2023) Aligning distillation for cold-start item recommendation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1147–1157. https://doi.org/10.1145/3539618.3591732
39. Chen H, Bei Y, Shen Q, Xu Y, Zhou S, Huang W, Huang X (2024) Macro Graph Neural Networks for Online Billion-Scale Recommender Systems. arXiv preprint arXiv:2401.14939. https://doi.org/10.48550/arXiv.2401.14939

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.