electronics

Article

Multi-Scale Fusion MaxViT for Medical Image Classification with Hyperparameter Optimization Using Super Beluga Whale Optimization

Jiaqi Zhao *, Tiannuo Liu and Lin Sun *

College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China; ltn@mail.tust.edu.cn

* Correspondence: 22103419@mail.tust.edu.cn (J.Z.); sunlin@tust.edu.cn (L.S.)

Abstract: This study presents an enhanced deep learning model, Multi-Scale Fusion MaxViT (MSF-MaxViT), designed for medical image classification. The aim is to improve both the accuracy and robustness of the image classification task. MSF-MaxViT incorporates a Parallel Attention mechanism for fusing local and global features, inspired by the MaxViT Block and Multihead Dynamic Attention, to improve feature representation. It also combines lightweight components such as the novel Multi-Scale Fusion Attention (MSFA) block, Feature Boosting (FB) Block, Coord Attention, and Edge Attention to enhance spatial and channel feature learning. To optimize the hyperparameters in the network model, the Super Beluga Whale Optimization (SBWO) algorithm is used, which combines bi-interpolation and adaptive parameter tuning, and experiments have shown that it has a relatively excellent convergence performance. The network model, combined with the improved SBWO algorithm, has an image classification accuracy of 92.87% on the HAM10000 dataset, which is 1.85% higher than that of MaxViT, proving the practicality and effectiveness of the model.

Keywords: Multi-Scale Fusion MaxViT; Beluga Whale Optimization; medical image classification; HAM10000 dataset

Academic Editor: Antoni Morell

Received: 12 February 2025
Revised: 22 February 2025
Accepted: 24 February 2025
Published: 25 February 2025

Citation: Zhao, J.; Liu, T.; Sun, L. Multi-Scale Fusion MaxViT for Medical Image Classification with Hyperparameter Optimization Using Super Beluga Whale Optimization. *Electronics* **2025**, *14*, 912. https://doi.org/10.3390/electronics14050912

1. Introduction

With the wide application of deep learning in medical image classification, hyperparameter optimization and the design of network architecture have become the key factors in improving the performance of the model. Deep neural networks (DNNs) are able to automatically learn rich features from a large amount of image data to assist doctors in making more accurate diagnoses. However, the effectiveness of deep learning not only relies on the network architecture but is also strongly influenced by the hyperparameter settings. Choosing the right hyperparameters enhances both the precision of the classification process and the overall applicability of the model.

Among many deep learning network architectures, the Max Vision Transformer (MaxViT) [1] performs well in vision tasks due to its powerful feature extraction capability. However, MaxViT itself still faces some challenges, such as the lack of fusion of local and global features, the generalization ability of the model, and the insufficient capture of detailed features. Ishak Pacal proposed a lightweight structure, introduced SE blocks after Stem's convolutional layers [2], and replaced traditional MLPs with Global Response Normalization (GRN)-based ones. But striking the best balance between lightweighting and model performance could necessitate numerous experiments and adjustments. Tao et al. proposed combining MaxVit with UNet to construct a cascade segmentation model [3].

However, this model involves relatively high algorithmic complexity and a large number of parameters. Ishak Pacal also developed a cervical cancer detection model by rescaling MaxViT [4], incorporating GRN-based MLP layers, and substituting ConvNeXtv2 blocks for MBConv blocks. Despite these advancements, issues such as inadequate feature fusion persist. Although all the above studies have made some progress in the application of MaxViT, key issues such as feature fusion are still challenges that need to be addressed. To remedy these deficiencies, this study proposes an improved network model based on the MaxViT network, named Multi-Scale Fusion MaxViT (MSF-MaxViT). By innovatively introducing the Parallel Attention module and MSFA module, MSF-MaxViT further enhances the fusion ability of local and global information based on the original MaxViT architecture, which improves the expression ability of image features and the generalization performance of the model.

Meanwhile, the choice of optimization algorithms is crucial for the determination of hyperparameters during the training process of deep neural networks. Although many optimization algorithms have been proposed in recent years, such as Arctic Puffin Optimization (APO) [5], GOOSE Algorithm [6], Red-billed Blue Magpie Optimizer (RBMO) [7], and Phalaropes Kingfisher Optimizer (PKO) [8], etc., and these new algorithms have achieved excellent performance improvement in some aspects, the Beluga Whale Optimization (BWO) algorithm [9] still has the advantage of dealing with complex function optimization and practical engineering applications. BWO simulates the group foraging behaviors of beluga whales, optimizes the parameter configurations through global search and local exploitation, and can strike a better balance between search efficiency and accuracy, especially in the solution of large-scale problems. Researchers have proposed various strategies to enhance its performance. Huang et al. added a dynamic Gaussian variation mechanism to BWO [10], aiming to help the algorithm jump out of the local optimal solution, but the method failed to significantly improve the convergence speed of the algorithm on complex, high-dimensionality problems. Horng and Lin introduced an elite inverse learning strategy (EOBL) and other optimization techniques in order to improve the algorithm's global search ability and stability [11], but the algorithm's convergence is still too slow in dynamic environments. Li et al. proposed a cyclone foraging strategy and elite population updating mechanism for the BWO algorithm's tendency to fall into local optimums [12], but the method may still suffer from the bottleneck of excessive computational resource consumption when dealing with large-scale engineering optimization problems. Yuan et al. proposed a spiral search strategy [13], a random perturbation of the t-distribution, and a cosine boundary adjustment mechanism by introducing a spiral search strategy, a random perturbation of the t-distribution and a cosine boundary adjustment mechanism that enables the population individuals to concentrate on the high potential region more effectively, but this method may still fall into local optimal solutions when facing complex, high-dimensionality optimization problems. Aiming at the above defects and problems, this study proposes the SBWO algorithm, which combines bi-interpolated optimization and adaptive parameter tuning, aiming to improve the performance of BWO in hyper-parameter optimization and to enhance its global exploration ability and local exploitation ability.

This study not only proposes innovative improvements in the MaxViT architecture, but also optimizes the hyperparameters of the MSF-MaxViT network through SBWO. Through experimental validation on the HAM10000 [14] dataset, the model in this study achieves a classification accuracy of 92.87% in the medical image classification task, which outperforms the existing mainstream deep learning models, proving the effectiveness of the MSF-MaxViT network architecture and the SBWO optimization algorithm. The primary contributions presented in this document include the following:

- The SBWO is proposed. To improve the performance of deep learning models in medical image classification, this study introduces the SBWO algorithm based on bi-interpolation optimization [15]. The algorithm combines adaptive parameter tuning and quadratic interpolation optimization to improve the convergence and global search capability of the algorithm. Compared to the traditional BWO, SBWO shows excellent performance in optimizing the hyperparameters of the MaxViT network.
- Combination of Parallel Attention Block based on MaxViT Block improvement and Multi-Scale Fusion Attention Block (MSFA) focusing on enhanced feature representation on MaxViT base model: An improved model, named MSF-MaxViT, is proposed in this paper. This model introduces a novel attention mechanism module, MSFA Block, into the MaxViT network, which enhances the network's generalization ability and feature expression ability by focusing on the combination of edge information enhancement, multi-scale fusion, and coordinate attention [16], which further improves the performance of medical image classification. Meanwhile, the Parallel Attention Block focuses on the dynamic fusion of multiple attention mechanisms of the original MaxViT Block, changing Block Attention and Grid Attention from serial to parallel, which can learn features at different scales and levels.
- Optimized hyperparameters of the MSF-MaxViT model: By optimizing the hyperparameters of MSF-MaxViT using SBWO, the classification accuracy on the HAM10000 dataset was improved to 92.87%. This result demonstrates the effectiveness of the proposed optimization strategy in improving the accuracy of the medical image classification task.
- Innovative and practical: The research in this paper effectively improves the performance of optimization algorithms and network modules. The outcomes of the experiments show that the method performs better in the HAM10000 medical image classification task and has the potential to be further extended to other datasets.

2. SBWO Optimization Algorithm

2.1. BWO Background

The beluga whale is a cetacean that lives in the ocean and is known as the 'canary of the sea' for its remarkable ability to produce a wide range of sounds. Beluga whales have a certain social behavior, often gathering in groups and moving and hunting by sound. They also engage in interactive behaviors, such as S-shaped swimming, releasing air bubbles, and playing with their own kind [17]. Beluga whales are also known as 'polar cuddlers'. Beluga whales are predominantly omnivorous and primarily hunt by sucking prey into their mouths through suction. Inspired by the behavioral characteristics of beluga whales, the original authors proposed a new meta-heuristic algorithm, BWO.

2.1.1. Initial Phase

The BWO algorithm operates on the premise that every beluga whale represents a candidate solution, undergoing constant updates during the optimization phase. The initialization position of the population is as follows:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \quad (1)$$

where n denotes the beluga whale population size, and d denotes variable dimension. The fitness values for all beluga whales are

$$F_X = \begin{bmatrix} f(x_{1,1}, x_{1,2}, \ldots, x_{1,d}) \\ f(x_{2,1}, x_{2,2}, \ldots, x_{2,d}) \\ \vdots \\ f(x_{n,1}, x_{n,2}, \ldots, x_{n,d}) \end{bmatrix} \tag{2}$$

The BWO algorithm can move from exploration to exploitation depending on the balancing factor B_f, which is given by

$$B_f = B_0\left(1 - \frac{T}{2T_{max}}\right) \tag{3}$$

$$W_f = 0.1 - \frac{0.05T}{T_{max}} \tag{4}$$

where T is the current number of iterations, T_{max} is the maximum number of iterations, and B_0 for each iteration varies randomly between (0, 1). The exploration phase occurs when the equilibrium factor occurs, $B_f > 0.5$, while the exploitation phase occurs when $B_f \leq 0.5$, and when the equilibrium is reached, $B_f < W_f$, the whale falling phase begins [18,19]. As the number of iterations T increases, the fluctuation range of the equilibrium factor B_f decreases from (0, 1) to (0, 0.5), indicating that the probability of the exploitation and exploration phases changes significantly, while the probability of the exploitation phase increases with the increase in the number of iterations T.

2.1.2. Global Exploration Phase

During the exploration phase, the algorithm modeled beluga whales swimming in synchronized or mirrored pairs. The behavior was mathematically modeled as

$$\begin{cases} X_{i,j}^{T+1} = X_{i,p_j}^{T} + \left(X_{r,p_1}^{T} - X_{i,p_j}^{T}\right)(1 + r_1)sin(2\pi r_2), & j = even \\ X_{i,j}^{T+1} = X_{i,p_j}^{T} + \left(X_{r,p_1}^{T} - X_{i,p_j}^{T}\right)(1 + r_1)cos(2\pi r_2), & j = odd \end{cases} \tag{5}$$

where T is the current iteration number, $X_{i,j}^{T+1}$ is the new position of the i beluga whale in dimension j, $p_j(j_j = 1, 2, \ldots\ d)$ is a random integer in dimension d, X_{i,p_j}^{T} is the position of the i beluga whale in dimension p_j, X_{r,p_1}^{T} is a randomly selected beluga whale in dimension p_1, r is a random integer, r_1 and r_2 denote a random number in the range of (0, 1), $sin(2\pi r_2)$ and $cos(2\pi r_2)$ simulats the reflection of beluga whales fins under the water surface, and *even* and *odd* denote an odd number and an even number, respectively.

2.1.3. Localized Development Phase

The development phase of the BWO simulates beluga hunting behavior, where beluga whales move and cooperate to forage based on the information they share with each other. The mathematical model is

$$X_i^{T+1} = r_3 X_{best}^{T} - r_4 X_i^{T} + C_1 \cdot L_F \cdot \left(X_r^{T} - X_i^{T}\right) \tag{6}$$

where r_3, r_4 is a random number between (0, 1), X_{best}^{T} is the optimal position of the beluga population, [20] C_1 is used to measure the random jump strength of Levi's flight, and L_F denotes the Levi's flight function, which is expressed mathematically as

$$C_1 = 2r_4\left(1 - \frac{T}{T_{max}}\right) \tag{7}$$

$$L_F = 0.05 \times \frac{u \times \sigma}{|v|^{\frac{1}{\beta}}} \tag{8}$$

where u and v are random numbers conforming to a normal distribution, $\beta = 1.5$, and the parameter σ is given by

$$\sigma = \left(\frac{\Gamma(1+\beta) \times sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{(1+\beta)}{2}\right) \times \beta \times 2^{\frac{(\beta-1)}{2}}}\right)^{\frac{1}{\beta}} \tag{9}$$

2.1.4. Whale-Fall Stage

After a beluga whale dies, its body slowly sinks to the bottom of the sea. In BWO, the algorithm enters the whale fall phase when the equilibrium factor $B_f < W_f$. This phase enhances the ability of the algorithm to jump out of the local optimum. The mathematical formula for this phase is

$$X_i^{T+1} = r_5 X_i^T - r_6 X_r^T + r_7 X_{step} \tag{10}$$

$$W_f = 0.1 - \frac{0.05T}{T_{max}} \tag{11}$$

where r_5, r_6, and r_7 are random numbers distributed within (0,1), and X_{step} denotes the step size of the whale fall. C_2 is the step factor, which is determined by the probability of whale fall W_f and the population size pop, and its formula is

$$X_{step} = (u_b - l_b)exp\left(-\frac{C_2 T}{T_{max}}\right) \tag{12}$$

where u_b and l_b are the upper and lower bounds, respectively.

$$C_2 = 2W_f \times pop \tag{13}$$

2.2. BWO Improvement Strategy

In the SBWO algorithm, various strategies are employed to enhance optimization performance and avoid falling into local optima. Specifically, the algorithm incorporates adaptive parameter tuning, inertia weighting strategies, and variational operations to improve the robustness and control of the algorithm.

2.2.1. Adaptive Parameter Tuning

Adaptive tuning parameters η and $inertia_{weight}$ are introduced into the algorithm, and these parameters change dynamically with the iterative process, which is a common optimization approach [21]. This tuning approach aims to better balance the exploration and development of the algorithm by flexibly changing the parameters during the search process to improve the algorithm's capacity to leap beyond the local optimal solution and improve the global search performance. The use of dynamic tuning can better adapt to changes in search space complexity over time.

Adaptive parameters η:

$$\eta = \frac{\eta_{max} - \eta_{min}}{1 + e^{\alpha(T - \frac{Max\ it}{2})}} + \eta_{min} \tag{14}$$

where a is the parameter controlling the rate of change, T is the current number of iterations, and Max_{it} is the maximum number of iterations.

Inertia Weights:

$$inertia_{weight} = inertia_{max} - \left(\frac{inertia\_max - inertia\_min}{Max_{it}} \right) \cdot T \tag{15}$$

This inertia weight is used to control the change in speed and helps to balance the two phases of exploration and development.

2.2.2. Enhanced Quadratic Interpolation Strategy

The new location update strategy is incorporated into the algorithm by enhancing the quadratic interpolation approach. Quadratic interpolation performs position updating by using three different individuals, x_{best}, x_{r_1}, and x_{r_2} which further improves the search accuracy [22]. Using this approach, the local optimal solution can be approximated faster while avoiding excessive falling into local optima during the local search. The Quadratic interpolation optimization formula is as follows:

$$x_p^* = \frac{(x_2^2 - x_3^2)f_1 + (x_3^2 - x_1^2)f_2 + (x_1^2 - x_2^2)f_3}{2(x_2 - x_3)f_1 + (x_2 - x_1)f_2 + (x_1 - x_2)f_3} \tag{16}$$

In this paper, at each beluga position update, the optimal individual X_{best}^t is selected to form a low-degree polynomial with the value of the objective function with the other two random individuals X_{rl}^t and X_{rr}^t, and the solution $X_{I,j}^{T+1}$ is found by using this formula. Substituting the above variables into the formula x_p^*, the following formula can be obtained.

$$X_{i,j}^{t+1} = \frac{1}{2} \times \frac{\left[\left({X_{rl,j}^t}^2 - X_{best}^t \right) f(X_{rr}^t) + \left(X_{best}^t - {X_{rr,j}^t}^2 \right) f(X_{rl}^t) + \left({X_{rr,j}^t}^2 - {X_{rl,j}^t}^2 \right) f(X_{best}^t) \right]}{\left[\left(X_{rl,j}^t - X_{best}^t \right) f(X_{rr}^t) + \left(X_{best}^t - X_{rr,j}^t \right) f(X_{rl}^t) + \left(X_{rr,j}^t - X_{rl,j}^t \right) f(X_{best}^t) + eps \right]} \tag{17}$$

where $f(X_{best}^t)$, $f(X_{rl}^t)$, and $f(X_{rr}^t)$ denote the fitness values of X_{best}^t, X_{rl}^t, and X_{rr}^t, respectively. By using this method, the local search ability of the beluga search algorithm is greatly improved, which, in turn, improves the capacity to leap beyond the local optimal solution in the high-dimensional space and the convergence performance.

2.2.3. Exploring and Developing Mechanisms

For better global exploration and local exploitation in the search space, the algorithm employs both the IBWO method and the Levy flight mechanism. The IBWO method drives new position updates through differences in the random individual positions, while the Levy flights use their random jumping properties to enhance local search capabilities. The combination of these two allows the algorithm to efficiently perform both global exploration at an early stage and local optimization at a later stage, ensuring better convergence and improving global search capabilities [23]. Exploration and exploitation are achieved in two ways:

The exploratory phase (IBWO approach) is driven by differences in randomized individuals as follows.

$$Newpos(i,j) = inertia_{weight} \cdot pos(i,j) + \eta \cdot (pos(RI,j) - pos(i,j)) \cdot (r_1 + 1) \cdot sin(r_2 \cdot 360) \tag{18}$$

Development phase (Levy Flight, combining IBWO and NIWBWO methods): the development of local search using the Levy Flight mechanism:

$$Newpos(i,j) = pos(i,j) + S \cdot (X_{best}(j) - pos(i,j)) \tag{19}$$

One of the steps that Levy Flight takes is through the following:

$$S = \frac{u}{|v|^{\frac{1}{\alpha}}}, \quad u \sim N\left(0, \sigma^2\right), \quad v \sim N(0,1) \tag{20}$$

where $\alpha = \frac{3}{2}$ is a parameter of the Levy distribution, and σ is the standard deviation calculated from the Gamma function.

2.2.4. Mechanisms of Variation

During the exploration and development phase, in order to avoid the algorithm from falling into a local optimum, a mutation mechanism is added to increase the diversity of solutions. By randomly perturbing individuals, the mutation operation expands the breadth of global exploration, enabling the algorithm to identify possible solution areas, even close to the boundary. The mutation rate is set to 0.05 to ensure that the mutation process can strike a balance between diversity and stability.

A variation mechanism is used to increase the diversity of solutions and thus avoid falling into local optima. The rate of variation mutationrate is 0.05, and the variation formula is

$$ewpos(i, mutation_{idx}) = lb(mutation_{idx}) + (lb(mutation_{idx}) - lb(mutation_{idx})) \cdot rand(0,1) \tag{21}$$

Diversity is enhanced by randomly selecting locations and introducing random variation within the boundaries.

2.2.5. Algorithmic Description

Adaptive tuning enables dynamic adjustment of the search strategy by varying η and inertia weights during the iteration process. The quadratic interpolation strategy enhances the accuracy of searching within the search space. The exploration and exploitation mechanism combines Levy flight and BWO techniques to enhance the ability of global search and local exploitation. The mutation mechanism increases the diversity of understanding and avoids falling into local optimums. The following is an overall description of the SBWO (Algorithm 1).

Algorithm 1: SBWO

1. **Input**: Population size N_{pop}, maximum iterations T_{max}, lower bound lb, upper bound ub, dimensions nD, objective function f_{obj}.
2. **Initialize**:

   The fitness array fit $= \infty$ for all solutions.
   Population positions pos within bounds lb and ub.
   Global best position x_{best} and fitness f_{best}.
   Adaptive parameters η_{max}, η_{min}, inertia max, inertia min, mutation rate.
3. **Output**: Global best position x_{best}, best fitness f_{best}, fitness curve Curve.
4. **For** i $= 1$ to N_{pop}:

   Calculate fitness $f(\text{pos}\,[i])$ and update x_{best} when better.
5. **End For.**
6. **While** $t \leq T_{max}$:

   Update adaptive parameters η and inertia weight.
   Compute whale fall probability WF and exploration − exploitation rate kk.

Algorithm 1 *Cont.*

7. **For** $i = 1$ to N_{pop}:
   Interpolation Strategy:
      Select two random whales rl, rr.
      Update temporary positions T_{pos} using quadratic interpolation.
      Update pos$[i]$ and x_{best} when better.
   Exploration/Exploitation Phase:
      If $kk\ [i] > 0.5$ (exploration):
         Update position using random peers and a non-linear inertia weight.
      Else (exploitation using Levy flight):
         Update position with Levy flight strategy.
      Mutation: Apply mutation occurs with probability $\text{mutation}_{\text{rate}}$.
      Boundary check and fitness update.
      Update fit$[i]$ and x_{best} when improved.
8. **Whale Fall Mechanism (EBWO):**
   For each individual whale, if $kk\ [i] \leq WF$:
      Update position using whale fall mechanism.
      Boundary check and fitness update.
      Update fit$[i]$ and x_{best} when improved.
9. Update the best fitness value f_{best} and position x_{best}.
10. **Record** Curve$[t] = f_{best}$.
11. Increment t.

2.3. Hyperparametric Optimization MSF-MaxViT

Although deep learning networks show great potential in dealing with complex problems, their performance is affected by a variety of factors, especially the network and the setting of hyperparameters. Among these factors, the choice of hyperparameters is crucial for optimizing the performance and generalization ability. To address this challenge, this study proposes the SBWO algorithm for MSF-MaxViT network improvement.

Well-tuned hyperparameters can alleviate overfitting and improve classification performance. Therefore, the hyperparameter optimization problem has become a research hotspot within the realm of deep learning [24,25]. The SBWO algorithm is used to dynamically optimize key hyperparameters, such as the learning rate and dropout, during the training process of the MSF-MaxViT network. The learning rate has a direct impact on both the convergence speed and classification accuracy in the model, and SBWO effectively improves the training performance of the model through the dynamic modification of the learning rate. dropout. As an important parameter of regularization, SBWO can accurately search and select the optimal dropout, thus further improving the classification performance of the model while avoiding overfitting.

By combining the SBWO algorithm with the MSF-MaxViT network, this study proposes an efficient hyperparameter optimization framework, which not only improves the training performance of the model but also provides a new solution for the application of deep learning networks in complex tasks.

3. Multi-Scale Fusion MaxViT Network Model

3.1. MaxViT Background

In the field of image classification, although existing convolutional neural network (CNN) models have achieved remarkable results, researchers are still exploring innovative architectures to further improve the performance and efficiency of these models. In 2020,

the Google team introduced the Transformer architecture from the field of natural language processing to computer vision for the first time and proposed the Vision Transformer (ViT) [26]. However, the performance advantage of ViT is accompanied by an increase in the number of model parameters and a heightened computational cost, which greatly limits its practical application in resource-constrained scenarios.

To further balance the performance and efficiency of ViT, recent studies have gradually combined the attention mechanism of the Transformer with traditional convolutional structures. In 2021, Microsoft Research proposed the Swin Transformer [27], a hierarchical visual Transformer architecture, which has achieved excellent results in several vision tasks. However, the high computational complexity of the Transformer architecture remains a major bottleneck for model scaling. To solve this problem, in 2022, the Google team proposed the MaxViT model [1], which further combines the local and global attention mechanisms and introduces both chunked sparse and axial attention to improve the efficiency and performance of the model. MaxViT not only achieves leading classification performance on the ImageNet dataset but also demonstrates excellent parameter efficiency and inference speed, providing more possibilities for real-world deployment scenarios.

The design of MaxViT is based on a layered architecture, with each layer consisting of multiple MaxViT Blocks. Each MaxViT Block efficiently processes information within the block by dividing the image into non-overlapping blocks using the block sparse attention mechanism. Subsequently, axial attention computes the dependencies between features along the horizontal and vertical axes separately, thus avoiding the waste of resources when computing the attention matrix globally [1]. In addition, MaxViT introduces a deep convolution operation within each block to further enhance the local feature extraction capability. This fusion design of convolution and attention mechanism makes MaxViT both efficient and flexible in multi-scale feature extraction and improves the network's expressive ability for complex visual tasks. Based on these advantages, MaxViT emerges as a key breakthrough in the visual transformer sector, providing a new direction for the further exploration of efficient deep learning models.

In this paper, MaxViT is chosen as the base network, combining its efficient hierarchical feature extraction capabilities and multi-scale mdelling features to further enhance the improved performance and efficiency of the medical image classification task. By optimizing the hyperparameters of MaxViT and introducing a custom attention module, this paper aims to explore its efficient application in clinical medical image classification tasks.

3.2. MaxViT Optimization Improvements

3.2.1. The Overall MSF-MaxViT Framework

The MSF-MaxViT model represents an improved model based on the MaxViT network, aiming to enhance the performance of medical image classification tasks. MSF-MaxViT combines an innovative parallel attention mechanism and an enhanced feature extraction module, which further optimizes the learning ability of image features and the generalization of the model on the basis of MaxViT through multi-level information fusion and feature expression enhancement.

The core architecture of MSF-MaxViT is the MaxViT model, which demonstrates powerful global and local feature extraction capabilities in vision tasks by combining convolutional operations with the Transformer's self-attention mechanism. The MaxViT architecture employs a multi-scale-based feature extraction scheme that aims to capture both detailed information and the macroscopic structure of an image. However, MaxViT still has some limitations in feature fusion and information flow, especially in the fusion of local and global information. Therefore, this study improves MaxViT by introducing Parallel Attention and MSFA modules to enhance its feature learning capability.

The Parallel Attention module is an important innovation in the MSF-MaxViT to enhance the attention mechanism of MaxViT. While the traditional MaxViT Block uses serial attention computation, Parallel Attention changes this serial processing to parallel computation to process local and global information in images more efficiently. The Parallel Attention module consists of two Grid Attention [1] and Block Attention parallel attention branches, each of which is responsible for information capture at different scales. In addition, a multi-head attention mechanism is introduced in the module. The module replaces the single attention mechanism in the traditional MaxViT, enabling the model to better learn features at different scales and levels, and dynamically adjusts the weights of different branches during feature fusion, thus improving the image representation. The overall framework of MSF-MaxViT is shown in Figure 1.

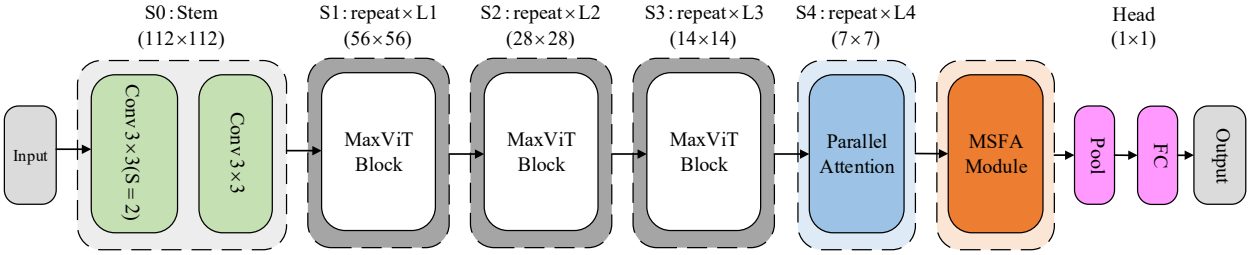


Figure 1. MSF-MaxViT.

The Multi-Scale Fusion Attention (MSFA) module is another core component of MSF-MaxViT, which is designed to enhance the model's ability to learn medical image features. The MSFA module combines several advanced attention mechanisms, including FB Block, CoordAttention, and Edge Attention, to further enhance the model's feature representation ability through the joint learning of the image. The joint learning of spatial and channel features further enhances the feature representation ability of the model. The design idea of the MSFA module is to strengthen the model's ability to perceive local details, edge information, and spatial structure by fusing multiple attention mechanisms. Different from the Parallel Attention module, the MSFA module focuses more on enhancing the network's ability at the feature expression level, especially when dealing with complex medical images, which can effectively improve the model's robustness and generalization ability.

After the feature extraction and enhancement module, MSF-MaxViT performs the final image classification through a Fully Connected Layer (FCL) as the classification head. In this layer, Global Average Pooling (GAP) is used to downscale the extracted features, and the dropout mechanism is introduced to prevent overfitting and improve the stability of the model.

In the overall process of MSF-MaxViT, firstly, the input image is a 224 × 224 RGB image; then, the input image is subjected to feature extraction through the MaxViT module, from which high-dimensionality features are extracted. After that, local and global feature information is computed and fused in parallel by the Parallel Attention module; then, the MSFA module is enhanced, and the feature representation is further enhanced by the MSFA module to capture the image information; finally, at the head of classification, the extracted features are classified by the fully-connected layer, and the final classification results are output.

3.2.2. Parallel Attention

In this study, the MaxViT architecture is improved by proposing the Parallel Attention module, which is based on the MaxViT Block for innovative design. MaxViT, recognized as an effective visual Transformer model, contains multiple computational modules of serial attention mechanism within its original structure. Within this structure, the attention

mechanism is sequentially applied to different image feature levels, and although this serial approach can gradually extract fine-grained features, it is insufficient in the fusion efficiency of local and global information. Therefore, based on the inspiration of MaxViT Block, the Parallel Attention module adopts the design concept of parallel computing to accelerate the feature fusion process by simultaneously performing information extraction at multiple attention levels. The Parallel Attention module contains the following main parts:

Grid Attention and Block Attention: these two distinct branches of attention process input feature maps through convolutional operations, respectively. Grid Attention is used to capture information about the grid structure of an image, whereas Block Attention focuses more on the details of local regions. Through parallel computation, these two can fuse information across multiple scales to improve the model's representation of different features.

The Multi-Head Attention Mechanism: To enhance the feature learning ability, Parallel Attention also incorporates Multi-Head Attention. This attention processes input features in parallel through multiple independent attention heads, allowing the model to capture complex features of an image from different perspectives, thus enhancing the diversity and richness of feature representation.

Dynamic Weight Fusion: In Parallel Attention, the outputs of Grid and Block Attention, as well as the results of multi-head attention, are fused through dynamic weights. Specifically, learnable weight parameters are assigned to each branch (including Grid, Block, and Multihead Attention). In this way, the model is able to automatically learn the importance of each attention module in a particular task, thus optimizing the information fusion process. The framework of Parallel Attention is shown in Figure 2.

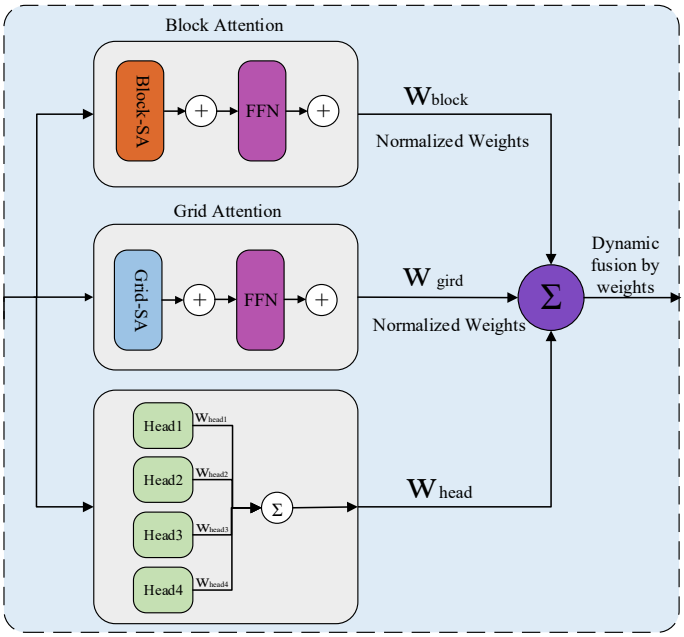


Figure 2. Parallel attention.

The Parallel Attention module is integrated into the core structure of MaxViT, replacing the original serial attention mechanism and forming the MSF-MaxViT network together with the MSFA module. In the structure of the model, Parallel Attention is located after the feature extraction module, which is responsible for further fusion and enhancement of the features extracted by MaxViT, thereby enhancing the model's capacity to handle intricate image data. With this design, the model effectively captures both local details and global structural information, which enhances the performance in medical image classification

tasks. Inspired by the MaxViT Block part, the Parallel Attention module improves the efficiency and effectiveness of information fusion by replacing serial computation with parallel processing. MaxViT's original single-attention mechanism is limited to layer-by-layer feature extraction; however, the parallel design is able to perform simultaneous feature extraction across various stages, thus accelerating the feature learning process of the network. This design not only enhances the computational efficiency of the model but also improves its generalization ability and performance in complex tasks.

3.2.3. MSFA Block

The Multi-Scale Fusion Attention (MSFA) module is a key component within the MSF-MaxViT network architecture, which is mainly used to enhance the learning ability of image features. The design of the MSFA module is inspired by the Position-Sensitive Attention (PSA) module [28]. The PSA module employs the fusion of spatial position information and channel information but fails to effectively capture details and global information due to its limited effectiveness in certain application scenarios, especially when processing complex medical images. Therefore, this study makes an innovative improvement based on PSA: the traditional Spatial-Channel Pooling (SPC) module is replaced by FB Block to enhance the model's expressive ability and feature extraction capability. Through the introduction of FB Block, the MSFA module can handle multi-scale information more accurately and show better performance in practical applications. The core idea of the MSFA module is to achieve more efficient feature fusion and stronger feature expression by introducing different attention mechanism modules and multi-scale feature extraction strategies. The MSFA module is shown in Figure 3.

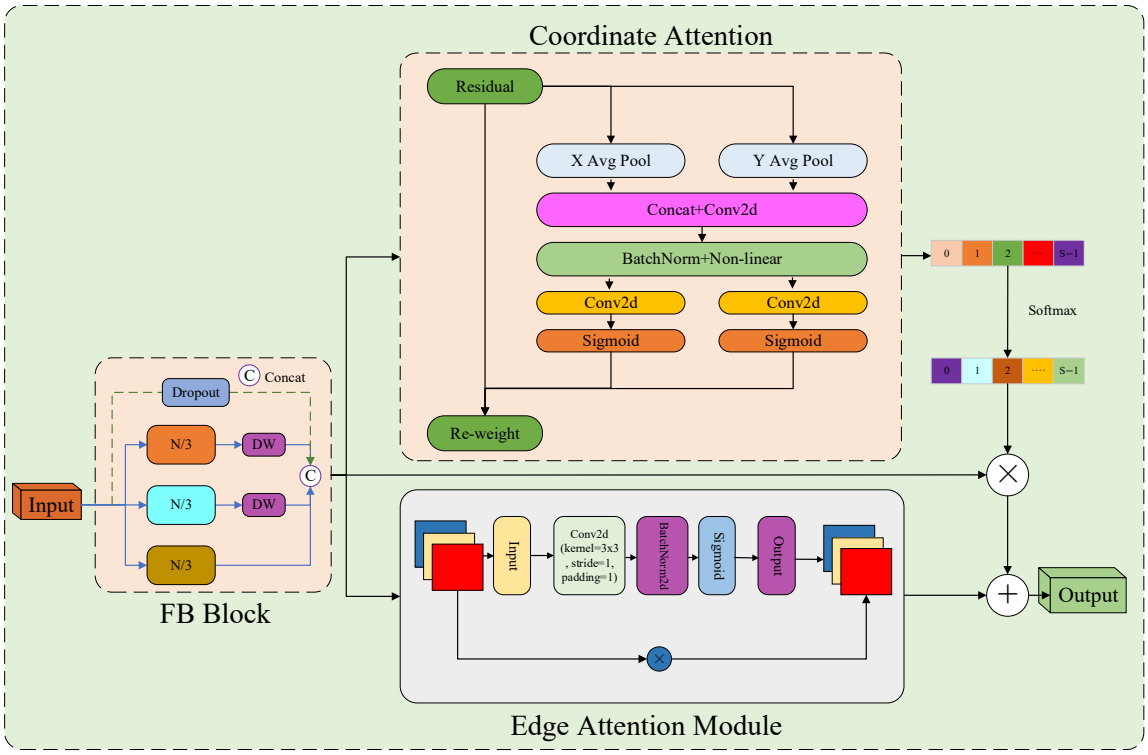


Figure 3. Multi-Scale Fusion Attention.

The workflow of the MSFA module can be divided into four main steps:

Multi-scale feature extraction: firstly, the MSFA module slices the input features through FB Block and performs feature extraction at different scales through multiple branches. FB Block adopts deep convolution operation, combined with a multi-branching

structure, which effectively extracts fine-grained local features from the image and enhances the model's ability to perceive local details. In FB Block, the main architecture is inspired by the M2C Block [29]. FB Block segments the feature map across the channel's dimension by a specific scaling factor into three distinct segments. Identity mapping stands as one of the subdivisions; the others use deep convolution at different scales to complete feature extraction and then along the channel dimension in order to fuse multi-scale and multi-dimensional feature information. In order to reduce the computational overhead of the module, instead of performing deep convolution operations on all channels, some of the channels are selected for deep convolution operations. In particular, the kernel function for deep convolution is set to $1 \times KS$ and $KS \times 1$ for the two parallel branches performing deep convolution, where KS denotes the default kernel size.

Channel Attention Generation: In the MSFA module, channel attention for a multi-scale feature map is extracted using coordinate attention [16] to effectively capture the global dependencies of salient features within an image. The process adaptively adjusts the weight of each channel via a channel-by-channel weighting mechanism to highlight the feature expressions in key regions, thus significantly enhancing the perceptual capability of the attention mechanism. The extracted multi-scale channel attention is subsequently normalized by the SoftMax function to achieve the fusion and interaction of multi-scale features. This step ensures the effective integration of information at each scale and the re-calibration of the channels within the feature map, which ultimately improves the model's representation ability in complex information scenarios.

Edge Attention Module: The core of the design of the Edge Attention Module is designed to enhance the perception of edge features in the spatial domain. Firstly, a convolutional kernel of 3×3 is used to perform a convolutional operation on the input features to capture the local edge features, thus achieving spatial weighting. Subsequently, the convolution results are normalized using Batch Norm to stabilize the training process and accelerate model convergence. Finally, a Sigmoid activation function is used to nonlinearly map the feature maps to generate weight distributions based on the edge information. The aim of this module is to improve the model's perception and representation of edge information, thus improving its ability to portray detailed features.

Weighted feature fusion: Finally, the MSFA module obtains the weighted feature maps by performing an element-by-element dot multiplication operation on the recalibrated attention weights and the corresponding feature maps. The weighting process allows for features with varying scales, thereby producing a more comprehensive feature representation. The feature map not only preserves the detailed information present in the original image but also integrates information from multiple scales, thereby enhancing the strength of the feature representation.

The innovative design of the MSFA module enhances the capacity for learning and the representation of image features when processing medical images. By combining attention mechanisms such as FB Block, CoordAtt, and Edge Attention, MSFA can perform feature fusion at multiple scales and levels and further enhance the performance of the model through the multi-head attention mechanism. Findings from the experiments reveal that the MSFA module improves the accuracy and robustness of the MSF-MaxViT model in medical image classification, demonstrating its strong potential for practical applications.

4. Experimental Analysis and Results

4.1. HAM10000 Dataset

The HAM10000 dataset includes 10,015 photographs depicting skin lesions specifically designed to train machine learning algorithms to accurately classify skin lesions. This

dataset is publicly available dataset for pigmented skin lesions and is widely used in research on skin lesion classification [30,31].

The images in this dataset consist of seven types of skin lesions: actinic keratoses and intraepithelial carcinomas (AKIEC), basal cell carcinomas (BCC), benign keratosis lentis (BKL), dermatofibromas (DF), melanomas (MEL), nevi with pigmented naevi (NV), and vasculitic lesions (VASC) [32].

Each image is thermoscopic, high-resolution, and accompanied by a label for the type of lesion. The images in the HAM10000 dataset were rigorously reviewed by several experienced dermatologists to ensure the accuracy of the data. There is a significant imbalance within this dataset, where the number of images in the 'NV' category far exceeds that of the other categories, and this imbalance may lead to a model bias in favor of the more numerous categories during classification. Therefore, when dividing the data training and test set, the different categories are classified in proportion to their numbers to avoid the problem of category imbalance in random classification.

Data Set Division and Enhancement

In the experiments of this paper, the HAM10000 dataset is divided into a training set and a validation set in the ratio of 80:20, containing 8012 and 2003 images, respectively [1]. Due to the imbalance of the dataset, data enhancement techniques are used in order to prevent the model from biasing the classification of a few classes during the training process. Specific enhancement methods include geometric transformations (e.g., random rotation and horizontal flip) and normalization. After data enhancement, this topic further balances the number of images of various types of skin lesions, which improves the classification performance of the model, reduces the overfitting phenomenon of the model, and increases the stability of the model during the learning process.

Six images of skin lesions from the HAM10000 dataset labeled as different types of skin lesions are shown in Figure 4. Image (a) is benign keratosis (bkl), image (b) is melanoma (mel), image (c) is basal cell carcinoma (bcc), image (d) is nevus (nv), image (e) is keratosis pilaris (akiec), and image (f) is a vascular lesion (vasc). These images provide important data support for the task of skin lesion classification.

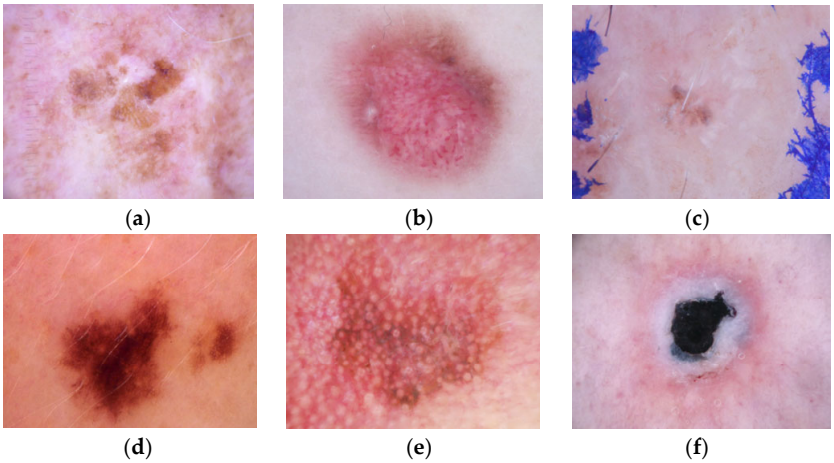


Figure 4. HAM10000 Image: (**a**) ISIC_0027419: bkl; (**b**) ISIC_0033603: mel; (**c**) ISIC_0031263: bcc; (**d**) ISIC_0029624: nv; (**e**) ISIC_0031228: akiec; (**f**) ISIC_0033749: vasc.

4.2. Testing of Optimization Algorithms

The CEC2005 test set is a standard set of test functions provided by the IEEE to evaluate the performance of optimization algorithms [20,33]. The test set consists of 30 single-objective optimization problems specifically designed to evaluate the performance

of the algorithms in different problem types, especially high-dimensionality and multi-peaked problems.

We compare SBWO with other optimization algorithms, including APO, GOOSE, RBMO, and PKO. The algorithms were run independently 30 times to obtain the optimal values, and the search results are shown in Table 1. The average convergence curves of some functions are shown in Figure 5. As can be seen in Figure 5, the optimal values of SBWO surpass the performance of alternative comparative algorithms for functions such as F1 and F10, reflecting the superior stability of the improved algorithm in searching for an optimal value. For other functions, although the evaluation index of the SBWO may not be the best, it shows some improvement when compared with other algorithms, which illustrates the success of the improvement strategy suggested in this document. Figure 6 shows the superior ability of the SBWO to jump out of the local optima. In conclusion, the test of the CEC2005 function reflects that SBWO has strong competitiveness and good comprehensive performance in solving complex problems.

Table 1. Results of benchmark functions for compared algorithms.

Test Function	BWO	SBWO	APO	GOOSE	RBMO	PKO
F1	1.14×10^{-267}	0	1.10×10^{-6}	5.15×10^{-3}	2.05×10^{-4}	3.30×10^{-4}
F2	1.15×10^{-134}	0	1.50×10^{-4}	1.17×10^{2}	1.20×10^{-2}	1.49×10^{-2}
F3	1.97×10^{-241}	0	2.74×10^{-2}	7.88×10^{3}	2.21×10^{2}	1.47×10^{3}
F4	4.05×10^{-130}	0	1.36×10^{-1}	3.77×10^{1}	3.67×10^{0}	1.01×10^{0}
F9	2.20×10^{-259}	0	1.08×10^{-7}	2.26×10^{-1}	1.21×10^{-5}	6.40×10^{-8}
F10	2.80×10^{-275}	0	2.91×10^{-16}	8.53×10^{-6}	1.67×10^{2}	2.11×10^{5}
F11	1.98×10^{-129}	0	2.02×10^{-3}	1.79×10^{2}	4.75×10^{-2}	3.35×10^{-2}
F12	0	0	4.76×10^{-24}	1.77×10^{-12}	6.03×10^{-12}	3.17×10^{-11}
F24	0	0	5.28×10^{0}	1.84×10^{2}	2.35×10^{1}	2.0019

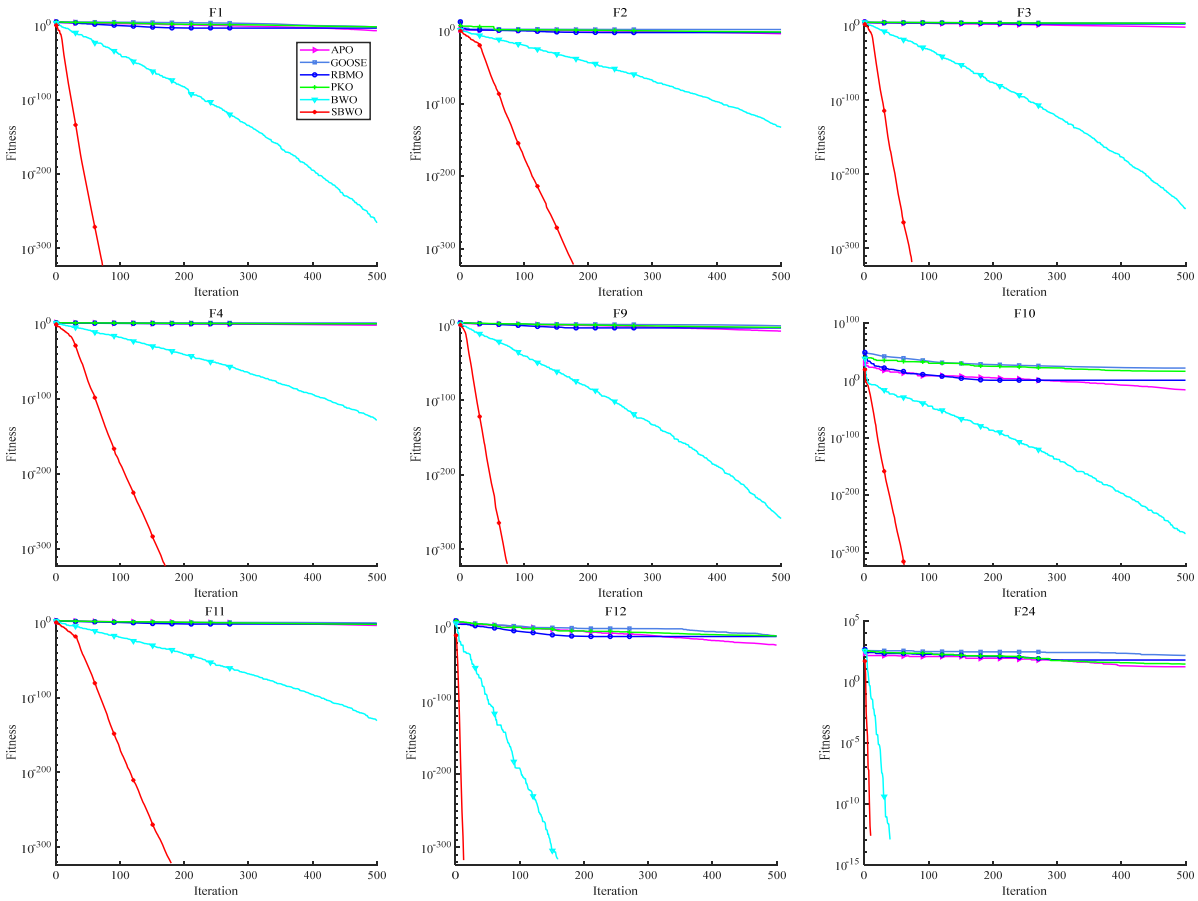


Figure 5. Algorithm comparison curve.

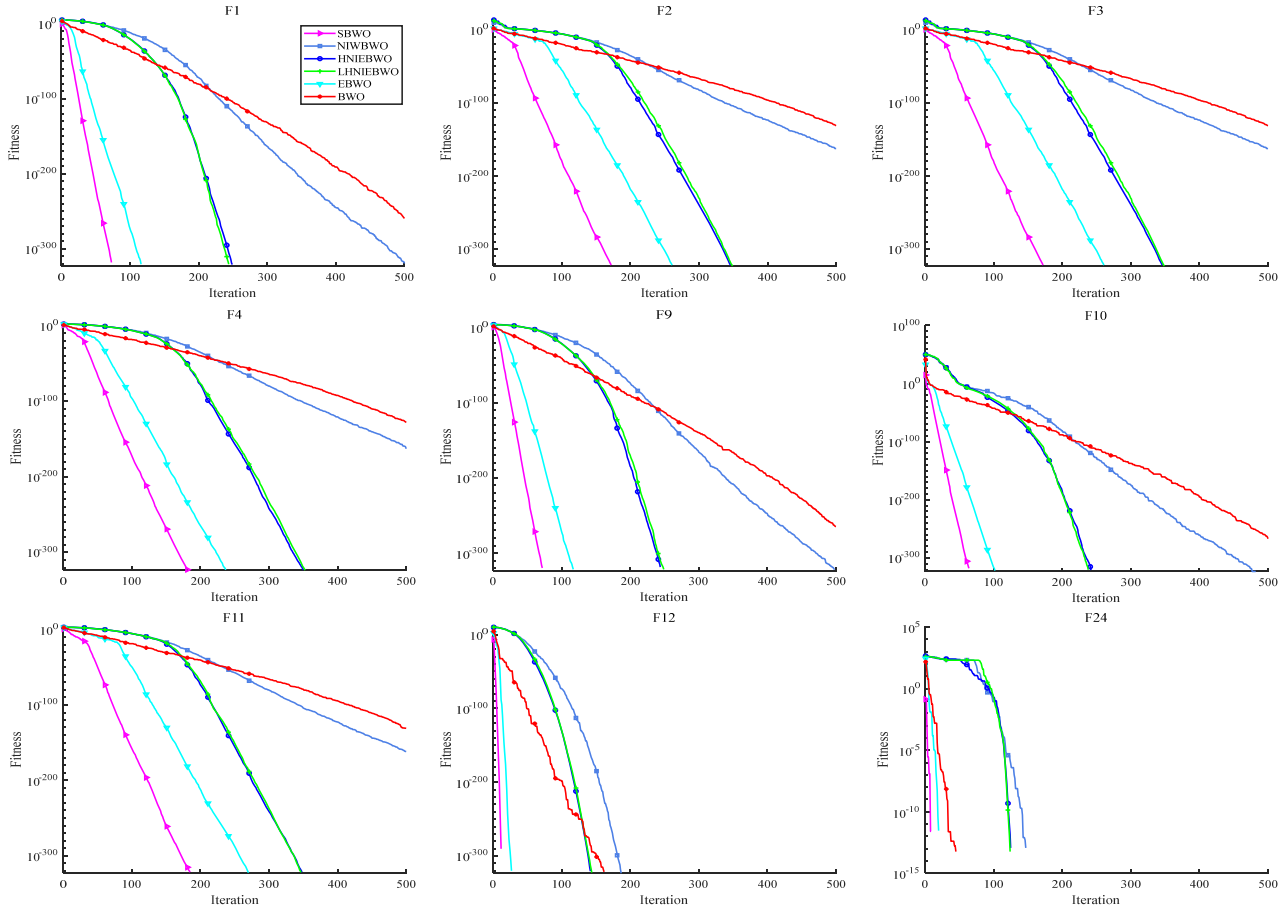


Figure 6. Convergence curve.

This experiment evaluates different versions of the BWO algorithm used to confirm the efficiency of every enhancement. Specific results are shown in Table 2. The base model BWO is the original BWO algorithm, which serves as a benchmark for comparison. The SBWO is the best model that combines several improvements (e.g., quadratic interpolation strategy, nonlinear inertia weight factors, etc.) and represents the final performance of the algorithm. In the ablation experiment charts, Enhanced Beluga Whale Optimization (EBWO), Hybrid Nonlinear Inertial Enhanced Beluga Whale Optimization (HNIEBWO), Latin Hypercube Nonlinear Inertial Enhanced Beluga Whale Optimization (LHNIEBWO), Nonlinear Inertial Weighting Beluga Whale Optimization (NIWBWO), and Improved Beluga Whale Optimization (IBWO) are all intermediate models that introduce different improvements step by step, demonstrating the impact of each improvement individually on the algorithm's performance. Among these methods, EBWO improves the search accuracy by introducing a quadratic interpolation strategy; HNIEBWO combines nonlinear inertial weighting factors to further balance global exploration and local exploitation; LHNIEBWO initializes the population using Latin Hypercubic Sampling, which increases the diversity of the population; whereas NIWBWO improves the search by enhancing the nonlinear inertial weighting factors' stability during the process. IBWO further balances population diversity and convergence speed throughout the exploration and development phases by introducing an adaptive tuning strategy and adjusting inertia weights. The combination of these strategies enables SBWO to avoid local optima more effectively and improve optimization performance. By comparing models such as BWO and SBWO, one can clearly see the contributions of each algorithmic improvement in solving the optimization problem.

Table 2. Comparative analysis of ablation experiments.

Test Function	BWO	SBWO	EBWO	NIWBWO	LHNIEBWO	HNIEBWO
F1	1.32×10^{-260}	0	0	8.82×10^{-321}	0	0
F2	4.55×10^{-132}	0	0	2.24×10^{-163}	0	0
F3	4.00×10^{-249}	0	0	1.17×10^{-305}	0	0
F4	6.38×10^{-129}	0	0	1.08×10^{-163}	0	0
F9	1.79×10^{-265}	0	0	0	0	0
F10	1.68×10^{-267}	0	0	0	0	0
F11	5.94×10^{-132}	0	0	8.67×10^{-163}	0	0
F12	0	0	0	0	0	0
F24	0	0	0	0	0	0

4.3. Experimental Configuration and Results

The entire experiment was performed in the Python 3.11 environment using a computer configured with a Windows 11 operating system, CUDA 12.6 computing platform, NVIDIA GeForce RTX 4070 SUPER graphics card for GPU, and Intel(R) i5-14600KF for the CPU. PyTorch 2.5.1+cu111 served as the framework for deep learning.

The hyper-parameter settings used during the training process include the initial learning rate, the minimum learning rate, the batch size, the number of training rounds, the loss function, and the relevant configuration of the optimizer. Specifically, the initial learning rate is set to 1.00×10^{-4}, the minimum learning rate is 1.00×10^{-5}, the batch size is 64, the number of rounds of training is 45, and the loss function uses Cross-Entropy Loss (CEL). The optimizer uses AdamW, where the learning rate is adjusted to 0.00076 after SBWO optimization, and the dropout rate is 0.49. In addition, the model uses pre-training weights and freezes the 'stem' and 'layer1' layers in the MaxViT model. The training process was enabled with Mixed-Precision Training and processed using GradScaler and autocast to improve computational efficiency.

As for Table 3, the table shows the performance of different models in terms of accuracy in the medical image classification task. The table lists the names of several models, the year they were proposed, as well as the corresponding classification accuracies. Specifically, the ConvNeXt_L (2022) [34] and TransSLC (2022) [35] models achieve 88.40% and 90.20% accuracy, respectively. Other classical models, such as NASNetLarge (2018) [36] and Xception (2017) [36], achieved 91.11% and 91.47% accuracy, respectively. In addition, the traditional DenseNet (2015) [37] model has an accuracy of 91.56%.

Table 3. Accuracy comparison.

Model Name	Year	Accuracy
ConvNeXt_L [34]	2022	88.40%
TransSLC [35]	2022	90.20%
An Enhanced CNN Model [38]	2022	90.55%
NASNetLarge [36]	2018	91.11%
Xception [36]	2017	91.47%
InceptionV3 [36]	2015	91.56%
DenseNet [37]	2017	88.20%
Resnet50, Restnet101 + KcPCA + SVM RBF [39]	2019	89.80%
Model Based on DenseNet and ConvNeXt Fusion [34]	2023	90.85%
MaxViT	2021	91.02%
MSF-MaxViT (Ours)	2024	92.87%

Notably, our proposed MSF-MaxViT (2024) model performs the best in terms of accuracy, achieving 92.87%, outperforming all the compared models. This result validates the effectiveness of our proposed MSF-MaxViT model in medical image classification and shows that the model is highly accurate and competitive in practical applications.

4.4. Ablation Experiments

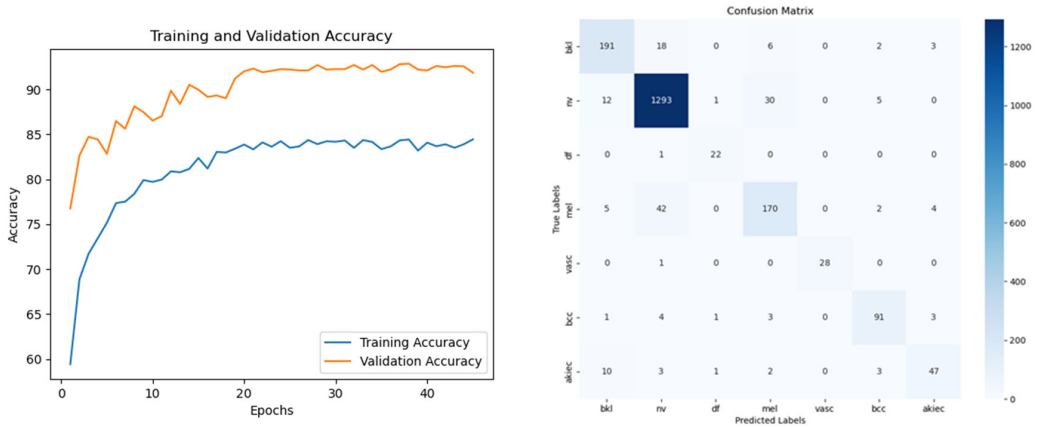
To verify the effectiveness of the MSF-MaxViT model and its improved modules suggested in this document for performance enhancement, multiple sets of ablation experiments were designed, and the independent contributions of each module were quantified. Experimental outcomes are displayed in Table 4., with the highest accuracy of 91.02% for the base model, MaxViT, and the final stable accuracy of 90.82% serves as a control benchmark. The change in accuracy is not significant after the introduction of the initial MSFA module alone. However, after further adding the Parallel Attention module on top of the MSFA module and integrating residual linkage, dynamic fusion weights, and edge information enhancement with multiscale convolution into the MSFA module, the model performance improves, with the highest accuracy of 91.37% and the final stable accuracy of 91.17%, which verifies the importance of multiscale information and edge enhancement in feature representation. Further analysis shows that after removing the multiscale convolution, the accuracy slightly decreases to 91.07%, but it is still higher than that of the MaxViT base model, indicating that although the multiscale convolution contributes, it is not a key determinant. Finally, combined with the SBWO optimization strategy, the model achieves a maximum accuracy of 92.87% and a final accuracy of 91.87%, which are 1.85% and 1.05% higher than the base model, which fully proves the synergistic effect between modules and the importance of the optimization strategy in improving the accuracy and stability of the model.

Table 4. Comparison of ablation experiments.

Model	Max Accuracy	Last Accuracy
MaxViT	91.02%	90.82%
MaxViT + MSFA	90.07%	89.88%
MaxViT + MSFA + Parallel Attention	89.03%	88.98%
MaxViT + MSFA (Residual + Dynamic Fusion Weights) + Parallel Attention	91.37%	91.17%
MaxViT + MSFA (Residual + Edge Infor Enhancement + Multiscale Convolution) + Parallel Attention	91.07%	90.97%
MaxViT + MSFA (SOTA)	91.42%	91.32%
SOTA + SBWO (ours)	92.87%	91.87%

Figure 7 presents the key results of the model training process, including training and validation accuracy changes and classification performance evaluation. Figure 7a presents the accuracy changes during the training process, where the blue curve represents the training accuracy, and the orange curve represents the validation accuracy. It can be seen in the figure that the training accuracy gradually improves as the training progresses, reaching a peak of 84.44%, while the validation accuracy also shows a steady improvement. This indicates that the model gradually converges during the training process and can perform stably on the validation set.

Figure 7b shows the confusion matrix, which shows how the model's predictions on each category compare with the true labels. As can be seen in the figure, most of the categories are very accurate, especially the category 'nv', which has a high prediction accuracy of 1293 correct predictions with almost no misclassification. Other categories, such as 'bkl' and 'mel', also have high classification accuracy, although there are some misclassifications (e.g., 18 misclassifications of 'bkl' to 'nv"). By analyzing the confusion matrix, it is possible to visually assess the performance of the model on different skin lesion types, helping to improve the model's classification ability.



(**a**) (**b**)

Figure 7. (**a**) Accuracy images; (**b**) confusion matrix.

4.5. Evaluation Indicators

In this experiment, in order to thoroughly evaluate of the efficacy of the suggested model, a variety of evaluation metrics are used, and the results are compared with those of the baseline model [40,41].

As shown in the data in Table 5, the MSF-MaxViT (Ours) model outperforms the other compared models, with excellent performance in all the metrics, especially in recall, precision, and F1 scores, reaching 92.00%. This result validates the advantages of our proposed model in the medical image classification task. Other models, such as the Enhanced CNN Model, also perform well in some metrics, but are slightly lower than MSF-MaxViT in overall performance.

Table 5. Comparison of model evaluation indicators.

Model Name	Recall	Precision	F1-Score
ConvNeXt_L [34]	76.47%	80.04%	78.24%
TransSLC [35]	85.00%	87.00%	85.00%
An Enhanced CNN Model [38]	91.85%	92.89%	92.37%
NASNetLarge [36]	86.00%	86.00%	86.00%
Xception [36]	88.00%	89.00%	88.00%
InceptionV3 [36]	89.00%	89.00%	89.00%
DenseNet [37]	84.61%	78.67%	81.53%
Resnet50, Restnet101 + KcPCA + SVM RBF [39]	89.71%	90.14%	89.92%
Model Based on DenseNet and ConvNeXt Fusion [34]	83.81%	83.75%	83.45%
MSF-MaxViT (Ours)	92.00%	92.00%	92.00%

The combined use of these formulas facilitates a thorough evaluation of model performance from various viewpoints, particularly in instances of category imbalance, where each metric offers a distinct perspective on performance.

5. Discussion and Conclusions

This study presents a novel deep learning methodology that integrates SBWO with a new module utilizing MaxViT, aimed at improving the accuracy of classifying medical images. In the MaxViT, the Parallel Attention module is introduced, which successfully achieves the effective fusion of local and global features through the parallel computation of the multi-head dynamic attention mechanism, thus enhancing the feature representation ability of the model and further improving the accuracy in classifying medical images. In addition, this study proposes the Multi-Scale Fusion Attention (MSFA) module, which effectively enhances the model's ability in spatial and channel feature learning and improves its generalization performance by integrating the FB Block, Coord Attention, and Edge Attention modules.

For hyperparameter optimization, the improved BWO algorithm is adopted, which combines the quadratic interpolation optimization strategy and adaptive parameter tuning to effectively enhance the stability and convergence of the MSF-MaxViT network and further optimize the network structure and hyperparameter configuration. Outcomes of the experiments indicate that the MSF-MaxViT model, optimized using the SBWO technique, attains a classification accuracy of 92.87% on the HAM10000 dataset. This represents an improvement of 1.85% relative to the original MaxViT model.

Future work will aim to further improve the performance of the model on large-scale medical image datasets, exploring more lightweight network architectures and efficient attention mechanisms to reduce the computational cost and improve the real-time performance of the model. In addition, considering the potential of multimodal data, future research could extend the model to handle different types of medical data, such as combining electrical impedance imaging (EIT) techniques with conventional medical imaging data [42]. With the introduction of diverse data and an increase in the number of samples, we expect to be able to effectively improve the classification accuracy and robustness of the model, thus facilitating its practical application in clinical aid diagnosis. With these enhancements, MSF-MaxViT is expected to show wider applicability in the field of medical image classification.

Author Contributions: Conceptualization, J.Z. and L.S.; Methodology, J.Z.; Software, J.Z.; Validation, J.Z.; Formal analysis, J.Z.; Investigation, J.Z. and L.S.; Resources, J.Z. and L.S.; Data curation, J.Z., T.L. and L.S.; Writing—original draft, J.Z.; Writing—review & editing, T.L. and L.S.; Visualization, J.Z.; Supervision, T.L. and L.S.; Project administration, T.L. and L.S.; Funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. MaxVit: Multi-Axis Vision Transformer. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 459–479. [CrossRef]
2. Pacal, I. Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model. *Expert Syst. Appl.* **2024**, *238*, 122099. [CrossRef]
3. Jiang, T.; Guo, J.; Xing, W.; Yu, M.; Li, Y.; Zhang, B.; Ta, D. A prior segmentation knowledge enhanced deep learning system for the classification of tumors in ultrasound image. *Eng. Appl. Artif. Intell.* **2025**, *142*, 109926. [CrossRef]
4. Pacal, I. MaxCerVixT: A novel lightweight vision transformer-based Approach for precise cervical cancer detection. *Knowl.-Based Syst.* **2024**, *289*, 111482. [CrossRef]
5. Wang, W.C.; Tian, W.C.; Xu, D.M.; Zang, H.F. Arctic Puffin Optimization: A Bio-Inspired Metaheuristic Algorithm for Solving Engineering Design Optimization. *Adv. Eng. Softw.* **2024**, *195*, 103694. [CrossRef]
6. Hamad, R.K.; Rashid, T.A. GOOSE Algorithm: A Powerful Optimization Tool for Real-World Engineering Challenges and Beyond. *Evol. Syst.* **2024**, *15*, 1249–1274. [CrossRef]
7. Fu, S.; Li, K.; Huang, H.; Ma, C.; Fan, Q.; Zhu, Y. Red-Billed Blue Magpie Optimizer: A Novel Metaheuristic Algorithm for 2D/3D UAV Path Planning and Engineering Design Problems. *Artif. Intell. Rev.* **2024**, *57*, 134. [CrossRef]
8. Bouaouda, A.; Hashim, F.A.; Sayouti, Y.; Hussien, A.G. Pied Kingfisher Optimizer: A New Bio-Inspired Algorithm for Solving Numerical Optimization and Industrial Engineering Problems. *Neural Comput. Appl.* **2024**, *36*, 15455–15513. [CrossRef]
9. Zhong, C.; Li, G.; Meng, Z. Beluga Whale Optimization: A Novel Nature-Inspired Metaheuristic Algorithm. *Knowl.-Based Syst.* **2022**, *251*, 109215. [CrossRef]
10. Huang, J.; Hu, H. Hybrid Beluga Whale Optimization Algorithm with Multi-Strategy for Functions and Engineering Optimization Problems. *J. Big Data* **2024**, *11*, 3. [CrossRef]
11. Horng, S.C.; Lin, S.S. Improved Beluga Whale Optimization for Solving the Simulation Optimization Problems with Stochastic Constraints. *Mathematics* **2023**, *11*, 1854. [CrossRef]

12. Li, J.; Zhou, X.; Zhou, Y.; Han, A. Optimal Configuration of Distributed Generation Based on an Improved Beluga Whale Optimization. *IEEE Access* **2024**, *12*, 31000–31013. [CrossRef]
13. Yuan, H.; Chen, Q.; Li, H.; Zeng, D.; Wu, T.; Wang, Y.; Zhang, W. Improved Beluga Whale Optimization Algorithm Based Cluster Routing in Wireless Sensor Networks. *Math. Biosci. Eng.* **2024**, *21*, 4587–4625. [CrossRef]
14. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Sci. Data* **2018**, *5*, 180161. [CrossRef] [PubMed]
15. Qaraad, M.; Amjad, S.; Hussein, N.K.; Farag, M.A.; Mirjalili, S.; Elhosseini, M.A. Quadratic Interpolation and a New Local Search Approach to Improve Particle Swarm Optimization: Solar Photovoltaic Parameter Estimation. *Expert Syst. Appl.* **2024**, *236*, 121417. [CrossRef]
16. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722. [CrossRef]
17. Paçacı, S. Improvement of Beluga Whale Optimization Algorithm by Distance Balance Selection Method. *Yalvaç Akad. Derg.* **2023**, *8*, 125–144. [CrossRef]
18. Jia, H.; Wen, Q.; Wu, D.; Wang, Z.; Wang, Y.; Wen, C.; Abualigah, L. Modified Beluga Whale Optimization with Multi-Strategies for Solving Engineering Problems. *J. Comput. Des. Eng.* **2023**, *10*, 2065–2093. [CrossRef]
19. Chen, X.; Zhang, M.; Yang, M.; Wang, D. NHBBWO: A Novel Hybrid Butterfly-Beluga Whale Optimization Algorithm with the Dynamic Strategy for WSN Coverage Optimization. *Peer-to-Peer Netw. Appl.* **2025**, *18*, 80. [CrossRef]
20. Yuan, X.; Hu, G.; Zhong, J.; Wei, G. HBWO-JS: Jellyfish Search Boosted Hybrid Beluga Whale Optimization Algorithm for Engineering Applications. *J. Comput. Des. Eng.* **2023**, *10*, 1615–1656. [CrossRef]
21. Agrawal, A.; Tripathi, S. Particle Swarm Optimization with Adaptive Inertia Weight Based on Cumulative Binomial Probability. *Evol. Intell.* **2021**, *14*, 305–313. [CrossRef]
22. Chen, H.; Wang, Z.; Wu, D.; Jia, H.; Wen, C.; Rao, H.; Abualigah, L. An Improved Multi-Strategy Beluga Whale Optimization for Global Optimization Problems. *Math. Biosci. Eng.* **2023**, *20*, 13267–13317. [CrossRef]
23. Li, Y.; Li, X.; Liu, J.; Ruan, X. An Improved Bat Algorithm Based on Lévy Flights and Adjustment Factors. *Symmetry* **2019**, *11*, 925. [CrossRef]
24. Zhang, L.; Qiao, Z.; Li, L. An Evolutionary Deep Learning Method Based on Improved Heap-Based Optimization for Medical Image Classification and Diagnosis. *IEEE Access* **2024**, *12*, 102745–102773. [CrossRef]
25. El-Bouzaidi, Y.E.I.; Hibbi, F.Z.; Abdoun, O. Optimizing Convolutional Neural Network Impact of Hyperparameter Tuning and Transfer Learning. In *Innovations in Optimization and Machine Learning*; IGI Global Scientific Publishing: Hershey, PA, USA, 2025; pp. 301–326. [CrossRef]
26. Dosovitskiy, A. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]
28. Zhang, H.; Zu, K.; Lu, J.; Zou, Y.; Meng, D. EPSANet: An Efficient Pyramid Split Attention Block on Convolutional Neural Network. *arXiv* **2021**, arXiv:2105.14447. [CrossRef]
29. Yue, W.; Liu, S.; Li, Y. Eff-PCNet: An Efficient Pure CNN Network for Medical Image Classification. *Appl. Sci.* **2023**, *13*, 9226. [CrossRef]
30. Maduranga, M.W.P.; Nandasena, D. Mobile-Based Skin Disease Diagnosis System Using Convolutional Neural Networks (CNN). *Int. J. Image Graph. Signal Process.* **2022**, *12*, 47. [CrossRef]
31. Adebiyi, A.; Abdalnabi, N.; Smith, E.H.; Hirner, J.; Simoes, E.J.; Becevic, M.; Rao, P. Accurate skin lesion classification using multimodal learning on the HAM10000 dataset. *medRxiv* **2024**. [CrossRef]
32. Shetty, B.; Fernandes, R.; Rodrigues, A.P.; Chengoden, R.; Bhattacharya, S.; Lakshmanna, K. Skin Lesion Classification of Dermoscopic Images Using Machine Learning and Convolutional Neural Network. *Sci. Rep.* **2022**, *12*, 18134. [CrossRef]
33. Min, C.; Zhang, M.; Zhang, Q.; Jiang, Z.; Zhou, L. A Two-Stage Adaptive Differential Evolution Algorithm with Accompanying Populations. *Mathematics* **2025**, *13*, 440. [CrossRef]
34. Wei, M.; Wu, Q.; Ji, H.; Wang, J.; Lyu, T.; Liu, J.; Zhao, L. A Skin Disease Classification Model Based on DenseNet and ConvNext Fusion. *Electronics* **2023**, *12*, 438. [CrossRef]
35. Sarker, M.M.K.; Moreno-García, C.F.; Ren, J.; Elyan, E. TransSLC: Skin Lesion Classification in Dermatoscopic Images Using Transformers. In *Annual Conference on Medical Image Understanding and Analysis*; Springer: Cham, Switzerland, 2022; pp. 651–660. [CrossRef]
36. Chaturvedi, S.S.; Tembhurne, J.V.; Diwan, T. A Multi-Class Skin Cancer Classification Using Deep Convolutional Neural Networks. *Multimed. Tools Appl.* **2020**, *79*, 28477–28498. [CrossRef]
37. Heller, N.; Bussmann, E.; Shah, A.; Dean, J.; Papanikolopoulos, N. Computer-Aided Diagnosis of Skin Lesions from Morphological Features. Available online: https://api.semanticscholar.org/CorpusID:52108303 (accessed on 23 February 2025).

38. Haider, K.M.M.; Dhar, M.; Akter, F.; Islam, S.; Shariar, S.R.; Hossain, M.I. An Enhanced CNN Model for Classifying Skin Cancer. In Proceedings of the 2nd International Conference on Computing Advancements, Dhaka, Bangladesh, 10–12 March 2022; pp. 456–459. [CrossRef]
39. Khan, M.A.; Javed, M.Y.; Sharif, M.; Saba, T.; Rehman, A. Multi-Model Deep Neural Network-Based Features Extraction and Optimal Selection Approach for Skin Lesion Classification. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 3–4 April 2019; pp. 1–7. [CrossRef]
40. Pacal, I.; Alaftekin, M.; Zengul, F.D. Enhancing Skin Cancer Diagnosis Using Swin Transformer with Hybrid Shifted Window-Based Multi-Head Self-Attention and SwiGLU-Based MLP. *J. Imaging Inform. Med.* **2024**, *37*, 3174–3192. [CrossRef] [PubMed]
41. Saheed, Y.K.; Misra, S. CPS-IoT-PPDNN: A New Explainable Privacy Preserving DNN for Resilient Anomaly Detection in Cyber-Physical Systems-Enabled IoT Networks. *Chaos Solitons Fractals* **2025**, *191*, 115939. [CrossRef]
42. Laganà, F.; Prattico, D.; De Carlo, D.; Oliva, G.; Pullano, S.A.; Calcagno, S. Engineering Biomedical Problems to Detect Carcinomas: A Tomographic Impedance Approach. *Eng* **2024**, *5*, 1594–1614. [CrossRef]