Automated Identification of Hip Arthroplasty Implants Using Artificial Intelligence

Zibo Gong

China Medical University

Yonghui Fu (✉ yonghuifu@hotmail.com)

China Medical University

Ming He

China Medical University

Xinzhe Fu

Massachusetts Institute of Technology

Research Article

Keywords: Machine learning, Deep learning, Artificial intelligence, hip arthroplasty, Neural networks

Posted Date: February 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1275278/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Objectives: The purpose of this study was to develop and evaluate the performance of deep learning methods based on convolutional neural networks (CNN) to detect and identify specific hip arthroplasty models.

Methods: In this study, we propose a novel deep learning-based approach to identify hip arthroplasty implants' design using anterior-posterior (AP) images of both the stem and the cup. We harness the pre-trained ResNet50 CNN model and employ transfer learning methods to adapt the model for implants identification task using a total of 714 radiographs of 4 different hip arthroplasty implant designs. Performance was compared with operative note and crosschecked with implant sheets. We also evaluate the difference of performance of models trained with the images of the stem, the cup or both.

Result: The training and validation data sets were comprised of 357 stem images and 357 cup radiographs across 313 patients and included 4 hip arthroplasty implants from 4 leading implant manufactures. After 1000 training epochs the model classified 4 implant models with very high accuracy. Our results showed that jointly using stem images and cup images did not improve the classification accuracy of the CNN model.

Conclusion: CNN can accurately distinguish between specific hip arthroplasty designs. This technology could offer a useful adjunct to the surgeon in preoperative identification of the prior implant. However, using both stem images and cup images to train the CNN is not more effective than using images from only one perspective.

Key Points

- Classification of specific hip arthroplasty designs on radiographs is excellent using CNN.
- Automatic classification of the hip implant models varies in accuracy and is inferior for cup images.
- Using both stem images and cup images to train the CNN is not more effective than using images from only one perspective.

Introduction

Total hip arthroplasty (THA) has been called the “Operation of the Century” because of its durability, reliability, and reproducibility in relieving pain and improving function in patients with coxarthrosis ^[1]. However, despite the clinical success of THAs, the number of revision THAs performed has increased with time. The etiology of the increase in the number of revision procedures is multifactorial. The increasing absolute number of primary arthroplasties, expansion of the indications to include younger and more active individuals, all likely contribute to overall revision rates. Projections based on population studies pointed to continued increases in the prevalence of revision procedures ^[2]. One of the critical steps in

preoperative planning for revision total hip arthroplasty is the identification of the failed implant. A recent survey of arthroplasty surgeons showed that surgeons spent approximately 20 minutes for each revision case to identify the implant preoperatively. 10% of implants could not be identified preoperatively with 2% not being identified intraoperatively. Failure to identify implant preoperatively resulted in additional requested implants, added surgical time, increased perioperative morbidity, increased healthcare cost [3].

Machine learning is an application of artificial intelligence (AI). It enables computers to find hidden insights without being explicitly programmed using algorithms that iteratively learn from the data. In the field of general imaging and computer vision, deep learning is the leading machine learning tool. Deep learning refers to techniques that build on developments in artificial neural networks in which multiple network layers are added to increase the levels of abstraction and performance [4]. Recently, deep learning, and in particular convolution neural networks (CNN), has shown groundbreaking results in a variety of general image recognition and computer-aided diagnosis tasks [5]. Within orthopedics, these powerful models already can reach human level performance in diagnosis of fractures [6] and staging knee osteoarthritis (OA) severity [7], which clearly indicates the possibility for using them in clinical practice in the near future.

We hypothesized that deep learning-based AI algorithms could facilitate the automated identification of hip arthroplasty implants, thereby aiding in preoperative planning and resulting in saved time and healthcare resources spent on this labor-intense task.

Methods

Study Design and Radiograph Acquisition

After institutional review board approval, we retrospectively collected all radiographs taken between June 1, 2011 and Dec 1, 2020 at one university hospital. We confirm that all methods were performed in accordance with the relevant guidelines and regulations. Images were collected from surgeries performed by 3 fellowship-trained arthroplasty surgeons to ensure a variety of implant manufacturers and implant designs. At the time of collection, images had all identifying information removed and were thus de-identified. Implant type was identified through the primary surgery operative note and crosschecked with implant sheets. Implant designs were only included in our analysis if more than 30 images per model were identified [8].

From the medical records of 313 patients, a total of 357 images were included in this analysis.

Although Zimmer and Biomet merged (Zimmer Biomet), these were treated as two distinct manufacturers. The following 4 designs from the four industry leading manufacturers were included: Biomet Echo Bi-Metric (Zimmer Biomet), Biomet Universal RingLoc (Zimmer Biomet) Depuy Corail (Depuy Synthes), Depuy Pinnacle (Depuy Synthes), LINK Lubinus SP II, LINK Vario cup, and Zimmer Versys FMT and Trilogy (Zimmer Biomet). Implant designs that did not meet the 30-implant threshold were not

included. Figure 1 demonstrated example of Cup and Stem anterior-posterior (AP) radiographs of each included implant design. The four types of implants are denoted as type A, type B, type C, and type D respectively in this paper.

Overview of Framework

We used convolutional neural network-based algorithms for classification of hip implants. As a CNN model typically involved millions of parameters while our data set only contained less than one thousand images, it was infeasible to train a CNN model from scratch using our data. Therefore, we adopted the transfer learning framework and the ResNet architecture^[9]. It started with a ResNet50 network pre-trained on ImageNet^[10], and did gradient descent (backward-propagation) using our training data only on the last two layers of the network. We manually cut the image of every anteroposterior (AP) view of the hip into two parts- the stem and the cup. We trained one network using the AP images of the stem, and another network using the AP images of the cup. Finally, we used a two-layer feed-forward network to combine the output of stem-trained network and cup-trained network to evaluate whether jointly considering stem images and cup images could improve the classification accuracy. Figure 2 demonstrates the overview of the framework of our deep learning-based method.

Dataset:

Our dataset contained 714 images from 4 different kind of implants.

Image Preprocessing:

We followed standard procedures to pre-process our training data so that it could work with a network trained on ImageNet. We rescaled each image to a size of 224*224, and normalized it according to ImageNet standards. We also performed data augmentation, i.e., random rotation, horizontal flips, etc., to increase the amount of training data and make our algorithm robust to the orientation of the images.

Dataset Division

We first divided the set of patients into three groups of sizes ~60% (group 1), ~30% (group 2), and ~10% (group 3). This split technique was used on a per-design basis to ensure the ratio of each implant remained constant. Next, we used the cup and stem images of patients in group 1 for training, those of patients in group 2 for validation, and those of patients in group 3 for testing. The validation set was used to compute cross-validation loss for hyper-parameter tuning and early stopping determination.

Model Training

We adopted the adaptive gradient method ADAM^[11] to train our models. Based on the cross-validation loss, we chose the hyper-parameters for ADAM as (learning rate = 0.001, weight_decay = 0). The maximum number of epochs was 1000 and the batch size was 16. The early stopping threshold was set to 8. During the training process of each network, the early stopping threshold was hit after around 50

epochs. As we mentioned above, we trained one network with the stem images and another with the cup images. We further combined the two models to see if jointly utilizing stem and cup images could yield superior testing results. The combination was done via the following logistic-regression based method. We collected the outputs of the stem network and the cup network (both are of the form of a 4-dimensional vector, with each element corresponding to the classification weight the network gives to the category of implants), and then fed them as the input to a two-layer feed-forward neural network, and trained the network with the data from the validation set. Note that the above construction relied on our dataset division procedure, where the training set, validation set, and testing set, each contained the stem and cup images of the same set of patients. We referred to the resulting network constructed from the outputs of stem networks and cup network as the “joint network”.

Model Testing

We tested our models (stem, cup, Joint) using the testing set. The prediction result on each testing image was a 4-dimensional vector, with each coordinate representing the classification confidence of the corresponding category of implants.

Statistical Analysis

Since we were studying a multi-class classification problem, we would directly present the confusion matrices of our methods on the testing data, and compute the operation characteristics generalized for multi-class classification.

Results

Training Progress

The following figure (Figure 3) showed the training progress of our method (we used the cup-network as an example). As the training proceeded, the network adjusted its parameters and learned the correct classification function as demonstrated by decrease in both training and validation loss.

Testing Performance

In this section, we evaluated our neural network models on the testing data sets. In the testing sets, we had in total 71 images. The classification results of the three neural network models (stem, cup, Joint) were presented in the form of tables. The column labels of each table indicated the ground-truth implant type, while row labels indicated the classification results. For example, the first row in the table of the stem-network indicated that the stem network classified 32 images with type-A plant as type A, and 3 image with type-A plant as type B.

Stem-Network

The classification results (Confusion Matrix) of the Stem-Network were shown in Table 1.

The stem-network achieved an overall classification accuracy of 91.5%. Although the overall accuracy was high, the accuracy on some implants with scarce training data was less satisfying. For example, its classification accuracy on implant type-B was about 69.6%.

Stem-Network

Table 1
Classification Results (Confusion Matrix) of the Stem-Network

Test Results Ground Truths	A	B	C	D
A	32	3	0	0
B	4	17	0	0
C	0	0	9	0
D	0	2	0	4

Cup-Network

The classification results (Confusion Matrix) of the Cup-Network were shown in Table 2.

The cup-network achieved a satisfying accuracy of 83.7%. However, its testing performance was dominated by that of the stem-network.

Cup-Network

Table 2
Classification Results (Confusion Matrix) of the Cup-Network

Test Results Ground Truths	A	B	C	D
A	28	6	1	0
B	6	15	0	0
C	0	0	9	0
D	0	3	0	3

Joint Network

The classification results (Confusion Matrix) of the Joint Network were shown in Table 3.

The join-network that essentially combined the stem-network and the cup-network had an overall accuracy of 88.8%. However, from the experiments, we did not see that combining stem images with cup images had the potential to achieve the best of both worlds, and improved the classification performance of only using one category of image data. This can be mainly attributed to that the cup-network's performance is dominated by the stem-network, which implies that the joint network cannot have better performance than the stem-network.

Joint Network

Table 3
Classification Results (Confusion Matrix) of the Joint Network

Test Results Ground Truths	A	B	C	D
A	30	5	0	0
B	5	16	0	0
C	0	0	9	0
D	0	2	0	4

Operation Characteristics

For a multi-class classification method, let M be its confusion matrix on the testing data. We used M_{ij} to denote the entry on the i th row and j th column of M . Then, for each row (class), the recall and precision of the algorithm were defined as:

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}}$$

1

$$Recall_i = \frac{M_{ii}}{\sum_j M_{ij}}$$

2

The precision (recall) of the algorithm was the average of its precision (recall) for each class.

Following the above definition, we had the following operating characteristics of the Stem-network, Cup-network and Joint-network (Table 4).

Table 4
Operation Characteristics of Our
Methods

	Precision	Recall
Stem-network	0.915	0.847
Cup-network	0.837	0.754
Joint-network	0.888	0.821

Discussion

Preoperative identification of arthroplasty implants prior to revision surgery is a difficult task and an essential step in preventing increases in perioperative morbidity and increased healthcare costs. The key finding of this study was that an artificial intelligence-based, deep learning CNN system could be trained to provide an automated identification of THA implant from radiographic images with near perfect accuracy.

To our knowledge, there are only 4 previous studies that have used deep learning for identifying hip arthroplasty implants from plain radiographs. One study by Kang et al. ^[12] used a small sample sized dataset (170 AP hip radiographs) to distinguish between 29 hip implants. In particular, in their dataset, 24 hip implants had fewer than 10 samples and no implant has more than 20 sample.

Borjali et al. ^[13] similarly applied a CNN to identify the design of nine different THA femoral implants from AP hip radiograph. Although limited by the volume of radiographs trained, this study showed that a deep-learning algorithm was capable of recognizing the unique femoral stem design features of a hip arthroplasty implant.

Karnuta et al. ^[8] reported in the other study that their deep learning model using 1766 AP hip radiographs could discriminate successfully 18 hip implant models.

The study by Murphy et al. ^[14] showed the impact artificial neural network architecture might play in test performance, ranging anywhere from 46.60–91.67% accuracy in test data depending on the architectures chosen.

However, all these four studies were only capable of identifying the femoral component of the implant, without paying attention to the identification ability of the deep-learning system to include the acetabular component of the implant.

In our study, the deep learning model developed from the AP view of hip radiographs represents the first advanced artificial intelligence model in identifying both the femoral component and the acetabular component of the implant from radiographs. Although the AP view of hip radiographs of different cup implants have similar appearances and are almost indistinguishable for human, the classification accuracy of our deep learning method is quite high (up to 83.7%). Furthermore, the accuracy of our method is higher than some of the previously proposed methods for stem implant identification.

This study has several limitations.

First, the data come from only one institution, the number of images for each brand implant was highly varied because the contractual relationship between the single institution and different manufactures has been changing over the past 10 years. This could result in an imbalance in trained implant data.

Second, our algorithm was trained with only 4 hip arthroplasty implants, therefore, the data was limited to these models alone and should not be generalized to identify other models.

Third, clearly the CNN remains a less commonly studied model as it is technologically demanding and requires access to numerous data, a team involving a surgeon, radiologist, and computer scientist to apply the deep learning models prior to implementing this type of solution in clinical practice.

Conclusion

In conclusion, the technology behind machine learning is exponentially advancing and the enthusiasm for AI in healthcare is growing. With this study, we have demonstrated that a deep learning algorithm can identify the design of 4 different hip arthroplasty implants from AP hip radiograph. These findings suggest that this technology has the potential to classify hip implants prior to revision surgery, thus saving significant time, and reducing perioperative morbidity and healthcare cost. It is hoped that it can be used to collect large-scale implant information and may be applied to mobile phone applications in the future.

Abbreviations

AI Artificial intelligence

CNN Convolutional neural networks

THA Total hip arthroplasty

OA Osteoarthritis

AP Anterior-posterior

Declarations

Ethical review committee statement

The institutional review board approved the study with a waiver of informed consent because all images were anonymized before the time of the study.

Authors' contributions

Zibo Gong¹ and Yonghui Fu² contributed equally to this work

Level of Evidence:

Level III, Study of a non-consecutive patients without a universally applied “gold” standard.

Data Availability Statement

The datasets generated during and analyzed during the current study are not publicly available due to the requirement of The Ethics Committee of Shengjing Hospital but are available from the corresponding author on reasonable request.

Credit authorship contribution statement

Yonghui Fu: Conceptualization, Methodology, Project administration, Writing - review & editing, Supervision.

Zibo Gong: Resources, Investigation, Data curation, Writing - original draft.

Ming He: Data curation, Resources, Writing - review & editing

Xinzhe Fu: Software, Formal analysis, Writing - review & editing.

Additional Information

The author(s) declare no competing interests.

References

1. Learmonth,I.D., Young,C., Rorabeck,C.(2007). The operation of the century: total hip replacement. *Lancet*.**370(9597)**:1508–19.
2. Bozic,K.J. et al (2015) Comparative epidemiology of revision arthroplasty: Failed THA poses greater clinical and economic burdens than failed TKA. *Clin. Orthop. Relat. Res.* **473**:2131–2138
3. Wilson,N.A., Jehn,M., York,S., Davis,C.M. (2014) Revision total Hip and knee arthroplasty implant identification: Implications for use of unique device identification 2012 AAHKS member survey results. *J. Arthroplasty* **29**:251–255
4. LeCun,Y., Bengio,Y., Hinton,G. (2015) Deep learning. *Nature* **521(7553)**:436–444

5. Esteva, A. et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**:115–118
6. Kim,D.H., MacKinnon,T. (2018) Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol.* **73(5)**:439–445
7. Abedin,J. et al (2019) Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images. *Sci Rep.* **9(1)**:5761
8. Karnuta,J.M. et al(2021). Artificial Intelligence to Identify Arthroplasty Implants from Radiographs of the Hip. *J Arthroplasty*.**36(7S)**:S290-S294.e1.
9. He,K.M., Zhang,X.Y., Ren,S.Q., Jian,S. (2016) Identity mappings in deep residual networks. In *European Conference On Computer Vision*, pp. 630-645. Springer
10. Krizhevsky, Alex, Ilya,S., Geoffrey,E., Hinton, (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**:1097–1105
11. Kingma, Diederik,P., Jimmy,B. (2015) Adam: A method for stochastic optimization, *International Conference on Learning and Representation*.
12. Kang,Y.J., Yoo,J.I., Cha,Y.H., Park,C.H., Kim,J.T.(2019). Machine learning-based identification of hip arthroplasty designs. *J Orthop Translat.***21**:13–17
13. Borjali,A. et al(2021) Comparing the performance of a deep convolutional neural network with orthopedic surgeons on the identification of total hip prosthesis design from plain radiographs. *Med Phys.* **48(5)**:2327–2336.
14. Murphy,M. et al(2021). Artificial intelligence accurately identifies total hip arthroplasty implants: a tool for revision surgery. *Hip Int.* 8:1120700020987526. doi:10.1177/1120700020987526

Figures

Figure 1

demonstrated example of Cup and Stem radiographs of each included implant design.

Figure 2

Overview of the framework of our deep learning-based method

Figure 3

Training and Validation Losses Curve of the cup-network