Physical and Engineering Sciences in Medicine (2022) 45:219–229
https://doi.org/10.1007/s13246-022-01106-6

SCIENTIFIC PAPER

Explainable machine learning for knee osteoarthritis diagnosis based on a novel fuzzy feature selection methodology

Christos Kokkotis[1,2] · Charis Ntakolia[3,4] · Serafeim Moustakidis[5] · Giannis Giakas[2] · Dimitrios Tsaopoulos[1]

Received: 2 August 2021 / Accepted: 19 January 2022 / Published online: 31 January 2022

Abstract

Knee Osteoarthritis (KOA) is a degenerative joint disease of the knee that results from the progressive loss of cartilage. Due to KOA's multifactorial nature and the poor understanding of its pathophysiology, there is a need for reliable tools that will reduce diagnostic errors made by clinicians. The existence of public databases has facilitated the advent of advanced analytics in KOA research however the heterogeneity of the available data along with the observed high feature dimensionality make this diagnosis task difficult. The objective of the present study is to provide a robust Feature Selection (FS) methodology that could: (i) handle the multidimensional nature of the available datasets and (ii) alleviate the defectiveness of existing feature selection techniques towards the identification of important risk factors which contribute to KOA diagnosis. For this aim, we used multidimensional data obtained from the Osteoarthritis Initiative database for individuals without or with KOA. The proposed fuzzy ensemble feature selection methodology aggregates the results of several FS algorithms (filter, wrapper and embedded ones) based on fuzzy logic. The effectiveness of the proposed methodology was evaluated using an extensive experimental setup that involved multiple competing FS algorithms and several well-known ML models. A 73.55% classification accuracy was achieved by the best performing model (Random Forest classifier) on a group of twenty-one selected risk factors. Explainability analysis was finally performed to quantify the impact of the selected features on the model's output thus enhancing our understanding of the rationale behind the decision-making mechanism of the best model.

Keywords KOA diagnosis · Machine learning · Clinical data · Explainability · Feature selection

Introduction

Knee Osteoarthritis (KOA) is one of the most common types of osteoarthritis and musculoskeletal disorder. Being the 11th highest cause of disability globally, KOA is a multifactorial disease that results from mechanical and constitutional factors [1]. Obesity, age, gender, knee injuries and lifestyle are likely risk factors of KOA as they have been highlighted in the relevant recent literature [2]. In addition, swelling, pain and stiffness have been characterized as typical symptoms of the disease with irreversible cartilage damage being KOA's main consequence [3–5]. KOA is closely associated with a huge economic burden for the healthcare system and an unbearable health burden of the patients and their families [6, 7]. Significant consequences of KOA are the social isolation and low quality of life of the individual [8, 9]. Furthermore, the quantification of KOA is performed with the Kellgren–Lawrence (KL) severity grading scale, which is the most commonly grading system (current gold standard) and consists of five severity grades, from 0 to 4 [10].

Despite the fact that the scientific community has put a lot of effort into KOA research, a major challenge remains with respect to early diagnosis, long-term diagnosis and treatment of KOA. The parallel increase in computing power along with the collection of big datasets combined with the need to address the above challenges has led many research teams to use artificial intelligence (AI) techniques in the field of

✉ Christos Kokkotis
chkokkotis@gmail.com

[1] Institute for Bio-Economy & Agri-Technology, Center for Research and Technology Hellas, 38333 Volos, Greece

[2] TEFAA, Department of Physical Education & Sport Science, University of Thessaly, 42100 Trikala, Greece

[3] University Mental Health Research Institute, 11527 Athens, Greece

[4] School of Naval Architecture and Marine Engineering, National Technical University of Athens, 15772 Athens, Greece

[5] AIDEAS OÜ, Narva mnt 5, 10117 Tallinn, Estonia

KOA [11]. In light of the above, several AI enabled studies have been proposed in the recent literature with the objective to diagnose or predict KOA. Yoo et al. used data from the Fifth Korea National Health and Nutrition Examination Surveys (KNHANES V-1) and the Osteoarthritis Initiative (OAI) to build an artificial neural network (ANN)-based a scoring system for the identification of KOA severity [12]. The proposed ANN model achieved an area under the curve (AUC) of 76% for the symptomatic KOA in an external validation with OAI data. In another study, Lim et al. proposed a method for early diagnosis of KOA based on clinical data from Korean National Health and Nutrition Examination Survey (KNHANES) [13]. They achieved a 76.8% AUC by using a deep neural network with scaled principal component analysis. In 2019, Christodoulou et al. investigated the deep learning capabilities in KOA diagnosis [14]. They used clinical data from OAI database and they achieved an 86.95% accuracy working on an aged subgroup (70 +).

In another study, Moustakidis et al. proposed a deep learning methodology for the recognition of participants being at high risk of developing KOA in at least one knee and participants with symptomatic KOA [15]. They employed self-reported data about disability, joint symptoms, general health and function from all individuals without or with KOA from the baseline visit (OAI) and they achieved accuracies up to 86.95%. Furthermore, Kwon et al. proposed an automatic classification of KOA severity that made use of gait analysis data and radiographic imaging (from Seoul National University Hospital) [16]). They employed Inception-ResNet-v2 for feature extraction from X-rays and a support vector machine for KOA diagnosis achieving AUC scores of 93%, 82%, 83%, 88% and 97% for the KL grades 0–4, respectively. In addition, Moustakidis et al. proposed a KOA classification approach with a focus on both accuracy and fairness [17]. They worked on different subgroups of participants from self-reported clinical data (OAI) and the dense neural networks methodology improved the accuracy up to 79.6% with fairness measured by balanced equalized odds (~ 92%) and demographic parity (98.5%) in the KOA case study.

Given that medical data and features can be subjective or difficult to interpret, medical decision making has a great potential to benefit from the use of fuzzy logic (FL). FL has been used to diagnose or facilitate decision making systems tackling many diseases, including OA. Hardi et al. proposed an expert system based on the fuzzy Tsukamoto method for OA diagnosis [18]. They treated symptoms of OA as fuzzy values that were further converted into firm value by using a weighted average demonstrating a 90% accuracy in the task of diagnosis of osteoarthritis disease. In general, various feature selection methods have integrated fuzzy logic in their internal mechanisms in order to handle the observed fuzziness and therefore improve the way that features are treated and combined. For instance, with emphasis to medical applications, the mutual information method combined with FL was used: (i) to select miRNAs in cancer [19]; (ii) to classify tumors [20]; and to select features for multilabel learning [21]. Similar studies include fuzzy entropy by using thresholds [22] for feature selection in various medical datasets and fuzzy rough sets [23, 24] for dimensionality reduction of feature space to prevent samples from misclassification.

In the aforementioned literature, various feature selection methods have been proposed for addressing the dimensionality reduction problem in the case of large medical datasets. Conventional approaches, such as filters, wrappers, and embedded methods, can be considered effective depending on the nature of the dataset. Therefore, the optimal feature selection method in each case can be decided after a thorough and comparative evaluation among various feature selection techniques and prediction models [25, 26]. This leads to excessive computational effort and time [27, 28]. To this end, it is of high importance to develop novel FS techniques that will be effective in large datasets and decrease significantly the dimensionality of feature space. Motivated by: (i) the ability of fuzzy logic to enhance the effectiveness of feature selection techniques, as it is shown in the related literature, and (ii) the effectiveness of the state-of-the-art feature selection technique [25, 26] that is based on a voting scheme, a fuzzy ensemble FS methodology is proposed in this paper that aggregates the results of several FS algorithms (filter, wrapper and embedded). To handle the multidimensional nature of the OAI dataset and to avoid bias and alleviate the defectiveness of single feature selection results, fuzzy logic is employed to combine multiple feature importance scores thus leading to a more robust selection of informative features. The proposed method contributes to the significant reduction of the initial OAI feature dimensionality and to a decrease in the computational complexity of the classification models employed. To prove the effectiveness of the proposed methodology, an extensive experimental setup was designed involving multiple competing FS algorithms and several well-known ML models. As a post-hoc explainability, SHapley Additive exPlanations (SHAP) model was finally employed to identify the contribution of the selected features and the rationale behind the decision-making mechanism of best performing model.

Methods

Dataset description

For the purpose of this study, data were obtained from the osteoarthritis initiative (OAI) database (available on https://nda.nih.gov/oai/). OAI is a prospective observational, multicenter and longitudinal study of KOA, which is sponsored by the National Institutes of Health (part of the Department of Health and Human Services). The goals of the OAI are to provide resources to enable a better understanding of prevention and treatment of knee osteoarthritis. OAI has enrolled 4796 women and men, aged 45–79 years. The present study used clinical evaluation tabular data (643 features in total either numeric or categorical) from the baseline visit from all participants with or without KOA. The features of clinical dataset were divided into seven categories as shown in Table 1. Furthermore, in the present study, Kellgren and Lawrence (KL) grades were used as the outcome for the classification task.

Methodology

The proposed AI methodology for KOA diagnosis consists of five processing steps: (i) data pre-processing, (ii) application of FS techniques, (iii) learning process, (iv) evaluation of the classification results and (v) explainability analysis, as illustrated in Fig. 1. An extensive explanation of the steps of the proposed methodology is given in the following subsections. The code was implemented in Python 3.6 by using scikit-learn 0.24.2.

Problem definition

In this study, we defined the KL-grade prediction task as a binary class classification problem. Furthermore, 3872 subjects in total were employed. From the 4976 subjects in the OAI database we dropped out the participants who had not KL ≥ 2 at baseline and the participants who had knee Osteoarthritis in any of their knees (participants who had not KL0 or KL1 grade at baseline). Furthermore, we excluded participants with missing values. Specifically, the subjects of the study were divided into two equal groups:

(i) KOA—participants who have KL ≥ 2 at baseline. Participants in the group who had KL grades equal (early diagnosis) or higher than 2 in at least one of the two knees or in both at baseline;
(ii) Non-KOA—participants who had KL0 or KL1 grade at baseline. Especially, this group of participants do not have KOA in any of their knees.

Data pre-processing

Mode imputation was employed to handle categorical and continuous missing values [29]. In our study, data were normalised to (0, 1) to build a common basis for the FS algorithms and learning techniques that follow [30]. Furthermore, to cope with the imbalance data problem a stratified strategy for data resampling was applied. In particular, the number of the subjects in the majority class was reduced in order to become equal to the number of samples on the minority class [31]. The stratified sampling was selected due to a number of benefits such as the smaller estimation error, effectiveness in measurements and better representation of all subgroups of the classes.

Proposed FS

The proposed Fuzzy logic-enhanced Feature Selection method (FLFS) combines the outputs of six well-known feature selection methods from three feature selection

Table 1 Main categories of the clinical evaluation data considered in this study

Category	Description
Medical history	Medications and health histories based on questionnaire results (not including medical imaging outcomes)
Symptoms	Arthritis symptoms or health-related disability and function based on questionnaire data
Subject characteristics	Includes variables which describe anthropometric parameters and personal information
Nutrition	Questionnaire based on block food frequency
Physical exam	Includes performance measures and knee and hand exams
Physical activity	Questionnaire results regarding living and leisure activities
Behavioral	Consists of variables which quantify the social behavior and the quality level of daily routine

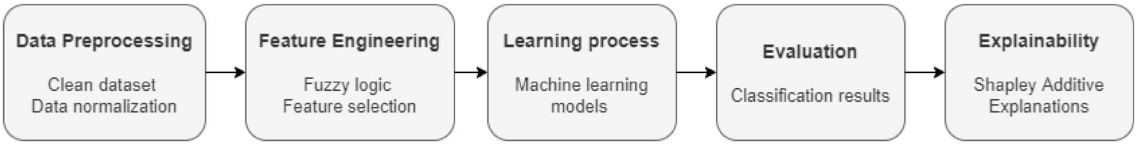
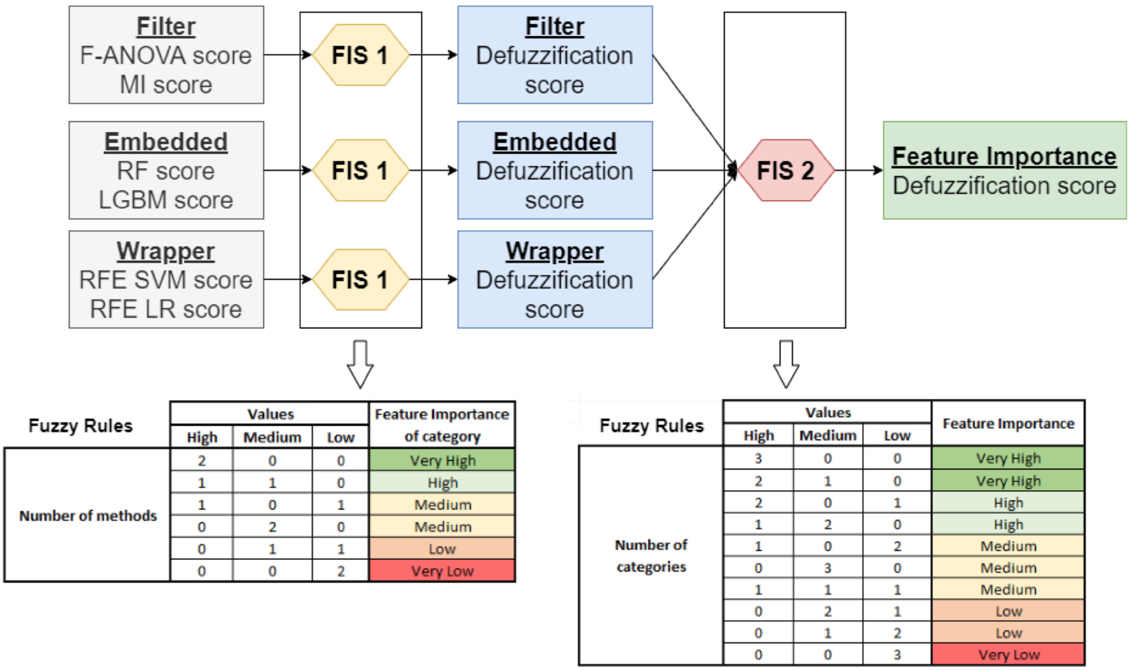


Fig. 1 The proposed AI methodology for KOA diagnosis



Fuzzy Rules

	Values			Feature Importance of category
	High	Medium	Low	
Number of methods	2	0	0	Very High
	1	1	0	High
	1	0	1	Medium
	0	2	0	Medium
	0	1	1	Low
	0	0	2	Very Low

Fuzzy Rules

	Values			Feature Importance
	High	Medium	Low	
Number of categories	3	0	0	Very High
	2	1	0	Very High
	2	0	1	High
	1	2	0	High
	1	0	2	Medium
	0	3	0	Medium
	1	1	1	Medium
	0	2	1	Low
	0	1	2	Low
	0	0	3	Very Low

Fig. 2 Feature Selection method based on Fuzzy logic flowchart

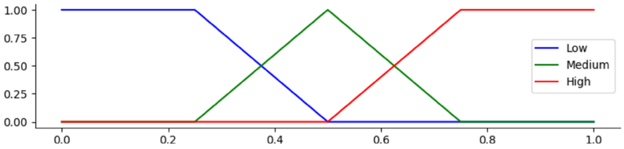


Fig. 3 Fuzzy set of input variables for FIS 1 and 2

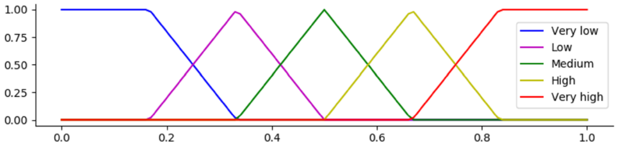


Fig. 4 Fuzzy set of output variable for FIS 1 and 2

categories (Filter, Wrapper and Embedded). Specifically, from the filter category, the mutual information [32] and the f-ANOVA [33] techniques were applied. From the wrapper category, we employed a recursive feature elimination (RFE) based on logistic regression [34] and an RFE based on support vector machine [35] techniques, respectively. Furthermore, from the embedded category, a LightGBM [36] and a random forest technique [37] were applied. To calculate the importance of a feature for each category, the scores of the associated FS techniques were used as input to the Fuzzy Inference System (FIS) 1 that was implemented with Mamdani inference methodology [38]. The output of the FIS 1 was the defuzzification value that represents the feature importance score for the specific feature selection category. Then, the defuzzification score of each category was used as input to the FIS 2 where the output defuzzification value represents the overall feature importance. Figure 2 illustrates the FSFL flowchart with the defined fuzzy rules for each FIS and the selected feature selection methods for this study. Figure 3 shows the fuzzy sets used in the presented methodology for the input variables for FIS 1 and FIS 2, while Fig. 4 shows the fuzzy sets of output variable for FIS 1 and 2.

Learning

In order to handle the demanding task of KOA classification, we investigated various ML models for their suitability and behavior in this problem, which are commonly used for medical applications. Specifically, random forest (RF) [39], multilayer perceptron (MLP) [40], logistic regression (LR) [41], support-vector machines (SVMs) [42], and k-nearest neighbors (KNN) [43] classifiers were tested. Furthermore, to avoid overfitting, and to optimize the performance of our models hyperparameter selection was applied individually per model.

Validation

For the experimental evaluation, a repeated stratified five-fold cross validation was used [44]. The performance of the classifiers was also evaluated in terms of the recall, f1-score and precision as additional evaluation criteria [45]. A brief description of these metrics is given below. Initially, the accuracy is the ratio of correctly predicted observations to the total observations and can be characterized as the most intuitive performance measure. Recall (or Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. Moreover, the ratio of correctly predicted positive observations to the total predicted positive observations is called precision or positive predictive value. F1-score is the weighted average of Precision and Recall.

Explainability

In the present work, we also examine how the risk factors have contributed to the final decision of KOA diagnosis. In order to achieve this, we used SHapley Additive exPlanations (SHAP), which is an approach to explain individual predictions based on Shapley Values of game theory and local explanations [46, 47]. In particular, we employed SHAP to rank features in terms of their impact on the final ML (Random Forest) outputs and to build a mini explainer model, which contributes to understanding the behavioral and the contribution of the risk factors in KOA diagnosis.

Results

In this section, we demonstrate the overall predictive performance of the models in relation to the first 100 selected features, and the highest metrics of the best models are also presented. Then, reference is made in the most important risk factors as they have been selected by the proposed Fuzzy FS methodology. Moreover, a comparative analysis is presented to prove the superiority of the proposed FS methodology compared to a number of well-known FS techniques. For the interpretation of the best model, an explainability analysis is employed to enhance our understanding of the reasoning behind its decision-making mechanism.

Predictive performance

This subsection presents the results of a comparative analysis over a number of well-known ML models on the diagnosis classification task by using the first 100 selected risk factors. Figure 5 shows the testing accuracy performance (%) of the competing ML models with respect to the number of selected features. Specifically, KNN failed in diagnosis task, recording low testing accuracy performances. The rest of the ML models had an upward trend in the range of the first 15 risk factors. Overall, the best overall performance was achieved by RF with a maximum of 73.55% at 21 features.

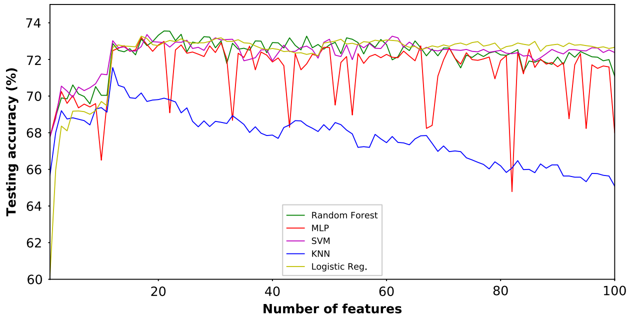


Fig. 5 Curves with testing accuracy scores with respect to the number of selected features for different ML models

Table 2 Summary of best metrics per model and number of selected features

Models	Accuracy	Precision	Recall	F1-Score	Num. of features
RF	73.55	73.82	73.64	73.59	21
MLP	73.20	73.48	73.20	73.13	17
LR	73.27	73.38	73.27	73.24	17
SVMs	73.36	73.68	73.36	73.27	18
KNN	71.55	71.74	71.55	71.49	12

Furthermore, the classification performance of the best performing ML models was further evaluated with respect to various validation metrics including class precision, recall, and f1-score. Table 2 demonstrates the best performance metrics of RF, MLP, LR, SVMs, and KNN models on the diagnosis task. In particular, RF achieved the best overall performance (73.55% accuracy) on the group of the twenty-one (21) risk factors. SVMs achieved the second-highest accuracy (73.36%). The rest of the ML models achieved lower accuracies.

Features selected

Figure 6 reveals more information about the origin of the 21 risk factors as selected by the chosen Fuzzy FS approach (refer to Appendix A for a detailed description of the selected features). As observed in Fig. 6, six features describing subject characteristics were among the selected risk factors e.g., the age of the participants, the body mass index (BMI), and the diastolic blood pressure. Moreover, five out of the 21 selected risk factors come from the symptom's category, representing clinical parameters related to stiffness, knee difficulty, swelling, and pain, demonstrating the indication of the existence of KOA. Four of the risk factors are related to physical exams, whereas another two medical history and two physical activity parameters were selected as relevant to KOA occurrence. A behavioural risk

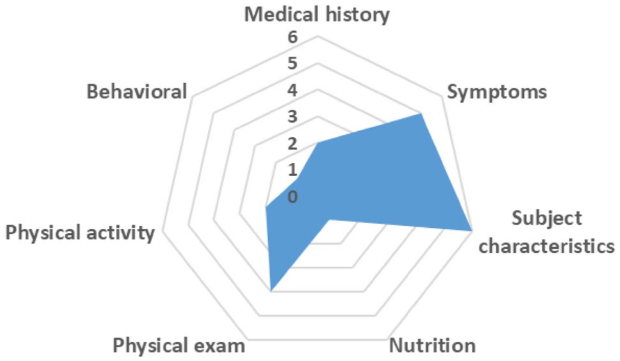


Fig. 6 The number of selected risk factors per category for the first 21 most informative features (a full description is given in Appendix A)

factor and a nutrition risk factor were also selected by the proposed Fuzzy FS approach.

Comparative analysis

The performance of the proposed FLFS methodology was compared with each one of the six FS techniques that were also implemented independently. Finally, another recently published FS technique was also selected as comparative in which the final feature ranking, is decided on the basis of a majority vote scheme [25, 26].

Table 3 shows the maximum achieved accuracy in the first selected 100 features of OAI dataset and the number of features where the maximum accuracy was reached for each feature selection method used in the experimental evaluation with the best performed model (RF). The last row in Table 3 shows the dimensionality reduction achieved with the proposed FS method compared to other competitive methods. Specifically, the metric DR was defined to quantify the difference (%) in dimensionality reduction compared to FSFL:

$$DR = 1 - \frac{Max\,number\,of\,features\,(FSFL)}{Max\,number\,of\,features\,(Compared\,method)} \quad (1)$$

The proposed FSFL method achieved the best trade-off between performance and dimensionality reduction being capable of reducing significantly the feature set dimensionality while achieving slightly higher or comparable prediction performance with the rest of the competing algorithms. Specifically, the proposed FSFL technique reaches the highest accuracy (73.55%) at 21 selected features while the second-best accuracy (73.51%) was achieved by LBGM Emb at 87 features. This shows that the proposed FSFL technique results to a 76% smaller set of selected features compared to the second-best performing technique. On the other hand, the second-best performer with respect to dimensionality reduction was RF Emb with 73.36% accuracy achieved on a considerably larger feature subset with more than double features (43) compared to FSFL (21).

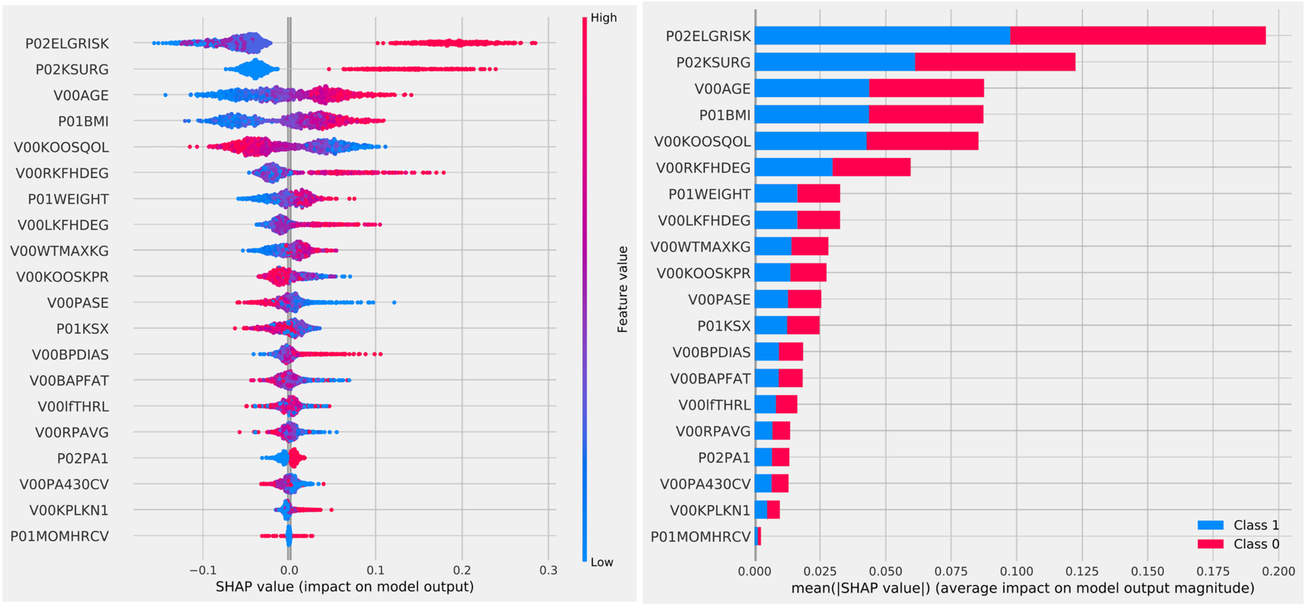
Explainability results

Figure 7a depicts how the features' impact shapes the output of the final model (RF) on the testing dataset. The features are sorted by the sum of SHAP value magnitudes over all testing subjects. Furthermore, the SHAP values are used to demonstrate the contribution of each risk factor (negative or positive) on the model's output. Specifically, blue color represents low feature values, whereas red color represents high values, respectively. In particular, a high value of PO2ELGRISK (knee symptoms, risk factors, or both status) increases the probability of the subjects to be assigned to class KOA. Similarly to PO2ELGRISK, the higher the values of risk factors V00AGE, P02KSRG, P01BM1, V00RKFHDEG, P01WEIGHT, V00LKFHDEG, V00WTMACKG, V00BRDIAS, V00KPLKN1, and P02PA1, the more probable for subjects to belong to class KOA. The rest of the selected risk factors in Fig. 7a have the opposite effect pushing the prediction output of the model to the class of healthy subjects. Figure 7b presents the SHAP global feature importance. The risk factors are sorted by the mean [|SHAP value|], which is the average impact on model output magnitude.

Figure 8 interprets locally the behavior of the model for the prediction output in a subject that suffers by KOA. P02ELGRISK (with a value of 2) and P01BMI (with a value of 29.8) push the predictions towards the class of KOA patients. Therefore, a high value of the aforementioned risk factors results to the increase of the output probability of the subject to be classified as KOA patient. On the contrary,

Table 3 Comparative analysis of FS methods

	FSFL	Vote FS	RF Emb FS	LGBM Emb FS	SVM RFE FS	LR RFE FS	Filter MI FS	Filter f-ANOVA FS
Maximum accuracy (%)	73.55	72.99	73.36	73.51	70.53	73.50	72.75	73.44
Number of selected features	21	76	43	87	96	60	91	53
DR (%)	–	+72%	+51%	+76%	+78%	+65%	+77%	+60%



(a)

(b)

Fig. 7 **a** Features' impact on random forest (21F) model output for the testing set of OAI dataset. **b** Features' average impact magnitude for testing instances

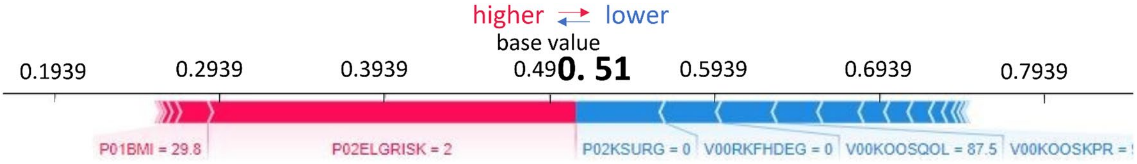


Fig. 8 Risk factors contributions to ML model output for a KOA status subject

increase of the risk factors P02KSURG, V00RKFHDEG, V00KOOSQOL, and V00KOOSKPR lowers the probability of a subject to be classified as KOA. Since, our prediction score = 0.51 > base value = 0.49, this subject has been positively classified, i.e., class KOA status.

Discussion

Handling the multidimensional nature of the OAI dataset, a novel fuzzy ensemble FS methodology was designed, implemented and tested in this paper. Its main novelty lies on the combination of several well-known FS algorithms based on a properly designed fuzzy inference mechanism that effectively aggregates their outputs. The superiority of the proposed FS technique was demonstrated through a thorough comparative investigation that included several state-of-the-art algorithms coming from different FS families (filter, wrapper, embedded and hybrid). Specifically, the proposed FS technique reached 73.55% accuracy at 21 features leading to the enhancement of the effectiveness of the initial voting scheme, while it contributes to the dimensionality reduction of OAI dataset compared to the competitive FS techniques.

Indeed, the proposed fuzzy FS methodology outperformed the aforementioned FS techniques achieving the best trade-off between dimensionality reduction and prediction accuracy. Working on a high-dimensional dataset of 643 features, twenty-one risk factors were selected for the objective of KOA diagnosis. Observing the nature of the selected risk factors, it was found that subject characteristics, symptoms, and physical exams are the most important risk factors contributing considerably to the KOA diagnosis. Overall, it was concluded that a combination of heterogeneous risk factors coming from different feature categories is needed for the effective diagnosis of KOA.

Some of the already published papers in the field of ML-based KOA diagnosis have reported higher classification accuracies (> 90% in some cases). This can be attributed to the fact that they have either utilized imaging data (e.g. X-ray) [16] or have focused on subgroups of patients (e.g.

elderly or overweight populations) [14, 48]. This study focuses on the application of an explainable ML pipeline for KOA diagnosis on non-imaging data making use of a novel Fuzzy FS algorithm followed by SoA ML classifiers and post-hoc SHAP analysis.

To sanity check the AI models beyond mere performance and further quantify the relevance of the selected risk factors, a post hoc explainability analysis was also conducted using SHAP. As observed by SHAP, P02ELGRISK, P02KSURG, age (V00AGE), BMI (P01BMI) and V00KOOSQOL are five risk factors that have a major impact to the prediction output, which are in line with the existing literature. Specifically, P02ELGRISK, that represents knee symptoms, is an important risk factor in the diagnosis of KOA, as it has been identified by Lespasio et al. [49]. The history of knee surgery (P02KSURG) has been recognised as an important risk factor of KOA by Katz et al. [50], whereas the age of the subjects was also characterized as crucial in the occurrence of KOA and therefore was considered in the development of a predictive model for KOA diagnosis [15]. The knee injury and osteoarthritis outcome (KOOS) is a well-known knee-specific instrument that has been widely employed to evaluate quality of life in patients with knee injuries and identify patients who are at risk of developing OA [51]. Moreover, high BMI is suggested to be a high-risk factor in the development of KOA. High BMI values lead to the increment of knee joint mechanical loading [52].

Although the proposed FSFL technique selects a subset of risk factors with a significant dimensionality reduction compared to popular FS techniques, the application of a post-hoc explainability is still important in order to identify the contribution of the selected features to prediction output of the model. The use of explainability analysis algorithms for the interpretation of the ML models increases the understanding of the principle of operation of each ML model and reveal the interactions that shape the diagnosis outcome.

The proposed methodology can be considered as computationally intensive; however, FS is considered here as an offline process and therefore the execution time does not play a crucial role. Future work will focus on the identification of easily measurable biomarkers and biomechanical parameters derived from musculoskeletal models, in combination with the already selected risk factors for the early diagnosis of KOA in the general population. Hence, to achieve this goal more advanced AI analytics tools in combination with the FSFL algorithm will be employed. Future work also includes the use of various regressors in the proposed ML pipeline to predict the KL grade and explain the factors that contribute to this prediction outcome. This methodology has been proven effective in solving the high-dimensional problem of KOA diagnosis. Modified versions of the proposed methodology have been also successfully applied to data problems with similar characteristics (JSN prediction [52] and KL progression [53, 54] and pain prediction [55]. Given its proven effectiveness, the methodology could be further updated and extended to solve classification or regression problems of high dimensionality in various domains (e.g. other knee injuries diagnosis or even prediction of COVID-19 infections).

Conclusion

To enforce the development of more reliable, powerful, and non-invasive diagnostic tools, this study focuses on the identification and interpretation of the risk factors that contribute to the diagnosis of KOA. The proposed methodology is based on a novel fuzzy logic-based feature selection followed by learning algorithms and subsequently a post-hoc explainability analysis. The proposed technique aggregates the results of several FS algorithms (filter, wrapper and embedded ones), whereas fuzzy logic was employed to combine multiple feature importance scores thus leading to a more robust selection of informative features. The results showed that the presented methodology was capable to select a subset of risk factors that increase the performance accuracy of various ML models, compared to popular FS techniques. This was achieved with a significant decrease on the feature dimensionality (up to 78%). SHAP was finally applied to enhance our understanding of the rationale behind the decision-making mechanism of the selected ML model and the impact of the used risk factors on the prediction output.

Appendix A

See Table 4.

Table 4 The 21 most informative selected risk factors as described in OAI database

Selected features	Description	Category
P02ELGRISK	Knee symptoms, risk factors, or both, status at IEI/SV	Symptoms
P01BMI	Body mass index	Subject characteristics
V00AGE	Age	Subject characteristics
P01WEIGHT	Average current scale weight (kg)	Subject characteristics

Selected features	Description	Category
V00LKFHDEG	Left knee exam: flexion contracture/hyperextension, degrees (contracture positive)	Physical exam
V00KOOSKPR	Right knee: KOOS Pain Score	Symptoms
P01MOMHRCV	Mother had hip replacement surgery	Medical history
P02PA1	Climb up total of 10 or more flights of stairs on most days	Physical activity
P01KSX	Frequent knee pain status by person	Symptoms
V00RKFHDEG	Right knee exam: flexion contracture/hyperextension, degrees (contracture positive)	Physical exam
V00WTMAXKG	Maximum adult weight, self-reported (kg)	Subject characteristics
P02KSURG	Either knee, history of knee surgery	Medical history
V00lfTHRL	Left Flexion MAX force high relaxation limit	Physical exam
V00BAPFAT	Block Brief 2000: daily % of calories from fat, alcoholic beverages excluded from denominator (kcal)	Nutrition
V00RPAVG	Radial pulse: average beats per minute	Subject characteristics
V00PASE	Physical Activity Scale for the Elderly (PASE) score	Physical activity
V00KOOSQOL	KOOS quality of life score	Symptoms
V00LFXCOMP	Isometric strength: left knee flexion, able to complete (3) measurements	Physical exam
V00BPDIAS	Blood pressure: diastolic (mm Hg)	Subject characteristics
V00PA430CV	How often lift or move objects weighing 25 pounds or more by hand during a typical week, past 30 days	Behavioral
V00KPLKN1	Left knee pain: twisting/pivoting on knee, last 7 days	Symptoms

Funding This research was funded by the European Community's H2020 Programme, under Grant agreement No. 777159 (OACTIVE).

Data availability Data from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/).

Declarations

Conflict of interest The authors declare no conflict of interest.

Institutional review board Not applicable.

Informed consent Not applicable.

Ethical approval Institutional Review Board Statement: Not applicable. The data are available upon request at https://nda.nih.gov/oai/.

References

1. Malanga G, Niazi F, Kidd VD et al (2020) Knee osteoarthritis treatment costs in the medicare patient population. Am Health Drug Benefits 13:144–153
2. Johnson VL, Hunter DJ (2014) The epidemiology of osteoarthritis. Best Pract Res Clin Rheumatol 28:5–15. https://doi.org/10.1016/j.berh.2014.01.004
3. Silverwood V, Blagojevic-Bucknall M, Jinks C et al (2015) Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis. Osteoarthritis Cartilage 23:507–515. https://doi.org/10.1016/j.joca.2014.11.019
4. Ackerman IN, Kemp JL, Crossley KM et al (2017) Hip and knee osteoarthritis affects younger people, too. J Orthop Sports Phys Ther 47:67–79. https://doi.org/10.2519/jospt.2017.7286
5. Toivanen AT, Heliövaara M, Impivaara O et al (2010) Obesity, physically demanding work and traumatic knee injury are major risk factors for knee osteoarthritis—a population-based study with a follow-up of 22 years. Rheumatology (Oxford) 49:308–314. https://doi.org/10.1093/rheumatology/kep388
6. London NJ, Miller LE, Block JE (2011) Clinical and economic consequences of the treatment gap in knee osteoarthritis management. Med Hypotheses 76:887–892. https://doi.org/10.1016/j.mehy.2011.02.044
7. Gupta S, Hawker GA, Laporte A et al (2005) The economic burden of disabling hip and knee osteoarthritis (OA) from the perspective of individuals living with this condition. Rheumatology 44:1531–1537. https://doi.org/10.1093/rheumatology/kei049
8. Mahir L, Belhaj K, Zahi S et al (2016) Impact of knee osteoarthritis on the quality of life. Ann Phys Rehabil Med 59:e159. https://doi.org/10.1016/j.rehab.2016.07.355
9. Farr Ii J, Miller LE, Block JE (2013) Quality of life in patients with knee osteoarthritis: a commentary on nonsurgical and surgical treatments. Open Orthop J 7:619–623. https://doi.org/10.2174/1874325001307010619
10. Kohn MD, Sassoon AA, Fernando ND (2016) Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop Relat Res 474:1886–1893. https://doi.org/10.1007/s11999-016-4732-4
11. Kokkotis C, Moustakidis S, Papageorgiou E et al (2020) Machine learning in knee osteoarthritis: a review. Osteoarthritis Cartilage Open 2:100069. https://doi.org/10.1016/j.ocarto.2020.100069
12. Yoo TK, Kim DW, Choi SB et al (2016) Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. PLoS ONE 11:e0148724. https://doi.org/10.1371/journal.pone.0148724
13. Lim J, Kim J, Cheon S (2019) A deep neural network-based method for early detection of osteoarthritis using statistical data. Int J Environ Res Public Health 16:E1281. https://doi.org/10.3390/ijerph16071281
14. Christodoulou E, Moustakidis S, Papandrianos N, et al (2019) Exploring deep learning capabilities in knee osteoarthritis case

study for classification. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). pp 1–6

15. Moustakidis S, Christodoulou E, Papageorgiou E et al (2019) Application of machine intelligence for osteoarthritis classification: a classical implementation and a quantum perspective. Quantum Mach Intell 1:73–86. https://doi.org/10.1007/s42484-019-00008-3
16. Kwon SB, Han H-S, Lee MC et al (2020) Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images. IEEE Access 8:120597–120603. https://doi.org/10.1109/ACCESS.2020.3006335
17. Moustakidis S, Papandrianos NI, Christodolou E et al (2020) Dense neural networks in knee osteoarthritis classification: a study on accuracy and fairness. Neural Comput Appl. https://doi.org/10.1007/s00521-020-05459-5
18. Hardi S, Triwiyono A, Amalia A (2020) Expert system for diagnosing osteoarthritis with fuzzy Tsukamoto method. J Phys Conf Ser 1641:012107. https://doi.org/10.1088/1742-6596/1641/1/012107
19. Pal JK, Ray SS, Pal SK (2017) Fuzzy mutual information based grouping and new fitness function for PSO in selection of miRNAs in cancer. Comput Biol Med 89:540–548. https://doi.org/10.1016/j.compbiomed.2017.08.013
20. Dai J, Chen J (2020) Feature selection via normative fuzzy information weight with application into tumor classification. Appl Soft Comput 92:106299. https://doi.org/10.1016/j.asoc.2020.106299
21. Lin Y, Hu Q, Liu J et al (2017) Streaming feature selection for multilabel learning based on fuzzy mutual information. IEEE Trans Fuzzy Syst 25:1491–1507. https://doi.org/10.1109/TFUZZ.2017.2735947
22. Jaganathan P, Kuppuchamy R (2013) A threshold fuzzy entropy based feature selection for medical database classification. Comput Biol Med 43:2222–2229. https://doi.org/10.1016/j.compbiomed.2013.10.016
23. Wang C, Qi Y, Shao M et al (2017) A fitting model for feature selection with fuzzy rough sets. IEEE Trans Fuzzy Syst 25:741–753. https://doi.org/10.1109/TFUZZ.2016.2574918
24. Qian Y, Wang Q, Cheng H et al (2015) Fuzzy-rough feature selection accelerator. Fuzzy Sets Syst 258:61–78. https://doi.org/10.1016/j.fss.2014.04.029
25. Ntakolia C, Kokkotis C, Moustakidis S, Tsaopoulos D (2021) Prediction of joint space narrowing progression in knee osteoarthritis patients. Diagnostics 11:285. https://doi.org/10.3390/diagnostics11020285
26. Kokkotis C, Moustakidis S, Giakas G, Tsaopoulos D (2020) Identification of risk factors and machine learning-based prediction models for knee osteoarthritis patients. Appl Sci 10:6797. https://doi.org/10.3390/app10196797
27. Jamshidi A, Leclercq M, Labbe A et al (2020) Identification of the most important features of knee osteoarthritis progressors using machine learning methods. Ann Rheum Dis 79:807–807. https://doi.org/10.1136/annrheumdis-2020-eular.1033
28. Shilaskar S, Ghatol A (2013) Feature selection for medical diagnosis: evaluation for cardiovascular diseases. Expert Syst Appl 40:4146–4153. https://doi.org/10.1016/j.eswa.2013.01.032
29. Silva-Ramírez E-L, Pino-Mejías R, López-Coello M, Cubiles-de-la-Vega M-D (2011) Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Netw 24:121–129. https://doi.org/10.1016/j.neunet.2010.09.008
30. Pihera J, Musliu N (2014) Application of machine learning to algorithm selection for TSP. In: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. pp 47–54
31. Japkowicz N (2000) Learning from imbalanced data sets: a comparison of various strategies. AAAI Press, Palo Alto, pp 10–15
32. Vergara JR, Estévez PA (2015) A review of feature selection methods based on mutual information. Neural Comput and Appl. https://doi.org/10.1007/s00521-013-1368-0
33. Suchwalko A, Buzalewicz I, Podbielska H (2013) Identification of bacteria species by using morphological and textural properties of bacterial colonies diffraction patterns. Proc SPIE Int Soc Opt Eng 8791:87911M. https://doi.org/10.1117/12.2020337
34. Zhao J, Karimzadeh M, Masjedi A, Wang T, Zhang X, Crawford MM, Ebert DS (2019) FeatureExplorer: interactive feature selection and exploration of regression models for hyperspectral images. In: 2019 IEEE Visualization Conference (VIS), pp 161–165
35. Ding J, Shi J, Wu F-X (2011) SVM-RFE based feature selection for tandem mass spectrum quality assessment. Int J Data Min Bioinform 5:73–88. https://doi.org/10.1504/ijdmb.2011.038578
36. Ye Y, Liu C, Zemiti N, Yang C (2019) Optimal feature selection for EMG-based finger force estimation using LightGBM model. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp 1–7
37. Mate Y, Somai N (2021) Hybrid feature selection and Bayesian optimization with machine learning for breast cancer prediction. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). pp 612–619
38. Gayathri BM, Sumathi CP (2015) Mamdani fuzzy inference system for breast cancer risk detection. In: 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). pp 1–6
39. Ram M, Najafi A, Shakeri MT (2017) Classification and biomarker genes selection for cancer gene expression data using random forest. Iran J Pathol 12:339–347
40. Parisi L, Biggs P, Whatling G, Holt C (2015) A novel comparison of artificial intelligence methods for diagnosing knee osteoarthritis. In: 25th Congress of the International Society of Biomechanics, Glasgow, United Kingdom
41. Ntakolia C, Kokkotis C, Moustakidis S, Tsaopoulos D (2020) A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). pp 934–941
42. Kubkaddi S, Ravikumar K (2017) Early detection of Knee Osteoarthritis using SVM Classifier. IJSEAT 5(3):259–262
43. Long NP, Park S, Anh NH et al (2019) Efficacy of integrating a novel 16-gene biomarker panel and intelligence classifiers for differential diagnosis of rheumatoid arthritis and osteoarthritis. J Clin Med 8:E50. https://doi.org/10.3390/jcm8010050
44. Wong T-T, Yeh P-Y (2020) Reliable accuracy estimates from k-fold cross validation. IEEE Trans Knowl Data Eng 32:1586–1594. https://doi.org/10.1109/TKDE.2019.2912815
45. Ghosh M, Sanyal G (2018) An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning. J Big Data 5:44. https://doi.org/10.1186/s40537-018-0152-5
46. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. Curran Associates, Inc
47. Nohara Y, Matsumoto K, Soejima H, Nakashima N (2019) Explanation of machine learning models using improved shapley additive explanation. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Association for Computing Machinery, New York, NY, USA, p 546
48. Lazzarini N, Runhaar J, Bay-Jensen AC et al (2017) A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. Osteoarthritis Cartilage 25:2014–2021. https://doi.org/10.1016/j.joca.2017.09.001

49. Lespasio MJ, Piuzzi NS, Husni ME et al (2017) Knee osteoarthritis: a primer. Perm J 21:16–183. https://doi.org/10.7812/TPP/16-183
50. Katz JN, Arant KR, Loeser RF (2021) Diagnosis and treatment of hip and knee osteoarthritis: a review. JAMA 325:568–578. https://doi.org/10.1001/jama.2020.22171
51. Roos EM, Lohmander LS (2003) The knee injury and osteoarthritis outcome score (KOOS): from joint injury to osteoarthritis. Health Qual Life Outcomes 1:64. https://doi.org/10.1186/1477-7525-1-64
52. Cooper C, Snow S, McAlindon TE et al (2000) Risk factors for the incidence and progression of radiographic knee osteoarthritis. Arthritis Rheum 43:995–1000. https://doi.org/10.1002/1529-0131(200005)43:5%3c995::AID-ANR6%3e3.0.CO;2-1
53. Tiulpin A, Klein S, Bierma-Zeinstra SMA et al (2019) Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci Rep 9:20038. https://doi.org/10.1038/s41598-019-56527-3
54. Kokkotis C, Moustakidis S, Baltzopoulos V et al (2021) Identifying robust risk factors for knee osteoarthritis progression: an evolutionary machine learning approach. Healthcare 9:260. https://doi.org/10.3390/healthcare9030260
55. Alexos A, Kokkotis C, Moustakidis S, et al (2020) Prediction of pain in knee osteoarthritis patients using machine learning: Data from Osteoarthritis Initiative. In: 2020 11th International Conference on Information, Intelligence, Systems and Applications IISA. pp 1–7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.