algorithms

MDPI

Article

Non-Imaging Differential Diagnosis of Lower Limb Osteoarthritis: An Interpretable Machine Learning Framework

Zhanel Baigarayeva [1,2,3], Assiya Boltaboyeva [1,2,3], Baglan Imanbek [2,*], Bibars Amangeldy [2], Nurdaulet Tasmurzayev [2], Kassymbek Ozhikenov [1,*], Assylbek Ozhiken [1], Zhadyra Alimbayeva [1] and Naoya Maeda-Nishino [4,5]

1 Institute of Automation and Information Technology, Satbayev University, Almaty 050013, Kazakhstan
2 Faculty of Information Technologies and Artificial Intelligence, Al Farabi Kazakh National University, Almaty 050040, Kazakhstan
3 LLP Kazakhstan R&D Solutions, Almaty 050056, Kazakhstan
4 Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto, CA 94305, USA
5 HAKUAI Medical Corporation, Osaka 573-1010, Japan
* Correspondence: imanbek.baglan@kaznu.kz (B.I.); k.ozhikenov@satbayev.university (K.O.)

Abstract

Background: Osteoarthritis (OA) is a prevalent chronic degenerative disorder, with coxarthrosis (hip OA) and gonarthrosis (knee OA) representing its most significant clinical manifestations. While diagnosis typically relies on imaging, such methods can be resource-intensive and insensitive to early disease trajectories. Objective: This study aims to achieve the differential diagnosis of coxarthrosis and gonarthrosis using solely routine preoperative clinical and laboratory data, benchmarking state-of-the-art machine learning algorithms. Methods: A retrospective analysis was conducted on 893 patients (617 with knee OA, 276 with hip OA) from a clinical hospital in Almaty, Kazakhstan. The study evaluated a diverse portfolio of models, including gradient boosting decision trees (LightGBM, XGBoost, CatBoost), deep learning architectures (RealMLP, TabDPT, TabM), and the pretrained tabular foundation model RealTabPFN v2.5. Results: The RealTabPFN v2.5 (Tuned) model achieved superior performance, recording a mean ROC–AUC of 0.9831, accuracy of 0.9485, and an F1-score of 0.9474. SHAP interpretability analysis identified heart rate (66.2%) and age (18.1%) as the dominant predictors driving the model's decision-making process. Conclusion: Pretrained tabular foundation models demonstrate exceptional capability in distinguishing OA subtypes using limited clinical datasets, outperforming traditional ensemble methods. This approach offers a practical, high-performance triage tool for primary clinical assessment in resource-constrained settings.

Keywords: osteoarthritis; coxarthrosis; gonarthrosis; machine learning; tabular foundation models; differential diagnosis; explainable AI; clinical decision support

Check for updates

Academic Editors: Irene Rechichi, Silvia Seoni and Sujit Kumar Sahu

Received: 22 December 2025
Revised: 14 January 2026
Accepted: 16 January 2026
Published: 20 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Osteoarthritis (OA) stands as the most prevalent form of arthritis globally, characterized by progressive degeneration of articular cartilage with associated bone and soft tissue changes [1]. The condition represents a chronic degenerative disorder affecting multiple joints, with particular prevalence in weight-bearing articulations. OA is fundamentally defined by the loss of articular cartilage accompanied by hypertrophic bone changes, osteophyte formation, and subchondral sclerosis [2]. The disease process involves

complex interactions between mechanical factors, inflammatory pathways, and molecular degradation mechanisms that collectively contribute to joint dysfunction and pain [3,4].

Coxarthrosis, specifically referring to osteoarthritis of the hip joint, and gonarthrosis, denoting OA of the knee joint, represent the most clinically significant manifestations of OA in terms of disability and quality of life impact [5]. Recent global burden estimates indicate that in 2021, there were approximately 466.3 million new OA cases worldwide, with an age-standardized incidence rate (ASIR) of around 535 per 100,000 population. The prevalence of OA reached about 606.9 million cases, with disability-adjusted life years (DALYs) rising to approximately 213 million [6]. The burden of OA is significantly higher in women compared to men, with knee osteoarthritis (KOA) remaining the most common form of OA globally [7].

International clinical criteria for knee and hip OA typically integrate age, pain characteristics, stiffness, functional limitation, and physical examination findings; imaging is recommended mainly to support the diagnosis and rule out alternative or co-existing pathologies. Recent evidence emphasizes that a purely radiographic definition of OA is inadequate for early disease detection and for understanding diverse disease trajectories, especially in coxarthrosis and gonarthrosis where clinical, biomechanical, and metabolic drivers differ between joints [8–11]. It is insensitive to early cartilage damage, cannot visualize key soft-tissue structures, and correlates poorly with symptoms [9,12]. Technical factors also introduce variability, and the K/L system tends to overemphasize osteophytes while missing early joint space loss. As a result, radiography is considered suboptimal for early diagnosis and precise phenotyping [13]. Given the limitations of radiograph, Magnetic resonance imaging (MRI) offers a more advanced option for assessing KOA and Hip osteoarthritis (HOA) [14,15]. However, its routine use is limited by high cost, lower availability, long scan times, and contraindications [16,17].

Non-imaging methods, such as biochemical biomarkers in blood, urine, and synovial fluid, offer a complementary approach to osteoarthritis assessment. These markers capture cartilage turnover, bone metabolism, synovial inflammation, and matrix remodeling. While systemic markers may be influenced by extra-articular factors, synovial fluid biomarkers more directly reflect intra-articular processes [18,19]. Recent work has begun to integrate these multi-analyte profiles with machine learning methods to derive diagnostic or staging models of KOA and related conditions [19].

Machine learning models utilizing non-imaging data have proven to be powerful and interpretable tools for both the diagnosis of Knee Osteoarthritis (KOA) and the prediction of related conditions. A diagnostic framework for KOA, employing an XGBoost model combined with the Boruta feature selection method, demonstrated robust performance, achieving an Area Under the Curve (AUC) of 0.758, G-means of 0.800, and an F1 score of 0.703 on an unseen test set [20]. Separately, for predicting the risk of sarcopenia in elderly Chinese KOA patients, the CatBoost algorithm was selected as the optimal model after achieving strong results on an independent test set: an accuracy of 0.8902, F1 score of 0.8627, and an AUC of 0.9697, alongside a low Brier score of 0.0691 that indicates dependable probability estimates [21].

Recent KOA-focused studies have shown that non-imaging clinical and laboratory data can accurately support diagnosis and risk prediction. In one study of 2594 patients, ML models (Logistic Regression (LR), Random Forest (RF), eXtreme Gradient Boosting (xGBoost), Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT)) using five key variables achieved validation AUCs over 0.82–0.96, outperforming single markers [22]. Another study applied deep learning (convolutional neural network (CNN) and artificial neural network (ANN)) to multi-year EMR data (diagnoses, prescriptions,

age, sex) and predicted one-year KOA risk with an AUROC of 0.97, without any imaging or questionnaire data [23].

A large UK Biobank study used interpretable ML models combining clinical data, routine biomarkers, questionnaires, and multi-omics from about 19,000 OA cases and matched controls to predict 5-year OA risk. For weight-bearing joints like the knee and hip, multi-modal models achieved AUCs around 0.73, outperforming simple models based on age, sex, and BMI [10]. In another study, two peripheral blood genes (BIRC2 and SEH1L) were identified via combined bioinformatics and MLs (least absolute shrinkage and selection operator (LASSO) logistic regression and support vector machine-recursive feature elimination (SVM-RFE)) as candidate diagnostic markers for KOA, and functional experiments in human chondrocytes supported their biological plausibility. These models explicitly integrate high-dimensional laboratory and bioinformatic data with statistical learning algorithms to derive interpretable diagnostic signatures with an AUC of 0.814–0.964 [24].

The scientific contribution of this study lies in its methodological approach, specifically the benchmarking of various state-of-the-art machine learning algorithms for tabular data prediction, referred to as "TabArena" methods [25]. While the clinical utility lies in achieving the differential diagnosis of coxarthrosis (Hip OA) and gonarthrosis (Knee OA) using solely routine preoperative clinical and laboratory data, the scientific contribution stems from the evaluation of three complementary families of high-performance models.

This systematic comparison included highly competitive gradient boosting decision tree models, such as LightGBM (tuned), XGBoost (tuned + ensembled), and CatBoost (tuned + ensembled), alongside a selection of recent deep learning architectures for tabular data. These deep models included the Transformer-based TabDPT (tuned + ensembled), the parameter-efficient TabM (tuned + ensembled), the deep learning baseline RealMLP (tuned + ensembled), and the pre-trained tabular foundation model RealTabPFN-v2.5, investigated at three complexity levels: (tuned + ensembled), (tuned), and (default). The study also benchmarked the Metric-Learning Pipeline ModernNCA (tuned + ensembled) and two AutoML baselines: AutoGluon 1.4 (best, 4 h) and AutoGluon 1.4 (extreme, 4 h). By subjecting this diverse portfolio of models to a unified evaluation framework without data leakage, this study establishes a high-performance benchmark demonstrating the capability of modern ML to perform complex differential diagnosis based on routine clinical data.

This approach is not intended to replace imaging-based gold standards but rather to serve as a practical triage tool during primary clinical assessment. It is designed for patients with non-specific lower limb pain, assisting clinicians in deciding whether to primarily order a hip or knee X-ray. Importantly, the model is not intended for primary diagnostic screening (distinguishing OA from healthy joints or other conditions), but is specifically designed to aid in differentiating between the two types of OA in patients who already present with symptoms and clinical suspicions.

2. Materials and Methods

2.1. Data Collection

The study population was recruited from the endoprosthetics department of the City Clinical Hospital of Almaty, Kazakhstan. Adhering to the ethical principles of the Declaration of Helsinki [26], the study was conducted in full compliance with ethical standards and received approval from the Bioethics Committee of the National Center for Scientific and Technical Orthopedics (NCSTO) named after Academician N. D. Batpenov (Protocol No. 1/4). In accordance with local legislation and the approved protocol, the requirement for informed consent was waived as the analysis utilized a strictly retrospective and completely anonymized dataset. Prior to dissemination, data were processed by

NCSTO; direct identifiers were removed, and indirect identifiers were generalized or pseudonymized following ISO 25237:2017 standards to ensure patient confidentiality on secure institutional servers [27].

Initially, clinical and laboratory indicators were collected from 2043 patients. Eligible participants were ≥50 years of age with clinical and radiologic proof of coxarthrosis or gonarthrosis (Kellgren–Lawrence grade ≥ II) [28], and normal C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR) levels [29]. These parameters ensured a consistent osteoarthritis cluster and excluded borderline or preclinical cases.

Exclusion criteria encompassed high levels of inflammatory markers (CRP, ESR, RF, ACPA, or uric acid), diagnosed inflammatory arthritis, aseptic necrosis, tumors, infections, recent joint injuries, and discomfort originating primarily from the spine. Patients with severe comorbidities, cognitive impairment, or incomplete records were also omitted to rule out confounding factors such as autoimmune diseases or metabolic disorders.

Following the application of these strict inclusion and exclusion criteria to the initial 2043 records, a final group of 893 patients was retained for analysis. The collected information focused exclusively on the preoperative phase and comprised demographic details (age), clinical indicators, hematological parameters, and biochemical markers. This approach facilitates rapid diagnosis and predictive modeling [30]. To ensure the homogeneity of the OA cohort and minimize confounding factors, strict inclusion and exclusion criteria were applied to the initial pool of 1792 patients. Inclusion criteria were: age 50 years or older, clinically and radiologically confirmed coxarthrosis or gonarthrosis (Kellgren–Lawrence grade ≥ II), and normal levels of C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR). The restriction to age ≥ 50 ensured the exclusion of juvenile or early-adult secondary OA, validating that the observed younger mean age in the hip cohort (relative to the knee cohort) reflects the natural disease trajectory of primary osteoarthritis rather than outlier inclusion. In the final cohort, the average age was 62.38 years, and the average resting heart rate was 76.67 ± 2.82 beats/min. Table 1 displays the characteristics of the subjects.

Table 1. Baseline demographic, clinical, and laboratory characteristics of the study population.

Variable	Overall (*n* = 893)	Knee OA (*n* = 617)	Hip OA (*n* = 276)
Age (years)	62.6 ± 11.6	65.9 ± 7.7	55.4 ± 15.1
Gender (Female)	660 (73.9%)	490 (79.4%)	170 (61.6%)
Gender (Male)	233 (26.1%)	127 (20.6%)	106 (38.4%)
BMI (kg/m^2)	26.3 ± 2.4	26.5 ± 2.5	26.1 ± 2.1
Heart Rate (bpm)	76.9 ± 2.7	78.0 ± 1.8	74.4 ± 2.8
Hemoglobin (g/L)	133.7 ± 18.8	133.2 ± 18.1	134.9 ± 20.4
Leukocytes (10^9/L)	6.7 ± 1.8	6.6 ± 1.8	6.9 ± 1.9
Platelets (10^9/L)	263.4 ± 67.4	264.4 ± 67.9	261.0 ± 66.4
ESR (mm/h)	14.1 ± 11.2	14.2 ± 11.4	13.8 ± 10.7
Creatinine (μmol/L)	72.9 ± 30.3	71.1 ± 19.9	76.8 ± 45.4
Glucose (mmol/L)	5.9 ± 1.5	6.1 ± 1.5	5.6 ± 1.3
Length of Stay (days)	8.8 ± 2.3	8.8 ± 2.4	8.7 ± 1.8

2.2. Data Preprocessing

The raw dataset exhibited substantial non-uniform missingness across several clinically essential variables, including weight (23.1%), AST (15.6%), ALT (15.1%), ESR (14.6%), and platelet count (11.2%). Given the physiological interdependence of these markers and the moderate dataset size, multiple imputation was not applied to avoid introducing artificial correlations and biased signal structure. Instead, a complete-case analysis was adopted, reducing the cohort from 2043 to 893 patients (~56%). This conservative strategy

prioritized data integrity, interpretability, and model stability across heterogeneous machine learning frameworks.

The dataset quality was perceived as a key determinant to later machine learning performance and interpretability [31]. Since the original dataset was from clinical practice, this data had missing values, variable scales which would fluctuate to different scales, and class imbalance. For this reason, a comprehensive preprocessing procedure was performed before the model was formulated.

Certain variables and clinical results including blood work and serum tests was a good example which had missing data by incomplete entry of data from EMR. So as not to introduce bias based on the assumptions that results should be entered and to prevent having an analysis of case which would be too complete and thereby improve the performance of the model, a complete case analysis was chosen as a method. All samples with at least one missing candidate predictor variable were eliminated. After the preliminary analysis of numeric variables, we considered the integrity of the numbers in the dataset and whether the values were within reasonable limits. The reduction in cases after implementing quality control measures was from 2043 cases to 893 cases. This improvement guaranteed that only complete and clinically reasonable cases remained in place for modeling.

As most of the data were numerical in character, we did not have a wide range of categories of interest. Sex, which was the only predictor variable that was categorical, was encoded into a binary number (0 = female, 1 = male), in order to keep gender non-ordinal and still enter fully into all machine-learning models. The variable to be predicted, diagnosis, was binary, thus was not encoded either (0 = gonarthrosis, 1 = coxarthrosis). To ensure that there is no information leakage, the preoperative risk score was not added to the feature set, as it is the integrated value of clinical features that share clinical attributes with the predictors that a model can be trained using.

Feature engineering was specifically done to amplify prediction power under the clinically interpretable conditions. BMI was calculated from the height and weight as previously proposed [32]:

$$BMI = \frac{Weight\ (kg)}{Height\ (m)2} \tag{1}$$

BMI was included as an additional predictor together with the value of the height and weight parameters. This process helped models model two forms of risk, adiposity-related risk through BMI and anthropometric-driven effects derived from biomechanical parameters. No additional features were derived after all to avoid complication and to fit common clinical measures. Following some preprocessing, exclusion of the risk variable and BMI derivation, the outcome dataset eventually included 20 numerical predictors which include demographic variables, vital signs, hematological indices, renal and hepatic function markers, glucose metabolism and hospitalization-associated markers with one binary diagnostic outcome variable.

Feature scaling [33] was used due to the heterogeneous units of measurement in clinical variables (e.g., age (years), hemoglobin (g/L), glucose (mmol/L), and platelet count ($\times 10^9$/L). Without scaling variables with much larger numeric ranges could disproportionately impact model training. This allowed for Z-score normalization to be used on all numerical predictors [34], defined as,

$$Z = \frac{x - \mu}{\sigma} \tag{2}$$

where X represents the observed value, μ the feature mean, and σ the standard deviation. Z-score normalization was preferred over min-max scaling because it preserves distribu-

tional properties and is less sensitive to extreme values, which are common in clinical data [34]. To prevent data leakage, scaling parameters were fitted exclusively on training data within each cross-validation fold and subsequently applied to the corresponding validation data [35]. This fold-specific scaling strategy preserved evaluation integrity and ensured reproducibility. This approach was explicitly adopted to prevent information leakage and to ensure that the reported ROC–AUC values reflect genuine predictive signal rather than preprocessing artifacts.

2.3. Machine Learning Algorithms for Classification

Dataset quality was considered a key factor determining the performance and interpretability of subsequent machine learning [30]. The final dataset consisted of complete clinical and laboratory records for 893 patients. Due to strict collection criteria, the dataset was characterized by high integrity, and preliminary analysis confirmed that all numerical values fell within valid physiological ranges, allowing all 893 cases to be used for modeling without the need to exclude any records.

The prediction task was described as a supervised binary classification problem that served to distinguish between coxarthrosis and gonarthrosis based on clinical and laboratory characteristics. Coxarthrosis was referred to as the positive class in this study, compared with gonarthrosis, the negative class. This formulation reflects a clinically relevant diagnostic setting where decisions must be made before surgical procedures using routinely available clinical data. Faced with this problem, a rich set of machine learning algorithms was purposely chosen to account for complementary inductive biases and modeling assumptions. To achieve extensive methodological coverage, a diverse array of classical ensemble-based methods, deep learning architectures, metric-learning approaches, and automated machine learning frameworks were integrated. Such a strategy allowed for comparisons of models with different levels of complexity, interpretability, and ability to accommodate nonlinear feature interactions that allowed for a robust assessment of the performance of algorithms on structured clinical data.

Gradient boosting decision tree models, including LightGBM, XGBoost, and CatBoost, were selected due to their strong performance on tabular datasets and their ability to model non-linear relationships and feature interactions [36]. These models construct additive ensembles of weak learners, iteratively minimizing a differentiable loss function:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad (3)$$

where $h_m(x)$ denotes the weak learner at iteration m and is the learning rate [37]. Hyperparameters controlling tree depth, number of leaves, learning rate, subsampling, and L1/L2 regularization were optimized using grid search combined with stratified five-fold cross-validation. Model selection was based on maximizing mean cross-validated ROC–AUC.

At this stage, we tested two configurations for LightGBM in order to investigate the impact of hyperparameter optimization and ensembling on predictive performance. The hyperparameters determining the model capacity and regularization in the tuned setting; number of boosting iterations, learning rate, maximum tree depth, number of leaves, subsampling ratios, and L1/L2 regularization terms were optimized via grid search with stratified cross-validation, using ROC–AUC as the model selection parameter. A tuned + ensemble configuration was implemented by averaging class-probability predictions from multiple LightGBM models trained with identical optimized hyperparameters but different random seeds, with performance evaluated via stratified cross-validation.

Extending from our previous approach, XGBoost was further evaluated with tuned + ensemble configurations under stratified five-fold cross-validation. Numerical features

were standardized and class imbalance was mitigated with a blend of class-weighted learning and SMOTE used only on the training folds. Hyperparameters were tuned using grid search with ROC–AUC as the selection metric. To give greater robustness, predictions from different models trained on different random seeds were averaged together for final outputs.

Likewise, CatBoost was also evaluated through tuned and tuned + ensemble settings. With the tuned setup, important hyperparameters including number of boosting iterations, tree depth, learning rate, and L2 regularization strength were optimized via grid search with stratified cross-validation. The ordered boosting strategy of CatBoost was kept in order to prevent target leakage during tree construction. In the tuned + ensemble condition, the optimized CatBoost model was mounted as part of a bagging implementation, which means that when you combine several models trained on different resampled subsets you have many converging predictions. This ensemble approach improved both robustness and accuracy, especially in the case that the model was under class imbalance.

In the gradient boosting decision tree models, class imbalance was consistently dealt with applying a class-weighted learning scheme together with SMOTE in training folds, and numerical features were standardized in order to ensure stable and comparable optimization performance. These arrangements together allowed for a regular comparison of optimized single models and their ensemble-based versions relative to predictive stability and generalization accuracy.

Following the evaluation of gradient boosting decision tree models, deep learning–based approaches were incorporated to assess the effectiveness of neural architectures on structured clinical data. RealMLP, a regularized multilayer perceptron tailored for tabular data [13–38], was evaluated under tuned and tuned + ensembled configurations. In the tuned setting, model capacity and optimization dynamics were deliberately constrained to mitigate overfitting on the limited dataset, employing shallow depth, low hidden dimensionality, batch normalization, high dropout, and L2 regularization via AdamW. Class imbalance was handled using class-weighted cross-entropy loss in combination with SMOTE applied exclusively to training folds, while early stopping based on validation loss controlled training duration. Hyperparameters were selected based on prior literature and empirical best practices for small-sample tabular learning rather than exhaustive grid search. In the tuned + ensembled configuration, robustness was further improved through deep ensembling, where multiple independently initialized RealMLP models were trained per cross-validation fold and their class-probability outputs averaged, reducing estimator variance and improving predictive stability and calibration.

After learning gradient boosting decision tree approaches, deep learning-based methods were used to evaluate the effectiveness of neural architectures on structured clinical data. RealMLP, a regularized multilayer perceptron for tabular data [38], was studied while utilizing tuned and tuned + ensembled configurations. Model capacity and optimization dynamics were purposefully limited in the tuned environment to suppress overfitting on small dataset by using shallow depth, low hidden dimensionality, batch normalization, high dropout and L2 regularization with AdamW. Class imbalance was dealt with by using class-weighted cross-entropy loss in combination with SMOTE applied exclusively to training folds, with early stopping based on validation loss for limiting training duration. Hyperparameters were selected based on past literature and best practice for small-scale tabular learning, rather than exhaustive grid search. In the tuned + ensembled setup, robustness was also achieved through deep ensembling, where several independently initialized RealMLP models were trained for each cross-validation fold and their class-probability outputs were averaged to reduce estimator variation, stabilizing prediction and improving the calibration.

Complementing the direct discriminative modeling strategy of RealMLP, ModernNCA was implemented as a metric learning approach rather than a conventional classifier, learning a low-dimensional embedding space optimized for class separation via a distance-based probabilistic objective [39]. Input features were standardized, and SMOTE was applied within training folds. ModernNCA was evaluated exclusively in a tuned + ensemble configuration, employing conservative architectural constraints and early stopping based on validation ROC–AUC as the primary mechanism for controlling overfitting in the learned embedding space; no exhaustive grid search was performed. All reported metrics are expressed as mean ± standard deviation across stratified cross-validation folds, with ensemble averaging applied at the prediction level within each fold. Together with RealMLP, these neural approaches provided complementary views of representation learning, contrasting direct discriminative modeling with distance-based embedding optimization.

To complement RealMLP, ModernNCA is performed as a metric learning, as opposed to a standard classifier, which learns a low-dimensional embedding space optimized for the separation of the classes through distance-based probabilistic objective [39]. Input features were standardized and SMOTE was implemented in training folds. ModernNCA was evaluated using only tuned + ensembled architecture and conservative architectural constraints, where the early stopping was controlled based on validation ROC–AUC as the key procedure to control overfitting of learned embedding space, but there was no exhaustive grid search. All reported metrics are presented as mean ± standard deviation with stratified cross-validation folds, and ensemble averaging is provided at the prediction level per fold. These neural approaches offered complementary perspectives about representation learning with RealMLP and offered a different perspective on direct discriminative modeling and distance-based embedding optimization.

Extending the analysis to attention-based deep learning paradigms, transformer-based tabular models, namely TabDPT and TabM, were employed to evaluate whether attention-based feature interaction modeling can improve discrimination in structured preoperative clinical data, where complex and non-linear dependencies between variables are expected [40,41]. Both architectures are built upon self-attention mechanisms of the form

$$Attention\ (Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where

1. Q (Query matrix) represents the features that are requesting information. Each row corresponds to a query vector;
2. K (Key matrix) represents the features against which the queries are;
3. V (Value matrix) contains the information that will be aggregated and passed forward;
4. QK^T computes the similarity (dot product) between each query and all keys;
5. d_k is the dimensionality of the key vectors; dividing by $\sqrt{d_k}$ stabilizes gradients and prevents excessively large attention scores;
6. $softmax(\cdot)$ converts the similarity scores into normalized attention weights that sum to one;
7. Multiplication V produces a weighted sum of the value vectors, yielding the final attention output.

In this mechanism, each input element dynamically attends to other elements by assigning higher weights to more relevant features. This enables transformer models to capture complex dependencies and interactions within tabular or sequential data without relying on a fixed feature ordering, allowing adaptive weighting of feature relationships and flexible modeling of inter-feature dependencies [42].

In the tuned TabDPT setup, numerical features were embedded individually and processed by a shallow transformer encoder with limited capacity. Overfitting was controlled using elevated dropout, AdamW regularization, and a reduced learning rate. Stratified five-fold cross-validation was employed, with class imbalance handled via SMOTE within training folds and a class-weighted loss. Early stopping based on validation loss was applied, and architectural parameters were fixed without exhaustive hyperparameter search.

The tuned + ensembled TabDPT configuration retained the same architectural and regularization settings while enhancing robustness through internal ensembling. A single transformer backbone produced multiple parallel prediction heads, each acting as an independent voter. During training, logits from all ensemble members were averaged prior to loss computation to encourage consensus, while inference relied on probability averaging across ensemble members to reduce variance without increasing model depth or training complexity.

For TabM, a related transformer-based architecture emphasizing internal ensembling, both tuned and tuned + ensemble configurations were evaluated. In the tuned TabM setting, numerical features were embedded into a lower-dimensional latent space and processed through a moderately deep transformer encoder (three layers) with reduced width, increased dropout (0.2), and L2 regularization. Mean pooling across token embeddings was used instead of a classification token, and early stopping controlled training duration. In the tuned + ensembled TabM configuration, robustness was further improved by producing multiple parallel class logits within the model head (ensemble size up to 32). Logits were averaged during training to stabilize optimization, and probability averaging was applied at inference to reduce estimator variance. Across all transformer-based configurations, feature standardization, SMOTE-based resampling, class-weighted loss functions, and fold-specific preprocessing were applied consistently within stratified cross-validation to ensure leakage-free and comparable evaluation.

AutoGluon v1.4 was employed as an automated machine learning benchmark using the best_quality and extreme presets [43]. These presets provide strong out-of-the-box performance by internally handling model selection, hyperparameter optimization, stacking, and ensembling across a diverse set of base learners [44]. Both configurations were evaluated using stratified five-fold cross-validation, with a fixed computational time budget of four hours. All hyperparameter optimization and ensemble construction were performed entirely within the AutoGluon framework, with no manual grid search or user-defined tuning, allowing AutoGluon to serve as a robust automated reference reflecting performance achievable without manual model engineering.

In addition, pretrained tabular foundation models were evaluated to assess inference-based generalization on a relatively small clinical dataset. Specifically, RealTabPFN v2.5 was included due to its design philosophy of minimizing task-specific training while leveraging priors learned from both synthetic tabular tasks and continued pre-training on curated real-world datasets [45]. Three configurations were examined. First, RealTabPFN (Default v2.5) was applied using its standard inference setting without any task-specific hyperparameter modification. Second, RealTabPFN (Tuned v2.5) employed an adjusted inference configuration by increasing the number of internal estimators to improve predictive stability and reduce variance at inference time. Third, RealTabPFN (Tuned + Ensembled v2.5) combined the tuned inference configuration with an external bagging ensemble, in which multiple TabPFN predictors were trained on bootstrap resamples within each cross-validation fold and their class-probability outputs averaged. For all TabPFN-based models, no traditional grid search or gradient-based fine-tuning was performed, consistent with their intended use as pretrained inference models rather than conventionally trained classifiers [46].

Overall, multiple machine learning models representing different learning paradigms were systematically evaluated under consistent and leakage-free experimental conditions. The following section presents and compares their predictive performance.

To comprehensively assess model performance and ensure clinical relevance, multiple complementary evaluation metrics were employed. The use of several metrics was necessary because no single measure fully captures predictive quality, particularly in the presence of class imbalance and asymmetric clinical risks.

Accuracy measures the overall proportion of correctly classified instances and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy provides an intuitive summary of overall correctness; however, in medical classification tasks, it may be misleading when class distributions are imbalanced, as high accuracy can be achieved by favoring the majority class.

Precision quantifies the reliability of positive predictions and is expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

This metric reflects the proportion of patients predicted as having the target condition who are correctly identified. High precision is particularly important in clinical settings where false-positive predictions may lead to unnecessary interventions, diagnostic procedures, or patient anxiety.

Recall, also known as sensitivity, measures the ability of a model to correctly identify positive cases:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

In a medical context, recall is crucial because false negatives correspond to missed diagnoses. A model with high recall reduces the likelihood of failing to identify patients who truly have the condition, which is especially important when delayed treatment may worsen outcomes.

The $F1$-score represents the harmonic mean of precision and recall:

$$F1 = P\,\frac{2TP}{2TP + FP + FN} \tag{8}$$

By balancing both false positives and false negatives, the F1-score provides a single metric that is particularly informative when class distributions are uneven. It is well-suited for evaluating clinical classifiers where both types of errors carry significant consequences.

ROC–AUC evaluates a model's ability to discriminate between classes across all possible classification thresholds. It is defined as the area under the ROC curve, which plots the true positive rate (recall) against the false positive rate:

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

An AUC value of 0.5 indicates random performance, whereas a value of 1.0 indicates perfect discrimination.

ROC–AUC was selected as the primary evaluation metric because it is threshold-independent and robust to class imbalance, making it particularly appropriate for comparing different models in a clinical prediction setting.

3. Results

Across all evaluated models, discriminative performance was generally strong, with most achieving high ROC–AUC values under stratified five-fold cross-validation. To fully capture model efficacy in a clinical diagnostic scenario, we utilized a comprehensive suite of metrics, prioritizing the tradeoff between false-positive and false-negative errors alongside overall accuracy. Among all assessed methods, RealTabPFN v2.5 (Tuned) achieved the superior overall performance, recording the best mean ROC–AUC (0.9831 ± 0.0053) as well as consistently high accuracy (0.9485 ± 0.0173), precision (0.9503 ± 0.0161), recall (0.9485 ± 0.0173), and F1-scores (0.9474 ± 0.0182).

While conventional tree-based ensemble models—such as LightGBM, XGBoost, and CatBoost—exhibited strong predictive capabilities, RealTabPFN distinguished itself through superior stability and robustness across folds. This advantage likely stems from its nature as a pretrained tabular foundation model, which utilizes rich prior structures and a probabilistic framework to handle class imbalance and generalization better than standard gradient boosting or neural networks. Although ensemble-based versions of other models helped decrease variance, they offered diminishing returns compared to the tuned RealTabPFN. Consequently, RealTabPFN v2.5 (Tuned) was selected as the main predictive model for further analysis. To ensure clinical interpretability and trust, post hoc explanations were applied to this final model, providing clear insights into the factors driving individual predictions in the decision-making process.

Performance of all models is given in Table 2, reported as mean ± standard deviation across stratified five-fold cross-validation.

Table 2. Performance comparison of evaluated models.

Model	Accuracy	Precision	Recall	F1-Score	ROC–AUC
RealTabPFN v2.5 (Tuned)	0.9485 ± 0.0173	0.9503 ± 0.0161	0.9485 ± 0.0173	0.9474 ± 0.0182	0.9831 ± 0.0053
RealTabPFN v2.5 (Default)	0.9474 ± 0.0160	0.9715 ± 0.0170	0.8554 ± 0.0512	0.9090 ± 0.0300	0.9831 ± 0.0045
AutoGluon 1.4 (Extreme, 4 h)	0.9429 ± 0.0186	0.9589 ± 0.0159	0.8518 ± 0.0486	0.9018 ± 0.0338	0.9795 ± 0.0032
AutoGluon 1.4 (Best, 4 h)	0.9451 ± 0.0144	0.9520 ± 0.0117	0.8662 ± 0.0384	0.9068 ± 0.0255	0.9793 ± 0.0049
RealTabPFN v2.5 (Tuned + Ensemble)	0.9463 ± 0.0166	0.9469 ± 0.0158	0.9463 ± 0.0166	0.9454 ± 0.0173	0.9791 ± 0.0119
LightGBM (Tuned)	0.9384 ± 0.0150	0.9051 ± 0.0096	0.8953 ± 0.0564	0.8991 ± 0.0280	0.9791 ± 0.0061
CatBoost (Tuned)	0.9373 ± 0.0184	0.9261 ± 0.0087	0.8663 ± 0.0579	0.8943 ± 0.0338	0.9790 ± 0.0064
XGBoost (Tuned + Ensemble)	0.9261 ± 0.0185	0.8755 ± 0.0226	0.8881 ± 0.0577	0.8807 ± 0.0322	0.9730 ± 0.0070
LightGBM (Tuned + Ensemble)	0.9362 ± 0.0228	0.9194 ± 0.0176	0.8701 ± 0.0703	0.8927 ± 0.0425	0.9781 ± 0.0055
CatBoost (Tuned + Ensemble)	0.9306 ± 0.0168	0.9178 ± 0.0088	0.8518 ± 0.0518	0.8828 ± 0.0307	0.9773 ± 0.0075
TabDPT (Tuned)	0.9026 ± 0.0282	0.8057 ± 0.0486	0.9060 ± 0.0460	0.8522 ± 0.0403	0.9651 ± 0.0120
TabDPT (Tuned + Ensemble)	0.9160 ± 0.0150	0.8291 ± 0.0402	0.9204 ± 0.0201	0.8717 ± 0.0198	0.9611 ± 0.0102
TabM (Tuned)	0.9127 ± 0.0215	0.8444 ± 0.0326	0.8806 ± 0.0514	0.8616 ± 0.0339	0.9613 ± 0.0105
TabM (Tuned + Ensemble)	0.9070 ± 0.0229	0.8149 ± 0.0331	0.9061 ± 0.0554	0.8575 ± 0.0361	0.9610 ± 0.0131
ModernNCA (Tuned + Ensemble)	0.9071 ± 0.0179	0.8443 ± 0.0333	0.8588 ± 0.0341	0.8511 ± 0.0270	0.9408 ± 0.0152
RealMLP (Tuned + Ensemble)	0.9048 ± 0.0157	0.8504 ± 0.0285	0.8406 ± 0.0345	0.8452 ± 0.0254	0.9334 ± 0.0213
RealMLP (Tuned)	0.9003 ± 0.0107	0.8311 ± 0.0236	0.8515 ± 0.0266	0.8408 ± 0.0165	0.9305 ± 0.0202

To further examine the robustness of the high ROC–AUC values, fold-wise performance metrics for the best-performing model, RealTabPFN v2.5 (Tuned), are reported in Table 3.

Table 3. Fold wise cross validation performance of RealTabPFN v2.5 Tuned.

Fold	Validation Accuracy	Validation ROC–AUC	Training Accuracy
1	0.9441	0.9905	0.9734
2	0.9609	0.9838	0.9692
3	0.9218	0.9792	0.9776
4	0.9663	0.9852	0.9790
5	0.9494	0.9769	0.9818
Mean ± SD	0.9485 ± 0.0173	0.9831 ± 0.0053	0.9762 ± 0.0048

As shown in these results, the model demonstrated high stability across all stratified cross validation folds, achieving a mean validation ROC AUC of 0.9831 with a low standard deviation of 0.0053. This consistency indicates that the model is capturing reproducible clinical patterns rather than noise or fold specific artifacts. Furthermore, the generalization gap defined as the difference between mean training accuracy of 0.9762 and validation accuracy of 0.9485 was limited to 0.0277. This narrow margin supports stable generalization behavior and contrasts with the high variance overfitting often suspected when ROC AUC values exceed 0.95 in clinical prediction tasks. The observed robustness is likely attributable to the architecture of RealTabPFN v2.5 which leverages Prior Data Fitted Networks to model complex non linear relationships while maintaining strong generalization on structured clinical data.

Figure 1 gives the aggregated receiver operating characteristic (ROC) curve for the best-performing model, RealTabPFN v2.5, obtained by pooling predictions across stratified five-fold cross-validation. The ROC curve shows the association of true positive rate (sensitivity) with false positive rate at every decision threshold in a threshold-independent manner, and therefore provides a measure of discriminative effect based on the ROC curve.

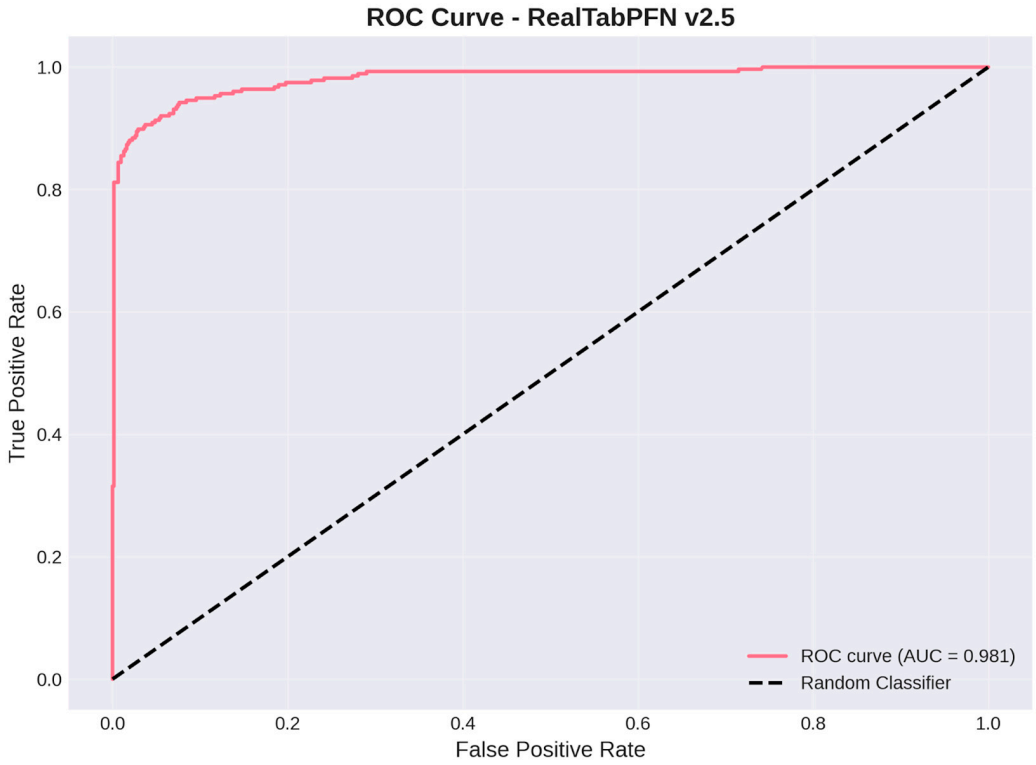


Figure 1. Illustration of Aggregated ROC Curve.

The model demonstrates strong class separability, with the ROC curve consistently lying well above the diagonal corresponding to random classification. The area under the ROC curve (AUC = 0.981) indicates excellent discrimination between coxarthrosis and gonarthrosis, reflecting model ability to correctly rank positive versus negative cases and the fact that it ranks the positive more highly than the negative for a wide range of thresholds. Significantly, the high true positives obtained at low false positives can imply robustness against class imbalance, which is supportive of reliable performance in clinically feasible operating areas.

Following the ROC analysis, Figure 2 presents the aggregated precision–recall (PR) curve for RealTabPFN v2.5, constructed by pooling predictions across stratified five-fold cross-validation. The PR curve characterizes the trade-off between precision and recall and is particularly informative in settings with class imbalance, where ROC curves alone may provide an overly optimistic assessment.

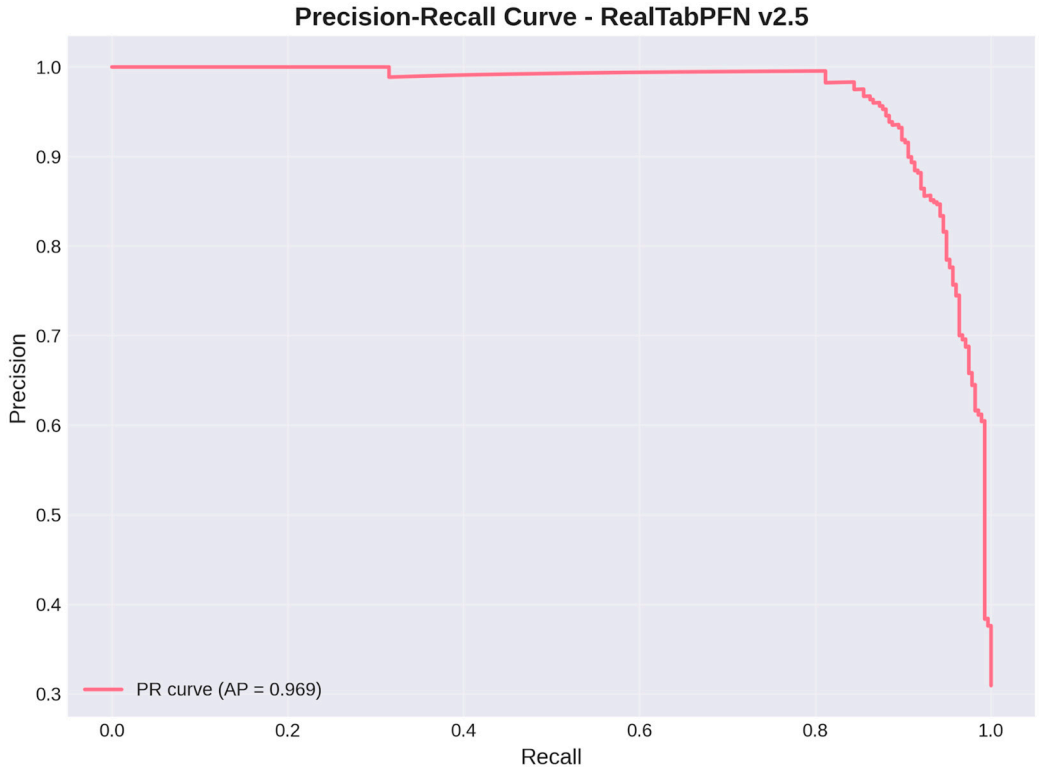


Figure 2. Illustration of Aggregated Precision–Recall (PR) Curve.

The model demonstrates exceptional robust performance with an Average Precision (AP) of 0.969, indicating a high success rate in identifying positive cases without a significant increase in false positives. Notably, the curve maintains precision near unity (1.0) until reaching a recall threshold of approximately 0.8, at which point a controlled trade-off begins. This stability in the high-precision/high-recall quadrant suggests that the model is particularly effective for risk-sensitive medical decision-making, where both reliability and sensitivity are critical.

The calibration curve for RealTabPFN v2.5 is shown in Figure 3, corresponding to the correlation among the model's average predicted probabilities and the observed fraction of positive cases, summed over stratified five-fold cross validation. The dashed diagonal denotes perfect calibration, wherein predicted risks correspond exactly to the empirical outcomes.

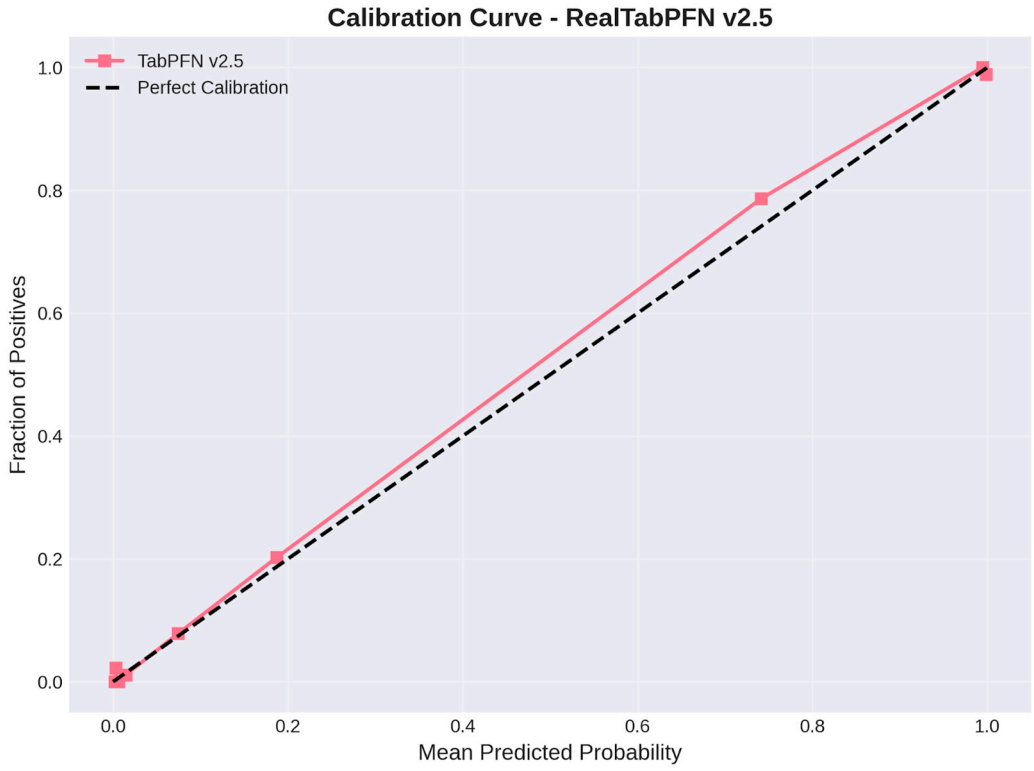


Figure 3. Calibration curve of the RealTabPFN v2.5 model.

The calibration curve of the model matches the reference line very closely in most probability bins. In fact, low-probability predictions (≈0.05–0.10) match very well with observed positive fractions that are near the same value which means it is correct to estimate low-risk cases as a given. Likewise with intermediate predictions, near 0.20–0.30, outcome rates observed are very near this predicted level, suggesting valid and reliable probabilistic estimates in this clinically significant decision domain.

At higher predicted probabilities (≈0.70–0.80), the model slightly overestimates risk, as the observed fraction of positives falls marginally below the predicted probability. However, this deviation is modest, and at the highest probability bin (≈1.0) the curve again closely matches the diagonal, indicating that predictions labeled as very high risk are strongly supported by observed outcomes. Overall, the calibration behavior demonstrates that RealTabPFN v2.5 provides well-aligned and interpretable probability estimates across the full risk spectrum, with only minor deviations that do not compromise its practical utility.

The model estimates risk at higher predicted probabilities (≈0.70–0.80) as the observed fraction of positives slightly is lower than the predicted probability. Nevertheless it is a small deviation, with the curve again corresponding to the diagonal in the highest probability bin (≈1.0); we can conclude that the very high risk predicted value is well confirmed by observed outcomes. More broadly, the calibration results reveal that the model offers clear and interpretable probability estimates of probabilities by the end of the risk spectrum and contains only slight deviations which do not threaten the performance of the model.

Following the calibration analysis, which demonstrated close agreement between predicted probabilities and observed outcome frequencies, Figure 4 provides a complementary view of model performance through the aggregated confusion matrix for TabPFN v2.5, where Class 0 represents gonarthrosis and Class 1 represents coxarthrosis.

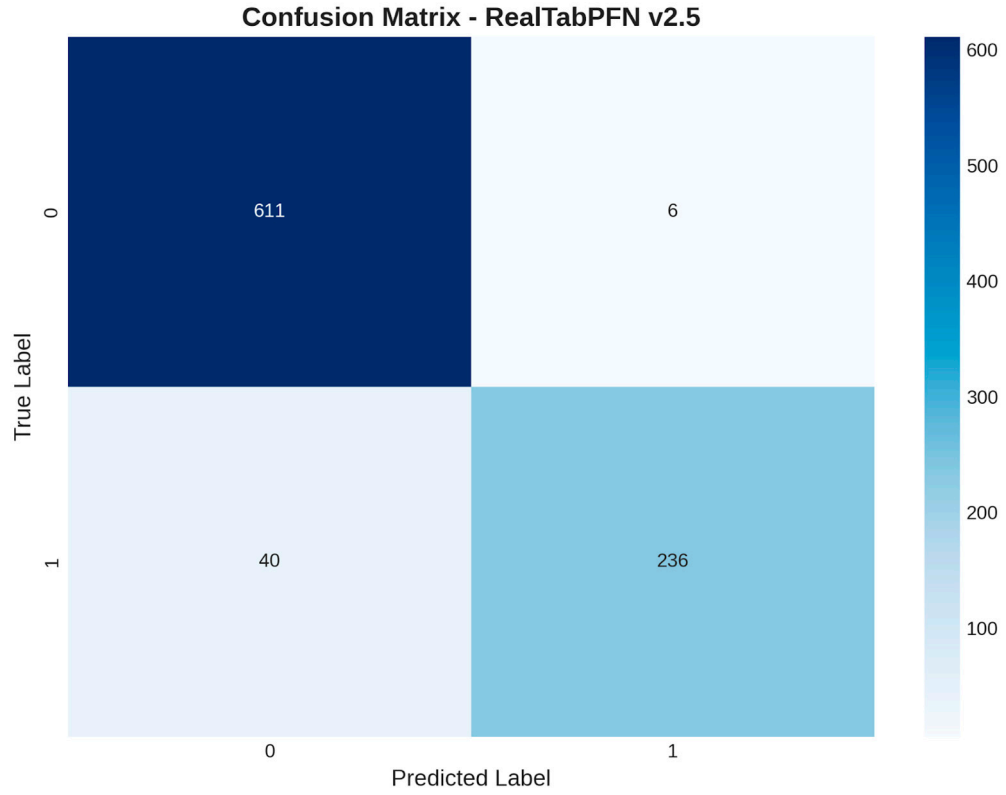


Figure 4. Illustration of Aggregated Confusion Matrix.

The model correctly classified 611 gonarthrosis cases as true negatives and 236 coxarthrosis cases as true positives, confirming strong overall predictive accuracy. Misclassifications were limited, comprising 6 false positives, in which gonarthrosis cases were incorrectly predicted as coxarthrosis, and 40 false negatives, where coxarthrosis cases were classified as gonarthrosis. Consistent with the precision–recall results, the error profile indicates a slightly higher incidence of false negatives than false positives, reflecting a conservative decision tendency while maintaining low rates of incorrect positive predictions.

Based on classification performance assessment from the confusion matrix, which represented the predictive accuracy and error distribution of the model, an interpretability analysis was carried out to determine what features guided these predictions. To this purpose, SHAP (SHapley Additive exPlanations) was used to offer a global, model-agnostic assessment of feature importance where it is combined with the performance metrics giving an all-around look into how the model was making these predictions.

Figure 5 shows the global feature importance ranking by SHAP value quantified by mean absolute SHAP value and represented in a relative percentage contribution. Heart rate (HR) is the major predictor, which explains about 66.2% of model significance; hence, model predictions are mainly guided by HR variations. Age is the second most important predictor at 18.1%, followed by hematocrit with a moderate (4.9%) significance. Combined, these three major features account for most of the model's predictive power.

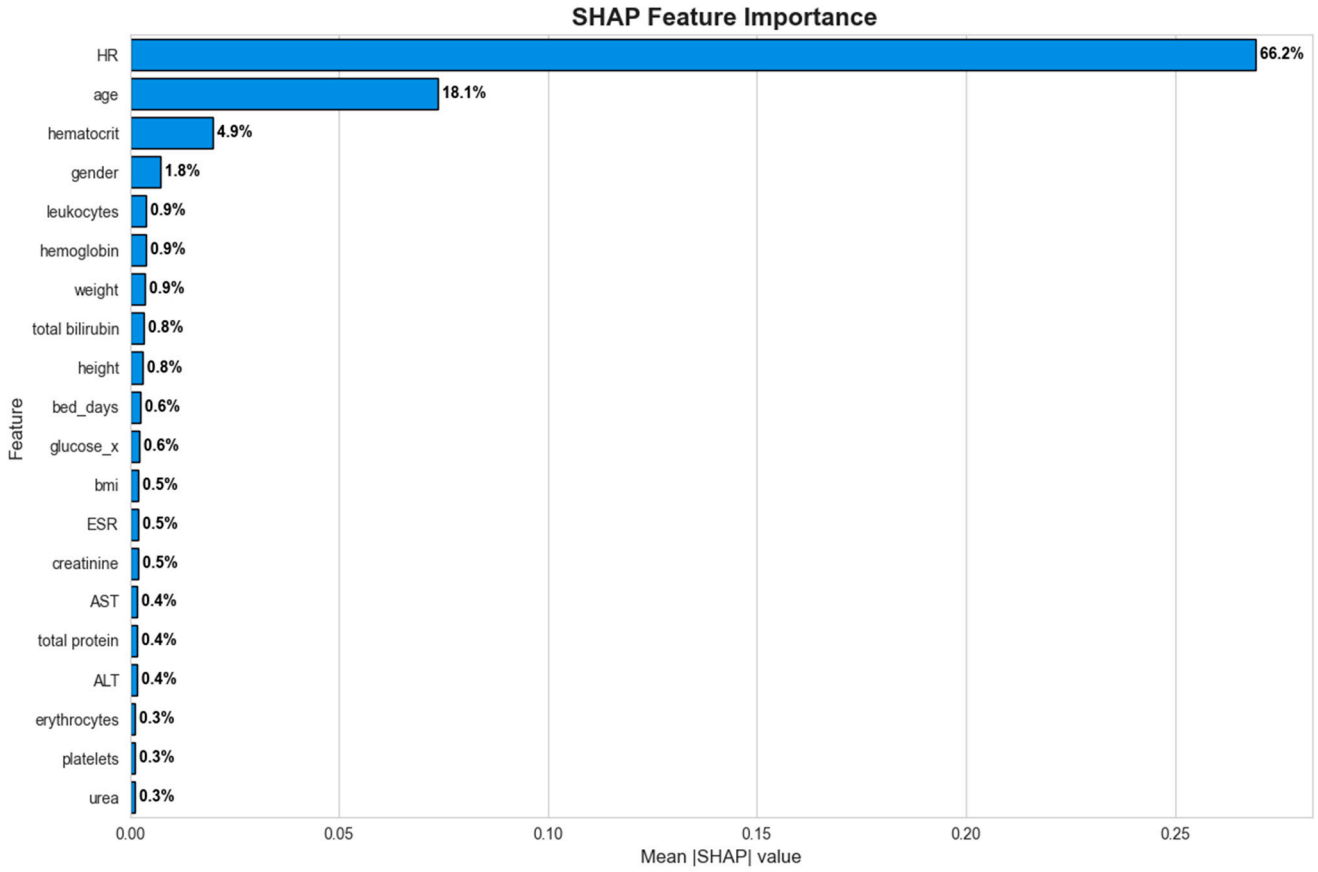


Figure 5. SHAP Feature importance of TabPFN v2.5 (tuned).

All other variables contribute only slightly, contributing less than 2% of the overall mean importance. Gender (1.8%), leukocytes, hemoglobin and weight (0.9%) show relatively modest, albeit reliable, influence, and biochemical and clinical variables, like bilirubin, glucose, BMI, ESR, creatinine, AST, ALT, urea, have limited impact. This steep drop in significance suggests the limited feature concentration is dominant, reiterating the interpretability of our model and pointing to the fact that only a small fraction of clinically appropriate variables can contain the majority of the prediction signal.

The global SHAP feature-importance ranking in Figure 6 does measure how much each single variable contributes to the model's prediction overall, but does not indicate which direction, or how large of an effect each feature has on the resulting prediction.

The global contribution of its features to the model's predictions is captured (Figure 6), which represents the SHAP summary plot. The most significant predictor was heart rate (HR), which is a variable whose higher values consistently lead to a higher predicted probability of the positive class and whose lower values act as protective factors. Age was ranked in second place overall, and with higher values of age positively shaping the model output. Hematocrit showed a relatively moderate and uniform contribution, while gender reported a smaller but systematically significant effect.

Other laboratory and clinical variables, such as leukocytes, hemoglobin, bilirubin, creatinine, and glucose, had rather slight effects. Anthropometric and biochemical measures including weight, height, BMI, AST, ALT, urea, platelets and erythrocytes were predominantly centered around zero reflecting low global.

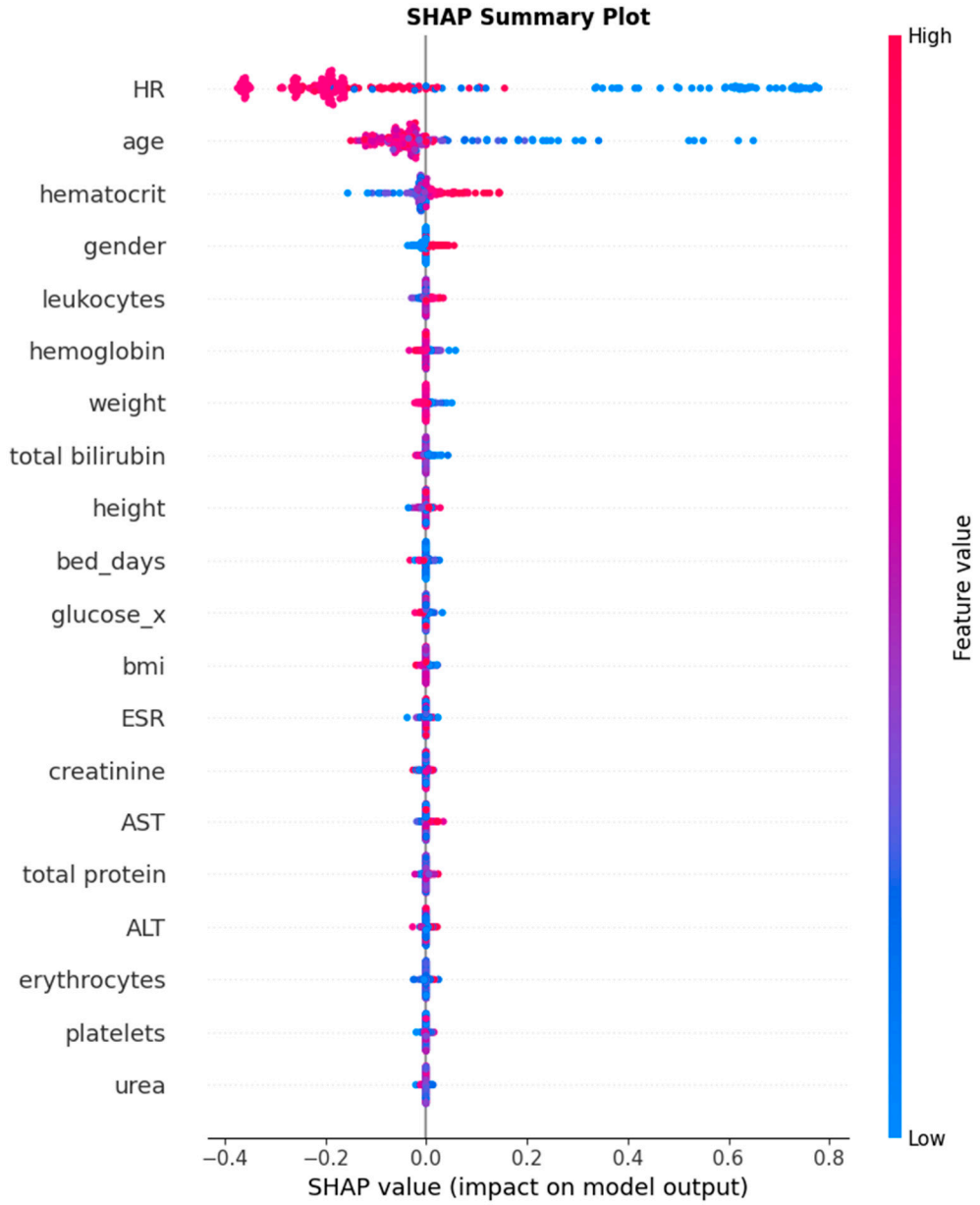


Figure 6. SHAP Summary Plot.

4. Discussion

It is not surprising that clinical data has become a lightning rod for machine learning, providing tabular clinical data and allowing coxarthrosis and gonarthrosis (which are both clinically defined diseases) to be classified. This work stands apart by being completely dependent on preoperative indicators collected prior to surgical intervention, providing early diagnostics without involving any advanced imaging modalities nor any invasive procedures. By comparing a broad range of machine learning paradigms, including gradient boosting ensembles, neural network architectures, transformer-based models, automated machine learning frameworks, and pretrained tabular foundation models.

The findings show that clinically meaningful and stable patterns distinguishing hip from knee osteoarthritis are present in routine preoperative data, and we support the use of modern tabular learning techniques to extract these patterns with high fidelity. The RealTabPFN v2.5 model had a strong performance, with favorable calibration and balanced precision–recall behavior, and suggests that pretrained tabular foundation models can exploit the prior structural knowledge acquired from large collections of tabular tasks with

superior generalization in moderate-sized clinical datasets. SHAP-based interpretability analysis demonstrated that heart rate and age dominated the model's decision-making process, with heart rate alone accounting for around 66% of the predictive contribution. This finding suggests that systemic physiological state, rather than isolated laboratory abnormalities, plays a critical role in differentiation between osteoarthritis subtypes at clinical presentation.

The identification of Resting Heart Rate (HR) as the dominant predictor, with a 66.2% SHAP contribution, is a striking and unexpected finding. While the dominance of a systemic marker like HR in diagnosing a structural joint disease initially appears counter-intuitive, several clinical and pathophysiological factors specific to this dataset likely drive this distinction. Importantly, we caution against interpreting HR as a direct causal determinant of joint-specific osteoarthritis. Instead, it should be viewed as a strong surrogate indicator or proxy capturing systemic and contextual differences between patient groups at preoperative presentation.

Clinically, this association likely reflects three key mechanisms. First, HR may capture pain-induced autonomic dysregulation. Chronic pain sensitization [47,48], which can differ in intensity and quality between coxarthrosis and gonarthrosis, is known to elevate resting heart rate via sustained sympathetic tone. If the KOA cohort experienced higher levels of preoperative acute pain or anxiety, this would manifest as a consistent elevation in HR. Second, the distinction likely reflects varying degrees of physical deconditioning. KOA is frequently associated with greater immobility and a more sedentary lifestyle compared to hip OA [49], potentially leading to faster cardiovascular deconditioning and a consequently higher resting heart rate. Third, differences in age distribution and medication exposure between the two cohorts may further contribute to systematic HR variation. In the present dataset, KOA patients were older on average. Age-related changes in cardiovascular regulation, combined with heterogeneous use of beta-blockers [50], analgesics [51], or anxiolytics [52] that were not explicitly captured in the dataset, may influence resting HR in a joint-specific manner. In this context, the model likely leverages HR not just as an isolated biomarker, but as a composite physiological feature integrating age, pain perception, and pharmacological modulation. We propose this finding as 'hypothesis-generating,' suggesting that systemic biomarkers like HR might hold untapped potential for phenotyping OA patients. However, these high SHAP values represent a statistical association rather than direct causation. Future targeted studies incorporating pain scores, medication data, and autonomic markers are necessary to disentangle these confounding factors from intrinsic disease pathology.

This study achieved a ROC-AUC of 0.9831 ± 0.0053 for binary classification of coxarthrosis versus gonarthrosis using preoperative clinical and laboratory biomarkers with RealTabPFN v2.5 (Tuned), representing a fundamentally novel approach that substantially differs from and outperforms existing osteoarthritis classification methods. While previous machine learning studies in OA research have primarily focused on single-joint severity grading or disease progression prediction, this work uniquely addresses inter-joint OA subtype differentiation using only routine preoperative clinical data accessible in any healthcare setting. Performance comparisons reveal several critical distinctions from prior approaches: hip OA severity classification [53] achieved AUC of 0.93 on external validation using deep learning (ResNet-50 with CBAM) on 20,745 radiographs, but this imaging-dependent approach classifies severity only within hip OA patients and requires specialized radiographic infrastructure unavailable in resource-constrained settings. Hip OA detection via gait analysis [54] achieved 98.21% binary accuracy using support vector machines on 3D motion capture data from 179 subjects, achieving comparable binary classification performance to the present study but requiring expensive motion capture

laboratory equipment ($50,000+) and specialized technicians, limiting applicability to research institutions only. Knee OA severity classification [55] achieved substantially lower accuracy of 76.7% using deep learning (NfNet-F1) on 14,607 radiographs for multi-class KL severity grading (5 classes), approximately 21 percentage points lower than the present study despite utilizing advanced imaging modalities.

Methodologically, class imbalance was addressed using a hybrid strategy combining SMOTE with class-weighted learning. While synthetic oversampling in clinical datasets carries a theoretical risk of generating physiologically implausible samples [56], several aspects of the present study mitigate this concern. SMOTE was applied strictly within the training folds, ensuring that all validation and test evaluations were performed exclusively on real, observed patient data and preventing information leakage. The feature space in this study predominantly consists of continuous, routinely measured clinical and laboratory variables constrained within physiologically plausible ranges. Under these conditions, SMOTE operates through local interpolation between nearest neighbors [57,58], meaning that synthetic samples represent convex combinations of existing patients and are therefore likely to correspond to realistic clinical phenotypes within the observed data manifold. SMOTE was preferred over class-weighted learning alone because cost-sensitive approaches adjust the loss function [59,60] but do not address sparsity of the minority class in feature space. In moderately sized tabular clinical datasets, this sparsity can lead to poorly defined decision boundaries and reduced sensitivity for the minority condition. Local densification of the minority class distribution through SMOTE improves boundary definition in underrepresented regions, while class-weighted learning simultaneously penalizes misclassification asymmetrically. The combined approach was therefore selected to enhance minority class sensitivity without compromising evaluation integrity.

While the obtained results are promising, this study has several limitations. First, the study sample was restricted to a single medical center in Almaty, Kazakhstan, which may limit the generalizability of the findings to other regions or healthcare institutions. In addition, site specific clinical practices, preoperative workflows, and correlated demographic characteristics inherent to a single center dataset may have contributed to the observed model performance. Second, a potential demographic bias is present due to a significant age difference between the study groups: knee osteoarthritis patients were approximately 10 years older on average than hip osteoarthritis patients (65.9 ± 7.7 vs. 55.4 ± 15.1 years). Given that age was identified as an influential predictive feature, this imbalance may partially contribute to the model's discriminative performance and should be considered when interpreting the results. Third, although heart rate (HR) was identified as a strong predictor, this association has not been validated in independent cohorts to establish its biological significance within the context of this disease. Fourth, we acknowledge that the binary classification methods employed may be insufficient for analyzing complex clinical cases involving polyarticular disease or early-stage, mild osteoarthritis. Fifth, while SHAP analysis elucidates feature contribution, these insights reflect statistical associations rather than causal mechanisms; therefore, interpretation requires caution. Finally, the study lacked imaging data, genetic markers, and detailed patient-reported outcomes, which could have enhanced both the predictive power and the depth of the model's insights.

An additional limitation arises from the reliance on complete case analysis, which resulted in the exclusion of more than half of the initially available cohort. This approach may introduce systematic selection bias, as patients with complete laboratory and clinical records may differ meaningfully from those excluded due to missing data. Consequently, model performance and generalizability may be affected. Future studies should therefore explore advanced missing data strategies, such as multiple imputation or model aware handling of missingness, to quantify their impact on bias, uncertainty, and predictive performance.

Regarding clinical implementation, the proposed framework is designed as a Clinical Decision Support System (CDSS) and a triage aid for primary healthcare, rather than a standalone diagnostic replacement. Its primary utility lies in identifying high-probability cases for specialist referral in resource-limited settings where imaging is not immediately available.

It is important to note that our model was trained on a cohort with definite OA (KL grade $\geq$ II). Consequently, its performance in patients with early-stage osteoarthritis, mixed hip-knee involvement, or polyarticular disease has not yet been established and requires separate validation. Before any clinical deployment, strict safeguards—including 'human-in-the-loop' verification by clinicians and external validation across diverse populations—must be implemented to ensure patient safety and algorithmic fairness.

The developed model requires validation across diverse populations and healthcare settings to assess generalizability and performance variability. In addition, future work will include explicit robustness analyses, such as sensitivity testing across demographic subgroups, feature ablation to assess predictor dependence, and calibration assessment under distributional shifts. Longitudinal studies could shed light on the relationship between preoperative predictors, postoperative outcomes, and recovery trajectories. Targeted investigation into the biological mechanisms linking HR to osteoarthritis subtypes may facilitate the identification of modifiable risk factors. Incorporating multi-modal data—imaging, genetics, biomarkers, and patient-reported outcomes—would enable the application of multimodal machine learning, potentially improving diagnostic accuracy. Clinical implementation will require interpretable clinical decision support tools validated for their impact on diagnosis and treatment. Furthermore, fine-tuning methodologies on medical tabular data and few-shot learning techniques could potentially enhance performance and facilitate broader clinical adoption.

This study demonstrates the feasibility and high accuracy of differential diagnosis for coxarthrosis and gonarthrosis based on machine learning methods using preoperative clinical parameters, highlighting the superiority of pre-trained tabular foundation models over traditional gradient boosting methods or neural networks in this context. The identification of HR and age as two primary independent predictors opens new avenues for research. Nevertheless, responsible integration of these findings into clinical practice requires significant effort, including extensive external validation, mechanistic studies, multimodal data integration, algorithmic fairness assessment, prospective outcome studies, and user implementation evaluation. The outstanding performance achieved using basic clinical data suggests that ML-enabled diagnostic tools could benefit resource-limited regions and primary healthcare settings by improving early detection, treatment planning, and outcomes for patients with osteoarthritis—one of the most prevalent and disabling musculoskeletal conditions worldwide. As machine learning technologies evolve and healthcare systems establish necessary governance frameworks for algorithmic decision support, this model—adaptable across various disciplines and diagnostic tasks—may serve as a benchmark for the development and validation of evidence-based clinical AI applications.

5. Conclusions

This study establishes the feasibility and high accuracy of utilizing machine learning for the differential diagnosis of coxarthrosis and gonarthrosis based exclusively on preoperative clinical parameters, identifying the pretrained tabular foundation model, RealTabPFN v2.5, as superior to traditional gradient boosting or neural network methods. The model achieved an ROC-AUC of 0.9831, demonstrating that robust diagnostic patterns exist within routine laboratory and physiological data without the immediate need for advanced

imaging. A significant finding was the identification of heart rate and age as the primary independent predictors, with heart rate alone accounting for approximately 66% of the predictive contribution. This suggests that systemic physiological states may play a critical, previously underappreciated role in differentiating osteoarthritis subtypes. While the study is limited by its retrospective nature and single-center population, the results highlight the potential of AI-assisted diagnostic tools to enhance value in resource-constrained and primary care settings. Future research should focus on extensive external validation, the exploration of biological mechanisms linking heart rate to OA subtypes, and the integration of multimodal data to further refine clinical decision support systems.

Author Contributions: Conceptualization, Z.B., B.I., A.B., B.A. and N.T.; methodology, Z.A., A.B., B.I., B.A., N.T., K.O. and N.M.-N.; software, B.A., N.T., A.O. and Z.A.; validation, B.I., K.O., A.O., N.M.-N. and Z.A.; formal analysis, B.I., K.O., A.O., N.M.-N. and Z.A.; investigation, Z.B., B.I., A.B., B.A. and N.T.; resources, Z.B., A.B., B.A. and N.T.; data curation, A.O., N.M.-N. and Z.A.; writing—original draft preparation, Z.B., A.B., B.A., N.T. and Z.A.; writing—review and editing, N.M.-N., B.I., K.O. and A.O.; visualization, Z.B., A.B., B.A. and N.T.; supervision, B.I., K.O., A.O. and N.M.-N.; project administration, B.I. and K.O.; funding acquisition, K.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992820).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Bioethics Committee of NCSTO named after academician N.D.Batpenov.

Informed Consent Statement: Informed consent for participation is not required as per local legislation [Protocol № 1/4].

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Conflicts of Interest: Authors Zhanel Baigaraeva and Assiya Boltaboyeva were employed by the company LLP "Kazakhstan R&D Solutions" and author Naoya Maeda-Nishino was employed by the company HAKUAI Medical Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OA	Osteoarthritis
ROC–AUC	Receiver Operating Characteristic—Area Under Curve
F1-score	F1 Metric (harmonic mean of precision and recall)
SHAP	SHapley Additive exPlanations
BMI	Body Mass Index
ESR	Erythrocyte Sedimentation Rate
CRP	C-reactive Protein
AUC	Area Under Curve
SMOTE	Synthetic Minority Over-sampling Technique
XGBoost	eXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
CatBoost	Category Boosting (Gradient Boosting Algorithm by Yandex)
RealTabPFN	RealTab Pretrained Foundation Model
TabDPT	Transformer-based Tabular Data Pretraining
TabM	Tabular Model

RealMLP	Real Multilayer Perceptron (for tabular data)
AutoGluon	Automated Machine Learning Framework
ModernNCA	Metric Learning-based Non-Convex Embedding
HR	Heart Rate
PR	Precision–Recall

References

1. Sen, R.; Hurley, J.A. *Osteoarthritis; StatPearls—NCBI Bookshelf*; National Library of Medicine: Bethesda, MD, USA, 2023. Available online: https://www.ncbi.nlm.nih.gov/books/NBK482326/ (accessed on 20 February 2023).
2. Allen, K.D.; Thoma, L.M.; Golightly, Y.M. Epidemiology of osteoarthritis. *Osteoarthr. Cartil.* **2021**, *30*, 184–195. [CrossRef] [PubMed]
3. Bi, A.S.; Sachs, J.P.; Mufti, Y.N.; Franzia, C.H.; Jesse, J.E.; Lemme, N.J.; Wimmer, M.A.; Cole, B.J. Why do osteoarthritic knees hurt? Leveraging the impact of load for symptom management. *J. Cartil. Jt. Preserv.* **2025**, 100271. [CrossRef]
4. He, Y.; Li, Z.; Alexander, P.G.; Ocasio-Nieves, B.D.; Yocum, L.; Lin, H.; Tuan, R.S. Pathogenesis of osteoarthritis: Risk factors, regulatory pathways in chondrocytes, and experimental models. *Biology* **2020**, *9*, 194. [CrossRef] [PubMed]
5. Hasan, M.; Shuckett, R. Clinical features and pathogenetic mechanisms of osteoarthritis of the hip and knee. *Br. Columbia Med. J.* **2018**, *52*, 393–398. Available online: https://bcmj.org/articles/clinical-features-and-pathogenetic-mechanisms-osteo%C2%ADarthritis-hip-and-knee (accessed on 9 May 2018).
6. Li, H.; Liang, X.; Sun, Y.; Jia, H.; Li, J.; Li, G. Global, regional, and national burdens of osteoarthritis from 1990 to 2021: Findings from the 2021 global burden of disease study. *Front. Med.* **2024**, *11*, 1476853. [CrossRef]
7. Chitnavis, J.; Sinsheimer, J.S.; Suchard, M.A.; Clipsham, K.; Carr, A.J. End-stage coxarthrosis and gonarthrosis. Aetiology, clinical patterns and radiological features of idiopathic osteoarthritis. *Rheumatology* **2000**, *39*, 612–619. [CrossRef]
8. Piccolo, C.L.; Mallio, C.A.; Vaccarino, F.; Grasso, R.F.; Zobel, B.B. Imaging of knee osteoarthritis: A review of multimodal diagnostic approach. *Quant. Imaging Med. Surg.* **2023**, *13*, 7582–7595. [CrossRef]
9. Roemer, F.; Guermazi, A.; Demehri, S.; Wirth, W.; Kijowski, R. Imaging in osteoarthritis. *Osteoarthr. Cartil.* **2021**, *30*, 913–934. [CrossRef]
10. Nielsen, R.L.; Monfeuga, T.; Kitchen, R.R.; Egerod, L.; Leal, L.G.; Schreyer, A.T.H.; Gade, F.S.; Sun, C.; Helenius, M.; Simonsen, L.; et al. Data-driven identification of predictive risk biomarkers for subgroups of osteoarthritis using interpretable machine learning. *Nat. Commun.* **2024**, *15*, 2817. [CrossRef]
11. Mourad, C.; Vande Berg, B. Osteoarthritis of the hip: Is radiography still needed? *Skelet. Radiol.* **2023**, *52*, 2259–2270. [CrossRef]
12. Newman, S.; Ahmed, H.; Rehmatullah, N. Radiographic vs. MRI vs. arthroscopic assessment and grading of knee osteoarthritis—Are we using appropriate imaging? *J. Exp. Orthop.* **2022**, *9*, 2. [CrossRef] [PubMed]
13. Wayne, B. Medical imaging of osteoarthritis of the knee in clinical trials. *CenterWatch*. 2024. Available online: https://www.centerwatch.com/insights/medical-imaging-of-osteoarthritis-of-the-knee-in-clinical-trials/ (accessed on 12 August 2024).
14. Hayashi, D.; Guermazi, A.; Kwoh, C.K. Clinical and translational potential of MRI evaluation in knee osteoarthritis. *Curr. Rheumatol. Rep.* **2014**, *16*, 391. [CrossRef] [PubMed]
15. Li, Q.; Amano, K.; Link, T.M.; Ma, C.B. Advanced Imaging in Osteoarthritis. *Sports Health* **2016**, *8*, 418–428. [CrossRef] [PubMed]
16. Liu, Y.; Leong, A.T.L.; Zhao, Y.; Xiao, L.; Mak, H.K.F.; Tsang, A.C.O.; Lau, G.K.K.; Leung, G.K.K.; Wu, E.X. A low-cost and shielding-free ultra-low-field brain MRI scanner. *Nat. Commun.* **2021**, *12*, 7238. [CrossRef]
17. Herrmann, J.; Afat, S.; Gassenmaier, S.; Koerzdoerfer, G.; Lingg, A.; Almansour, H.; Nickel, D.; Werner, S. Image quality and diagnostic performance of accelerated 2D hip MRI with deep learning reconstruction based on a deep iterative hierarchical network. *Diagnostics* **2023**, *13*, 3241. [CrossRef]
18. Boffa, A.; Merli, G.; Andriolo, L.; Lattermann, C.; Salzmann, G.M.; Filardo, G. Synovial fluid biomarkers in knee osteoarthritis: A systematic review and quantitative evaluation using BIPEDs criteria. *Cartilage* **2021**, *13*, 82S–103S. [CrossRef]
19. Jayadev, C.; Hulley, P.; Swales, C.; Snelling, S.; Collins, G.; Taylor, P.; Price, A. Synovial fluid fingerprinting in end-stage knee osteoarthritis: A novel biomarker concept. *Bone Jt. Res.* **2020**, *9*, 623–632. [CrossRef]
20. Fan, Z.; Song, W.; Ke, Y.; Jia, L.; Li, S.; Li, J.J.; Zhang, Y.; Lin, J.; Wang, B. XGBoost-SHAP-based interpretable diagnostic framework for knee osteoarthritis: A population-based retrospective cohort study. *Arthritis Res. Ther.* **2024**, *26*, 213. [CrossRef]
21. Wang, Z.; Zhou, Y.; Zeng, X.; Zhou, Y.; Yang, T.; Hu, K. An explainable machine learning-based prediction model for sarcopenia in elderly Chinese people with knee osteoarthritis. *Aging Clin. Exp. Res.* **2025**, *37*, 67. [CrossRef]
22. Chen, W.; Zheng, H.; Ye, B.; Guo, T.; Xu, Y.; Fu, Z.; Ji, X.; Chai, X.; Li, S.; Deng, Q. Identification of biomarkers for knee osteoarthritis through clinical data and machine learning models. *Sci. Rep.* **2025**, *15*, 1703. [CrossRef]
23. Ningrum, D.N.A.; Kung, W.-M.; Tzeng, I.-S.; Yuan, S.-P.; Wu, C.-C.; Huang, C.-Y.; Muhtar, M.S.; Nguyen, P.-A.; Li, J.Y.-C.; Wang, Y.-C. A deep learning model to predict knee osteoarthritis based on nonimage longitudinal medical record. *J. Multidiscip. Healthc.* **2021**, *14*, 2477–2485. [CrossRef] [PubMed]

24. Wang, X.; Liu, T.; Sheng, Y.; Qiu, C.; Zhang, Y.; Liu, Y.; Wu, C. Exploration and verification of circulating diagnostic biomarkers in osteoarthritis based on machine learning. *Front. Genet.* **2025**, *16*, 1513675. [CrossRef] [PubMed]
25. TABArena—A Hugging Face Space by TabArena. Available online: https://huggingface.co/spaces/TabArena/leaderboard (accessed on 10 January 2026).
26. Kohn, M.D.; Sassoon, A.A.; Fernando, N.D. Classifications In brief: Kellgren-Lawrence classification of osteoarthritis. *Clin. Orthop. Relat. Res.* **2016**, *474*, 1886–1893. [CrossRef] [PubMed]
27. *ISO 25237:2017*; Health Informatics—Pseudonymization. International Organization for Standardization (ISO): Geneva, Switzerland, 2017.
28. Bay-Jensen, A.-C.; Bihlet, A.; Byrjalsen, I.; Andersen, J.R.; Riis, B.J.; Christiansen, C.; Michaelis, M.; Guehring, H.; Ladel, C.; Karsdal, M.A. Serum C-reactive protein metabolite (CRPM) is associated with incidence of contralateral knee osteoarthritis. *Sci. Rep.* **2021**, *11*, 6583. [CrossRef] [PubMed]
29. Lex, J.R.; Abbas, A.; Mosseri, J.; Toor, J.S.; Simone, M.; Ravi, B.; Whyne, C.; Khalil, E.B. Using machine learning to predict-then-optimize elective orthopedic surgery scheduling to improve operating room utilization: Retrospective study. *JMIR Med. Inform.* **2025**, *13*, e70857. [CrossRef]
30. World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA* **2013**, *310*, 2191–2194. [CrossRef]
31. Mohammed, S.; Budach, L.; Feuerpfeil, M.; Ihde, N.; Nathansen, A.; Noack, N.; Patzlaff, H.; Naumann, F.; Harmouch, H. The effects of data quality on machine learning performance on tabular data. *Inf. Syst.* **2025**, *132*, 102549. [CrossRef]
32. Zierle-Ghosh, A.; Jan, A. *Physiology, Body Mass Index; StatPearls—NCBI Bookshelf*; National Library of Medicine: Bethesda, MD, USA, 2023. Available online: https://www.ncbi.nlm.nih.gov/books/NBK535456/ (accessed on 5 November 2023).
33. Scikit-Learn. Importance of Feature Scaling. Available online: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html (accessed on 10 January 2026).
34. Blog, M. Feature Scaling: Standardization vs. Normalization and Various Types of Normalization. *Minkyung's Blog*. Available online: https://mkang32.github.io/python/2020/12/27/feature-scaling.html (accessed on 27 December 2020).
35. Singh, D.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524. [CrossRef]
36. Scikit-Learn Developers. Common Pitfalls and Recommended Practices. Scikit-Learn. Available online: https://scikit-learn.org/stable/common_pitfalls.html (accessed on 3 July 2024).
37. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? *arXiv* **2022**, arXiv:2207.08815. [CrossRef]
38. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
39. Holzmüller, D.; Grinsztajn, L.; Steinwart, I. RealMLP: Advancing MLPs and Default Parameters for Tabular Data. OpenReview. Available online: https://openreview.net/forum?id=fwajDrDy89 (accessed on 10 January 2026).
40. Ye, H.; Yin, H.; Zhan, D.; Chao, W. Revisiting nearest neighbor for tabular data: A deep tabular baseline two decades later. *arXiv* **2024**, arXiv:2407.03257. [CrossRef]
41. Gorishniy, Y.; Kotelnikov, A.; Babenko, A. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. *arXiv* **2024**, arXiv:2410.24210. [CrossRef]
42. Ma, J.; Thomas, V.; Hosseinzadeh, R.; Kamkari, H.; Labach, A.; Cresswell, J.C.; Golestan, K.; Yu, G.; Caterini, A.L.; Volkovs, M. TABDPT: Scaling Tabular Foundation Models on Real Data. *arXiv* **2024**, arXiv:2410.18164. [CrossRef]
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
44. Erickson, N.; Mueller, J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li, M.; Smola, A. AutoGluon-Tabular: Robust and accurate AutoML for structured data. *arXiv* **2020**, arXiv:2003.06505. [CrossRef]
45. AutoGluon-Tabular Hyperparameters—Amazon SageMaker AI. Available online: https://docs.aws.amazon.com/sagemaker/latest/dg/autogluon-tabular-hyperparameters.html (accessed on 10 January 2026).
46. Garg, A.; Ali, M.; Hollmann, N.; Purucker, L.; Müller, S.; Hutter, F. Real-TabPFN: Improving tabular foundation models via continued pre-training with Real-World data. *arXiv* **2025**, arXiv:2507.03971. [CrossRef]
47. TabPFN-2.5 performance on the standard TabArena-lite benchmark. *arXiv* **2025**, arXiv:2511.08667. [CrossRef]
48. Reynolds, C.A.; Minic, Z. Chronic Pain-Associated Cardiovascular Disease: The Role of Sympathetic Nerve Activity. *Int. J. Mol. Sci.* **2023**, *24*, 5378. [CrossRef]
49. Karri, J.; Zhang, L.; Li, S.; Chen, Y.; Stampas, A.; Li, S. Heart Rate Variability: A Novel Modality for Diagnosing Neuropathic Pain after Spinal Cord Injury. *Front. Physiol.* **2017**, *8*, 495. [CrossRef]
50. Ma, X.; Zhang, K.; Ma, C.; Zhang, Y.; Ma, J. Physical Activity and the Osteoarthritis of the Knee: A Mendelian Randomization Study. *Medicine* **2024**, *103*, e38650. [CrossRef]

51. Diaconu, C.C.; Marcu, D.R.; Bratu, O.G.; Stanescu, A.M.A.; Gheorghe, G.; Hlescu, A.A.; Mischianu, D.L.; Manea, M. Beta-Blockers in Cardiovascular Therapy: A Review. *J. Mind Med. Sci.* **2019**, *6*, 216–223. [CrossRef]
52. Cleveland Clinic. Analgesics Class of Medications (Painkillers & NSAIDs). 2026. Available online: https://my.clevelandclinic.org/health/drugs/21483-analgesics (accessed on 8 January 2026).
53. Simone, C.G.; Bobrin, B.D. Anxiolytics and Sedative-Hypnotics Toxicity. *StatPearls*. 2023. Available online: https://www.ncbi.nlm.nih.gov/books/NBK562309/ (accessed on 13 January 2023).
54. Xu, S.; Guo, C.; Zuo, J.; Yang, M.; Chen, B.; Li, S.; Xiao, J.; Tang, X.; Qin, Y. Development and validation of a deep learning model for automatic severity grading of hip osteoarthritis: A multi-center study. *Ann. Med.* **2025**, *57*, 2584361. [CrossRef]
55. Yang, Q.; Ji, X.; Zhang, Y.; Du, S.; Ji, B.; Zeng, W. An interpretable machine learning approach for predicting and grading hip osteoarthritis using gait analysis. *BMC Musculoskelet. Disord.* **2025**, *26*, 580. [CrossRef]
56. Yayli, S.B.; Kılıç, K.; Beyaz, S. Deep learning in gonarthrosis classification: A comparative study of model architectures and single vs. multi-model methods. *Front. Artif. Intell.* **2025**, *8*, 1413820. [CrossRef]
57. Alkhawaldeh, I.M.; Albalkhi, I.; Naswhan, A.J. Challenges and Limitations of Synthetic Minority Oversampling Techniques in Machine Learning. *World J. Methodol.* **2023**, *13*, 373–378. [CrossRef]
58. Bellinger, C.; Drummond, C.; Japkowicz, N. Manifold-Based Synthetic Oversampling with Manifold Conformance Estimation. *Mach. Learn.* **2018**, *107*, 605–637. [CrossRef]
59. Tourt, D.; Booker, Q.; Rebman, C.; Ahmed, S.; Patel, R.; Liu, Y.; Batista, G.; Prati, R.; Monard, M.; Bunkhumpornpat, C.; et al. Strategic Application of SMOTE and Its Variants to Enhance AI-Driven Healthcare Classification in Imbalanced Datasets. *Issues Inf. Syst.* **2025**, *26*, 70–85. [CrossRef]
60. Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.