

Supplementary Document

1. Model Training

An AI model combining Resnet18 and TabNet with multi-modal cross-attention fusion was trained to predict the year of TKR surgery within a 9-year timeframe. We used data split as: 70% for training, 10% for validation, and 20% for testing. Horizontal flipping and random crop were used for data augmentation. To improve model generalizability, random cropping of input image size to 300x300x160 was implemented for DESS MR scans. Adam optimizer was used with a learning rate and a weight decay of 10-4. The model with the best validation accuracy was selected as the best model. The second last layer of Resnet18 DL model, the output of global max pooling layer before fully connected one provided 512 features for each image modality.

2. Model prediction evaluation metrics

Accuracy and macro-AUC were used as estimation evaluation metrics. Accuracy was calculated as:

$$ACC = 100 \times \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (1)$$

where:

- ACC: the accuracy of the TKR time prediction model
- N_{correct} : the number of patients whose predicted TKR time falls within ± 1 year of the actual TKR time ($|y - \hat{y}| \leq 1$),
- N_{total} : the total number of patients in the study.

We compute the macro-AUC for a 10-class classification task, where each class represents one year to TKR (0–9 years). Since our model originally predicts 30 bins, each corresponding to 4-month intervals, we aggregate every 3 consecutive bins to obtain probabilities for 10 yearly bins before computing the macro-AUC using a One-vs-Rest (OvR) strategy. The output of our model $M \in \mathbb{R}^{B \times 30}$, where B is the batch size and 30 bins correspond to 4-month intervals. Since the model outputs log-probabilities, we apply the softmax function to obtain probabilities, $P = \exp(M)$, where P represents the probability distribution across 30 bins. To convert 30 bins (4-months each) into 10 bins (1-year each), we sum every 3 consecutive bins:

$$P_j^{(\text{year})} = \sum_{k=1}^3 P_{(3j+k)} \quad (2)$$

for $j = 0, 1, \dots, 9$. This gives us a new probability matrix, $P^{(\text{year})} \in \mathbb{R}^{B \times 10}$ where each column represents a 1-year probability. Let the true labels be y , where each ground truth

y_i (for the i^{th} sample) represents the true time to TKR in years. The labels are discrete values, $y \in \{0, 1, \dots, 9\}$ where each class corresponds to a yearly bin. The macro-AUC is computed using a One-vs-Rest (OvR) strategy, which involves computing AUC for each class k (treating it as a binary classification problem: Class k vs. all others) and averaging the AUC scores across all 10 classes. The macro-AUC is given by:

$$\text{Macro-AUC} = \frac{1}{10} \sum_{k=0}^9 \text{AUC}(P_k^{(\text{year})}, y_k) \quad (3)$$

where:

- $P_k^{(\text{year})}$ represents the predicted probability of class k ,
- y_k is the true label transformed into a binary format for the One-vs-Rest approach,
- AUC is the area under the receiver operating characteristic (ROC) curve.

3. Ablation study

To justify image encoder choice in our end-to-end trained multi-modal model, we evaluated ResNet18, ResNet34, ResNet50, and Med3D using MRI-only data. ResNet18 provided the best prediction accuracy for our DESS MRI data from the OAI dataset, as provided in Table 1.

Model	ACC (%)
ResNet18	57.9
ResNet34	53.1
ResNet50	53.3
Med3D	55.8

Table 1: Performance comparison of AI models in predicting the year of TKR.

We compared the performance of our end-to-end trained model with commonly used traditional machine learning (ML) models for TKR prediction. Specifically, we extracted features from the image encoder and concatenated them with the selected tabular data, then evaluated the performance of a random forest (RF) model, XGBoost, and a multi-layer perceptron (MLP) using the combined dataset. Table 2 demonstrate that the end-to-end trained model outperformed these traditional ML models, highlighting the advantage of joint feature extraction and optimization in a unified framework.

Model	ACC (%)	MAE
RF	59.0	1.56
XGBoost	52.9	1.69
MLP	52.2	1.83
Our Model	63.4	1.33

Table 2: Performance comparison of ML models and our proposed end-to-end trained multimodal model in predicting the year of TKR.