NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2024; 13(2), 704-712

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Performance comparison of visual transformer based models for shoulder implant classification

Omuz implantı sınıflandırmasında görü dönüştürücü tabanlı modellerin performans karşılaştırması

Elif Baykal Kablan[1,*], Yavuz Kablan[2]

[1] *Karadeniz Technical University, Software Engineering Department, 61080, Trabzon, Türkiye*
[2] *Karadeniz Technical University, Electrical and Electronics Department, 61080, Trabzon, Türkiye*

Abstract

Total shoulder arthroplasty (TSA) is a surgical procedure addressing severe pain and restricted shoulder joint movement. During TSA surgery, X-ray images guide the selection of the prosthetic implant suitable for the patient from a variety of models produced by different manufacturers. However, prostheses may wear or loosen over time, thus requiring periodic evaluation and replacement. Currently, the process involves taking new X-ray images from patients, resulting in variability in expert opinions on implant types. Therefore, there is a need for highly accurate automated diagnostic systems to help recognize unknown implants. In this study, we present a performance comparison of vision transformer (ViT) based models for automatic shoulder implant classification from X-ray images. Fine-tuning of pre-trained ViT models on a publicly available shoulder X-ray dataset showed high success in terms of accuracy, precision, sensitivity, and F-measure metrics. The Swin-B model yielded the highest results with 93.84\% accuracy, 88.15\% precision, and 85.52\% recall. These results showed that ViT based models can help improve treatment planning by providing reliable identification of shoulder implant manufacturers and model information and time efficiency, especially for specialists.

Keywords: Total shoulder arthroplasty, Shoulder implants, X-ray, Classification, Vision transformer

Öz

Total omuz artroplastisi (TSA), şiddetli ağrı ve kısıtlı omuz eklemi hareketini ele alan cerrahi bir prosedürdür. TSA ameliyatı sırasında röntgen görüntüleri, farklı üreticiler tarafından üretilen çeşitli modeller arasından hastaya uygun protez implantın seçimine rehberlik etmektedir. Bununla birlikte, protezler zamanla aşınabilir veya gevşeyebilir, bu nedenle periyodik değerlendirme ve değiştirme gerektirmektedir. Halihazırda bu süreç, hastalardan yeni röntgen görüntülerinin alınmasını gerektirmekte ve implant tiplerine ilişkin uzman görüşlerinde değişkenliğe neden olmaktadır. Bu nedenle, bilinmeyen implantları tanımaya yardımcı olacak yüksek doğrulukta otomatik teşhis sistemlerine ihtiyaç vardır. Bu çalışmada, X-ray görüntülerinden otomatik omuz implantı sınıflandırması için görü dönüştürücü (ViT) tabanlı modellerin performans karşılaştırması sunulmaktadır. Önceden eğitilmiş ViT modellerinin herkese açık bir veri kümesi üzerinde ince ayarı doğruluk, hassasiyet, duyarlılık ve F-ölçümü metriklerinde yüksek başarı göstermiştir. Swin-B modeli %93.84 doğruluk, %88.15 kesinlik ve %85.52 duyarlılık ile en yüksek sonuçları vermiştir. Bu sonuçlar, ViT tabanlı modellerin omuz implantı üreticilerinin ve model bilgilerinin güvenilir bir şekilde tanımlanmasını ve özellikle uzmanlar için zaman verimliliği sağlayarak tedavi planlamasının iyileştirilmesine yardımcı olabileceğini göstermiştir.

Anahtar kelimeler: Total omuz artroplastisi, Omuz implantları, X-ray, Sınıflandırma, Görü dönüştürücü

1 Introduction

The shoulder is one of the most complex and commonly used joints in the body. Total shoulder arthroplasty (TSA) is a surgical procedure performed on patients who experience severe pain and restricted movement in the shoulder joint due to conditions such as osteoarthritis and rheumatoid arthritis [1-3]. During TSA surgery, the dysfunctional joint is removed, and a prosthetic implant is inserted in its place. X-ray images are used to determine the appropriate type of prosthesis. This surgical intervention aims to alleviate pain and improve joint function, allowing patients to regain mobility and enhance their overall quality of life.

Nowadays, there are several manufacturers of shoulder prosthetics, each offering different models tailored to individual patients. Despite the advancements in prosthetic materials and surgical techniques, prostheses may experience wear or loosen over time. Therefore, it is essential to periodically evaluate the condition of the prosthesis after the surgery and replace it if necessary [4]. Additionally, in cases of accidents or injuries, the existing prosthesis might get damaged and require replacement. To ensure a rapid and successful process of prosthesis improvement or replacement, it is crucial for both the patient and the doctor to be aware of the manufacturer and model information. However, incomplete medical records may lead to uncertainty in such

* Sorumlu yazar / Corresponding author, e-posta / e-mail: ebaykal@ktu.edu.tr (E. Baykal Kablan)
Geliş / Recieved: 05.12.2023 Kabul / Accepted: 12.03.2024 Yayımlanma / Published: 15.04.2024
doi: 10.28948/ngumuh.1400666

situations. Consequently, a thorough examination by medical experts is conducted by obtaining new X-ray images from the patient to verify the manufacturer and model information of the prosthesis. Due to the reliance on expert opinions, time consumption, and the taxing nature of this examination, there is a demand for automated systems to aid in this process.
The contributions of automated systems are as follows:

- Fast and efficient results: Automated systems quickly analyze X-ray images and provide instant results. This reduces waiting times for patients and accelerates the treatment process.
- High precision and accuracy: Machine learning algorithms work with large datasets to achieve high precision and accurate results. This ensures the correct identification of shoulder implants.
- Objective decision-making: Automated systems can make objective decisions without being influenced by human factors. This eliminates the risk of misinterpretation or personal biases.
- Continuous improvement: Automated systems can be updated with new datasets and continue learning. This allows the system to improve its performance over time.
- Reduces expert workload: Automated systems alleviate the burden on experts by eliminating the need to manually review X-ray images continuously. It enables experts to use their time more efficiently.

In summary, automated systems offer rapid, accurate, and efficient results, leading to better healthcare services and increased patient satisfaction. Moreover, they help healthcare professionals by reducing their workload and improving the overall efficiency of the treatment process.

Traditional machine learning methods, such as support vector machines and decision trees, often necessitate manual feature engineering and struggle with complex, high-dimensional data. The reduced utilization of classical machine learning approaches in biomedical image processing can be primarily attributed to their limitations in automatically extracting intricate features and patterns from extensive datasets. In recent years, however, the ability of deep learning to automatically learn from data, perform hierarchical feature extraction, and demonstrate versatility has positioned it as a powerful and preferred approach in biomedical image processing tasks. This has contributed to remarkable advancements in medical diagnosis and patient care. For a more detailed discussion and comparison, readers are referred to research papers [5, 6]. In the literature, several deep learning-based studies have been proposed for the detection and classification of shoulder implants from X-ray images [4, 7-12]. To get an idea about the effectiveness and applicability of deep learning in addressing the complexity of shoulder implant detection and classification from X-ray images, we can summarize these studies as follows.

Urban et al. fine-tuned six different pre-trained deep learning CNN architectures (VGG-16, VGG-19, ResNet-50, ResNet-152, NASNet, DenseNet-201) for the classification of shoulder implants on shoulder implant dataset [7]. The dataset consists of 597 X-ray images of shoulder implants, involving 16 different models from four different implant manufacturers. Through 10-fold cross-validation, they achieved the highest classification performance with an accuracy of 80.4% using the NASNet architecture. Yi et al. used a different dataset containing 482 X-ray images of shoulder implants belonging to five different implant models [8]. Instead of training a single classifier for all implant models, they trained five separate ResNet-152 architectures as binary classifiers for each implant model. Through 20-fold cross-validation, they achieved AUC-ROC values ranging from 0.86 for Solar to 1.0 for Zimmer, indicating varying levels of performance for different implant models. However, using multiple classifiers may add complexity to the overall system and require more computational resources. Yilmaz proposed a multi-channel CNN model that introduces a novel channel selection layer for choosing the most prominent feature filters [9]. By applying effective feature selection among channels for each image, the model significantly improved the accuracy rate. The proposed method achieved a higher performance with an accuracy rate of 97.2%. Sultan et al. proposed a CNN-based method for the classification of shoulder implants [10]. They used rotational data augmentation to increase the training dataset by 36 times. Modified ResNet and DenseNet network models were combined in depth to form the DRE-Net network ensemble architecture. Through 10-fold cross-validation, they achieved an accuracy of 85.92%, an F1 score of 84.69%, a precision of 85.33%, and a recall of 84.11%. Efeoglu et al. evaluated 12 different classifiers to classify shoulder implants from three different manufacturers [11]. By employing 10-fold cross-validation, they found that the K-NN algorithm achieved the highest accuracy of 74%, outperforming the other algorithms. Sivari et al. proposed ten distinct hybrid classifier models by combining deep learning and machine learning algorithms and then subjected the models to statistical testing [4]. According to the experimental findings, the DenseNet201 + Logistic Regression model achieved an impressive accuracy of 95.07%. In contrast to other methods, Karaci utilized the YOLOv3 object detection model to detect the head region of shoulder implants and then fed these regions as inputs to various CNN architectures [12]. As a result, focusing on the head region of the implant using the YOLOv3 object detection model improved the classification accuracy. By combining YOLOv3 with the DenseNet201 model, they achieved an accuracy of 84.76%. However, it's worth noting that using the YOLOv3 object detection model may introduce additional computational complexity and require more resources than other methods. These studies collectively highlight the effectiveness and versatility of deep learning methods, especially CNN architectures, for the detection and classification of shoulder implants from X-ray images. The advancements made in this area hold great promise for enhancing medical imaging and facilitating implant-related decision-making in clinical settings.

Convolutional neural networks (CNNs) have demonstrated significant success in various medical image analysis tasks, including the classification of shoulder implants, thanks to their ability to learn complex representations from data. However, the limited local receptive field in the convolution operation restricts the capture of long-range pixel dependencies [13]. To overcome this limitation, transformer architectures, inspired by their remarkable achievements in natural language processing [14], have been integrated into CNN architectures to encode long-range dependencies and learn more efficient feature representations [15]. The vision transformer (ViT) model proposed by Dosovitskiy et al. marked the first application of

transformer-based architectures to images, formulating the image classification task as predicting a sequence of image patches [16]. Following this pioneering work, ViT-based approaches [17,18] have been proposed and demonstrated state-of-the-art performance across various datasets [19-22].

In this study, the use of vision transformer models, which have recently become a popular area in deep learning architectures, is proposed for effective shoulder implant classification. Most of the studies in the literature are based on classical machine learning and CNN-based approaches. However, adapting transformer models for images has been shown to yield more effective results than CNNs. Due to their high accuracy and often superior performance compared to CNNs, many new studies have started incorporating vision transformer models. This research investigates the success of transformer models in shoulder implant classification and provides detailed comparisons with other CNN-based methods. To ensure objectivity and comparability with other approaches, a publicly available dataset is used. The experimental results show that vision transformers offer highly successful performance and generalization ability.

The contributions of this work can be summarized as follows:

- We present a pioneering approach for shoulder implant classification utilizing ViTs, showcasing the use of this emerging technology in the medical imaging domain.
- Through the fine-tuning of pre-trained ViT models, we achieve enhanced performance compared to traditional CNN-based models such as ResNet-50, Inceptionv3, and MobileNetv3. Despite challenges such as variable image resolution and class imbalance in a small dataset, the ViT-based approach outperforms individual CNN models.
- We achieve results that are competitive with the latest studies in the current literature. This underlines the significance of the proposed ViT-based approach in keeping pace with, and potentially surpassing, the state-of-the-art in shoulder implant classification from X-ray images.

The remaining sections of the paper are as follows: Section 2 provides a detailed explanation of the vision transformer structures, the publicly available shoulder implant dataset, and evaluation metrics. Section 3 presents the experimental results and discussion. Finally, Section 4 offers recommendations and conclusions.

2. Material and methods

All experiments were conducted on a computer equipped with an Intel(R) Core(TM) i9-11900K 3.50 GHz CPU and an NVIDIA GeForce RTX 3080 12GB GPU. Python programming language and PyTorch deep learning framework have been used for implementing the proposed vision transformer-based and CNN-based models in this study.

2.1 Dataset

In this study, a publicly available dataset consisting of 597 X-ray images of shoulder implants [7] was utilized. The dataset was selected due to its currency, having been introduced in 2020 and widely adopted in recent literature. The dataset encompasses X-ray images from four distinct shoulder implant manufacturers: 83 images from Cofield, 294 from Depuy, 71 from Tornier, and 149 from Zimmer. Sample X-ray images representing the four classes in the dataset can be found in Figure 1. The dataset poses several challenges, including variable and relatively low image resolution. Additionally, there are variations in image contrast and an imbalanced distribution of samples among the classes.

For a fair comparison, we used the same setting as in the state-of-the-art study by Sivari et al. [4], where the dataset was divided into two subsets: 90% for training and 10% for testing. Furthermore, for 10-fold cross-validation, the training set was further partitioned into ten groups, with nine sets used for training and one set for validation in each fold. During the dataset split, random images were selected from each class. Additionally, measures were taken during each cross-validation fold to prevent scenarios where all selected validation samples belonged to a single class or where certain classes had no representative samples.

Figure 1. Some sample images from different manufacturers in the dataset (a) Cofield (b) Depuy (c) Tornier (d) Zimmer

2.2 Vision transformer (ViT) model and its variants

In recent years, Transformers, introduced by Vaswani et al. [14], have emerged as highly successful deep learning architectures in Natural language processing (NLP) tasks. The superior performance of Transformers in NLP has motivated researchers to adapt them for computer vision tasks. The vision transformer (ViT), proposed in 2020 [16], has garnered significant attention for its promising results in computer vision tasks.

The ViT architecture adopts an attention mechanism that weights the importance of each part of the input image differently, in contrast to the traditional convolutional layers used in CNNs. The general architecture of the vision transformer model is shown in Figure 2.

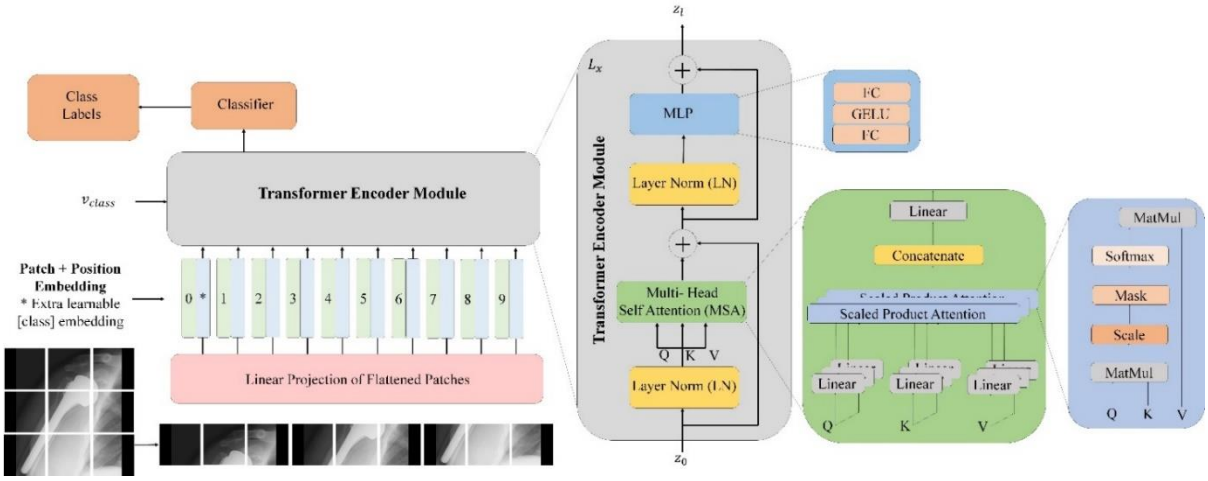


Figure 2. The general architecture of the Vision Transformer (ViT) model

In the ViT model, the input image, represented by $x \in R^{H\times W\times C}$, is transformed into a smaller array of image patches. Each reshaped image patch, denoted as p, is mapped to the z_0 vector using a learned embedding matrix E with Equation (1).

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{pos} \quad (1)$$

where $E \in R^{(P^2\times C)\times D}$, $E_{pos} \in R^{(N+1)\times D}$ and $x_p \in R^{N\times(P^2\times C)}$. N represents the number of extracted image patches and C indicates the number of channels. The x_{class} represents the signal added to the image patches and is used for classification purposes.

The transformer encoder module consists of identical layers denoted by L_x. Each layer is composed of multi-head self-attention (MSA), multi-layer perceptron (MLP), and layer normalization (LN) along with residual connections. To achieve the output map z_l of the same length as the input z_0, the following steps are applied with Equation (2-3).

$$z_l' = MSA\big(LN(z_{l-1})\big) + z_{l-1} \quad (2)$$

$$z_l = MLP\big(LN(z_l')\big) + z_l' \quad (3)$$

where $l = 1\ldots L$ represents each ViT encoder layer, and the output of each layer is fed as input to the next layer. The final MLP layer is placed on the output of the L identical encoder layers, and along with the softmax classifier, it is used for classifying the learned image representations.

The self-attention mechanism is employed to predict the relationship between elements in the input. In essence, an attention layer combines the total information from all input elements to update each component in the sequence. The interaction obtained from the inputs is transformed by three learnable weight matrices, query (W^Q), key (W^K), and value (W^V) matrices. For a given input array X, the input is first projected into these weight matrices as follows: $Q = XW^Q$, $K = XW^K$, and $V = XW^V$. The output of the self-attention layer is calculated with Equation (4).

$$Self - attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (4)$$

where d_k is a scaling factor and the result of self-attention is combined with the input using the softmax function.

In the realm of ViT models, two noteworthy advancements have emerged, namely, the data-efficient image transformer (DeiT) and the Swin Transformer. These innovations address crucial challenges in large-scale training datasets and processing efficiency for ViT models.

2.2.1 *Data-efficient image transformer*

DeiT [17] addresses the need for large-scale training datasets in ViT by incorporating an additional distillation token in its input, enabling it to achieve success with fewer data. This model interacts with class and patch tokens through self-attention layers and undergoes learning through backpropagation with the distillation token. The distillation token employs a teacher-student formulation within a knowledge distillation framework.

2.2.2 *Swin transformer*

Swin transformer [18] introduces a novel hierarchical design that divides the input image into non-overlapping "windows" to efficiently process large-scale images. Unlike traditional image transformers that operate on fixed-size patches, Swin Transformer adopts a "shifted window" mechanism, aiming to capture overlapping information between adjacent windows. Swin Transformer's key feature is its self-attention and shifted window mechanisms, which allow large images to be processed in parallel. The hierarchical self-attention mechanism provides better feature representation by capturing both local and global dependencies. The sliding window mechanism allows the network to be less sensitive to the order of the input sequence, thus contributing to the robustness and generalization of the model.

2.3 *Evaluation metrics*

We used accuracy (ACC), precision (PRE), recall (REC), and f-measure (FM) to evaluate the performance of the proposed shoulder implant classification models [23]. ACC is calculated as the ratio of correctly predicted instances to

the total number of instances in the test set and indicates the overall accuracy of the classifier. PRE is the ratio of correct positive predictions to the total number of positive predictions made by the classifier. REC measures the ability of the classifier to correctly identify positive examples. The FM is the harmonic mean of precision and recall and provides a more balanced assessment of the classifier's performance. These metrics are calculated as given in Equation (5).

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$
$$PRE = \frac{TP}{TP + FP}$$
$$REC = \frac{TP}{TP + FN} \quad (5)$$
$$FM = \frac{2 \times PRE \times REC}{PRE + REC}$$

We also used a confusion matrix to illustrate the relationship between actual class values and the class values predicted by the classifier. $k\,x\,k$ confusion matrix was created for the k-class classification problem. The $[i;\ j]$ cells of the confusion matrix ($i = 1, \dots, k;\ j = 1, \dots, k$) indicate the frequencies of observations related to the actual class C_i and the predicted class C_j. Table 1 shows an example confusion matrix representation for a 3-class problem.

Table 1. A confusion matrix representation for the multi-class problem

		Predicted Class		
		C_1	C_2	C_3
Actual Class	C_1	a	b	c
	C_2	d	e	f
	C_3	g	h	i

3. Results and discussion

In this section, the performance of the proposed ViT-based models designed for shoulder implant classification is meticulously compared with the latest CNN-based models. To ensure optimal accuracy during the training of all networks, a transfer learning strategy is employed, significantly reducing both time and data resource requirements. Throughout the learning process, the AdamW optimizer was used with a learning rate of 1e-5 and a weight-decay regularization value of 1e-8. The cross-entropy loss function is employed as the designated error function for optimal model learning. Uniform parameters are maintained across all models to ensure a fair and comparable assessment of both ViT-based and CNN-based models in their classification performance for shoulder implants. The batch size for each model is set at 4, and the learning process epoch number is set to 60.

The results presented in Table 2 provide a detailed overview of the classification outcomes achieved by each variant of the ViT-based models and three prominent CNN-based models, showcasing the superior performance of ViT models in the context of shoulder implant classification. Notably, the Swin Transformer model versions, ranging from Tiny (T), Small (S), Base (B), and Large (L), exhibit the highest classification accuracies among all models considered. Specifically, the Swin-B model stands out with an impressive accuracy of 93.84%. This accuracy is accompanied by a precision value of 88.15% and a recall value of 85.52%, underscoring the robustness and reliability of the Swin-B model in correctly classifying shoulder implants. Comparatively, the results obtained with all ViT models surpass those achieved by three well-established CNN-based models, namely ResNet50 [24], Inceptionv3 [25], and MobileNetv3 [26].

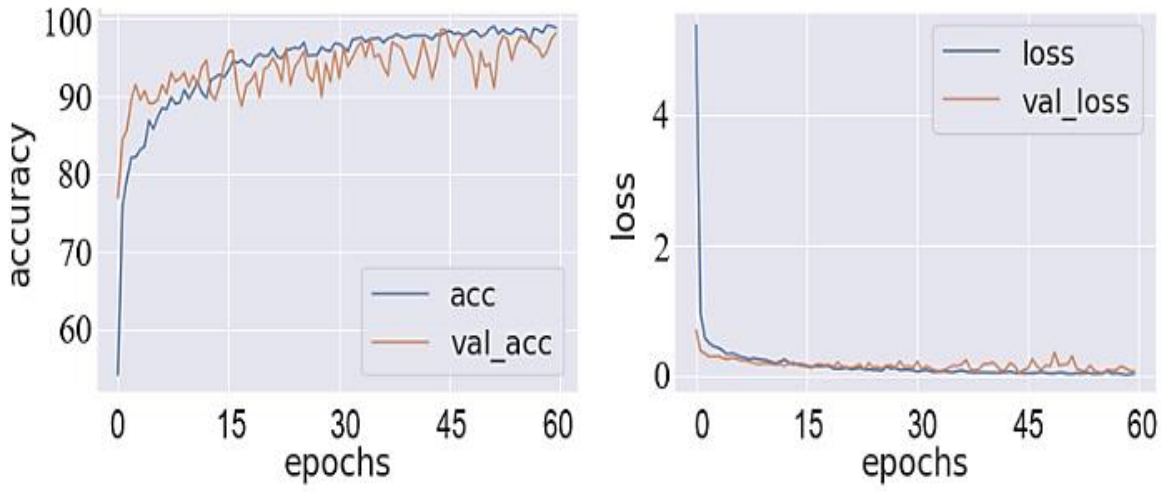
Table 2. Performance comparison of the proposed ViT-based models with CNN-based models

Model	ACC	PRE	REC	FM
ResNet-50	72.30	-	-	-
Inceptionv3	87.69	73.36	71.46	71.09
MobileNetv3	78.46	53.09	52.06	52.32
ViT-T	82.3	64.21	65.81	63.15
DeiT-T	86.15	70.13	63.64	64.31
Swin-T	90.76	80.33	77.84	77.76
ViT-S	87.69	74.67	68.59	69.49
DeiT-S	89.23	74.65	73.02	70.71
Swin-S	91.53	86.38	76.38	77.90
ViT-B	93.07	83.6	85.52	84.39
Deit-B	91.53	82.99	77.44	78.69
Swin-B	**93.84**	**88.15**	**85.52**	**85.93**
ViT-L	91.53	85.61	79.89	81.5
DeiT-L	93.84	86.60	86.95	85.36
Swin-L	93.84	87.36	83.69	83.94

While the highest accuracy attained by the Inceptionv3 model is 87.69%, the Swin-B model significantly outperforms it with a remarkable accuracy of 93.84%. This observation underscores the substantial advancements offered by vision-transformer models in the domain of shoulder implant classification. To validate the accuracy and loss variations during the training of the Swin-B model for 60 epochs, we present the Figure 3.

In Figure 3(a), both training and validation accuracies demonstrate swift convergence towards 100%, indicating robust learning. Correspondingly, in Figure 3(b), the loss values converge rapidly towards 0, underscoring the effective optimization of the model.

These results affirm the potential contributions of vision-transformer models, highlighting their capacity to significantly enhance the accuracy and efficiency of shoulder implant classification. Furthermore, the results reinforce the role of ViT models in advancing the applications of medical image analysis, suggesting a promising avenue for further improvements.



(a) (b)

Figure 3. The accuracy and loss variation obtained during the training of the Swin-B model for 60 epochs.

Table 3 also presents a comprehensive comparison of our proposed model for detecting and classifying shoulder implant manufacturers with previous studies conducted on the same dataset [4,7,9-12]. Noteworthy, prior studies have put forth diverse methodologies, encompassing individual machine learning, individual deep learning approaches, and hybrid strategies, showcasing the evolution of techniques in the field. A critical observation from this comprehensive review is that recent studies, leveraging both deep learning and machine learning approaches, consistently yield accurate and reliable classification results.

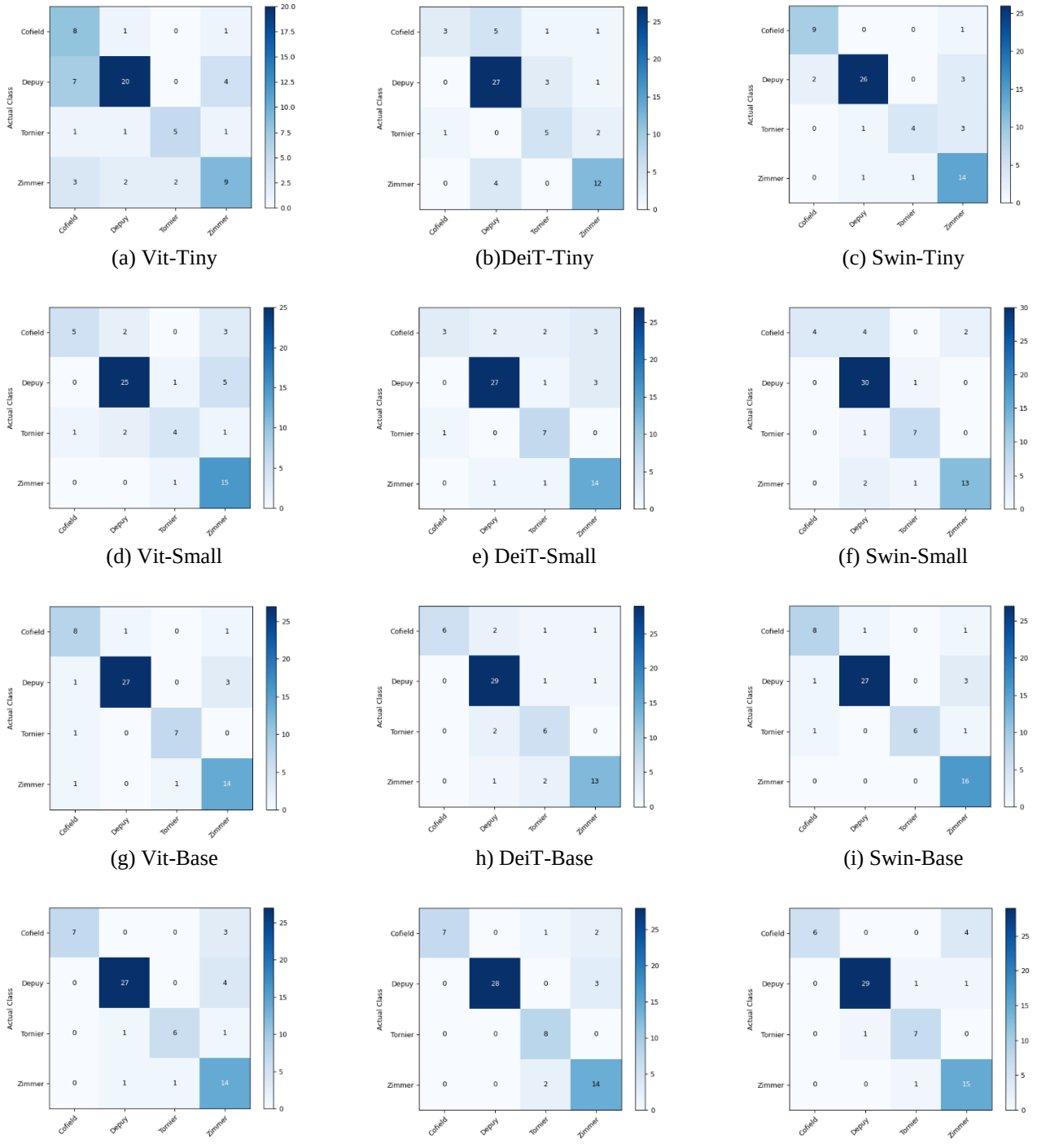
Urban et al. [7] proposed CNN-based NasNet architecture and achieved a classification accuracy of only 80%, which appears to be relatively low. However, the authors attributed this to the variable and relatively low image resolution, as well as the class imbalance issues present in the dataset. Subsequently, it was observed that the other proposed CNN-based models [4,9,12] were more resilient to these challenges. In particular, the highest cross-validation accuracy, reaching an impressive 97.2%, was reported in a recent study [9]. It's important to note that the study conducted by Sivari et al. [4] achieved the second-highest result at 95.07%, but it's essential to highlight that the dataset in the leading study [9] wasn't divided into test and validation sets. Instead, only 5-fold cross-validation results were provided, raising considerations about the robustness of the evaluation methodology. Addressing this concern, Sivari et al. [4] emphasized the importance of a separate validation dataset to fine-tune model hyperparameters and assess its generalization capabilities. This aligns with best practices in model evaluation.

In our proposed method, we adhere to these principles, employing a comprehensive test set for unbiased result evaluation, and as a result, we achieved a remarkable 93.84% accuracy. It is also noted that the ViT-based methods proposed in this study are more robust than CNN models against datasets containing the aforementioned challenges.

Table 3. Comparison with other studies on the same dataset

Study	Method	ACC
Urban et al. [7]	NasNet	80.0
Sultan et al. [10]	DRE-Net	85.92
Yilmaz [9]	Multi-channel CNN	97.2
Efeoglu et al. [11]	K-NN	74.0
Karaci [12]	YOLOv3 + DenseNet201	84.76
Sivari et al. [4]	DenseNet201 + LR	95.07
This study	Swin-B	**93.84**

In the study, confusion matrices were obtained to generate a summary by comparing the predicted class labels by the classification model with the actual class labels. Figure 4 displays the confusion matrices obtained sequentially for the tiny, small, base, and large models. It is observed that as the model complexity increases, the number of correctly identified examples (TP) also increases. For instance, in the ViT model, the number of TP was 42 in the tiny version, while it increased to 54 in the large version. On the other hand, for the Swin model, the TP count was 53 in the tiny version, and it rise to 57 in the large version. The Swin model was found to provide the highest results due to its multi-scale processing capabilities, which enabled more comprehensive feature extraction compared to other models. This is due to Swin model's unique self-attention mechanism and architectural design and its ability to effectively capture long-range dependencies in image data. Therefore, the proposed model is thought to help expert pathologists who will make the final decision on patients. Comparing the outcomes of our proposed approach with other machine learning and deep learning studies in the literature, we observe that our results are not only comparable but also highly successful. Moreover, they signify a promising foundation for further improvements in the field. The adaptability and potential for enhancements make our proposed methodology a valuable addition to the existing body of knowledge in the realm of shoulder implant classification.



(a) Vit-Tiny (b)DeiT-Tiny (c) Swin-Tiny

(d) Vit-Small e) DeiT-Small (f) Swin-Small

(g) Vit-Base h) DeiT-Base (i) Swin-Base

(j) Vit-Large (k) DeiT-Large (l) Swin-Large

Figure 4. Confusion matrices of the ViT models (ViT: first column, DeiT: second column, and Swin: third column) obtained sequentially for the Tiny, Small, Base, and Large versions on 4-class shoulder implant dataset

4. Conclusions

Vision transformer (ViT) models adapt the Transformer architecture, originally designed for large-scale text processing, to the field of image processing. These models tend to outperform conventional approaches in image processing tasks, particularly when faced with large datasets and complex structures. A notable advantage of ViT models is their ability to address object relationships in a broader context, making them ideal for tasks requiring long-range dependencies to be addressed. Additionally, when pre-trained on large-scale training datasets, ViT models exhibit significant improvements in feature transfer and generalization capabilities. In this paper, we present a novel

approach using ViT for shoulder implant classification from X-ray images. The proposed approach has shown promising results in automating this process by enabling efficient and accurate identification of shoulder implant manufacturers and models. By fine-tuning pre-trained ViT models, we achieved superior performance than CNN-based individual models on a small dataset despite challenges such as variable image resolution and class imbalance within a small dataset. This automated approach will reduce the reliance on expert opinion and our approach will save valuable time and effort, enabling faster and more accurate decision making. However, it is worth noting that while the results are encouraging, the precision and recall values are still areas for improvement. Therefore, future research directions could include studying larger datasets and proposing new Vision Transformer-based architectures, aiming to further improve the precision and recall of the system.

Conflict of interest

The authors declare that there is no conflict of interest.

Similarity rate (iThenticate): 20%

References

[1] R. H. Cofield, Total shoulder arthroplasty with the Neer prosthesis. JBJS, 66(6), 899-906, 1984. https://doi.org/10.2106/00004623-198466060-00010

[2] J. Sanchez-Sotelo, Total shoulder arthroplasty. The open orthopaedics journal, 5, 106, 2011. https://doi.org/10.2174/1874325001105010106

[3] C. Sukjamsri, The effect of implant misalignment on shoulder replacement outcomes (Doctoral dissertation, Imperial College London), 2015. https://doi.org/10.25560/28581

[4] E. Sivari, M. S. Güzel, E. Bostanci and A. Mishra, A Novel Hybrid Machine Learning Based System to Classify Shoulder Implant Manufacturers. In Healthcare (Vol. 10, No. 3, p. 580), MDPI, 2022. https://doi.org/10.3390/healthcare10030580

[5] D. P. Sahoo, M. Rout, P. K. Mallick and S. R. Samanta, Comparative Analysis of Medical Images using Transfer Learning Based Deep Learning Models. In 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) (pp. 1-8), IEEE, 2022. https://doi.org/10.1109/ASSIC55218.2022.10088373

[6] B. Sistaninejhad, H. Rasi and P. Nayeri, A Review Paper about Deep Learning for Medical Image Analysis. Computational and Mathematical Methods in Medicine, 2023. https://doi.org/10.1155/2023/7091301

[7] G. Urban, S. Porhemmat, M. Stark, B. Feeley, K. Okada and P. Baldi, Classifying shoulder implants in X-ray images using deep learning. Computational and structural biotechnology journal, 18, 967-972, 2020. https://doi.org/10.1016/j.csbj.2020.04.005

[8] P. H. Yi, T. K. Kim, J. Wei, X. Li, G. D. Hager, H. I. Sair and J. Fritz, Automated detection and classification of shoulder arthroplasty models using deep learning. Skeletal radiology, 49, 1623-1632, 2020. https://doi.org/10.1007/s00256-020-03463-3

[9] A. Yılmaz, Shoulder implant manufacturer detection by using deep learning: Proposed channel selection layer. Coatings, 11(3), 346, 2021. https://doi.org/10.3390/coatings11030346

[10] H. Sultan, M. Owais, C. Park, T. Mahmood, A. Haider and K.R. Park, Artificial intelligence-based recognition of different types of shoulder implants in X-ray scans based on dense residual ensemble-network for personalized medicine. J. Pers. Med, 11, 482, 2021. https://doi.org/10.3390/jpm11060482

[11] E. Efeoğlu and T. U. N. A. Gürkan, Radyografi Görüntüleri Ve Sınıflandırma Algoritmaları Kullanılarak Omuz Protezlerinin Üreticilerinin Belirlenmesi. Kırklareli Üniversitesi Mühendislik ve Fen Bilimleri Dergisi, 7(1), 57-73, 2021. https://doi.org/10.34186/klujes.906660

[12] A. Karaci, Detection and classification of shoulder implants from X-ray images: YOLO and pretrained convolution neural network based approach. J. Fac. Eng. Archit. Gazi Univ, 37, 283-294, 2022. https://doi.org/10.17341/gazimmfd.888202

[13] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan and H. Fu, Transformers in medical imaging: A survey. Medical Image Analysis, 102802, 2023. https://doi.org/10.1016/j.media.2023.102802

[14] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman and J. Shlens, Scaling local self-attention for parameter efficient visual backbones. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12894-12904), 2021. https://doi.org/10.48550/arXiv.2103.12731

[15] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. https://doi.org/10.48550/arXiv.2010.11929

[17] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jégou, Training data-efficient image transformers & distillation through attention. In International conference on machine learning (pp. 10347-10357), PMLR, 2021.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang and B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022), 2021.

[19] S. Tummala, J. Kim and S. Kadry, BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of swin

transformers. *Mathematics*, *10*(21), 4109, 2022. https://doi.org/10.3390/math10214109

[20] S. Ayas, Multiclass skin lesion classification in dermoscopic images using swin transformer model. *Neural Computing and Applications*, *35*(9), 6713-6722, 2023. https://doi.org/10.1007/s00521-022-08053-z

[21] A. Alotaibi, T. Alafifi, F. Alkhilaiwi, Y. Alatawi, H. Althobaiti, A. Alrefaei and T. Nguyen, ViT-DeiT: An Ensemble Model for Breast Cancer Histopathological Images Classification. In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)* (pp. 1-6). IEEE, 2023. https://doi.org/10.1109/ICAISC56366.2023.10085467

[22] S. Regmi, A. Subedi, U. Bagci and D. Jha, Vision Transformer for Efficient Chest X-ray and Gastrointestinal Image Classification. arXiv preprint arXiv:2304.11529, 2023.

[23] D. M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061, 2020. https://doi.org/10.48550/arXiv.2010.16061

[24] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778), 2016. https://doi.org/10.48550/arXiv.1512.03385

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826), 2016. https://doi.org/10.48550/arXiv.1512.00567

[26] A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan and H. Adam, Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324), 2019. https://doi.org/10.48550/arXiv.1905.02244