PROJECT DELIVERABLE REPORT

Project Title:
Advanced personalised, multi-scale computer models preventing osteoarthritis
SC1-PM-17-2017 - Personalised computer models and in-silico systems for well-being

Deliverable number	D6.5
Deliverable title	Design and implementation of personalised predictive models
Submission month of deliverable	M42
Issuing partner	CERTH
Contributing partners	LJMU
Dissemination Level (PU/PP/RE/CO):	PU
Project coordinator	University of Nicosia (UNIC)
Tel:	**+357 22 841 528**
Fax:	**+357 22 357481**
Email:	felekkis.k@unic.ac.cy & giannaki.c@unic.ac.cy
Project web site address	www.oactive.eu

Revision History

Version	Date	Responsible	Description/Remarks/Reason for changes
1.0		CERTH, LJMU	First Draft
1.1		CERTH, LJMU	Review of First Draft
1.2		CERTH, LJMU	Review of Second Draft

Contents

1. Summary

The first aim of this deliverable (Task 6.5) is to present the working prototypes of the personalised predictive OACTIVE models used either for prevention, diagnosis or even during the intervention stage.

The rest of this deliverable is organised as follows. Section 2 gives an introduction about the OACTIVEs targets and the approaches, which are presented in this Deliverable (D6.5). In Section 3, the proposed methodologies about the prediction of Knee Osteoarthritis (KOA) progression are presented. Diagnosis approaches are given in Section 4. Personalized interpretable models are provided in Section 5. Conclusions and future work are finally drawn in Section 6.

This report refers to Deliverable 6.5, which relates to the OACTIVE WP 6, "Hyper-modelling framework empowered by big data and deep learning" led by CERTH. The objective of WP6 is to develop the hyper-modelling framework of OACTIVE which will include: 1) data management mechanisms to ensure a high level of data quality and accessibility for the big data analytics applications 2) development of data pre-processing algorithms to improve data quality and consequently facilitate the efficiency of the data mining task, 3) development of data mining techniques for knowledge discovery, 4) development of the ICT deep learning infrastructure, 5) design and implementation of personalized predictive models, 6) an ontology-based framework for data standardization and 7) mechanisms for increased privacy and security.

2. Introduction

OACTIVE prioritises the development of a number of computational efficient 'local' predictive/diagnostic models that address specific OA stages in the disease continuum of a patient. Advanced pattern recognition models will be employed to model the KOA disease onset and further progression. The training process of the models will be based on the significant risk factors recognized on the previous evaluation analysis (Identify step). The outcome will be the generation of different local personalised decision models for diagnosis and prediction of KOA progression. Various classification models generated in the previous task (such as logistic regression, decisions trees, support vector machines, and deep learning neural networks) investigated for their appropriateness in providing accurate and robust decisions. The best model will be selected to accomplish the complex problem of KOA diagnosis and severity assessment. Moreover, to analytically assess the information content of each risk factor family, the proposed model will be separately applied on every feature family. The final outcome (diagnosis) will be derived by applying fusion techniques on the individual decisions allowing bio-medical researchers to investigate the influence of environmental factors on OA occurrence and their interactions with other health factors.

KOA is a multifactorial disease that causes low quality of life, poor psychology and resignation from life. Furthermore, KOA is a big data problem in terms of data complexity, heterogeneity and size as it has been commonly considered in the literature with most of the reported studies being limited in the amount of information they can adequately process. In this Deliverable, to cope with prediction of KOA progression in the first work, we propose a methodology (i) To provide a robust feature selection (FS) approach that could identify important risk factors which contribute to the prediction of KOA (Deliverable 6.3) and (ii) to develop machine learning (ML) prediction models for KOA. The current work considers multidisciplinary data from the osteoarthritis initiative (OAI) database, the available features of which come from heterogeneous sources such as questionnaire data, physical activity indexes, self-reported data about joint symptoms, disability and function as well as general health and physical exams' data. The novelty of the proposed FS methodology lies on the combination of different well-known approaches including filter, wrapper and embedded techniques, whereas feature ranking is decided on the basis of a majority vote scheme to avoid bias. The validation of the selected factors was performed in data subgroups employing seven well-known classifiers in five different approaches.

In the second approach the main goal is to build a prognostic tool that will predict the progression of pain in KOA patients using data collected at baseline. For this task we investigated two different methodologies. Initially, in the first work we leverage a feature importance voting system (Deliverable 6.3) for identifying the most important risk factors and various machine learning algorithms to classify, whether a patient's pain with KOA, will stabilize, increase or decrease. These models have been implemented on different combinations of feature subsets. The proposed methodology demonstrated unique potential in identifying pain progression at an early stage therefore improving future KOA prevention efforts. In the second work, the proposed methodology relies on an innovative evolutionary ML methodology capable of achieving state-of-the-art accuracy results. The prediction task is decomposed into local binary classification problems, which are treated separately with tailored ML models trained on selected feature subsets, whereas the final prediction is derived by fusing the outputs of these local models. The nature of the selected risk factors is discussed and the superiority of the proposed methodology is finally demonstrated compared to well-known ML algorithms.

Furthermore, in the third task, two approaches are presented to predict the progression of knee joint space narrowing (JSN) in each knee and in both knees combined. A machine learning approach is proposed with the use of multidisciplinary data from the osteoarthritis initiative database. The proposed methodology employs: (i) A clustering process to identify groups of people with progressing and non-progressing JSN (Deliverable 6.3) (ii) a robust feature selection (FS) process consisting of filter, wrapper, and embedded techniques that identifies the most informative risk factors (Deliverable 6.3) and (iii) a decision making process based on the evaluation and comparison of various classification algorithms towards the selection and development of the final predictive model for JSN.

Finally, we worked to increase the generalization of the personalized prediction models. Specifically, this work contributes to the identification of risk factors for KOA progression via a robust feature selection (FS) methodology that overcomes two crucial challenges: (i) the observed high dimensionality and heterogeneity of the available data that are obtained from the Osteoarthritis Initiative (OAI) database and (ii) a severe class imbalance problem posed by the fact that the KOA progressors class is significantly smaller than the non-progressors' class. The proposed feature selection methodology relies on a combination of evolutionary algorithms and machine learning (ML) models We investigated the effectiveness of the proposed approach in a comparative analysis with well-known FS techniques with respect to metrics related to both prediction accuracy and generalization capability. The proposed FS methodology may contribute to the development of new, efficient risk stratification strategies and identification of risk phenotypes of each KOA patient to enable appropriate interventions.

On the other hand, to develop personalised diagnosis models, we worked on two different approaches. Initially, the aim of the first work is to provide a data mining approach that could identify important risk factors which contribute to the diagnosis of KOA. Data were obtained from the osteoarthritis initiative (OAI) database enrolling people, with non-symptomatic KOA and symptomatic KOA or being at high risk of developing KOA. The current work considered multidisciplinary data from heterogeneous sources such as questionnaire data, physical activity indexes, self-reported data about joint symptoms, disability and function as well as general health and physical exams' data from individuals with or without KOA from the baseline visit. For the data mining part, a robust feature selection methodology was employed consisting of filter, wrapper and embedded techniques whereas feature ranking was decided on the basis of a majority vote scheme (Deliverable 6.3). The results are the basis for the development of easy-to-use diagnostic tools for clinicians for the early detection of KOA.

Furthermore, KOA is the most common form of arthritis in the knee that comes with a variation in symptoms' intensity, frequency and pattern. Knee OA (KOA) is often diagnosed using invasive and expensive methods that can measure changes in joint morphology and function. Early and accurate

identification of significant risk factors in clinical data is of vital importance in diagnosing KOA. A machine intelligence approach is proposed here to enable automated, non-invasive identification of risk factors from self-reported clinical data about joint symptoms, disability, function and general health. The proposed methodology was applied to recognize participants with symptomatic KOA or being at high risk of developing KOA in at least one knee. Different machine learning and deep learning algorithms were tested and compared in terms of multiple criteria e.g., accuracy, per class accuracy and execution time. Deep learning was proved to be the most effective in terms of accuracy with classification accuracies up to 86.95%, evaluated on data from the osteoarthritis initiative study. Insights about ten different feature subsets and their effect on classification accuracy are provided. The proposed methodology was also demonstrated in subgroups defined by gender and age. The results suggest that machine intelligence and especially deep learning may facilitate clinical evaluation, monitoring and even prediction of knee osteoarthritis. Apart from the classical implementation of the proposed methodology, a quantum perspective is also discussed highlighting the future application of quantum computers in KOA diagnosis.

Interpretable models provide decisions which are made with clarity and that the processes that go into making decisions about a person's health are easily explainable to the patient and understood by doctors. For this task insights about the baseline presentation with and without clinical manifestation of osteoarthritis were derived from statistical and machine learning models. This produced models for diagnosis of KOA (radiological index KL 2+) and prognostic model for propensity to develop KOA within 5 years of first presentation. Both models are displayed as app interfaces with nomograms for transparency and to enable scenario analysis for the effects of modifiable risk factors.

3. Personalised Prediction of KOA progression

Knee Osteoarthritis (KOA) is the most common type compared with other types of osteoarthritis (OA). KOA results from a complex interplay of constitutional and mechanical factors, including mechanical forces, local inflammation, joint integrity, biochemical processes and genetic predisposition. The specific disease causes significant problems when it occurs. In recent years, it has been also realized that KOA is closely associated with obesity and age [1]. Moreover, KOA is diagnosed in the young and athletes following older injuries [2]. The particularity of this disease is that the knee osteoarthritic process is gradual with a variation in symptoms intensity, frequency and pattern [3]. Due to the multifactorial nature of KOA, disease pathophysiology is still poorly understood and prognosis prediction tools are under current investigation.
Prognosis and treatment of KOA is a challenge for the scientific community. Increasing data collection has led to an increasing number of studies employing big data and AI analytics applied in the KOA research. As a result of this, several techniques have been reported in the literature in which ML models were used to predict KOA [4]. In 2017, Lazzarini et al. developed five (5) ML models that can be used to predict the incidence of knee OA in overweight and obese women. By integrating a wide variety of biomedical data in their models, they showed that using a small subset of the available information is possible to accurately predict the incidence of KOA by using Random Forest (RF) [5]. In another study, Halilaj et al. aimed to characterize different clusters of KOA progression and build models to predict these clusters early [6]. LASSO regression models were used to predict joint space narrowing and pain progression which are the most widely used surrogates of structural and symptomatic disease status. Furthermore, Pedoia et al. [7] used MRI and multidimensional biomechanics data attempting to meet the existing gap in multidimensional data analysis for precision medicine in KOA. They achieved large-scale integration of compositional imaging and skeletal biomechanics by using logistic regression as the ML model.

In 2019, Abedin et al. built two different prediction models, which achieved comparable accuracy with the aforementioned studies. In this study elastic net and RF were used along with a convolution neural network.

The aim of this work was to explore whether the prediction accuracy of a statistical model based on the patient's questionnaire data is comparable to the prediction accuracy based on X-ray image-based modeling to predict KOA severity [8]. In another study, in 2019 Nelson et al. applied innovative ML approaches (e.g., K- means, t-SNE), specialized for a high dimension, low sample size setting, to phenotyping in KOA in order to better define progression phenotypes that may be more homogeneous and responsive to potential disease modifying interventions [9]. Moreover, in 2019 Tiulpin et al. proposed a novel method based on ML that directly utilizes raw radiographic data, physical examination, patient's medical history, anthropometric data and, optionally, a radiologist's statement (Kellgren and Lawrence (KL)-grade) to predict structural KOA progression by using logistic regression and gradient boosting machine. They demonstrated that a knee X-ray image alone is already a very powerful source of data to predict whether a particular knee will have OA progression or not [10]. Futhermore, in the same year, Widera et al. used several ML models (e.g., logistic regression, K-nearest neighbor, SVC (linear kernel), SVC (RBF kernel) and RF) in combination with clinical data and X-ray image assessment metrics to develop predictive models for patient selection that outperform the conventional inclusion criteria used in clinical trials [11]. However, few studies have tried to apply ML models for the prediction of KOA. There is still a lack of knowledge on the contribution of self-reported clinical data on the KOA prognosis and their impact on the training of the associated ML predictive models [12–17].

According to our knowledge, identification of risk factors for developing and especially predicting KOA has been limited by an absence of non-invasive methods to inform clinical decision making and enable early detection of people who are most likely to progress to severe KOA. Hence, we worked on four approaches: (i) personalized prediction based on risk factors that are relevant with Kellgren-Lawrence (KL) progression, (ii) personalized prediction based on risk factors that are relevant with pain progression, (iii) personalized prediction based on risk factors that are relevant with joint space narrowing (JSN)) progression and (iv) increasing generalization using an evolutionary Machine Learning approach based on risk factors that are relevant with Kellgren-Lawrence (KL) progression. In the first approach, the main purpose is twofold: (i) The prediction of KOA through the identification of risk factors that are relevant with KL progression from a big pool of risk factors available in the osteoarthritis initiative (OAI) database and (ii) the development of machine learning-based models that can predict long-term KL progression. To accomplish the aforementioned targets, a robust ML pipeline that involves a hybrid feature selection technique and well-known ML models was implemented. Moreover, this work also explores three different options with respect to the time period within which data should be considered in order to reliably predict KOA progression. Finally, a discussion on the nature of the selected features is also provided.

Subsequently, to cope with the purpose of the second task we investigated two approaches. The aim of the first approach is: (i) to identify different clusters of KOA pain progression, (ii) to identify informative parameters that are relevant with pain progression from a big pool of risk factors that are available in osteoarthritis initiative (OAI) database and (iii) to build ML models that can predict long-term pain progression using baseline data. To accomplish the aforementioned targets, we built a ML empowered methodology capable of achieving state-of-the-art accuracy results with the minimum possible number of features. By using a relatively small number of features, and at the same time not sacrificing test set performance, we can run the algorithm faster at inference time, and implement it in portable devices (e.g., a smartphone). The dataset, as described has 726 features. By reducing this number to a relatively small number features, for instance 25 we could create more possibilities for the implementation of an algorithm in a small mobile device; or test the algorithm faster in the subjects by requiring less computational power. In order to do this, we have developed a hybrid technique, in which we derive the feature importance from different Feature Selection (FS) algorithms via a common voting system. Afterwards, we explored the suitability of different ML algorithms in an extensive comparative experimentation, to distinguish the one that produces the best results for our prediction. The aim of the other one is: (i) the identification of

different clusters of KOA pain progression, (ii) the selection of informative and robust parameters that are relevant with pain progression and (iii) the development of AI-powered predictive models that could be used for patient-specific prediction of pain progression. To accomplish the aforementioned targets, we rely on an innovative evolutionary ML methodology capable of achieving state-of-the-art accuracy results. One of the novelties of the proposed methodology is that the prediction task has been decomposed into local binary classification problems. Each of the local problems is treated separately (with custom ML models trained on selected feature subsets) and the final prediction is derived by fusing the outputs of these local models. The nature of the selected risk factors is discussed and the superiority of the proposed methodology over well-known ML algorithms is also demonstrated.

The third approach aims towards the accurate prediction of JSN on Medial compartment (JSM) progression via the development of a novel machine learning approach. This ML approach handles the heterogeneity among a plethora of features (725) deriving from various feature categories, including diagnosis from medical examination and medical imaging outcomes, among others. In this work the effectiveness of two strategies is investigated for predicting the JSN progression of KOA patients by: (i) Developing predictive models that are trained on data from the left knee and right knee separately and (ii) developing predictive models that combine KOA patients' data for both the right and left knee. For each strategy the same steps were followed. Initially, a clustering approach is applied for the identification of patients groups with and without JSN progression. Then the risk factors are identified based on a voting scheme that incorporates various categories of feature selection techniques. The prediction stage was implemented with the use of well-known ML models in an extensive comparative experimentation. End, the fourth work increasing generalization using an evolutionary Machine Learning approach based on risk factors that are relevant with Kellgren-Lawrence (KL) progression. Hence, we propose an FS technique that incorporates a number of characteristics towards the identification of robust risk factors that generalize well over the whole dataset. The proposed FS methodology, termed GenWrapper in this work, is an evolutionary genetic algorithm (GA)-based wrapper technique that differentiates from the classical GA-based FS techniques in terms of the following: (i) GenWrapper applies random under-sampling at each individual solution, forcing the GA to converge to solutions (feature subsets) that generalize well regardless of the applied data sampling; (ii) It ranks features with respect to the number of times that they have been selected in all the individual solutions for the final population. The combined effect of the aforementioned GenWrapper characteristics leads to selected features that consistently work well at any possible data sample and, thus, have increased generalization capacity with respect to KOA progression. An extensive comparative analysis has been performed to prove the superiority of GenWrapper over well-known FS algorithms with respect to both prediction accuracy and generalization.

3.1 Personalised Prediction of KL progression

Data description

Data were obtained from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/). Specifically, the current work only includes clinical data from: (i) The baseline; (ii) the first follow up visit at month 12 and (iii) the next follow up visit at month 24 from all individuals being at high risk to develop KOA or without KOA. Eight feature categories were considered as possible risk factors for the prediction of KL as shown in Table 1. Furthermore, our work was based on Kellgren and Lawrence (KL) grade as the main indicator for assessing the clinical status of the participants. Specifically, the variables 'V99ERXIOA' and 'V99ELXIOA' were used to assign participants into subgroups (classes) of participants whose KOA status progresses or not (during labelling process).

Table 1. *Main categories of the feature subsets considered in this work.*

		Timeline of Visit		
Category	**Description**	**Baseline**	**12 Months**	**24 Months**
Subject characteristics	Anthropometric parameters including height, weight, BMI, abdominal circumference, etc.	●	●	●
Behavioural	Participants' social behaviour and quality level of daily routine	●	●	●
Medical history	Questionnaire data regarding a Participant's arthritis-related and general health histories and medications	●	-	-
Medical imaging outcome	Medical imaging outcomes (e.g., osteophytes and joint space narrowing)	●	-	-
Nutrition	Block Food Frequency questionnaire	●	-	-
Physical activity	Questionnaire results regarding leisure activities, etc.	●	●	●
Physical exam	Physical measurements of participants, including isometric strength, knee and hand exams, walking tests and other performance measures	●	●	●
Symptoms	Arthritis symptoms and general arthritis or health-related function and disability	●	●	●

As described in Deliverable 6.3, we consider KL grades prediction as a two-class classification problem. Specifically, the participants of the study were divided into two groups:

- Non-progressors: Healthy participants (KL grade 0 or 1) that remained healthy throughout the whole duration of the OAI study (eight years) and
- (2) KOA progressors: Healthy participants who developed OA (KL > 1) during the curse of the OAI study.

Hence, the main objective of this work is to build ML models that could discriminate the two aforementioned groups and therefore be able to decide whether a new testing sample (healthy participant) will develop OA (assigned in the progressors' class) or not (assigned to the non-progressors' class). Secondary objectives of the work are to: (i) Identify which of the available risk factors contribute more to the classification output and as result can be considered as contributing factors in the prediction of OA and (ii) explore three different options (a single visit, two visits within a year and two visits within two years) with respect to the time period within which data should be considered in order to reliably predict KOA progression. To achieve these targets, we have worked on five different approaches in which different data subsets were considered comprising features from the baseline combined (or not) with features from visits 1 (at month 12) and 2 (month 24). The motivation behind this is to investigate whether data from the baseline are sufficient to predict the progression of KOA or additional data from subsequent visits should be also included in the training to increase the predictive accuracy of the proposed techniques. Data resampling was applied at each of the five datasets to cope with the problem of class size imbalance and generate dataset in which classes are represented by an equal number of samples.

A short overview through visualization of the aforementioned data subsets investigated in this research work is given in what follows.

- Dataset A (FS1): Progressors vs non-progressors using data from the baseline visit. This dataset only contains data from the baseline (724 features). After data resampling, the participants were divided into two equal categories (Figure 1), as follows:

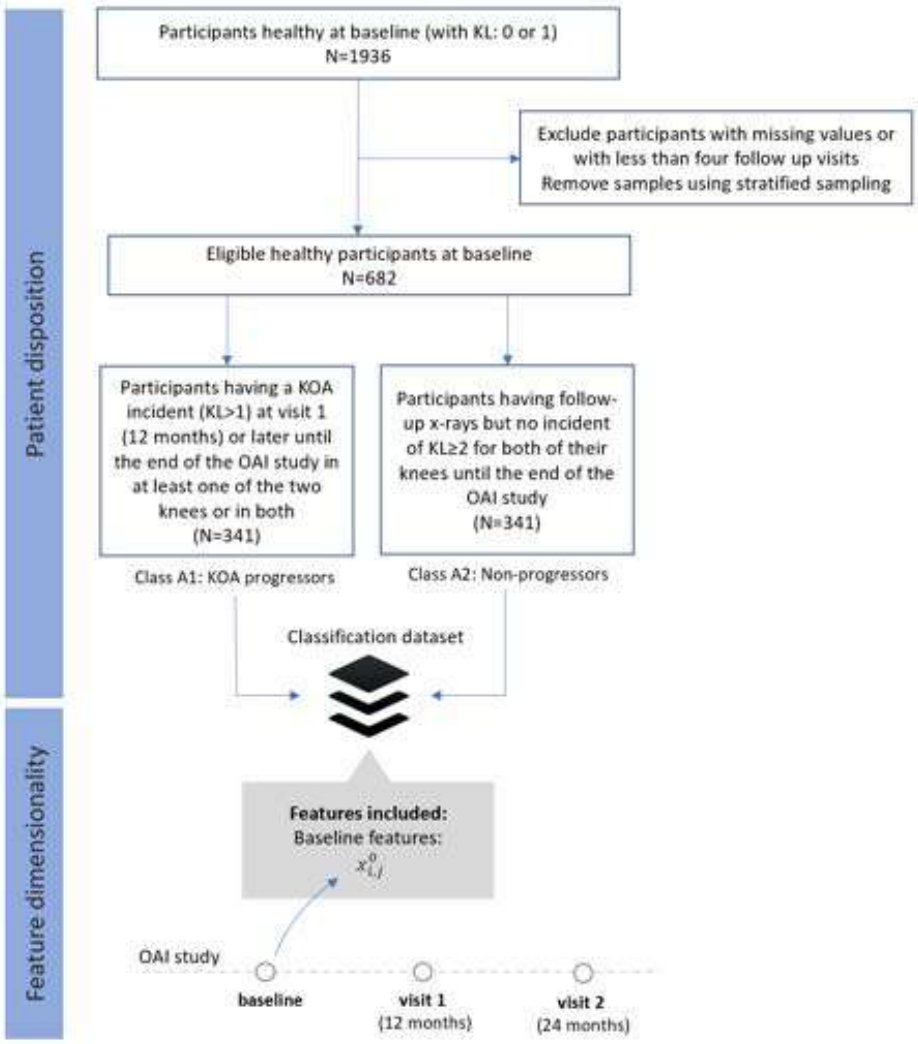


Figure 1. *Flow chart of study design for dataset A.*

- Dataset B (FS2): Progressors vs non-progressors using progression data within the first 12 months. Dataset B contains data that declares the features' progression within the first 12 months. Specifically, after data resampling, the following two classes of participants were created (Figure 2), as follows:

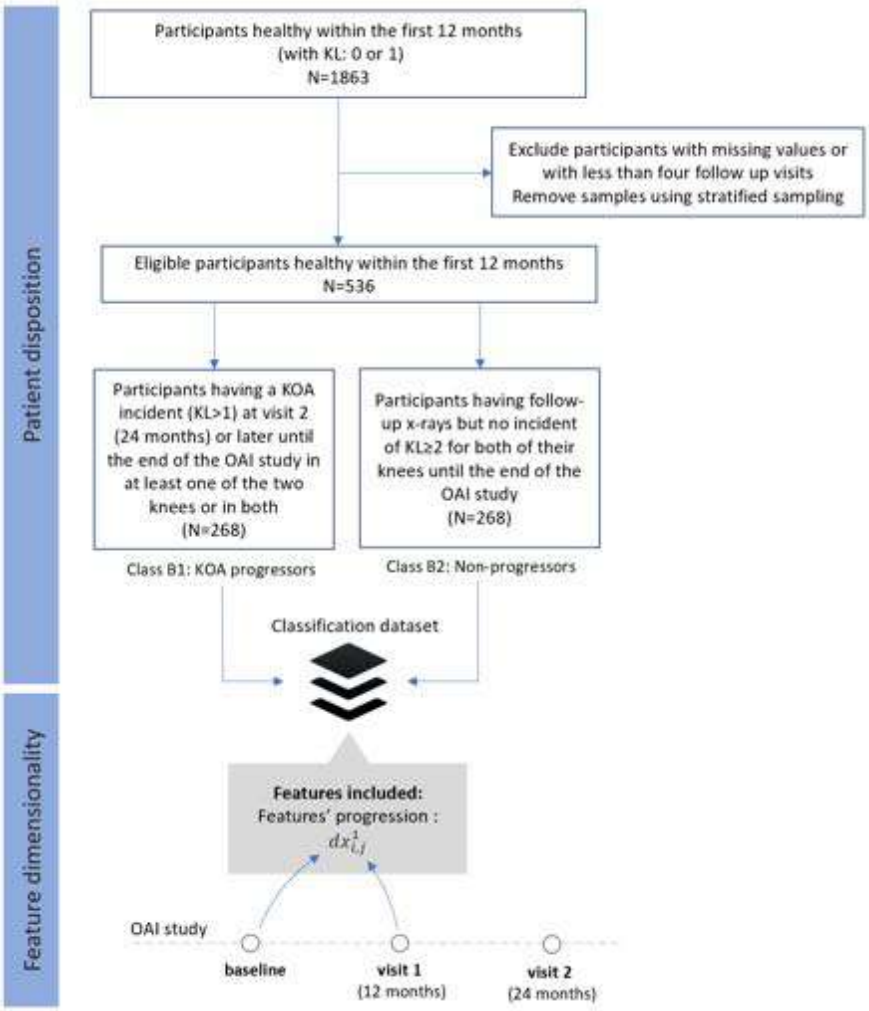


Figure 2. *Flow chart of study design for dataset B.*

- Dataset C (FS3): Progressors vs non-progressors using progression data within the first 24 months. Dataset C contains data that declares the features' progression within the first 24 months (until visit 2). The participants were divided into two equal categories (Figure 3), as follows:

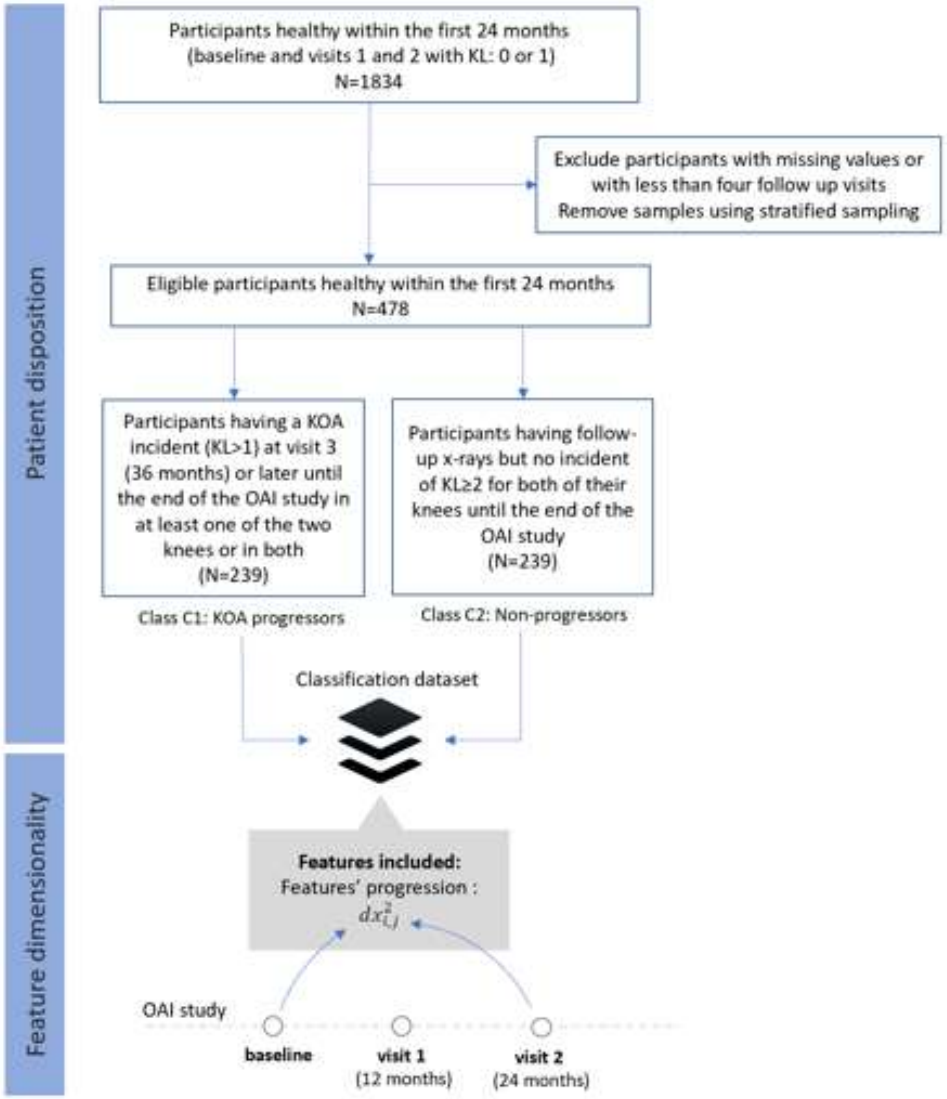


Figure 3. *Flow chart of study design for dataset C.*

- Dataset D (FS4): Progressors vs non-progressors using data from the baseline visit along with progression data within the first 12 months. After the application of data sampling, the participants were divided into two equal categories (Figure 4), as follows:

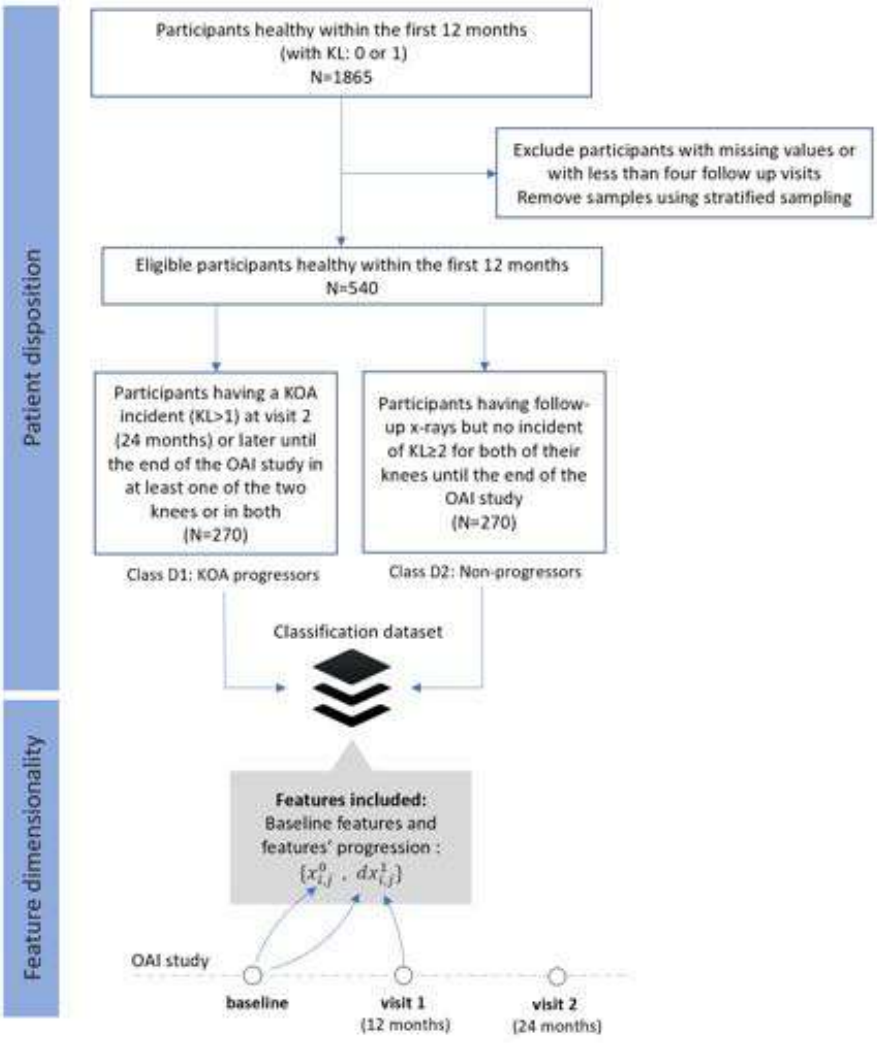


Figure 4. *Flow chart of study design for dataset D.*

- Dataset E (FS5): Progressors vs non-progressors using data from the baseline visit along with progression data within the first 24 months. Similarly, participants were divided into two equal categories (Figure 5), as follows:

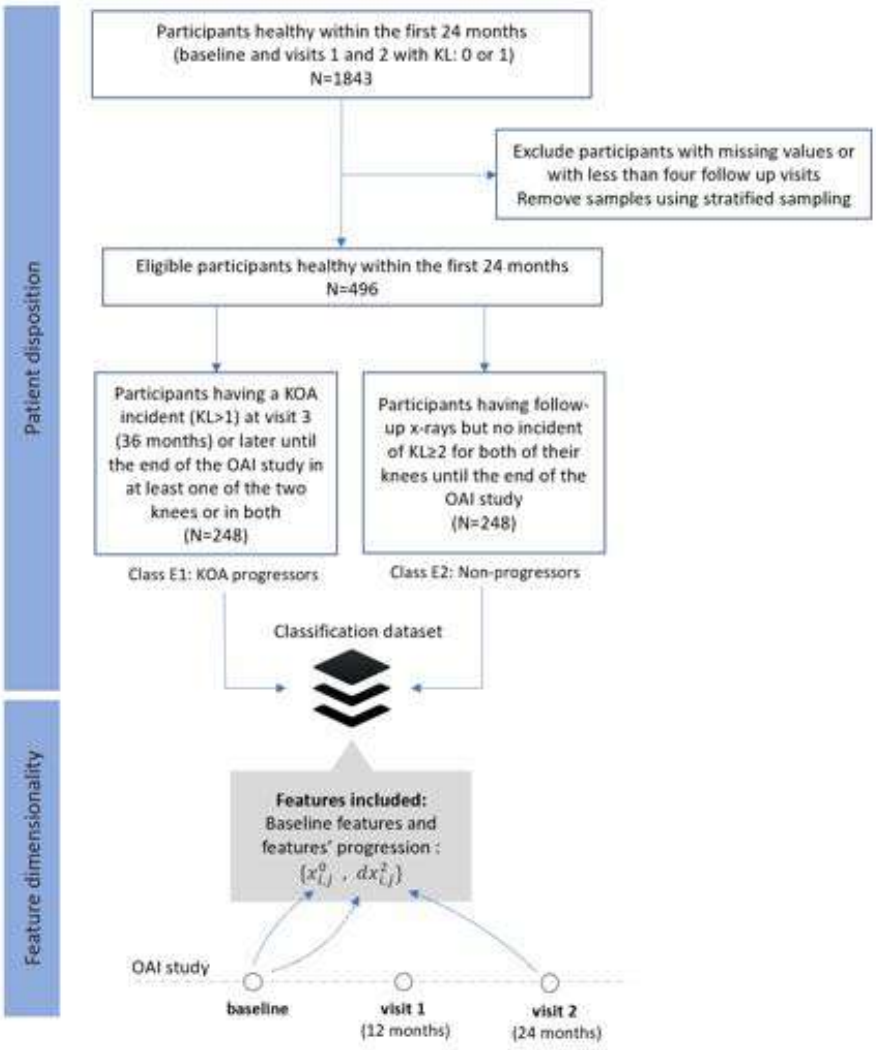


Figure 5. *Flow chart of study design for dataset E.*

Methodology

The proposed in this work ML methodology for KOA prediction includes four processing steps: (1) data pre-processing of the collected clinical data (Deliverable 6.3), (2) feature selection using the proposed approach (Deliverable 6.3), (3) learning process via the use of well-known ML models and (4) evaluation of the classification results. More details about the proposed methodology are presented in the following sections.

Pre-Processing and Feature Selection (FS)

The steps of the proposed methodology have described in Deliverable 6.3.

Learning Process

Various ML models were evaluated for their suitability in the task of KOA prediction. A brief description of these models is given below.

We tested logistic regression [18] which is likely the most commonly used algorithm for solving classification problems. Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. The interpretation of the weights in logistic regression differs from the interpretation of the weights in linear regression, since the outcome in logistic regression is a probability between 0 and 1. We also evaluated

decision trees (DTs) [19] which are a non-parametric supervised learning method used for classification and regression. They are simple to understand and to interpret. DTs require little data preparation and perform well even if their assumptions are somewhat violated by the true model from which the data were generated.

K-Nearest Neighbor (KNN) [20] as well as non-linear support vector machines (SVM) algorithms [21], which can deal with the overfitting problems that appear in high-dimensional spaces. In the classification setting, the KNN algorithm essentially boils down to forming a majority vote between the K most similar instances to a given "unseen" observation. Similarity is defined according to a distance metric between two data points. A popular one is the Euclidean distance method. Furthermore, SVMs are a set of supervised learning methods used for classification, regression and outlier's detection. They are effective in high dimensional spaces and still effective in cases where the number of dimensions is greater than the number of samples.

The ensemble technique Random Forest (RF) [22] was also evaluated using DT models as weak learners. RF classifier creates a set of decision trees from randomly selected subsets of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. XGboost [23] and naive Bayes [24] algorithms were also considered. XGboost model is a sum of CART (tree) learners which try to minimize the log loss objective and the scores at leaves. These scores are actually the weights that have a meaning as a sum across all the trees of the model. Furthermore, they are always adjusted in order to minimize the loss. Moreover, naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Naive Bayes learners and classifiers can be extremely fast. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution.

Hyperparameter selection was implemented to optimize the performance of our models and to avoid overfitting and bias errors. Each model was optimized with respect to a number of preselected hyperparameters (Table 2). Specifically (i) 'gamma': [0,0.4,0.5,0.6], 'maximal depth': [1,2,3,4,5,6,7,8], 'minimum child and weight': [1,3,4,5,6,8] were optimized for XGboost, (ii) 'criterion': ['gini', 'entropy'], 'minimum samples leaf': [1,2,3], 'minimum samples split': [3,4,5,6,7] and 'number of estimators': [10,15,20,25,30] for random forest, (iii) 'maximal features': ['auto', 'sqrt', 'log2'], 'minimum samples leafs': [1,2,3,4,5,6,7,8,9,10,11] and 'minimum number of decision splits': [2,3,4,5,6,7,8,9,10,11,12,13,14,15] for decision trees, (iv) 'C': [0.001,0.01,0.1,1,2,3,4,5,6,7,8,9,10] and 'kernel': ['linear','sigmoid','rbf','poly'] for SVMs, (v) 'k-parameter': [5,7,9,12,14,15,16,17] for KNN and (vi) 'penalty': ['l1', 'l2'] and 'C': [100, 10, 1.0, 0.1, 0.01] for logistic regression.

Table 2. *Hyperparameters description.*

ML Models	Hyperparameters	Description
XGboost	Gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree.
	Maximal depth	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
	Minimum child and Weight	Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning.
Random Forest	Criterion	The function to measure the quality of a split.
	Minimum samples leaf	The minimum number of samples required to be at a leaf node.

Decision Trees	Number of estimators	The number of trees in the forest.
	Maximal features	The number of features to consider when looking for the best split.
	Minimum samples split	The minimum number of samples required to split an internal node
	Minimum number of leafs	The minimum number of samples required to be at a leaf node.
SVMs	C	Regularization parameter. The strength of the regularization is inversely proportional to C.
	Kernel	Specifies the kernel type to be used in the algorithm.
KNN	k-parameter	Number of neighbors to use by default for k neighbors queries.
Logistic Regression	Penalty	Used to specify the norm used in the penalization.
	C	Inverse of regularization strength; must be a positive float.

Validation

A hold out 70–30% random data split was applied to generate the training and testing subsets, respectively. Learning of the ML was performed on the stratified version of the training sets and the final performance was estimated on the testing sets.

Results

In this section, we present the most important risk factors as they have been selected by the proposed hybrid FS methodology. Moreover, the overall performance of the models is presented in relation to the number of selected features and then reference is made to the models with the highest accuracies. Results are initially given per dataset and an overall assessment is provided at the end. The efficacy of the proposed FS methodology is also compared with the performance of the six individual FS criteria.

Prediction Performance

The proposed ML methodology was applied on each of the five datasets. Specifically, the proposed FS was executed on the pre-processed versions of the datasets ranking the available features with respect to their relevance with the progression of OA. Then the proposed ML models were trained on feature subsets of increasing dimensionality (with a step of 5). These feature subsets were generated by sorting the features according to the selected ranking. This means that the proposed ML models were trained to classify KOA progressors and non-progressors based on the first (5, 10, 15, etc.) most informative features and the testing classification accuracies were finally calculated until the full feature set has been tested. The classification results on the five datasets are given below.

- Dataset A

Figure 6 depicts the testing performance (%) of the competing ML models with respect to the number of selected features for dataset A. In particular, DTs failed in this task, recording low testing performances (in the range of 42.44–65.85%). In contrast, the other models had an upward trend in the first 20–60 features, followed by a steady testing performance in most of the cases. Specifically, the logistic regression model showed an upward trend with respect to selected features in the first 30–50 features, with a maximum of 71.71% at 50 features (which was the overall best performer). The inclusion of additional features led to a small reduction in the accuracies achieved.

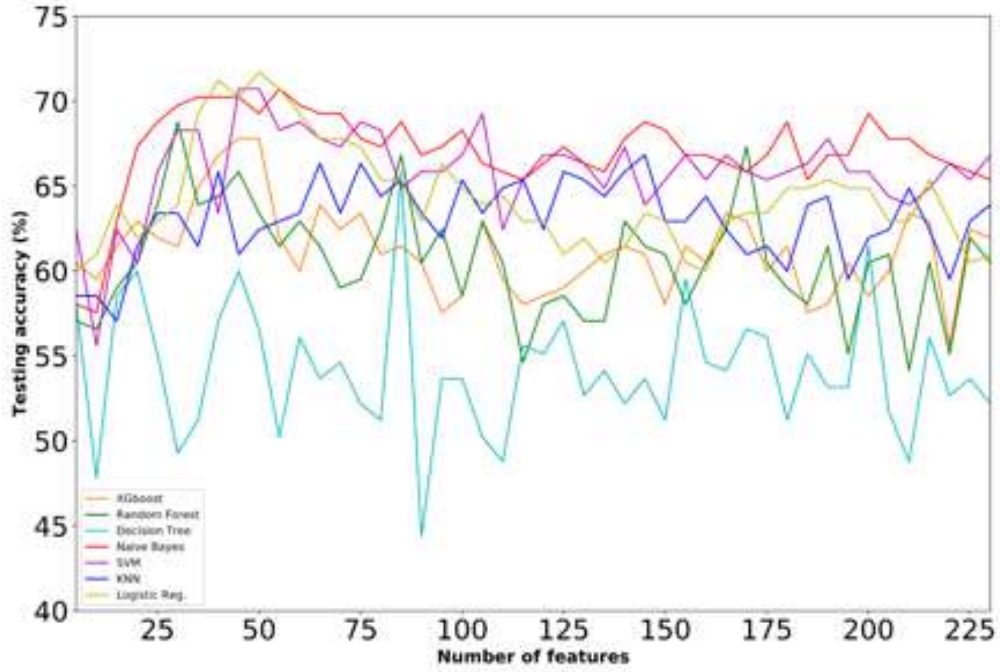


Figure 6. *Learning curves with testing accuracy scores on dataset A for different machine learning (ML) models trained on feature subsets of increasing dimensionality.*

- Dataset B

Figure 7 demonstrates the testing performance (%) of the competing ML models with respect to the number of selected features for dataset B. The following remarks could be extracted from Figure 8: (i) Considerably lower accuracies were achieved by all the competing ML models compared to the ones received in dataset A; (ii) LR and NB gave the maximum testing performance of approximately 64% at 25 features (which was the overall best performer in dataset B). The addition of more features did not increase the testing performance of the model but led to a reduction in the accuracies achieved. (iii) Low testing performances were accomplished by the rest of the ML models (in the range of 42.24–62.11%).

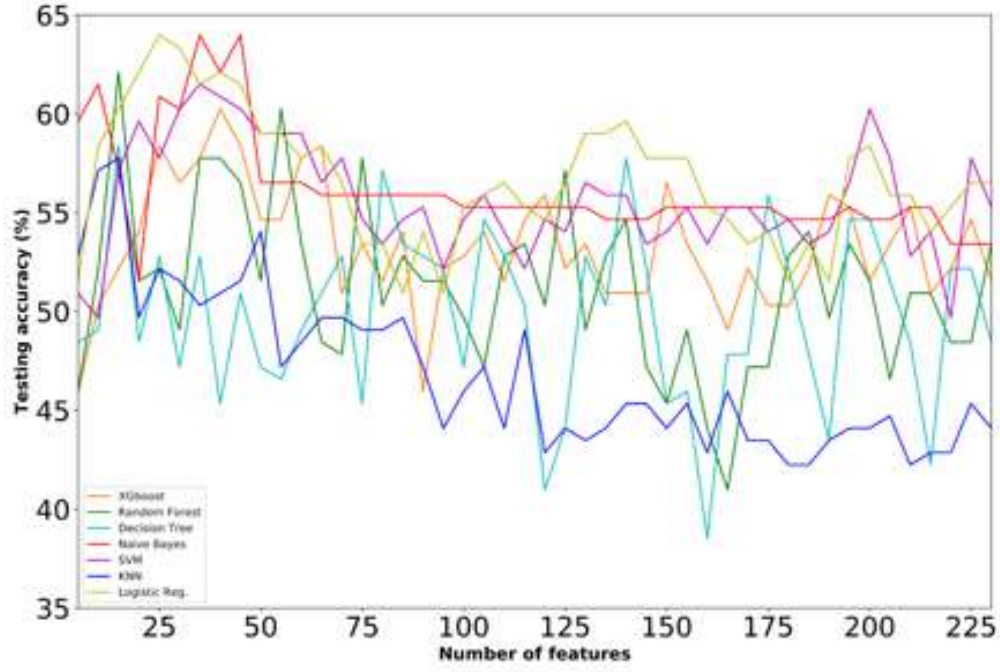


Figure 7. *Learning curves with testing accuracy scores on dataset B for different ML models trained on feature subsets of increasing dimensionality.*

- Dataset C

Less informative features with small generalization capacity are contained in dataset C, as reported in Figure 8. Unlike the previous two datasets, the best testing performance for dataset C was received at 225 features using DTs (66.67%). In general, unstable and low testing performances were observed for the majority of the employed ML models. The second highest accuracy was received for SVM (65.28%), whereas lower accuracies were obtained by the rest of the models. A significant number of features (more than 100) was also required in five out of the seven FS approaches highlighting the inability of dataset C features to provide useful information for the progression of KOA.

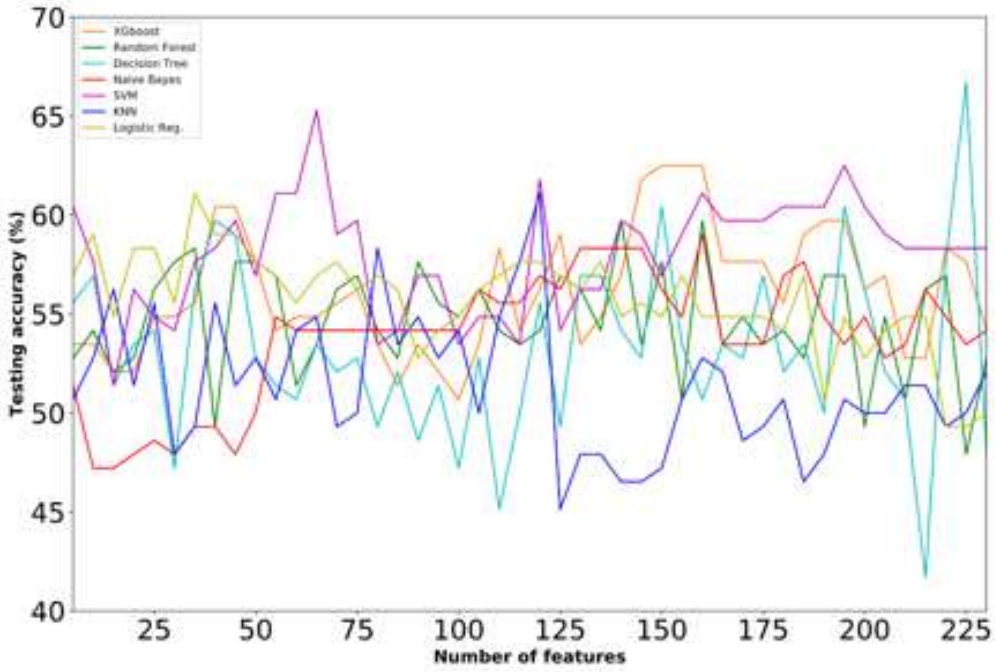


Figure 8. *Learning curves with testing accuracy scores on dataset C for different ML models trained on feature subsets of increasing dimensionality.*

- Dataset D

The combination of datasets A and B proved to be beneficial in the task of predicting KOA progression. Specifically, the following conclusions are drawn from the results reported in Figure 9: (i) The best performance (74.07%) was achieved by the SVM on the group of the fifty-five selected risk factors with linear kernel penalty and C = 0.1 (Dataset D). (ii) The second highest accuracy was received for the logistic regression (72.84%), whereas lower accuracies were obtained by the rest of the models. (iii) SVM and LR followed a similar progression in the reported accuracies with respect to the number of selected features with an upward trend in the first 20–55 features, followed by a slight performance decrease as the number of features increases. (iv) KNN gave moderate results with a maximum testing performance of 71.6% at 75 selected features. (v) Low testing accuracies were obtained by RF, XGboost and DT in the range of 42.59–66.67%.

- Dataset E

In dataset E, the SVM-based approach exhibited an upward trend with respect to selected features in the first 20–70 features, with a maximum of 71.81% at 70 features (which was the best in the category). The inclusion of additional features led to a small reduction in the accuracies achieved (Figure 10). Similarly to

SVM, LR gave the second highest accuracy (71.14%) for less features (55). XGboost also gave a comparable performance (70.47%) in a subset of 45 selected features. Lower testing accuracies were received by the rest of ML models.

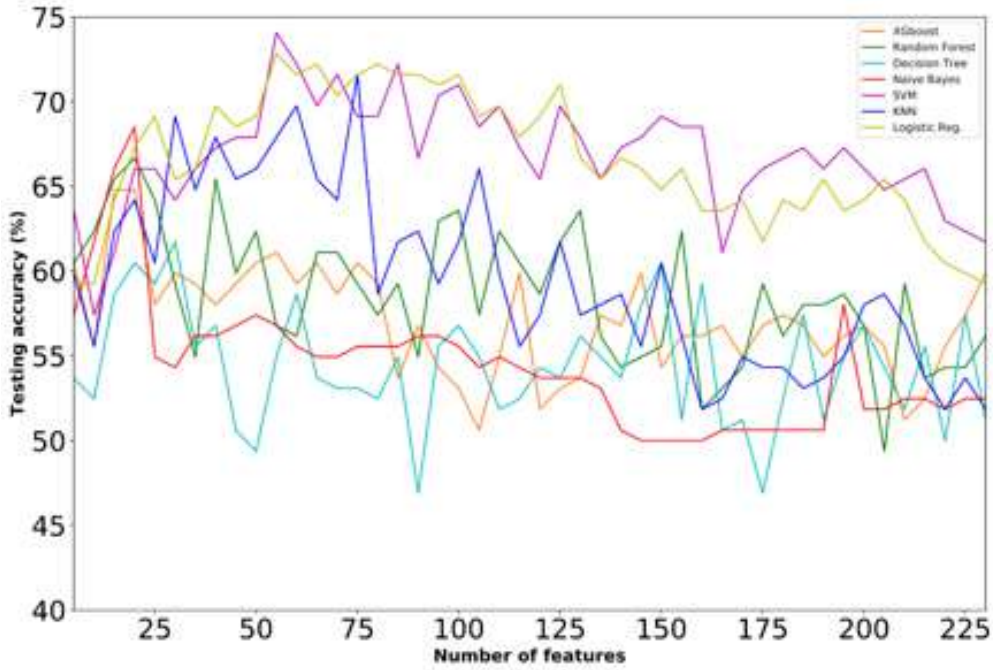


Figure 9. *Learning curves with testing accuracy scores on dataset D for different ML models trained on feature subsets of increasing dimensionality.*

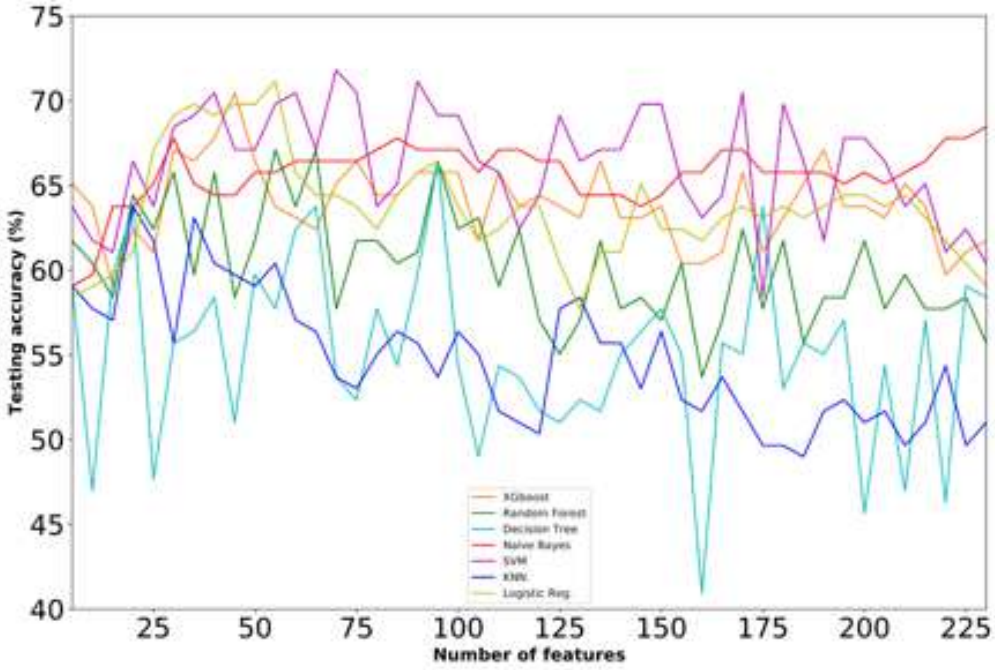


Figure 10. *Learning curves with testing accuracy scores on dataset E for different ML models trained on feature subsets of increasing dimensionality.*

Selected Features

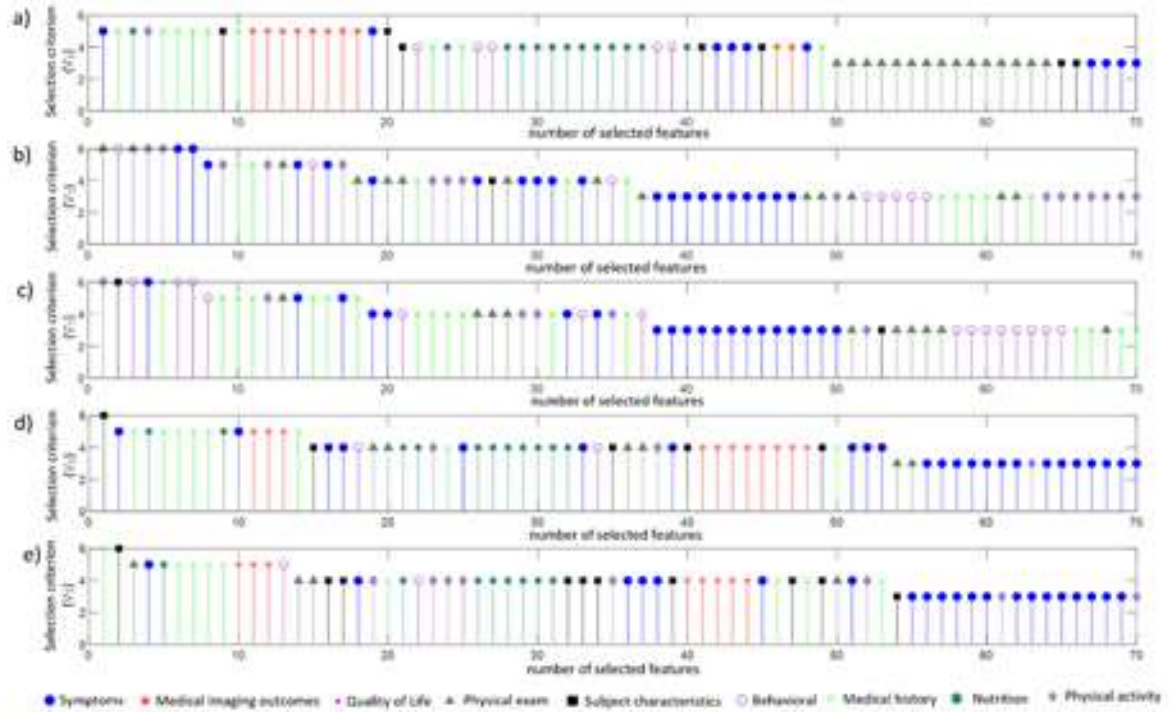


Figure 11. *Features selected in datasets A to E in (a–e), respectively. Axis y (selection criterion) denotes how many times a feature has been selected (6 declares that a specific feature has been selected by all six FS techniques and so on). Features have been ranked based on the selection criterion Vj and are visualised with different colors each one representing a specific feature category.*

Figure 11 shows the first 70 features selected by the proposed FS approach for datasets A to E. Features are visualised with different colors and marks depending on the feature category they belong. The following conclusions could be drawn from the analysis of Figure 12: (i) Symptoms and medical imaging outcomes seem to be the most informative feature categories in dataset D in which the overall best performance was achieved. Specifically, eleven medical history outcomes and ten symptoms were selected in the first 55 features that gave the optimum prediction accuracy; (ii) nutrition and medical history characteristics were also proved to be contributing risk factors since approximately 20 out of the first selected 55 features were from these two feature categories (in dataset D). The full list of selected features for dataset D is provided in the appendix; (iii) similar results with respect to the selected features were extracted from the analyses in datasets A and E (in Figure 12 a, e) that gave comparative prediction results (close to 72%); (iv) a different order in the selected features was observed in datasets B and C (as depicted in Figure 12 b, c). The low accuracies recorded in these datasets (less than 67%) verify that the contained in these datasets features are less informative; (v) overall, it was concluded that a combination of heterogeneous features coming from almost all feature categories is needed to predict KL progression highlighting the necessity of adopting a multi-parametric approach that could handle the complexity of the available data.

Discussion of results

This work focuses on the development of a ML-empowered methodology for KL grades prediction in healthy participants. The prediction task has been coped as a two-class classification problem where the participants of the study were divided into two groups (KOA progressors and non-progressors). Various ML models were employed to perform the binary classification task (KOA progressors versus non-progressors) where accuracies up to 74.07% (Dataset D) were achieved. Within the secondary objectives of the work were to identify informative risk factors from a big pool of available features that contribute more to the classification output (KOA prediction). Moreover, we explored different options with respect to the time period within which data should be considered in order to reliably predict KOA progression.

Three different options were investigated as far as the time period within which data should be considered in order to reliably predict KOA progression. To accomplish this, we worked with 5 different datasets. We first examined whether baseline data (dataset A) could solely contribute in predicting KOA progression. Going one step further, the features 'progression within the first 12 months or 24 months was also considered as an alternative source of information (datasets B and C). The aforementioned analysis in Section 4 revealed that: (i) a 71.71% prediction performance can be achieved using features from the baseline, (ii) features' progression cannot solely provide reliable KOA predictions and (iii) a combination of features is required to maximize the prediction capability of the proposed methodology. Specifically, the overall best accuracy (74.07%) was obtained by combining datasets A and B that contain features from the baseline visit along with their progression over the next 12 months. Considering a longer period of time (24 months) in the calculation of features' progression resulted to lower prediction accuracies (71.81%).

The proposed FS methodology outperformed six well-known FS techniques achieving the best tradeoff between prediction accuracy and dimensionality reduction. From the pool of approximately 700 features of the OAI dataset, fifty-five were finally selected in this work to predict KOA. As far as the nature of the selected features, it was concluded that symptoms, medical imaging outcomes, nutrition and medical history are the most important risk factors contributing considerably to the KOA prediction. However, it was also extracted that a combination of heterogeneous features coming from almost all feature categories is needed to effectively predict KL progression.

Seven ML algorithms were evaluated for their suitability in implementing the prediction task. Table 7 with the summary of all reporting result indicates that LR and SVM were proved to be the best performing models. The good performance of SVM could be attributed to the fact that SVM models are particularly well suited for classifying small or medium-sized complex datasets (both in terms of data size and

dimensionality). LR was the second-best performer providing the highest prediction accuracy in datasets A and B and the second highest in datasets D and E. The fact that a generalized linear model such as LR accomplishes high performances indicates that the power of the proposed methodology lies on the effective and robust mechanism of selecting important risk factors and not so much on the complexity of the finally employed classifier. Identifying important features from the pool of heterogeneous health-related parameters (including anthropometrics, medical history, exams, medical outcomes, etc.) that are available nowadays is a key to increase our understanding of the KOA progression and therefore to provide robust prediction tools.

3.2 Personalised Prediction of Pain progression

1st Approach:

Dataset description

Data from the OAI database was used in this work in order to validate our approach. This database was designed for 2 specific reasons:(i) to identify the factors that cause KOA, (ii) to promote the research in the area of KOA, which is going to create a better quality of life for patients with KOA. The OAI database was launched in 2002, and its data is from patients in the ages 45-79 years old, either with symptomatic KOA, or being on the verge of developing it, in at least one knee. The study that produced this database had taken place in four medical centers in the US. In total 4796 patients were enrolled in the study, which lasted for 8 years. The most significant thing about this database is that it had a more than 90% follow up for the first 4 years. In this work though, we have not used all of the features. We have developed a voting system for assessing feature importance using only baseline data. WOMAC pain data from the first four visits was utilized to identify the different clusters of pain progression, whereas the selected feature subsets, as generated by the application of proposed FS methodology on baseline features, were used to train the ML models and finally produce the predictions.

Methodology

The proposed, in this work methodology, comprises of the following components: (i) a fitting technique for grouping/ labeling of the data, (ii) a hybrid and robust Feature Selection technique employing a number of feature ranking algorithms to avoid bias, (iii) Machine Learning models for decision making and (iv) Validation.

Grouping/Labeling, Data Pre-Processing and Feature Selection

As described in Deliverable 6.3 the available data was grouped into three clusters, each one representing a different pain progression condition:1) cluster 1: pain decline, 2) cluster 2: no significant pain change and 3) cluster 3: pain increase. Furthermore, we applied the data pre-processing steps and the hybrid feature selection methodology, which have described in Deliverable 6.3.

Machine Learning Algorithms

Six (6) Machine Learning models were explored for their suitability in predicting pain progression on feature subsets of varying dimensionality, in order to see which one produces the best results. In this subsection we give a brief overview of the models that were employed in order to tackle the pain prediction problem.

1) Decision Trees: Decision Tree [25] is one of the most famous algorithms for supervised learning for classification problems. It uses a lot of if-then-else decision rule statements in order to come to a decision. Its structure is a branch structure which breaks the data into data subsets, and then it produces decision

and leaf nodes. Every node has a minimum of two branches, and every leaf node is for classification or a decision prediction.

2) k Nearest Neighbors: k Nearest Neighbors (kNN) [26] is a non-parametric, lazy learning algorithm. The classification prediction of a sample datapoint, is achieved with the use of data, which are class-separated. The algorithm presumes that similar datapoints are close to each other. More specifically, this algorithm loops over every datapoint in the data and calculates the distance between every datapoint and the chosen datapoint. The distances are sorted in an ascending order and then the algorithm chooses the first k entries.

3) Support Vector Machines: Support Vector Machines (SVM) [27] is an algorithm which finds a line that separates the datapoints, that belong to different classes. The datapoints that are closest to the line play a crucial role in the learning process e (the so-called support vectors). Then the distance between the line and every datapoint is calculated, with an overall target to maximize the distance between classes. In case a non-linear separation is needed, kernels are applied in order to project the datapoints into higher dimensional spaces.

4) Random Forest: Random Forest is an algorithm consisted of many decision trees algorithms [28]. Its characteristics are the randomness in the sampling of datapoints when building the trees; and the randomness in the feature's subsets, when splitting nodes. Every tree in the algorithm learns from a random sample of data. These samples of data are being used several times by the trees, which means that the trees take them with replacement. So, every tree has high variance because of this fact, but the random forest has lower variance in overall. It is worth noting that the decisions are the average of the predictions of all the trees in the random forest.

5) XGBoost: XGBoost or eXtreme Gradient Boosting [29], is a parallel tree boosting that solves data science problems in a fast and accurate way. After constructing the boosted trees the algorithm calculates the importance score of every feature of the dataset. This score is an indicator of how useful is its feature to the construction of the trees inside the algorithm. The calculation of this score is achieved by the amount that each feature point split improves the performance for the model for the data that the node is responsible for. A popular measure of performance is the Gini index which selects the split points. More specifically the Gini coefficient is a statistic which quantifies the amount of inequality that exists in a population. It is a number between 0 and 1, with 0 representing perfect equality and 1 perfect inequality. XGBoost in fact ranks the features of the data by comparing them to each other.

6) Naive Bayes: Naive Bayes is a probabilistic classifier that uses the Maximum A Posteriori decision rule in a Bayesian setting and is included in supervised learning [30]. The main idea behind this method is the Bayes Theorem. Bayes theorem approximates the probability of an event given the probability of a past event. The Naive Bayes predicts membership of probabilities for every class, such as the probability that the given data point belongs to a particular class. The data point belongs to the class with the highest probability score.

Validation

We validated the results by performing a 70%-30% train-test split. Learning of the algorithms was achieved on the stratified version of the train and the final performance was calculated on the test data.

Results

Tables 10 and 13 present the feature ranking exploration of the first 10 Features of the whole dataset for the left and the right knee, respectively. The feature ranking was decided on the basis of a majority vote

scheme by using the proposed feature selection methodology, as discussed in Deliverable 6.3. We also note, that features related to symptoms were selected.

A. Results on Left Leg

1) Feature Selection Results: Table 3 shows in order of importance the features for the left knee after the FS implementation. It was noted that the features that occupied the first positions, concern self-reported data about pain, difficulties in daily life and quality of life in knee-related functions. The following features were selected due to the direct correlation of these symptoms with the presence or imminent development of KOA, a finding that emerges from the literature survey. We can observe that these features are directly related to pain on the left leg.

Table 3. *Most important features for the left leg*

Features	Description
V00WPLKN5	Left knee pain: standing, last 7 days
V00WPLKN4	Left knee pain: sit or lie down, last 7 days
V00WPLKN3	Left knee pain: in bed, last 7 days
V00WPLKN2	Left knee pain: stairs, last 7 days
V00WPLKN1	Left knee pain: walking, last 7 days
V00WOMKPL	Left knee: WOMAC Pain Score
V00P7LKFR	Left knee pain: how often
V00KQOL4	Quality of life: how much difficulty with knee(s)
V00DIRKN7	Right knee difficulty: in car/out of car, last 7 days
V00DILKN6	Left knee difficulty: walking, last 7 days

2) Performance: Table 4 cites the results of various algorithms applied on different combinations of feature subsets as they have been ordered by the proposed FS methodology. It was observed that RF achieved the best accuracy score, which is 84.3% at the first 25 features, whereas the inclusion of additional features led to a progressive decline in the accuracies achieved. The rest of the ML models achieved inferior results, with SVM producing the second-best results with 80.83% accuracy score. In overall, as we add more features to the aforementioned models, we observe that their accuracy scores decrease.

Table 4. *Left leg: features and model accuracy scores (%)*

Features	DT	KNN	NB	Rf	SVM	XGB
5	64.46	57.02	70.25	71.07	66.12	63.64
10	71.9	71.07	73.55	79.34	73.55	74.38
15	71.9	76.86	76.03	77.69	75.21	80.99
20	71.9	76.86	73.55	81.82	72.73	83.47
25	66.94	79.34	69.42	84.3	78.51	80.99
50	68.6	69.42	71.07	75.21	78.51	74.38
100	73.55	73.55	69.42	76.86	78.51	83.47
150	67.77	67.77	68.6	78.51	76.86	79.34
200	58.68	67.77	71.9	76.03	78.51	77.69
250	61.16	71.9	65.29	74.38	76.86	76.86
300	67.77	66.12	61.16	75.21	81.82	79.34
350	64.46	61.16	63.64	76.03	76.86	76.03
400	60.33	61.98	62.81	74.38	75.21	76.03

450	56.2	66.12	61.98	79.34	78.51	77.69
500	58.68	58.68	62.81	69.42	77.69	76.86
550	61.98	64.46	62.81	77.69	76.86	75.21
600	55.37	63.64	60.33	76.86	71.07	75.21
650	61.98	61.98	58.68	72.73	74.38	73.55
700	50.41	56.2	58.68	70.25	72.73	76.03
750	74.38	60.33	58.68	71.9	71.9	75.21

B. Results on Right Leg

1) Feature Selection Results: Table 5 depicts in order of importance the features for the right knee after the FS implementation. It was observed that 7 out of the 10 first selected features were the same with the ones selected for the left leg. This finding indicated the selected features that lead to the prediction of KOA for each leg have uniformity and mainly concern self-reported data on pain, stiffness and quality of life.

Table 5. *Most important features for the right leg*

Features	Description
V00P7RKFR	Right knee pain: how often
V00WPRKN5	Right knee pain: standing, last 7 days
V00WPRKN4	Right knee pain: sit or lie down, last 7 days
V00WPRKN3	Right knee pain: in bed, last 7 days
V00WPRKN2	Right knee pain: stairs, last 7 days
V00WPRKN1	Right knee pain: walking, last 7 days
V00WOMKPR	Right knee: WOMAC Pain Score (calc)
V00KSXLKN2	Left knee symptoms last 7 days
V00KPRKN3	Right knee pain: bending knee fully, last 7 days
V00DIRKN3	Right knee difficulty: stand from sitting, last 7 days

2) Performance: For the right leg, Table 6 shows the results of the machine learning algorithms that we have applied on different combinations of feature subsets, created by the FS methodology. The best performing algorithm for the right leg is Random Forest with an accuracy score of 84.3%, for 20 features; and as you can see the addition of additional extra features has produced inferior results for our prediction. Table 15 shows the confusion matrix of the Random Forest for the best prediction score that it has produced. It is observed that the other algorithms have achieved inferior results as observed from the Tables 3 and 5, similar results are obtained on both legs; indicating the repeatability and robustness of the proposed methodology.

Table 6. *Right leg: features and model accuracy scores (%)*

Features	DT	KNN	NB	Rf	SVM	XGB
5	63.33	68.33	73.33	69.17	70.0	68.33
10	75.83	75.83	75.83	76.67	75.0	75.0
15	73.33	75.0	75.0	76.67	79.17	70.83
20	65.0	70.83	76.67	82.5	80.0	74.17
25	69.17	65.0	77.5	75.0	77.5	75.0
50	72.5	66.67	73.33	78.33	79.17	75.83

100	65.0	64.17	70.83	78.33	78.33	78.33
150	67.5	58.33	66.67	77.5	80.83	77.5
200	60.0	60.83	62.5	77.5	78.33	79.17
250	75.0	58.33	63.33	77.5	78.33	73.33
300	68.33	58.33	64.17	78.33	80.0	78.33
350	60.0	45.83	63.33	73.33	76.67	78.33
400	60.83	53.33	63.33	74.17	75.83	75.0
450	59.17	60.0	60.83	75.0	77.5	75.0
500	55.83	58.33	59.17	74.17	77.5	73.33
550	52.5	53.33	55.0	71.67	77.5	75.0
600	65.0	50.83	55.0	69.17	74.17	76.67
650	65.83	50.83	54.17	69.17	74.17	74.17
700	60.83	51.67	52.5	75.83	77.5	74.17
750	65.0	50.83	50.83	71.67	75.83	73.33

Discussion of results

In this work we have proposed a methodology in which we identified three different clusters of KOA pain progression along with the most informative parameters towards the development of prognostic ML models that can predict long-term pain progression. In order to achieve this, we have developed a voting system for feature importance, in which 6 different methods are used to show the most important features in the dataset. Then we applied 6 different models in various subsets of data, which procedure has proved that XGB achieves a state-of-the-art accuracy score by using only a small number of features. As you can see on Tables 11 and 14, we present the results of our analysis on various combinations of models and numbers of features. Tables 10 and 13 present the 10 most important features for KOA pain progression on the right and the left leg respectively. Summing up we have used for this work only data from the baseline and not from future visits for our prediction. Moreover, we detect the basic trends in pain progression so that we can construct the 3 classes of patients. More specifically we have achieved an 84.3% for the prediction of pain on the left leg, and an 82.5% on the right leg. An important observation here is that these high accuracy scores were achieved by using a relatively small subset of features (25 features for the left leg, and 20 for the right leg) that share similar characteristics. It was also observed from the Tables 10 and 13 that the most important features for the pain progression prediction are related directly with the pain on each leg respectively. These accuracy scores, with the combination of a small number of features, can set the foundation, for the development of robust tools capable of identifying pain progression at an early stage therefore improving future KOA prevention efforts. Our ultimate goal is to improve the quality of life for people with KOA. For our future work, we are planning to also consider imaging data and associated image-based biomarkers that are expected to further improve the predictive capacity of the proposed methodology.

2nd Approach:

Dataset description

Data were obtained for this study from the OAI database (available at https://nda.nih.gov/oai/). OAI is a multi-center, ten-year observational study of men and women with ultimate objective to provide resources to enable a better understanding of prevention and treatment of KOA. In this work, we considered only clinical data from the baseline study of OAI. OAI is a big pool of risk factors, which is characterized by heterogeneity and high dimensionality. In total, 649 features were considered as possible risk factors for the prediction of pain progression (Table 7). Specifically, we divided the available clinical data from the baseline visit into 7 categories: (i) subject characteristics, including variable e.g., BMI and height (ii) symptoms, which are related to swelling, knee difficulty, stiffness and pain (iii) behavioural, including

participants' quality level of social status and daily routine (iv) medical history, which includes variables regarding a participant's medications and general health histories (v) nutrition, which includes variables from Block Food Frequency questionnaire (vi) physical activity, which consists of questionnaire results regarding activities during a typical week or the last 7 days and (vii) physical exam, which are related to physical measurements of participants.

Table 7. *Characteristics of OAIs risk factors*

Feature category	Description of Categories	Number of features
Subject characteristics	Anthropometric parameters of participants, e.g., BMI and height	34
Symptoms	Symptoms related to stiffness, swelling, knee difficulty and pain (only for experiment B)	122
Behavioural	Questionnaire results regarding participants' social status and quality level of daily routine	39
Medical history	Questionnaire results regarding a participant's medications and general health histories	138
Nutrition	Variables which derived from Block Food Frequency questionnaire.	205
Physical activity	Questionnaire data regarding activities during a typical week or the last 7 days	41
Physical exam	Participants' physical measurements and performance measures	70
	Total number of features:	649

Pain prediction has been formulated as a three-class classification problem. Specifically, the participants of the study were divided into three groups: (1) participants who experienced decline in pain intensity; (2) participants that had no significant change in pain intensity and (3) participants who experienced increase in pain intensity during the curse of the study (Figure 12). The main objective of the study is to build ML models that could discriminate the three aforementioned groups and therefore be able to decide whether a participant sample will experience any pain progression in the future. A secondary objective is to identify which of the available risk factors contribute more to the classification output and as result can be considered as contributing factors in the prediction of pain.

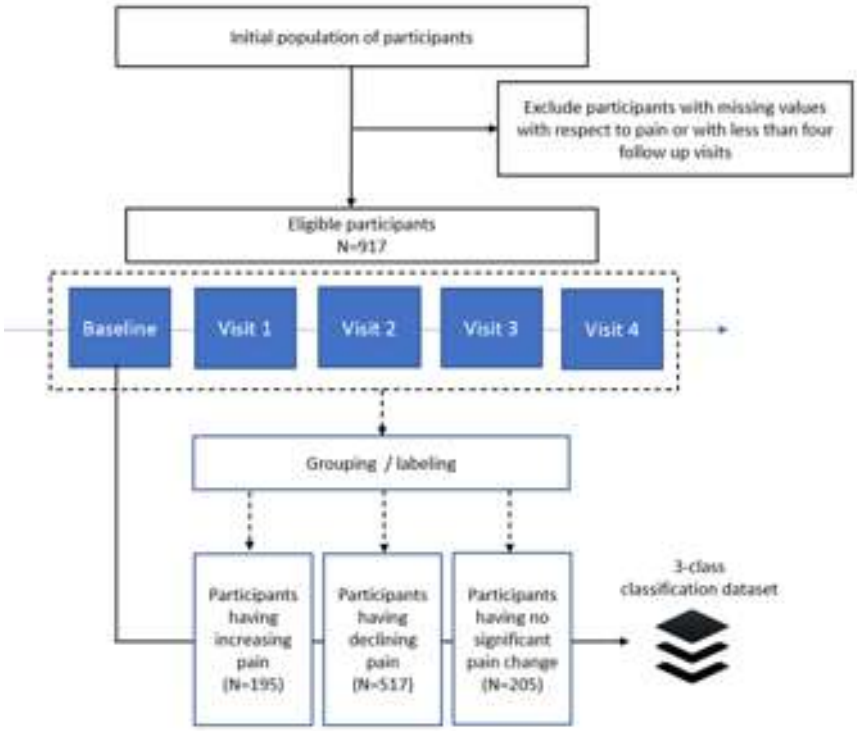


Figure 12. *Stratification of patients in the current study.*

Methods

The proposed methodology comprises of the following components: (i) grouping/labeling of the available data employing a linear fitting technique to model the progression of pain, (ii) an evolutionary feature selection technique that selects robust predictive risk factors, (iii) ML for decision making and (iv) a well-known validation scheme.

Feature Selection

A recently published FS technique, termed GenWrapper (Section 3.4) was employed to identify the most informative risk factors from the baseline visit that could be used to discriminate the three aforementioned clusters of pain progression (Figure 13). The employed FS is an evolutionary genetic algorithm-based wrapper technique that selects features which consistently work well at any possible data sample and thus has increased generalization capacity with respect to KOA progression. GenWrapper also performs well on imbalanced datasets and this is another reason why it was selected in the current study where the number of samples per class varies considerably. The 3-class problem was formulated as a group of three binary classification problems and GenWrapper was applied separately at each one. This leads to the creation of three selected feature subsets (FS_{12}, FS_{13} and FS_{23}) on which three local ML models were trained (M_{12}, M_{13} and M_{23}, respectively). Specifically:

- FS_{12} comprises selected risk factors that are sensitive to the discrimination of classes 1 (pain decline) and 2 (stable pain or no pain);
- FS_{13} comprises selected risk factors that can effectively discriminate classes 1 (pain decline) and 3 (pain increase) and
- FS_{23} comprises selected risk factors that can effectively discriminate classes 2 (stable pain or no pain) and 3 (pain increase).

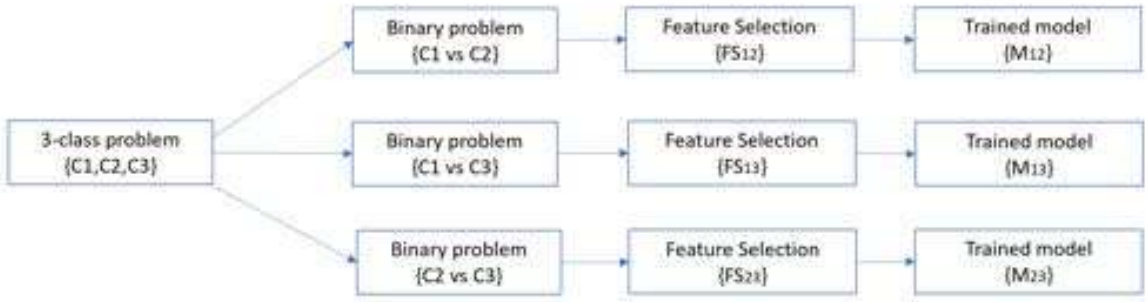


Figure 13. *The proposed FS methodology*

Machine Learning

Decision making was based on the combination of the outputs of the three trained models M_{12}, M_{13} and M_{23} as shown in **Error! Reference source not found.**. Overall, the process of assigning a decision on a sample x_i can be described as follows:

- x_i is provided as input into the three different ML pipelines.
- Three variants of x_i are created (x_i^{12}, x_i^{13} and x_i^{23}), each one comprising features as they have been selected from the FS applied on the three binary problems.
- Three decisions d_i^{12}, d_i^{13} and d_i^{23} are produced from the trained models M12, M13 and M23 after supplying them with x_i^{12}, x_i^{13} and x_i^{23}, respectively.
- Decision fusion is performed and the decisions with the majority counts are considered as the final predicted outputs. For example, if d_i^{12}=1, d_i^{13}=1 and $d_i^{23} = 2$ then the final predicted output of the model is class 1 as it has been produced by the majority of the models.

Various classification models were explored for the suitability in implementing the learning task. Given their effective application in previous KOA studies, Support Vector Machines (SVMs) were finally selected because of their capacity to handle high dimensional feature spaces and their high generalization performance.

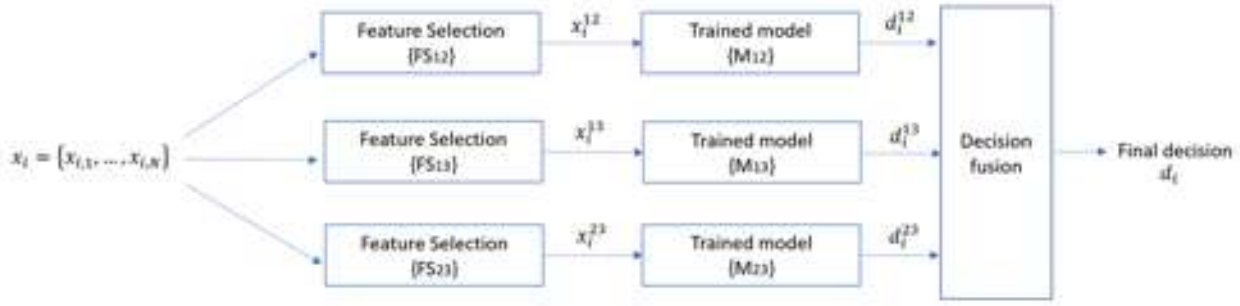


Figure 14. *Decision making as a combination of binary classifiers trained on selected feature subsets*

Results and discussion

This section presents the performance of the proposed predictive modelling methodology. The predictive capacity of each of the three binary classifiers is initially demonstrated with respect to the number of selected risk factors. Two different approaches were investigated: (i) in the first one all pain-related variables were

omitted from the available feature set and (ii) in the second experimentation, the pain-related variables of the baseline were retained.

Predictive capacity of the binary classifiers without pain related variables

The experimentation in this subsection focuses on a subset of the initial feature space in which all the pain-related variables were excluded. The list of the excluded variables is given in the appendix. Figure 15 shows the performance (10FCV) of the M_{12} classifier with respect to the number of features as they have been selected by the employed FS algorithm. As it is observed, the accuracy increases rapidly for the 15 first selected whereas the inclusion of additional features leads to a slight performance increase. The overall best performance (88.82%) was achieved at 32 risk factors that we were finally selected to be included in the FS_{12} subset.

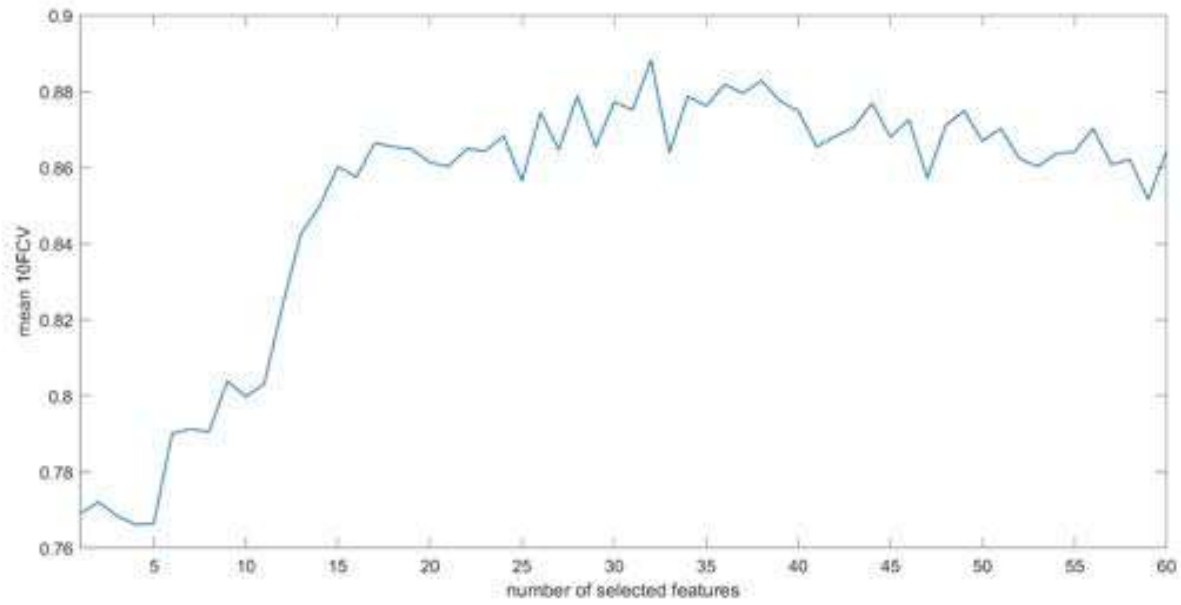


Figure 15. *Performance of the binary classifier M12 (declining pain versus no pain progression) with respect to the number of selected risk factors. Pain related variables have been excluded from the feature set.*

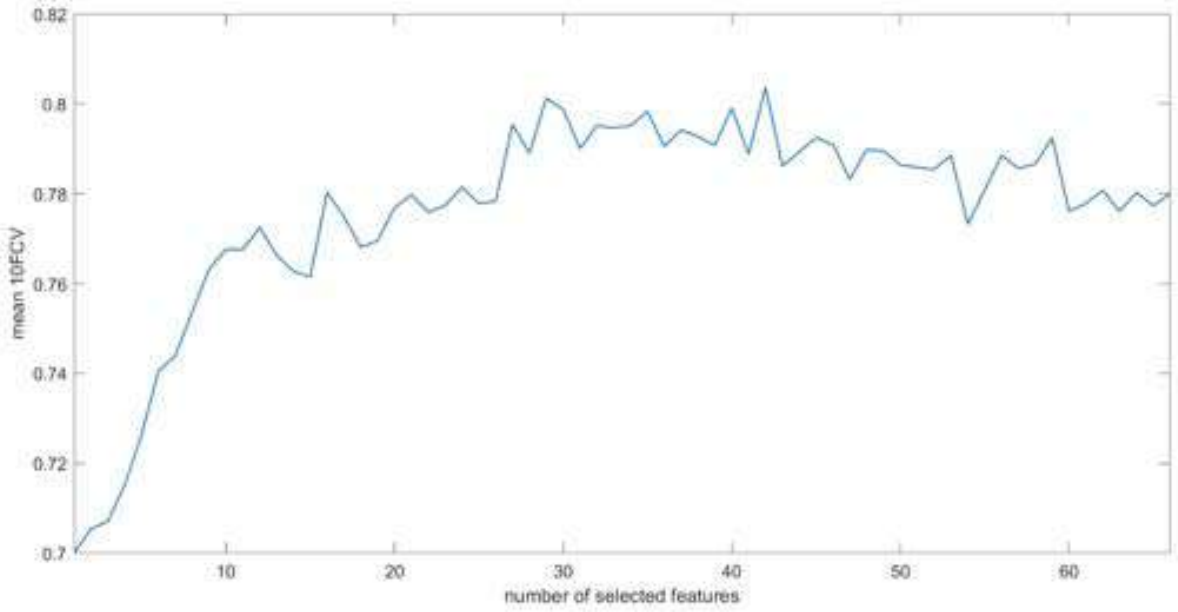


Figure 16. *Performance of the binary classifier M_{23} (no pain progression versus increasing pain) with respect to the number of selected risk factors. Pain related variables have been excluded from the feature set.*

The 10FCV performance of the M_{23} classifier is depicted in Figure 16. Similarly to M_{12}, the predictive accuracy of M_{23} increases with respect to the number of features that are added in the training feature set and reaches the maximum of 80.36% at 42 selected features. A slight decrease on the accuracy is then observed with the addition of more features.

A larger number of features were selected to maximize the performance of the M_{13} classifier as shown in Figure 17. Specifically, the maximum 10FCV accuracy of 84.51% was achieved at 58 selected features. The necessity of including more features indicates that the discrimination between classes 1(declining pain) and 3(increasing pain) is a more difficult task compared to the tasks where class 2 (no significant pain change) is being discriminated from classes 1 and 3.

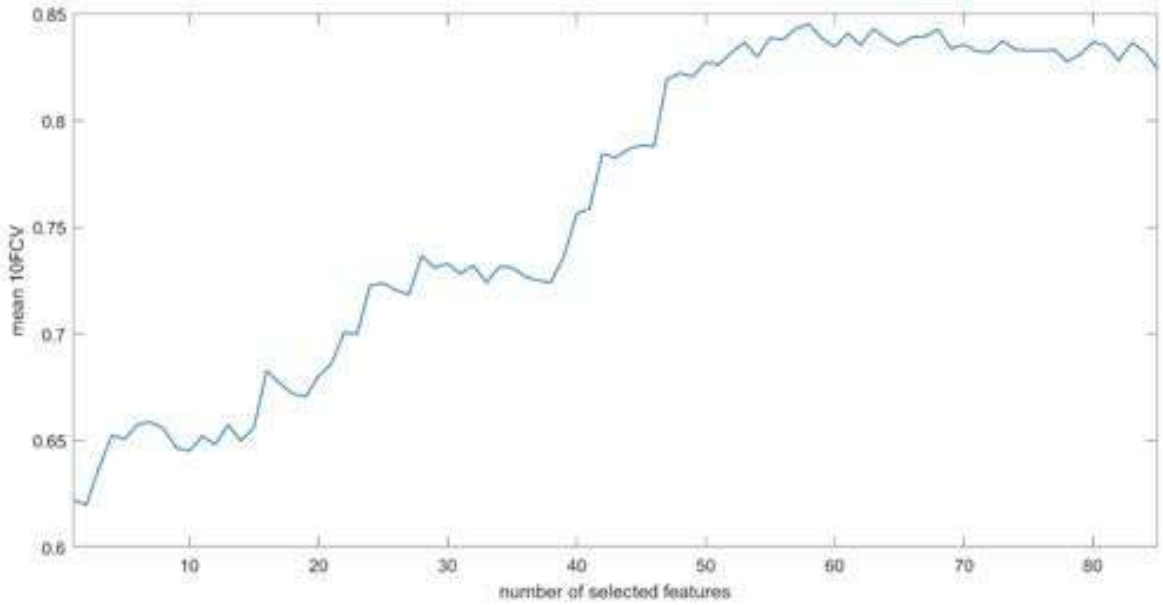


Figure 17. *Performance of the binary classifier M13 (declining pain versus increasing pain) with respect to the number of selected risk factors. Pain related variables have been excluded from the feature set.*

Predictive capacity of the binary classifiers with pain variables

Figures 18-20 show the performance of the three binary classifiers M_{12}, M_{23} and M_{13} which are trained on risk factors selected from the full feature set available at the baseline visit (with the pain variables included).

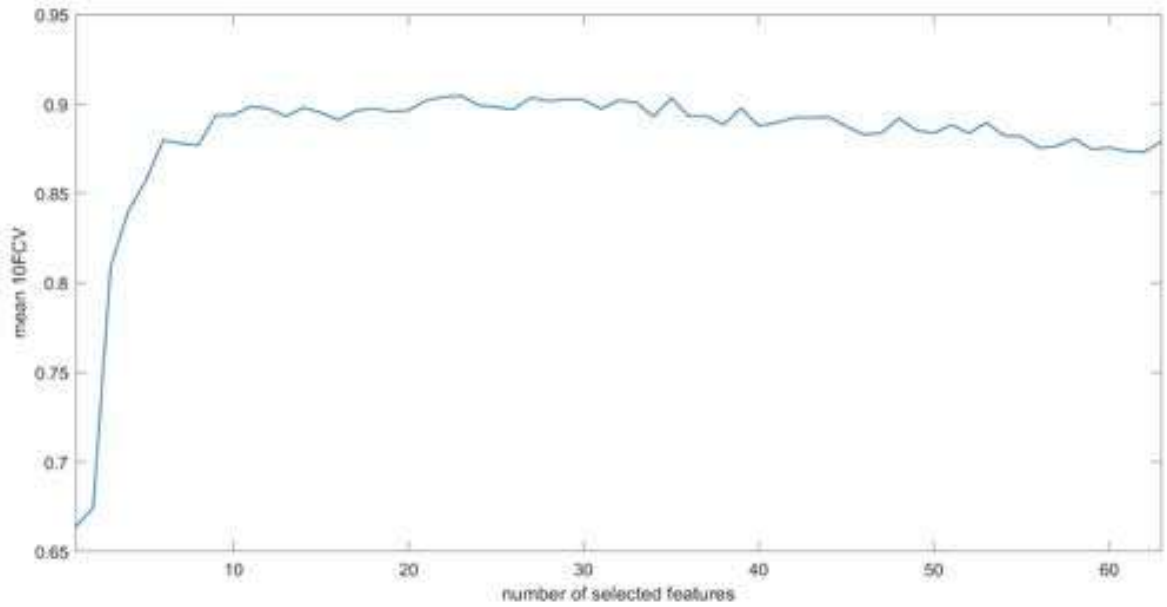


Figure 18. *Performance of the binary classifier M_{12} (declining pain versus no pain progression) with respect to the number of selected risk factors. Pain related variables have been retained in the feature set.*

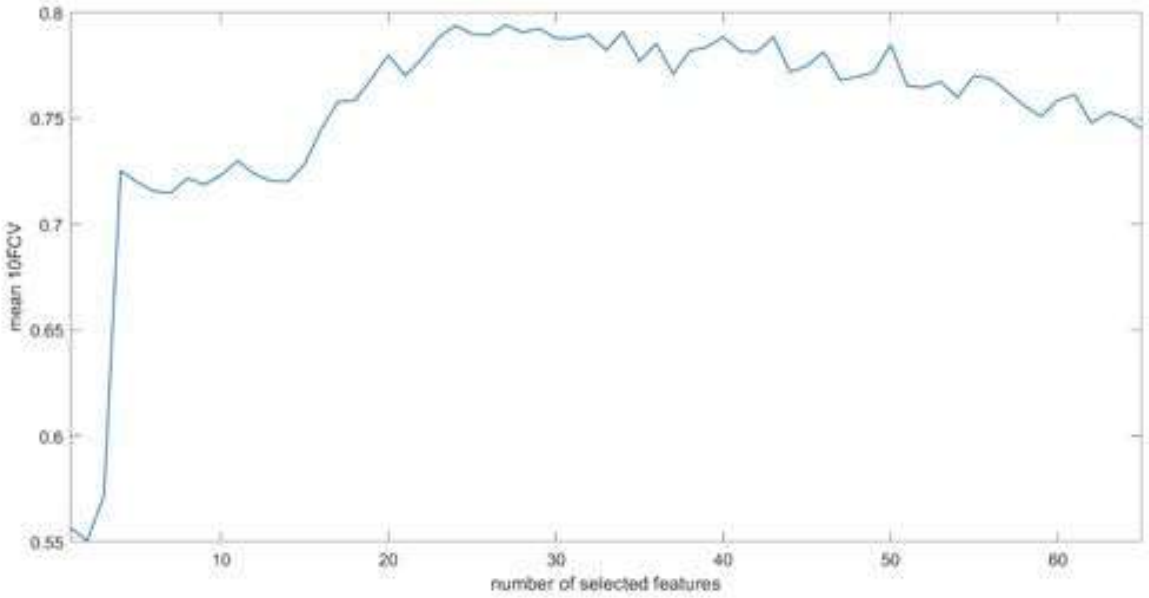


Figure 19. *Performance of the binary classifier M_{23} (no pain progression versus increasing pain) with respect to the number of selected risk factors. Pain related variables have been retained in the feature set.*

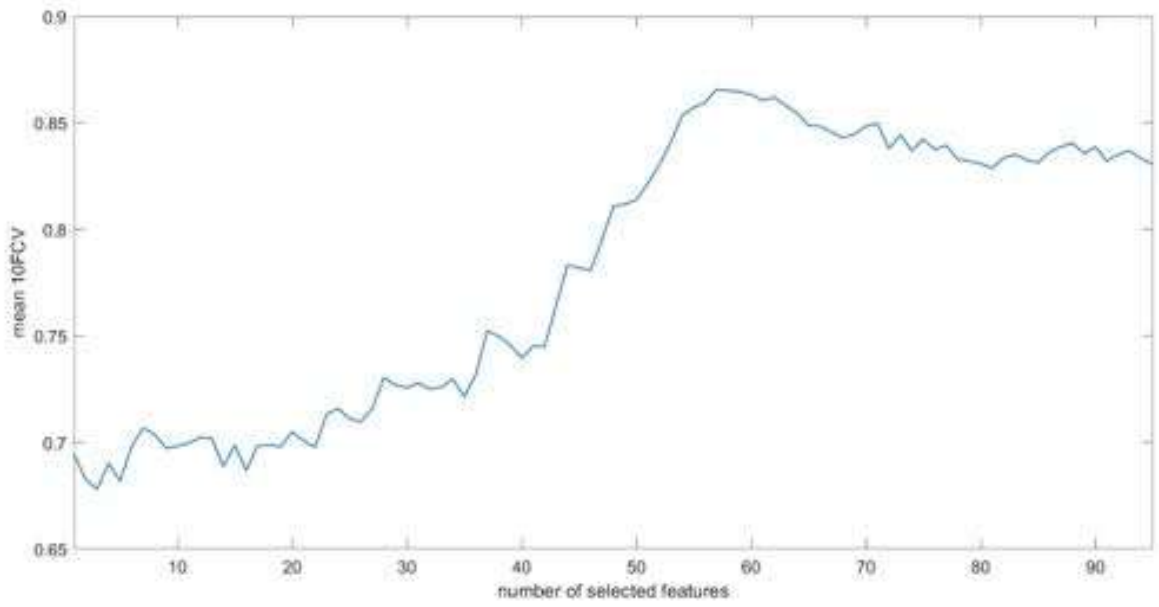


Figure 20. *Performance of the binary classifier M_{13} (declining pain versus increasing pain) with respect to the number of selected risk factors. Pain related variables have been retained in the feature set.*

The following remarks can be extracted from Figures 18-20: (i) The maximum performance (90.46%) was achieved by the M_{12} classifier, the M_{13} classifier also gave a high accuracy (86.56%) whereas a 79.41% discrimination was accomplished by M_{23} classifier. (ii) As far as the number of selected features, 23, 27 and 57 risk factors are needed to maximize the predictive accuracy of M_{12}, M_{23} and M_{13}, respectively. (iii) Overall, it was concluded that the inclusion of pain related variables leads to higher accuracy on less features compared to the experimentation of the previous subsection where the pain related variables had been excluded.

Feature selection results

Table 8 cites the number of selected features for each one of the local binary models (with and without pain-related features). The following conclusions could be drawn from the analysis of Table 8 with respect to the first experimentation without pain-related risk factors): (i) Symptoms and nutrition seem to be the most informative feature categories. Specifically, 30 symptoms and 42 nutrition parameters were selected in the experiment without pain variables, demonstrating that the contribution of these 2 categories to the prediction output is significant; (ii) Behavioral data, medical history and physical exam variables were also selected by the proposed FS (13, 19 and 15 respectively) providing complementary valuable information to the ones mentioned above (symptoms and nutrition); (iii) Subject characteristics and physical activity had a smaller impact on the prediction output (with 5 features selected per category).

Table 8. *Number of features selected per category in both experiments with and without pain-related features*

Feature category	**Without pain-related features**			**With pain-related features**		
	C1 vs C2	C2 vs C3	C1 vs C3	C1 vs C2	C2 vs C3	C1 vs C3
Subject characteristics	-	1	4	-	-	2
Symptoms	10	11	9	14	7	18

Behavioural	5	4	4	2	2	3
Medical history	4	5	10	2	9	9
Nutrition	9	15	18	1	6	15
Physical activity	-	2	3	-	1	4
Physical exam	4	4	7	4	2	6
Total number of features	32	42	58	23	27	57

Similar findings were observed in the experimentation with the whole feature set (including the pain-related variables). The main difference between the two experiments (with and without pain variables) was that the inclusion of pain variables led to the selection of less features in total (107 compared to 132). Moreover, the number of selected features for the symptoms' category was increased (39 in total). This could be attributed to the fact the initial state of pain at baseline is obviously a significant indicator of pain future progression and therefore a significant number of pain variables and other similar risk factors from the baseline visit were selected by the proposed FS. Overall, it was concluded that a combination of heterogeneous features coming from almost all feature categories is needed to predict pain progression highlighting the necessity of adopting a multi-parametric approach that could handle the complexity of the available data.

3.3 Personalised Prediction of JSN progression

Methodology

A machine learning approach was developed in this work by taking advantage of the combination of predictive and descriptive techniques, such as clustering, FS, and classification. The proposed methodology for predicting JSN consists of 4 main steps: (i) Data pre-processing, (ii) data clustering, (iii) feature selection and (iv) data classification. In the first step, data cleaning and normalization are performed to remove noise and bring all the variables to the same range. Then the samples are clustered based on their JSN progression using well-known clustering algorithms. Then, a selection of features is realized based on the identified clusters (that are considered as classes in our case). The selected features are used to develop prediction models for the KOA progression of patients (Figure 22).

Figure 22. *Methodology flowchart.*

In this work two strategies were investigated: (i) In the first one, two predictive models were developed using data from the right and the left knee, separately and (ii) the second strategy focuses on the development of a unique predictive model using data from both knees of KOA patients.

Data Pre-Processing, Data Clustering and Feature Selection

The steps for Data Pre-Processing, Data Clustering and Feature Selection have described in Deliverable 6.3. The clustering process (Deliverable 6.3) that was followed is presented in Figure 23.

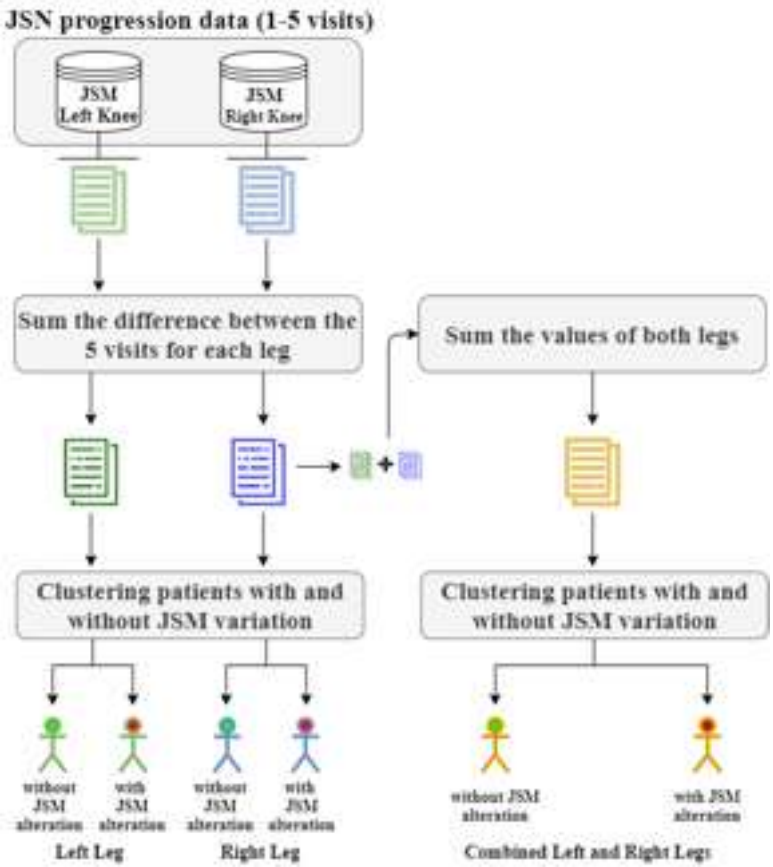


Figure 23. *Clustering process of the proposed methodology. JSM: Joint space narrowing on medial compartment.*

Data Classification

Six well-known classification algorithms were tested for the identification of the optimum model that achieves the highest accuracy on the test data:

- Gradient boosting model (GBM) is an ensemble ML algorithm, which can be used for classification or regression predictive tasks. Weak learners are used from GBM to produce strong learners through a gradual, additive, and sequential process. Hence, for the development of a new improved tree a modified version of the initial training data set is fitted in GBM [31].
- Logistic regression (LR) describes the relationship of data to a dichotomous dependent variable. LR is based on the logistic function. This model is designed to describe the data with a probability in the range of 0 and 1 [32]:

$$f(x) = \frac{1}{1+e^{-x}}, \text{ where } x \in (-\infty, +\infty) \text{ and } 0 \leq f(x) \leq 1;$$

- Neural networks (NNs), both shallow and deep NNs were employed. NNs are based on a supervised training procedure to generate a nonlinear model for prediction. They consist of layers (e.g., input layer, hidden layers, and output layer). Following a layered feedforward structure, the information is transferred unidirectionally from the input layer to output layer through the hidden layers [17,33,34].
- Naïve Bayes Gaussian (NBG) employs the Bayes theorem. This probabilistic classifier presents strong independence assumptions between the variables/features given the class. Furthermore, this model embraces the assumption that the data follow the Gaussian distribution [35,36].
- Random forest (RF) belongs in the ensemble learning methods and is based on decision trees. This model constructs a large number of decision trees. Every decision tree denotes a class prediction. Thus, the class with the most votes represents the model's prediction [37,38].

- Support vector machines (SVMs) are another supervised learning model [39,40]. SVMs target to create the hyperplane, which is a decision boundary between two classes that enables the prediction of labels from one or more feature vectors. The main aim of SVMs is to maximize the class margin that is actually the distance between the closest points (support vectors) of each class [41].

Evaluation

Medical Data

Data from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/) were used in this study. Specifically, only clinical data from the baseline from all individuals without or being at high risk to develop KOA in at least one knee were included. In total, 725 features from 9 feature categories were considered as possible risk factors for the prediction of JSN as shown in Table 16. Clustering was performed on the JSN progression represented by the JSM measures (especially using the variables V00XRJSM, V01XRJSM, V03XRJSM, V05XRJSM, and V06XRJSM of the OAI from the first five visits) to group patients into two clusters (non-progressing patients and those whose JSN changes over time).

The available data from the baseline visit were divided into 9 categories (Table 9): (i) Anthropometrics, (ii) behavioral, (iii) symptoms, (iv) quality of life, (v) medical history, (vi) medical imaging outcomes, (vii) nutrition, (viii) physical exam, and (ix) physical activity. The first category contains anthropometric characteristics, such as body mass index, weight, and height. The behavior category concerns the habits and sociability of the participant. The symptoms category also contains all features that are associated with pain and any dysfunction. The quality-of-life category refers to variables that represent the participation of the individual to social events and activities. The medical history category includes features related to the medical history of the participants and of their family and whether they have received a medical prescription in specific time periods. Another category is the medical imaging outcomes which come after clinical evaluation with imaging such as X-rays. In addition, in the category of physical examination, we included all the characteristics related to the examination of a participant (such as hand and knee exam), various biomechanical measurements, and field tests. Finally, the category of physical activity includes all variables that relate to the individual activity, such as household activities and leisure activities.

Table 9. *Maim categories of the feature subsets considered in the proposed methodology.*

Category	Description	Number of Features
Anthropometrics	Includes measurements of participants such as height, weight, BMI (body mass index), etc.	37
Behavioral	Questionnaire results which describe the participants' social behaviour	61
Symptoms	Includes variables of participants' arthritis symptoms and general arthritis or health-related function and disability	108
Quality of life	Variables which describe the quality level of daily routine	12
Medical history	Questionnaire results regarding a participant's arthritis-related and general health histories and medications	123
Medical imaging outcome	Variables which contain medical imaging outcomes (e.g., osteophytes and joint space narrowing (JSN))	21
Nutrition	Variables resultfrom the use of the modified Block Food Frequency questionnaire	224

Physical exam	Variables of participants' measurements, performance measures, and knee and hand exams	115
Physical activity	Questionnaire data results regarding household activities, leisure activities, etc.	24
	Total number of features:	725

Evaluation Methodology

The proposed methodology was applied in the context of predicting the JSN progression in patients with KOA by using the medical data derived from the dataset. Initially, the methodology was applied for each leg separately and, consecutively, for both legs combined.

The proportion of 70–30% was chosen for splitting the data set to training set and testing set, respectively, with normalization upon the features. The models evaluation was performed on the medical dataset. Hyper parameter tuning was applied to most of the aforementioned models with grid search and 3-fold cross validation. Specifically, the involved hyper parameters are presented in Table 10 for each model. The prediction models were evaluated in subsets of features with increasing dimensionality.

Table 10. *Hyper parameter settings for tuning. GBM: Gradient Boosting Model; LR: Logistic Regression; NN: Neural Networks; NBG: Naïve Bayes Gaussian; RF: Random Forest; SVM: Support Vector Machine.*

Classification Model	**Hyper parameters tuning**
GBM	The number of boosting stages to perform from 10 to 500 with 10 step size The maximum depth of the individual regression estimators from 1 to 10 with 1 step size The minimum number of samples required to split an internal node: 2, 5 and 10 The minimum number of samples required to be at a leaf node: 1, 2 and 4 The number of features to consider when looking for the best split: $\sqrt{n_{features}}$ or $\log_2(n_{features})$
LR	The inverse of regularization strength was tested on 0.001, 0.01, 0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 Algorithm to use in the optimization problem was set to 4 different solvers that handle L2 or no penalty, such as 'newton-cg', 'lbfgs', 'sag' and 'saga' A binary problem is fit for each label or the loss minimized is the multinomial loss fit across the entire probability distribution, even when the data is binary With and without reusing the solution of the previous call to fit as initialization
NN	Both shallow and deep structures were investigated Hidden layers varying from 1 to 3 with different number of nodes per layer (50, 100, 200) Activator function: Relu and $tanh$

	Solver for weight optimization: adam, stochastic gradient descent, stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba and an optimizer in the family of quasi-Newton methods L2 penalty (regularization term) parameter: 0.0001 and 0.05 The learning rate schedule for weight updates was set as a constant learning rate given by the given number and as adaptive by keeping the learning rate constant to the given number as long as training loss keeps decreasing.
NBG	-
RF	The number of trees in the forest from 10 to 500 with 10 step size The maximum depth of the tree from 1 to 10 with 1 step size The minimum number of samples required to split an internal node: 2, 5 and 10 The minimum number of samples required to be at a leaf node: 1, 2 and 4 The number of features to consider when looking for the best split: $\sqrt{n_{features}}$ or $\log_2(n_{features})$ With and without bootstrap
SVM	The regularization parameter was tested on 0.001, 0.01, 0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 Kernel type was set to linear, polynomial, sigmoid and radial basis functions

Results and Discussion

Clustering Results

For this task we use the clustering results from Deliverable 6.3. We ought to mention that K-means achieved better clustering among patient groups. Regarding the identified clusters, the large one includes patients with stable JSN progression or patients that did not present KOA at all in their left and/or right leg, while the second one includes patients with alterations to JSN measures.

Feature Selection Results

Figure 14 illustrates the first 100 features that are selected based on the proposed FS approach separately for the first strategy as well as for the second strategy. From the analysis of the results (Figure 24), we have concluded that the feature categories with the highest contribution seem to come from the symptoms' category and the category of medical imaging outcomes. Indeed, in all cases there is a feature or two from the symptoms' category that were selected first. Then, three imaging outcomes were selected on all three cases. In total, 21, 19, and 20 features of the first 40 selected in the left knee, right knee, and both knees combined, respectively, come from either the symptoms or the imaging outcomes category. Other contributing factors proved to be the nutrition and physical exam outcomes since approximately 20 out of the 100 features were selected in each case. Features from the anthropometrics and medical history categories were selected in all cases. Overall, the main outcome of this analysis is that a combination of heterogeneous features from almost all feature categories is necessary for an accurate prediction of JSN. This highlights that there is a need for a multi-parametric approach in order to handle the complexity and heterogeneity of the available data.

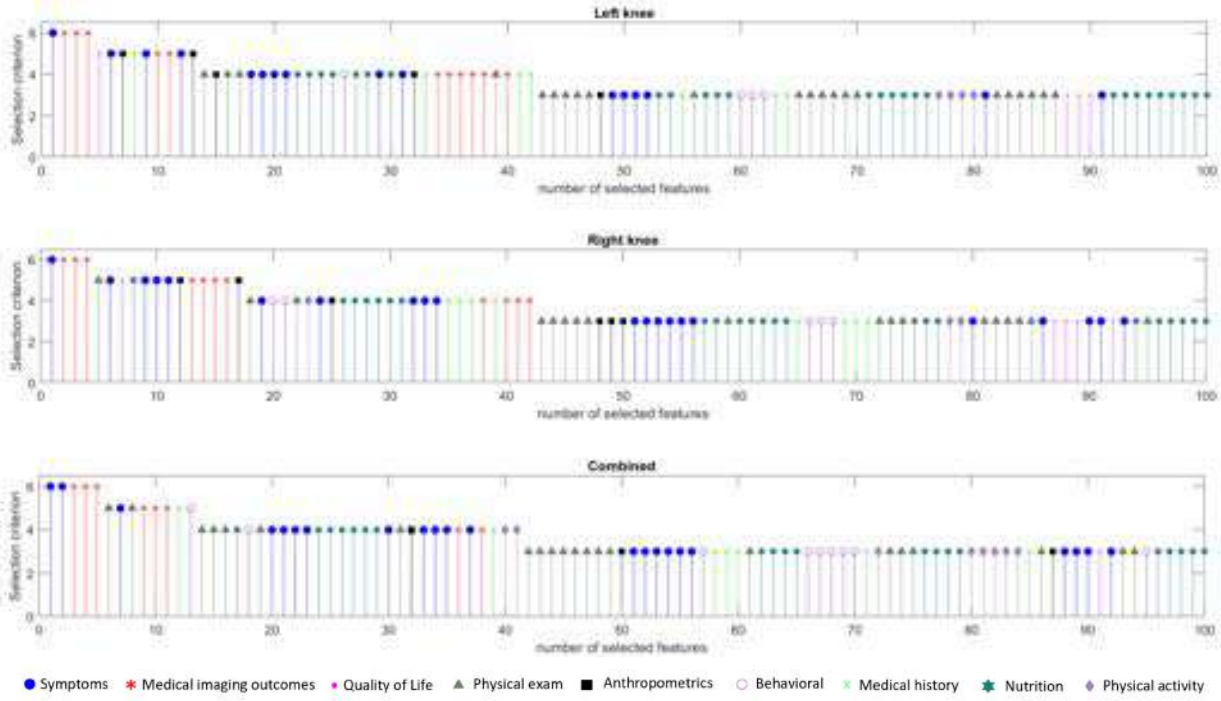


Figure 24. *The first 100 features selected for the left (**top**), the right knee (**middle**), and both legs (**down**).*

Classification Results

Figure 25 shows the alterations in the achieved accuracy over the test set with respect to the number of features (with a step of 2) for the left leg.

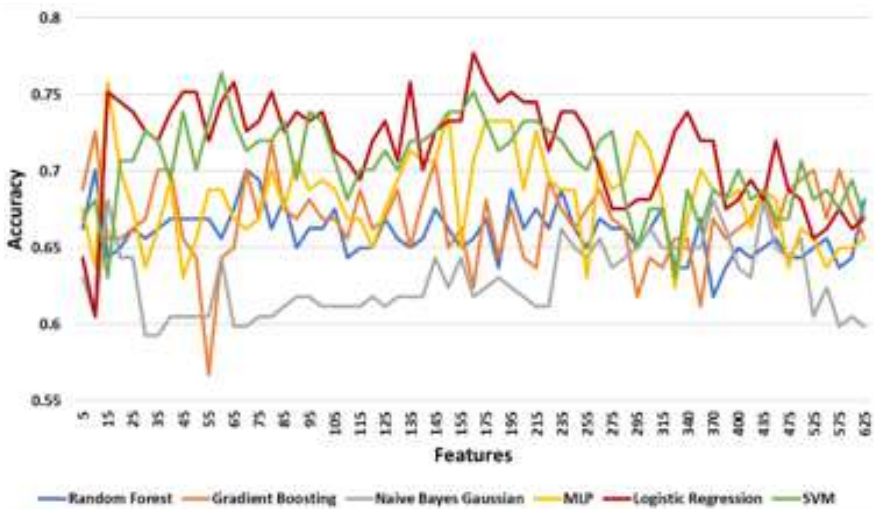


Figure 25. *The accuracy of models over test set for increasing number of features for the left leg. Results are shown with a step size of 5 (two features added at each step).*

Figure 26 shows the alterations in the achieved accuracy over the test set with respect to the number of features with a step of 2 for the right leg.

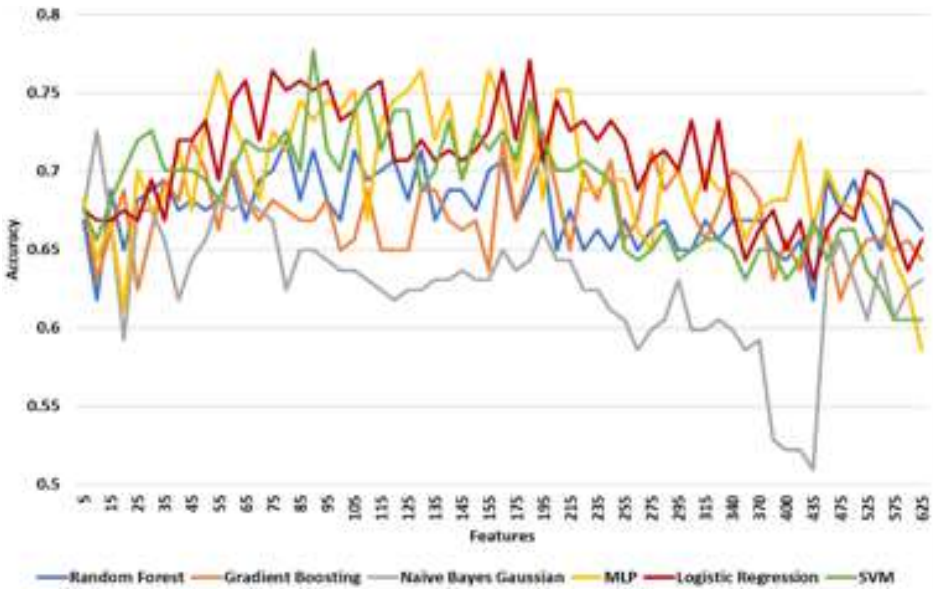


Figure 26. *The accuracy of prediction models over test set for various number of features for the right leg. Results are shown with a step size of 5 (two features added at each step).*

Figure 27 shows the alterations in the achieved accuracy over a test set for various number of features for both right and left legs combined.

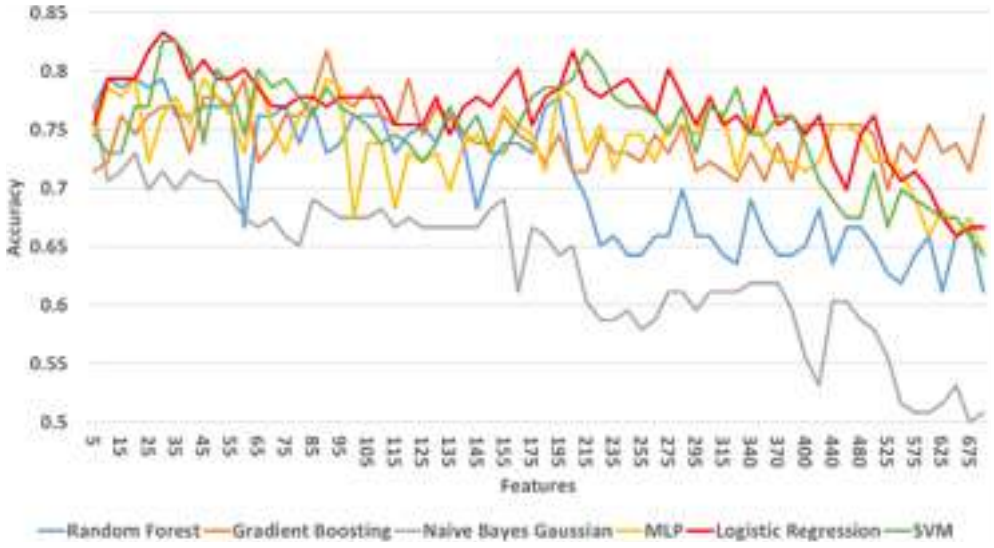


Figure 27. *The accuracy of prediction models over test set for various number of features for the left and right legs combined. Results are shown with a step size of 5 (two features added at each step).*

From the aforementioned classification results on the two proposed strategies (analysis on separate legs and combined) the following remarks can be drawn. Training predictive models using data from one of the two legs leads to inferior results compared to the performance of the model that is trained on data coming from both legs. This can be attributed to the fact that a predictive model trained on data only from the right leg ignores any JSN progression that might happen to the left leg. Due to complex interactions that occur in the dynamics of both legs, predictive models that are trained on data from a single leg are based on partial knowledge of the problem and thus lead to inferior results while requiring a larger number of features.

The need for applying data under-sampling on the dataset could be considered as a limitation of our study. Alternative data resampling algorithms (including more advanced data augmentation techniques such as

generative adversarial networks) have been identified as a future research direction. The use of additional evaluation metrics (other than accuracy) such as precision, recall, or F score would also be beneficial for dealing with the observed data imbalance problem.

3.4 Increasing generalization using an evolutionary Machine Learning approach

Dataset Description

Data were obtained from the Osteoarthritis Initiative (OAI) database (available upon request at https://nda.nih.gov/oai/, accessed on 18 June 2020), which include clinical evaluation data, a biospecimen repository and radiological (magnetic resonance and X-ray) images from 4796 women and men aged 45–79 years. The features considered in this work for the prediction of KL are shown in Table 11. The current study included clinical data from the baseline and the first follow-up visit at month 12 from all individuals being at high risk to develop KOA or without KOA. Specifically, the dataset contains 957 features from eight different feature categories, as shown in Table 11. In addition, our study was based on the Kellgren and Lawrence (KL) grade as the main indicator for assessing the OA clinical status of the participants. Specifically, the variables "V99ERXIOA" and "V99ELXIOA" were used to assign participants into subgroups (classes) of participants whose KOA status progressed or not.

Table 11. *Main categories of the feature subsets considered in this work. A brief description is given along with the number of features considered per category and for each of the two visits.*

Category	Description	Number of Features from Baseline	Number of Features from Visit 1
Subject characteristics	Includes anthropometric parameters (Body mass index (BMI), height, etc.)	36	9
Symptoms	Questionnaire data regarding arthritis symptoms and general arthritis or health-related function and disability	120	80
Behavioral	Includes variables of participants' quality level of daily routine and social behavior	61	43
Medical history	Questionnaire results regarding a participant's arthritis-related and general health histories and medications	123	51 (only medications)
Medical imaging outcome	Medical imaging outcomes (e.g., joint space narrowing and osteophytes)	21	-
Nutrition	Block Food Frequency questionnaire	224	-
Physical activity	Questionnaire data regarding leisure activities, etc.	24	24
Physical exam	Participants' measurements, including knee and hand exams, walking tests and other performance measures	115	26
	Number of features (subtotal):	724	233
	Total number of features:	957	

Problem Definition

In this work, we consider KL grade prediction as a two-class classification problem. Specifically, the participants of the study were divided into two groups: (a) Non-progressors—healthy participants with KL0 or 1 at baseline with no further incidents in both of their knees until the end of the OAI data collection; (b) KOA progressors—participants who were healthy during the first 12 months (with no incident at baseline and the first visit) and then they had an incident (KL ≥ 2) recorded at their second visit (24 months) or later, until the end of the OAI study (Figure 28).

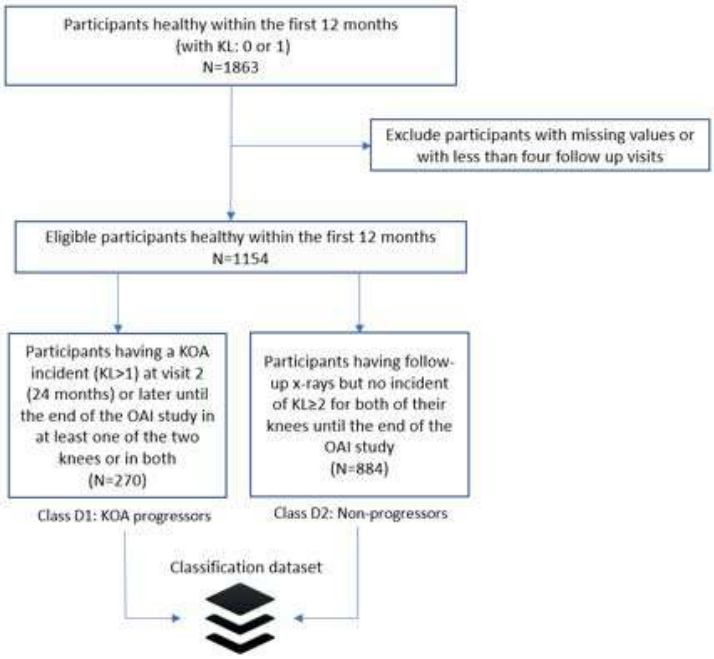


Figure 28. *Stratification of the patients in our study and formulation of the training dataset. Inclusion/exclusion criteria are presented along with the definition of the two data classes (knee osteoarthritis (KOA) progressors and non-progressors).*

Data Pre-Processing

Initially, data cleaning was performed by excluding the columns with more than 20% missing values compared to the total number of subjects. Afterwards, data imputation was performed to handle missing values. As an imputation strategy, mode imputation was implemented to replace missing values of the numerical or categorical variables by the most frequent value of the non-missing variables [42]. Standardization of a dataset is a common requirement for many ML estimators [43]. In our work data were normalized by removing the mean and scaling to unit variance to build a common basis for the machine learning algorithms that followed. After application of the exclusion criteria, classes 1 (KOA progressors) and 2 (non-progressors) comprised 270 and 884 samples, respectively.

Feature Selection

Class imbalance is among the major challenges encountered in health-related predictive models, skewing the performance of ML algorithms and biasing predictions in favor of the majority class. To alleviate this problem, a novel evolutionary feature selection is proposed in this work that overcomes the class imbalance problem and increases the generalization capacity of the finally employed ML algorithm.

The proposed FS is a genetic algorithm-based approach inspired by the procedures of natural evolution (Figure 29). It operates on a population of individuals (solutions), and at each generation, a new population is created by selecting individuals according to their level of fitness in the problem domain (KOA progression in our case). The individuals are then recombined using operators borrowed from natural genetics (selection, reproduction and mutation). This iterative process leads to the evolution of populations of individuals that are better suited to the problem domain. Here, each individual in the population

represents an ML model trained on a specific feature subset to discriminate the aforementioned classes (KOA progressors versus non-progressors). Genes are binary values and represent the inclusion or not of particular features in the model. The number of genes is the total number of input variables in the dataset. Concatenating all genes, a so-called individual or chromosome is formulated that represents a possible solution (feature subset) in our FS problem.

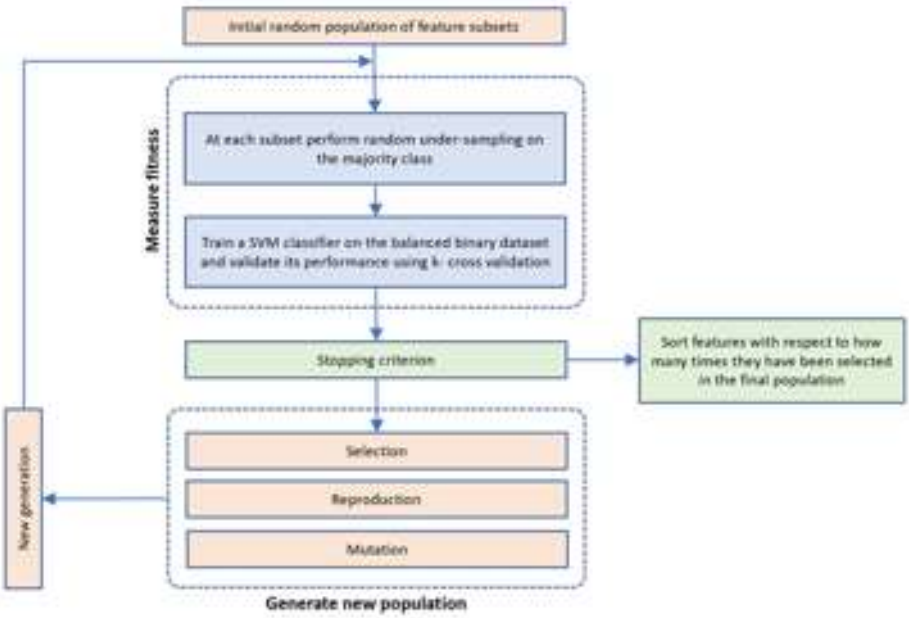


Figure 29. *The proposed GenWrapper feature selection (FS) methodology that includes all the involved processing steps: (i) generation of the initial population; (ii) fitness measurement approach; (iii) stopping criterion; (iv) evolution mechanisms and (v) final feature ranking after the termination of the genetic algorithm (GA).*

The Optimization Toolbox of MATLAB 2020b was used for the implementation of GenWrapper. The proposed FS algorithm proceeds along the following steps:

The Optimization Toolbox of MATLAB 2020b was used for the implementation of GenWrapper. The proposed FS algorithm proceeds along the following steps:

- Step1. Initialization

A group of k chromosomes are randomly generated, forming the initial population of individuals.

- Step2. Fitness assignment

A fitness value is assigned to each chromosome in the population. Specifically, the process of measuring fitness in GenWrapper can be summarized as follows. The following 3-step process (Figure 30) is repeated for each of the chromosomes of the population:

Step 2.1. From the training dataset, we keep only the features that have a value of 1 in the current chromosome. This creates a truncated training set.
Step 2.2. Random undersampling on the majority class is performed on the truncated training set. This action leads to a balanced variant of the truncated training set.
Step 2.3. A classifier is trained on the newly produced balanced dataset. Linear support vector machines (SVMs) have been chosen as the main classification criterion due to their generalization capability.
Step 2.4. A k-fold cross-validation scheme is employed to validate the classifier performance that is finally assigned as a fitness value to the specific individual.

- Step3. Termination condition

The algorithm stops if the average relative change in the best fitness function value over K generations is less than or equal to a pre-determined threshold.

- Step4. Generation of a new population

In case the termination criterion is not satisfied, a new population of individuals is generated by applying the following three GA operators:

Selection operator: The best individuals are selected according to their fitness value.
Crossover operator: This operator recombines the selected individuals to generate a new population.
Mutation operator: Mutated versions of the new individuals are created by randomly changing genes in the chromosomes (e.g., by flipping a 0 to 1 and vice versa).

- Step 6. The algorithm returns to step 2.
- Step 7: Final feature ranking determination

Upon termination of the GA algorithm, the features are ranked with respect to the number of times that they have been selected in all the individuals (chromosomes) of the final population.

Step 7.1. A feature gets a vote when it has a value of 1 in a chromosome of the final generation.
Step 7.2. Step 7.1 is repeated for all the chromosomes of the final generation and the features' votes are summed up.
Step 7.3. Features are ranked in descending order with respect to the total number of votes received.

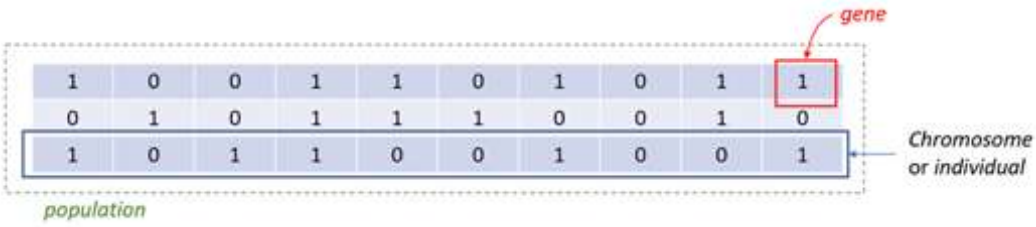


Figure 30. *Definition of genes, chromosomes and population.*

GenWrapper evaluates the fitness of each chromosome (feature subset) by firstly applying random undersampling at the associated dataset (in step 2.2) and then by training an SVM classifier on it (Figure 31). The k-fold cross-validation (CV) performance of the SVM is considered as the fitness of the specific individual. The best individuals (feature subsets that maximize the fitness value) are then selected and combined to generate the new population. This procedure forces the GA to converge to solutions (feature subsets) that generalize well regardless of the specific sampling that has been applied. If a specific resampling process had been applied universally on the dataset before the application of the GA-based FS, then this would lead to overfitting, since the GA algorithm would try to select the best features that fit to the specific data sample. The proposed technique integrates a random sampling mechanism when evaluating each individual, leading to features that generalize well on the whole population. Moreover, the choice of k-fold cross-validation as a validation scheme guarantees that the selected features have high predictive capacity over the whole dataset considered. Another characteristic of the proposed evolutionary FS is the way that features are selected/ranked in the final population. Instead of selecting features from the best individual in the final population, the proposed selection criterion relies on the general performance of features over the whole final population. The best solution (the one with the highest fitness value in the final population) corresponds only to the maximum possible accuracy that can be achieved by a selected feature subset on a specific subset of the whole sample. However, this does not necessarily mean that the best solution generalizes well in the whole sample. Therefore, to achieve the best possible generalization, the proposed FS ranks features with respect to the number of times that they have been selected in all the individuals of the final population. The parameters of the proposed GA-based FS have been properly selected and are cited in Table 12 below.

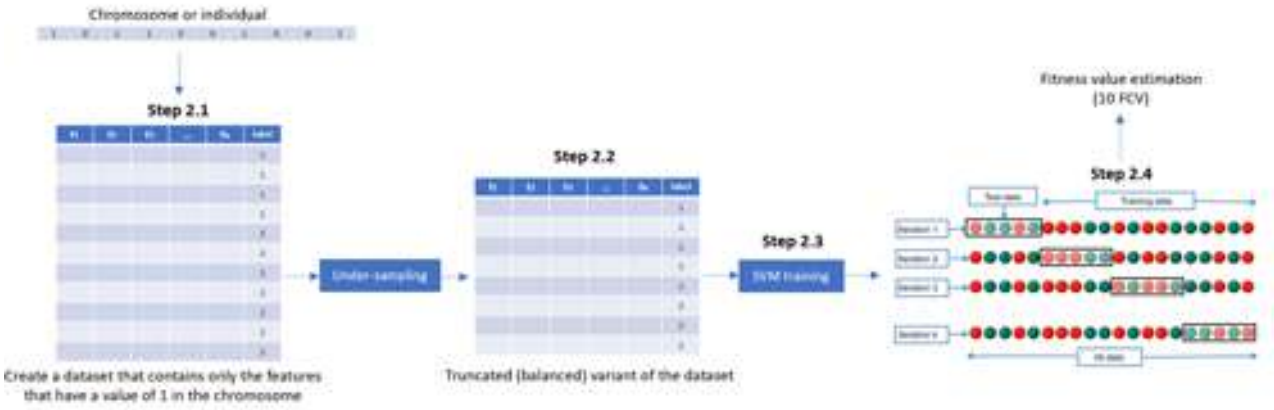


Figure 31. *Proposed mechanism for estimating the fitness of each chromosome within a generation.*

Table 12. *Hyperparameters of the optimized GenWrapper algorithm. A brief description of each hyperparameter is provided along with the finally selected value.*

Parameter	Description	Selected Value
Population size	Number of individual solutions in the population	50
Number of generations	Maximum number of generations before the algorithm halts	100
Mutation rate	Probability rate of being mutated	0.1
Crossover Fraction	The fraction of the population at the next generation, not including elite children, that the crossover function creates.	0.8
Elite Count	Positive integer specifying how many individuals in the current generation are guaranteed to survive into the next generation	5
StallGenLimit	The algorithm stops if the weighted average change in the fitness function value over StallGenLimit generations is less than Function tolerance	50
Tolerance		1×10^{-3}

Learning

Given that the main objective of study is the identification of robust risk factors, two well-known linear ML models (linear regression (LR) and linear SVM) were utilized to evaluate the predictive capability of the selected features. The reason for employing linear models is because (i) they are computationally efficient, so they can be executed multiple times within a repetitive process such as the GA-based algorithm that is proposed in this work, and (ii) they generalize well and, therefore, can be used to assess the generalization performance of the selected features. A brief description of these models is given below.

LR is the most commonly used algorithm for solving classification problems [44]. It is an extension of the linear regression model for classification problems and it models the probabilities for classification problems with two possible outcomes. SVMs are supervised learning models for classification, regression and outlier detection but are more commonly used in classification problems [45]. SVMs are effective in high-dimensional spaces and are still effective in cases where the number of dimensions is greater than the number of samples.

Validation

To evaluate the predictive capacity of the selected feature subset, a repeated cross validation process was adopted using the aforementioned classifiers. Specifically, the validation approach proceeds with the following steps

- Step 1. Random undersampling is applied on the majority class, and the retained samples along with those from the minority class form a balanced binary dataset.
- Step 2. A classifier is built on the balanced binary dataset and its accuracy is calculated using 10-fold cross-validation (10FCV).
- Step 3. Steps 1 and 2 are repeated 10 times, each one using a different randomly generated balanced dataset.
- Step 4. The final performance is calculated by averaging the obtained 10FCV classification accuracies. The resulting final performance will be referred to here as mean 10FCV.

By adopting this repeated validation approach, we guarantee that the selected features are not only suitable for a specific data sample but that they generalize well over the whole dataset. The calculated mean 10FCV performance aggregates the accuracies from 100 training runs (10 repetitions of 10FCV) on different randomly created data samples, forming a reliable measure for estimating the predictive capacity of the selected features.

Results

In this section, we demonstrate the efficiency of the proposed feature selection algorithm in comparison with other well-known FS techniques. The most significant risk factors, as selected by the proposed FS methodology, are also presented.

Selection Criterion

Figure 32 shows the evolution of the proposed fitness value with respect to the number of generations. As it was discussed in Section 2, the mean fitness value is calculated by averaging the fitness values of all the 50 individual solutions in each generation. Each individual fitness value represents the performance of the employed ML model (SVM in our case) on a new, randomly generated balanced dataset (after downsampling the majority class) using k-fold cross-validation. Thus, the mean fitness value aggregates the performance of 50 employed ML models that were trained on slightly different versions of the initially available dataset. As it is observed in Figure 22, the mean fitness value decreases with the number of generations, meaning that the FS converges to a pool of selected feature subsets that have increased classification capacity, regardless of any specific data sampling.

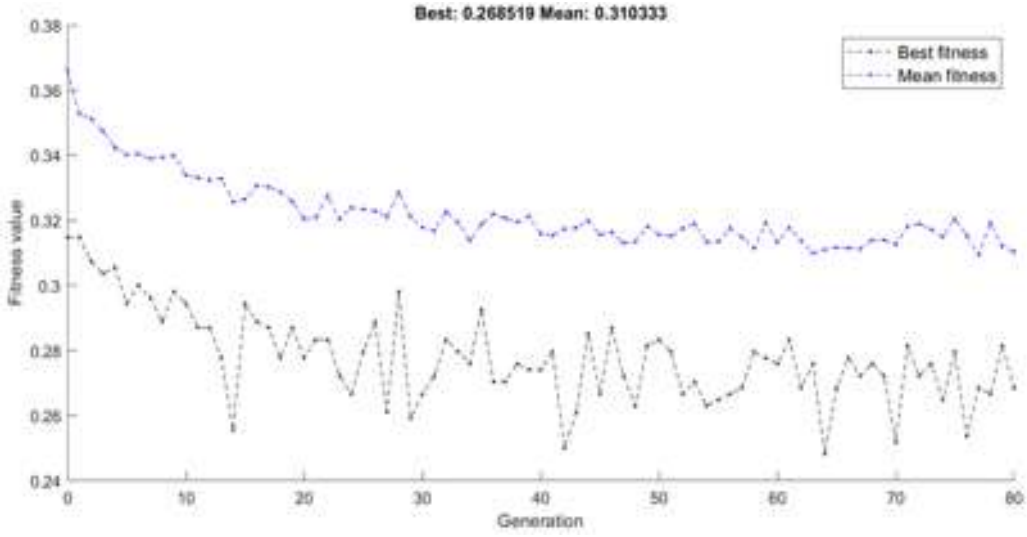


Figure 32. *Fitness with respect to number of generations for GenWrapper. The black and blue dashed lines show the best and the mean fitness achieved at each generation, respectively.*

The dashed black line in Figure 32 represents the minimum fitness values received at each generation of the algorithm. However, as it was noted, the best fitness value (0.26818 in our case) corresponds to a selected feature subset that has been decided based on its performance on a part of the available sample. The proposed scheme, instead of selecting the "best" feature subset of the final generation, proceeds by ranking the available features with respect to the times they have been selected in the 50 different individual solutions of the final generation. Figure 33 illustrates an example of such a ranking where seven features have been selected in all 50 individual solutions, another nine have been selected in 49 individual solutions and so on. The highly ranked features are the ones that are consistently selected by all individual solutions that are generated on different data samples.

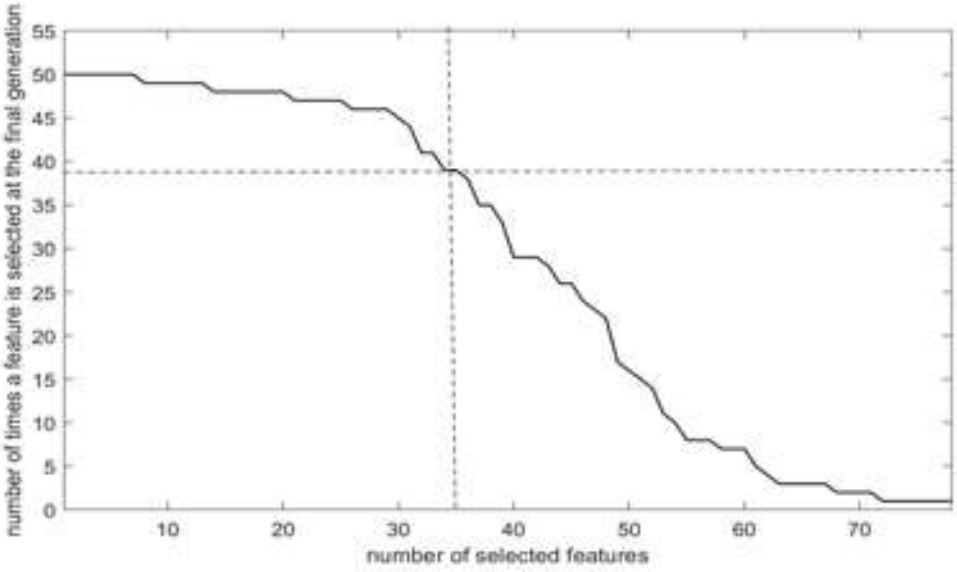


Figure 33. *Feature ranking produced by the proposed FS (the dashed line indicates the number of features that were finally selected).*

To prove the superiority of the proposed feature selection criterion over the "best" individual solution, we performed the following experimentation. Two competing feature subsets were initially extracted: (a) the proposed one that has been selected after selecting the top 35 highly ranked features and (b) the feature subset extracted from the "best" individual solution of the final GA generation (comprising 42 features). The generalization capacity of both features subsets was assessed by employing the repetitive validation approach proposed in this work and the results are shown in Table 13. The proposed feature ranking led to higher accuracy (in terms of mean performance, minimum and maximum accuracies), employing less features (35) compared to the ones selected in the "best" individual solution (42).

Table 13. *Comparative analysis with respect to the final selection of features: proposed feature ranking versus the feature subset of the best individual solution in the final generation.*

FS Criterion	10FCV Accuracy Performed 10 Times				
	Average	**Min**	**Max**	**Std**	**No. of Features**

Feature subset extracted from the "best" individual solution of the final generation	70.10%	67.59%	72.04%	1.13%	42
Proposed feature ranking	71.25%	69.22%	73.33%	1.57%	35

Features Selected

Table 14 cites the 35 features selected by the chosen GenWrapper FS approach. A short description of the features and the categories in which they belong are presented. Seven out of the 35 selected risk factors come from the symptoms category, representing parameters related to pain, swelling, stiffness and knee difficulty, demonstrating the relevance of symptoms in the occurrence and progression of KOA. Moreover, eight features represent diet and nutrition-related parameters that also constitute an important risk factor category. Nine of the features are related to physical activity or exams, whereas another five behavioral risk factors were selected as relevant to KOA progression. Medical history or status estimated through subjective (three self-reported risk factors) or more objective metrics (medical imaging outcomes such as the existence of osteophytes) were also selected by the proposed FS approach. Finally, two parameters describing subject characteristics were among the selected risk factors (specifically the patient's body mass index (BMI) and height).

Table 14. *Characteristics of the 35 most informative risk factors as selected by the proposed GenWrapper.*

Selected Features	Feature Category	Description
P01BMI, P01HEIGHT	Subject characteristics	Anthropometric parameters including height and BMI
KSXRKN1, V00WOMSTFR, KPLKN1, V00WPLKN2, DIRKN16, V00KOOSYML, V00INCOME	Symptoms	Symptoms related to pain, swelling, stiffness and knee difficulty
V00EDCV, V00KQOL4, V00KQOL2, V00CESD9, CEMPLOY	Behavioral	Participants' quality level of daily routine and social behavior and social status
V00RXCHOND, V00RA, V00CHNFQCV	Medical history	Questionnaire data regarding a participant's general health histories and medications
P01SVLKOST	Medical imaging outcome	Medical imaging outcomes (e.g., osteophytes)
V00SUPCA, V00FFQ59, V00FFQSZ13, V00FFQ33, V00SUPB2, V00FFQ12, V00SUPFOL, V00FFQ19	Nutrition	Block Food Frequency questionnaire for daily average, how much each time or for past 12 months
PASE2, PASE6, V00PA130CV	Physical activity	Questionnaire results regarding activities during typical week or past 7 days

RKALNMT, V00lfmaxf, V00rfTHPL, V00lfTHPL, STEPST1, V00rkdefcv	Physical exam	Physical measurements of participants, including tests and other performance measures

Comparative Analysis

The performance of the proposed FS methodology was compared with eight well-known FS techniques in the recent literature. The selected techniques along with their main characteristics are briefly presented below.

A classical wrapper FS was employed in which the feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset. It follows a time-consuming search approach by evaluating all the possible combinations of features against the evaluation criterion. The evaluation criterion is simply a performance measure which depends on the type of problem. Infinite latent feature selection (ILFS) is a probabilistic latent feature selection approach that performs the ranking step by considering all the possible subsets of features, bypassing the combinatorial problem [47]. Unsupervised graph-based filter (Inf-FS) is another FS algorithm proposed, again, by Roffo et al. (2015) [48]. In Inf-FS, each feature is a node in a graph, a path is a selection of features and the higher the centrality score, the most important the feature. It assigns a score of importance to each feature by taking into account all the possible feature subsets as paths on a graph. Correlation-based feature selection (CFS) sorts features according to pairwise correlations [49], whereas LASSO, proposed by Hagos et al. (2017), applies a regularization process that penalizes the coefficients of the regression variables while setting the less relevant ones to zero with respect to the constraint on the sum [50]. In LASSO, FS is a consequence of this process, when all the variables that still have non-zero coefficients are selected to be part of the model. Minimum redundancy maximum relevance (Mrmr) [51] is another well-known FS algorithm that systematically performs variable selection, achieving a reasonable trade-off between relevance and redundancy. A hybrid FS methodology was also employed that combines the outcomes of six FS techniques: two filter algorithms (Chi-square and Pearson correlation), three embedded ones (LightGBM, logistic regression and random forest) and one wrapper (with logistic regression) [52]. In this approach, all six FS techniques are applied separately, with each one resulting in a selected FS, and the final feature ranking is decided on the basis of a majority vote scheme. PCA is a well-known feature reduction method that reduces the dimensionality of data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs. The MATLAB-based feature selection library FSLib 2018 (https://www.mathworks.com/matlabcentral/fileexchange/56937-feature-selection-library, accessed on 30 January 2021) was used for the implementation of the competing FS algorithms on a research workstation with Intel Core i7-7500 processor, 2.70 GHz CPU (16 GB RAM).

Figure 34 depicts the results of the comparison between the proposed GenWrapper FS and a classical wrapper FS technique. Specifically, the obtained mean 10FCV accuracies are shown with respect to the number of features as they have been ranked by the two compared approaches using two classifiers (LR and SVM). The following remarks can be extracted from Figure 24:

- GenWrapper significantly outperforms the classical wrapper FS, especially for a small number of selected features (up to 20). This superiority is proven for both SVM and LR;
- GenWrapper employing SVM gives the best overall performance (71.25% at 35 selected features).

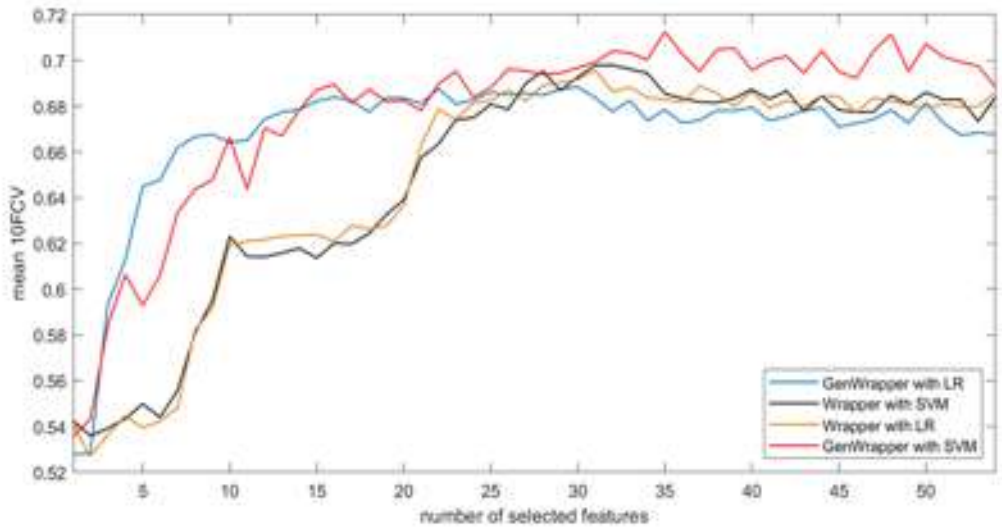


Figure 34. *Accuracy (mean 10-fold cross-validation (10FCV)) with respect to selected features (curves): GenWrapper versus a classical wrapper using two classifiers (support vector machine (SVM) and logistic regression (LR)).*

Figure 35 shows the progression of the mean 10FCV accuracy with respect to the number of selected features for the proposed FS and the other seven competing FS techniques (CFS, ILFS, Inf-FS, Lasso, Mrmr, PCA and hybrid). In this comparative analysis, a linear SVM classifier were employed by all techniques since it proved to be the most efficient ML model. GenWrapper is the best-performing technique, achieving high accuracies (3.4% higher than the second best). Hybrid FS and Mrmr were the second and third best performers, achieving accuracies of 67.85% and 67.29%, respectively. Mrmr was very successful at the first 10 selected features but then it reached a threshold within the range of 67–68%, whereas the inclusion of further features had a minor or even negative effect on the classification performance. The rest of the FS techniques had moderate performances (61.97–65.11%).

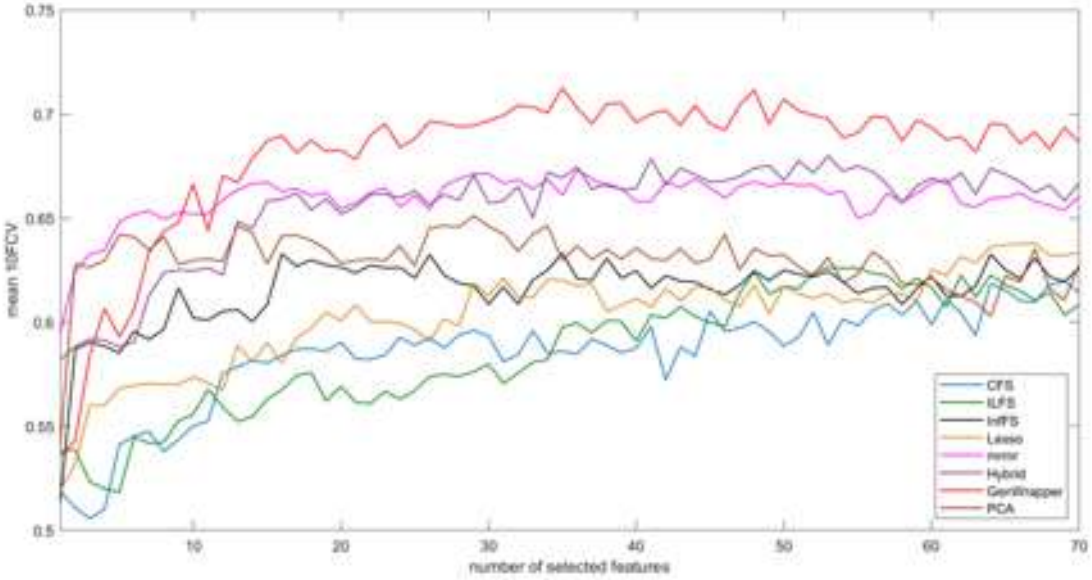


Figure 35. *Accuracy (mean 10FCV) with respect to selected features: GenWrapper versus the remaining competing FS techniques. SVM was used for the classification task for all eight FS techniques.*

Discussion of results

Predicting KOA onset and its further progression is among the best strategies to reduce the burden of the disease. Risk factors for incident OA may differ from those for OA progression given that the incidence

and progression of radiographic knee OA may involve different processes [53,54]. Several risk factors have been reported to be associated with the incidence of knee OA [55,56]. However, our understanding about predictive risk factors associated with KOA progression is limited due to the fact that the number of studies, in which risk factors and incidence of knee OA have been investigated longitudinally, is relatively small. This study contributes to the identification of robust risk factors for knee OA progression as a first, but very important, step toward achieving the goal of developing preventive strategies and intervention programs and finally reducing the incidence and associated morbidity of knee OA.

Identifying important features from an imbalanced data set is an inherently challenging task, especially in the current KOA prediction problem with limited samples and a massive number of features. Feature selection algorithms employing data resampling have been typically utilized to reduce the feature dimensionality and at the same time to overcome the class imbalance challenge. Oversampling algorithms randomly replicate examples from the minority class which in some scenarios can facilitate the FS process but is also prone to overfitting [57]. In data under-sampling, examples from the majority class are randomly discarded in order to rectify the disparities between classes. However, informative samples might be discarded from the final training set, reducing the generalization capabilities of the finally selected risk factors. New approaches are needed to address the intersection of the high dimensionality and imbalanced class problems due to their complicated interactions.

To cope with all the aforementioned challenges, the proposed FS technique incorporates a number of features aiming towards the identification of robust risk factors (with increased generalization capacity) extracted from a highly imbalanced dataset. GenWrapper relies on a stochastic method for function optimization based on the mechanics of natural genetics and biological evolution. This stochastic search is employed to identify a globally optimal feature subset, compared to a costly search that makes local decisions. The proposed FS performs better than traditional feature selection techniques, can manage datasets with many features and does not need any specific knowledge about the problem under study. Compared to traditional GA-based FS algorithms, GenWrapper applies random undersampling at each individual solution, forcing the GA to converge to solutions (feature subsets) that generalize well regardless of the applied data sampling. K-fold cross-validation is utilized to measure the fitness of each individual solution, guaranteeing that the selected features have high predictive capacity over the whole dataset considered. Finally, instead of selecting the "best" individual of the final population, the proposed FS ranks features with respect to the number of times that they have been selected in all the individual solutions of the final population. This leads to selected features that consistently work well at any possible data sample and, thus, have increased generalization capacity with respect to KOA progression.

Linear classifiers were employed on this study, and this choice can be attributed to the fact that evidence of linear separability between the two classes (progressors versus non-progressors) was identified in previous studies of the authors on the same problem. Specifically, as it was reported in [52], LR and linear SVMs outperformed all the competing non-linear models (including Random Forest, XGboost, KNN and decision trees) on the same problem of predicting KOA. This finding highlight that the power of the proposed technique lies on the selection of robust and informative risk factors, whereas the complexity of the finally employed classification models plays a less crucial role.

The performance of the proposed FS methodology was compared with eight well-known FS techniques in the recent literature. GenWrapper employing SVM led to the overall best performance (71.25% at 35 selected features), significantly outperforming all the competing algorithms. Specifically, it proved to be more accurate than the classical wrapper FS (which was the second-best approach), and this superiority was more evident for a small number of selected features (up to 20). GenWrapper was also much more effective (at least 3.4% more accurate) than the other seven competing FS techniques (CFS, ILFS, Inf-FS, LASSO, Mrmr, PCA and hybrid). Finally, apart from being the most accurate approach, GenWrapper was prove to also be the most consistent FS technique, with the great majority of the obtained 10FCV accuracies being

higher than 70%, whereas all the other competing FS algorithms led to inferior and less consistent accuracies.

4. Personalised Diagnosis models

Knee OA has a higher prevalence rate compared with other types of OA. In recent years, it has been realized that KOA results from a multifactorial, complex interplay of constitutional and mechanical factors, including joint integrity, mechanical forces, local inflammation, genetic predisposition and biochemical processes. Furthermore, KOA is closely associated with obesity and age [1]. The specific disease causes significant problems when it occurs. The main consequences are: 1) low quality of life, due to severe pain and stiffness, 2) social exhaustion due to low public participation and 3) low levels of psychology and resignation from life [3]. Due to the multifactorial nature of KOA, disease pathophysiology is still poorly understood and diagnosis tools are under current investigation.

Diagnosis, prognosis, or treatment of KOA is a challenge for the scientific community. Increasing data collection has led to an increasing number of studies employing big data and AI analytics applied in the KOA research. As a result of this, several techniques have been reported in the literature in which Machine learning (ML) models were used to diagnose KOA [4]. In 2016, Yoo et al. [58] developed a new easy-to-calculate self-assessment scoring system making use of a large population dataset. The main finding was the identification of patients at high risk of knee OA who need treatment before aggravation. An ANN model was constructed with AUC of 0.66-0.88. In another study, Long et al. [59] used outcome scores (KOOS) and biomechanical gait parameters for the identification of parameters which are associated with functional and quality of life outcomes for injury and knee OA in comparison with health subjects. Furthermore, Lim et al. [60] developed a deep learning (DL) model for the detection of KOA by using demographic, personal characteristics, lifestyle and health status related variables. In additional, Moustakidis et al. by using self-reported clinical data (such as symptoms, disability, function and general health) in subsets developed different ML models as well as DL architectures for the KOA diagnosis [17]. Moreover, in 2019 Christodoulou et al. used self-reported clinical data for the investigation of the DL capabilities in diagnosis of KOA. The potential of this approach was demonstrated by classifying different subgroups of control participants from self-reported clinical data [61]. According to the literature, there is a lack of knowledge on the self-reported clinical data contribution of the diagnosis and training of classifiers.

Identification of risk factors for developing Knee OA has been limited by an absence of non-invasive methods to inform clinical decision making and enable early detection of people who are most likely to progress to severe KOA. The first approach of this section contributes to the diagnosis of KOA through the identification of risk factors based on a robust feature selection technique and well-known ML models of machine learning. Furthermore, the second approach in this section makes a contribution towards KOA diagnosis through the application of various machine intelligence models on self-reported clinical data (such as symptoms, disability, function and general health) from the osteoarthritis initiative study. Different machine learning models as well as deep learning architectures were tested with respect to their ability to recognise participants with symptomatic KOA or being at high risk of developing KOA in at least one knee. The effect of various feature subsets was also investigated. These feature categories are related to (i) the temporal occurrence of symptoms, (ii) symptoms' type and (ii) participants' quality of life status. WOMAC and KOOS features were also evaluated for their capacity to diagnose KOA. Finally, the best performing approach (deep learning) was demonstrated in subgroups defined by gender and age. A quantum perspective of the application of deep learning techniques for the task of OA diagnosis is also given in the discussions.

4.1 Diagnosis of KOA based on KL grade

Data description

Data was obtained from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/). Specifically, the current study only includes clinical data from baseline from all individuals with or without KOA (Table 15). Furthermore, our study was based on Kellgren and Lawrence (KL) grade as outcome for the classification.

Table 15. Main categories of the feature subsets considered in this work.

Category	Description
Accelerometry	Variables that describe whether a person is physically active
Biomarkers	Variables which describe the collection of biospecimens
Joint symptoms/function	Questionnaire results regarding arthritis symptoms and general arthritis or health-related function and disability
Medical history	Questionnaire data regarding a participant's arthritis-related and general health histories
Nutrition	Variables which collected using the modified Block Food Frequency questionnaire
Physical exam, measurements	Variables which contain physical measurements of participants, including height, weight, BMI, abdominal circumference, blood pressure, isometric strength, knee and hand exams, walking tests, and other performance measures
Subject characteristics, risk factors	Variables which contains demographic information and other descriptive information about enrolled OAI participants

Subsequently, the 4796 samples of the dataset were divided into two equal categories, Classes became equal by reducing the number of the majority class to the number of the minority, at random as follows:

- Class 1: KOA: This class comprises of 1936 participants who have KL >=2 at baseline. These participants had KL grades equal or higher than 2 in at least one of the two knees or in both.
- Class 2: Non-KOA: This class involves 1936 participants with KL0-1 at baseline. In particular, these participants do not have KOA in any of their knees.

Methodology

The proposed in this work ML methodology for KOA diagnosis includes five processing steps: data pre-processing of the collected clinical data (707 features in total), feature selection, learning process and evaluation of the classification results. More details about the proposed methodology are presented in the following sections.

Pre-processing and Feature Selection (FS)

The steps for pre-processing and Feature selection have described in Deliverable 6.3.

Learning Process

Various ML models were evaluated for their suitability in the task of KOA classification. A brief description of these models is provided below.

- XGboost is a popular and efficient implementation of the Gradient Boosted Trees algorithm. It is a supervised learning method that is based on function approximation by optimizing specific loss functions as well as applying several regularization techniques. Specifically, this model is a sum of CART (tree) learners which try to minimize the log loss objective and the scores at leaves. These scores are actually the weights that have a meaning as a sum across all the trees of the model. Furthermore, they are always adjusted in order to minimize the loss [23].
- Random Forest classifier is ensemble algorithm. Ensembled algorithms are those which combines more than one algorithm of same or different kind for classifying objects. Random forest classifier creates a set of decision trees from randomly selected subsets of training set. It then aggregates the votes from different decision trees to decide the final class of the test object [22].
- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. DTs are simple to understand and to interpret. They require little data preparation and perform well even if their assumptions are somewhat violated by the true model from which the data were generated [19].
- Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Naive Bayes learners and classifiers can be extremely fast. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution [24].
- Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier's detection. They are effective in high dimensional spaces and still effective in cases where the number of dimensions is greater than the number of samples. Furthermore, SVMs use a subset of training points in the decision function, called support vectors [45].
- K-Nearest Neighbor (KNN) is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given "unseen" observation. Similarity is defined according to a distance metric between two data points. A popular one is the Euclidean distance method [62].
- Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. The interpretation of the weights in logistic regression differs from the interpretation of the weights in linear regression, since the outcome in logistic regression is a probability between 0 and 1 [44].

Hyperparameter selection was implemented to optimize the performance of our models and to avoid the overfitting and the bias error. Each model was optimized with respect to a number of parameters. Specifically (i) gamma, maximal depth, minimum child and weight were optimized for XGboost, (ii) criterion, minimum samples leaf, minimum samples split and number of estimators for Random Forest, (iii) maximal features, minimum samples and minimum number of decision splits for Decision Trees, (iv) C and kernel for SVMs, (v) leaf size and k-parameter for KNN and (vi) penalty and C for Logistic Regression.

Validation

A 70%-30% random data split was applied to generate the training and testing subsets, respectively. Learning of the ML was performed on the stratified version of the training sets and the final performance was estimated on the testing sets.

Results

In this section, we present the most important risk factors as they have been selected by the proposed hybrid FS methodology. Moreover, the overall performance of the models is presented in relation to the number of features and then reference is made to the models with the highest accuracies. Finally, an explainability analysis is performed on the model that was chosen as the best.

Table 16. *First ten selected features in order of importance, their description and the number of appearances in their selection criteria.*

Feature	Description	Criterion
V00LKPFCRE	Left knee exam: patello-femoral crepitus present on exam	6/6
V00KSXRKN1	Right knee symptoms: swelling, last 7 days	6/6
V00KSXLKN5	Left knee symptoms: bend knee fully, last 7 days	6/6
V00KSXLKN1	Left knee symptoms: swelling, last 7 days	5/6
P02ELGRISK	Knee symptoms, risk factors, or both, status at IEI/SV	6/6
P01KSURG	Right knee, ever have surgery or arthroscopy	6/6
V00WTMAXKG	Maximum adult weight, self-reported (kg)	5/6
V00RKPFCRE	Right knee exam: patello-femoral crepitus present on exam	5/6
V00KQOL1	Quality of life: how often aware of problems with knee(s)	5/6
V00ABCIRC	Abdominal circumference (cm)	5/6

Table 16 gives a short description of the first ten (10) selected features along the associated votes that were assigned to each one. Most of the selected features are related to symptoms, weight, medical history (e.g., surgery) and data from the physical examination of the subjects.

Figure 36 depicts the testing performance (%) of the competing ML models with respect to the number of selected features. In particular, DTs failed in this task, recording low testing performances. In contrast, the other models had an upward trend in the first 20-40 features, followed by a steady testing performance. Specifically, the Logistic Regression model with respect to selected features showed an upward trend in 30-40 features, with a maximum at 40 features. Then with the addition of new features, it showed a relatively stable testing performance. Best overall performance was achieved by Logistic Regression at 40 features

whereas the inclusion of additional features led to a small reduction in the accuracies achieved.

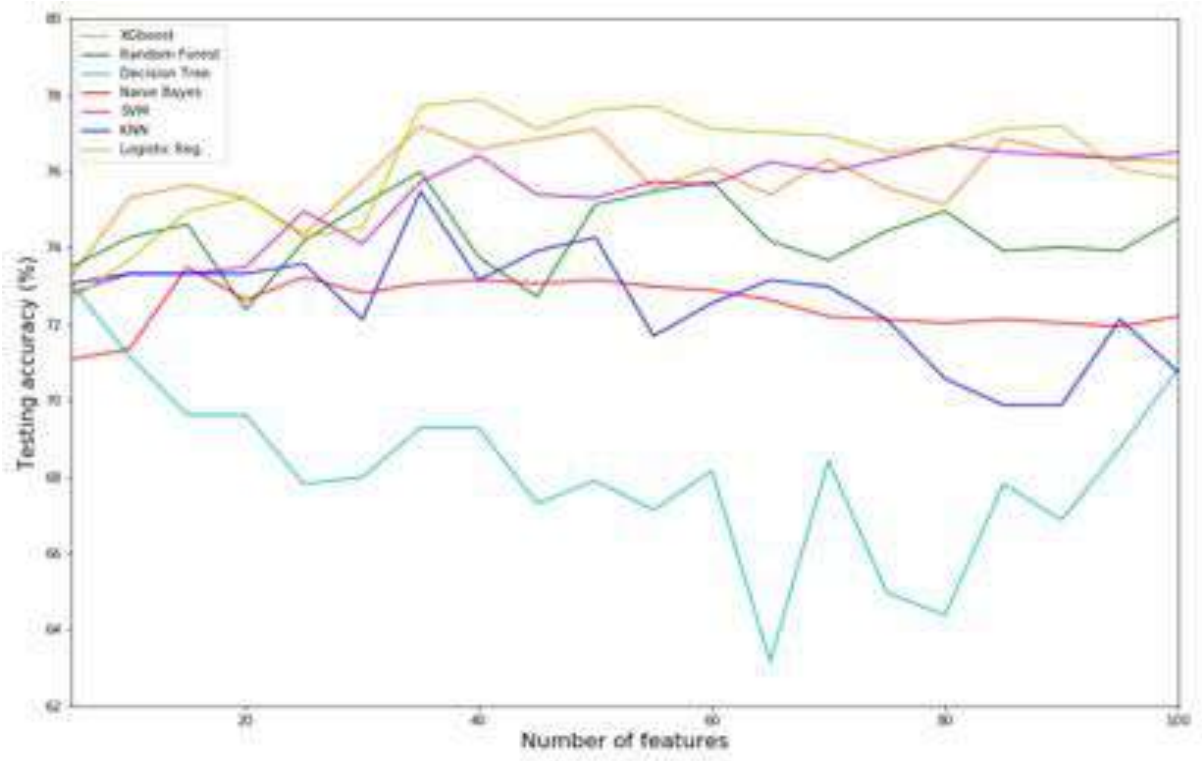


Figure 36. *Learning curves with testing accuracy scores for different ML models trained on feature subsets of increasing dimensionality.*

Discussion of results

The main finding of this work is that explainability of ML models can play an important role on identifying the impact of different risk factors in the diagnosis of KOA. Based on the results of a thorough comparative investigation of various ML models on the KOA diagnosis, it was observed that the correct choice of features plays a key role in the specific problem, while the capacity of the classifiers comes in second place. In light of this, the proposed ML workflow for diagnosis of knee osteoarthritis was focused not only on the detection accuracies achieved but also on the post-hoc explainability of the generated ML models (Deliverable 9.3).

4.2 Machine Learning and Deep Learning Diagnosis of KOA with focus on patients subgroups

Data Description

Data was obtained from the osteoarthritis initiative (OAI) database which is a multi-centre prospective longitudinal cohort study designed to identify risk factors associated with the incidence and progression of KOA [63]. Launched in 2002, OAI began enrolling people, aged 45–79 years, with symptomatic KOA or being at high risk of developing KOA in at least one knee in four US medical centres. In total 4796 participants were recruited and followed over an 8-year period with a follow-up rate of more than 90% over the first 48 months.

The current study only includes self-reported data about joint symptoms, disability, function and general health from all individuals with or without KOA from the baseline visit. The selected dataset, that comprises of 141 risk factors from 4796 participants, was further separated into 10 overlapping feature subsets with different characteristics. Three subsets are relevant with the temporal occurrence of symptoms, four subsets refer to different types of self-reported symptoms and one involves features related to health, emotional problems, lifestyle and psychology. Hybrid metrics related to WOMAC and KOOS have been also

considered as separate sets. The effect of each feature subset on the KOA diagnosis was investigated in the following sections of the work, providing insights about their clinical significance. Table 17 cites the main characteristics of the 10 feature subsets considered in our work.

Table 17. Main characteristics of the feature subsets considered in this work.

Category	**Num. of features**	**Feature category**	**Description**
Temporal occurrence of symptoms	68	past week	Any type of symptoms over the past 7 days
	10	past month	Any type of symptoms over the past 30 days
	13	past year	Any type of symptoms over the past 12 months
Type of symptoms	64	Pain	Features related to pain in various activities for both knees, hips and joints in all time intervals
	27	Stiffness	Features related to stiffness in all the time intervals
	37	Knee difficulty	Knee difficulty on either right or left leg on various activities in all time intervals
	12	Other symptoms	Symptoms such as swelling, grinding sensation, knee catch or hang up in all time intervals
Quality of life	15	Quality of life	Features related to health, emotional problems, lifestyle, psychology
Hybrid metrics	8	WOMAC	Indexes which consist a score of questions about pain, symptoms and quality of life for both of knees
	5	KOOS	Indexes which consist a score of questions about pain, stiffness and disability for both of knees

Furthermore, the 4796 samples of the dataset were divided into three categories as follows:

- Class 1: Incidence: This class comprises of 3284 participants who do not have symptomatic tibiofermoral knee OA at the screening clinic visit in at least one knee, but who do meet the risk factor eligibility criteria for their age group.
- Class 2: Progression: This class involves 1390 participants with frequent knee symptoms, which are defined as "pain, aching or stiffness in or around the knee on most days". These participants had knee symptoms on most days of 1 month of the preceding year and radiographic tibiofemoral knee OA (Osteoarthritis Research Society International (OARSI) atlas grades 1–3) on a fixed-flexion radiograph at recruitment in at least one knee.

- Class 3: Non-exposed control group: 122 participants have been assigned in this class without any knee symptoms in either knee, who do not have any of the eligibility risk factors and who have OARSI grade 0 in both tibiofermoral compartments for osteophytes.

The following 2 classification problems were investigated in this work: (a) a *2-class problem* with the objective to discriminate participants belonging to class 1 (*progression*) and class 2 (*incidence*), (b) a *3-class problem* that is a multi-class classification problem where all three classes were considered in the training and testing datasets. It should be noted that class 3 is much smaller than the other two thus setting a highly imbalanced data challenge.

Methodology

The proposed in this work machine intelligence methodology for OA classification includes three processing steps: data pre-processing to handle missing values and normalise the collected clinical data, a learning process for training, and evaluation of the classification results, as illustrated in Figure 37. The proposed methodology is thoroughly presented in the following sections.

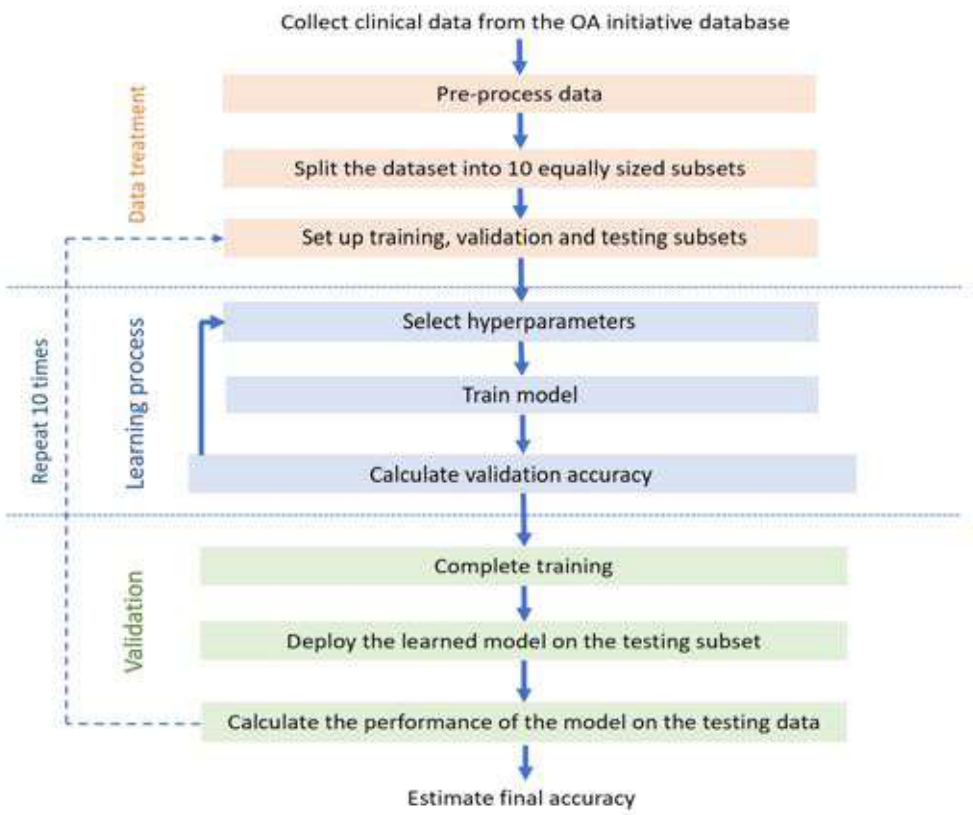


Figure 37. *Flowchart of the proposed machine intelligence methodology.*

Pre-processing

Mean imputation was performed to handle missing values. Specifically, for numerical features missing values were replaced by the mean feature value. In case of categorical features, the most frequent category was used to replace NaNs. Since activation functions of DNNs do not generally map into the full spectrum of real numbers, we first standardized our data to be drawn from N(0; 1). Normalization also allowed us to

compute more precise errors in this standardized space, rather than in the raw feature space. Data resampling was employed to cope with the class imbalance problem.

Learning process

Various machine intelligence models were evaluated for the suitability in the task of OA classification. Both machine learning and deep learning techniques were investigated, as described below.

Machine learning models: We tested linear discriminant analysis (LDA) [64] to provide a baseline for comparisons with more advanced models. We also evaluated decision trees [65,66] driven by Gini's diversity index, KNN and weighted KNN [67], as well as non-linear support vector machines (SVM) algorithms with Gaussian kernel [68], which can deal with the overfitting problems that appear in high-dimensional spaces. The ensemble techniques AdaBoost [69] and Random Forest [70] were also evaluated using DT models as weak learners. Three fuzzy based algorithms were also tested including Fuzzy K-Nearest Neighbors (FKNN) and Fuzzy Nearest Prototype classifier (Fuzzy NPC) by Keller and Gray [71] as well as Condensed Fuzzy K-Nearest Neighbors (CFKNN) by Zhai [72].

Deep learning models: Deep learning [73] holds great promise to fulfil the challenging needs of various industries including data-driven healthcare. It performs human-like reasoning and extracts compact features which embody the semantics of input data. Deep neural networks are stacked layer models in which a series of layers is connected together including an input layer, an output layer and a few hidden layers placed between them. The number of nodes in the input and output layers correspond to the dimensionality of the input and the target data, respectively. The nonlinear relationship between the DNN layers is indicated by the following equations:

$$z_j^l = \sum_i w_{i,j}^l \, x_i^{l-1} + b_j^l$$

$$h_{W,b}(x) = f\left(z_j^l\right) = f\left(\sum_i w_{i,j}^l \, x_i^{l-1} + b_j^l\right)$$

where x_j^l is the activation value of neuron *j* in layer l; z_j^l is a linear activation combination of neurons in the previous layer; b_j^l is the bias value of neuron *j* in layer *l*; $w_{i,j}^l$ is the weight parameter between node *i* in layer *l−1* and node *j* in layer *l*; and f(.) is the activation function.

Our DNN models use fully connected, dense neural layers where the output of one layer serves as the input for the next layer. A number of different DNN structures were investigated in this work with varying: (i) input dimensionality (as described in Table I), (ii) number of hidden layers and (iii) number of nodes per hidden layer. Rectified linear activation was selected given that it has demonstrated high performance on a variety of recognition tasks, and is a more biologically accurate model of neuron activations [74]. The final neural layer reduces the dimensionality to either 2 or 3 nodes using Softmax as activation function. Adaptive learning rate was employed with ADADELTA [75] that automatically combines the benefits of learning rate annealing and momentum training to avoid slow convergence. Weight initialisation was performed using uniform distribution. Early stopping was implemented based on the convergence of the *logloss* metric.

Validation

Ten-fold cross validation (10FCV) was used to evaluate the effectiveness of the learned classification models. The dataset was split into 10 subsets, called folds. The train-test method was applied iteratively by

using each of the 10 folds for testing, while the learning model was trained with the remaining nine. The performance was calculated by averaging the individual ten test scores. To achieve a fair comparison between the different approaches, hyperparameter selection was performed for each one of the investigated machine and deep learning algorithms. A validation subset was held out from the training sets (a randomly selected 10%) as a criterion for selecting the optimum hyperparameters by means of a grid search process.

Results and Comparisons

Comparative analysis

This subsection cites the results of a comparative analysis over a number of well-established machine learning and deep learning models on the problems of 2-class and 3-class classification using the entire feature sets. Cross validated results are shown in Table 18, whereas the optimal hyperparameters are highlighted per model. Each model was optimized on the validation subsets with respect to the following parameters: (i) minimum leaf size and maximal number of decision splits for Decision Trees, (ii) C and sigma for SVMs, (iii) k-parameter for KNN, Fuzzy KNN, Fuzzy NPC and CFKNN, (iv) number of weak learners and weak learner type for Adaboost and Random Forest and (v) number of hidden layers and number of nodes per layer for Deep Neural Networks.

The best overall performance on the 2-class problem (80.74%) was achieved by the DNN model with 3 hidden layers and 50 nodes per layer. DNN also outperformed all the rest ML models in the 3-class problem demonstrating at the same time the highest level of accuracy stability over the 10 testing folds (79.5% overall accuracy with a standard deviation of 1.2). However, this accuracy comes with an increase of computational complexity since DNN was the slowest in its execution with 31.5s and 36.4s training time for the 2-class and 3-class problems, respectively. KNN was the fastest algorithm with 0.016s and 0.03s execution time for the 2- and 3-class problems achieving moderate performances. Statistical significance analysis was also performed by applying t-tests at the confidence level of 5% on the accuracies obtained on the 10 CV data folds. The results of DNN were significantly different from the majority of the rest models for both 2-class and 3-class problems. No significant differences were obtained on the results of DNN, SVM, Adaboost and Random Forest in the 2-class problem and the results of DNN and SVM for the 3-class problem.

Table 18. Comparative analysis between the best DNN models and state-of-the-art ML models

Model type	**10 fold cross validation accuracy (%)**			
	2 classes		**3 classes**	
	Overall (std)	**Time (s)**	**Overall (std)**	**time**
Decision trees (minimum leaf size: 5, Split criterion: Gini's index, Maximal number of decision splits: 7)	79.3* (2.1)	0.22	77.7* (2.0)	0.26
Linear Discriminant	80.1* (2.3)	0.07	76.2* (2.8)	0.08
SVM Gaussian (C=1, sigma =0.15)	80.2 (1.05)	2.8	79.1 (1.34)	3.2

KNN (k=9)	79.1* (1.8)	**0.016**	76.9* (2.2)	**0.03**
Fuzzy KNN (k=11)	79.2* (1.33)	0.034	77.39* (1.45)	0.06
Fuzzy NPC (k=5)	77.8* (1.24)	0.09	72.4* (1.9)	0.11
CFKNN (K=9)	78.6* (1.06)	0.1	73.6* (2.05)	0.14
Adaboost (number of weak learners: 130, Maximal number of decision splits: 1024, weak learner: DT)	80.6 (1.33)	25.6	78.6* (1.2)	28.7
Random Forest (number of weak learners: 130, Maximal number of decision splits: 4, weak learner: DT)	80.1 (1.1)	5.1	77.7* (1.86)	5.5
Deep Learning (Adam optimization, ReLU functions, adaptive learning rate, 3 hidden layers, 50 nodes per layer)	**80.7 (1.1)**	31.5	**79.5 (1.2)**	36.4

*Significantly different from DNN ($p < 0.05$) by applying t-tests on the 10FCV accuracies over the 10 data folds

The classification performance of the best performing models (in which no significant statistical differences were identified) was further evaluated with respect to various validation metrics including confusion matrix, class precision, sensitivity and specificity.

Table 19. Confusion matrix for the best DNN architecture on the 2-class problem using the entire feature set

Model		**Incidence**	**Progression**	**Precision**	**Sensitivity**	**Specificity**	**Overall accuracy**
DNN	**Incidence**	2593	691	78.96%	**92.54%**	63.08%	**80.74%**
	Progression	209	1181	**84.96%**			
SVM	**Incidence**	2633	651	80.17%	90.54%	63.13%	80.19%
	Progression	275	1115	80.21%			
RF	**Incidence**	2720	564	**82.82%**	88.25%	**64.57%**	80.1%
	Progression	362	1028	73.95%			

ADA	**Incidence**	2617	667	79.68%	91.59%	63.29%	80.59%
	Progression	240	1150	82.73%			

Table 20: Confusion matrix for the best DNN architecture on the 3-class problem using the entire feature set

Model		**Incidence**	**Progression**	**Non-exposed**	**Per class accuracy**	**Overall accuracy**
DNN	**Incidence**	2813	431	40	**85.65%**	**79.50%**
	Progression	442	948	0	68.20%	
	Non-exposed	70	0	52	**42.62%**	
SVM	**Incidence**	2767	472	45	84.25%	79.08%
	Progression	375	1014	1	**72.94%**	
	Non-exposed	110	0	12	9.83%	
RF	**Incidence**	2740	544	0	83.43	77.68%
	Progression	452	936	2	67.33	
	Non-exposed	70	2	50	40.98	
ADA	**Incidence**	2807	436	41	85.47%	78.58%
	Progression	467	921	2	66.25%	
	Non-exposed	79	2	41	33.60%	

Table 19 demonstrates the results of DNN, SVM, RF and Adaboost on the 2-class problem. Apart from being the best model in terms of overall accuracy, DNN achieved the highest sensitivity (92.54%) as well as the highest precision for the class 'progression' (84.96%) maintaining a 78.96% precision for class 'incidence'. The highest specificity was achieved by RF that although did not perform well on the progression class (having a class precision of 73.95% that was the lowest among the four models). On the

3-class problem (Table 20), DNN accomplished the highest accuracies for two of the three classes (incidence and non-exposed) having also the second highest accuracy for the progression class. SVM achieved the highest-class accuracy for 'progression' samples and it failed in recognizing the non-exposed class with only 9.83% correct assignments. Overall, DNN was proved to be the most effective model in terms of the overall accuracy and the rest of the validation metrics. Despite being the most computationally intensive, DNN was selected for the rest of the experimentation on this work given that its execution time was not prohibitive for performing multiple runs.

Effect of feature categories on the classification performance

DNN is utilized in this subsection as a criterion for evaluating the discriminating capability of different feature categories. Results are demonstrated in different DNN architectures to assess the effect of the DNN structure on the classification performance. Figure 38 shows the performance of different DNN architectures on the 2-class problem using feature subsets that correspond to symptoms occurred at different time periods before the visit. The best accuracy (79.35%) was obtained for the feature subset '*past month*' using an architecture of 3 hidden layers (with 100 nodes at each layer) applied after data resampling. The feature subsets 'past week' and 'past year' were proved to be slightly less informative achieving accuracies marginally higher than 78%.

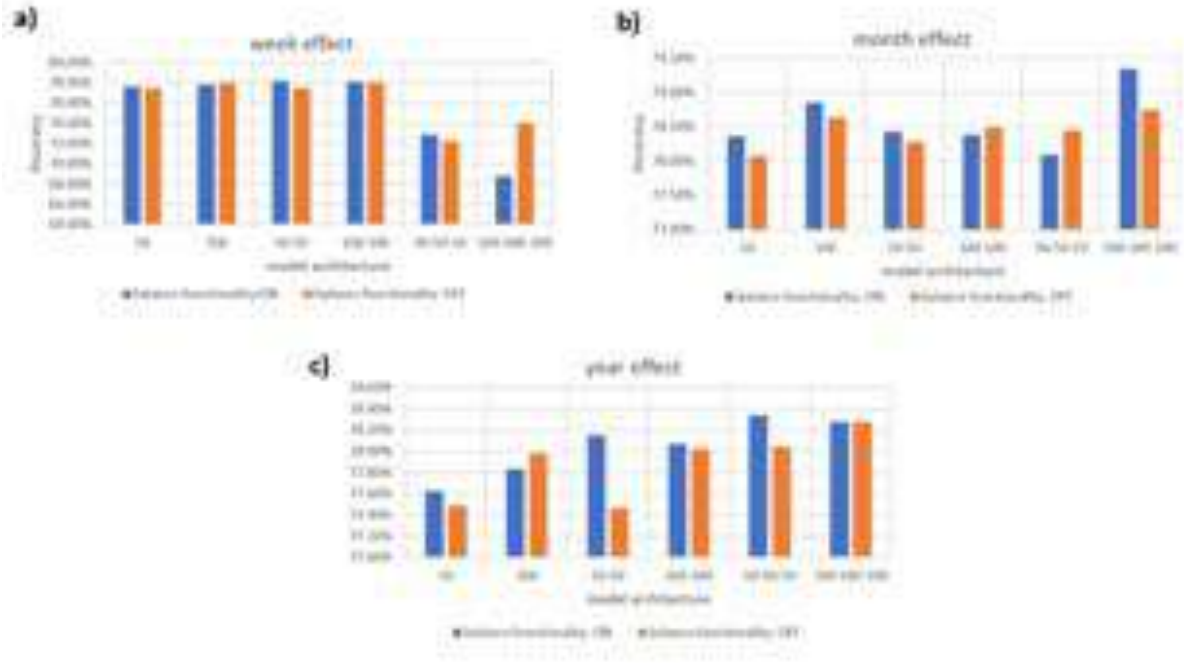


Figure 38. *Results for different DNN architectures for the 2-class classification problem using features that corresponds to symptoms occurred over the: (a) last week, (b) last month and (c) last year.*

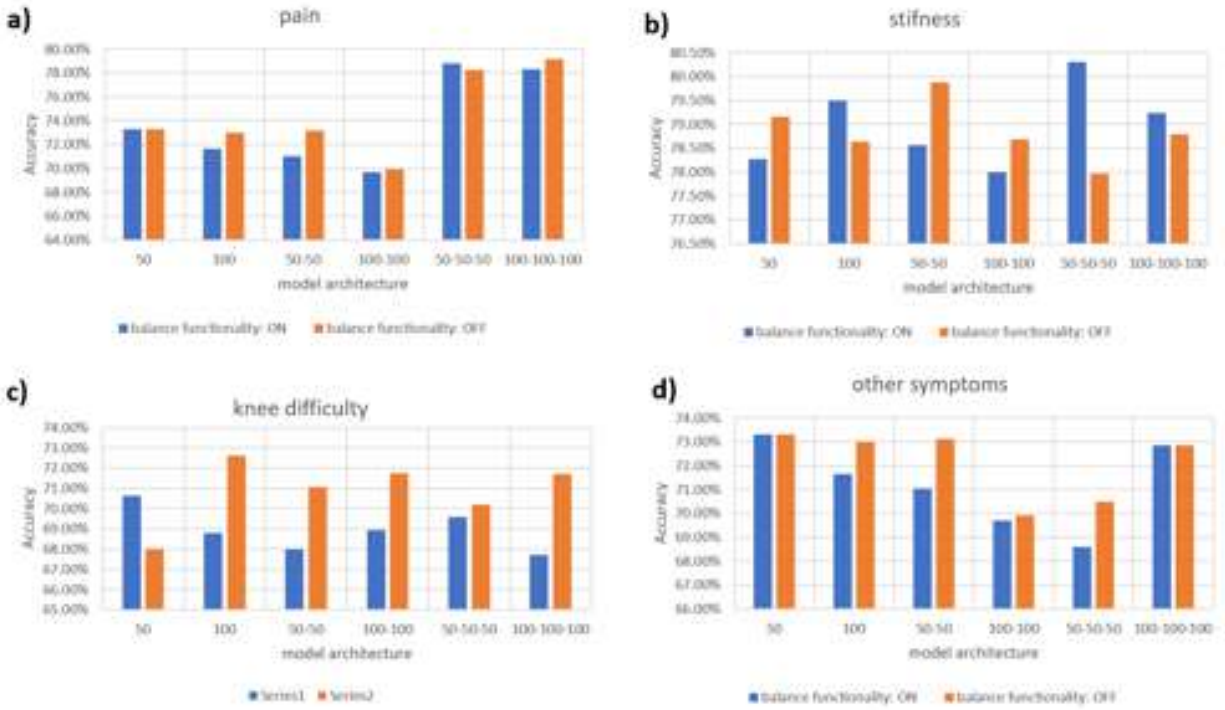


Figure 39. *Results for different DNN architectures for the 2-class classification problem using features that corresponds to symptoms related to: (a) pain, (b) stiffness, (c) knee difficulty and (d) other symptoms.*

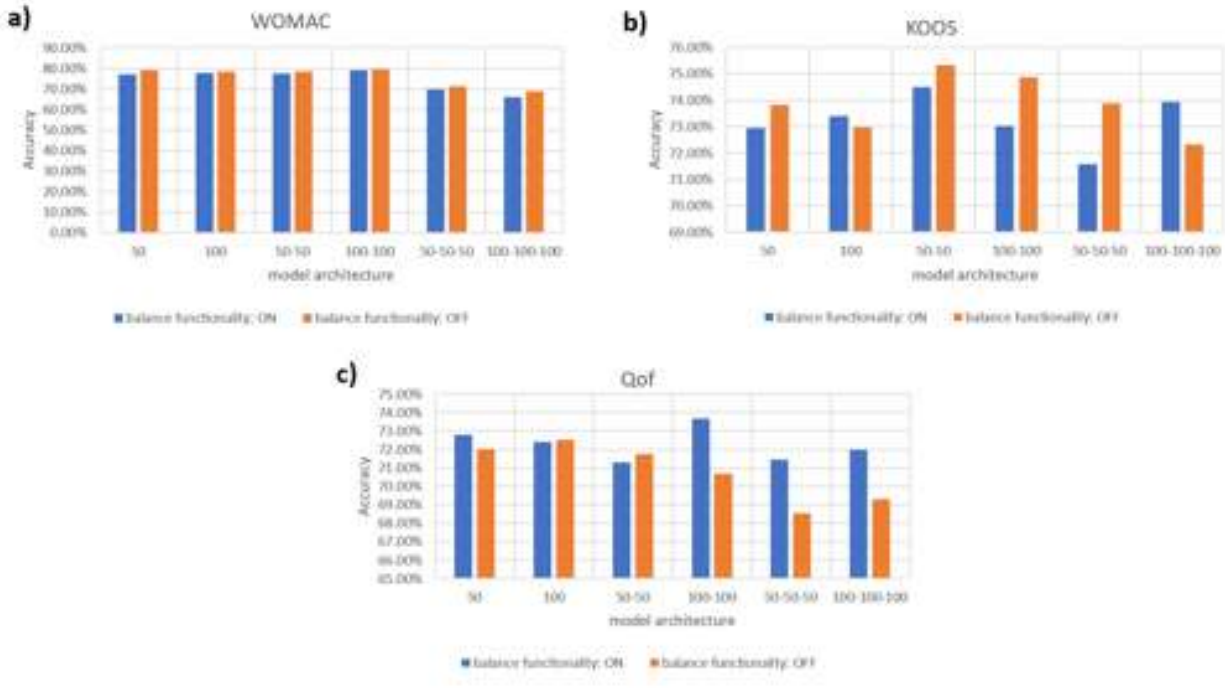


Figure 40. *Results for different DNN architectures for the 2-class classification problem using: (a)WOMAC features, (b) KOOS features and (c) features related with participants' quality of life.*

The effect of symptoms' type on the diagnosis of KOA was also investigated. Figure 39 depicts the performance of DNN using features that correspond to symptoms related to pain, stiffness, knee difficulty and other symptoms such as swelling and grinding. Stiffness was proved to be the most informative symptom with the maximum accuracy of 80.3% achieved by the best DNN using only features related to pain. It is worth to notice that this accuracy is very close to the best accuracy achieved using the entire feature set. Pain-related features were the second best that led to accuracies of 78.2%-79.2% using the deepest DNN models. The rest of symptom types achieved lower performances in the range of 73%. Figure 40 shows the performance obtained using WOMAC-based, KOOS-based features and risk factors related to health and quality of life. A 10FCV performance of approximately 80% was received using DNN models trained on WOMAC features, whereas KOOS and QoF features led to accuracies up to 75.33% and 73.68%, respectively.

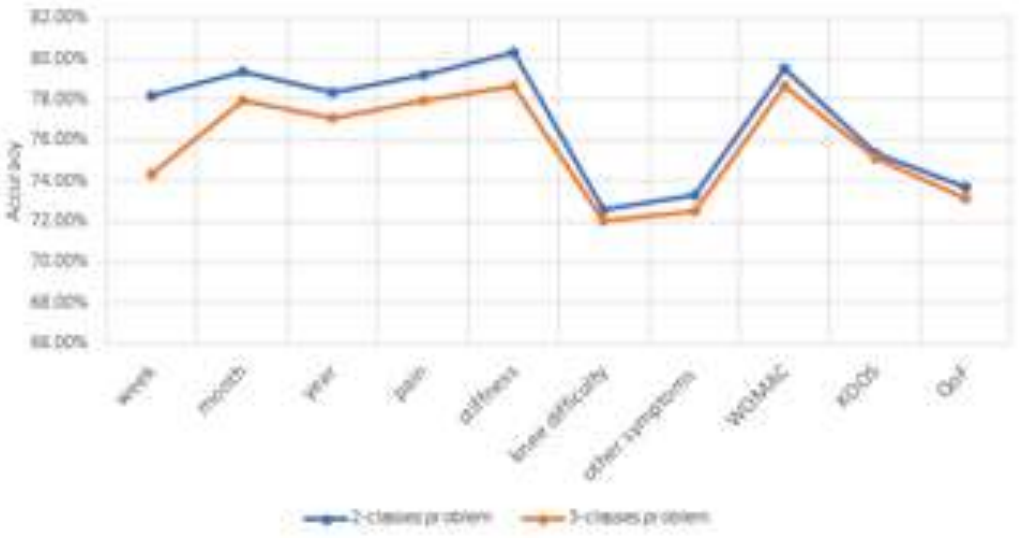


Figure 41. *Results of the best performing DNN architectures for the 2-class (blue line) and 3-class (orange line) classification problem using different feature subsets.*

Figure 41 summarizes all the classification accuracies obtained from the best performing DNN architectures trained on the 2-class problem (blue line) using the proposed 10 feature subsets. The same analysis was performed on the 3-class problem and the best accuracies per feature category are shown in same figure in orange. It was concluded that the addition of the third class led to a small decay in all the performances received over all the feature subsets. As far as the class accuracy of the non-exposed participants, the following remarks could be drawn from Figure 42: (i) WOMAC features provided an almost perfect (99.19%) identification of class 3, (ii) the feature subsets 'stiffness' and 'pain' accomplished a moderate performance, classifying correctly only 51.64% and 47.55% of the control participants, respectively and (iii) from the rest of the feature subsets, only QoF features contributed with a 12.3% per-class accuracy for class 3. The remaining feature subsets, that do not appear in Figure 32, did not contribute at all in the identification of class 3 participants.

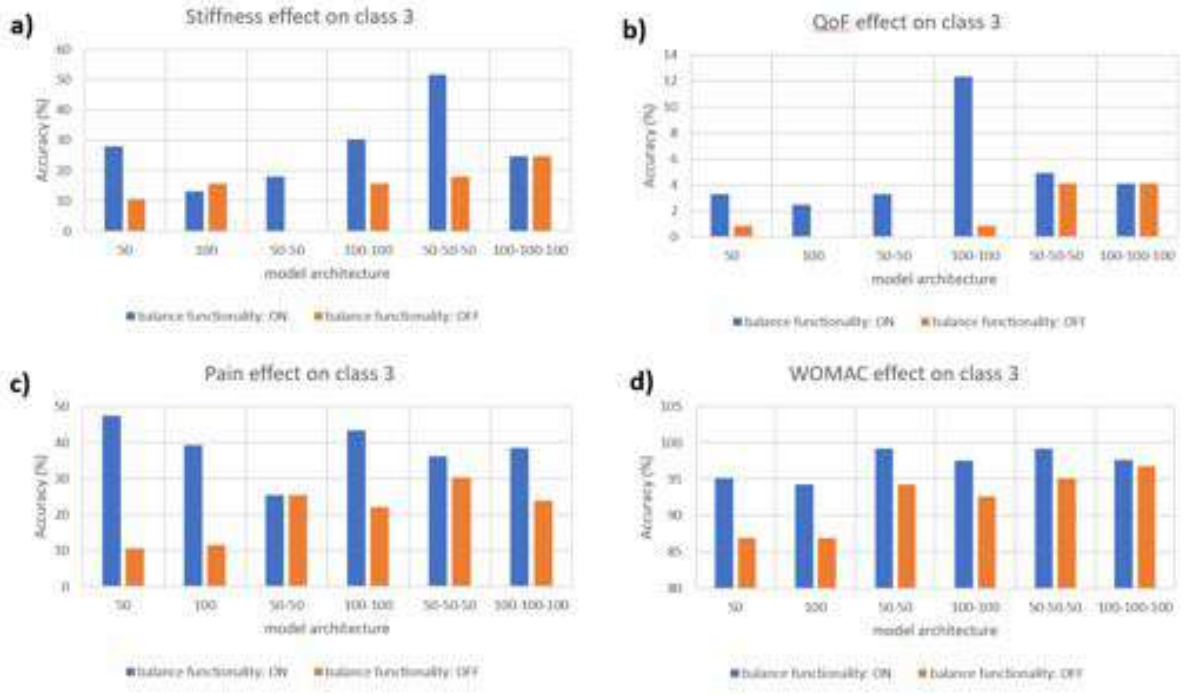


Figure 42. *Accuracy rates for the participants belonging to class 3 (control) for different DNN architectures using features related to (a) stiffness, (b) quality of life, (c) pain and (d) WOMAN features.*

KOA diagnosis with respect to gender and age

Table 21 cites classification accuracies obtained by the proposed methodology trained on data subgroups with the full feature set. The following four subgroups were considered: (i) participants older than 70 years, (iii) participants under 70 years, (iii) male participants and (iv) female participants. Significant difference was observed between the two age subgroups. Specifically, a performance of 86.95% was achieved on the KOA recognition for older participants, whereas the KOA diagnosis accuracy of the 70- age subgroup (80.81%) was closer to the overall accuracy taken on the entire dataset. Accuracies of ~81% and a negligible difference of approximately 0.5% were received for the male and female subgroups suggesting that gender is not a factor that could considerably differentiate the diagnosis capacity of the DNN models.

Table 21. *Classification accuracies on different data subgroups.*

subgroups			**problem**	**accuracy**	**Best model**	
					DNN architecture	**Data sampling**
age	70+		3-class	86.95%	3 hidden layers (100-100-100)	on
	70-		3-class	80.81%	3 hidden layers (50-50-50)	off

gender	Male		3-class	81.37%	2 hidden layers (100-100)	on
	Female		3-class	81.81%	2 hidden layers (100-100)	off

Discussion of results

An overall performance of 80.74% was achieved in the 2-class problem by the best DNN model trained on the entire feature set, whereas a small accuracy decay was observed when the third class of control participants was added. This decay can be attributed to the inability of the model to cope with the data imbalance issue where class 3 is much smaller in size than classes 1 and 2. Specifically, only half of the control participants were correctly classified indicating the difficulty to differentiate them from participants of high risk to develop KOA. The inclusion of data resampling contributed to better accuracies for class 3 participants outperforming the performance of all the DNN models trained on the original datasets (without data resampling). Finally, the proposed DNN was compared with well-known machine learning techniques and the results verified the superiority of deep learning in the KOA diagnosis task in terms of accuracy while being more computationally intensive. As far as the architecture of the selected DNN model, it was concluded that adding more layers (apart from increasing computational complexity to the training and testing phases), allowed for more easy representation of the interactions within the input data and therefore led to the highest accuracy in the case of using the full feature set. Acting as a universal approximator, DNN architectures with 2 hidden layers also gave high accuracy during the evaluation of the different feature subsets.

As far as the effect of the symptoms' type on the diagnosis of KOA, stiffness was proved to be the most informative symptom leading to an accuracy of 80.3%. Accuracies in the range of 78.2% - 79.2% were received by the deepest DNN models (with 3 hidden layers) trained on pain-related features revealing the importance of pain as a critical risk factor in KOA diagnosis. The rest of symptom types achieved lower performances at the level of 73%. WOMAC also had a significant effect on the KOA diagnosis as demonstrated by the approximately 80% accuracy of the DNN models trained only with WOMAC features. KOOS and features related with quality of life led to lower accuracies (up to 75.33% and 73.68%, respectively). Small difference in accuracy was observed between the three feature categories that were defined by the temporal occurrence of symptoms (last week, month and year). In the challenging task of discriminating control participants from those in high risk, WOMAC features provided an almost perfect (99.19%) identification of class 3, whereas the DNN models trained on the feature categories 'stiffness' and 'pain' classified correctly only 51.64% and 47.55% of the control participants, respectively. The rest of the feature subsets had a minor or no effect on the identification of class 3 participants. The application of the proposed method in subgroups revealed that it is possible to build even more accurate diagnostic models that work for specific populations. The model built on the aged subgroup (70+) accomplished an 86.95% accuracy that was the highest reported in this work. This finding implies that local models trained on more focused populations could outperform the global one. The model trained on the 70- subgroup provided an accuracy (80.81%) closer to the performance of the global model. No significant difference was received in the accuracies from male and female subgroups except to a slight increase for both in the range of approximately 2% compared to the global model.

Quantum Classification Perspective for Osteoarthritis Classification

Machine and deep learning have recently achieved impressive results in various sectors including healthcare. This can be attributed to the increased computational power and data availability, as well as algorithmic advances. However, we have almost reached the physical limits of the current solutions in terms of their speed whereas the size of the available datasets is still increasing. Given the above challenges, quantum computers may be useful for accelerating the training process of existing learning models as well as providing a way to learn more about complex patterns in physical systems that conventional computers cannot in any reasonable amount of time.

Recent findings by Havlíček, Córcoles et al. [76] set new horizons on the effective combination of machine and deep learning with quantum computing altering how computations are performed to address previously untenable problems without requiring fundamentally new algorithms. Quantum computing is expected to give AI such a boost that it would be able to discover hidden patterns within huge datasets alleviating the computational burden of the existing deep learning algorithms. Significant progress has been recently made in this area towards a better understanding of quantum computers' power for learning tasks. Quantum Neural Networks (QNN) have been proposed by Farhi and Neven [77] investigating how a popular classification task might be carried out on quantum processors. Despite being primarily theoretical, this study envisions the practical implementation of QNN in the near future. Issues related with the robust training of such networks have been also discussed by McClean et al. [78] with the aim of guiding future strategies for initializing and training QNNs.

The results of this work on the task of OA classification revealed that DL offers the best solution which unfortunately comes with an increase of the computational complexity and therefore the execution time that is required for training. However, the advent of quantum computing brings a new perspective alleviating the computational burden of all the existing learning techniques that are physically limited by the current chip fabrication approaches. The arrival of full-scale quantum computers is expected to accelerate and boost the currently employed deep learning technique, letting the proposed AI system to find unexplored hidden patterns in the multi-dimensional OA database and thus provide more robust diagnosis.

5. Interpretable models

Developing interpretable models for use in healthcare is vital. It is important that decisions are made with clarity and that the processes that go into making decisions about a person's health are easily explainable to the patient and understood by doctors. For a long time, logistic regression models have been the models of choice in medical statistics for these reasons. Every variable, or individual risk factor, have a weighting that can be used to explain that features input into the model making the model one of the easier types to use in clinical circumstances. Several of these risk prediction models make use of logistic regression [79], [80], [83], one uses the less interpretable, but arguably more accurate artificial neural network [82] and one carries out probability analysis on features and predictors that are known to be influential in OA modelling [81].

Risk prediction models, even when used solely for research, have the power to help with new insights, and show the type of models that are able to be developed and used for helping individuals reduce their risk to a given disease, or at a population level to help promote change [80]. A lot of the prediction power is to do with the data used in the study, and on the external validation datasets GOAL performed better than the OAI cohort on the same model, which may give information as to the way the dataset was collected or the information it contains. Many risk prediction models for OA have only made use of easily obtainable information, such as simple biomarkers or data from questionnaires along with demographic information. One model showed that these extra information points offer little insight into risk prediction over what simple demographics alone can provide [79]. The biggest predictor into progression has shown to be minor radiographic changes, where interventions are still able to slow progression of the disease. The use and availability of risk calculators can help to educate people at risk of developing the disease on ways that they can reduce their risk. Providing people with a calculator that provides insight into the effect that

interventions focused on risk reduction can have on their susceptibility to a disease is a powerful tool in both education and successful management of the disease [81]. Developing a model that utilises a more complex technique in a risk calculator resulted in a performance improvement compared with the more simple logistic regression [82]. Adding unnecessary extra terms in a model results in a model hat is harder to understand, however in some situations adding the extra term may reduce error and can therefore be of benefit, especially in a situation where medical interventions can be the result [83].

Data Description

Osteoarthritis Initiative

The data used in this analysis is from the Osteoarthritis Initiative (OAI) [84]. The data is available for public access at https://nda.nih.gov/oai/ . The specific dataset used in this analysis is *AllCinical00*.

The Osteoarthritis Initiative was a multi-centre study, conducted over a 10-year period in America starting in 2004. The dataset is initially made up of 4796 subjects recruited based on their likelihood to develop knee OA. A list of inclusion criteria was advertised in various places for people to refer themselves to be part of the study. Figure 43 shows the advert that was used to recruit to the Osteoarthritis Initiative.

Figure 43. *The advert used to recruit people into the OAI Observational study.*

In the study, clinical examinations, questionnaires and telephone interviews were conducted at varying intervals and the results were recorded. For the initial analysis, only the primary recordings were required, but future analysis will make use of the data from other time intervals [85].

Table 22 lists the variables in the subset used as a result of the pooled investigation that looked to identify variables of key importance to the clinical issue relating to diagnosis of knee osteoarthritis at first presentation based on a list of symptoms.

Table 22. *Variables used in the final analysis, with the definitions and the possible outcome values.*

Variable	Definition	Outcome
Age	How old is the subject on the day of investigation?	45-50 (1) 50-55 (2) 55-60 (3) 60-65 (4) 65+ (5)
BMI	What is the subjects BMI?	BMI less than 25 (0) BMI more than 25 (1)
B.LINE_SYMP	Does the subject present with knee pain today?	Yes (1) No (0)
Gender	Is the subject male or female?	Male (0) Female (1)
P01KPACT30	Has the subject limited their activities in the last 30 days due to knee pain?	Yes (1) No (0)
P01BP30	Has the subject had back pain in the last 30 days?	Yes (1) No (0)
Knee_swell	Has the subject had swelling in the knee in the past 7 days?	Yes (1) No (0)
Diff_upstr	Has the subject climbed up at least 10 flights of stairs in a single day in the past 30 days?	Yes (1) No (0)
Knee_stiff_day_limit	How many days has the subject limited their activity due to knee pain, aching or stiffness in the past 30 days?	No limit (1) Limited 1-7 days (2) Limited 7-14 days (3) Limited 14-21 days (4) Limited 21+ days (5)
Stairs.freq	How often does the subject climb at least 10 flights of stairs in a typical week, during the last 30 days?	None (0) One day per week or less (1) 2-3 days per week (2) 4-5 days per week (3) Nearly every day or every day (4)
Lift.freq	How often does the subject lift or move objects weighing at least 25 pounds by hand in a typical week, during the last 30 days?	None (0) One day per week or less (1) 2-3 days per week (2) 4-5 days per week (3), Nearly every day or every day (4)

Data Subsets

To develop the dataset that will be used, auxiliary investigations were first required. The initial breakdown of this process is shown in figure 44 and figure 45.

Clinical and Demographic data

Removing any subjects that have missing values in variables resulted in a useable data set of 4,226 subject. Missing values are defined as non-imputed values, except in cases where more than one primary column gives rise to the new variable used in the analysis. One such example of this is the General Arthritis variable in the analysis which was made up of a series of questions relating to different types of arthritis. In this situation if any single option was present, in either a yes or no capacity then that answer was recorded for

general arthritis with the only exception being if all were missing then the new variable would be recorded as missing, and that subject removed from the analysis. These criteria resulted in the removal of a further n = 281 participants, leaving a sample of n = 4,226 as the updated maximum for any analysis of complete case subjects.

Seventeen variables were identified using the extracted OAI data. The variables include the age, gender and BMI of the individual, along with information of family history, previous injuries and diagnoses of osteoarthritis in other joints and general arthritis in the body. Within the analysis of the clinical data, the training sample contained 2,145 subjects and the test sample comprised 2,081 subjects.

Self-Reported data

The self-reported data is made up from subject's answers to questionnaires relating to their symptoms and how they are impacted, recorded at the first presentation meeting. Similar to the clinical data, many of the variables needed to compress in order to be suitable. One such example is left and right sides of the body. For the purposes of this analysis, it was not important if the affected joint was on the left or the right, so to combat the duplication of cases the same approach was taken as for the clinical data variables. If a person presented in one side of the body or the other then that would be the value taken forward, if they presented in both sides, the most severe measurement would be taken for categorical variables, and the average taken for numerical variables, if they were problem free for that variable on both sides of the body, that response was taken, and if missing values were present in both cases that was also recorded. The complete case analysis had a total sample size of n = 3711, giving a training set of n = 1868 and a test set of n = 1843.

Self-Reported Physical Activity Data

In a similar approach to the Self-Reported dataset, the Self-Reported Physical Activity data is made up of answers on questions about how much they take exercise and how this impacts them. This set of data on its own appears to be the most modifiable in terms of lifestyle changes that a person can make.

Like the clinical and self-reported data, only cases that had a definite KL grade and no missing values were included. In the complete case any row with missing data was to be removed. The complete case analysis had a total sample size of n = 3309, giving a training set of n = 1691 and a test set of n = 1618.

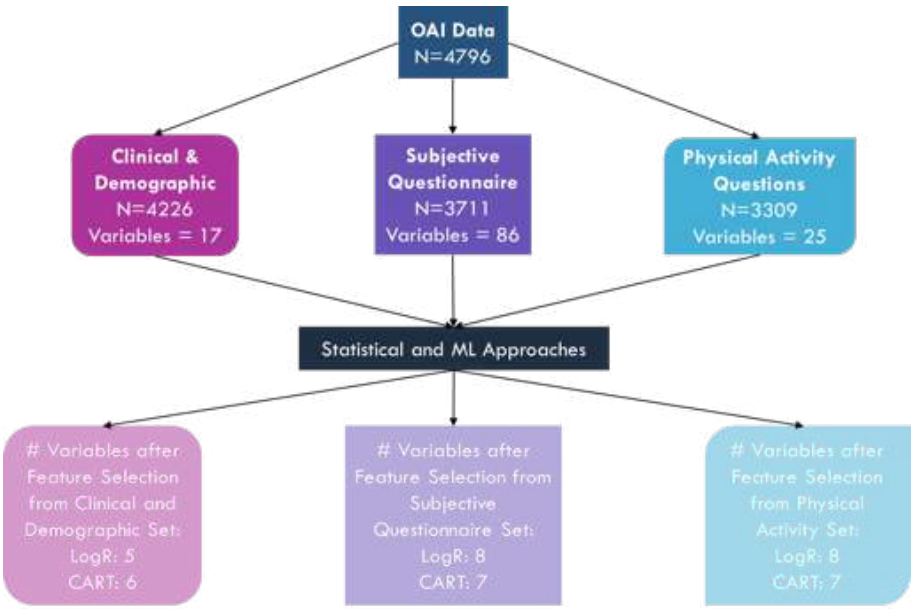


Figure 44. *Visual representation of how the data is split, in terms of cohort size and variables after feature selection.*

Pooled Data Model

This approach looked at selecting the important features from the whole dataset available. To perform this analysis, all of the variable subsets were pooled together, and the models run on this data. The resultant feature set provided the subset of variables used in the main analysis. Figure 35 illustrates the process used.

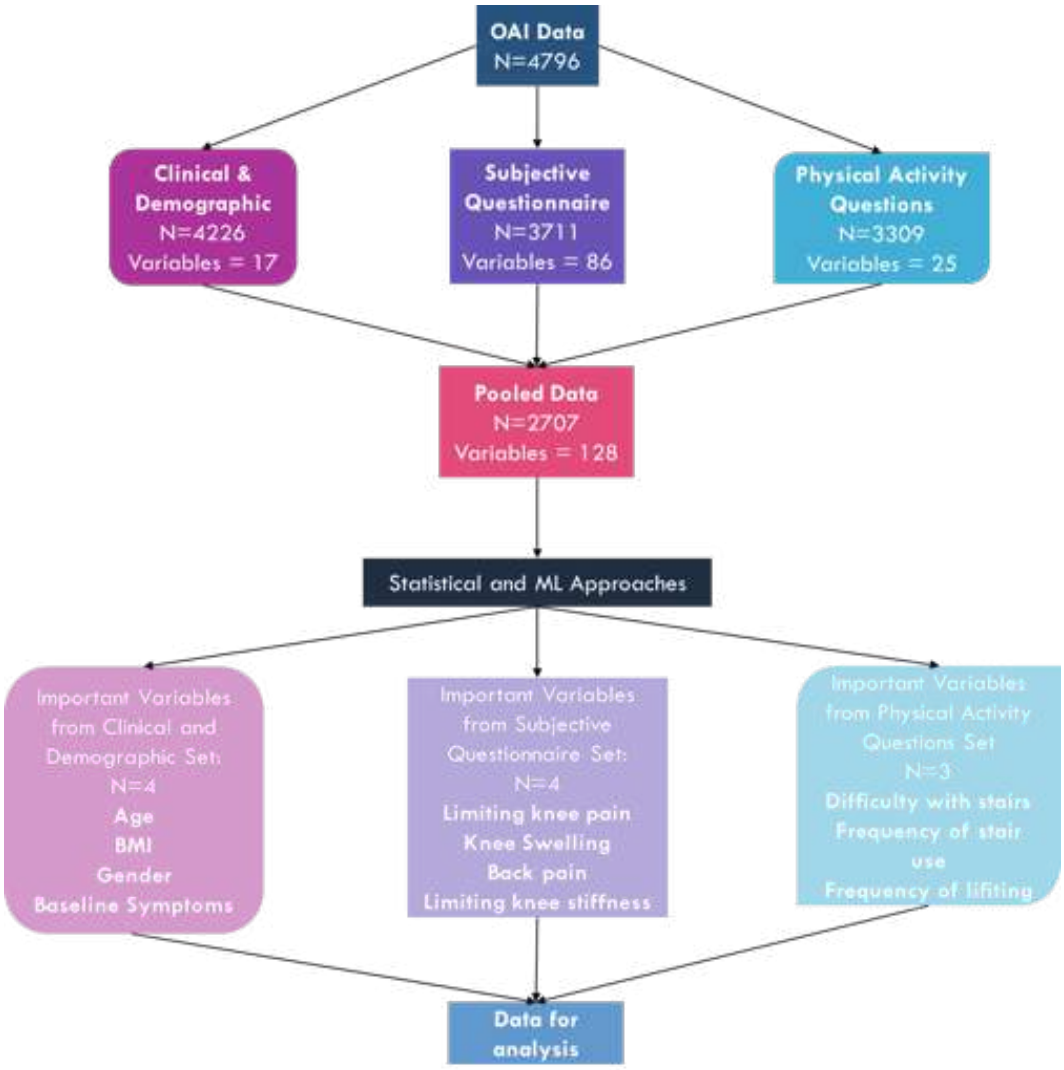


Figure 45. *A visual representation of how the data required for the analysis was selected, and how it is made up of the data cohorts.*

Sample Selection Criteria

The cohort considered in this analysis was only subjects without any missing values for the selected variable set. This is a complete case analysis [86]. In the preliminary steps of the analysis, not detailed in this report, a complete case and imputed analysis were used and compared. The cohort, including all of the missing values had a sample size of 4796. After reducing the sample by removing those who have no Kellgren-Lawrence grade leaves a sample of 4507 subject. Finally, removing those subjects who have missing values in any portion of the variable sets leaves a usable cohort of 2707 subjects in the complete case analysis.

In order to create one model that incorporates the clinical, self-reported and Self-Reported Physical Activity features these variable sets need to be consolidated. The final variable set from each data subset was derived

from analysing the results from the two analyses and selecting variables that have been selected by logistic regression and CART.

The approach taken in the analysis is to combine the datasets with the Subject ID and utilise the variable subset dataset. This dataset would then be used in the machine learning models. The resultant analysis incorporates the clinical, demographic, self-reported questionnaires and the physical activity data. All subjects included in this analysis have a determined Kellgren-Lawrence grade attributed to them as the outcome variable. The variables in this subset are listed and defined in Table 22.

Methods

The methods used in the analysis performed here are machine-learning methods. Machine learning is an approach that can provide the ability to automatically learn and progress without being programmed explicitly. The type of machine learning used in these analyses are supervised machine learning. This is where there are previously labelled data that can train a model to make predictions given a set of predefined variables.

Throughout this analysis, four main approaches are used.

Classification and Regression Trees

CART is a rule induction approach that determines univariate cut points. This machine learning approach can be classification or regression-based. In this decision-making, the classification approach is the most suitable option for the data. In clinical situations, this can be used to develop a set of questions that can aid clinicians in a decision-making process before invasive investigative tests are undertaken.

When trying to choose a machine learning approach to use a number of things are taken into consideration at each step. Many of the decisions are made regarding how easy the models are to use and understand. CART analysis has the advantage of being very interpretable and easy to understand. This is, in part, due to being able to represent the results in different form, such as graphically or with the tree diagram itself. Another advantage to this method is that it is highly interpretable. Conditions to class membership are clearly explained meaning that the explanation about how the decisions are made are easily demonstrated, removing the 'black box' nature of machine learning. Another reason this approach is a favourite is due to the way the decisions made closely mirror those made by humans.

Logistic Regression

This is the most commonly used statistical model in medical decision support. Although it is linear-in-the-parameters, careful discretisation of continuous variables creates a piecewise linear model with the capability to model highly non-linear data, which are typical in clinical medicine. As a result, logistic regression models are often very competitive in discrimination accuracy compared with neural networks and other machine learning methods, except when interactions between variables have a significant role in decision-making, in which case rule induction may be preferred.

Logistic regression is a preferred method as it can also be translated into nomograms for easy clinical use and interpretation. The use of a nomogram can turn otherwise complex mathematical models into easy-to-understand graphics that can show the real implications of changing behaviours to those seeking advice. For example, nomograms could be of particular use in educating a subject seeking medical advice how best to change their lifestyle in order to prevent developing OA or slow down their risk of progression.

$$\frac{\mathrm{P(class|X)}}{1-\mathrm{P(class|X)}} = \prod_{i=0}^{n} e^{\beta_i x_i} = e^{\beta_1 x_1} . e^{\beta_2 x_2} \ldots . e^{\beta_n x_n} . e^{\beta_0 x_0}$$

Equation 0-1: Logistic regression odds ratio formula

Equation 0-1 illustrates an expression where for binary covariates $\{x_i\}$ the exponentials show explicitly the size of the effect of the variable on the odds-ratio.

Lasso

Lasso (least absolute shrinkage and selection operator) is a shrinkage method used in statistics and machine learning to perform both variable selection and regularisation to aid in prediction accuracy and model interpretability. The shrinkage relates to the ability to discard variables that are not as useful in the model. This approach is preferred over subset selection as they are more continuous and therefore have lower variability.

Lasso uses L_1 penalisation, as in Equation 0-2. This means that by adding a penalty equal to the absolute sum of the coefficients the method will shrink some parameters to zero, so some variables will not play any role in the model. Using Lasso in this way is one approach to select features in a model. The penalty performs a continuous variable selection process in the model.

$$\hat{\beta}_{lasso} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Equation 0-2: L1 penalisation term

Multilayer Perceptron Automatic Relevance Determination (MLP-ARD)

A multilayer perceptron (MLP) is a type of artificial feed-forward neural network [87]. The MLP is made up of at least three layers: an input layer, a hidden layer and an output layer. The output layer in this case is a binary classifier.

For the MLP-ARD configuration, a standard MLP is used in the first instance [88]. The ARD is useful when it is important to know what variables are contributing the most to the classification. The ARD is to determine the most relevant features in the data.

For the ARD approach to work, a separate hyper parameter α_i is assigned to each group of weights spanning from the i^{th} input variable. The hyperparameter is re-estimated through each iteration of the tuning process. At the end of the training stage and hyperparameters with large values indicate that an input has little impact on the final model. This highlights what features can be dropped from the final model.

Partial Response Network

The partial response network, PRN, is a method to open the black box approach of the MLP [89]. The end product results in non-linear univariate and bivariate partial responses from the MLP. When the performance of the PRN is compared with a fully connected MLP, there is usually performance improvements because of the PRN implementation. The bivariate responses come from modelling pairwise interactions in the network. Interactions are modelled up to pairwise, and all others are categorised under the residual modelled in the network. The PRN implementation mimics models of deep learning but offers the advantage of being highly interpretable.

The way in which the PRN works can be explained in six steps, shown in Figure 46.

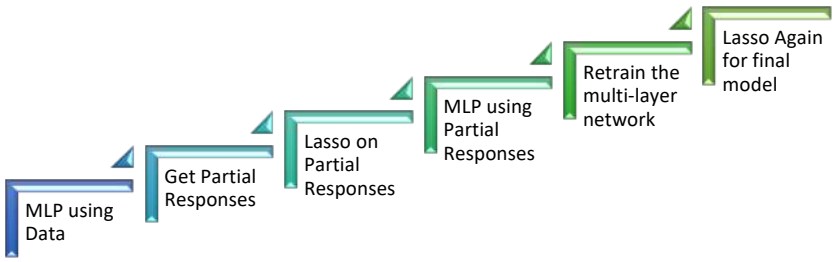


Figure 46. *The six steps used to develop the partial response network.*

Preliminary Results

Results and initial discussions relating to findings are presented for the four models used in the analysis.

Table 23. *A table of performance metrics for the different models used in the analysis, giving the area under the curve (AUC), sensitivity, specificity and positive predictive value (PPV).*

	AUC	**Sensitivity**	**Specificity**	**PPV**
CART	0.719	0.600	0.776	0.641
LogR	0.763	0.674	0.716	0.613
MLP-ARD	0.778	0.677	0.676	0.576
PRN-Lasso	0.793	0.698	0.697	0.599

The receiver operating characteristic curve (ROC curve) is a plot that graphically indicates the ability of a model to correctly classify binary outcomes as a threshold is altered. The area under the curve (AUC) is equal to the probability that a classifier will rank a random positive instance higher than a randomly chosen negative one [90]. In the AUC a value of 0.5 indicates a guess, with greater than this being deemed better than a guess, and lower than o.5 being worse than a guess.

Sensitivity (Equation 0-3), specificity (Equation 0-4) and positive predictive value (PPV) (Equation 0-5) are all statistical measures of the performance of binary classification tests. The sensitivity measures the proportion of actual positives that are correctly identified. The specificity measures the proportion of actual negatives correctly identified.

Equation 0-3: Sensitivity formula

$$Sensitivity = \frac{TP}{TP + FN}$$

Equation 0-4: Specificity formula

$$Specificity = \frac{TN}{TN + FP}$$

Equation 0-5: Positive predictive value formula

$$Positive\ Predictive\ Value = \frac{TP}{TP + FP}$$

Where:

- TP is true positive result,
- TN is true negative result,

- FP is a false positive result, and
- FN is a false negative result.

CART Results

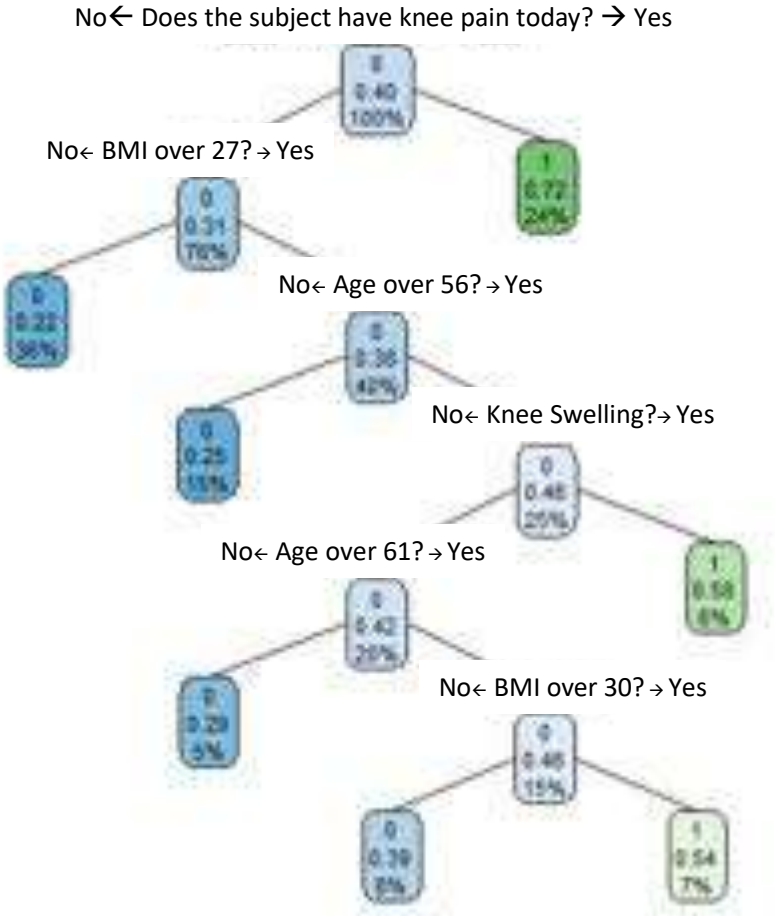


Figure 47. *Tree diagram showing the stages in the process to determining based on a set of questions if a subject has KOA.*

The tree diagram in Figure 47 shows the splitting criteria in a highly interpretable way that could be transformed into a question set for clinicians to use as a signposting tool. In the diagram, the middle number is the prevalence of people in that group that have knee osteoarthritis by following the conditions to arrive at that node. The bottom number is the percentage of the population that is covered by the node criteria.

From Table 23 CART analysis seems to be the worst performing model, but only slightly. However, the drop in performance can be traded off for simplicity and ease of interpretation that the model provides. The performance is important, but where decisions impact people it is imperative that results can be explained, and the CART model offers a high level of interpretability.

Logistic Regression Results

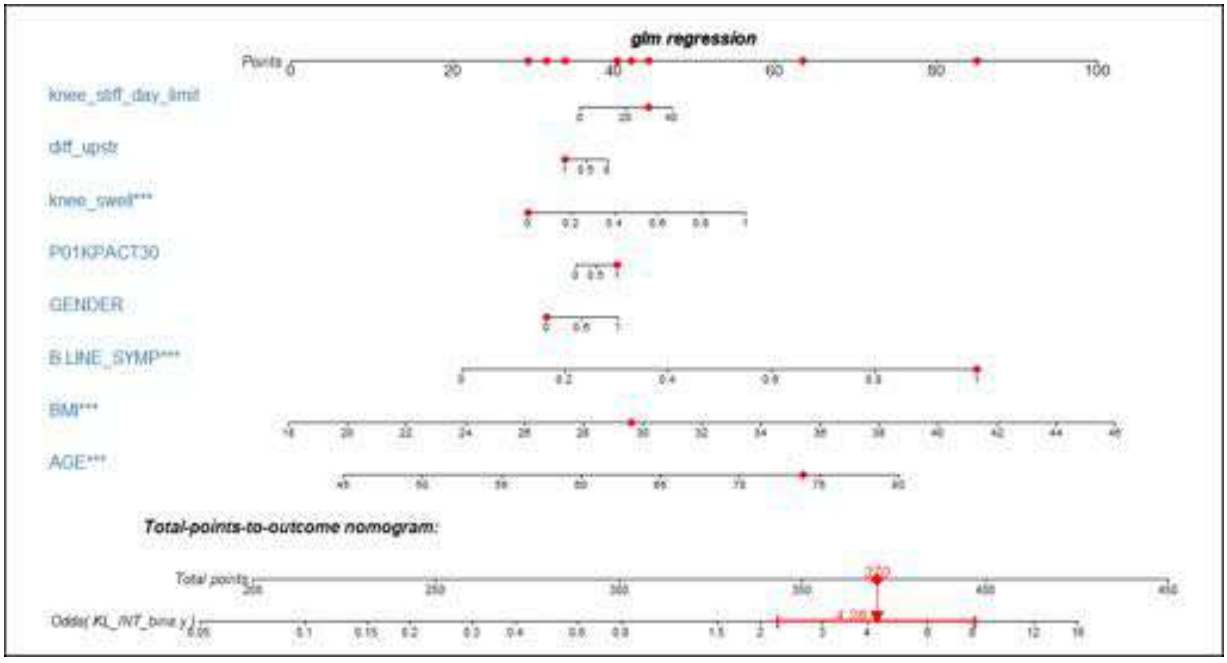


Figure 48. *A nomogram produced from the LogR model indicating the likelihood of having osteoarthritis based on a set of features at different values.*

The nomogram depicted in Figure 48 related to the performance of the LogR model. The LogR model is a baseline indicator as it is used in clinical practise, as this is the preferred method for binary classification for a variety of healthcare problems. The LogR model produces an interpretable nomogram that gives every value a point score that relates to the chance of having the disease in question, in this instance the disease is KOA. The nomogram also indicates a possible confidence interval where the symptom scores could fall, giving another reason that this type of approach is preferred in the medical arena. The results from the LogR analysis indicate that the data is well suited to this type of modelling approach.

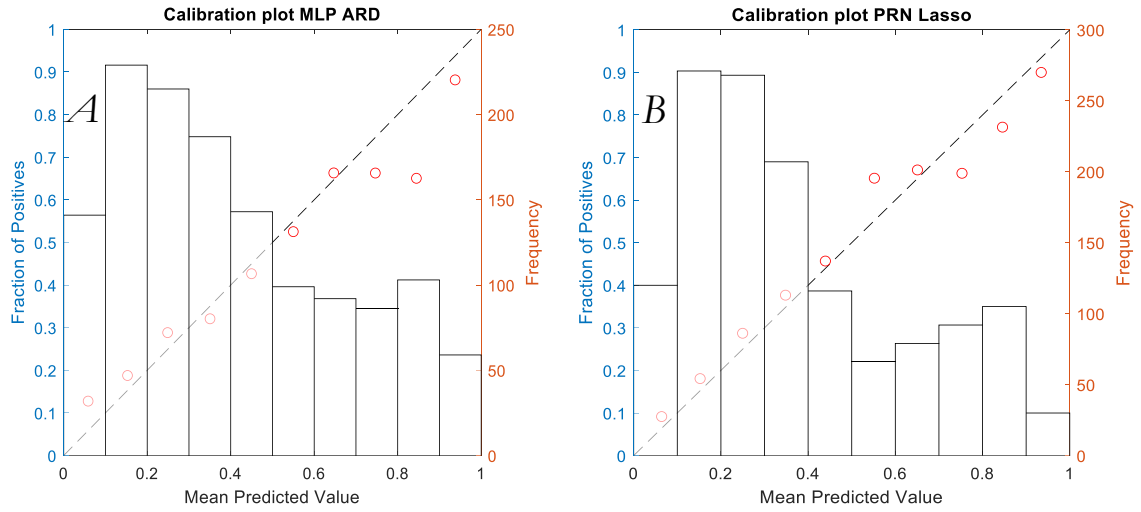


Figure 49. *Calibration plots for the MLP-ARD (A) and the PRN-Lasso (B), showing how well the models are calibrated to the data.*

MLP-ARD and PRN-Lasso Results

The calibration plots, shown in Figure 49, show how the model fit is near the pattern present in the data. This shows that at each step of the model, the model is adequately trained to perform predictions within the applicability domain. The models are well calibrated as the points are all close to the diagonal line. The PRN-Lasso is the best-calibrated model.

Features after PRN-Lasso

After the initial MLP-ARD, the lasso model selects the most important features in the data. For this dataset, the main features are five univariate and six bivariate effects. The features still important after the PRN-lasso are the ones that are in the final model. In the final model, there are four univariate effects. These are: age, BMI, baseline symptoms (presenting with pain) and knee swelling. The effects can be shown in Figure 50.

As age increases, the more the effect age has on the logit, so the contribution to the outcome, the presence of KOA, increases. The contribution to the logit is nearly linear until about the age of 70 where a similar pattern can be shown with the BMI and its effects to the presence of KOA. As the BMI increases, the contribution to the logit also increases in a nearly linear pattern. The presence of pain symptoms at the investigation will increase the contribution to the logit. The subject experiencing knee swelling will increase the contribution to the logit as this symptom would indicate the presence of knee osteoarthritis. Both of these statements are in line with what is presented in the literature.

The ROC curves show how well the models predicted the binary categories. The models vary in simplicity, with the most complex model being the MLP-ARD and the most parsimonious model is in fact the PRN-Lasso, which only uses four of the eleven input variables with the structure of a Generalise Additive Model, which is self-explanatory.

A further brief analysis was carried out about the relationship between the standard radiological indicator of OA (KL score) and reference measures of pain. It is sometimes commented that the KL score does not appear to correlate at all with the experience of pain.

In order to elucidate this relationship, we analysed the observed and inferred KL score against both KOOS and WOMAC. The results are similar so we report them for WOMAC. The results shown in Figure 52 indicate that while observed KL and WOMAC seem entirely uncorrelated, this may be due to subjective effects that add noise to the estimated likelihood of clinical KL, which is the inference made by the statistical model from the overall population, as the log-odds-ratio (logit) shows a clear correlation with WOMAC.

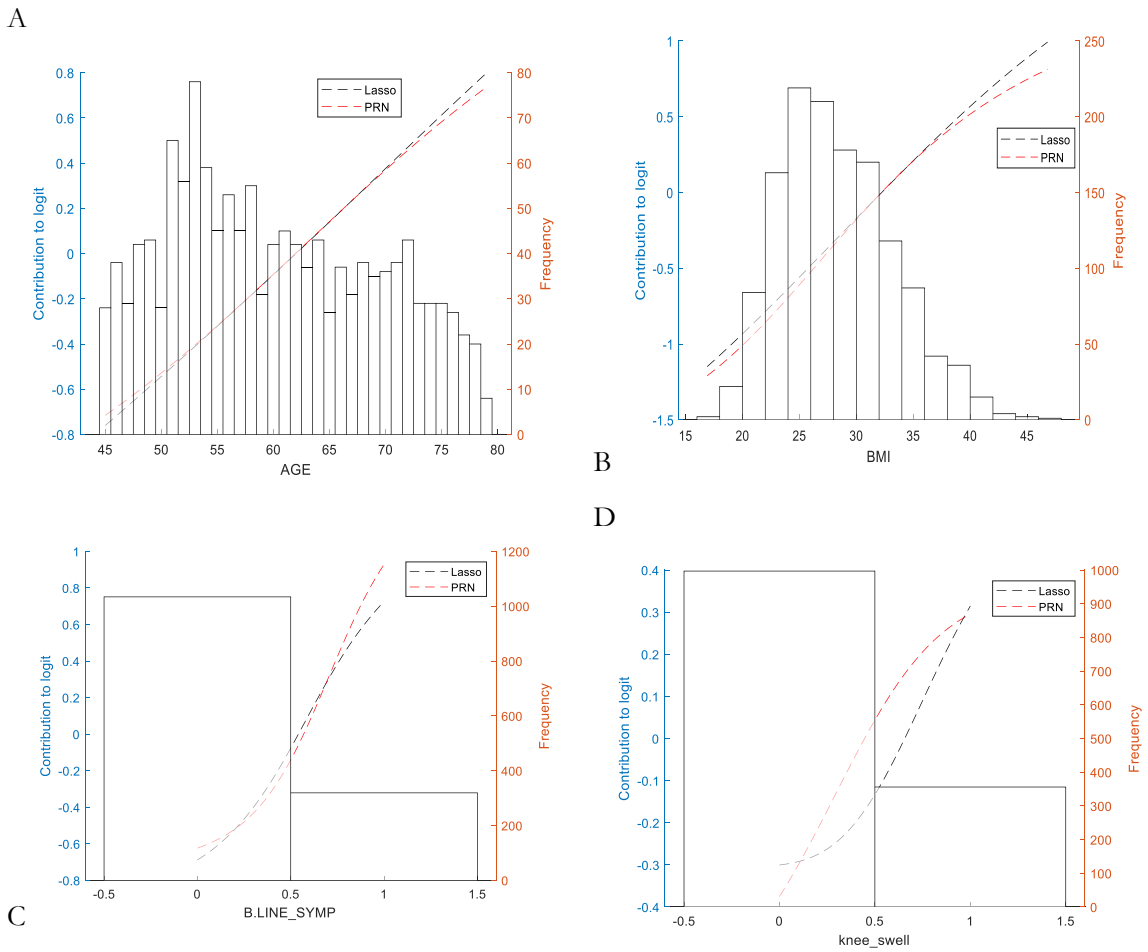


Figure 50. *Partial response graphs for the variables in the final model as generated by the PRN. Graph A depicts age, B is BMI, C is presence of pain at initial investigation and D is knee swelling.*

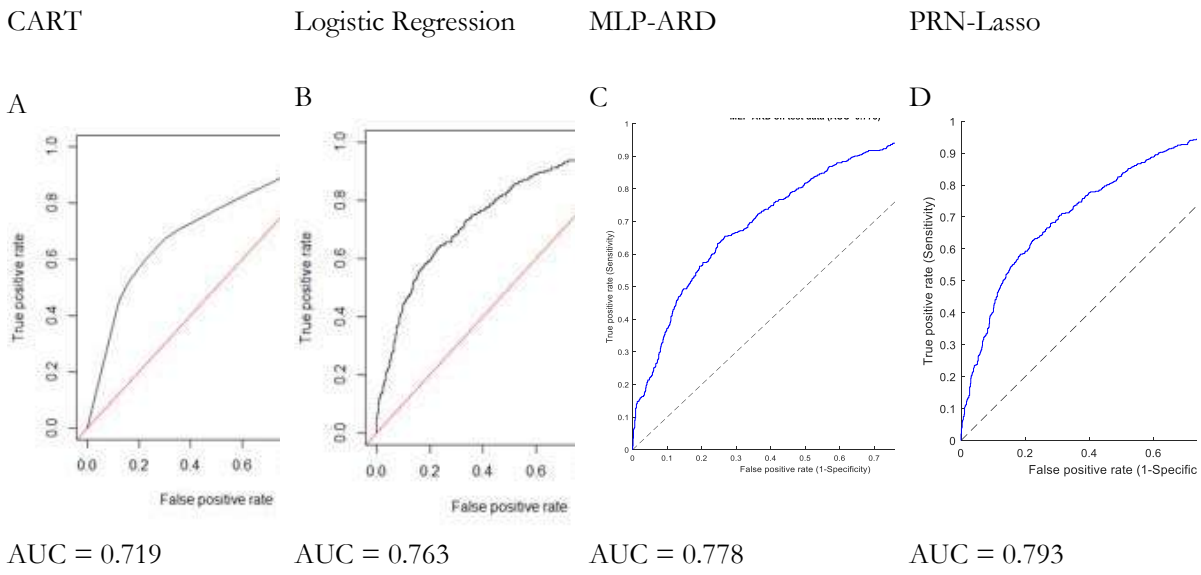
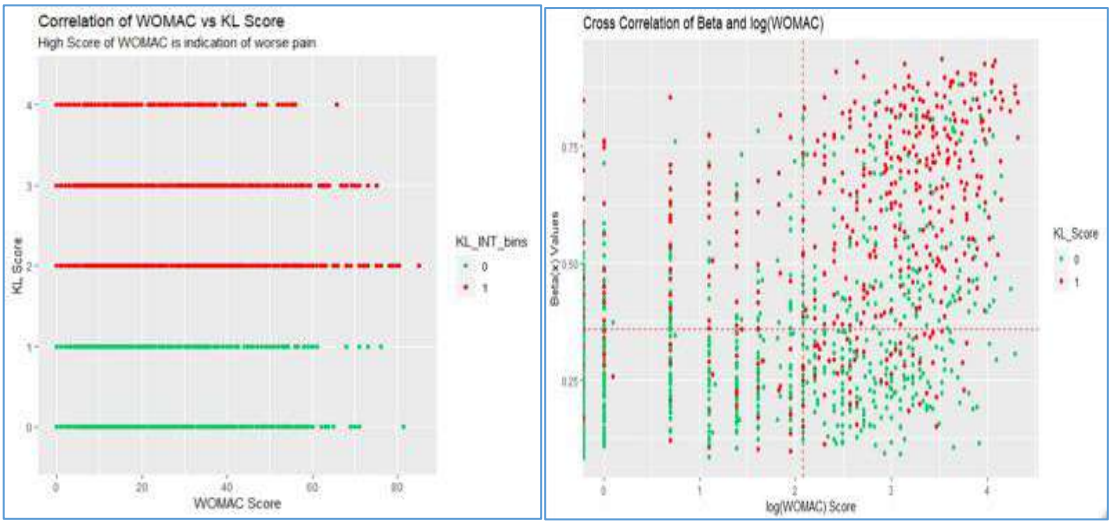


Figure 51. *ROC curves for the different approaches used in the analysis. A is the CART approach, B is LogR, C is the MLP-ARD and D is the PRN-Lasso.*



(a) (b)

Figure 52. *Relationship between radiological OA measured by the KL score and perception of pain measured by the WOMAC score. a) Scatterplot of observed KL vs. WOMAC. b) Scatterplot of the logit of inferred KL derived from logistic regression (this is the sum of beta*x scores) against the logarithm of self-reported WOMAC.*

The diagnostic risk model was validated externally using data from the Multicenter Osteoarthritis Study (MOST) (n=831 after pre-processing). This verified the accuracy of the model with clinical variables, although some of the variables needed to be mapped to those used by the MOST data set

Table 24. *Mapping of variables from OAI to the MOST data set.*

Variable	Definition
knee_stiff_day_limit	In the last 30 days, how many days has the subject limited activity due to knee pain/aching/stiffness? 0 days/ 1-7 days/ 7-14 days/ 14-21 days/ 21 or more days
Diff_upstr	Does the subject have difficulty getting upstairs? Yes/ No
Knee_swell	In the past 7 days, has the subject had knee swelling? Yes/ No
P01KPACT30	In the past 30 days, has the subject limited their activity due to knee pain? Yes/ No
Gender	What is the subject's gender?Male/ Female
B.LINE_SYMP	Does the subject present with knee pain today? Yes/ No
BMI	What is the subject's BMI? Less than 25/ 25 or more
AGE	How old is the subject at the assessment? 45-50/ 50-55/ 55-60/ 60-65/ 65+

Table 25. *Summary statistics of external validation of the diagnostic model for the MOST data set. The cut-point is the prevalence of KL2+ in the OAI data.*

Measure	Value
Accuracy	0.6859
Sensitivity	0.9052
Specificity	0.2353
PPV	0.5421
NPV	0.7128
AUROC [CI]	0.6697 [0.631, 0.708]

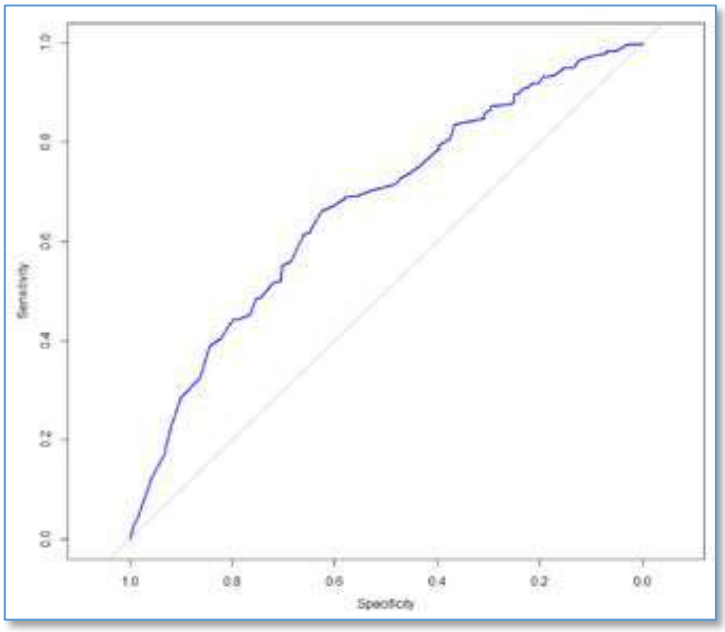


Figure 53. *ROC curve for validation of the diagnostic model with data from the MOST study.*

Prognostic modelling

Prognostic modelling of the OActive is beyond the timescale of the project. However, insights from such models developed using the OAI data are relevant to deeper understanding of the phenomenology of KOA.

Proportional hazards regression was applied to the OAI data, specifically the recruitment cohort comprising subjects with KL 0 or 1 at first presentation, followed-up for the occurrence of KL 2 or above, which is the event of interest. Our study was framed with a timescale of 5 years following which all surviving cases are censored. The term survival in this study means the period of time when the subject is disease free. Stepwise model selection was again applied, as in the case of logistic regression, with appropriate penalisation for model complexity, and the risk index i.e., the time-independent linear covariate effects in

the Cox model, was stratified with the log-rank test, resulting in two statistically significant populations, one with high and the other with moderate survival, which we term low and high risk, respectively.

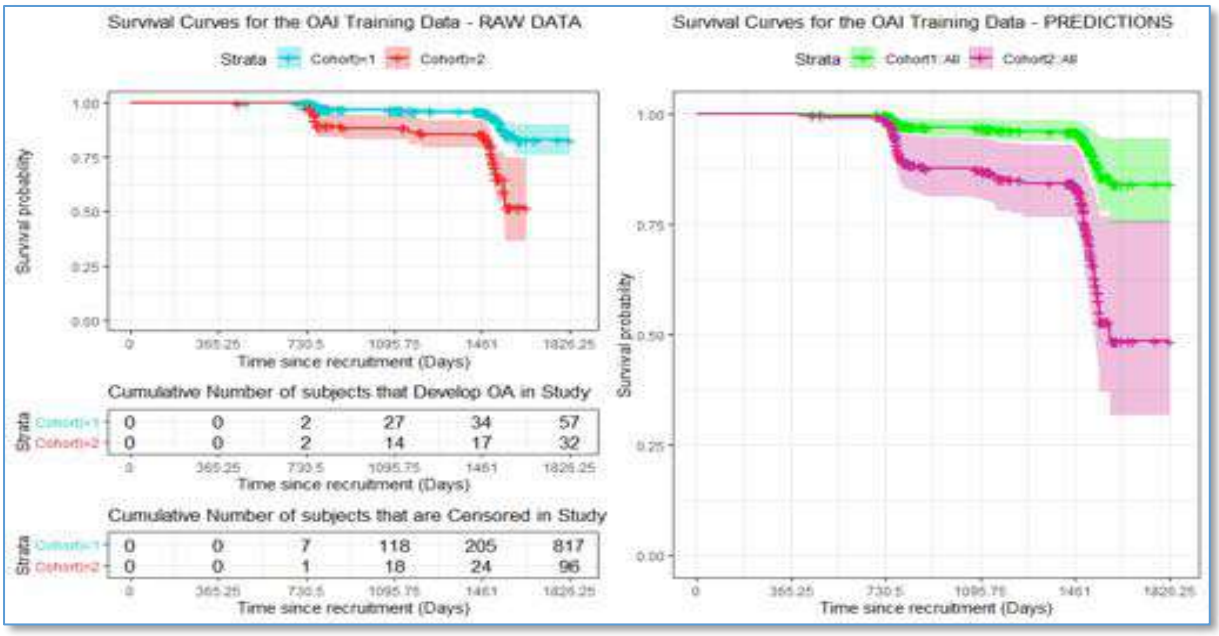


Figure 54. *Observed and predicted time-to-event curves for the OAI data, modelling time to KL2+ from KL) or 1 at first presentation.*

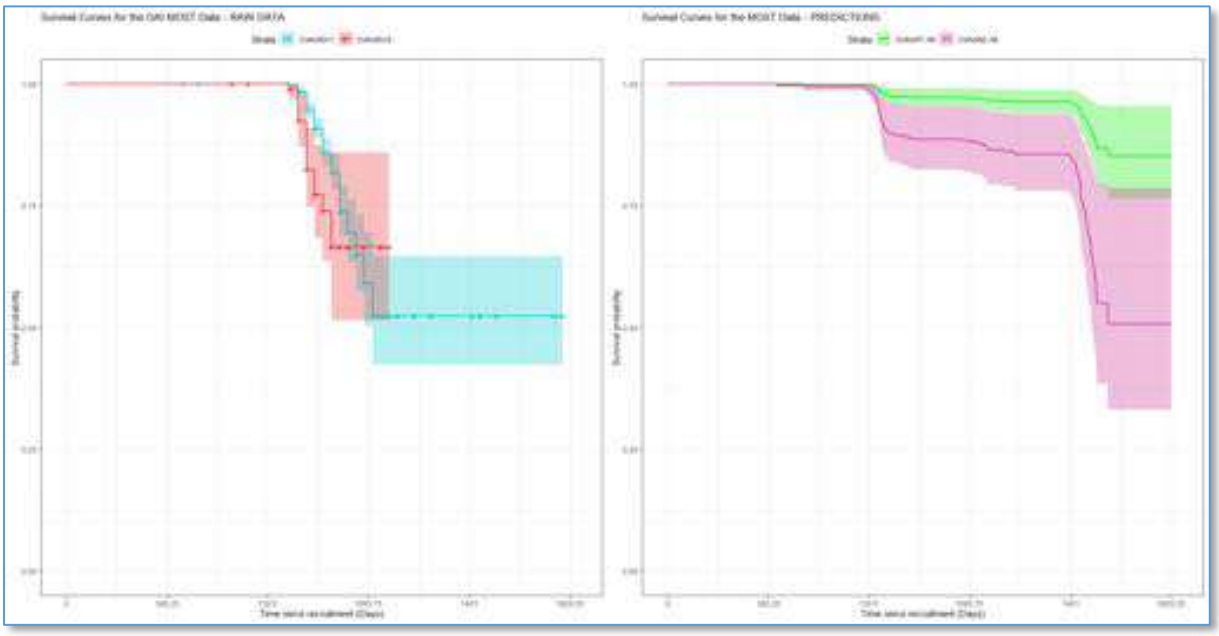


Figure 55. *Corresponding figs to the previous figure, for external validation of the prognostic modelling by application to the MOST data set.*

Nomograms

A key part of risk model development is transparency to the end-user. To this purpose linear models were applied, not just because they are interpretably but also because their performance is statistically comparable with those of machine learning models for the OAI data set. This is most likely the result of the relatively

high levels of noise in this observational clinical data set. Indeed, the linearity of dependence on covariates was explicitly verified by modelling with Partial Response Network.

Visualisation of the diagnostic and prognostic models was made available through the app interfaces with nomograms in figs 56-57.

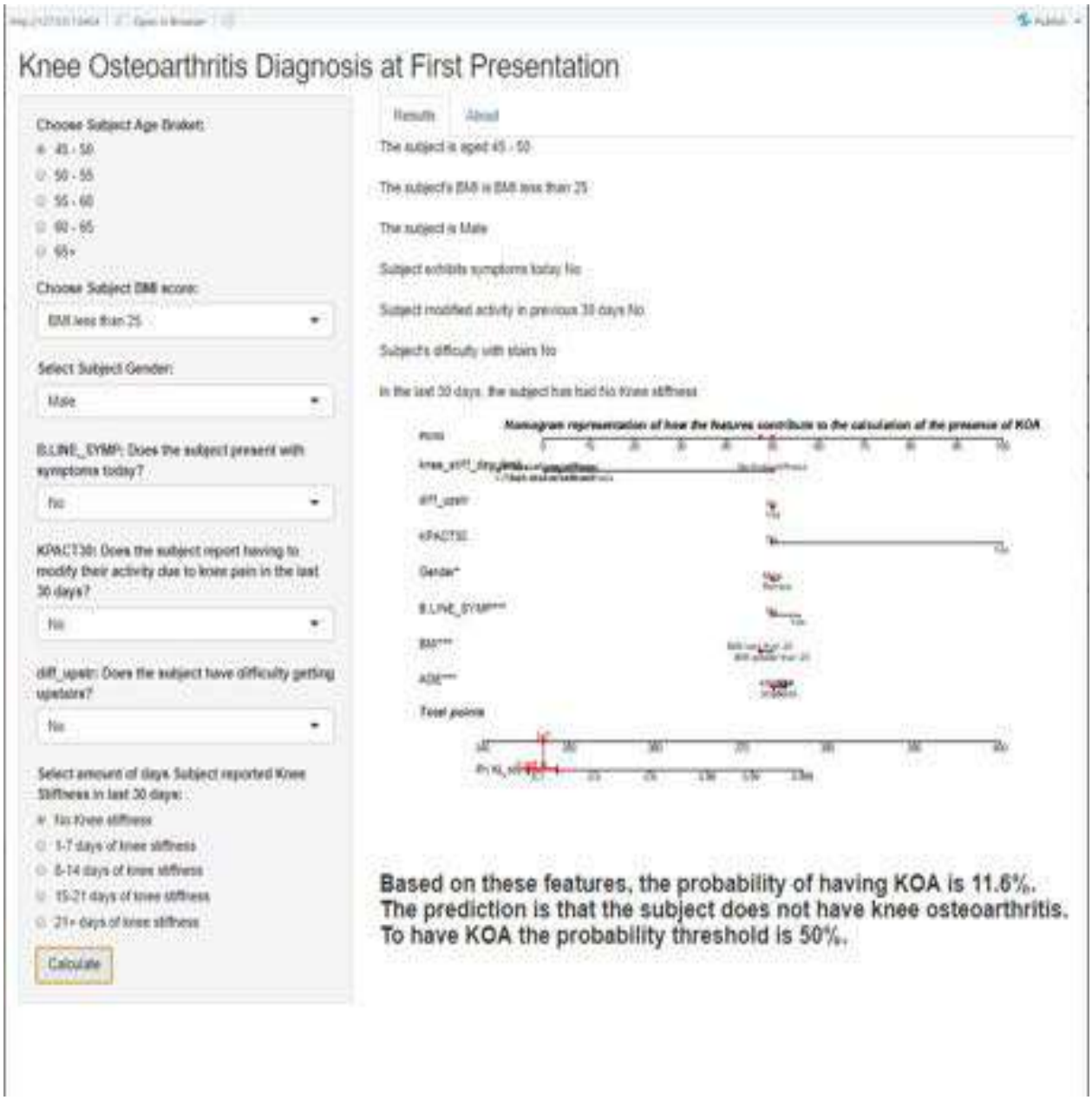


Figure 56. *Snapshot of the app with nomogram feature for the diagnostic model. The red points represent the component scores for the example subject, whose details are entered in the column on the left of the figure. The total score is also show, mapping to a predicted probability at the bottom of the figure, where the cut-point represents the prevalence of KL2+ in the OAI data set.*

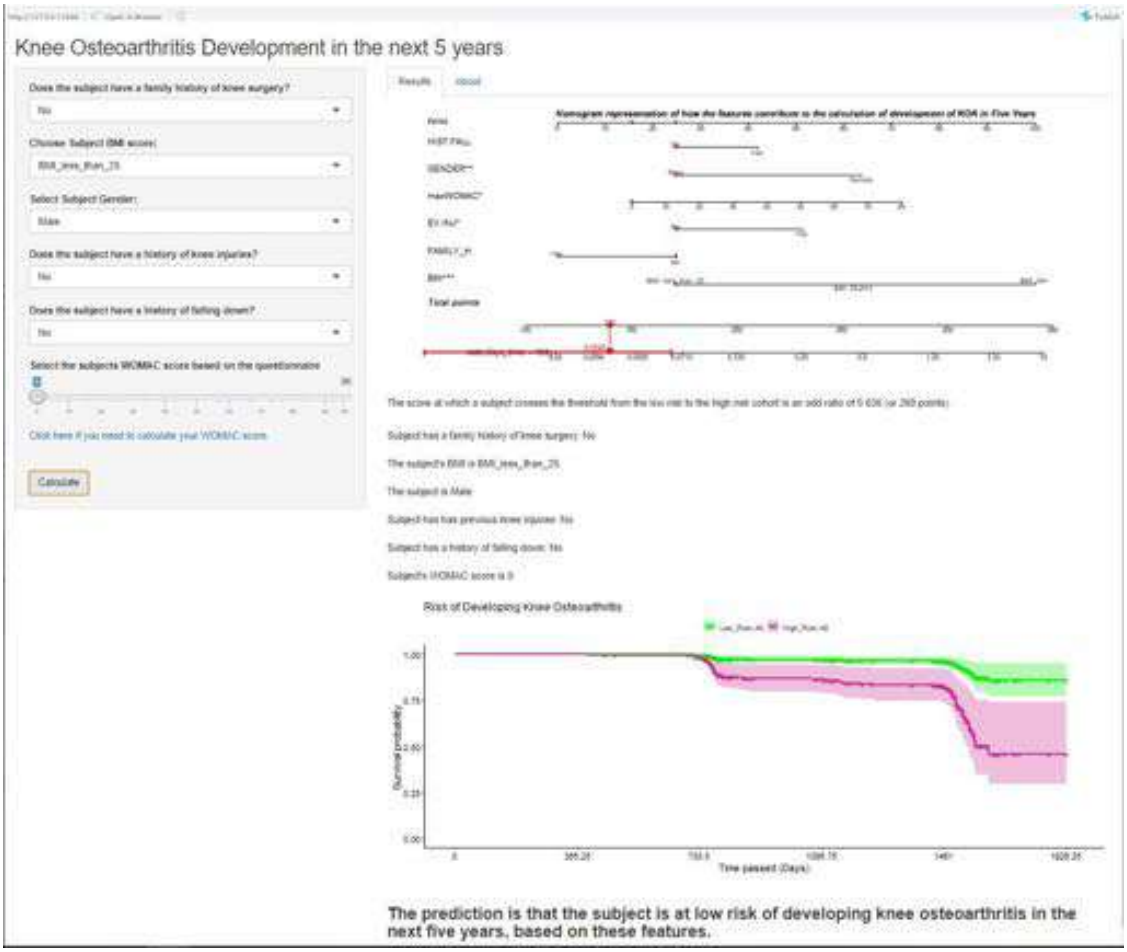


Figure 57. *Snapshot of the app with nomogram feature for the prognostic model. The red points represent the component scores for the example subject, as in figure 46. This time the subject is allocated to the high (purple) or low (green) risk cohort according to the cut-point in the scale at the bottom of the figure. In both nomograms, the graphics clearly show the contribution of each risk factor to the final inference. Changing any of the risk factors carries out scenario analysis, for instance to see the effect of different categories of BMI.*

6. Conclusions

In this Deliverable (D6.5) are presented the working prototypes of the personalised predictive OACTIVE models used either for prevention, diagnosis or even during the intervention stage. Specifically, Section 3 presents four approaches of personalized prediction models. In Section 4 methodologies for non-invasive OA diagnosis are presented and finally in Section 5 Interpretable models are shown.

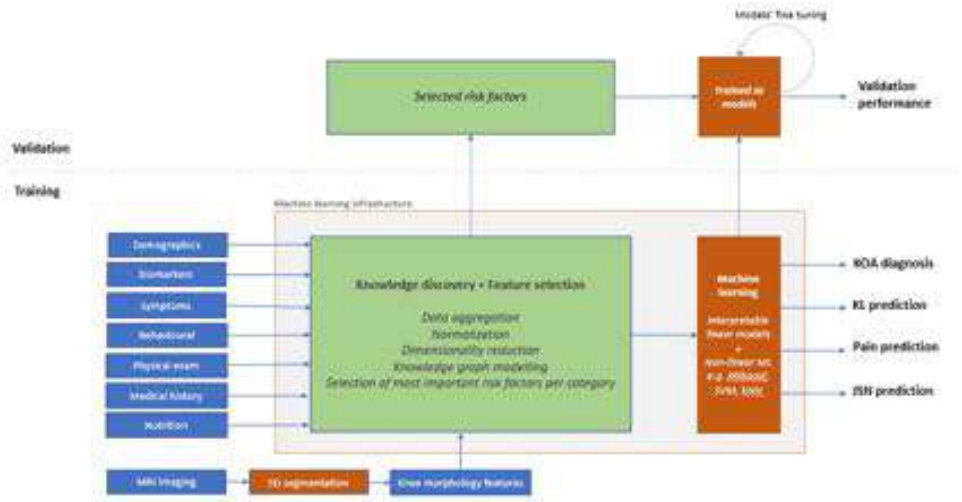


Figure 58. *Hyper-modelling framework of OACTIVE.*

The first approach of the personalized prediction models focuses on the development of a ML-based methodology capable of (i) predicting KOA progression (and specifically KL grades progression) and (ii) identifying important risk factors which contribute to the prediction of KOA. The proposed FS methodology combines well-known approaches including filter, wrapper and embedded techniques whereas feature ranking is decided on the basis of a majority vote scheme to avoid bias. Finally, a variety of ML models were built on the selected features to implement the KOA prediction task (treated as a two-class classification problem where a participant is classified to either the class of KOA progressors or to the non-progressors' class). Apart from the selection of important risk factors, this work also explores three different options with respect to the time period within which data should be considered in order to reliably predict KOA progression. The nature of the selected features was also discussed to increase our understanding of their effect on the KOA progression. After an extensive experimentation, a 74.07% classification accuracy was achieved by SVM on a group of fifty-five selected risk factors (in dataset D).

Furthermore, the second approach consists of two works. The purpose of the first work is: (i) to identify different clusters of KOA pain progression, (ii) to identify informative parameters that are relevant with pain progression from a big pool of risk factors that are available in osteoarthritis initiative (OAI) database and (iii) to build ML models that can predict long-term pain progression using baseline data. To accomplish the aforementioned targets, we built a ML empowered methodology capable of achieving state-of-the-art accuracy results with the minimum possible number of features. Specifically, we have achieved an 84.3% for the prediction of pain on the left leg, and an 82.5% on the right leg. An important observation here is that these high accuracy scores were achieved by using a relatively small subset of features (25 features for the left leg, and 20 for the right leg) that share similar characteristics. It was also observed that the most important features for the pain progression prediction are related directly with the pain on each leg respectively. Furthermore, the second work focuses on the development of patient-specific models of KOA prediction with a special emphasis on pain progression. The presented study contributes to the identification of different clusters of KOA pain progression, the selection of informative and robust parameters that are relevant with pain progression and the development of AI-powered predictive models that could be used for patient-specific prediction of pain progression.

Moreover, the main objective of the third approach was the accurate prediction of JSN in KOA patients based on a machine learning pipeline trained on multimodal data from the OAI (725 features in total were considered). To identify and group patients with and without JSN progression a clustering process (Deliverable 6.3) was initially performed on the JSN progression based on the JSM outcomes of patients

over the first five visits. Afterwards, for the identification of the most important features for the related clusters discrimination (progressing versus non-progressing patients), a hybrid feature selection technique was employed. Finally, the selected features were employed for the training of various ML models in order to predict JSN in KOA patients. The outcome of the ML models indicated that the LR model achieved the best performance for the left leg with a 78.3% accuracy for 164 features, while for the right leg, the SVM model dominated with a 77.7% accuracy for 88 and 90 features. However, the best overall performance was achieved by the second strategy where the data from both legs were combined. Specifically, the LR model achieved an 83.3% accuracy for a significantly lower number of features (29). The overall validation and the interpretation of the models will be taking place in Deliverable 9.3.

The fourth work has the aim to increase the generalization using an evolutionary Machine Learning approach. Specifically, this work focuses on the identification of important and robust risk factors which contribute to KOA progression. The proposed FS methodology relies on an evolutionary machine learning methodology that leads to the selection of a relatively small feature subset (35 risk factors) which generalizes well on the whole dataset (mean accuracy of 71.25%). We investigated the effectiveness of the proposed approach in a comparative analysis with well-known FS techniques with respect to metrics related to both prediction accuracy and generalization capability.

In Section 4 of diagnosis models, in the first task we built classification models with aim to present a machine learning workflow for diagnosis of KOA with a focus on post-hoc explainability (Deliverable 9.3). Overall, understanding the inner workings of ML algorithms is the most important. So, as next step we will validate and will use the SHAP theory, because explainability refers to being able to trace and follow the logic ML algorithms use to form their conclusions. The second approach presents a methodology, which shows potential for non-invasive OA diagnosis. Here we demonstrated its potential to reliably identify informative risk factors from self-reported clinical data and recognize at a certain level participant with symptomatic KOA or being at high risk of developing KOA in at least one knee. A quantum computing perspective of the future application of the proposed methodology is also discussed highlighting the potential to massively speed up certain types of classification problems. Our method may promote future development and clinical implementation of non-invasive tools for KOA diagnosis and prediction. Future work includes the development of machine learning and deep learning models that could predict the progression of the disease using selected risk factors. More emphasis will be given to local prediction models that will be trained on data subgroups defined by parameters such as body mass index combined with demographics and social indicators. The methodology will be finally extended to include parameters from more disciplines including nutrition, medical history, biomarkers and physical measurements of participants performed in the clinic. Research at the intersection of machine learning and clinical research offers great promise for improving OA related research, advancing clinical decision-making and accelerating intervention programs. To enhance appropriate use of machine/deep learning techniques and stay abreast of new developments in advance analytical techniques, open data and scientific tools must be dynamically encouraged within the OA research community.

Interpretable models provide decisions which are made with clarity and that the processes that go into making decisions about a person's health are easily explainable to the patient and understood by doctors. For this task insights about the baseline presentation with and without clinical manifestation of osteoarthritis were derived from statistical and machine learning models. This produced a new risk model based on clinical indicators and questionnaire reports at first presentation. Three interpretable models were presented, rule based, logistic regression and the partial response network which is a self-explaining neural network, with the finding reported at the IDEAL 2019 conference [91]. Further, the inferences made about KL scores were found to correlate reasonably with self-reported pain score WOMAC, indicating that it may be possible to separate the expected effect of radiological disease on pain from an additional subjective

additive component. The KL models are transparent so can represented by nomograms which indicate the weight of evidence in making individual inferences for each subject.

7. References

[1] Silverwood, V., Blagojevic-Bucknall, M., Jinks, C., Jordan, J. L., Protheroe, J., & Jordan, K. P. (2015). Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis. Osteoarthritis and cartilage, 23(4), 507-515.

[2] Ackerman, I. N., Kemp, J. L., Crossley, K. M., Culvenor, A. G., & Hinman, R. S. (2017). Hip and knee osteoarthritis affects younger people, too. journal of orthopaedic & sports physical therapy, 47(2), 67-79.

[3] Lespasio, M. J., Piuzzi, N. S., Husni, M. E., Muschler, G. F., Guarino, A. J., & Mont, M. A. (2017). Knee osteoarthritis: a primer. The Permanente Journal, 21.

[4] Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G., & Tsaopoulos, D. E. (2020). Machine learning in knee osteoarthritis: A review. Osteoarthritis and Cartilage Open, 100069.

[5] Lazzarini, N., Runhaar, J., Bay-Jensen, A. C., Thudium, C. S., Bierma-Zeinstra, S. M. A., Henrotin, Y., & Bacardit, J. (2017). A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. Osteoarthritis and cartilage, 25(12), 2014-2021.

[6] Halilaj, E., Le, Y., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018). Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative. Osteoarthritis and cartilage, 26(12), 1643-1650.

[7] Pedoia, V., Haefeli, J., Morioka, K., Teng, H. L., Nardo, L., Souza, R. B., ... & Majumdar, S. (2018). MRI and biomechanics multidimensional data analysis reveals R2-R1ρ as an early predictor of cartilage lesion progression in knee osteoarthritis. Journal of Magnetic Resonance Imaging, 47(1), 78-90.

[8] Abedin, J., Antony, J., McGuinness, K., Moran, K., O'Connor, N. E., Rebholz-Schuhmann, D., & Newell, J. (2019). Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images. Scientific reports, 9(1), 1-11.

[9] Nelson, A. E., Fang, F., Arbeeva, L., Cleveland, R. J., Schwartz, T. A., Callahan, L. F., ... & Loeser, R. F. (2019). A machine learning approach to knee osteoarthritis phenotyping: data from the FNIH Biomarkers Consortium. Osteoarthritis and cartilage, 27(7), 994-1001.

[10] Tiulpin, A., Klein, S., Bierma-Zeinstra, S. M., Thevenot, J., Rahtu, E., van Meurs, J., & Saarakkala, S. (2019). Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Scientific reports, 9(1), 1-11.

[11] Widera, P., Welsing, P. M., Ladel, C., Loughlin, J., Lafeber, F. P., Dop, F. P., ... & Bacardit, J. (2020). Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data. Scientific reports, 10(1), 1-15.

[12] Alexos, A., Moustakidis, S., Kokkotis, C., & Tsaopoulos, D. (2020, May). Physical activity as a risk factor in the progression of osteoarthritis: a machine learning perspective. In International Conference on Learning and Intelligent Optimization (pp. 16-26). Springer, Cham.

[13] Ashinsky, B. G., Bouhrara, M., Coletta, C. E., Lehallier, B., Urish, K. L., Lin, P. C., ... & Spencer, R. G. (2017). Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. Journal of Orthopaedic Research, 35(10), 2243-2250.

[14] Donoghue, C., Rao, A., Bull, A. M., & Rueckert, D. (2011, March). Manifold learning for automatically predicting articular cartilage morphology in the knee with data from the osteoarthritis initiative (OAI). In

Medical Imaging 2011: Image Processing (Vol. 7962, p. 79620E). International Society for Optics and Photonics.

[15] Marques, J., Genant, H. K., Lillholm, M., & Dam, E. B. (2013). Diagnosis of osteoarthritis and prognosis of tibial cartilage loss by quantification of tibia trabecular bone from MRI. Magnetic resonance in medicine, 70(2), 568-575.

[16] Yoo, T. K., Kim, S. K., Choi, S. B., Kim, D. Y., & Kim, D. W. (2013, July). Interpretation of movement during stair ascent for predicting severity and prognosis of knee osteoarthritis in elderly women using support vector machine. In 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 192-196). IEEE.

[17] Moustakidis, S., Christodoulou, E., Papageorgiou, E., Kokkotis, C., Papandrianos, N., & Tsaopoulos, D. (2019). Application of machine intelligence for osteoarthritis classification: a classical implementation and a quantum perspective. Quantum Machine Intelligence, 1(3), 73-86.

[18] Rockel, J. S., Zhang, W., Shestopaloff, K., Likhodii, S., Sun, G., Furey, A., ... & Kapoor, M. (2018). A classification modeling approach for determining metabolite signatures in osteoarthritis. PloS one, 13(6), e0199618.

[19] Kobayashi, T., Kannari, T., Horiuchi, H., Matsui, N., Ito, T., Nojin, K., ... & Yamanaka, M. (2019). Predictors affecting balance performances in patients with knee osteoarthritis using decision tree analysis. Osteoarthritis and Cartilage, 27, S243.

[20] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[21] Gornale, S. S., Patravali, P. U., Marathe, K. S., & Hiremath, P. S. (2017). Determination of osteoarthritis using histogram of oriented gradients and multiclass SVM. International Journal of Image, Graphics and Signal Processing, 9(12), 41.

[22] Kotti, M., Duffell, L. D., Faisal, A. A., & McGregor, A. H. (2017). Detecting knee osteoarthritis and its discriminating parameters using random forests. Medical engineering & physics, 43, 19-29.

[23] Torlay, L.; Perrone-Bertolotti, M.; Thomas, E.; Baciu, M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform. 2017, 4, 159–169.

[24] Du, Y., Shan, J., & Zhang, M. (2017, November). Knee osteoarthritis prediction on MR images using cartilage damage index and machine learning methods. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 671-677). IEEE.

[25] Quinlan, J. R. (1986). Induction of Decision Trees. Mach. Learn.

[26] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[27] Cortes, C., & Vapnik, V. (1995). Support-vector networks Machine learning (pp. 237–297), Vol. 20. Boston, MA: Kluwer Academic Publisher.

[28] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[29] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[30] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[31] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. In The elements of statistical learning (pp. 337-387). Springer, New York, NY.

[32] Kleinbaum, D. G., & Klein, M. (2010). Logistic regression, statistics for biology and health. Retrieved from DOI, 10, 978-1.

[33] Taud, H., & Mas, J. F. (2018). Multilayer perceptron (MLP). In Geomatic Approaches for Modeling Land Change Scenarios (pp. 451-455). Springer, Cham.

[34] Ntakolia, C., Diamantis, D. E., Papandrianos, N., Moustakidis, S., & Papageorgiou, E. I. (2020, December). A Lightweight Convolutional Neural Network Architecture Applied for Bone Metastasis Classification in Nuclear Medicine: A Case Study on Prostate Cancer Patients. In Healthcare (Vol. 8, No. 4, p. 493). Multidisciplinary Digital Publishing Institute.

[35] Perez, A., Larranaga, P., & Inza, I. (2006). Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. International Journal of Approximate Reasoning, 43(1), 1-25.

[36] Jahromi, A. H., & Taheri, M. (2017, October). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. In 2017 Artificial Intelligence and Signal Processing Conference (AISP) (pp. 209-212). IEEE.

[37] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[38] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[39] Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.

[40] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[41] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics-Proteomics, 15(1), 41-51.

[42] Dodge, Y., & Commenges, D. (Eds.). (2006). The Oxford dictionary of statistical terms. Oxford University Press on Demand.

[43] Shanker, M., Hu, M. Y., & Hung, M. S. (1996). Effect of data standardization on neural network training. Omega, 24(4), 385-397.

[44] Rockel, J. S., Zhang, W., Shestopaloff, K., Likhodii, S., Sun, G., Furey, A., ... & Kapoor, M. (2018). A classification modeling approach for determining metabolite signatures in osteoarthritis. PloS one, 13(6), e0199618.

[45] Gornale, S. S., Patravali, P. U., Marathe, K. S., & Hiremath, P. S. (2017). Determination of osteoarthritis using histogram of oriented gradients and multiclass SVM. International Journal of Image, Graphics and Signal Processing, 9(12), 41.

[46] Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

[47] Roffo, G., Melzi, S., Castellani, U., & Vinciarelli, A. (2017). Infinite latent feature selection: A probabilistic latent graph-based ranking approach. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1398-1406).

[48] Roffo, G., Melzi, S., & Cristani, M. (2015). Infinite feature selection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4202-4210).

[49] Shahbaz, M. B., Wang, X., Behnad, A., & Samarabandu, J. (2016, October). On efficiency enhancement of the correlation-based feature selection for intrusion detection systems. In 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 1-7). IEEE.

[50] Hagos, D. H., Yazidi, A., Kure, Ø., & Engelstad, P. E. (2017, March). Enhancing security attacks analysis using regularized machine learning techniques. In 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA) (pp. 909-918). IEEE.

[51] Nguyen, H. T., Franke, K., & Petrovic, S. (2010, August). Towards a generic feature-selection measure for intrusion detection. In 2010 20th International Conference on Pattern Recognition (pp. 1529-1532). IEEE.

[52] Kokkotis, C., Moustakidis, S., Giakas, G., & Tsaopoulos, D. (2020). Identification of Risk Factors and Machine Learning-Based Prediction Models for Knee Osteoarthritis Patients. Applied Sciences, 10(19), 6797.

[53] Cooper, C., Snow, S., McAlindon, T. E., Kellingray, S., Stuart, B., Coggon, D., & Dieppe, P. A. (2000). Risk factors for the incidence and progression of radiographic knee osteoarthritis. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology, 43(5), 995-1000.

[54] Hartley, A.; Hardcastle, S.A.; Paternoster, L.; McCloskey, E.; Poole, K.E.; Javaid, M.K.; Aye, M.; Moss, K.; Granell, R.; Gregory, J. Individuals with High Bone Mass have increased progression of radiographic and clinical features of knee osteoarthritis. Osteoarthr. Cartil. 2020, 28, 1180–1190

[55] Blagojevic, M., Jinks, C., Jeffery, A., & Jordan, 1. (2010). Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. Osteoarthritis and cartilage, 18(1), 24-33.

[56] Heidari, B. (2011). Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I. Caspian journal of internal medicine, 2(2), 205.

[57] Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. ieee ComputatioNal iNtelligeNCe magaziNe, 13(4), 59-76.

[58] Yoo, T. K., Kim, D. W., Choi, S. B., Oh, E., & Park, J. S. (2016). Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. PloS one, 11(2), e0148724.

[59] Long, M. J., Papi, E., Duffell, L. D., & McGregor, A. H. (2017). Predicting knee osteoarthritis risk in injured populations. Clinical biomechanics, 47, 87-95.

[60] Lim, J., Kim, J., & Cheon, S. (2019). A deep neural network-based method for early detection of osteoarthritis using statistical data. International journal of environmental research and public health, 16(7), 1281.

[61] Christodoulou, E., Moustakidis, S., Papandrianos, N., Tsaopoulos, D., & Papageorgiou, E. (2019, July). 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). IEEE.

[62] Navale, D. I., Hegadi, R. S., & Mendgudli, N. (2015, December). Block based texture analysis approach for knee osteoarthritis identification using SVM. In 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 338-341). IEEE.

[63] Eckstein, F., W. Wirth and M. C. Nevitt (2012). "Recent advances in osteoarthritis imaging--the osteoarthritis initiative." Nat Rev Rheumatol 8(10): 622-630.

[64] Duda, R. O., P. E. Hart and D. G. Stork (2012). Pattern classification, John Wiley & Sons.

[65] Belson, W. A. (1959). "Matching and Prediction on the Principle of Biological Classification." Journal of the Royal Statistical Society. Series C (Applied Statistics) 8(2): 65-75.

[66] Witten, I. H., E. Frank, M. A. Hall and C. J. Pal (2016). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.

[67] Atkeson, C. G., A. W. Moore and S. Schaal (1997). "Locally Weighted Learning." Artificial Intelligence Review 11(1): 11-73.

[68] Scholkopf, B. (1997). "Support vector learning." Ph. D. thesis, Technische Universitat Berlin.

[69] Freund, Y. and R. E. Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of computer and system sciences 55(1): 119-139.

[70] Breiman, L. (2001). "Random forests." Machine learning 45(1): 5-32.

[71] Keller, J., Gray, M., & Givens, J. (1985). A fuzzy K-nearest neighbor algorithm. IEEE Transactions On Systems, Man, And Cybernetics, SMC-15(4), 580-585. doi: 10.1109/tsmc.1985.6313426.

[72] Zhai, J., Zhai, M., & Kang, X. (2014). Condensed fuzzy nearest neighbor methods based on fuzzy rough set technique. Intelligent Data Analysis, 18(3), 429-447.

[73] LeCun, Y., Y. Bengio and G. Hinton (2015). "Deep learning." Nature 521(7553): 436-444.

[74] LeCun, Y., S. Chopra, R. Hadsell, M. Ranzato and F. Huang (2006). "A tutorial on energy-based learning." Predicting structured data 1(0).

[75] Zeiler, M. D. (2012). "ADADELTA: an adaptive learning rate method." arXiv preprint arXiv:1212.5701.

[76] Havlíček, V., Córcoles, A., Temme, K., Harrow, A., Kandala, A., Chow, J., & Gambetta, J. (2019). Supervised learning with quantum-enhanced feature spaces. Nature, 567(7747), 209-212. doi: 10.1038/s41586-019-0980-2.

[77] Farhi, E. & Neven, H. (2018). Classification with quantum neural networks on near term processors. Preprint at https://arxiv.org/abs/1802.06002.

[78] McClean, J., Boixo, S., Smelyanskiy, V., Babbush, R., & Neven, H. (2018). Barren plateaus in quantum neural network training landscapes. Nature Communications, 9(1). doi: 10.1038/s41467-018-07090-4.

[79] Kerkhof, H. J. M., Bierma-Zeinstra, S. M. A., Arden, N. K., Metrustry, S., Castano-Betancourt, M., Hart, D. J., ... & Uitterlinden, A. G. (2014). Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. Annals of the rheumatic diseases, 73(12), 2116-2121.

[80] Zhang, W., McWilliams, D. F., Ingham, S. L., Doherty, S. A., Muthuri, S., Muir, K. R., & Doherty, M. (2011). Nottingham knee osteoarthritis risk prediction models. Annals of the rheumatic diseases, 70(9), 1599-1604.

[81] Losina, E., Klara, K., Michl, G. L., Collins, J. E., & Katz, J. N. (2015). Development and feasibility of a personalized, interactive risk calculator for knee osteoarthritis. BMC musculoskeletal disorders, 16(1), 1-12.

[82] Yoo, T. K., Kim, D. W., Choi, S. B., Oh, E., & Park, J. S. (2016). Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. PloS one, 11(2), e0148724.

[83] Joseph, G. B., McCulloch, C. E., Nevitt, M. C., Neumann, J., Gersing, A. S., Kretzschmar, M., ... & Link, T. M. (2018). Tool for osteoarthritis risk prediction (TOARP) over 8 years using baseline clinical data, X-ray, and MRI: Data from the osteoarthritis initiative. Journal of Magnetic Resonance Imaging, 47(6), 1517-1526.

[84] NIMH, "OAI," nda.nih.gov. [Online]. Available: https://nda.nih.gov/oai/. [Accessed: 17-Oct-2019].

[85] NDA, "OAI," nda.nih.gov. [Online]. Available: https://nda.nih.gov/oai/study-details/schedule-of-assessments.html. [Accessed: 20-Oct-2019].

[86] Pigott, T. D. (2001). A review of methods for missing data. Educational research and evaluation, 7(4), 353-383.

[87] Kindermans, P. J., Schütt, K. T., Alber, M., Müller, K. R., Erhan, D., Kim, B., & Dähne, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. arXiv preprint arXiv:1705.05598.

[88] Van Calster, B., Timmerman, D., Nabney, I. T., Valentin, L., Van Holsbeke, C., & Van Huffel, S. (2006, August). Classifying ovarian tumors using Bayesian Multi-Layer Perceptrons and Automatic Relevance Determination: a multi-center study. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 5342-5345). IEEE.

[89] Lisboa, P. J. G., Ortega-martorell, S. Cashman, S., and Olier, I. (2019). "The Partial Response Network".

[90] Fawcett, T. (2006). "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874.

[91] McCabe, P. G., Olier, I., Ortega-Martorell, S., Jarman, I., Baltzopoulos, V., & Lisboa, P. (2019, November). Comparative Analysis for Computer-Based Decision Support: Case Study of Knee Osteoarthritis. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 114-122). Springer, Cham.