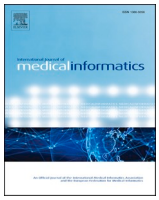
International Journal of Medical Informatics 196 (2025) 105802

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



Hip prosthesis failure prediction through radiological deep sequence learning

Francesco Masciulli [a], Anna Corti [a], Alessia Lindemann [a], Katia Chiappetta [b], Mattia Loppini [b,c,d], Valentina D.A. Corino [a,e,*]

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

ARTICLE INFO

Keywords:
Artificial intelligence
Hip replacement
Image classification
Temporal dependency

ABSTRACT

Background: Existing deep learning studies for the automated detection of hip prosthesis failure only consider the last available radiographic image. However, using longitudinal data is thought to improve the prediction, by combining temporal and spatial components. The aim of this study is to develop artificial intelligence models for predicting hip implant failure from multiple subsequent plain radiographs.
Methods: A cohort of 224 patients was considered for model development and a balanced cohort of 14 patients was used for external validation. A sequence of two or three anteroposterior radiographic images per patient was considered to track the prosthesis over time. A combination of a convolutional neural network (CNN) and a recurrent section was used. For the CNN, a pretrained autoencoder, a pretrained RadImageNet DenseNet and a pretrained custom DenseNet were considered. The recurrent section was implemented using either a single Gated Recurrent Unit (GRU) layer or a Long Short-Term Memory block.
Results: Considering 3 images as input provided a positive predictive value (PPV) of 0.966 and an f1 score of 0.933 on the validation set. Regarding the 2-image models, using the postoperative and the last image resulted in PPV of 0.933 and f1 score of 0.918, whereas using the second-to-last image with the post-operative one reached a PPV of 0.882 and f1 score of 0.923. On the external validation set, the 3-image model reached an accuracy of 0.786.
Conclusion: This study demonstrated the potential of the developed models, based on a series of plain radiographs, to predict hip prosthesis failure.

1. Introduction

Total hip arthroplasty (THA) is a widely performed procedure worldwide, offering highly effective treatment for various hip conditions in individuals of different age groups. More than 3.1 million THAs are performed every year in Europe [1] to increase quality of life and improve the mobility of patients to enhance their life expectancy. Although more common among the elderly, approximately 25% of joint replacement recipients are under the age of 65. With the aging population and an increasing number of procedures in younger patients, the placement of primary prosthetic implants is expected to rise by 2030, along with a corresponding increase in revision surgeries [2,3]. The main causes of THA failure are aseptic loosening, bearing surface wear, and osteolysis. Conducting regular radiographic follow-ups can help identifying complications and failures earlier, enabling more conservative revisions that lead to favorable functional outcomes. However, detecting loosening remains challenging, and a definitive diagnosis often requires revision surgery.

Some preliminary studies analyzed plain radiographs also in combination with other imaging techniques, finding either very low inter-observer agreement in radiograph evaluation [4,5] or very low specificity [6]. Leveraging Artificial Intelligence (AI), particularly the Deep Learning (DL) approach, to automatically evaluate X-ray images for monitoring patients with hip arthroplasties can improve the

* 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.2025.105802
Received 14 May 2024; Received in revised form 20 December 2024; Accepted 21 January 2025
Available online 22 January 2025

diagnostic accuracy for THA failure. Automatic classification algorithms, such as DL-based Convolutional Neural Networks (CNN), are powerful tools that extract valuable information from large datasets. DL is a branch of research in machine learning that employs hierarchical levels and different concept representations, allowing for complex processing structures. CNNs have already demonstrated significant success in various orthopedic applications, including the identification of fractures in the wrist, elbow, ankle, and hip, as well as the classification of knee osteoarthritis [7,8]. Recent studies have demonstrated the feasibility to automatically detect hip prosthetic failure from plain radiographs through a DL approach, based on CNN [9] and on a DL-ML procedure [10], using the last available radiograph before revision surgery. Other studies investigated DL models including as input not only the radiograph but also patient historical and comorbidity information [11], or a stacking classifier using three different CNN models to classify aseptic hip loosening [12]. Moreover, the evolution of the periprosthetic bone and implant's position throughout the entire follow-up period has shown the potential to be more relevant in outcome prediction than simply considering the latest radiographic images. Recently, an AI-based model employing four evolutional parameters derived from conventional radiological features has been developed to predict hip prosthesis failure, with an advance between a few months and more than a year [13].

However, to the best of our knowledge, no studies have used sequential radiographic images as input of a DL model to predict prosthesis failure. Also, no studies have assessed the diagnostic performance of plain radiograph in prosthesis failure prediction at different time points during the follow-up period. Thus, the aim of this study is to enhance the prediction of failed prosthesis in THA by combining multiple radiographic images in sequences as input of a neural network. Our hypothesis is that, by including a temporal component as input, in addition to the spatial features contained in the individual images, an earlier identification of failure will be achieved.

2. Materials and methods

2.1. Population

2.1.1. Model building

We considered a consecutive cohort of 224 patients undergoing primary THA for primary hip osteoarthritis without associated procedures at our tertiary referral institution between January 2010 and January 2020. Among these patients, 91 needed a total or partial revision surgery due to implant failure and were then included in the failed group, while the remaining constituted the non-failed group. The main reason for revision surgery was represented by aseptic loosening (59 %), followed by recurrent dislocations (18.6 %), periprosthetic fractures (17.4 %), ileo-psoas conflict (2.5 %) and septic loosening (2.5 %). The 133 non-failed subjects were verified to be still not at risk after a median of 132 months (IQR: 120–132 months) from the last considered visit.

For each patient, a series of 2D radiographic images acquired during clinical follow-up was available, consisting in the anteroposterior (AP) view, catching the whole pelvis area. Thus, each available image consisted of a radiography corresponding to a single patient at a specific follow-up visit. The number of visits performed by each patient was highly heterogeneous, representing a key feature characterizing the dataset. In most of the failed patients, the available images were less than four, a number much lower than in non-failed case (median [interquartile range] for failed group: 3 [2-4], for non-failed group 6 [5-7]). In addition, as expected, the patients from the non-failed group have a longer maximum time elapsed between the primary surgery and the last available visit. To overcome this heterogeneity, a cutoff at 60 months after the primary surgery was applied and only follow-up visits during this period were included in the analysis. Thus, the final dataset was composed of 205 patients (129 non-failed and 76 failed), with a variable number of visits as described in the results section. Demographics and main characteristics of the analyzed patients are presented in Table 1. Eventually, 2291 unique images, from these patients, were available in DICOM (Digital Imaging and COmmunication in Medicine) form, a standard typically used for communicating and managing clinical images.

2.1.2. Model external validation

A second dataset, including 14 patients (7 failed and 7 non-failed) was considered and used as external validation set to test the best performing model. All patients had three available images.

All included patients provided their informed consent, and this clinical study was approved by the Ethical Committee of Humanitas Research Hospital on 25/06/2020 (code: ICH-ORT-2019-01).

2.2. Methods

Image quality was appropriate for the analysis of all images after preprocessing that was performed similarly to [9] and [10]. In particular, the following pre-processing steps were performed to enhance image contrast and detail information [14,15]: i) mist effect reduction and gamma correction, ii) sigmoidal cutoff for contrast improvement, iii) adaptive histogram equalization, and iv) Gaussian filtering. Moreover, images of bilateral implants were split so that images from both hips were included in the analysis and images from the same patients (left and right hip) were included only in one set, either training or validation.

The training phase of the model was designed to ensure robust performance while minimizing overfitting. Data augmentation was applied during the training phase to address the challenges posed by potential noise in radiographic imaging and the limited size of the dataset. Data augmentation involved generating variations of the input images at each epoch through random transformation functions. These transformations included width and height translations (up to 10 %), zoom-in and zoom-out (zoom range: 0.8 to 1.2), and brightness adjustments (brightness range: 0.8 to 1.2). Horizontal image flipping and other transformations were omitted for consistency with clinical imaging standards.

The intention was to simulate realistic scenarios in which radiographs may be affected by noise or variability, reflecting the dataset's composition of everyday check-up images. By exposing the model to these variations, its generalization capabilities were enhanced.

The model was trained using Adam optimizer with a learning rate of 10^{-5}, and the binary cross-entropy loss function, which is suitable for determining the likelihood of failure in the hip arthroplasty binary classification task. To account for the dataset imbalance, class weights were incorporated into the loss function. These weights ensured that each class's contribution to the loss was inversely proportional to its frequency in the training set. This adjustment helped the model to balance the importance of both classes, reducing the risk of bias towards the majority class and improving its ability to correctly predict both outcomes. The training was performed using back-propagation, with the model's performance monitored on a separate cross-validation set. Early stopping was implemented as a regularization strategy to limit overfitting. Specifically, training was halted when the validation loss did not

Table 1
Patient characteristics.

Variable	Failed (n = 76)	Non-failed (n = 129)
Median age (IQR)	62 (55–75)	66 (58–74)
Sex (male), n (%)	21 (65.6)	73 (56.2)
Acetabular shell revised	17 (22 %)	
Acetabular liner revised	29 (40 %)	
Femoral component revised	33 (46 %)	
Femoral head revised	46 (64 %)	
Time of the first image	1 [1–1]	1 [1-2]
Time of the second image	11 [4.5–12]	24 [11.75–36]
Time of the third image	14 [7.5–36]	48 [36–60]

improve for a predefined number of epochs. Additionally, the model retained the weights from the epoch with the best validation performance, ensuring that the final model captured the optimal state achieved during training. This approach prevented the model from overoptimizing on the training set and ensured its ability to generalize effectively.

The following performance metrics were used to evaluate the networks: sensitivity, specificity, Positive Predictive Value (PPV), f1 score and the Area Under the Receiver Operating Characteristic Curve (AUC).

This study focused on providing a sequence of consecutive images to be exploited by an architecture that can track the prosthesis over time. To develop such architecture, inspired by previous works [3,8], we considered a combination of a CNN, to extract spatial features from each image separately, and a recurrent section, to learn the temporal correlation from the produced latent representations. The CNN analyzes one image at the time while the recurrent section update is based on the CNN output sequence. When backpropagating the gradient, the loss is computed independently for each sequence timestamp and combined to update the CNN in one step.

Two different networks (model 1 and model 2 detailed below) were developed and applied to two or three images: the post-operative radiography was always selected, representing a starting point from which the prosthesis evolution over time was modelled, then the last and/or second-to-last image were added. In particular, first the 3-image model was developed, and then the best performing networks with the 3-image case were applied to the 2-image problem. A training-test split was performed (75–25 %).

1) Baseline

The first analysis consisted in using a pre-trained DenseNet network as in [9,10] to a single image, to understand the relative importance of each image in failure prediction, unfreezing the last 8 layers.

2) Model 1

Fig. 1(a) represents the proposed architecture, highlighting the main components. The input image is fed to a CNN to extract a latent

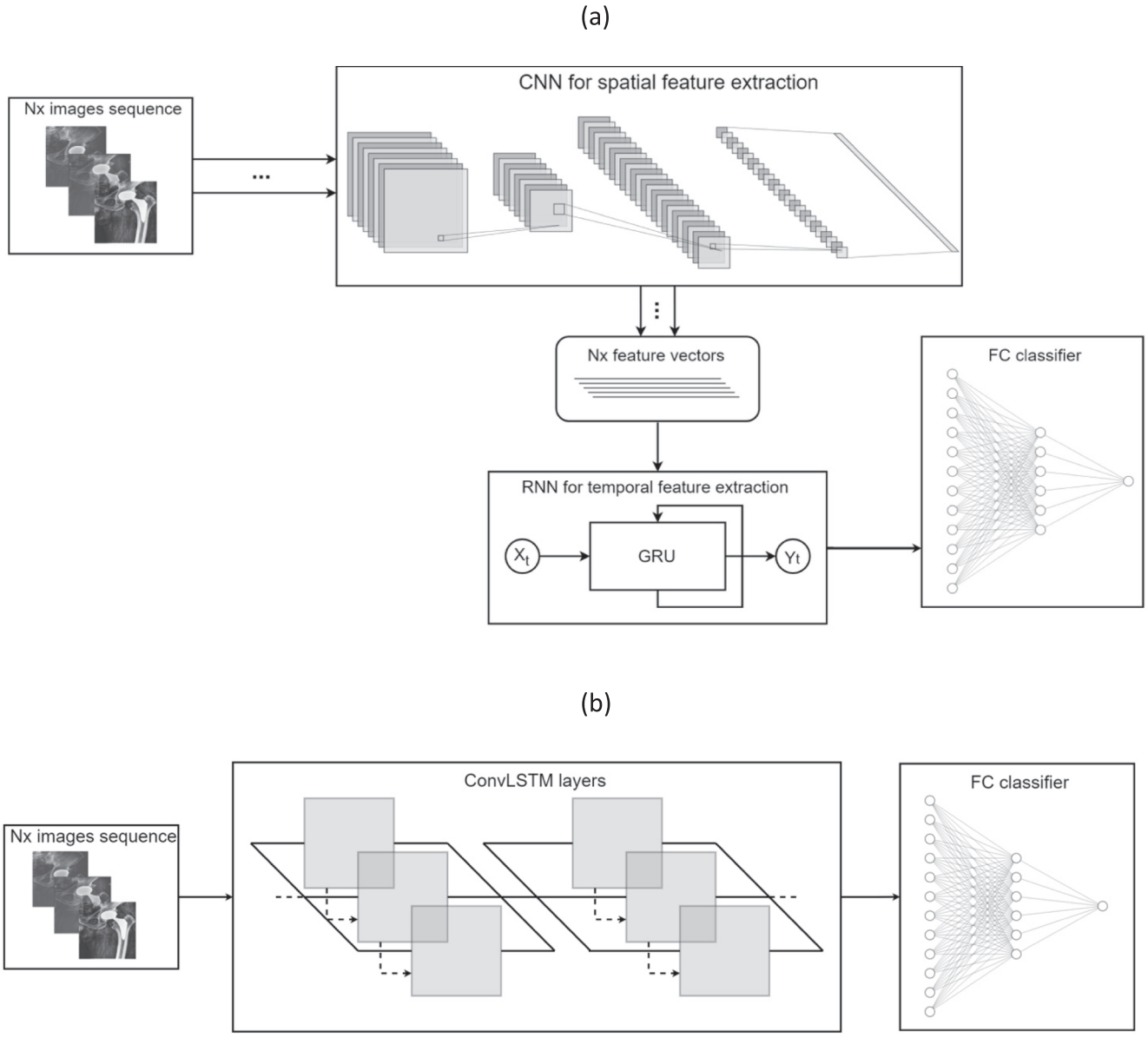


Fig. 1. Schematic representation of the network architecture for (a) model 1 and (b) model 2. Nx: number of images composing each sequence; CNN: convolutional neural network; RNN: recurrent neural network; FC: fully-connected.

representation from each image, thus modeling each radiography's spatial features. The peculiarity of this class of architecture is that the convolutional section can be any sort of CNN, giving this class a task customization property. In this study, we tested three different CNN models: a pretrained autoencoder (AE) [16], a pretrained RadImageNet Densenet [17] (named DenseNet1), and a pretrained custom DenseNet [9] (named DenseNet2), as detailed in the following.

The recurrent section was implemented with a single Gated Recurrent Unit (GRU) layer. The GRU is composed of a number of units equal to the CNN output vector dimension. To limit the model complexity, the number of units was empirically set to 512. Finally, the output of the RNN was flattened and fed into a fully-connected classifier (composed of two dense layers with, respectively, 32 and 1 neurons and a ReLU and Sigmoid activation functions), that predicts the actual class for each sequence based on the received information.

3) Model 2

The other implemented architecture exploits the paradigm proposed by Shi et al. [18] of the Convolutional LSTM (ConvLSTM) block (Fig. 1 (b)), exploiting the ConvLSTM block as a spatial–temporal feature extractor, following the architecture proposed in [19]. The architecture implemented in this study consisted in an initial convolutional backbone that acts as a spatial feature extractor of the single images with the purpose of reducing the dimensionality of the input. As initial convolutional backbone, the same architectures as in model 1 were evaluated (a pretrained RadImageNet Densenet [17] (named DenseNet1), and a pretrained custom DenseNet [9] (named DenseNet2). The extracted latent representations were subsequently fed to a bidirectional ConvLSTM block, which learns how to extract the spatial–temporal correlations between the x-rays composing each sequence. The output was finally flattened through a global max-pooling layer, and a simple fully-connected classifier was placed on top of the neural network, which is composed of two dense layers with, respectively, 32 and 1 neurons activated by a ReLU and a Sigmoid function.

To compare the diagnostic performance of the different models, the Hanley&McNeil test [20] was used.

3. Results

3.1. Available images

205 patients (129 non-failed and 76 failed) satisfied the conditions of a minimum of two images at the 60-month cutoff. When requesting three or four images, the number of patients was reduced to 160 (112 non-failed and 48 failed) and 103 (77 non-failed and 26 failed), respectively. Due to the very low number of patients with four images, the following analyses were performed considering only two or three images.

3.2. Validation results

When using the DenseNet network as in [9,10] to a single image, the last available x-ray image was the most informative one with an AUC of 0.839, while the second-to-last or the post-surgery ones achieved a much lower AUC of 0.785 and 0.704, respectively. The additional metrics shown in Table 2 also confirm the better performance obtained using the last available x-ray, even if good performance can be obtained using the second-to-last, with a f1 score of 0.833.

Table 3 shows the results obtained using three images with all the combinations of CNN and RNN networks. It can be noted that, as CNN, the DenseNet2 is the one achieving the best results, with an average increase in the performance metrics of 0.132 ± 0.053. Regarding the RNN block, a less evident trend can be observed, with the GRU achieving slightly higher performance. The best AUC was obtained combining DenseNet2 with the GRU block, being its value significantly higher than the ones obtained with the AE (p = 0.05) or with DenseNet1 (GRU p = 0.09, cLSTM p = 0.06). Supplementary Fig. S1 shows the calibration curve for the best performing model (DenseNet2 combined with GRU block).

Table 4 shows the results obtained using DenseNet2 and different combinations of two images with either GRU or ConvLSTM as RNN block. In terms of global performance (evaluated by f1 score, AUC and PPV), using the last available x-ray improved the prediction with respect to using the second-to-last image. The performance of these models was inferior to the one obtained when using three images. In particular, using the post-operative and the last images, performance is slightly inferior (p not significant), while it is almost significant (p = 0.07 and p = 0.09) when using the second-to-last image. Finally, a trend of slight improvement was obtained compared to the use of a single image, though not significant. Supplementary Fig. S2 and S3 show the calibration curves for the2-image DensNet2-GRU models, considering the post-operative and second-to-last, and the post-operative and last images, respectively.

Table 5 shows a comparison of the results obtained using DenseNet2 with two or three images and GRU. Results show that including the last image to a 2-image model significantly improves the results. Results of the 3-image model are also significantly better than the 2-image model without the last image. The model with the three images showed

Table 2
Results using a single image on the validation set using the baseline model.

	Post-operative x-ray	Second-to-last available x-ray	Last available x-ray
Sensitivity	0.742	0.806	0.936
Specificity	0.583	0.667	0.667
PPV	0.821	0.862	0.879
F1 score	0.780	0.833	0.906
g-mean	0.658	0.733	0.790
AUC	0.704	0.785	0.839

Table 3
Performance metrics obtained for each combination of CNN and RNN applied to three images (post-operative, second-to-last and last available) on the validation set.

	AE-GRU	AE-cLSTM	D1-GRU	D1-cLSTM	D2-GRU	D2-cLSTM
Specificity	0.742	0.677	0.710	0.710	0.903	0.936
Sensitivity	0.833	0.833	0.750	0.750	0.917	0.833
PPV	0.920	0.913	0.880	0.880	0.966	0.936
F1 score	0.821	0.778	0.786	0.786	0.933	0.936
g-mean	0.786	0.751	0.730	0.730	0.910	0.883
AUC	0.820	0.796	0.839	0.812	0.957	0.900

AE: autoencoder, GRU: Gated Recurrent Unit, cLSTM: convolutional Long Short-Term Memory; D: DenseNet; PPV: positive predictive value; AUC: area under the curve.

Table 4
Performance metrics obtained using DenseNet2 and RNN applied to two over three images (among post-operative, second-to-last and last available) on the validation set.

	Post-operative + second-to-last		Post-operative + last	
	GRU	cLSTM	GRU	cLSTM
Specificity	0.839	0.968	0.903	0.936
Sensitivity	0.667	0.667	0.833	0.750
PPV	0.867	0.882	0.933	0.906
F1 score	0.853	0.923	0.918	0.921
g-mean	0.748	0.804	0.867	0.838
AUC	0.831	0.841	0.895	0.882

GRU: Gated Recurrent Unit, cLSTM: convolutional Long Short-Term Memory; PPV: positive predictive value; AUC: area under the curve.

Table 5
Comparison between the best models (DenseNet2 + GRU) through bootstrapping.

	2 images: Post-operative + second-to-last	2 images: Post-operative + last	3 images: Post-operative + second-to-last + last
Specificity	0.839 (0.827–0.851)	0.915 (0.904–0.9253)*	0.909 (0.899–0.918)*
Sensitivity	0.656 (0.627–0.685)	0.825 (0.805–0.846)*	0.908 (0.892–0.924)*$^{+}$
F1 score	0.735 (0.721–0.750)	0.859 (0.847–0.871)*	0.887 (0.877–0.898)* $^{+}$
g-mean	0.737 (0.720–0.755)	0.867 (0.854–0.879)*	0.907 (0.898–0.920)* $^{+}$
AUC	0.825 (0.810–0.840)	0.890 (0.877–0.902)*	0.955 (0.947–0.963)* $^{+}$

GRU: Gated Recurrent Unit, cLSTM: convolutional Long Short-Term Memory; PPV: positive predictive value; AUC: area under the curve. * $p < 0.05$ vs. 2 images: Post-operative + second-to-last; $^{+}$ $p < 0.05$ vs. 2 images: Post-operative + last.

significantly better results when compared to the 2-image model using the last image for all metrics (but specificity).

Fig. 2 shows the ROC curve for the best performing method, i.e., DenseNet2 + GRU, with the confidence interval obtained by bootstrap 100 iterations. In the validation set, only 4 patients were incorrectly classified (one false negative and three false positive), highlighting the good PPV of the proposed algorithm. In addition, the median probability of predicting a true failed patient is 0.971 (interquartile range: 0.801 – 0.997), and of predicting a true non-failed patient is 0.927 (interquartile range: 0.784 – 0.971). On the contrary, the median probability of the misclassified patients is 0.704 (interquartile range: 0.652 – 0.822), showing the less robust prediction.

Fig. 3 shows the single probability of all the patients in the validation set for the best performing method, i.e., DenseNet2 + GRU. As described by the population results, it can be observed that the probability of being correctly classified as failed tends to be higher than the probability of being correctly classified as non-failed.

3.3. *External validation results*

The best performing method, i.e., DenseNet2 + GRU, was used also on the external validation set. The model correctly classified all the failed patients, but only four over seven non-failed patients. The achieved performance metrics are: accuracy = 0.786, precision = 0.85, recall = 0.786.

4. Discussion

In the current clinical practice, assessments during follow-up after hip replacement predominantly rely on traditional x-rays. These evaluations aim to highlight issues like component malalignment, subsidence, prosthesis loosening, and polyethylene wear, being all major contributors to failure. However, the early identification of these complications based on two-dimensional x-ray images presents substantial challenges for clinicians.

Recent studies in the field have witnessed the successful implementation of DL methods on x-ray of hip arthroplasty implants, effectively differentiating between various implant types [21,22]. Nevertheless, only a limited number of studies have applied these methods to differentiate between healthy and pathological implants following THA [9,10,11,23]. Moreover, only one x-ray is commonly considered, namely the last available before the revision surgery for

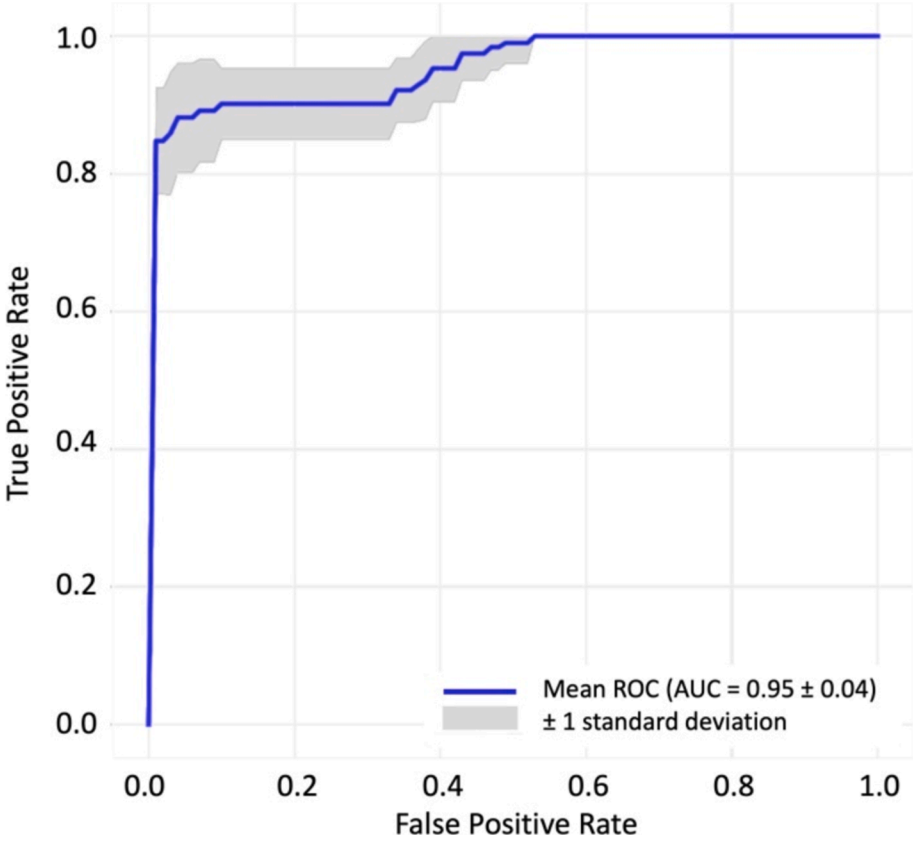


Fig. 2. ROC curve using three images and DenseNet2 (custom pretrained) + Gated Recurrent Unit. Confidence interval obtained by bootstrapping.

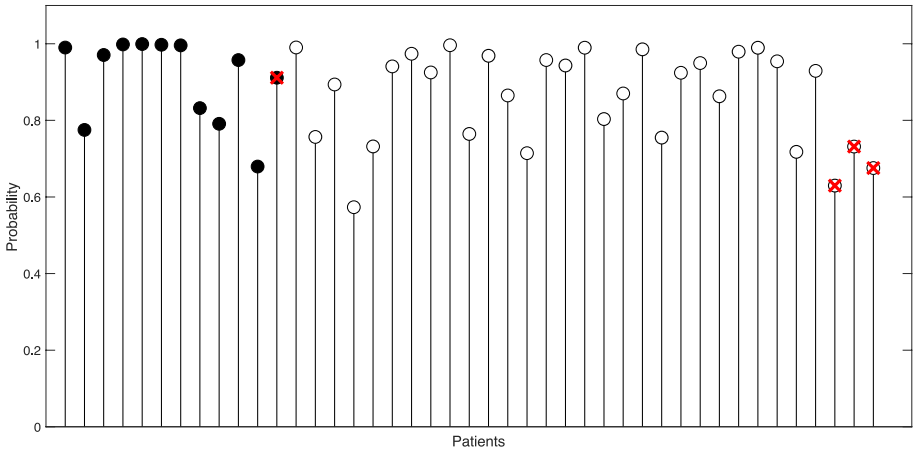


Fig. 3. Probability of the single patients using three images and DenseNet2 (custom pretrained) + Gated Recurrent Unit. Black and white dots represent failed and non-failed patients, respectively. The red crossed are patients wrongly classified.

patients with failed prosthesis.

In this study, we investigated the hypothesis that incorporating a temporal aspect alongside the spatial characteristics from the individual images will improve the detection of failure. In particular, images at different follow-up times were given as input of the network together with the x-ray immediately after surgery, taken as reference image as in clinical practice. The main findings of this study were: i) the possibility to successfully use sequential x-ray images to predict prosthesis failure, ii) the possibility to get an early prediction of failure by considering two images, even considering a follow-up image months before the revision surgery. The best results were obtained considering a 3-image model compared to a 2-image model. However, it is worth noting that the 2-image model considering the post-operative and second-to-last images was able to correctly predict failure with an f1 score of 0.923. These results are better than the ones obtained considering a single image, even the last available one (f1 score of 0.906).

Compared to the previous investigations [9,10], this study represents the first analysis in which temporal variation in hip prosthesis characteristics was considered. While in [9,10] the last available x-ray before revision surgery was used, here the temporal evolution was taken into account by considering two or three sequential images. Moreover, a well-performing model (f1 score = 0.923) was developed considering information up to the second-to-last x-ray, that on average was 8 months before the last visit.

Preliminary clinical studies evaluated the diagnostic accuracy to identify aseptic femoral component loosening of four imaging techniques, namely, plain radiography, subtraction arthrography, nuclear arthrography, and bone scintigraphy [4,5]. They found that plain radiography had a high diagnostic performance for femoral component failure, with a sensitivity and specificity of 81 % and 74 %. However, plain radiography had very low interobserver agreement evaluated by the Intraclass Correlation Coefficient that was only 0.47, highlighting the need for an automatic tool for failure prediction. Another study found radiographic criteria, visually evaluated by the clinician, highly effective in predicting cases that would require revision [6] with a sensitivity reaching 100 %, but with a drop in specificity (45 %). These studies included a manual analysis of the radiographs, with the drawback of being time-consuming and operator dependent.

Other investigations [11,12,23] corroborated the strong performance of DL methods and their efficiency compared to manual classifications. However, compared to these studies, our work provided a step forward, by addressing some limitations and improving the performance. First, the study by Borjali et al. [23] had a serious limitation related to a very small dataset of 40 patients without an entirely separate dataset for testing. Differently, the approach presented herein is more robust due to the larger dataset size, the presence of an external validation set and the inclusion of a series of images from the same patients to improve the prediction. Second, compared to the study by Shah et al. [11], the present work achieved a higher accuracy. Indeed, Shah et al. [11] reported an accuracy of 70 % when using the image-based model, that increased to 88 % only when adding clinical information. Moreover, in the study by Shah et al. [11] an external validation cohort for model evaluation was lacking, while it was considered in this study. Rahman et al. [12] similarly to [10] combined DL and ML approach, by a stacking approach. They trained eight state-of-the-art CNN models, and the top-performing three models were used as base learners, with their predictions used to train ten different machine learning classifiers as meta learners to make the final decision, obtaining an f1 score of 0.94. They performed a five-fold cross validation, while did not have an external validation cohort for model evaluation as in our study. In all the mentioned studies, there was no inclusion of both temporal and spatial characteristics from the individual images, that on the contrary is taken into account in this study, reflecting clinical routine evaluation.

The results of this study may be viewed as an additional resource for orthopedic surgeons, to be integrated into the overall clinical assessment during THA follow-up. It is important to note that a prediction of failure from this algorithm for an asymptomatic patient should not automatically lead to revision surgery, but rather prompt a more vigilant follow-up. Moreover, future investigations should explore AI algorithms integrating both radiological and clinical data over time. Early detection of failure may offer the possibility of less invasive procedures over extensive surgeries, leading to reduced complications, improved clinical outcomes, lower costs, and enhanced post-operative recovery.

A distinctive feature of this study is that all images were retrospectively collected from patients who underwent partial or total hip replacement revision, as well as from patients with primary THA who showed no clinical or radiographic signs of implant failure. The definition of THA “failure” as requiring revision surgery is both a strength and a limitation of the study. On one hand, it offers a comprehensive view of failure, extending beyond the commonly considered implant loosening. On the other hand, the decision to perform a THA revision can be influenced by various factors unrelated to radiographic findings. For instance, implant failure due to infection or pain might present as a fixed and well-positioned prosthesis on radiographs, which could lead the DL model to misclassify such cases as non-failed THA. As a result, the model may face challenges in distinguishing between radiographically similar images of failed and non-failed THA. Moreover, the algorithm was not evaluated in patients who exhibited clinical or radiographic signs suggesting implant failure but did not undergo surgery. Future research efforts will be necessary to verify the model ability to classify such patients as well as to include clinical variables whose evolution over time

has already been proved to be useful in failure prediction [13].

Software availability: the software is available on Zenodo https://doi.org/10.5281/zenodo.13765017.

Funding

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

AC is funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

This research received funding from the European Union - Next generation EU, Missione 4, Componente 1, CUP D53D23013910006 - ID 2022YME9N3 Development and validation of an artificial intelligence algorithm based on clinical and radiographic data for the "virtual follow-up" of patients with total hip arthroplasty.

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. So far, hip prosthesis failure prediction is performed from the last available radiographic image, while longitudinal studies are lacking.
What was already known?	• Deep learning can successfully analyze x-ray images. • The automatic detection of hip prosthetic loosening through a deep learning approach is feasible. • Changes in the periprosthetic bone and implant's position can be indicator of hip prosthesis failure, with an advance of months/years.
What does this study add to our knowledge?	• The analysis on sequential radiographs provides greater insights in hip prosthesis failure, highlighting the effectiveness of using longitudinal data. • When using a single radiograph, the last available radiograph has the best accuracy with AUC = 0.839, followed by the second-to-last available (AUC = 0.785) and the post-operative one (AUC = 0.704). • A sequence of three plain radiographs can be used to predict hip prosthesis failure, including both temporal and spatial characteristics from the individual images, with an AUC of 0.957 and a positive predictive value of 0.966. • When using only two radiographs, AUC decreases (range 0.831–0.895).

CRediT authorship contribution statement

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

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2025.105802.

References

[1] A. Lübbeke, A.J. Silman, C. Barea, D. Prieto-Alhambra, A.J. Carr, Mapping existing hip and knee replacement registries in Europe, *Health Policy* 122 (5) (May 2018) 548–557, https://doi.org/10.1016/j.healthpol.2018.03.010.

[2] 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 (1) (Jan. 2009) 128–133, https://doi.org/10.2106/JBJS.H.00155.

[3] S. Kurtz, K. Ong, E. Lau, F. Mowat, M. Halpern, Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030, *J. Bone. Joint. Surg. Am* 89 (4) (Apr. 2007) 780–785, https://doi.org/10.2106/JBJS.F.00222.

[4] O.P.P. Temmerman, et al., A comparison of radiographic and scintigraphic techniques to assess aseptic loosening of the acetabular component in a total hip replacement, *JBJS* 86 (11) (Nov. 2004) 2456.

[5] O.P.P. Temmerman, et al., 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 (5) (Jul. 2006) 316–323, https://doi.org/10.1007/s00402-006-0120-y.

[6] C. Khalily, L.A. Whiteside, Predictive value of early radiographic findings in cementless total hip arthroplasty femoral components: An 8- to 12-Year Follow-up, *J. . Arthroplasty* 13 (7) (Oct. 1998) 768–773, https://doi.org/10.1016/S0883-5403(98)90028-6.

[7] Y.-H. Kim, S.-H. Oh, J.-S. Kim, Primary total hip arthroplasty with a second-generation cementless total hip prosthesis in patients younger than fifty years of age, *J. Bone. Joint. Surg. Am* 85 (1) (Jan. 2003) 109–114, https://doi.org/10.2106/00004623-200301000-00017.

[8] K.A. Thomas, et al., Automated classification of radiographic knee osteoarthritis severity using deep neural networks, *Radiol. Artif. Intell* 2 (2) (2020), https://doi.org/10.1148/ryai.2020190065 e190065.

[9] 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 (7) (Jun. 2022) 288, https://doi.org/10.3390/bioengineering9070288.

[10] F. Muscato, A. Corti, F. Manlio Gambaro, K. Chiappetta, M. Loppini, and V. D. A. Corino, Combining deep learning and machine learning for the automatic identification of hip prosthesis failure: Development, validation and explainability analysis, *Int. J. Med. Inform.*, vol. 176, p. 105095, Aug. 2023, doi: 10.1016/j.ijmedinf.2023.105095.

[11] R. F. Shah, S. A. Bini, A. M. Martinez, V. Pedoia, and T. P. Vail, Incremental inputs improve the automated detection of implant loosening using machine-learning algorithms, *Bone Joint J*, vol. 102-B, no. 6_Supple_A, pp. 101–106, Jun. 2020, doi: 10.1302/0301-620X.102B6.BJJ-2019-1577.R1.

[12] T. Rahman, et al., HipXNet: Deep Learning Approaches to Detect Aseptic Loos-Ening of Hip Implants Using X-Ray Images, *IEEE Access* 10 (2022) 53359–53373, https://doi.org/10.1109/ACCESS.2022.3173424.

[13] M. Bulloni, F.M. Gambaro, K. Chiappetta, G. Grappiolo, V. Corino, M. Loppini, AI-based hip prosthesis failure prediction through evolutional radiological indices, *Arch. Orthop. Trauma. Surg* (Oct. 2023), https://doi.org/10.1007/s00402-023-05069-5.

[14] 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), https://doi.org/10.1155/2014/495729 e495729.

[15] R. Zhang, Making Convolutional Networks Shift-Invariant Again, arXiv, arXiv: 1904.11486, Jun. 2019. doi: 10.48550/arXiv.1904.11486.

[16] "Image Classification Using Deep Autoencoders | IEEE Conference Publication | IEEE Xplore." Accessed: Apr. 09, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8524276.

[17] X. Mei, et al., RadImageNet: An open radiologic deep learning research dataset for effective transfer learning, *Radiol. Artif. Intell* 4 (5) (2022), https://doi.org/10.1148/ryai.210315 e210315.

[18] X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. WOO, Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2015. Accessed: Oct. 08, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html.

[19] A. Hanson, K. PNVR, S. Krishnagopal, and L. Davis, Bidirectional Convolutional LSTM for the Detection of Violence in Videos, in: Computer Vision – ECCV 2018 Workshops, L. Leal-Taixé and S. Roth, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 280–295. doi: 10.1007/978-3-030-11012-3_24.

[20] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* 143 (1) (Apr. 1982) 29–36, https://doi.org/10.1148/radiology.143.1.7063747.

[21] 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 (7) (Jul. 2020) 1465–1471, https://doi.org/10.1002/jor.24617.

[22] J. M. Karnuta et al., Artificial Intelligence to Identify Arthroplasty Implants From Radiographs of the Hip, J. Arthroplasty, vol. 36, no. 7S, pp. S290-S294.e1, Jul. 2021, doi: 10.1016/j.arth.2020.11.015.

[23] A. Borjali, A. F. Chen, O. K. Muratoglu, M. A. Morid, and K. M. Varadarajan, "Detecting mechanical loosening of total hip replacement implant from plain radiograph using deep convolutional neural network," arXiv:1912.00943 [cs, eess], Dec. 2019, Accessed: Jul. 09, 2021. [Online]. Available: http://arxiv.org/abs/1912.00943.