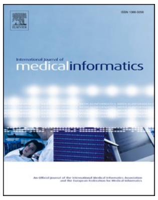
International Journal of Medical Informatics 176 (2023) 105095

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



# Combining deep learning and machine learning for the automatic identification of hip prosthesis failure: Development, validation and explainability analysis

Federico Muscato [a], Anna Corti [b], Francesco Manlio Gambaro [c], Katia Chiappetta [c], Mattia Loppini [c,d,e], Valentina D.A. Corino [a,f,*]

[a] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Golgi 39, 20131 Milan, MI, Italy
[b] Laboratory of Biological Structure Mechanics (LaBS), Department of Chemistry, Materials and Chemical Engineering "Giulio Natta", Politecnico di Milano, Milan, Italy
[c] Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, MI, Italy
[d] IRCCS Humanitas Research Hospital, Via Alessandro Manzoni 56, 20089 Rozzano, MI, Italy
[e] Fondazione Livio Sciutto Ricerca Biomedica in Ortopedia—ONLUS, Via A. Magliotto 2, 17100 Savona, SV, Italy
[f] Cardio Tech-Lab, Centro Cardiologico Monzino IRCCS, Via Carlo Parea 4, 20138 Milan, Italy

ARTICLE INFO

*Keywords:*
Artificial Intelligence
Hip Replacement
Image classification
Feature extraction
PCA algorithm

ABSTRACT

*Aim:* Revision hip arthroplasty has a less favorable outcome than primary total hip arthroplasty and an understanding of the timing of total hip arthroplasty failure may be helpful. The aim of this study is to develop a combined deep learning (DL) and machine learning (ML) approach to automatically detect hip prosthetic failure from conventional plain radiographs.

*Methods:* Two cohorts of patients (of 280 and 352 patients) were included in the study, for model development and validation, respectively. The analysis was based on one antero-posterior and one lateral radiographic view obtained from each patient during routine post-surgery follow-up. After pre-processing, three images were obtained: the original image, the acetabulum image and the stem image. These images were analyzed through convolutional neural networks aiming to predict prosthesis failure. Deep features of the three images were extracted for each model and two feature-based pipelines were developed: one utilizing only the features of the original image (original image pipeline) and the other concatenating the features of the three images (3-image pipeline). The obtained features were either used directly or reduced through principal component analysis. Both support vector machine (SVM) and random forest (RF) classifiers were considered for each pipeline.

*Results:* The SVM applied to the 3-image pipeline provided the best performance, with an accuracy of 0.958 ± 0.006 in the internal validation and an F1-score of 0.874 in the external validation set. The explainability analysis, besides identifying the features of the complete original images as the major contributor, highlighted the role of the acetabulum and stem images on the prediction.

*Conclusions:* This study demonstrated the potentialities of the developed DL-ML procedure based on plain radiographs in the detection of the failure of the hip prosthesis.

## 1. Introduction

Total hip arthroplasty (THA) is a highly effective treatment for several hip diseases in both young and elderly people and it is extremely widespread worldwide. In Italy, the number of primary hip replacements increased from 66,560 in 2001 to 97,263 in 2016 with an average increase of 3.1% per year [1]. A further increase in primary prosthetic implants placement is foreseen within 2030, due to progressive population aging and the growing number of procedures in younger patients. For the same reasons, a significant increase of implant revisions has to be expected [2,3], thus, a properly conducted radiographic follow-up aims to ensure an earlier identification of potential complications and failure, associated with more likely manageable complications, conservative revisions and favorable functional outcomes. However, the early detection of THA failure still remains a challenge, and the final diagnosis is often confirmed at the time of

[*] Corresponding author at: Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Golgi 39, 20131 Milan, MI, Italy.
*E-mail address:* valentina.corino@polimi.it (V.D.A. Corino).

https://doi.org/10.1016/j.ijmedinf.2023.105095
Received 13 March 2023; Received in revised form 26 April 2023; Accepted 12 May 2023
Available online 18 May 2023

revision surgery.

Artificial intelligence, particularly Deep Learning (DL) and Machine Learning (ML) can be used for the automatic evaluation of X-ray imaging for monitoring patients with hip arthroplasties, thus enhancing the diagnostic accuracy for THA failure. DL methods have already been applied to X-ray with a high degree of success in different orthopedic applications, such as identification of fractures [4] and classification of knee osteoarthritis [5]. A recent study has demonstrated the feasibility to automatically detect hip prosthetic failure through a DL approach, based on Convolutional Neural Networks (CNN) [6].

The aim of the present study is to develop a robust combined DL-ML pipeline for the automatic detection of hip prosthetic failure. In particular, in addition to the original radiograph (i.e., the one comprising the entire prosthesis), two additional images, with the acetabulum only and with the stem only, were considered to enhance the classification performance. To test the aforementioned hypothesis, both the baseline end-to-end DL approach and the feature-based DL-ML approach (based on the original image or on 3 images) will be applied on the same dataset. Finally an explainability analysis will be performed to identify the main contributors of the model.

## 2. Materials and methods

### 2.1. Study population

The study included two cohorts of patients who had undergone hip replacement surgery.

The first cohort, consisting of 280 patients, was used to develop the model. This group was divided equally into two sub-groups: those who had experienced failure and underwent revision surgery (140 patients), and those who did not require revision surgery and served as control group (140 patients), clinical characteristics are shown in Table 1.

In this study, failure included as loosening, bearing surface wear and osteolysis, malpositioning and dislocation. In particular, the THA failure in the two cohorts involved only the acetabular component in 48% of cases, the femoral component in 9% of cases and both the components in 43% of cases. Almost all failures were due to aseptic loosening (95.4%), while a minority of patient prosthesis failed for polyethylene wear (3.6%), repetitive luxation (0.5%) and periprosthetic infection (0.5%).

To be eligible for inclusion in the study, patients had to have one antero-posterior (AP) and one lateral (LAT) radiographic view of the implant before the revision surgery for the failed group or during post-operative follow-up for the non-failed group. Thus, the dataset used to develop the model included a total of 560 images.

**Table 1**
Clinical characteristics of the study populations.

| | Population | Failed | Non-failed |
|---|---|---|---|
| **Cohort 1** | | | |
| **Patients** | 280 | 140 | 140 |
| **Age (years)** | 66 ± 11 | 67 ± 11 | 65 ± 12 |
| **Sex (males)** | 104 (37%) | 48 (34%) | 56 (40%) |
| **Time from implant (months)** | – | 16 (0.1–100) | 42 (0.03–120) |
| **No Comorbidities** | 23 (8%) | 13 (9%) | 10 (7%) |
| **Dyslipidemia** | 16 (6%) | 10 (7%) | 6 (4%) |
| **Hypertension** | 47 (17%) | 27 (19%) | 20 (14%) |
| **Diabetes** | 9 (3%) | 8 (6%) | 1 (0.7%) |
| **Osteoporosis** | 7 (3%) | 5 (4%) | 2 (1%) |
| **Cardiomyopathy** | 16 (6%) | 10 (7%) | 6 (4%) |
| **Cohort 2** | | | |
| **Patients** | 352 | 275 | 77 |
| **Age (years)** | 67 ± 13 | 68 ± 11 | 61 ± 18 |
| **Sex (males)** | 125 (36%) | 93 (34%) | 32 (42%) |
| **Time from implant (months)** | – | 17 (0.2 – 102) | 41 (0.10–129) |
| **No Comorbidities** | 52 (15%) | 42 (15%) | 10 (13%) |
| **Dyslipidemia** | 19 (5%) | 17 (6%) | 2 (3%) |
| **Hypertension** | 65 (18%) | 56 (20%) | 9 (12%) |
| **Diabetes** | 9 (3%) | 6 (2%) | 3 (4%) |
| **Osteoporosis** | 7 (3%) | 5 (4%) | 2 (1%) |
| **Cardiomyopathy** | 16 (5%) | 11 (4%) | 5 (6%) |

The second cohort of patients, which served as the external validation group, consisted of 275 patients who had experienced failure and 77 patients who had not experienced failure. This dataset included either AP or LAT images, with a total of 771 images available for analysis, comprising of 589 failed and 182 non-failed images.

Patients included in this study were retrospectively collected from the digital medical records at Humanitas Research Hospital, Italy, between 2009 and 2019. All the radiographic images were provided by the Clinical and Radiographic Arthroplasty Register of Livio Sciutto Foundation Biomedical Research in Orthopaedics – ONLUS. The study was approved by the Institutional Ethical Committee of Humanitas Research Hospital (prot. 408/19, approved on June25, 2019), Italy, and all patients gave their written informed consent.

### 2.2. Pre-processing

In this work pre-processing was used for segmentation, to improve the resolution capability and ensure similar enough images. All images (DICOM format) were pre-processed to reduce noise and to have the same pixel range (between 0 and 1) and the same size. AP images were split vertically into two parts, each of them including only one limb (Fig. 1(a)).

The following pre-processing steps were performed: i) the mist like effect was reduced and brightness was enhanced through a gamma power transformation (Fig. 1(b)) [7]; ii) a sigmoidal function was used to improve contrast (Fig. 1(c)), thus emphasizing the prosthesis with respect to bone structures; iii) contrast-limited adaptive histogram equalization (CLAHE) method enabled contrast enhancement (Fig. 1(d)) [8]; iv) a low pass-filtering was performed through a 2-D Gaussian smoothing kernel, so that all the frequencies above the cutoff frequency, representing noise, were eliminated.

Finally, the image was resized to a standard input dimension (224x224) and standardized by z-score. After the above mentioned pre-processing of the X-rays images, a segmentation algorithm (based on Canny algorithm [9]) was used to identify the presence of the implant (some AP images had a unilateral implant). The segmentation of the prosthesis was also used to extract the image with only the acetabulum region (upper third of the image) and the one with the stem region.

All the pre-processing steps were performed using MATLAB R2018b (MathWorks, USA).

### 2.3. Baseline pipeline

The Densenet169 [10] pretrained for Imagenet [11] was used as baseline end-to-end DL model (*baseline model* in the following). The Fully Connected layers of the original structure were replaced with a Global Average Pooling, a 128-Dense, a Dropout and a 2-Dense layers and transfer learning with a fine-tuning approach was performed.

Data augmentation was applied enabling the resulting model to be more robust to non-relevant sources of variability, including suboptimal positioning of patients within the radiograph and suboptimal exposure settings. The deep model was compiled choosing the "binary cross entropy" as loss function and the training automatically stopped after 25 epochs in which the validation loss did not decrease. Moreover, if the validation loss was not decreasing in 10 epochs, the learning rate was reduced by a factor of 0.1 until a minimum value of $1x10^{-9}$ was reached. A Dropout rate of 0.5 was set to avoid overfitting; the batch size was 32 and the maximum number of epochs was 150. To assess the network performance, parallel evaluation of both accuracy and loss along epochs was considered.

### 2.4. Feature-based DL-ML pipeline development

Fig. 2 provides a schematic of the developed DL and DL-ML pipelines,

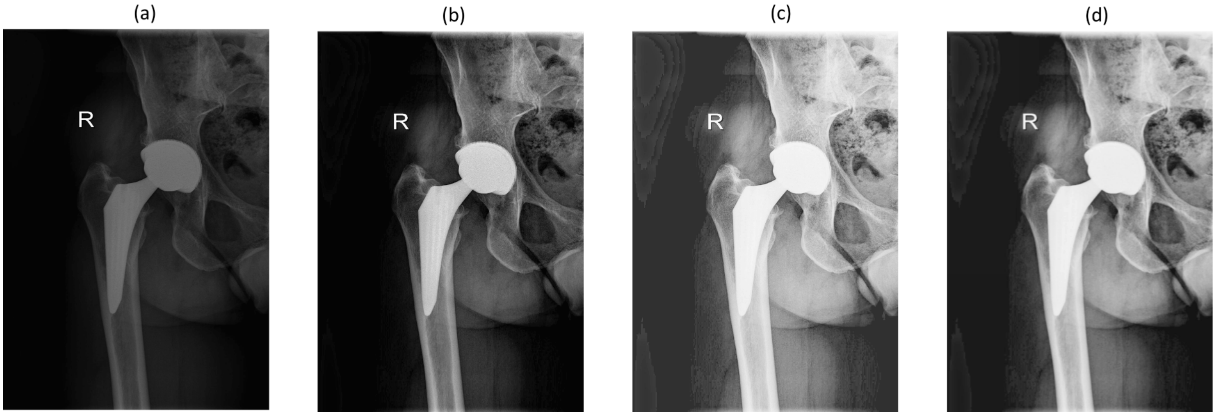


**Fig. 1.** Pre-processing steps. (a): Initial image; (b) mist effect reduction; (c) contrast enhancement; (d) contrast-limited adaptive histogram equalization and low-pass filtering.

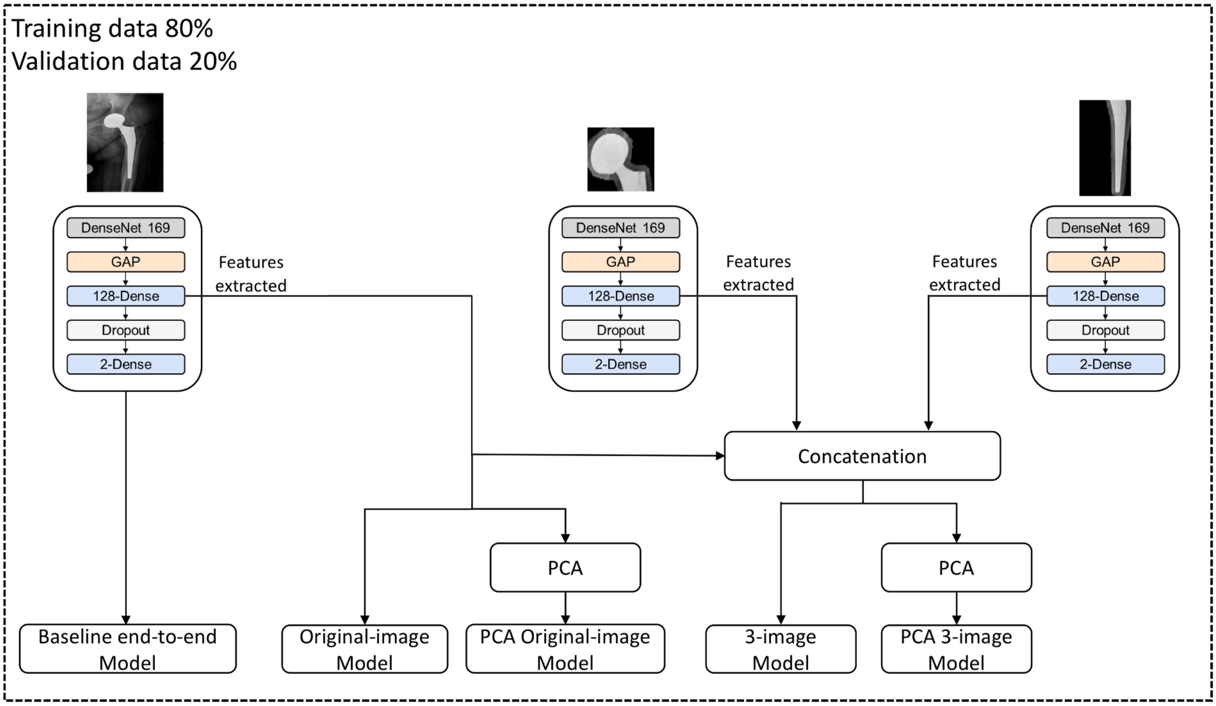


**Fig. 2.** Workflow of the baseline DL and feature-based DL-ML pipelines.

based on data from cohort 1. All the pipeline was performed using Python 3.7 (TensorFlow and Scikit-learn).

Starting from a baseline end-to-end DL pipeline, two combined DL-ML feature-based pipelines were developed, namely original-image and 3-image pipelines. A vector of 128 deep features (DFs) was extracted from the 128-Dense layer output. These features, representing the compressed information used by the network for final classification, were fed into a ML model to develop the original-image pipeline and classify failed and non-failed images.

The 3-image pipeline came from the intuition of using different image resolutions of DeepMedic [12]. Specifically, we considered the original image and the images containing the acetabular and the stem components, respectively. From these additional images, two distinct different deep models, with the same architecture as for the baseline model, were trained and 128 DFs were extracted similarly. Thus, a vector of 384 DFs was obtained combining the information of the three images, merging the DFs from the global original image with local information.

For both DL-ML pipelines, two methods were employed, namely: i) considering all the features, and ii) reducing the dimensionality through Principal Component Analysis (PCA), keeping 90, 95 or 97% of explained variance.

Finally, classification was performed with two ML models: a support vector machine (SVM) and a random forest (RF) model.

A stratified data split procedure according to the patients was employed to the first cohort of patients to develop the model (80% of patients for the training and the remaining 20% for validation). Thus, the same proportion of the failed and non-failed groups was maintained in the two groups and each patient was included in only one subset. This procedure was repeated 50 times for cross-validating the results and for the optimal model choice.

### *2.5. Model validation*

After testing the models with cross-validation, the final model was trained on the whole dataset of cohort 1 and externally validated on the

771 images of cohort 2.

### 2.6. *Model explainability*

In order to interpret the ML models and their prediction, the SHapley Additive exPlanations (SHAP) method [13] was used. Starting from a ML model, the SHAP method calculates each feature contribution by determining the alteration in the model expected prediction based on a specific feature. Our features, being extracted from a deep model, do not have a direct association with the images. Nevertheless, the SHAP method is thought to be valuable in this scenario, potentially enabling us to examine which groups of DFs significantly impact the classification and the extent of their influence.

### 2.7. *Statistical analysis*

To compare the difference in the accuracy obtained by the proposed pipelines, the Mann-Whitney *U* test was applied in case of comparison between two groups. The statistical analyses were performed in Matlab.

## 3. Results

### 3.1. *Model development*

#### 3.1.1. *Baseline pipeline*

Training and validation behaviors as a function of number of epochs for the baseline model are reported for a single repetition (over 50) in Fig. 3. The plateau reached in the accuracy during validation is very similar to the one obtained during training, meaning that overfitting was avoided. This model achieved a validation accuracy of 0.936 ± 0.010 averaged on the 50 repetitions.

#### 3.1.2. *Feature-based DL-ML pipeline*

In Table 2, the results in terms of accuracy on the validation set averaged on the 50 repetitions for the various feature-based approaches are shown. Since there were no observable differences in using different percentages of kept variance when using the PCA, only results with the 90% of kept variance are presented. It could be observed that there was no degradation in the accuracy between the baseline model and the original-image pipeline that fed a ML model (either SVM or RF) with the 128 DFs. Overall, the SVM slightly outperformed the RF, particularly in the 3-image pipeline. The 3-image pipeline (with and without performing PCA) with the SVM provided the best performance, with a significant difference in the obtained accuracy with respect to the baseline pipeline ($p < 0.001$) and the original-image pipeline ($p < 0.001$). Fig. 4 shows the ROC curves for the various feature-based approach for SVM and RF models.

### 3.2. *Model validation and explainability*

Given the higher performance of the 3-image pipeline using SVM and direct feature combination, this approach has been applied on the external validation set (cohort 2). The following performance metrics (mean and confidence interval obtained by bootstrapping) were achieved: specificity 0.863 (0.815–0.917), recall 0.919 (0.895–0.937), precision 0.956 (0.941–0.974), AUC 0.961 (0.953–0.973), F1-score 0.874 (0.850–0.897), balanced accuracy 0.861 (0.831 – 0.887).

**Table 2**
Performance in the validation cohort.

| SVM | Original image | PCA original image | 3-image | PCA 3-image |
|---|---|---|---|---|
| **Accuracy** | 0.945 ± 0.009 | 0.939 ± 0.009 | **0.958 ± 0.006** | 0.957 ± 0.005 |
| **Specificity** | 0.961 ± 0.009 | 0.957 ± 0.014 | 0.948 ± 0.010 | 0.948 ± 0.010 |
| **Recall** | 0.929 ± 0.014 | 0.921 ± 0.016 | 0.968 ± 0.010 | 0.966 ± 0.010 |
| **Precision** | 0.955 ± 0.009 | 0.955 ± 0.014 | 0.949 ± 0.009 | 0.949 ± 0.010 |
| **F1 score** | 0.945 ± 0.009 | 0.939 ± 0.009 | 0.958 ± 0.006 | 0.957 ± 0.005 |
| **AUC** | 0.982 ± 0.005 | 0.981 ± 0.005 | 0.986 ± 0.003 | 0.983 ± 0.005 |
| **RF** | **Original image** | **PCA original image** | **3-image** | **PCA 3-image** |
| **Accuracy** | 0.938 ± 0.009 | 0.938 ± 0.009 | 0.943 ± 0.011 | 0.943 ± 0.011 |
| **Specificity** | 0.954 ± 0.012 | 0.954 ± 0.014 | 0.945 ± 0.0125 | 0.934 ± 0.011 |
| **Recall** | 0.923 ± 0.016 | 0.923 ± 0.016 | 0.941 ± 0.024 | 0.952 ± 0.016 |
| **Precision** | 0.947 ± 0.010 | 0.952 ± 0.015 | 0.943 ± 0.013 | 0.934 ± 0.013 |
| **F1 score** | 0.938 ± 0.009 | 0.938 ± 0.009 | 0.943 ± 0.011 | 0.943 ± 0.011 |
| **AUC** | 0.981 ± 0.005 | 0.963 ± 0.013 | 0.986 ± 0.004 | 0.979 ± 0.008 |

SMV: support vector machine, RF: random forest, PCA: principal component analysis, AUC: area under the curve.

Table 3 shows the number and percentage of wrongly classified images as a function of location and cause of failure. It can be noted that the percentage of wrongly classified images is not dependent on the location of failure as the proportion of acetabular, femoral and both locations is similar (and not significantly different) in the whole image set and in the wrongly classified images. Although not statistically significant, the higher percentage of wrongly classified images with failure involving the femoral component with respect to the one in the complete dataset, might be related to lower number of cases with this location of failure: thus the algorithm may be not very well trained on this group. On the contrary, it might be observed that polyethylene wear is the cause of failure in only 10 images, they were all correctly classified by the algorithm.

SHAP analysis was carried out to investigate the feature contribution for the classification on the 3-image pipeline using SVM and direct feature combination. Thus, the examined total feature vector was composed by 384 features (128 for every type of image). Fig. 5 displays

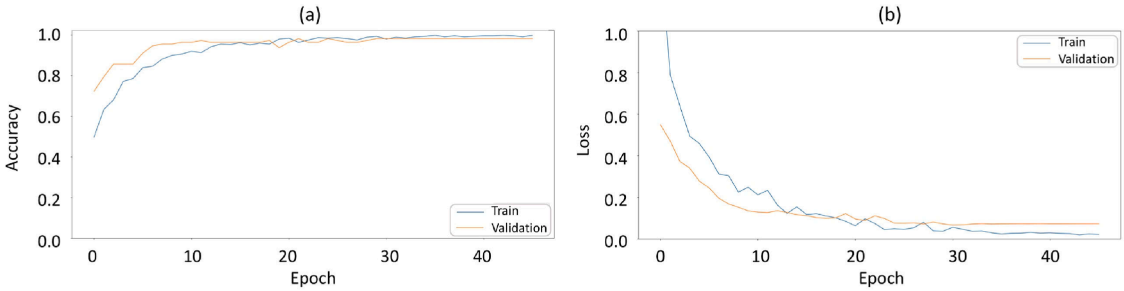


**Fig. 3.** Trend of accuracy (a) and loss (b) in one fold for the baseline pipeline.

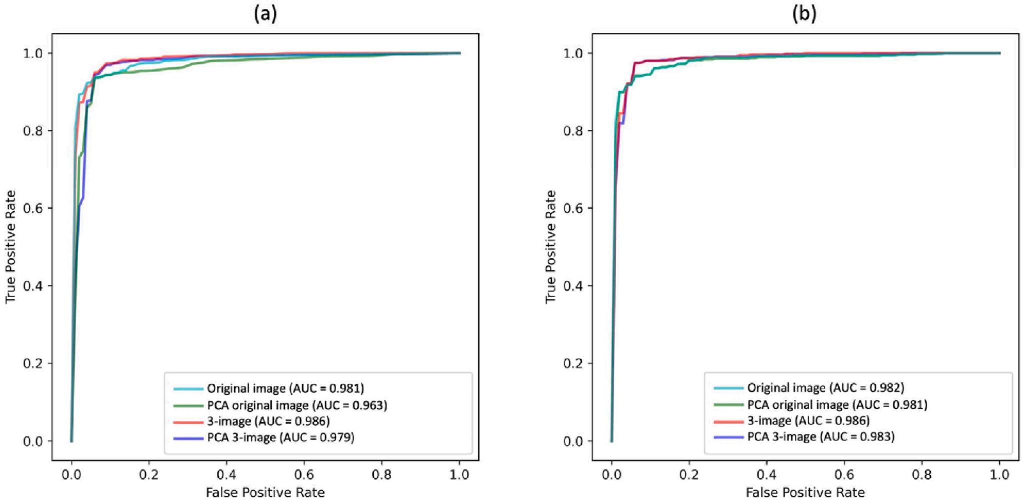


**Fig. 4.** ROC curves for (a) random forest and (b) support vector machine models, for the different feature-based approaches in the validation set (cohort 1).

**Table 3**
Performance in the external validation set (cohort 2).

| | Total number of images | Wrongly classified images | p-value Chi-square test |
|---|---|---|---|
| **Location of failure** | | | |
| Acetabular | 294 (50%) | 17 (50%) | 0.95 |
| Femoral | 37 (6%) | 5 (15%) | 0.06 |
| Acetabular and femoral | 258 (44%) | 12 (35%) | 0.33 |
| **Cause of failure** | | | |
| Aseptic loosening | 577 (98%) | 47 (98%) | 0.98 |
| Polyethylene wear | 10 (1.7%) | 0 (0%) | 0.36 |
| Repetitive luxation | 2 (0.3%) | 1 (2%) | 0.09 |

the SHAP values for the top 20 features that mostly contributed to the SVM model. Fig. 5(a) presents the features sorted in descending order of importance, with their corresponding average absolute SHAP value. In the summary plot (Fig. 5(b)), the importance of each feature is combined with its corresponding effects. Each point on the plot corresponds to a SHAP value of a feature for an instance, where the color indicates the feature value, ranging from low to high. Therefore, the relationship between the value of a feature and the impact on the prediction could be observed. The positive SHAP value indicates the contribution to the classification towards the failed class, negative towards the non-failed. As expected from the small difference in terms of results between the original image and the 3-images, the most impactful features identified by the SHAP analysis belong to the original image, and only two features not belonging to the original image were identified, namely (DF 255 from the acetabulum and DF 370 from the stem).

Fig. 5(c) and (d) show two plots that depict the 20 most significant features, sorted by their SHAP values, for one failed subject (Fig. 5(c)) and one non-failed subject (Fig. 5(d)). These plots show, locally, the relationships between the value of a specific DF of an instance and its contribution to the prediction. In these graphs, the value of the feature is represented next to its name. It can be observed that, coherently with the global analysis, for both the failed and non-failed cases, the major contribution to the prediction was given by features of the original image, but also some features belonging to the acetabulum and stem images were identified among the 20 most significant features. In both the failed and non-failed cases 13 over 20 features were common to the ones identified in the global analysis and their values was consistent with the global impact on the model output. For example, DF 17 was identified as the most impactful feature in both cases as in the global analysis, and its high/low value respectively in the failed/non-failed case is coherent with the relation reported in the summary plot of Fig. 5(b). Moreover, DF 255 for both the failed and non-failed cases and DF 370 for the failed case were identified to contribute to the prediction.

## 4. Discussion

In the clinical practice, the follow-up assessment of hip replacement is currently done with conventional X-rays, and it mainly aims to detect component malalignment, subsidence, prosthesis loosening and polyethylene wear (being the major causes of failure). However, the early detection of these complications on the basis of two-dimensional images (as X-rays are) can be highly challenging for clinicians.

The use of DL methods on radiographs of hip arthroplasty implants has been recently successfully applied for discrimination between different implants [14,15] but only few studies have used it to discriminate healthy from pathologic implants after THA [6,16,17]. Accordingly, the objective of this study was to develop a combined DL-ML pipeline to discriminate failed from non-failed hip prosthesis implants through X-rays images. The results showed that the combination of the global features from the original image with local information extracted from the acetabulum and the stem improved the performance of the classifier, underlying the effectiveness of features concatenation. Furthermore, the proposed DL-ML pipeline reduced the dimension of image features, and their memory-consuming, without affecting the classification accuracy. The similarity of the results obtained from the models trained on the original images and on the 3-images, suggested that features extracted from the original images have a great role in failure identification. This hypothesis was confirmed by the SHAP analysis, which identified the features of the complete original images as the major contributor. Moreover, among the 20 most impactful features, there are one derived from the acetabulum and one from the stem, thus highlighting the role of the acetabulum and stem images on the

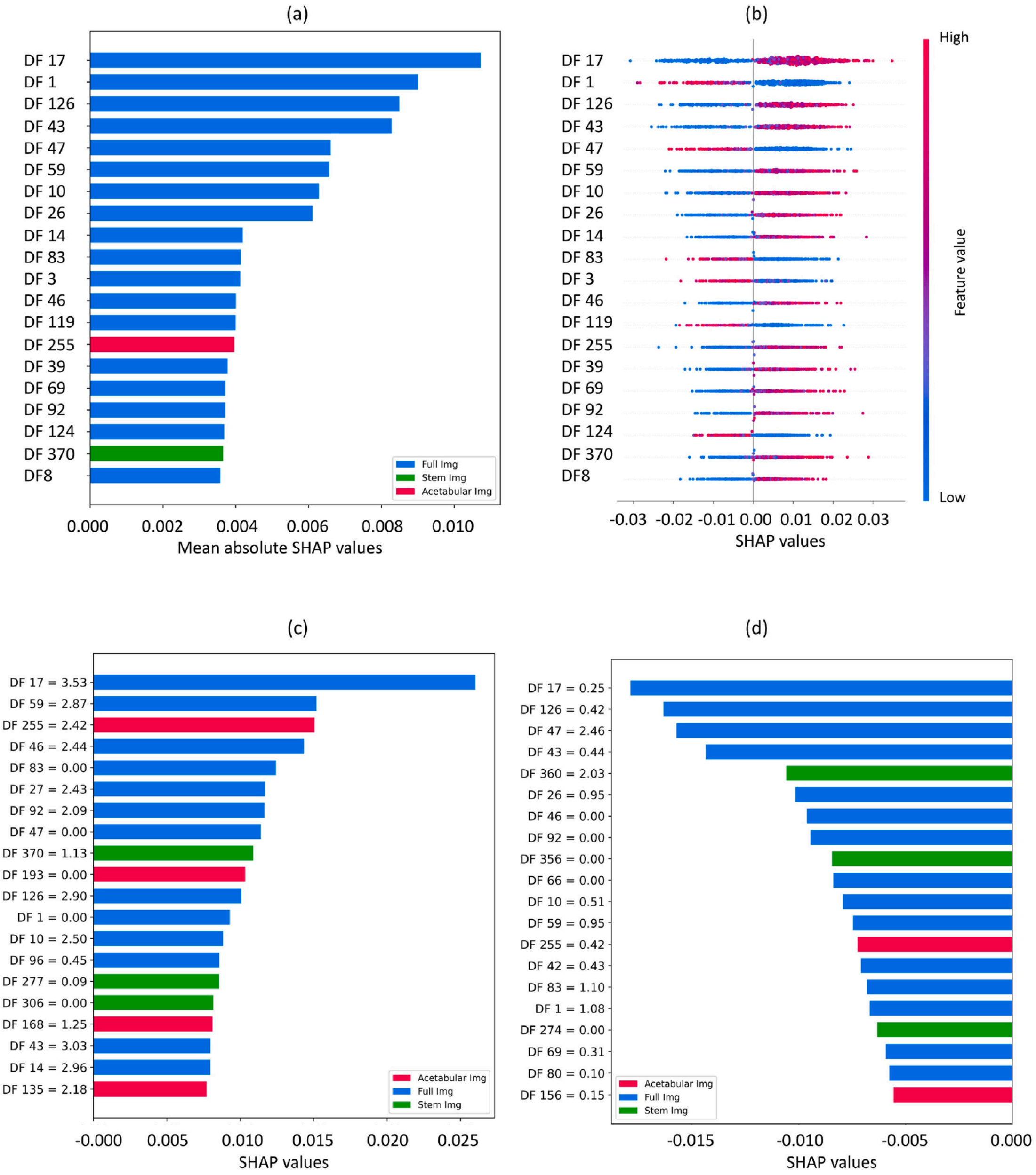


**Fig. 5.** Mean SHAP values for Deep Features (DF) of SVM (a) and SHAP values with feature values (b). Example of SHAP values for DFs of SVM for a (c) failed and (d) non-failed images.

prediction.

Overall, a high accuracy (>94%) was obtained in the present study by both the baseline and the DL-ML pipelines, with the highest accuracy of 96% achieved by the 3-image feature-based pipeline. These results demonstrated that the developed DL-ML approach outperformed the diagnostic ability of the orthopedic surgeon in detecting THA implant mobilization on plain radiographs, that Temmerman et al. estimated to be respectively 81% and 74% with poor inter-observer agreement [18]. However, future studies are needed to compare the discriminating performance of the DL-ML approach *vs.* orthopedic surgeon expertise and to quantitatively evaluate the incrementing effect in detecting THA implant mobilization that the DL-ML approach can provide to the surgeon, if used as a tool for decision support.

Compared to the previous investigation [6], in this study a more comprehensive and robust analysis approach was proposed. Indeed, while in [6] data were not balanced and only the baseline pipeline on the original image was applied, herein (i) data were balanced in the development cohort, (ii) an external validation cohort of 352 patients was tested, (iii) a feature-based approach was proposed, by applying ML classification methods on the features extracted from the last DenseNet layer of the baseline pipeline, (iv) two feature-based DL-ML pipelines were developed, one considering the original image and one considering also the acetabulum and stem images, (v) the impact of PCA versus direct concatenation of the features was explored and (vi) an explainability analysis was performed.

Similar to our studies, other investigations [16,17] confirmed the good performance of DL methods and their efficiency compared to manual classifications. However, compared to these studies, our work allowed either addressing some limitations or improving the performance. First, the study by Borjali et al. [16] had the serious limitation of using a very small dataset of 40 patients without an entirely separate dataset for testing and only AP images were used. Differently, the approach presented herein is more robust due to the larger dataset size, along with the use of the LAT images improving the flexibility of the algorithm with the benefit of additional views, and the external dataset used for validation. Second, compared to the study by Shah et al. [17], herein a higher accuracy was obtained. Indeed, Shah et al. [17] reported an accuracy of 88%, reached by the model based on both x-ray image features and clinical information, whereas a decreased accuracy of 70%, obtained by the image-based model. Moreover, in the study by Shah et al. [17] an external validation cohort for model evaluation was lacking, while it was considered in this study.

In clinical practice the output of this algorithm shall be considered as an additional tool for the orthopedic surgeon to be integrated in the patient's clinical picture during THA follow-up. Indeed, a negative outcome of this algorithm in an asymptomatic patient shall not drive straightforward to a revision surgery but to a stricter follow-up.

Indeed, future studies should investigate AI algorithms combining both radiological and clinical data as joint-specific and general health scores. This effort in improving our tools for the correct follow-up after THA should be put in the context of the significant advantages that an earlier identification of failure could give: that is to prevent extensive surgery in favor of less invasive ones, hence reduced complications, better clinical outcomes for the patients, lower costs and better post-operative recovery [19].

A limitation of this study is that all images were retrospectively collected from patients undergoing partial or total hip replacement revision and patients who underwent primary THA without any clinical or radiographic signs to suspect the failure of the implant. Therefore, the algorithm was not tested in patients who had a clinical or radiographic suspicion of failure but did not have surgery. Future efforts will be needed to confirm the ability of the network to classify those patients too.

## 5. Conclusion

A combined DL-ML approach allowed to detect the failure of the hip prosthesis from plain radiographs with a very high degree of precision. Moreover, including features extracted from the acetabular and the stem components improved the performance of the classifier. The proposed approach might be used in the follow-up of patients with hip replacement as a tool for the identification of implant failure.

## 6. Summary table

| | |
|---|---|
| **Problem** | The progressive population aging, and the growing number of hip prosthesis procedures call for the need of automatic detection of hip prosthetic failure to optimize the patient follow-up. |
| **What is already known** | Deep learning can successfully analyze x-ray images, and a preliminary study showed that the automatic detection of hip prosthetic loosening through a deep learning approach is feasible. |
| **What this paper adds** | A robust combined deep learning and machine learning approach can detect the failure of the hip prosthesis by enhancing the role of the stem and acetabular component analysis, with a recall of 0.92 in an external validation group of 352 patients. |

## Funding

This research received funding from the Ministry of Health (grant number: GR-2018–12367275.

## CRediT authorship contribution statement

**Federico Muscato:** Software, Formal analysis, Methodology, Writing – original draft. **Anna Corti:** Visualization, Writing – review & editing. **Francesco Manlio Gambaro:** Data curation, Writing – review & editing. **Katia Chiappetta:** Data curation, Writing – review & editing. **Mattia Loppini:** Conceptualization, Data curation, Writing – review & editing. **Valentina D.A. Corino:** Conceptualization, Methodology, Writing – original draft, Supervision.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Mattia Loppini reports financial support was provided by Ministry of Health. Valentina D.A. Corino reports financial support was provided by Ministry of Health.].

## References

[1] Report Annuale Riap 2019 e Compendio, RIAP. (2020). https://riap.iss.it/riap/it/attivita/report/2020/10/19/report-riap-2019/ (accessed July 9, 2021).

[2] S. Kurtz, K. Ong, E. Lau, F. Mowat, M. Halpern, to 2030, J. Bone Joint Surg. Am. 89 (4) (2007) 780–785.

[3] K.J. Bozic, S.M. Kurtz, E. Lau, K. Ong, T.P. Vail, D.J. Berry, The epidemiology of revision total hip arthroplasty in the United States, J. Bone Joint Surg. Am. 91 (2009) 128–133, https://doi.org/10.2106/JBJS.H.00155.

[4] D.H. Kim, T. MacKinnon, Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks, Clin. Radiol. 73 (2018) 439–445, https://doi.org/10.1016/j.crad.2017.11.015.

[5] K.A. Thomas, Ł. Kidziński, E. Halilaj, S.L. Fleming, G.R. Venkataraman, E.H.G. Oei, G.E. Gold, S.L. Delp, Automated Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks, Radiol. Artif. Intell. 2 (2) (2020) e190065.

[6] M. Loppini, F.M. Gambaro, K. Chiappetta, G. Grappiolo, A.M. Bianchi, V.D. A. Corino, Automatic Identification of Failure in Hip Replacement: An Artificial Intelligence Approach, Bioengineering (Basel). 9 (2022) 288, https://doi.org/10.3390/bioengineering9070288.

[7] Y. Ren, S. Wu, M. Wang, Z. Cen, Study on Construction of a Medical X-Ray Direct Digital Radiography System and Hybrid Preprocessing Methods, Comput. Math. Methods Med. 2014 (2014) 1–7.

[8] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: Graphics Gems IV, Academic Press Professional Inc, USA, 1994, pp. 474–485.

[9] J. Canny, A Computational Approach to Edge Detection, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8 (6) (1986) 679–698.

[10] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, in, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017 (2017) 2261–2269, https://doi.org/10.1109/CVPR.2017.243.

[11] ImageNet Large Scale Visual Recognition Challenge | SpringerLink, (n.d.). https://link.springer.com/article/10.1007/s11263-015-0816-y (accessed July 9, 2021).

[12] K. Kamnitsas, C. Ledig, V.F.J. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78, https://doi.org/10.1016/j.media.2016.10.004.

[13] C. Molnar, Interpretable Machine Learning, n.d. https://christophm.github.io/interpretable-ml-book/ (accessed March 8, 2023).

[14] A. Borjali, A.F. Chen, O.K. Muratoglu, M.A. Morid, K.M. Varadarajan, Detecting total hip replacement prosthesis design on plain radiographs using deep convolutional neural network, J. Orthop. Res. 38 (2020) 1465–1471, https://doi.org/10.1002/jor.24617.

[15] J.M. Karnuta, H.S. Haeberle, B.C. Luu, A.L. Roth, R.M. Molloy, L.M. Nystrom, N. S. Piuzzi, J.L. Schaffer, A.F. Chen, R. Iorio, V.E. Krebs, P.N. Ramkumar, Artificial Intelligence to Identify Arthroplasty Implants From Radiographs of the Hip, J Arthroplasty. 36 (7) (2021) S290–S294.e1.

[16] A. Borjali, A.F. Chen, O.K. Muratoglu, M.A. Morid, K.M. Varadarajan, Detecting mechanical loosening of total hip replacement implant from plain radiograph using deep convolutional neural network, ArXiv:1912.00943 [Cs, Eess]. (2019). http://arxiv.org/abs/1912.00943 (accessed July 9, 2021).

[17] R.F. Shah, S.A. Bini, A.M. Martinez, V. Pedoia, T.P. Vail, Incremental inputs improve the automated detection of implant loosening using machine-learning algorithms, Bone Joint J. 102-B (2020) 101–106, https://doi.org/10.1302/0301-620X.102B6.BJJ-2019-1577.R1.

[18] O.P.P. Temmerman, P.G.H.M. Raijmakers, J. Berkhof, E.F.L. David, R. Pijpers, M. A. Molenaar, O.S. Hoekstra, G.J.J. Teule, I.C. Heyligers, Diagnostic accuracy and interobserver variability of plain radiography, subtraction arthrography, nuclear arthrography, and bone scintigraphy in the assessment of aseptic femoral component loosening, Arch. Orthop. Trauma Surg. 126 (2006) 316–323, https://doi.org/10.1007/s00402-006-0120-y.

[19] M. Loppini, F.M. Gambaro, R.G.H.H. Nelissen, G. Grappiolo, Large variation in timing of follow-up visits after hip replacement: a review of the literature, EFORT Open Rev. 7 (2022) 200–205, https://doi.org/10.1530/EOR-21-0016.