J BONE ONCOL 43 (2023) 100508

Contents lists available at ScienceDirect

Journal of Bone Oncology

journal homepage: www.elsevier.com/locate/jbo

A fusion of VGG-16 and ViT models for improving bone tumor classification in computed tomography

Weimin Chen [a], Muhammad Ayoub [b], Mengyun Liao [b], Ruizheng Shi [c], Mu Zhang [d], Feng Su [d], Zhiguo Huang [d], Yuanzhe Li [e], Yi Wang [e], Kevin K.L. Wong [a,f,*]

[a] School of Information and Electronics, Hunan City University, Yiyang 413000, China
[b] School of Computer Science and Engineering, Central South University, Changsha 410083, Hunan, China
[c] National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha 410008, Hunan, China
[d] Department of Emergency, Xiangya Hospital, Central South University, Changsha 410008, Hunan, China
[e] Department of CT/MRI, The Second Affiliated Hospital of Fujian Medical University, Quanzhou 362000, China
[f] Department of Mechanical Engineering, College of Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada

HIGHLIGHTS

- We introduce a novel VGG-16-ViT fusion model to enhance bone tumor classification in computed tomography, leveraging the strengths of both architectures.
- Our proposed algorithm addresses limitations in CNNs' global perception ability, improving the accuracy of classifying diverse bone tumor types.
- The fusion model demonstrates an impressive 97.6% classification accuracy, with an 8% increase in sensitivity and specificity, surpassing traditional methods.
- Investigating the effect of secondary migration across three models shows potential for enhancing system performance, contributing to more accurate results.
- Our joint VGG-16 and Vision Transformer network proves effective in classifying bone tumors, promising improved early detection and prognosis for bone tumor patients.

ARTICLE INFO

Keywords:
Vision Transformer
VGG-16
ViT
Bone tumors diagnosis
Deep learning
Orthopedics image classification

ABSTRACT

Background and Objective: Bone tumors present significant challenges in orthopedic medicine due to variations in clinical treatment approaches for different tumor types, which includes benign, malignant, and intermediate cases. Convolutional Neural Networks (CNNs) have emerged as prominent models for tumor classification. However, their limited perception ability hinders the acquisition of global structural information, potentially affecting classification accuracy. To address this limitation, we propose an optimized deep learning algorithm for precise classification of diverse bone tumors.

Materials and Methods: Our dataset comprises 786 computed tomography (CT) images of bone tumors, featuring sections from two distinct bone species, namely the tibia and femur. Sourced from The Second Affiliated Hospital of Fujian Medical University, the dataset was meticulously preprocessed with noise reduction techniques. We introduce a novel fusion model, VGG16-ViT, leveraging the advantages of the VGG-16 network and the Vision Transformer (ViT) model. Specifically, we select 27 features from the third layer of VGG-16 and input them into the Vision Transformer encoder for comprehensive training. Furthermore, we evaluate the impact of secondary migration using CT images from Xiangya Hospital for validation.

Results: The proposed fusion model demonstrates notable improvements in classification performance. It effectively reduces the training time while achieving an impressive classification accuracy rate of 97.6%, marking a significant enhancement of 8% in sensitivity and specificity optimization. Furthermore, the investigation into secondary migration's effects on experimental outcomes across the three models reveals its potential to enhance system performance.

Conclusion: Our novel VGG-16 and Vision Transformer joint network exhibits robust classification performance on bone tumor datasets. The integration of these models enables precise and efficient classification,

* Corresponding author at: School of Information and Electronics, Hunan City University, Yiyang 413000, China.
E-mail address: kelvin.wong@hncu.edu.cn (K.K.L. Wong).

https://doi.org/10.1016/j.jbo.2023.100508
Received 9 May 2023; Received in revised form 14 August 2023; Accepted 20 September 2023
Available online 2 November 2023

accommodating the diverse characteristics of different bone tumor types. This advancement holds great significance for the early detection and prognosis of bone tumor patients in the future.

1. Introduction

The incidence of bone tumors is low, but primary malignant bone tumors rank third among the causes of death among cancer patients under 20 years of age. There are great differences in the clinical treatment regimen of different types of bone tumors (e.g. benign, malignant and intermediate). In practice, benign bone tumors have stable biological behavior, and more treatment regimens of local curettage or follow-up observation in the lesion are adopted, while malignant bone tumors are highly aggressive, and the treatment strategy of early adjuvant chemotherapy and extensive surgical resection can improve the survival rate of patients. There are many kinds of bone tumors, including more than ten different types of bone tumors around the knee joint, and the imaging manifestations are complex. As such, it is difficult for diagnostic experts or radiologists, especially intern doctors, to make accurate diagnosis due to lack of sufficient clinical experience, which obviously affects the clinical treatment effect and prognosis. For patients under 20 years old, primary malignant bone tumors rank as one of the leading causes of cancer-related deaths. Despite being relatively rare compared to other types of cancers, these aggressive bone tumors can have a significant impact on the health and well-being of young patients. Early detection, accurate classification, and timely treatment are crucial in improving the prognosis and overall survival rates for individuals affected by primary malignant bone tumors. In clinical diagnosis, digital radiography (DR), computed tomography (CT) and magnetic resonance imaging (MRI) are routine imaging methods. Artificial intelligence (AI) is a comprehensive frontier discipline that includes computer science, cybernetics, information theory, mathematics and other disciplines penetrating into each other. It involves studying and simulating human intelligence in order to expand it. In the field of medical imaging, the application of AI technology improves the speed of image interpretation and diagnosis, enhances accuracy and quality. Convolutional neural network (CNN) as a representative of it, can apply the image itself to the learning process. Therefore, there is no need to perform feature extraction before the learning process, and the most important function is that it can be automatically learned.

Deep learning is a branch of machine learning whose information transfer structure is similar to human neuron connections, with thousands of nodes from the shallow to the deep layers. It was originally derived from the artificial neural network (ANN) model. In the early stage, the ANN was limited by the computing power, so it could only input limited data and construct shallow neural network. However, with the breakthrough of computing power and the explosive growth of data, the constructed neural network becomes deeper, and the learning ability is getting stronger. Now it has been widely implemented in the fields of computer vision, natural language processing, speech recognition, etc. The most representative example is the impact of CNNs in the field of computer vision. The originator of convolutional neural networks is LeNet, which was proposed by Lecun in 1998 [1]. Subject to the hardware level at that time, LeNet is only a very small network, but it defines the basic structure of CNNs. In 2012, Krizhevsky and Hinton's Alex-Netwon [2] ImageNet's large-scale Computer Recognition Challenge by 10.9%, a huge success that brought CNNs back into the limelight. More advanced networks were developed, such as VGG [3], ResNet [4], InceptionNet [5] and DenseNet [6].

Machine learning comes from the earlier artificial intelligence, which needs to learn features extracted in advance to complete tasks. In some traditional big data analysis fields, machine learning is still the main analysis tool. At present, the research of AI in the imaging field of bone and joint system mainly focuses on the following aspects: (1) Bone age measurement: Related AI products for evaluating children's bone age based on deep learning technology have been widely researched. The accuracy of each product model in evaluating bone age is similar to or even better than that of radiologists, which can be completed in seconds from film reading to diagnostic report output; (2) Recognition and prediction of bone fracture: Foreign research teams focus on fracture interpretation and anatomical location in X-ray and CT examination, or combined with bone structure and bone density analysis to predict the risk of fracture, as well as to predict the risk of fracture in patients with cancer bone metastasis; (3) Osteoporosis: Multiple AI methods can help screen people at risk for osteoporosis or fracture, and AI can more accurately identify the risk of osteoporosis in postmenopausal women than traditional decision-making tools; (4) Identification of articular cartilage lesions: automatic segmentation of cartilage, detection of middle cartilage lesions (including chondromalacia, fibrosis, local defects, diffuse thinning caused by cartilage degeneration and acute cartilage injury) can be realized by using deep learning.

However, until now, few studies have been studied in the diagnosis of bone tumors. The reasons may be: (1) Bone tumors are relatively rare and the number of cases is insufficient; (2) There are many pathogenic sites, complicated disease types, poor data consistency and difficult model construction. All these have become obstacles to the promotion of AI-assisted diagnostic tools in the clinical application of osteoarthrosis. Therefore, it is of great clinical significance to construct an accurate and reliable auxiliary diagnostic tool for bone tumors. This topic takes bone tumors as the research object, and intends to provide a reliable auxiliary diagnostic tool for clinicians, especially junior physicians, by constructing a classification model for bone tumors, so as to improve the diagnostic efficiency. In practice, the Vision Transformer (ViT) is a deep neural network based on attention model, whose main feature is that it can effectively store global structure information of images. The overall contribution of our manuscript is below:

- We introduce the VGG16-ViT fusion model, effectively combining VGG-16 and Vision Transformer strengths to overcome CNN limitations and improve bone tumor classification performance.
- Our VGG16-ViT joint network achieves exceptional 97.6% accuracy in classifying bone tumors, offering precise identification of different tumor types for tailored treatment strategies.
- Through rigorous verification, we discover that secondary migration enhances system performance, providing valuable insights into the behavior and potential improvements of the proposed joint network.

Therefore, in this paper, the CNN and Vision Transformer network are combined to realize the classification of bone tumor CT images. This paper also explore the great potential of CNN and Vision Transformer in bone tumor image classification. Machine learning and deep learning techniques were used to construct a classification model of bone tumor, combined with clinical information, and film reading experiments were conducted.

2. Related work

The incidence of bone tumors is low, but primary malignant bone tumors rank third among the death [7] causes of cancer patients under 20 years old [8].

In 2020, the World Health Organization (WHO) released the fifth edition of its classification of bone and soft tissue tumors, marking a significant milestone in the field [9]. The primary objectives of this edition were to rectify grammar mistakes, address sentence errors, and adhere to scientific standards, thereby ensuring precision and clarity in

the classification process. The WHO aimed to enhance the classification's scientific rigor and practical applicability, aligning it with the latest advancements in tumor biology and pathology. This updated edition was meticulously crafted to integrate cutting-edge scientific knowledge, reflecting the constantly evolving landscape of tumor research. By establishing a standardized and globally accepted classification system, the fifth edition aimed to foster consistent reporting and communication among healthcare professionals, researchers, and pathologists. This consistency plays a pivotal role in elevating the quality of patient care and treatment outcomes, ultimately benefiting the medical community and the individuals they serve. Various types of bone tumors exhibit distinct biological characteristics. Benign bone tumors, for instance, are non-aggressive and typically do not recur or progress locally. Complete cure is often achievable through procedures like local excision or shaving, as seen in cases of bone cysts and osteochondromas.

Intermediate bone tumors, on the other hand, display localized invasive growth, causing destruction in the surrounding area. Incomplete local resection or curettage can lead to easy local recurrence, and there is a potential for malignant transformation, as observed in atypical chondrogenic tumors. Some intermediate tumors, like giant cell tumors of bone, may exhibit low-probability distant metastasis, with challenging histopathological predictions. These tumors possess characteristics intermediate between benign and malignant tumors, showing some degree of invasiveness and, albeit at a lower rate than malignant tumors, the potential to metastasize. They often exhibit cellular atypia and increased mitotic activity, indicating their intermediate nature in terms of cellular behavior. Histologically, their features may overlap with both benign and malignant tumors, making classification complex.

The prognosis for patients with intermediate bone tumors varies significantly, with some cases requiring more aggressive treatments than benign tumors. Accurate diagnosis and classification are paramount for determining appropriate clinical management. Recent advancements in deep learning techniques, including Explainable Deep Learning, have shown promise in enhancing the classification and understanding of these complex tumors. These advancements provide valuable insights, aiding in improved patient care and outcomes. In contrast, malignant bone tumors possess strong invasive capabilities and have a high likelihood of distant metastasis. They tend to recur after treatment, and their malignancy can further escalate upon recurrence, resulting in a poor prognosis. Examples of such tumors include osteosarcoma and Ewing's sarcoma [9].

Medical imaging is playing an increasingly important role in clinical diagnosis. It can provide medical imaging experts with direct or indirect information, including physiological structure, functional state, histological structure and pathological results of human tissues. In the image, benign bone tumors usually present as swelling growth masses with clear margins and visible sclerotic edges, generally without periosteal reaction and peripheral bone marrow edema, rarely breaking through the bone cortex, and soft tissue changes are not obvious.

However, malignant tumors show invasive growth with blurred edges, and acicular or flocculent tumor bone formation can be seen in some tumors. Because the rate of bone destruction is greater than the rate of repair, there is generally no osteosclerosis edge, and periosteal reaction often occurs, especially the scallion skin periosteal reaction, which breaks through the periosteal of bone cortex and destroys the proliferative periosteal, forming Codman triangle at both ends of the destruction. Surrounding bone marrow and soft tissue edema, tumor cell infiltration, the change is obvious. Intermediate bone tumors are between benign and malignant, and some tumors may have either sclerotic edges or no sclerotic edges, and periosteal reaction or no periosteal reaction [10], depending on their invasive ability [11]. Bone tumors are classified into the above three categories for the ultimate purpose of assisting and guiding clinical decision-making.

Therefore, there are great differences in the clinical treatment of different types of bone tumors. The biological behavior of benign bone tumors is stable. Intermediate tumors are locally aggressive, and doctors may choose more aggressive treatments such as extended excision to prevent local recurrence. As malignant bone tumors are highly invasive, comprehensive treatment is required, such as early chemotherapy [12], extensive surgical resection and postoperative radiotherapy, which can provide patients with a higher survival rate [13]. There are various incidence sites and diseases of bone tumors, and a variety of different types of bone tumors can occur just around the knee joint, with complex imaging manifestations. Due to the lack of sufficient clinical experience, it is difficult for diagnostic doctors, especially junior physicians, to make accurate diagnosis, which affects clinical treatment.

With the advent of the era of big data, the explosive growth of data and the rapid improvement of computer computing power make artificial intelligence gradually enter people's work and life. The field of AI was proposed as early as the 1950s. Its goal is to simulate human behavior, such as learning, thinking, reasoning, planning, etc. It is a science that integrates computer science, psychology, philosophy, linguistics and other disciplines. However, limited by the scientific and technological development level at that time, artificial intelligence has not been well developed. Since 2010, with the breakthrough of data storage technology and computing power, the field of artificial intelligence has made great progress. Various machine learning and deep learning models emerge endlessly. Its application also covers all aspects of work and life, including but not limited to visual recognition, natural language processing, intelligent search, reasoning, planning, etc. In the field of medical imaging, the application of AI technology can significantly improve the speed, accuracy and quality of image interpretation and diagnosis [14]. CNN as a representative, can apply the images themselves in the learning process [15], and the most important function is that they can learn automatically, without the need for feature extraction before the learning process [16].

Traditional machine learning algorithms can be traced back to the 1950s when Alan Turing suggested building learning machines. With the development of research in recent decades, abundant machine learning models have been developed. Compared with deep learning, its biggest feature is that it requires manual design and extraction of features in original data, that is, feature engineering. It requires people to find important features before model training, and therefore requires a lot of knowledge in related fields, which is a major bottleneck in machine learning. Compared with deep learning, traditional machine learning models are more suitable for relatively small, structured data. Common traditional machine learning models include logistic regression, decision tree, random forest, Naïve Bayes, K-nearest neighbor, support vector machine, ANN, etc.

According to the complexity of the model classification can be divided into simple model and complex model. Simple model is generally simple in calculation, weak in learning ability, but easy to understand, strong in interpretation, easy to cause the problem of underlearning. The representative models are decision tree and logistic regression model. Complex models generally have strong learning ability, but possess complex computation, poor interpretability, which is easy to lead to overfitting problems. The representative model is support vector machine model. Ensemble learning is a method to improve the learning ability of the model. It is to conduct random repeated sampling of the data, build multiple different weak learning models, and finally make fusion decisions on the results of multiple weak learners, such as minority obeying majority, so as to build a strong learner model. The representative model is random forest model. The learning ability of the same model is different, but not all models have better learning ability. Different models have different adaptability in different fields and data distribution. According to no free lunch theorem, no learning model can always be the most accurate in all fields, so the selection of models should be based on specific problems and specific analysis.

The most representative example is the impact of CNNs in the field of computer vision. The research of CNN begins to show an exponential

growth trend, and the model develops toward deeper layers and stronger learning ability. Compared with traditional machine learning, the biggest feature of these deep learning models is that they can automatically extract features without manual intervention, and their stronger learning ability makes them more suitable for larger scale data. Its disadvantages include large demand for data, long training cycle, easy overfitting when data is insufficient, and complex deep learning model, which is a "black box" model with poor interpretability. Therefore, to make reasonable use of artificial intelligence technology, it is necessary to have a sufficient understanding and familiarity with its applicable conditions, advantages and disadvantages.

Over the past few decades, CNN seems to have become the standard technique for medical image classification. With its high-quality classification accuracy, CNN is superior to recognition techniques based on traditional feature extraction, especially on large-scale datasets [17]. Khatamizad et al. [18] studied the use of VGG-16 and the v3 network to detect lesions and the results showed that VGG-16 was more effective. VGG [19] was proposed by the Visual Geometry Research Group at the University of Oxford. Its main contribution is to prove that the performance of the network can be improved by increasing the depth of the network. There are two types of VGG structures, which are the VGG16 and VGG19. While VGG-16 is more popular, VGG-19 has a deeper network, including 16 convolutional layers and 3 fully connected layers, that can extract more advanced features. Therefore, the VGG-16 network was chosen as the classification network for bone tumor CT images. The biggest feature of VGG networks is the use of smaller convolution kernels (33) instead of convolution kernels (55) and stacks of small convolution nuclei is used to achieve the same receptive field effect as large convolution kernel, but at the same time reduce the amount of computation. Although the performance of CNN is excellent, its limited local receptive field limits its performance, while the ViT network can extract global information well.

When the ViT network proposed by Dosovitskiy et al. [20] is trained on large-scale data sets, its image classification accuracy is better than the most advanced CNN network performance. Although the ViT network lacks the inductive bias of the CNN network, the result can be competitive with the most advanced CNN network. However, a large number of data and computing resources are the main reason limiting the development of ViT. Therefore, Touvron et al. [21] introduced distillation mechanism to train ViT network and improve the classification performance of ViT. Considering the high performance of ViT network, researchers have made attempts in various fields. In the field of object detection, Carion et al. [22] proposed a new object detection system architecture and tested it on COCO data of public data set. The results are comparable to the performance of the most advanced CNN method. In the field of image segmentation, a novel network U-Net Transformer network shows a competitive advantage, they use Transformer and CNN respectively for encoder and decoder parts. Dai et al. [23] combined CNN and Trans former network, used CNN to extract local features and Transformer's self-attention mechanism extracts global features and applies the algorithm to multi-mode image classification. The results show that its performance is better than the most advanced CNN network. Although many efforts have been made to improve models based on ViT, the fine-tuning of different model sizes and weights is still a problem to be solved when applying ViT models [24].

3. Materials and methods

3.1. Construction of the dataset

In pursuit of a rich and diverse dataset, we procured a comprehensive collection of 786 computed tomography (CT) images of bone tumors. This dataset comprises sections from two prominent bone species, the tibia and femur, allowing for a comprehensive examination of bone tumor cases in different anatomical regions. To ensure a reliable and representative dataset, the images were meticulously sourced from The Second Affiliated Hospital of Fujian Medical University, a renowned medical institution renowned for its expertise in bone tumor research and diagnosis. Our dataset encompasses an extensive variety of bone tumor cases, providing a wide range of instances for analysis and study. It includes CT images of various bone tumor types, including osteosarcomas, which pose unique challenges and complexities in clinical diagnosis and management. By encompassing diverse pathologies and anatomical locations, our dataset facilitates a thorough investigation of bone tumor characteristics and enables the development of robust classification models.

Acoustic image datasets often suffer from various types of noise, such as speckle noise, random noise, and Gaussian noise, which can degrade the quality and accuracy of the images. To address the issue of noise in our acoustic image dataset, we applied spatial filtering techniques. Spatial filtering is a common method used to enhance image quality by removing noise while preserving important features. In our study, we employed median filtering and Gaussian filtering to reduce the impact of speckle noise and random noise, respectively. Median filtering is effective in smoothing the image while preserving edges, making it suitable for speckle noise reduction. On the other hand, Gaussian filtering is effective in reducing random noise and blurring the image slightly. By applying these spatial filtering techniques, we aimed to improve the overall quality of the acoustic images and enhance the performance of our deep learning model in bone tumor classification.

3.2. Image preprocessing

We reviews and frames the region of interest (ROI) on the bone tumors images. The range of ROI is the uppermost, lowermost and both sides of the tumor; We cut out the lesion taken by the frame, then scaled the long side of ROl to 512 pixels, the short side proportional to the long side, and then filled them to 512 pixels using the pixel minimums and made ROI uniform size of 512 × 512, and then stacked the image 2 layers repeatedly so that it become a 3-channel image; Finally, the ROI pixel values are normalized and scaled to a mean of 0 and a variance of 1.

3.3. The VGG architecture

To extract spatial features from the slices of each sample and convert them into high-level feature representation tokens, a VGG-16 based CNN is employed as shown in Fig. 1, which illustrates the architecture of the VGG-16 based CNN. The model integrates 13 convolutional layers and 5 pooling layers from VGG-16 and introduces an additional convolutional layer at the beginning to expand the slice's channels and another at the end to facilitate the mapping from feature map to token.

Following dimension expansion, the slices are resized to 112×112×3. Within the VGG-16 framework, a series of convolution, activation, and pooling operations generate an output feature map of size 3×3×512. Lastly, a 3×3 convolution layer with input-channel of 512 and output-channel of 256 carries out the crucial mapping from feature map to token. The standard convolution calculation process formula of the feature map I to the feature map O is given by

$$O_j = ReLU\left(\sum_{i \in M_j} I_i * K_{ij} + b_j\right) \tag{1}$$

In the context of this expression, O_j represents the j^{th} channel of the feature map O, and I_i denotes the i^{th} channel of the feature map I. The set M_j comprises the channels in the feature map I. The term K_{ij} corresponds to the convolutional kernel related to I_i and O_j, while b_j stands for the bias offset of O_j after the convolution operation. The convolutional calculation is denoted by *. The activation function utilized is ReLU, which effectively converts negative inputs to 0, while leaving positive values unaffected. Through the convolution operation, the VGG-16

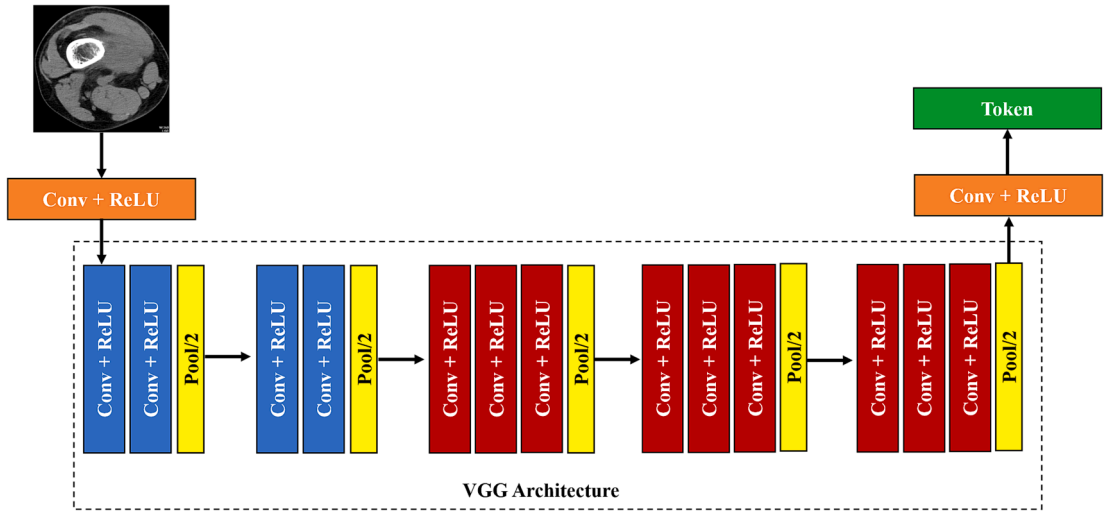

Fig. 1. VGG-16 based CNN architecture.

model efficiently extracts essential features from the input image. Notably, each convolution kernel captures distinct traits, and the convolution layer groups progressively extract 64, 128, 256, and 512 local features. The final convolutional layer plays a significant role by filtering 256 out of the 512 features, thereby establishing the feature representation for the specific slice.

By conducting convolution mapping on *N* slices from both slice series T1 and slice series T2, we obtain two token series: token series T1 and token series T2. As a result, each sample comprises a total of 2 N tokens. For incorporating temporal position information within each pair of tokens in token series T1 and T2, sinusoidal position encoding [22] is employed. Moreover, spatial position information is embedded separately within token series T1 and token series T2. Following the spatial and temporal position embedding, the 2 N tokens progress to the first temporal attention block.

3.4. The model based on VGG-16 and Vision Transformer

In the context of image classification, CNNs have some limitations, particularly in capturing global structural information. CNNs are known to prioritize learning from majority class samples in imbalanced datasets, leading to biased classifications and potentially overlooking crucial information from minority class samples. In our study, we addressed these limitations by proposing an optimized deep learning algorithm that combines the strengths of VGG16 and ViT. This fusion model, named VGG16-ViT, effectively overcomes the limited perception ability of traditional CNNs and provides a more comprehensive understanding of global information in bone tumor images. By leveraging the advantages of VGG16 and ViT, our model achieved improved classification accuracy, mitigated the impact of data imbalance, and enhanced the performance of the classification task, demonstrating its potential in advancing image classification in the biomedical domain.

The VGG16 structure comprises 16 convolutional layers and 3 fully connected layers, making it a deep convolutional neural network. The output of the convolutional layer and the fully connected layer can be expressed as:

$$a^{[l]} = g^{[l]}\left(w^{[l]}a^{[l-1]} + b^{[l]}\right) \tag{2}$$

The ViT is a kind of visual converter based on self-attention mechanism. Its basic structure is composed of multiple Transformer modules. Each Transformer module is composed of multi-head attention mechanism and feedforward neural network, which can be expressed as:

$$h^{[l]} = \mathrm{MultiHeadAttention}\left(x^{[l]}\right) + x^{[l]} \tag{3}$$

$$x^{[l]}x^{[l+1]} = \mathrm{LayerNorm}\left(h^{[l]}\right) + \mathrm{FFN}\left(h^{[l]}\right) \tag{4}$$

$$h^{[l+1]} = \mathrm{MultiHeadAttention}\left(x^{[l+1]}\right) + x^{[l+1]} \tag{5}$$

$$y = \mathrm{softmax}\left(x^{[L]}\right) \tag{6}$$

Fig. 2 depicts the architecture of the VGG-TSwinformer presented in this study. Each slice is mapped to a high-level feature representation token by a VGG-16-based CNN. Finally, the first temporal attention block receives token series T1 and token series T2 representing slice series T1 and slice series T2, respectively. The token series T1 and the token series T2 in VGG-TSwinformer each contain 10 attention blocks, of which the token series T1 and the token series T2 share 5 temporal attention blocks and have 5 spatial attention blocks in common. The first four spatial attention blocks of token series T1 and series T2 are alternatively designed into right-sliding window (RSwin) attention block and left-sliding window (LSwin) attention block in order to better incorporate local features and attempt to avoid dividing the same redundant tokens.

To enhance the extraction of local longitudinal features, we implement feature fusion on the axial slices corresponding to CT imaging data. Illustrated in Figure 2, this process involves employing a modified sliding-window attention mechanism (SWA) to the tokens within the corresponding positions of token series derived from CT images.

Our proposed SWA allows indirect feature fusion of tokens within a specific window range of the CT image series. Consequently, the feature fusion in CT imaging extends from the corresponding 2D slice to the local 3D space. The correlations between features at similar spatial distances often hold more significance than those at different spatial distances. For CT datasets, the suggested temporal attention and sliding-window attention mechanisms simplify the model's ability to detect changes in local characteristics during limited iterative training. This enhanced capability enables the model to exploit these changes in prediction, contributing to the accuracy of the overall analysis. The complete process of RSwin and LSwin calculation is shown in Algorithm 1 and 2.

As depicted in Fig. 2, MSA is carried out for each token in token series T2 during the final block of spatial attention. The calculation procedure for token series T1 is the same as the calculation procedure for token series T2. Table 1 displays the configuration information for four sliding-window attention blocks from token series T1 and token series T2. To obtain the final prediction, the classifier receives an average of the output tokens from the last block of spatial attention for token series T1 and token series T2.

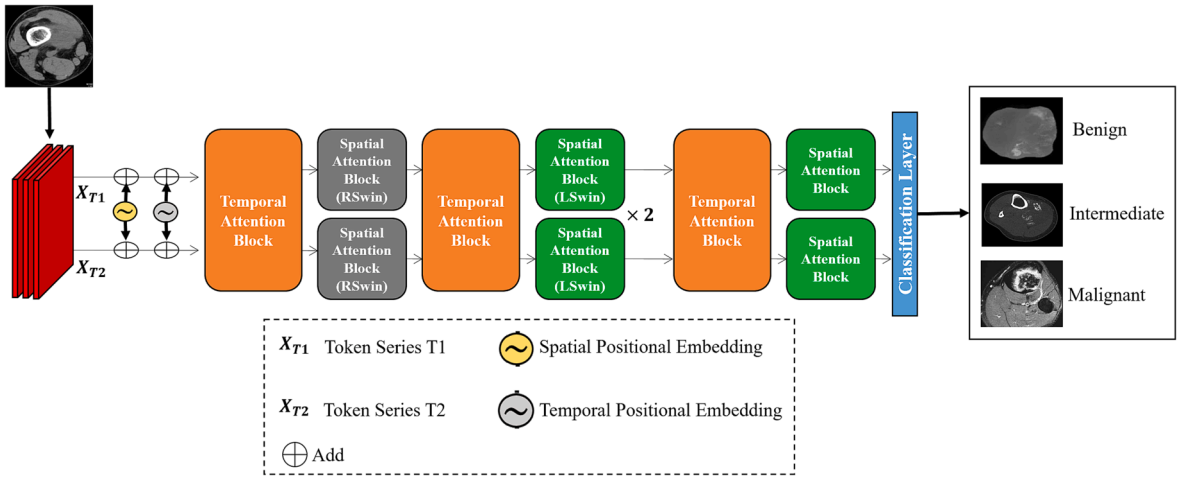

Fig. 2. VGG-TSwinformer model architecture.

Algorithm 1. RSwin block calculation process.

Input: $X^l_{T2} = (X^l_{(T2,1)}; \cdots; X^l_{(T2,N)})$

Output: X^{l+1}_{T2}

Num = ceil($\frac{N-w+1}{s}$)

M← zero(num + 1,N,C)

i← 1

while(i≤ *num*) **do**

W_i←MSA($X^l_{(T2,(i-1)s+1)}; \cdots; X^l_{(T2,(i=1)s+w)}$)

M[i-1, (i-1)s: (i-1) s + w-1,:] ←W_i

i←i + 1

end while

if((num-1)s + w!=N) **then**

W_i←MSA($X^l_{(T2,(i-1)s+1)} X^l_{(T2,(i-1)s+1)}; \cdots; X^l_{(T2,(i=1)s+w)}$)

M[i-1, num*s:N-1,:]←W_i

end if

count← (M[:,:,C]!=zero(C)).sum(axis = 0)

t←M.sum(axis = 0)

X^l_{T2}←MLP(LN(t)) + t

return X^l_{T2}

For the overall architecture model, we performed the following. Firstly, the image features or image data are divided into multiple patches for input, then the patches are flattened, and the location and category labels are added to the patches, which are then sent to the encoder part of Transformer network. Finally, the output results are sent to the MLP module for weighted summing classification.

Guo et al., [24] provides many ViT models, including ViT-B /32, ViT-B /16, ViT-L /32, ViT-L /16 and ViT-H /14. The order of these models is from small to large. Considering the small data set, ViT version based on ViT-B/16 was selected to complete the corresponding task of bone CT classification.

In this study, we propose a novel VGG16-ViT model for bone tumor classification by combining the VGG16 and ViT architectures. The process involves selecting 27 features from the third layer feature set of VGG16, which contains a total of 13,696 features, to achieve the best classification performance. The selected features are obtained from the first convolution layer. The overall architecture of the VGG16-ViT model is as follows: Firstly, the image data is divided into multiple patches [25], which are then flattened and provided with location and category labels. These patches are fed into the encoder part of the Transformer network. Finally, the output results from the Transformer network are sent to the Multi-Layer Perceptron (MLP) module for weighted summing classification. Given the relatively small dataset, we chose the ViT version based on ViT-B/16 for bone CT classification.

3.5. *Training details*

In the model training, the results of the single VGG16 model are not ideal, which may be due to the existence of redundant or irrelevant information in the extracted features. In order to solve this problem,

Algorithm 2. LSwin block calculation process.

Input: $X^{l}_{T2} = (X^{l}_{(T2,1)}; \cdots; X^{l}_{(T2,N)})$

Output: $X^{l+1}_{(T2,)}$

Num = ceil($\frac{N-w+1}{s}$)

M← zero(num + 1,N,C)

i← 1

while(i≤ *num*) **do**

W_i←MSA($X^{l}_{(T2,N-(i-1)s-w+1)}; \cdots; X^{l}_{(T2,N(i=1)s)}$)

M[i-1, N-(i-1)s-w: N-(i-1) s-1,:]←W_i

i←i + 1

end while

if(N-(num-1)s-w + 1!=1) **then**

W_i←MSA($X^{l}_{(T2,1);\cdots;X^{l}_{(T2,N-(i-1)s)}}$)

M[i-1, 0:N-num*s-1,:]←W_i

end if

count← (M[:,:,C]!=zero(C)).sum(axis = 0)

t←M.sum(axis = 0)

X^{l+1}_{T2}←MLP(LN(t)) + t

return X^{l+1}_{T2}

Table 1
Configuration details of four sliding-window attention blocks of token series T1 and token series T2.

Block num	Block name	Window sliding orientation	Window size	Sliding stride
2	RSwin	right	5	4
4	LSwin	left	9	8
6	RSwin	right	17	16
8	LSwin	left	33	32

feature selection should be carried out for the extracted features, keeping the features with good classification performance and discarding the features with poor classification performance. The traditional method is to use search to achieve enumeration, such as wide search and deep search. However, when the number of features is relatively large [26,27], such as the number of features in VGG-16 is 13, 696, the search will be time-consuming. Therefore, the mRMR (the minimum redundancy maximum correlation) method of feature selection proposed by Peng et al., [28] is selected to find the optimal combination of features. Firstly, the features should be ranked, and the top *M* features should be selected to form a feature group. Then, the Matthews correlation coefficient is used to evaluate the performance of the feature group. Finally, the minimum number of feature groups is found as candidate feature groups. In the second stage, it is necessary to use SVM classification to complete the further screening of features. The screening method is search method. After feature selection in the first stage, the number of feature groups is greatly reduced.

As the window size increases, the accuracy of SVM classification decreases, with better performance observed for smaller window sizes. Results of SVM using different kernel functions are not significantly different when using the same window scale, but polynomial kernel SVM performs slightly lower than linear kernel and radial basis kernel SVM. Additionally, sample training was carried out with C =1, 10, 100, 1000, 10,000 and σ^2 =0.01, 0.125, 0.5, 1.5, 10. After conducting experiments, it was discovered that the optimal range for parameter C is between 100 and 1000, while the range for σ^2 is between 0.01 and 0.125, resulting in the highest cross-validation accuracy.

The reduction of parameters in our proposed model was achieved through a two-step process. Firstly, we performed feature selection to identify the most informative and relevant features from the original dataset. This step involved evaluating the importance of each feature using techniques such as correlation analysis and feature ranking. The selected features were then retained for further processing, while irrelevant or redundant features were discarded. This feature selection process significantly reduced the dimensionality of the input data, leading to a more efficient and streamlined model.

Secondly, we employed a compression technique to further reduce the number of parameters in the model. Specifically, we utilized a quantization approach to approximate the weights and activations of the neural network. This technique allowed us to represent the model parameters using fewer bits, thereby reducing the memory footprint and computational complexity of the model. By employing both feature selection and compression, we were able to achieve a significant reduction in the number of parameters without compromising the model's performance.

3.6. Evaluation index of classification

In binary classification, the performance is quantitatively determined by sensitivity, specificity, and accuracy. The classification accuracy of a single VGG16 or ViT network, as well as the VGG16-ViT network combination, is measured by comparing the F-value. Equation (7) indicates the accuracy rate of the test set, reflecting the proportion of samples that are correctly classified. Equation (8) shows the sensitivity rate, which represents the proportion of accurately classified positive samples. Equation (9) shows the specificity rate, which represents the proportion of accurately classified negative samples. Additionally, transfer learning is incorporated to enhance the learning ability of the model, as shown in Equation (10). These metrics are calculated using

True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \tag{7}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FN} + \text{FP}} \times 100\% \tag{9}$$

$$\text{F1} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{10}$$

In this paper, TP is the count of correctly classified positive samples, and in this case the number of accurately classified CT images of bone tumors. FP indicates the number of negative samples that were misclassified as positive, or in other words, the number of CT of bone tumors that were misclassified. TN, on the other hand, indicates the number of negative samples that were correctly classified, or the number of images of non-bone tumors that were accurately classified. Finally, FN refers to the number of positive samples that were misclassified as negative, in which case it corresponds to the number of non-bone tumor images that were misclassified.

In addition to the aforementioned evaluation metrics, the Receiver Operating Characteristic (ROC) curve was also employed as an evaluation criterion. The model's effectiveness was evaluated by comparing its Area Under Curve (AUC) values, which were obtained by plotting a ROC curve. The ROC curve is a popular method for assessing classifier performance by balancing true positive and false positive error rates. AUC is a common performance metric that is derived from the ROC curve. In practice, AUC is indicative of a classifier's ability to differentiate between samples.

4. Experimental results and discussion

4.1. Datasets

The dataset used in this study contained image data and clinical data from 568 patients from the Second Affiliated Hospital of Fujian Medical University. In the model training and validation testing process, we used bone tumor 786 images from these patients. In the testing and verification phase of this experiment, we used 286 images based on 244 patients obtained from the Xiangya Hospital. After inclusion and exclusion screening, according to the fifth edition of the World Health Organization's preliminary classification of bone tumors, 786 images were classified as benign, malignant and intermediate bone tumors (see Fig. 3), and all lesions were pathologically confirmed.

The first hospital group of 568 patients were randomly assigned to the training set and validation set according to the ratio of 9:1, the training set was used to update the weight parameters of the model, the verification set was used to monitor the training effect of the model, and the model was saved as a reference, and the ROI of each patient only existed in one of the three datasets, and there were no duplicate cases between the three.

The ROI is enhanced online before entering the model, randomly flipping the ROI horizontally or vertically, and randomly rotating

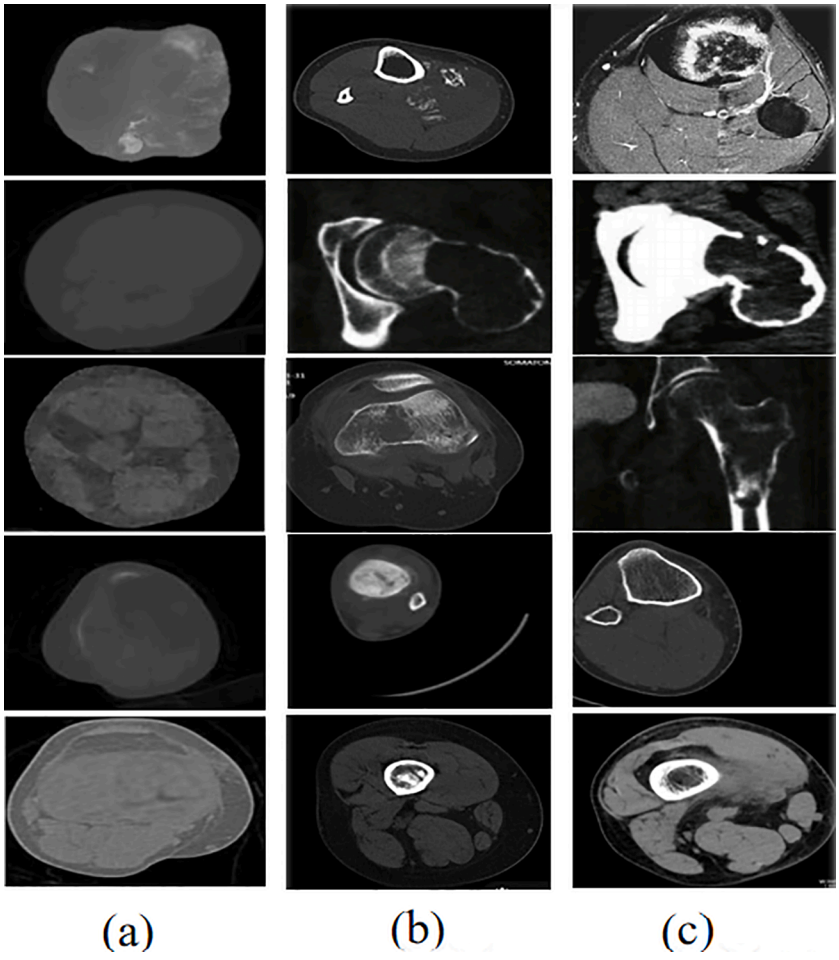

Fig. 3. The datasets containing CT images of the tibia and femur are partially shown. The (a) part indicates benign bone tumors; the (b) part indicates malignant bone tumors; the (c) part indicates intermediate tumors.

Table 2
Results of the model on dataset based on VGG16 and ViT-B/16.

Model	Accuracy rate	Sensitivity	Specificity	AUC value
VGG-16	91 %	89 %	83 %	0.91
ViT-B/16	93.1 %	91.5 %	85.6 %	0.93

Table 3
Results of the ViT model on the enhanced dataset.

Model	Accuracy rate	Sensitivity	Specificity	AUC value
ViT-B/16	94.1 %	92 %	88.7 %	0.95

between 0° and 25° to mitigate the overfitting of the model and enhance its generalization ability. A total of 244 patients and 286 CT images from the second hospital group were included in this study, and assigned for verification.

4.2. Results of VGG-16

This experiment completed the binary task, that is, the classification task of malignant tumor and benign tumor. The data set was divided into training set and test set, in which training set accounted for 80 % and test set accounted for 20 %. In order to ensure fair comparison, all experiments were conducted on this data set. The average value of 10 times cross validation was used as the final evaluation result. Among them, advanced VGG16 network was selected for CNN network, and the model pre-training weight provided in the literature was used to initialize the network. Adam optimizer was selected for the optimizer, and 100 epochs were trained for the model. The ViT model employed the cross entropy loss function to address the imbalanced nature of the dataset. The SGD optimizer was selected, the momentum was set to 0.9, the batch size was set to 128, the initial learning rate was set to 0.001, and 100 epochs were trained. The final results of the separate training network are shown in Table 2.

As can be seen from Table 1, based on the convolutional network VGG-16 model, the accuracy is 91 % and the AUC value is 0.91. The accuracy of the ViT model is more than 91 %, and the AUC value is greater than 0.91, which exceeds that of the VGG-16 model of convolutional network. However, compared to the ViT model in the original document [24], the dataset in this paper is much smaller than the original document. The reason for the better effect may be that, unlike natural images, the spatial or global information of tumor in CT of bone tumors seems to be more important. In order to further prove the above conjecture, the authors completed the expansion of the dataset and enhanced the dataset accordingly through subsequent experiments, including cropping, rotation, brightness and contrast changes. The results, shown in Table 3, show that data augmentation does not improve the performance of the ViT model, but rather affects its performance. (See Table 4).

4.3. VGG-16 model fusion with ViT

The ten-fold cross-validation method was adopted in the experiment, and the average value was taken as the final result. The results show that the feature set of the third layer has the highest classification performance, with an accuracy of 95.2 %. In fact, the feature selection process can process 13,696 features including the third layer feature set, from which 27 features are selected for combination, thus achieving the highest classification performance. The 27 feature combinations come from the first convolution layer.

In order to integrate the respective advantages of VGG and ViT models, the framework is used to select 27 features from the third layer features in VGG16 as patch input and input them into the ViT encoder for training.

Table 5
Results of VGG16-ViT model on dataset.

Model	Accuracy rate	Sensitivity	Specificity	AUC value
VGG16-ViT	97.6 %	93 %	90.7 %	0.97

Table 6
Results providing a comparison of the three models

Model	Accuracy rate	Sensitivity	Specificity	AUC value
VGG-16	91 %	89 %	83 %	0.91
ViT-B/16	93.1 %	91.5 %	85.6 %	0.93
VGG16-ViT	97.6 %	93 %	90.7 %	0.97

Table 5 suggests that the VGG16-ViT model achieved higher accuracy on the dataset, which may be attributed to the relatively small sample size. This suggests that the convolutional network's inductive bias can facilitate faster convergence of the model. See (Table 6).

In this section, we present the evaluation of our model's performance through accuracy and loss metrics. We plot the accuracy against the loss to gain insights into the model's convergence and overall performance, as shown in Fig. 4. Additionally, we assess the model's classification performance using a confusion matrix, which provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions for each class, offering a comprehensive view of the model's effectiveness.

4.3.1. Comparison with other models

In this section, we conduct a thorough evaluation of three models: VGG-16, ViT-B/16, and our proposed framework. We compare their performance using key metrics such as accuracy, sensitivity, and specificity, which provide valuable insights into their classification capabilities. The results of this comparative analysis are presented in Fig. 5, where we analyze the models' accuracy in making correct predictions, sensitivity in correctly identifying positive cases, and specificity in accurately recognizing negative cases.

Fig. 5 demonstrates that networks without transfer learning exhibit poorer performance compared to networks with transfer learning. In particular, Fig. 5 (c) indicates that the VGG16-ViT fusion network outperforms the VGG16 or ViT network alone. Notably, there is no sign of overfitting in the network with transfer learning, as the loss function is reduced on both the training and test sets. Therefore, employing a deeper network on a small dataset and incorporating transfer learning can enhance the system's classification performance.

Table 4
Feature selection classification performance.

Feature level	Feature set	The number of features selected	Specific characteristics	Accuracy rate
Third layer feature	CONV	27	CONV3_3(A feature) CONV4_2(Four features) CONV4_3 (Two features) CONV4_4(A feature) CONV5_1 (Five features) CONV5_4 (Fourteen features)	95.2 %

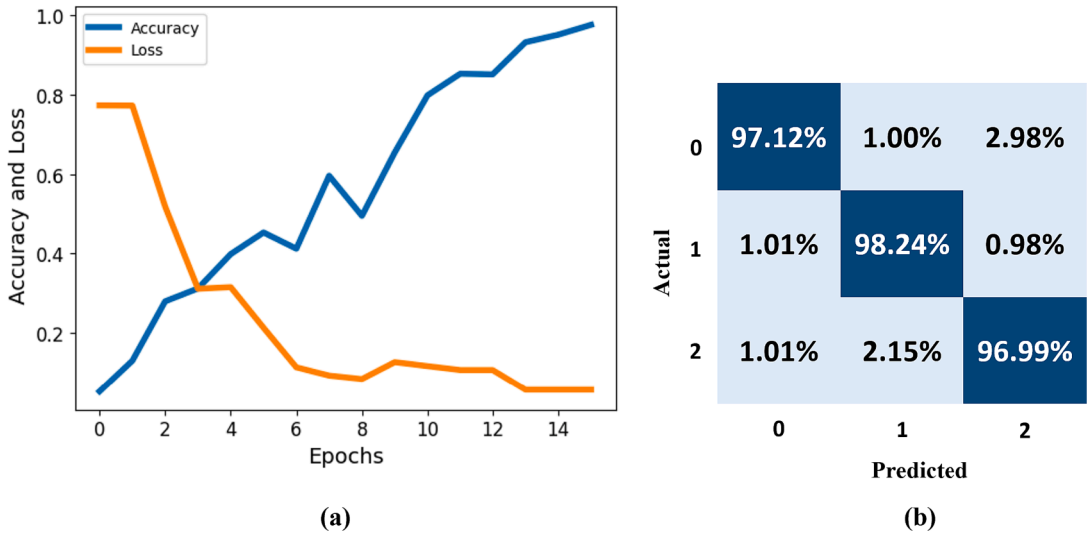

Fig. 4. Performance evaluation of proposed framework. Whereby (a) represents the accuracy and loss analysis, while (b) is the confusion metrics.

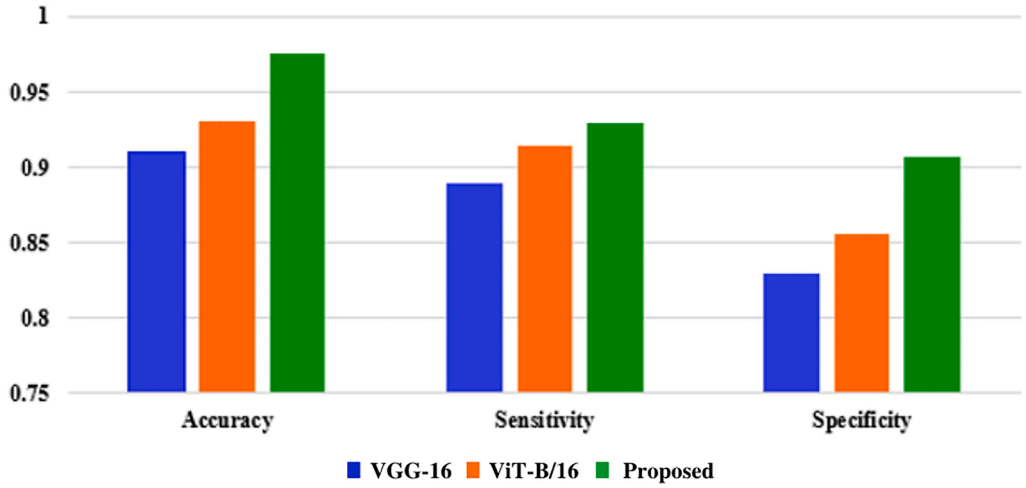

Fig. 5. The comparitive analysis among three models with different evaluation metrics.

5. Discussion

The characteristics of bone tumors varies widely, and different biological behaviors often bring different prognosis and affect their treatment options. Since the fourth edition of the WHO Classification of Bone and Soft Tissue Tumors published in 2013, it has clearly classified bone tumors as benign, intermediate and malignant, and divided different types of bone tumors according to their biological behaviors, reflecting the research progress of scientists in recent decades on bone and soft tissue tumors and tumor-like lesions in clinical, pathological, molecular biology and prognosis. Based on this, the fifth edition of the WHO Classification of Bone and Soft Tissue Tumors published in 2020 modified the classification of some types of tumors to further understand bone tumors, making them more conducive to guiding clinicians to choose appropriate treatment options for lesions. This paper analyzes the ability of CNN in image classification. The limited local sensitivity field of CNN restricts its ability to obtain global information, which makes the performance of these convolutional networks poor in obtaining global structural information and limits their visual recognition ability. Therefore, a network of ViT was introduced. The ability of the network to capture global information was improved, and finally the VGG-16 network was combined with the ViT model to complete the classification task of bone tumors images. Sensitivity, specificity, accuracy and ROC curves were selected for comparison and analysis with the above CNN structure model. The training method combining VGG-16 network and ViT model can improve the accuracy of image classification. Experimental results show that the proposed method has good universality. Compared to other methods in recent years, the results show that this method has the best performance in end-to-end fully automated networks. Although deep learning has been widely used in the field of medical image processing, there are still many bottlenecks and improvements that need to be broken through in the medical field.

First of all, although deep learning has strong generality, it also has high requirements on data and needs a lot of data due to the privacy and professionalism of medical images, the development of this field is more difficult. When the amount of data is sufficient, the deeper the network is, the better the performance will be. However, the higher the complexity of the network, the higher the parameter number and calculation consumption will increase, the training speed will be greatly slowed down, and the requirements on hardware equipment will also be increased. In order to apply the algorithm to clinical practice generally, more efficient and lightweight networks need to be designed in future studies.

6. Conclusion

The proposed VGG-16 and visual transformer joint network, known as VGG16-ViT, demonstrates promising classification performance on bone tumor datasets, achieving an accuracy rate of 97.6 % and effectively distinguishing different types of bone tumors. The optimization of sensitivity and specificity leads to an 8 % improvement, while the study reveals that secondary migration can further enhance the system's performance. One limitation is that the dataset used for training and verification is collected from a specific hospital, which may introduce bias and limit the generalizability of the model to other medical institutions. The small size of the dataset (786 CT images) could also affect the algorithm's robustness, and there might be a need for a more extensive and diverse dataset for further validation. Moreover, the study does not compare the performance of the proposed VGG16-ViT model with other state-of-the-art algorithms in bone tumor classification, which could provide a more comprehensive evaluation of its effectiveness. Future work should focus on acquiring larger and more diverse datasets, comparing the algorithm with other state-of-the-art methods, and incorporating advanced deep learning techniques to enhance classification performance. Prospective studies on a larger cohort of bone tumor patients will provide valuable insights into the algorithm's practical utility in clinical settings, enabling early detection and precise prognosis.

Ethical approval

All human subjects in this study have given their written consent for the participation of our research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Lecun, L. Bottou, Gradient-based learning applied to document recognition [J], Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[2] Deng, Tiancan. "A survey of convolutional neural networks for image classification: Models and datasets." In 2022 International Conference on Big Data, Information and Computer Network (BDICN), pp. 746-749. IEEE, 2022.

[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[4] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[5] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.

[6] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017.

[7] J.S. Bienrmann, W. Chow, D.R. Reed, et al., NCCN guidelines insights: bone cancer, version 2.2017 [J], J. Natl. Compr. Canc. Netw. (2017:) 1540–11413.

[8] Miller, Kimberly D., Leticia Nogueira, Angela B. Mariotto, Julia H. Rowland, K. Robin Yabroff, Catherine M. Alfano, Ahmedin Jemal, Joan L. Kramer, and Rebecca L. Siegel. "Cancer treatment and survivorship statistics, 2019." CA: a cancer journal for clinicians 69, no. 5 (2019): 363-385.

[9] WHO Classification of Tumours Editorial Board. Soft tissue and bone tumours. Lyon (France): International Agency for Research on Cancer 2020 [J]. WHO classification of tumours series, 2020 vol 3.

[10] T. Miller, Bone tumours and tumorlike conditions: analysis with conventional radiography [J], Radiology 246 (3) (2008) 662–674.

[11] C.M. Costelloe, J.E. Madewell, Radiography in the initial diagnosis of primary bone tumors [J], AJR Am. J. Roentgenol. 200 (1) (2003) 3–7.

[12] H. Fritzsche, K.D. Schaser, S. Hofbauer, Benign tumours and tumour-like lesions of the bone: general treatment principles [J], Orthopade 46 (6) (2017 Jun) 484–497.

[13] C.J. Gutowski, A. Basu-Mallick, J.A. Abraham, Management of bone Sarcoma [J], Surg. Clin. North Am. 96 (5) (2016 Oct) 1077–1106.

[14] J.C. Gore, Artificial intelligence in medical imaging [J], Magn. Reson. Imaging 68 (2020) A1–A4.

[15] Y.L. Cun, B. Boser, J. Denker, et al. Handwritten digit recognition with a backprogation network [J]. Advances in Neural Information Processing System, 1990.

[16] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, O. Abe, Deep learning with convolutional neural network in radiology [J], Jpn. J. Radiol. 36 (4) (2018) 257–272.

[17] Y. Zhou, J. Xu, Q. Liu, et al., A radiomics approach with CNN for shear-wave elastography breast tumor classifification, IEEE Trans. Biomed. Eng. 65 (9) (2018) 1935–1942.

[18] A. Hatamizadeh, Y. Tang, V. Nath, et al., Unetr: transformers for 3d medical image segmentation[C]//Proceedings of the IEEE/CVF, in: Winter Conference on Applications of Computer Vision, ACM, New York, 2022, pp. 574–584.

[19] Q. Guan, Y. Wang, B. Ping, et al., Deep convolutional neural network VGG-16 model for diffffferential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study, J. Cancer 10 (20) (2019) 4876–4884.

[20] W. Al-Dhabyani, M. Gomaa, H. Khaled, et al., Deep learning approaches for data augmentation and classification of breast masses using ultrasound images, Int. J. Appl. Math. Comput. Sci. 10 (5) (2019) 1–11.

[21] H. Touvron, M. Cord, M. Douze, et al., Training data effificient image transformers & distillation through attention[C], in: International Conference on Machine Learning. PMLR, IEEE, Piscataway, NJ, 2021, pp. 10347–10357.

[22] N. Carion, F. Massa, G. Synnaeve, et al., End-to-end object detection with transformers[C], in: European Conference on Computer Vision, Springer, Berlin, 2020, pp. 213–229.

[23] Y. Dai, Y. Gao, F. Liu, Transmed: transformers advance multi-modal medical image classifification, Diagnostics 11 (8) (2021) 1384–1399.

[24] M.H. Guo, T.X. Xu, J.J. Liu, et al., Attention mechanisms in computer vision: a survey, Computational Visual Media (2022) 1–38.

[25] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.

[26] A. Liu, J. Ghosh, C.E. Martin, Generative oversampling for mining imbalanced datasets. [C]. In DMIN, 2007: 66-72.

[27] R. Longadge, S. Dongre. Class imbalance problem in data mining review . arXiv preprint arXiv:1305.1707, 2013.

[28] A. El Akadi, A. Amine, A. El Ouardighi, D. Aboutajdine, A two stage gene selection scheme utilizing MRMR filter and GA wrapper, Knowl. Inf. Syst. 26 (3) (2011) 487–500.