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	D9.3
Deliverable title	Evaluation of OACTIVE models in big data registries
Submission month of deliverable	M42
Issuing partner	LJMU
Contributing partners	ALL
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	2/4/2021	LJMU, CERTH	First Draft
1.1	16/4/2021	UNIC, HULAFE	Review of First Draft
1.2	28/4/2021	LJMU, CERTH	Review of Second Draft
1.3	30/4/2021	LJMU	Final Version Review & Submission

Contents

1. Summary

The first aim of this deliverable (Task 9.2) is to present the results and outcomes of the long-term evaluation of OACTIVE using data from big data registries. The rest of this deliverable is organised as follows. Section 2 presents the main concept of this deliverable. Section 3 Baseline performance with PCA presents an extensive analysis of the OAI database. The validation of the proposed OActive models is given in Section 4. In Section 5 Machine Learning and Deep Learning Diagnosis models with focus on accuracy and fairness are presented. Section 6 presents an approach for the quantification of MRI impact. The Interpretable models are presented in Section 7. The deliverable ends with Section 8 Conclusions, which mentions the future expectations and advantages of the proposed works in knee osteoarthritis.

This report refers to Deliverable 9.3, which relates to the OACTIVE WP 9, "Technology assessment and full system validation" led by LJMU. The objective of WP9 is to validate the integrated OACTIVE system by employing a comprehensive methodology that involves: (i) Clinical studies in human populations and (ii) validation of the system using big data registries. The ethical, legal, and social challenges that need to be met in order for the scientific advances to be responsibly applied will be finally investigated.

2. Introduction

The integrated OACTIVE hyper-model validated using a big data registry, namely Osteoarthritis Initiative (OAI). This initiative has collected substantial amounts of imaging, lifestyle, and biochemical data and other complementary data streams on the healthy subjects, and patients affected by Osteoarthritis (OA). In OACTIVE, we combined all of this information allowing for the first time the simultaneous exploration of multiple risk factors in big human populations involving thousands of patients. In task, the big data methodology, developed in WP6, will search through massive amounts of information, analysing it to predict outcomes for individual patients. That information will include data from past treatment outcomes with the outcome not only to predict but also to reveal surprising associations in data that our human brains would never suspect. In terms of the validation, patient-specific data with follow-ups of more than 100 months will be used as follows: the personalised models will be built using data from the first months and the efficiency of OACTIVE will be tested using the latest data available against the following criteria: prediction accuracy and maximum prediction timeframe. The personalised models will be finally progressively updated incorporating more information from subsequent months and the predictive performance of the models will be estimated per case.

In this report, we will present an extensive analysis of the OAI database. We start our analysis by presenting the main characteristics of this database, in terms of correlation, with Pearson's correlation coefficient technique, and in terms of variable importance analysis, by utilizing the Random Forests Variable Importance. Afterwards, we visualise the data into two dimensions by utilising the PCA technique, and we colour each sample with the class it represents. Finally, in order to validate the datasets, we apply several classification models, such as Multinomial Logistic Regression, Linear Discriminant Analysis, k-Nearest Neighbours, Random Forests, and XGBoost.

Then, we proceeded to the evaluation of the personalized models, which were developed in Deliverable 6.3 and Deliverable 6.5. The training and the validation of the personalized models was based on the Osteoarthritis Initiative (OAI) database. Initially, we validated the personalized prediction of KL progression model. The novelty of the proposed FS methodology lies in 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. A 74.07% classification accuracy was achieved by SVM on the group of the first fifty-five selected risk factors. Furthermore, the effectiveness of the proposed approach was evaluated in a comparative analysis with

respect to classification errors and confusion matrices to confirm its clinical relevance. Then, two approaches validated for the personalized prediction model of the Pain progression. The main goal of the study was to build a prognostic tool that will predict the progression of pain in KOA patients using data collected at baseline. 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, and results up to 84.3% have been achieved with only a small amount of features. The proposed methodology demonstrated unique potential in identifying pain progression at an early stage, therefore, improving future KOA prevention efforts. The second approach relies on an innovative evolutionary Machine Learning 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.

Subsequently, the personalised prediction of knee joint space narrowing (JSN) progression was validated in each knee and in both knees combined. The proposed methodology employs: (i) A clustering process to identify groups of people with progressing and non-progressing JSN; (ii) a robust feature selection (FS) process consisting of filter, wrapper, and embedded techniques that identifies the most informative risk factors; (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; and (iv) post-hoc interpretation of the features' impact on the best performing model. The results showed that bounding the JSN progression of both knees can result in more robust prediction models with higher accuracy (83.3%) and with fewer risk factors (29) compared to the right knee (77.7%, 88 risk factors) and the left knee (78.3%, 164 risk factors), separately.

To increase the generalization, we used and validated an evolutionary Machine Learning approach as described in Deliverable 6.5. This work 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, leading to the selection of a relatively small feature subset of 35 risk factors that 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. The impact of the selected risk factors on the prediction output was further investigated using SHapley Additive exPlanations (SHAP).

Furthermore, we validated the robust data mining approach that could identify important risk factors which contribute to the diagnosis of KOA and their impact on model output, with a focus on post-hoc explainability. The validation of the extracted factors was performed in subgroups employing seven well-known classifiers. A 77.88 % classification accuracy was achieved by Logistic Regression on the group of the first forty selected (40) risk factors. We investigated the behavior of the best model, with respect to classification errors and the impact of used features, to confirm their clinical relevance. The interpretation of the model output was performed by SHAP.

At this point, we ought to emphasize that this work makes a contribution towards KOA diagnosis through the application of DNN models on self-reported clinical data**.** To the best of our knowledge, this work contains original content in the first-ever validation of machine and deep learning models with respect to fairness in the KOA classification research. Through this study, different DNN architectures were tested for their ability to recognise participants with symptomatic KOA or being at high risk of developing KOA

in one knee at least. Different subgroups were investigated defined by gender, age, and obesity. The subgroups considered are (i) participants older than 70 years, (ii) participants under 70 years old, (iii) male participants, (iv) female participants, (v) non-obese, and (vi) obese participants. The performance of the proposed DL methodology was validated in terms of both accuracy and fairness calculated using the aforementioned subgroups. Finally, a comparative analysis was conducted with various benchmark machine learning algorithms aiming to show the superiority of the proposed DNN structure for the knee OA classification task.

Quantifying the MRI impact, we present an extensive analysis of the OAI MRI dataset. To perform a complete analysis of this data we distinct it in 9 variables' categories dataset, and we analyse them separately. We start our analysis by presenting the main characteristics this database. In terms of correlation, with Spearman's correlation coefficient technique. In terms of variable importance analysis by utilizing the Random Forests Variable Importance. We do an extensive analysis of both the Principal Components Analysis and the Random Projections techniques to apply dimensionality reduction in our data. Afterwards, we visualise the data into two dimensions by utilising the PCA technique, and we colour each sample with the class it represents. Finally, in order to validate the datasets, we apply several classification models, such as Multinomial Logistic Regression, Linear Discriminant Analysis, k-Nearest Neighbours, Random Forests and XGBoost. End, predictions for clinical data were also validated on the Multicenter Osteoarthritis Study (MOST), which has over 2000 observations acquired with a different protocol than the OAI study. This allows an extensive external validation to be carried out for the diagnostic and prognostic statistical models developed by OACTIVE.

3. Baseline performance with PCA

Clinical Data Presentation

The entire OAI data set consists of 1187 variables, one response variable, and 4796 observations. From the 1187 variables, for the classification we are about to execute, 727 variables were suggested for use to be relevant to the classes we are about to predict.

From the 4796 observations, we removed the observation with no value in the response variable, this way only 1936 observations were left in the data set. Finally, from the initial analysis of the data, it was determined that from the 727 variables, the 566 variables are categorical and the remaining 161 are numerical variables.

The response variable is a two classes categorical variable. The first class, represented with "0", is the class of a person without osteoarthritis which will not present osteoarthritis in the near future. The second class, represented with "1", is the class with a person who in the near future will present osteoarthritis in one or both of his knees. From now on the second class represented with "1" will be considered as the positive class.

Correlation analysis of the selected variables saw that there are highly correlated variables inside the data set [1]. In the figure below we can see how many variable pairs are correlated (positively or negatively) and to what degree.

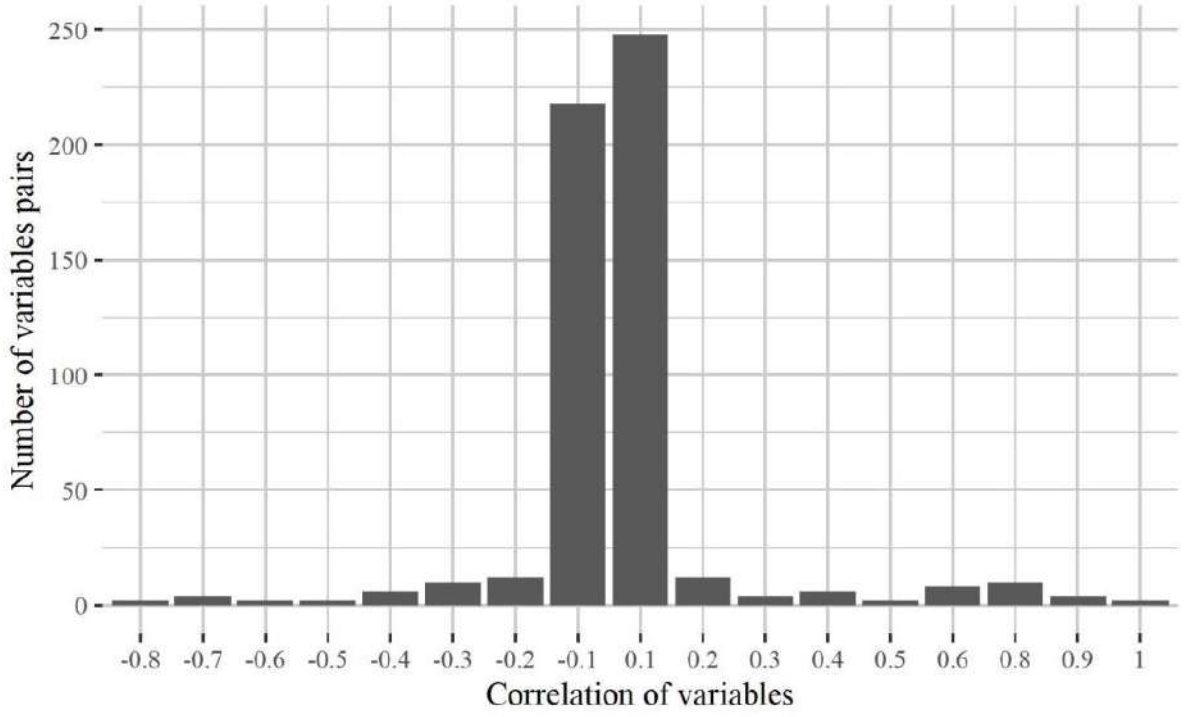


Figure 1. *Correlation of variables.*

To produce the above figure, a correlation matrix was created with the Pearson's correlation coefficient metric. From this matrix, we removed the values of the major diagonal because it represents pairs of the same variables which correlate 1.

Data preprocessing

The first step of the preprocessing of the data is the removal of the highly correlated variables. The variables were removed where variables with absolute correlation value above 0.7. As a result, 10 variables were removed leaving 717 variables. The variables removed are "V00WPLKN5", "V00WOMKPL", "V00WPLKN1", "V00KOOSYML", "V00FFQFLG5", "V00FFQFLG1", "V00NSKIP", "V00FFQFLG2", "P01SVXRRKR" and "P01SVRKMI".

Feature selection (Variable Importance)

One more preprocessing step executed for the data set is the Variable Importance Analysis with the use of the Random Forest Algorithm (RF). By using the Mean Decrease in Accuracy (MDA) measure. The resulted variables' selection was NOT used in the analysis. Its purpose is a better understanding of the data set.

To calculate the MDA with the RF algorithm the permuted out-of-bag (OOB) data were used. Specifically, it was done by recording the prediction error on the OOB portion of data, for each tree. The same process is repeated after permuting each predictor variable. The difference between the two (Decreases in Accuracy of Trees) is then averaged over all trees, and normalized by the standard deviation of the differences.

$$\text{MDA} = \frac{Mean(Decreases\ in\ Accuracy\ of\ Trees)}{StandardDeviation(Decreases\ in\ Accuracy\ of\ Trees)}$$

For the calculation of the importance of the OActive's dataset variables, an RF model was created with the use of 717 raw variables. This model creation was possible because RF is a decision tree-based algorithm.

This means that there is no need to convert the categorical variables to numeric with the use of dummy variables.
That way the following figure was created. In this figure, the 100 variables with the highest MDA score, are presented in descending order, from the most important to the less important.

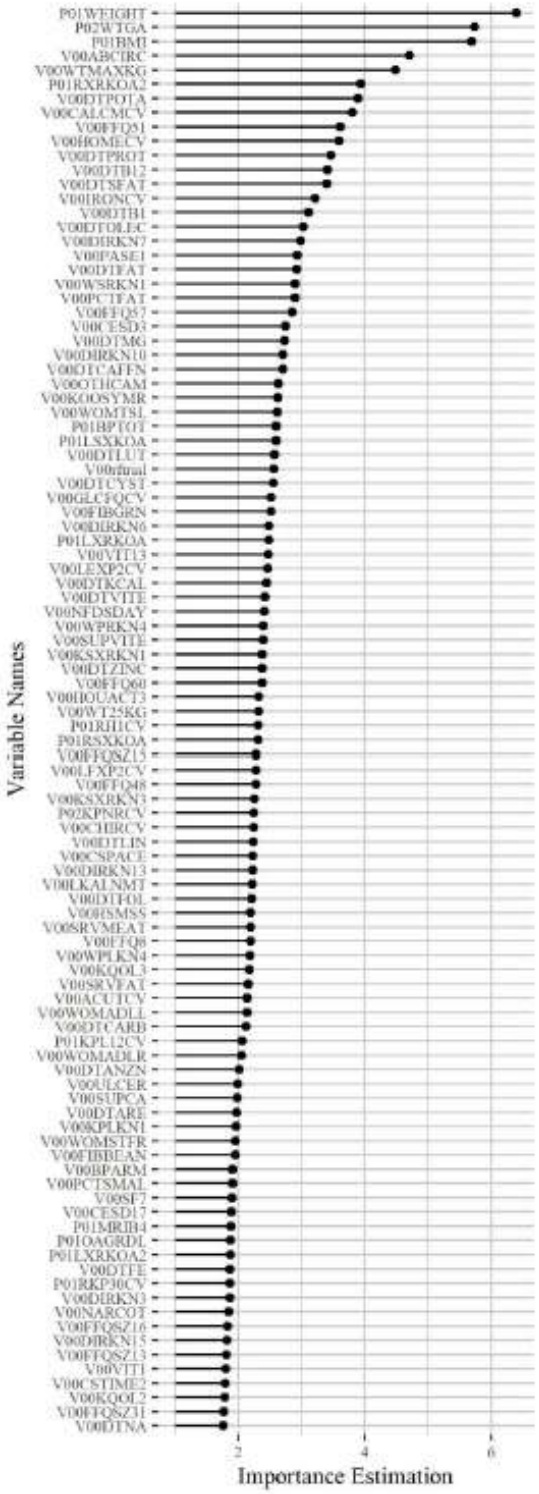


Figure 2. *Variables with the highest MDA score.*

Missing values treatment methods

The next preprocessing step of the analysis is the treatment of the missing values. The final count of values in the data set is 1,409,408 and from those 49,180 values are missing consisting the 3.49% of missing values. After careful examination of the data set, it was determined that the type of missing values mechanism, is that the values are missing completely at random. Furthermore, the missing values are scattered within the data set.

For this reason, no special imputation method was used in the data set. The imputation methods used were mean imputation for the numerical variables and simple dummy variables creation for the categorical.

Data normalization

For the normalization of the data, the technique of min-max scaling or min-max normalization was used. This technique consists of rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$\chi' = \frac{x - min(x)}{max(x) - min(x)},$$

where x is an original value, and x ' is the normalized value. After the scaling of the data, the ranges of the 100 first variables are sawn in the figure below.

Dimensionality reduction

From the preprocessing of the data so far, a huge number of new predictor variables was created, enough so that even state-of-the-art classification techniques would strangle to produce acceptable results. The solution to that challenge is the use of a dimensionality reduction technique [2,3].

The dimensionality reduction technique used for this data set is the technique of Random Projection. Specifically, a projection matrix was created with the use of Gaussian Distribution. The dimensions of the resulted vector space are 370 as it was recommended from the implementation method RanPro of the Random Projections technique in R. This recommendation was produced by using the Johnson-Lindenstrauss (JL) lemma with an error tolerance of 0.5.

Visualization of the data

For the visualization of the data, the Principal Components Analysis was utilized. The projection of the data on the first 2 principal components is given in Figure 3 below.

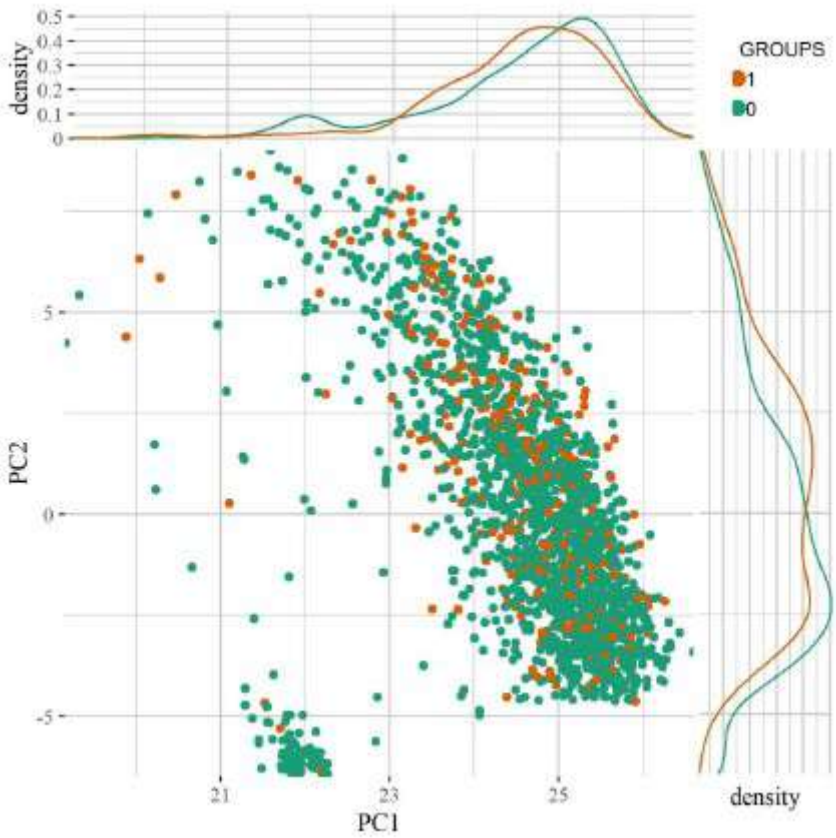


Figure 3. *Principal Components Analysis for OAI data.*

As it is immediately apparent from the representation of the data, the Classes "0" and "1" aren't segregated at all. As the density plots of each principal component represent, they have almost identical distribution in the space. Form this visualization the first conclusion that is drawn is that we have an almost impossible classification problem.

Classification tasks

The classification algorithms used for this data set are the Multinomial Logistic Regression (MLR), the Linear Discriminant Analysis (LDA), the k-Nearest Neighbors (kNN), the Random Forest (RF), and finally the XGBoost.

The results of the classifications can be seen in Table 1 below. The measures used are accuracy, sensitivity, and specificity.

Table 1. *Classification results.*

	ACCURACY	SENSITIVITY	SPECIFICITY
MLR	0.838	0	1
LDA	0.703	0.159	0.864
KNN	0.765	0.022	0.986
RF	0.770	0	1
XGBOOST	0.760	0	0.986

As it becomes apparent from the results of all the classification algorithms, even though they have good accuracy, they cannot predict the positive class well enough or not at all. This resulted accuracy is an

immediate consequence of the amount of "0" contained response variable. Specifically, 82.4% of the results belong to the class "0" and only 17.6% of the results belong to the class "1". In Figures 4-8 below, the Ro Curves illustrating the above problem are presented:

MLR

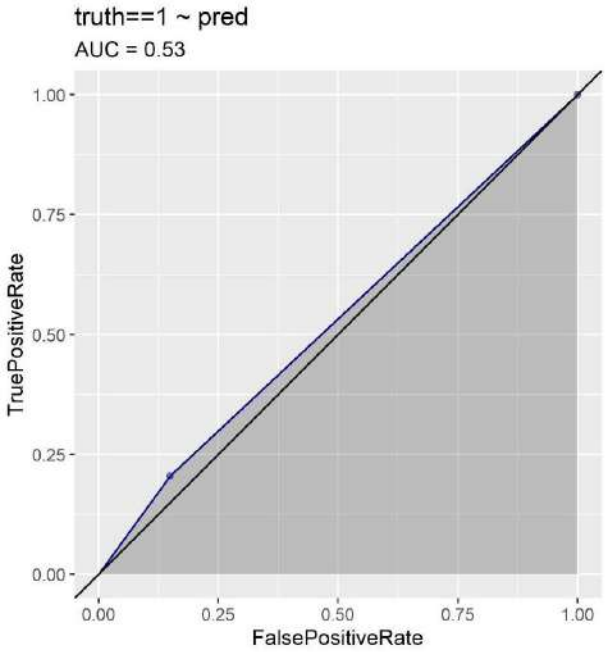


Figure 4. *AUC for MLR model.*

LDA

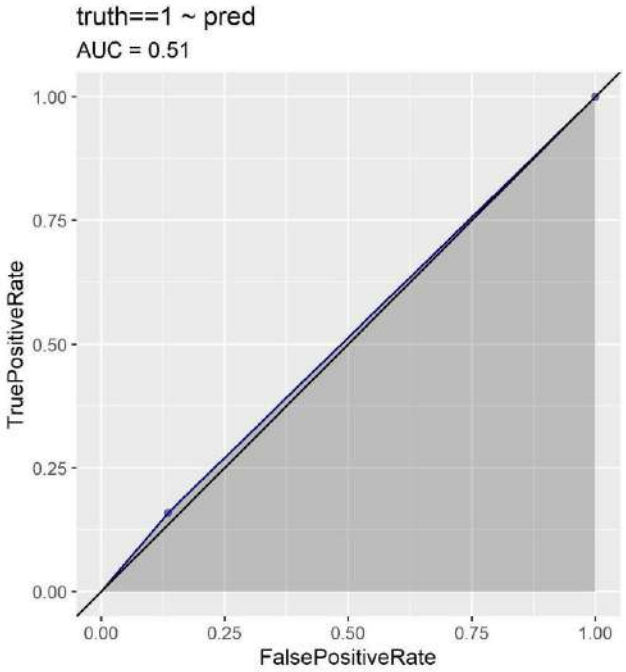


Figure 5. *AUC for LDA model.*

KNN

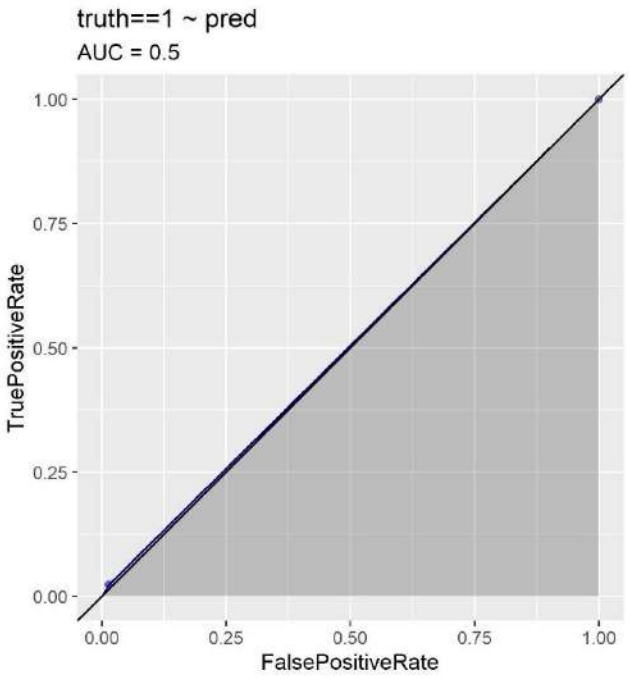


Figure 6. *AUC for KNN model.*

RF

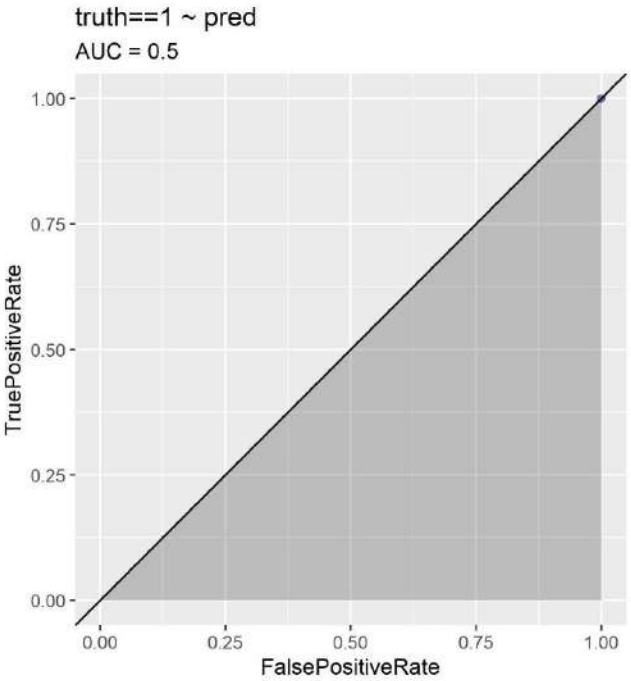


Figure 7. *AUC for RF model.*

XGBoot

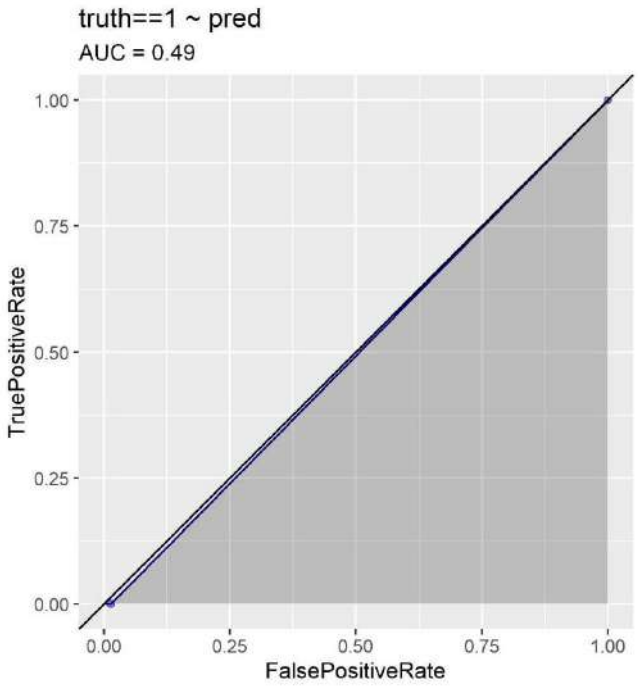


Figure 8. *AUC for XGBoost model.*

The extensive analysis presented in this report shows us the high quality of OAI's database. The findings of this analysis gave us valuable insight into this database. We found that it is a dataset that contains a large number of correlated variables. This results in a not-so-separable dataset, as shown even in all classification tasks. In the first look, the resulted accuracy is over 0.7 in all cases. But if we count the unbalanced nature of the data in combination with the sensitivity, which is below 0.16 in all cases, we can conclude that the positive class is not predicted by the classification models. To cope with this limitation of the data we proposed several technics which have described in Deliverable 6.3 and Deliverable 6.5 and validated below.

4. Validation of OACTIVE Models

In this section, we provide the evaluation criteria, the performance, and the interpretation of the personalised models [4-8].

4.1 Personalised Prediction of KL progression

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. We also evaluated the classifiers performance in terms of the confusion matrix as an additional evaluation criterion.

Confusion matrix is a way to evaluate the performance of a classifier. Specifically, a confusion matrix is a summary of prediction results on a classification problem (Table 2). To be created the confusion matrix, the number of correct (true) and incorrect (false) predictions are summarized with count values and broken down by each class.

Table 2. *Confusion matrix.*

		Actual Classes	
		Positive	Negative
Predicted classes	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Prediction Performance

The proposed ML methodology was applied to each of the five datasets. Specifically, the proposed FS was executed on the pre-processed versions of the datasets ranking the available features for 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

Table 3 summarizes the results of logistic regression, XGboost, SVM, random forest, KNN, naive Bayes and DT on the two-class problem. A moderate number of features (in the range of 30–55) was finally selected by the majority of the ML models (in five out of the seven), whereas the overall maximum was achieved by LR on a group of fifty selected (50) risk factors. KNN and DTs selected more features (145 and 85, respectively) leading to low accuracies. The second highest accuracy was received for SVM and Naive Bayes (70.73% in both), whereas lower accuracies were obtained by NB, RF, and XGboost.

Table 3. *Best testing accuracies achieved for ML model along with the confusion matrix, the optimum number of features, and the hyperparameters of the ML models employed. A1 and A2 denote classes 1 and 2 of dataset A, respectively.*

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
Logistic Regression	71.71		A1	A2	50	Penalty: l1, C: 1.0
		A1	73	28		
		A2	30	74		
Naive Bayes	70.73		A1	A2	55	GaussianNB
		A1	72	29		
		A2	31	73		
SVM	70.73		A1	A2	45	C = 2, kernel = sigmoid
		A1	75	26		
		A2	34	70		
KNN	66.83		A1	A2	145	leaf_size: 1, n_neighbors: 12, weights: distance
		A1	78	23		
		A2	45	59		
Decision Tree	65.85		A1	A2	85	max_features: log2, min_samples_leaf: 4, min_samples_split: 11
		A1	68	33		
		A2	37	67		
Random Forest	68.78		A1	A2	30	criterion: gini, min_samples_leaf: 3, min_samples_split: 7, n_estimators: 15
		A1	71	30		
		A2	34	70		
XGboost	67.8		A1	A2	45	

		A1	69	32	gamma: 0, max_depth: 1,
		A2	34	70	min_child_weight: 4

- Dataset B

The accuracies and confusion matrixes reported in Table 4 verify the aforementioned results. In all the competing models, the best accuracies were recorded using a relatively small number of selected risk factors (less or equal to 40).

Table 4. *Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features, and the hyperparameters of the ML models employed. B1 and B2 denote classes 1 and 2 of dataset B, respectively.*

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
Logistic Regression	63.98		B1	B2	25	Penalty: l1, C: 1.0
		B1	48	22		
		B2	36	55		
Naive Bayes	63.98		B1	B2	35	GaussianNB
		B1	50	20		
		B2	38	53		
SVM	61.49		B1	B2	35	C: 6, kernel: linear
		B1	46	24		
		B2	38	53		
KNN	57.76		B1	B2	15	leaf_size: 1, n_neighbors: 16, weights: uniform
		B1	63	7		
		B2	61	30		
Decision Tree	58.39		B1	B2	15	max_features: auto, min_samples_leaf: 1, min_samples_split: 6
		B1	41	29		
		B2	38	53		
Random Forest	62.11		B1	B2	15	criterion: gini, min_samples_leaf: 2, min_samples_split: 7, n_estimators: 30
		B1	48	22		
		B2	39	52		
XGboost	60.25		B1	B2	40	gamma: 0.4, max_depth: 7, min_child_weight: 5
		B1	44	26		
		B2	38	53		

- Dataset C

Less informative features with small generalization capacity are contained in dataset C, as reported in Table 5. 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.

Table 5. *Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features, and the hyperparameters of the ML models employed. C1 and C2 denote classes 1 and 2 of dataset C, respectively.*

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
Logistic Regression	61.11		C1	C2	35	Penalty: l1, C: 1.0
		C1	49	15		
		C2	41	39		
Naive Bayes	59.03		C1	C2	160	GaussianNB
		C1	23	41		
		C2	18	62		
SVM	65.28		C1	C2	65	C: 5, kernel: rbf
		C1	48	16		
		C2	34	46		
KNN	61.11		C1	C2	120	leaf_size: 1, n_neighbors: 5, weights: uniform
		C1	55	9		
		C2	47	33		
Decision Tree	66.67		C1	C2	225	max_features: auto, min_samples_leaf: 2, min_samples_split: 8
		C1	44	20		
		C2	28	52		
Random Forest	59.72		C1	C2	140	criterion: gini, min_samples_leaf': 1, min_samples_split: 5, n_estimators: 25
		C1	37	27		
		C2	31	49		
XGboost	62.5		C1	C2	150	n_estimators = 100, max_depth = 8, learning_rate = 0.1, subsample = 0.5
		C1	44	20		
		C2	34	46		

- 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 Table 6: (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). This performance was the overall best one achieved in all five datasets. (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%.

Table 6. *Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features and the hyperparameters of the ML models employed. D1 and D2 denote classes 1 and 2 of dataset D, respectively.*

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
Logistic Regression	72.84		D1	D2	55	Penalty: l1, C: 1.0
		D1	54	27		
		D2	17	64		
Naive Bayes	68.52		D1	D2	20	GaussianNB
		D1	44	37		
		D2	14	67		
SVM	74.07		D1	D2	55	C: 0.1, kernel: linear
		D1	56	25		
		D2	17	64		
KNN	71.6		D1	D2	75	algorithm: auto, leaf_size: 1, n_neighbors: 17, weights: uniform
		D1	55	26		
		D2	20	61		
Decision Tree	61.73		D1	D2	30	max_features: auto, min_samples_leaf: 3, min_samples_split: 10
		D1	56	25		
		C2	37	44		
Random Forest	66.67		D1	D2	20	criterion: gini, min_samples_leaf: 3, min_samples_split: 3, n_estimators: 25
		D1	47	34		
		D2	20	61		
XGboost	64.81		D1	D2	15	gamma: 0.6, max_depth: 1, min_child_weight: 8
		D1	51	30		
		D2	27	54		

- Dataset E

In dataset E, the SVM-based approach exhibited a maximum of 71.81% at 70 features (which was the best in the category). Similar to SVM, LR gave the second-highest accuracy (71.14%) for fewer 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 the ML models (Table 7).

Table 7. *Best testing accuracies achieved for each ML model along with the confusion matrix, the optimum number of features, and the hyperparameters of the ML models employed. E1 and E2 denote classes 1 and 2 of dataset E, respectively.*

Models	Accuracy (%)	Confusion Matrix			Features	Parameters
Logistic Regression	71.14		E1	E2	55	Penalty: l1, C: 1.0
		E1	50	17		
		E2	26	56		
Naive Bayes	68.46		E1	E2	230	GaussianNB
		E1	48	19		
		E2	28	54		
SVM	71.81		E1	E2	70	C: 1, kernel: sigmoid
		E1	50	17		
		E2	25	57		
KNN	63.76		E1	E2	20	algorithm: auto, leaf_size: 1, n_neighbors: 16, weights: uniform
		E1	48	19		
		E2	35	47		
Decision Tree	66.44		E1	E2	95	max_features: auto, min_samples_leaf: 2, min_samples_split: 12
		E1	45	22		
		E2	28	54		
Random Forest	67.11		E1	E2	55	criterion: gini, min_samples_leaf: 1, min_samples_split: 3, n_estimators: 30
		E1	42	25		
		E2	24	58		
Xgboost	70.47		E1	E2	45	gamma: 0.6, max_depth: 2, min_child_weight: 1
		E1	43	24		
		E2	20	62		

Table 8 cites the best accuracies achieved in each of the five datasets. The combined effect of baseline features (dataset A) and progression data (dataset B) had a positive effect on the prediction capacity of the proposed methodology, as clearly shown in Table 7 where the testing accuracy in dataset D is increased by 2.36% compared to the result obtained in dataset A. A minor difference (0.1%) is observed on the accuracies reported for datasets A and E, demonstrating that progression data have a negligible effect on the predictive capacity of the proposed methodology and therefore could be omitted. The accuracies received in datasets B and C reveal that the baseline features are crucial for predicting KOA progression.

Table 8. *Summary of all reported results.*

Dataset	Data Used in the Training			Best Testing Performance (%) Achieved	Best Model	Num. of Selected Features
	Baseline	M12 Progress Wrt Baseline	M24 Progress Wrt Baseline			
A	•			71.71	Logistic Regression	50
B		•		63.98	Logistic Regression	25
C			•	66.67	Decision Tree	225

D	•	•		74.07	SVM	55
E	•		•	71.81	SVM	70

Discussion

This work focuses on the development of an 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 three different options with respect to the time 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 (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 results 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.

A few studies have recently addressed the problem of predicting KOA progression from different perspectives and employing different data sources. A weighted neighbor distance classifier was presented by Ashinsky et al. to classify isolated T2 maps for the progression to symptomatic OA with 75% accuracy. Progression to clinical OA was defined by the development of symptoms as quantified by the WOMAC questionnaire 3 years after baseline evaluation. MRI images and PCA were employed by Du et al. to predict the progression of KOA using four ML techniques. For KL grade prediction, the best performance was achieved by ANN with AUC = 0.761 and F-measure = 0.714. An MRI-based ML methodology has been also proposed by Marques et al. to prognose tibial cartilage loss via quantification of tibia trabecular bone where an odds ratio of 3.9 (95% confidence interval: 2.4–6.5) was achieved. X-ray combined with pain scores have been utilized by Halilaj et al. to predict the progression of joint space narrowing (AUC = 0.86 using data from two visits spanning a year) and pain (AUC = 0.95 using data from a single visit). Similarly, another two studies (Tiulpin et al. and Widera et al.) made use of X-ray images along with clinical data to predict KOA progression using either CNN or ML approaches achieving less accurate results. The current work is the only one employing exclusively clinical non-imaging data and also contributes to the identification of important risk factors from a big pool of available features. The proposed methodology achieved comparable results with studies predicting KL grades progression demonstrating its uniqueness in facilitating the prognosis of KOA progression with a less complicated ML methodology (without the need for big imaging data and image-based deep learning networks).

Among the limitations of the current study is the relatively large number of features (55) that were finally selected as possible predictors of KOA. The selected features come from almost all feature categories highlighting the necessity of adopting a rigorous data collection process to formulate the input feature vector that is needed for the ML training. Moreover, the ML models employed are opaque (black boxes) and therefore they are insufficient to provide explanations on the decisions (inability to explain how a certain output has been drawn). To overcome the aforementioned challenges, it is important for AI developers to build transparency into their algorithms and/or enhance the explainability of existing ML or DL networks.

4.2 Personalised Prediction of Pain progression

Figure 9 demonstrates 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. Table 9 shows the confusion matrices of the best performing model RF. The rest of the ML models achieved inferior results, with SVM producing the second-best results with an 80.83% accuracy score. Overall, as we add more features to the aforementioned models, we observe that their accuracy scores decrease.

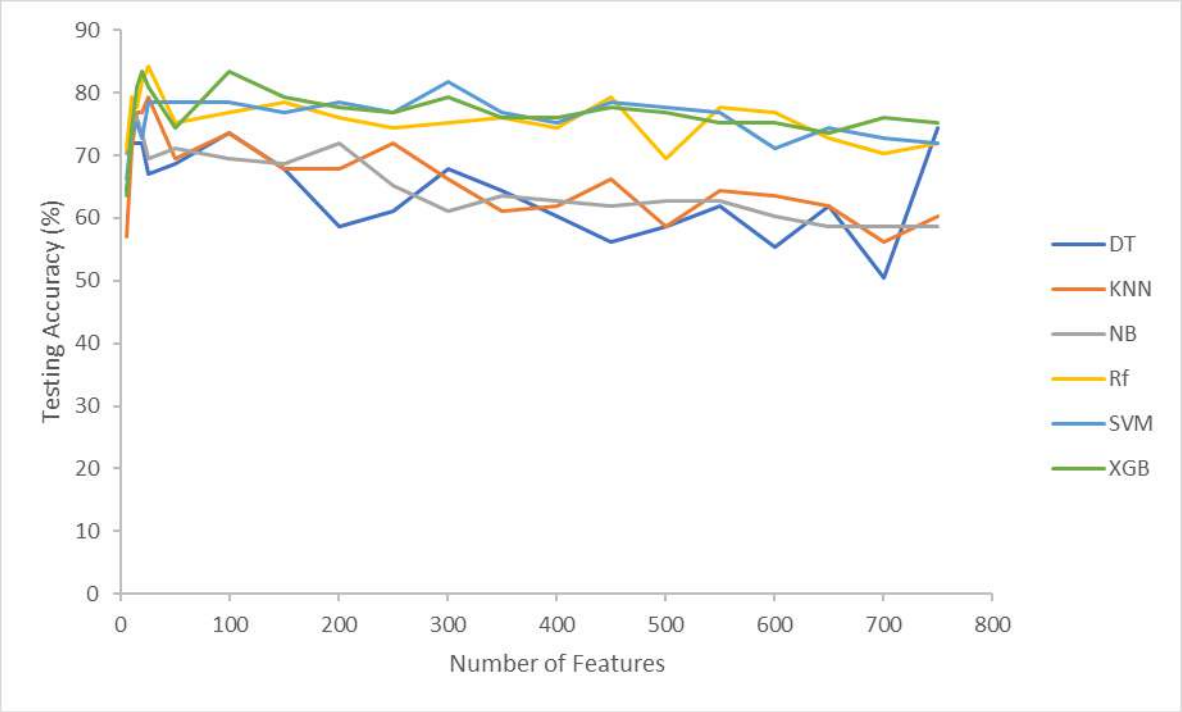


Figure 9. *Left leg: features and model accuracy scores (%)*

Table 9. *Random forest: confusion matrix (Left Leg)*

	Class 1	**Class 2**	**Class 3**	**Per class accuracy**
Class 1	28	3	12	65.12%
Class 2	0	33	0	100%
Class 3	7	0	38	79.17%

For the right leg, Figure 10 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 extra features has produced inferior results for our prediction. Table 10 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 Figure 11, similar results are obtained on both legs; indicating the repeatability and robustness of the proposed methodology.

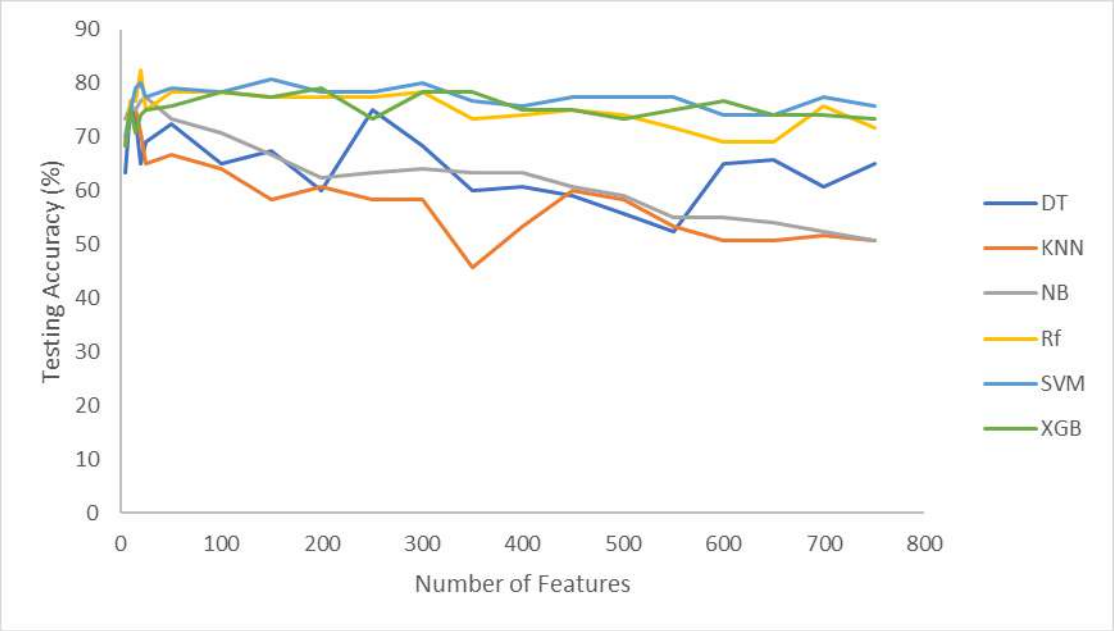


Figure 10. *Right leg: features and model accuracy scores (%)*

Table 10. *Random forest: confusion matrix (Right Leg).*

	Class 1	Class 2	Class 3	Per class accuracy
Class 1	28	5	7	70%
Class 2	0	39	0	100%
Class 3	8	1	32	78%

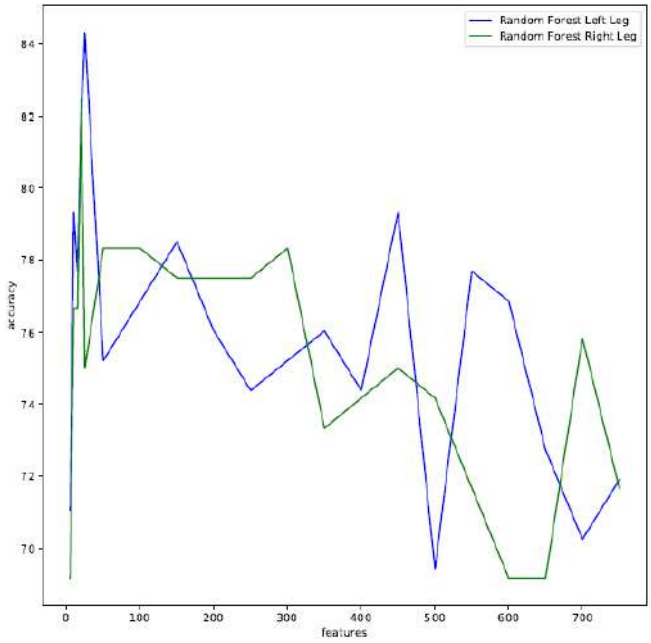


Figure 11. *The comparison of the performance of Random Forest on both legs in accordance to the number of features.*

Discussion of results

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 D6.3 and D6.5 that the most important features for the pain progression prediction are related directly to 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.

4.3 Personalised Prediction of JSN progression

Evaluation Methodology

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 model's 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 (Deliverable 6.5).

Post-Hoc Interpretation/Explainability

In this work, the SHapley Additive exPlanations (SHAP) were employed to rank features in terms of their impact on the final ML outputs. SHAP builds a mini explainer model for a single row-prediction pair that explains how this prediction was achieved. It is based on optimal shapley values from coalitional game theory that indicate how to fairly distribute the impact on the model's prediction among the features.

Classification Results

Table 11 shows the maximum, minimum, and mean accuracy along with the standard deviation achieved by the models over the test set for increasing the number of features for the left leg. For the left leg, the LR model performed better than the others with a maximum accuracy of around 77.7% for 165 features (Table 11). However, NNs and SVM had a comparative performance with a 75.8% and 76.4% maximum accuracy, respectively. To identify the exact number of features where the prediction accuracy is maximized, the two best performing models (LR and SVMs) were tested in the range of 155 and 175 features with a step of 1 with the results shown in Figure 12. The LR model performed best ($\cong$ 78.3%) at 164 features. For this performance the following hyperparameters were used: Maximum number of iterations 100, intercept scale 1, L2 penalty, Newton-cg solver with reuse of the previous solution as initial one, and tolerance 0.0001.

Table 11. *Maximum, minimum, and mean accuracy of prediction models over the tested set for the left leg. The best results are indicated in bold.*

Prediction Model	Maximum Accuracy	Minimum Accuracy	Mean Accuracy	Standard Deviation
Gradient Boosting	0.72611	0.56688	0.66707	0.02622
Logistic Regression	**0.77707**	0.60510	**0.71540**	0.03353
NNs (Neural Networks)	0.75796	0.62420	0.68234	0.02933

Naïve Bayes Gaussian	0.68153	0.59236	0.62794	0.02301
Random Forest	0.70064	0.61783	0.65989	0.01616
SVM	0.76433	0.63057	0.70377	0.02783

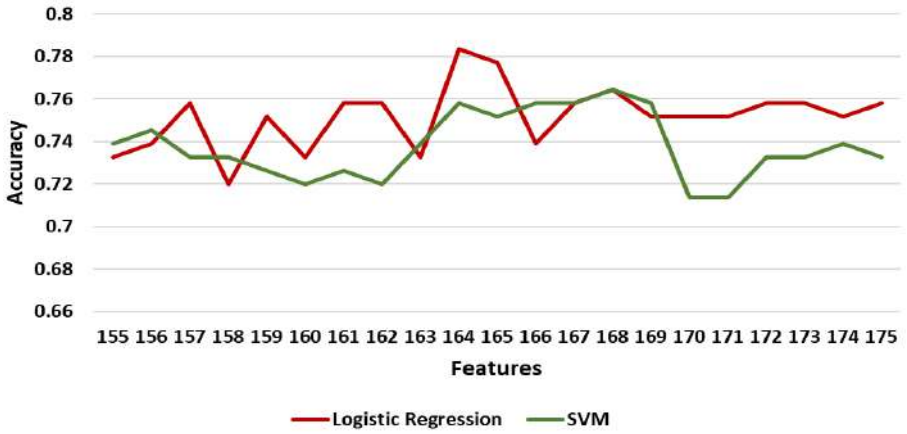


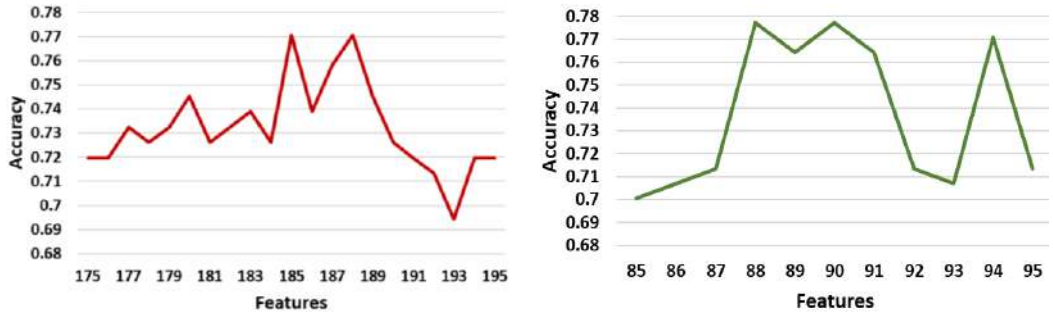
Figure 12. *The accuracy of LR (logistic regression) and SVM at [155, 175] features over the test set for the left leg. Results are shown with a step size of 1 (one feature added at each step).*

For the right leg a similar approach was adopted. Table 12 shows the maximum, minimum, and mean accuracy with the standard deviation achieved by the models over the test set for increasing the number of features for the right leg.

The SVM model presented the best performance by achieving the maximum accuracy $(\cong 77.7\%)$ for 90 features (Table 12). However, the LR and NNs models accomplished an adequate performance (Figure 13). Specifically, the LR model achieved a higher mean accuracy $(\cong 70.7\% \pm 0.036)$ with a lower standard deviation compared to the results of the model $(\cong 68.6\% \pm 0.039)$. To this end, these two models were re-evaluated for features in the neighborhoods, $\mathcal{U}_{LR}(185,10)$ and $\mathcal{U}_{SVM}(90,5)$ with a step of 1 feature at a time. LR achieved its best performance $(\cong 77.1\%$ accuracy$)$ at 185 and 188 features while the SVM model reached its maximum accuracy of 77.7% with 88 and 90 features (Figure 13). The SVM model's hyperparameters that achieved the best performance are the following: A linear kernel, regularization parameter at 0.1, tolerance at 0.001, and cache size at 200.

Table 12. *Maximum, minimum, and mean accuracy of prediction models over the tested set for the right leg. The best results are indicated in bold.*

Prediction Model	Maximum Accuracy	Minimum Accuracy	Mean Accuracy	Standard Deviation
Gradient Boosting	0.72611	0.61783	0.67172	0.02445
Logistic Regression	0.77070	0.63057	**0.70691**	0.03560
NNs	0.76433	0.58599	0.69983	0.03858
Naïve Bayes Gaussian	0.72611	0.50955	0.62774	0.03926
Random Forest	0.71975	0.61783	0.67577	0.02217
SVM	**0.77707**	0.60510	0.68598	0.03929



(**a**) (**b**)

Figure 13. *The performance evaluation of (**a**) LR in the range of 175–195 features and (**b**) SVM in the range of 85–95 features. Results are shown with a step size of 1 (one feature added at each step).*

Table 13 shows the maximum, minimum, and mean accuracy achieved among with the standard deviation by the models over the test set for various number of features for both right and left legs combined. The results show that the LR model performed better compared to the other models by reaching the maximum accuracy $(\cong 83.3\%)$ for 30 features, as illustrated in Table 13. Nevertheless, SVM and RF showed a comparative performance. The aforementioned three models are re-evaluated in order to find the number of features that maximize the accuracy. Hence, the models are tested in the neighborhood where all three models achieved their best performance $\mathcal{U}(30,5)$. From a more detailed analysis, LR remained the predictive model with the best performance $(\cong 83.3\%)$ for 29 features (Figure 14).

Overall, LR presented a stable performance (Figure 15, $77\% \pm 0.04$) reaching the maximum accuracy at 29 features (83.3%). The hyperparameters of the LR model with the best performance were identical to the ones presented for the case of the first strategy. A generalized linear model such as LR has accomplished the best performance in our study, indicating that the power of the proposed methodology is not so much dependent on the complexity of the learning model but lies on the effective and robust mechanism of selecting important risk factors. Identifying robust predictive risk factors from a high dimensional feature space (such as the OAI dataset) is crucial since it enhances our understanding of KOA progression and therefore contributes to the development of robust prediction tools.

Table 13. *Maximum, minimum, and mean accuracy of prediction models over the tested set for the left and right legs combined. The best results are indicated with bold.*

Prediction Model	Maximum Accuracy	Minimum Accuracy	Mean Accuracy	Standard Deviation
Gradient Boosting	0.81746	0.69841	0.74591	0.02449
Logistic Regression	**0.83333**	0.65873	**0.76503**	0.03725
NNs	0.79365	0.64286	0.73870	0.03470
Naïve Bayes Gaussian	0.76984	0.50000	0.63300	0.06331
Random Forest	0.79365	0.61111	0.70755	0.05645
SVM	0.82540	0.64286	0.74928	0.04223

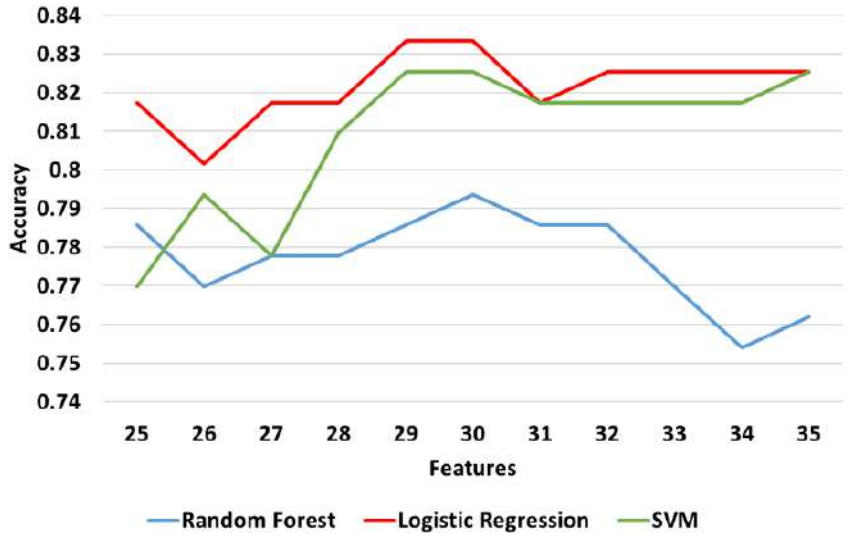


Figure 14. *The accuracy of LR, RF (random forest), and SVM from 25 to 35 features over the test set for left and right legs combined. Results are shown with a step size of 1 (one feature added at each step).*

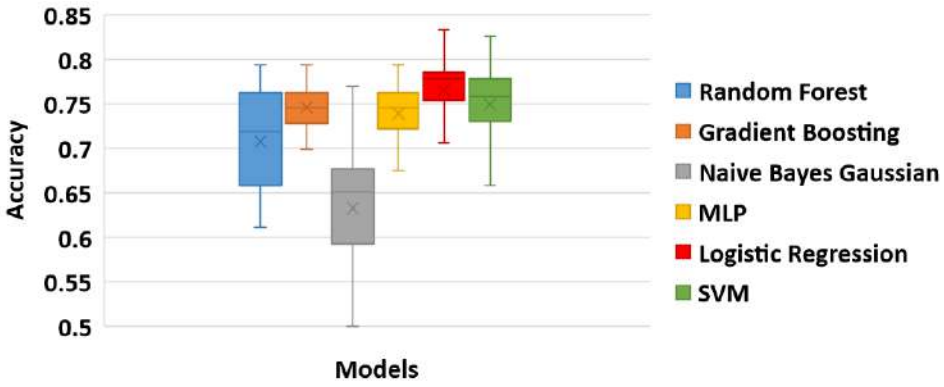


Figure 15. *The box plot of the prediction models based on their performance for the right and left legs combined.*

The strategy of this work takes into account information from both legs and therefore leads to a well-defined data classification problem in which the non-progressors do not experience any JSN progression in any of their legs, whereas the progressors' class incudes data from patients that experience JSN progression in at least one of their legs or both. This data problem proved to be more effectively handled by the proposed methodology with an 83.33% prediction accuracy at the first 29 features.

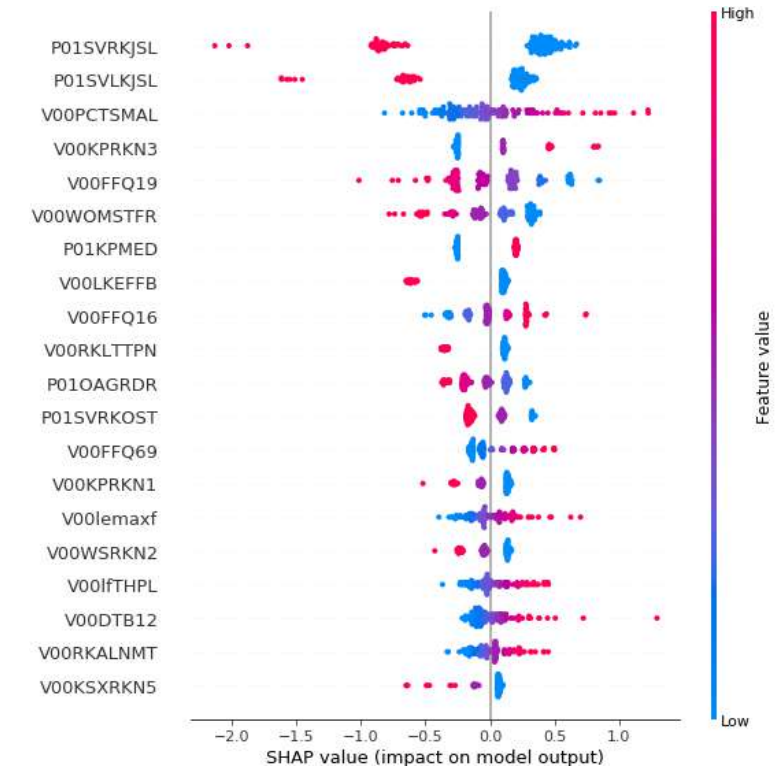


Figure 16. *The distribution of the features' impact on LR model output for the OAI (osteoarthritis initiative) dataset with 29 features across all instances.*

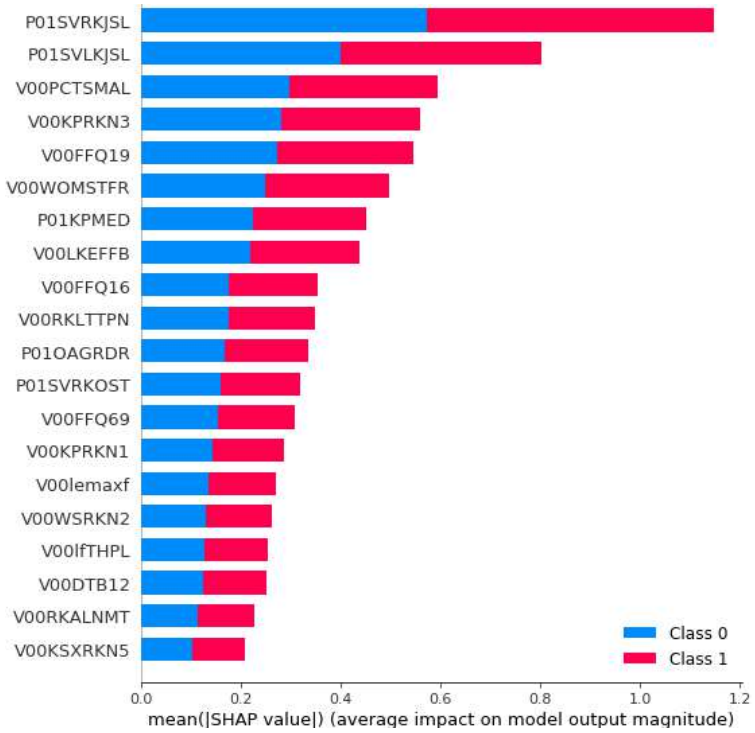


Figure 17. *The average impact magnitude of 29 features on the LR model output for the OAI dataset for all instances.*

Post-Hoc Explainability Results

To explain the impact of the selected features on the outcomes of the employed best prediction model, the SHAP method was used. SHAP was applied to the LR model which was trained on the selected 29 features

that come from both legs. In Figure 16, the features are sorted by the sum of SHAP value magnitudes over all samples. The SHAP values are used to indicate the distribution of each feature's impact on the model's output. Specifically, the feature value is represented by color, with the red color corresponding to a high impact while the blue to a low impact. For instance, a high P01SVRKJSL value (evidence of knee lateral joint space narrowing) lowers the predicted status of the subjects. The features P01SVLKJSL, V00FFQ19, V00WOMSTFR, V00LFEFFB, V00RKLTTPN, P01OAGRDR, P01SVRKOST, V00KPRKN1, V00WSRKN2, and V00KSXRKN5 present similar behavior. On the contrary, V00FFQ16 (how often the patient ate dishes with rice in the past 12 months) has a positive effect on the prediction outcome. Similar behavior was identified for the features V00PCTSMAL, V00KPRKN3, P01KPMED, V00FFQ69, V00lemaxf, V00lfTHPL, V00DTB12, and V00RKALNMT. Figure 17 illustrates the mean absolute value of SHAP values for each feature as a standard bar plot, which depicts the SHAP global feature importance. We observe that each feature has the same impact on both classes. Furthermore, the most important features that affected the prediction output were the P01SVRKJSL, P01SVLKJSL, and V00PCTSMAL.

4.4 Increasing generalization using an evolutionary Machine Learning approach

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.

Explainability

To further assess the impact of the selected features on the classification outcome, SHapley Additive exPlanations (SHAP) were considered. SHAP is a game-theoretic approach that explains the output of any machine learning model and achieves the connection of the optimal credit allocation with local explanations using the classic Shapley values from game theory that come with desirable properties. In this study, Kernel SHAP is used, which is a specially weighted local linear regression to estimate SHAP values for any model (e.g., SVM in a two-class classification problem). The optimization of loss function L in Kernel SHAP is described below (in Equation), where g is the explanation linear model that is trained on training data Z, $f(:)$ is the original prediction function to be explained and z' is a vector of 1s and 0s called a coalition. Here, 1s indicate the presence of the corresponding feature, while 0 indicates its absence. $h_x(z')$ maps a feature coalition to a feature set on which the model can be evaluated, whereas $\pi_x(z')$ is the SHAP kernel.

$$(f, g, \pi_x) = \sum_{z' \in Z} \left[f\left(h_x(z')\right) - g(z') \right]^2 \pi_x(z')$$

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, whereas their impact on the classification result is discussed employing SHAP.

Table 14 also shows the best accuracies achieved by each technique and the number of features for which the best accuracy was achieved. GenWrapper achieved its best accuracy at a relatively small number of features (35), whereas the rest had inferior performances and, in most of the cases, at a higher number of features. The classical wrapper FS was the only one that selected slightly fewer features (31). A statistical comparison was finally conducted, verifying that the accuracies obtained by the proposed GenWrapper were significantly different (higher) to the ones of all the competing FS algorithms ($p < 0.001$).

Table 14. *Best performance (mean 10FCV) was achieved by all competing FS techniques employing SVM along with the number of selected features in which this accuracy was accomplished.*

Approach	Best Accuracy (Mean 10FCV)	Number of Features	Statistical Comparison *	Execution Time (sec) **
GenWrapper	71.25	35	-	311.6
Wrapper	69.79	31	$p < 0.001$	10.2
CFS	61.97	69	$p < 0.001$	0.1
ILFS	63.63	82	$p < 0.001$	0.5
Inf-FS	63.32	35	$p < 0.001$	0.1
Lasso	64.41	94	$p < 0.001$	21.2
Mrmr	67.29	36	$p < 0.001$	2.3
Hybrid	67.85	41	$p < 0.001$	15.5
PCA	65.11	29	$p < 0.001$	<0.1

* Statistical comparison with the proposed GenWrapper. ** All the algorithms were executed on an Intel Core i7-7500 processor, 2.70 GHz CPU (16 GB RAM) using MATLAB 2020b.

The last part of the conducted comparative analysis focuses on a different performance metric—that is, the consistency of the obtained accuracies during the proposed repetitive validation process. As explained in the previous sections, the predictive capacity of the selected features is validated multiple times (10). In each of the ten repetitions, 10FCV is employed on a different, randomly selected balanced data sample. A feature subset could be considered robust when it consistently leads to high accuracies over the ten repetitions. Figure 18 is a bar graph that visualizes (i) the mean 10FCV accuracies, (ii) the standard deviation of the 10FCV accuracies, (iii) the range ([min, max]) of the 10FCV accuracies, and (iv) any outliers that deviate from the distribution of the 10FCV accuracies. GenWrapper was the most accurate approach (71.25%) and, at the same time, it proved to be the most consistent FS technique, with the great majority of obtained 10FCV accuracies being higher than 70%. The classical wrapper FS was also consistent over the ten repetitions but it was considerably less effective than the proposed GenWrapper. It should be noted that the hybrid FS approach achieved accuracies up to 72.5%; however, it does not generalize well given that it leads to a quite enlarged min–max range as well as an increased standard deviation, with the minimum accuracy being less than 60%. Mrmr has led to both moderate mean accuracy and moderate consistency (ranging between 66% and 70%) over the repetitions of the employed validation process. The rest of the competing FS approaches led to much lower 10FCV accuracies that ranged between 58% and 68%.

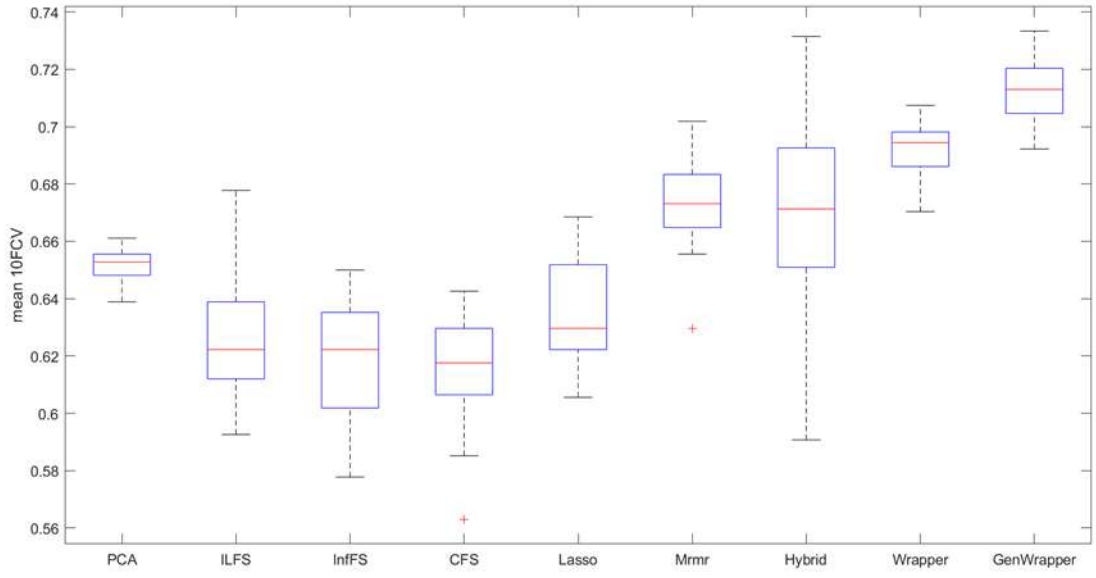


Figure 18. *Bar graph comparison for the best models (SVMs trained on the optimum number of selected features per case). Red lines correspond to the mean 10FCV, blue boxes visualize the standard deviation of the obtained accuracies, dashed black lines show the min–max range and the red crosses depict outliers (if any).*

Explainability Results

Figure 19a illustrates the features' impact on the output of the final model (SVM) on the OAI dataset. It sorts features by the sum of SHAP value magnitudes over all samples and uses SHAP values to show the contribution of each feature (positive or negative) on the model's output. The color represents the feature value (blue—low; red—high). This reveals, for example, that a high P01BMI (body mass index of the participants) increases the predicted status of the participants. Similar to BMI, the features P01SVLKOST, V00SUPCA, V00CHNFQCV, V00WOMSTFR, V00FFQSZ13, V00KQOL4, V00rkdefcv, KPLKN1, and V00PA130CV have a positive effect on the prediction outcome (their increase drives the output to increase), whereas the rest have the opposite effect. Figure 19b demonstrates the mean absolute value of

the SHAP values which represents the SHAP global feature importance. It should be noted that the features P01SVLKOST, BMI, V00SUPCA, and V00EDCV were the most important variables that significantly affected the prediction output (Appendix A).

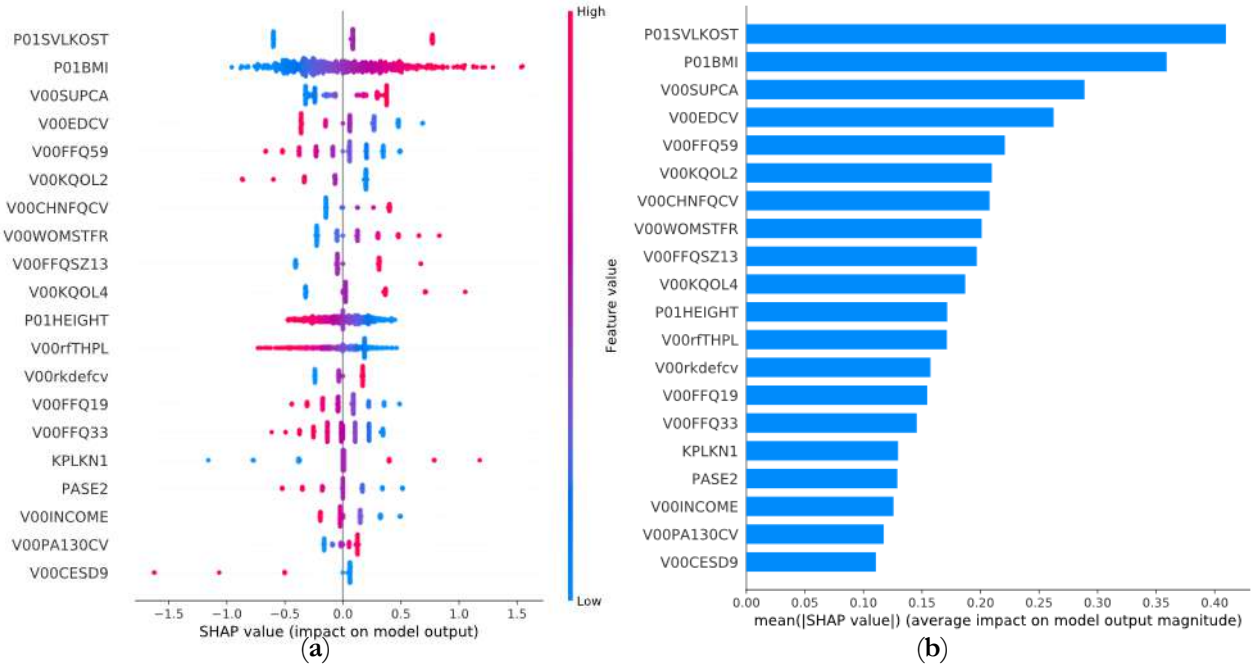


Figure 19. This figure depicts: (a) the SHAP summary plot and; (b) the SHAP feature importance for the SVM trained on the features selected by the proposed GenWrapper.

Discussion of results

During our work, we utilized multimodal data and we managed to identify the variables that mainly contributed to the predictive ability of our models. Important predictive risk factors selected by our models included assessments of pain and function, qualitative assessments of X-rays, assessments of behavioral characteristics, medical history, and nutrition from the Center for Epidemiologic Studies Depression Scale (CES-D) and Block Brief 2000 questionnaires. The strongest indicator variables are reporting on knee baseline radiographic OA status (P01SVLKOST), anthropometric characteristics (P01BMI), and on nutritional (V00SUPCA) and behavioral habits (V00KQOL4). Previous studies have also reported similar key predicted variables for KOA progression. Our findings suggest that early functional, behavioral and nutritional interventions should be encouraged and implemented for the prevention or slowing down of KOA progression.

Genetic algorithms might be costly in computational terms since the evaluation of each individual requires the training of a model. Due to its stochastic nature, the proposed FS takes a longer time to converge, and this could be considered a limitation. However, the identification of risk factors for KOA progression is, in principle, an offline approach, and therefore, its current execution time (~5 min) is not prohibitive. In the current study, time execution is not considered as crucial as the predictive capability of the finally selected features that can be used to enhance our understanding of whether a patient is at increased risk of progressive KOA. GenWrapper improves the current state of the art by identifying risk factors that are more accurate compared to the ones selected by eight well-known FS algorithms (by at least 3.4%) and, most importantly, more robust in terms of their performance on the entire population of subjects (as it has been validated with an extensive validation mechanism that involved 100 training runs on different data samples). This stated improvement could (i) allow preventive actions to be planned and implemented and

(ii) enable more personalized treatment pathways and interventions for treatment, targeting specific risk factors. From a different perspective, being able to identify non-progressors could also prevent over-investigations and over-treatment.

4.5 Diagnosis of KOA based on KL grade

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.

Interpretation / Explainability

SHapley Additive exPlanations (SHAP) is a game-theoretic approach to explain the output of any ML model (e.g., XGBoost, LightGBM, CatBoost, scikit-learn, and pyspark tree models). It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. In this work, we employed SHAP to rank features in terms of their impact on the final ML outputs and to build a mini explainer model for a single row-prediction pair that explains how this prediction was achieved.

Results

Table 15 summarizes the results of Logistic Regression, XGboost, SVM, Random Forest, KNN, Naïve Bayes, and DT on the 2-class problem. The best overall performance on the 2-class problem (77.71%) was achieved by the Logistic Regression on the group of the forty selected (40) risk factors with L2 penalty and C = 1.0. The second highest accuracy was received for XGboost (77.31%), whereas lower accuracies were obtained by the rest of the models.

Table 15. *Best testing accuracies achieved for each model along with the selected number of features and the hyperparameters of the ML model.*

Classifiers	Accuracy	Num. of Features	Selected hyperparameters
Logistic Regression	77.71%	40	Penalty: l2, C: 1.0
XGboost	77.31%	40	gamma: 0, max_depth: 2, min_child_weight: 8
SVM	76.93%	70	C: 8, kernel: linear
Random Forest	76.50%	35	criterion: entropy, min_samples_leaf: 2, min_samples_split: 6, n_estimators: 30
KNN	73.58%	10	leaf_size: 1, n_jobs: -1, n_neighbors: 15
Naïve Bayes	72.37%	45	GaussianNB

Decision Tree	72.63%	5	max_features: auto, min_samples_leaf: 8, min_samples_split: 2

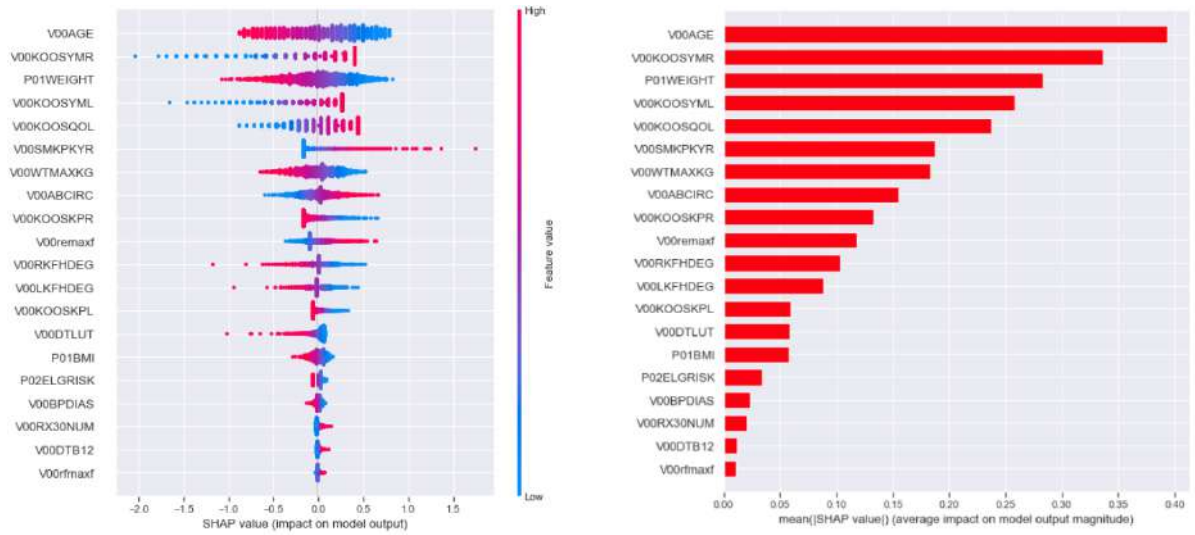


Figure 20. *Features' impact on Logistic Regression (40F) model output for the OAI dataset. The left panel, shows the distribution of the impact of a feature value on the model output across all instances. The right panel, shows the average impact magnitude for all instances.*

Figure 20 illustrates the features' impact on the output of the final model (Logistic Regression 40F) on the OAI dataset. For the certain subset, the left panel sorts feature by the sum of SHAP value magnitudes over all samples and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red for high, blue for low). This reveals for example that a high V00AGE (age of the participants) lowers the predicted status of the participants. Furthermore, from the right panel, we take the mean absolute value of the SHAP values for each feature to get a standard bar plot. In addition, it should be noted that in both panels the features are ordered by their total impact.

Discussion of results

It was also observed that there is a partial agreement between the order of the features' importance as selected by the proposed FS methodology and the features' impact on the prediction outcomes. The most significant difference between the outcomes of the FS and explainability analysis was observed in the variables VOOAGE, P01KSX, and P01BMI that were not selected on the first 10 features but were proven to contribute significantly shaping the KOA predictions (Figure 20). The likely reason for the aforementioned observation is that the FS and the explainability of the ML models perform different tasks. Specifically, FS selects a subset of relevant features (variables, predictors) for use in model construction and to improve prediction performance. On the contrary, the SHAP assigns on each feature an important value for a particular prediction and has the main task to explain the output of any ML model. For this reason, there is this mismatch in the order of importance of the features that resulted from FS and the use of SHAP. According to SHAP, these three variables contribute to the interpretation of predictions, which is in line with the existing literature. Age is an important factor in the occurrence of KOA, as evidenced by Moustakidis et al. in developing a model for diagnosing of KOA. Also, the presence of knee pain is a factor that leads to the diagnosis of a person with KOA. In addition, obesity (high BMI) is suggested to be a high-risk factor in the development of KOA due to the increased mechanical loading that is applied on the knee

joints. Furthermore, observing the processes for individual predictions, it was concluded that each group of subjects was reinforced by different features. This finding further indicates the need of applying explainability analysis algorithms on the generated ML models in order to enhance our understanding of them and identify the key elements that contribute and shape their predictions.

5. Machine Learning and Deep Learning Diagnosis models with focus on accuracy and fairness

Osteoarthritis dataset

In this research work, the dataset was selected from the osteoarthritis initiative (OAI) database designed to identify risk factors associated with the incidence and progression of knee OA. Osteoarthritis initiative study (https://oai.epi-ucsf.org/datarelease/) was launched in 2002, enrolling people aged 45–79 years, with symptomatic knee OA 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 8 years with a follow-up rate of more than 90% over the first 48 months. The current work only includes self-reported data related to joint symptoms, disability, function and general health from all individuals with or without KOA from the baseline visit.

Dataset characteristics

The selected dataset comprises 141 risk factors from 4796 participants. Next, we divide this dataset into six subgroups of participants. These subgroups are (i) participants older than 70 years, (ii) participants under 70 years, (iii) male participants, (iv) female participants, (v) non-obese, and (vi) obese participants.

Table 16. *Dataset characteristics*

	Category	Num. of features	Feature category	Description
Feature characteristics	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 of a score of questions about pain, symptoms, and quality of life for both knees
	5	KOOS	Indexes which consist of a score of questions about pain, stiffness, and disability for both knees

Sample characteristics

Groups		**Total number of samples**	**Samples in progression class**	**Samples in incidence class**
Weight	Obese	1761	681	1080
	Non-obese	2909	706	2203
Age	Over 70	1119	329	790
	Under 70	3559	1063	2496
Gender	Males	1945	597	1348
	Females	2729	793	1936

The 68 out of 141 features describe any type of symptoms over the past 7 days, such as any back pain, how often bothered by back pain, limited activities due to back pain, the number of days stayed in bed due to back pain, etc. Ten (10) out of 141 features describe any type of the same symptoms over the past 30 days; 13 out of 141 features, any type of the same symptoms over the past 12 months. Next, 64 out of 141 features are related to pain in various activities for both knees, hips, and joints in all time intervals, 27 out of 141 features are related to stiffness in all the time intervals, 37 out of 141 features are related to the knee difficulty on either right or left leg on various activities in all time intervals, 12 out of 141 are symptoms such as swelling, grinding sensation, knee catch or hang up in all time intervals, 15 out of 141 features are related to health, emotional problems, lifestyle, psychology, 8 are indexes which consist of a score of questions about pain, symptoms, and quality of life for both of knees, and 5 are indexes which consist of a score of questions about pain, stiffness, and disability for both knees.

The 4796 samples of the dataset were divided into two categories as follows:

- *Class 1*: Incidence: This class comprises 3284 participants who do not have symptomatic knee OA, 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".

Control samples or samples with missing data and outliers were excluded from the datasets of the current study.

To evaluate the predictive performance of the proposed methodology on different populations, the dataset was organized into the following subgroups with respect to Body Mass Index (BMI), age, and gender:

1) *Obese subgroup* consisting of subjects with BMI higher or equal to 30

2) *non-obese subgroup* with BMI<30
3) *over 70 subgroup (aging)* consisting of subjects that are more than 70 years old
4) *under 70 subgroup* with subjects younger than 70 years
5) *male subgroup,* and the
6) *female subgroup.*

The dataset characteristics (including a description of features and the number of samples per subgroup and per class) are presented in Table 16.

Methodology

The proposed DNN-based method for KOA classification includes three processing steps: data pre-processing to handle missing values and normalise the collected clinical data, a learning process for DNN training, and evaluation of the classification results. The proposed methodology is presented below.

Preprocessing

For handling missing values, mean imputation was performed. Specifically, for numerical features, missing values were replaced by the mean feature value. In the 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. Specifically, a variant of SMOTE (SMOTE-SVM) was utilized providing borderline over-sampling especially designed for imbalanced data classification problems. In SMOTE-SVM, a borderline area is approximated by the support vectors obtained after training a standard SVMs classifier on the original training set. New instances are then randomly created along the lines joining each minority class support vector with a number of its nearest neighbors using interpolation.

Dense Neural Networks

A DNN is actually a fully connected ANN. Concerning the learning process, DNNs use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation as well as can learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners. A DNN consists of a series of fully connected layers. A fully connected layer is a function from $\mathcal{R}^m$ to $\mathcal{R}^n$. Let $x \in \mathcal{R}^m$ represent the input to a fully connected layer. Let $y_j \in \mathcal{R}$ be the *j*-th output from the fully connected layer. Then y_j is computed as follows: $y_j = g\left(\sum_{i=1,..m} w_{ij} x_i\right)$ where g is a predefined function known as the activation function and w_{ij} are learnable parameters in the network. This transformation is iterated from layer to layer until we reach the final layer where a Softmax function is applied. For this work, we used H2O that is an open-source library widely used for constructing and learning DNNs in prediction and classification tasks. The design space of a DNN is practically infinite severely depending on the number of layers of the DNN and the number of neurons in each of those layers.

Due to the limited available computational power, the size of a DNN needs to be adjusted according to each problem's characteristics. In this study, we used fully connected, dense neural layers where the output of one layer serves as the input for the next layer. We investigated several different DNN architectures with varying: (i) number of hidden layers, (ii) number of nodes per hidden layer. The 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. The final neural layer reduces the dimensionality

to 2 nodes using 'Softmax' as an activation function. The adaptive learning rate was employed with ADADELTA that automatically combines the benefits of learning rate annealing and momentum training to avoid slow convergence. Weight initialisation was performed by using a uniform distribution. Early stopping was implemented based on the convergence of the logloss metric.

Validation

The performance of the proposed methodology was validated in terms of both accuracy and fairness. Accuracy was estimated using a 70% (training) - 30% (testing) split of the dataset. The proposed methodology was trained and optimised using the training set and the final predictive performance was estimated as the accuracy on the testing set. Fairness was calculated by employing the metrics that are presented below.

Definition 1 (*Demographic Parity*) is also known as statistical parity. A predictor satisfies demographic parity if the likelihood of a positive outcome is the same regardless of whether the person is in the protected (e.g., female) group.

$$DP\ (\%) = 100 - std(ACU_i), \forall\ i = 1 \ldots .6 \qquad (1)$$

where ACU_i denotes the overall accuracy of a predictor on the samples of a subgroup *i*. DP receives its maximum value (100) when all subgroup accuracies are equal.

Definition 2 (*Balanced Equalized Odds*) All groups (protected and unprotected) should have equal rates for true positives (TP) and true negatives (TN). This fairness definition combines two criteria: (i) equalized odds between groups (e.g. $TP_{males} = TP_{females}$ and $TN_{males} = TN_{females}$) and (ii) equalized odds between classes (e.g. $TP_{males} = TN_{males}$ and $TP_{females} = TN_{females}$). The proposed Balanced Equalized Odds (BEO) criterion is defined as follows:

$$BEO\ (\%) = 100 - std([TP_1, TN_1, \ldots, TP_K, TN_K]) \qquad (2)$$

where *K* the number of subgroups (*K*=6 in our work). BEO receives 100 in the ideal case in which $TP_i = TN_i, \forall\ i$.

Validation using benchmark machine learning algorithms

To effectively use the developed algorithm for classifying OA categories, it needs to be assured that the algorithm achieves its goal, with advantages compared with other benchmark machine learning algorithms. By comparing the results achieved by the developed algorithm with those presented by other algorithms, one can assess the viability, applicability, and quality of the classification algorithm. The methods selected for comparison purposes are decision trees, SVMs, kNN (with k=1 and 5), Adaboost, and Random Forest that are typically recommended for classification problems.

Results and discussion

Accuracy performance on the full dataset

This section reports the results of the conducted experiments with different DNN architectures on the full dataset. The proposed DNN models were applied on the 2-class problem and the obtained classification accuracies along with associated confusion matrixes and class accuracies are given in Table 17 with the without data resampling, respectively.

Best accuracies in the majority of the DNN architectures were received without the application of data resampling, whereas the best overall performance (79.6%) was achieved by the DNN model with 1 hidden layer and 50 nodes per layer (see Table 16). In regards to the effectiveness of the SMOTE-SVM resampling mechanism, the following remarks can be extracted:

The reported confusion matrixes (gray area in Table 17) reveal the inability of the proposed methodology (without data resampling) to recognize participants in the progression class that receives moderate class accuracies (from 62.63% to 69.19%). The application of data resampling on the training sets leads to increased class accuracies for the progression class (from 68.18 to 76.52%) and consequently more balanced confusion matrixes (Table 17a). Nevertheless, this increase in the class accuracies comes with a small reduction in the overall accuracies of the models in Table 17b (best accuracy observed: 78.81%).

Table 17. a) *Overall testing performance of the proposed DNN methodology for different network architectures and b) Overall testing performance of the proposed DNN methodology with SMOTE for different network architectures.*

Hidden layers	Num. of nodes		*progression*	*Incidence*	Class accuracy	Overall accuracy
1	50	*progression*	274	122	69.19	**79.60**
		Incidence	163	838	83.72	
1	100	*progression*	273	123	68.94	79.24
		Incidence	167	834	83.32	
2	50	*progression*	273	123	68.94	77.81
		Incidence	187	814	81.32	
2	100	*progression*	248	148	62.63	78.10
		Incidence	158	843	84.22	
3	50	*progression*	274	122	69.19	77.88
		Incidence	187	814	81.32	
3	100	*progression*	267	129	67.42	79.03
		Incidence	164	837	83.62	

a)

Hidden layers	Num. of nodes		*progression*	*Incidence*	Class accuracy	Overall accuracy
1	50	*progression*	303	93	76.52	
		Incidence	215	786	78.52	77.95
1	100	*progression*	281	115	70.96	
		Incidence	198	803	80.22	77.59
2	50	*progression*	270	126	68.18	
		Incidence	180	821	82.02	78.10
2	100	*progression*	283	113	71.46	
		Incidence	196	805	80.42	77.88
3	50	*progression*	290	106	73.23	
		Incidence	190	811	81.02	**78.81**
		progression	288	108	72.73	

3	100	*Incidence*	195	806	80.52

b)

(i) Overall, SMOTE-SVM had a positive effect on the classification of the smaller class (4.47% average increase) and a slightly negative effect on the overall accuracy (0.79% reduction).

Table 18. *Best performance achieved on subgroups.*

subgroup	Hidden layers	Num. of nodes		*progression*	*Incidence*	Class accuracy	Overall accuracy
Males	2	50	*progression*	74	342	82.21	78.58
			Incidence	108	55	66.26	
Females	3	50	*progression*	178	59	75.11	
			Incidence	116	468	80.14	78.68
Over 70	2	50	*progression*	77	33	70.00	
			Incidence	25	201	88.94	82.74
Under 70	1	100	*progression*	180	119	60.20	
			Incidence	113	659	85.36	78.34
Obese	1	100	*progression*	154	58	72.64	
			Incidence	52	265	83.60	79.21
Non - obese	3	50	*progression*	127	84	60.19	
			Incidence	76	593	88.64	81.82

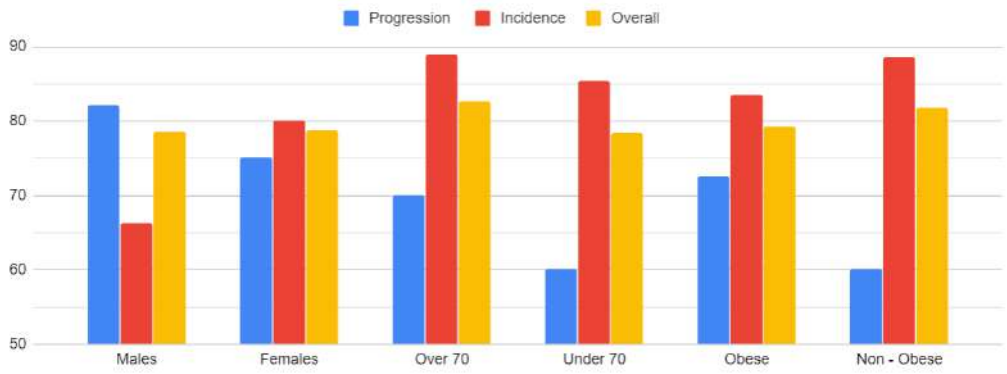


Figure 21. *Best class- and overall accuracy obtained on subgroups*

Results on subgroups

Next, the proposed DNN architectures were trained on data from six subgroups of participants: (i) participants older than 70 years, (ii) participants under 70 years, (iii) male participants, (iv) female participants, (v) non-obese and (vi) obese participants. Table 18 cites classification accuracies obtained by the proposed methodology (without data resampling) trained on the aforementioned data subgroups with the full feature set. Significant differences were observed between these subgroups and the entire dataset. In the following subsections, the results of each subgroup are analyzed and explained.

Results from gender effect in the diagnosis

Overall accuracies of ~78.6% and a negligible difference of approximately 0.1% 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.

With regards to class accuracies, both progression and incidence classes were classified with accuracies higher than 75% in females, whereas a significant difference between the two classes was observed in the class accuracies on the male subgroup (82.21% and 66.26% for progression and incidence classes, respectively).

Results from age subgroups

Table 19. *Performance achieved by the proposed DNN methodology with and without SMOTE.*

		Best accuracy on Full set (DNN architecture: 1 hidden layer of 50 nodes)				Best accuracy on full Set with SMOTE (DNN architecture: 3 hidden layers of 50 nodes)			
		Progression	Incidence	Class accuracy	Overall	Progression	Incidence	Class accur	Overall
obese	*progres*	133	52	71.89		139	46	75.14	
	Inciden	51	271	84.16	79.68	57	265	82.30	79.68
non-	*progres*	141	70	66.82		151	60	71.56	
	Inciden	112	567	83.51	79.55	133	546	80.41	78.31
over70	*progres*	61	29	67.78		66	24	73.33	
	Inciden	26	172	86.87	80.90	25	173	87.37	82.99
under	*progres*	213	93	69.61		*224*	*82*	73.20	
	Inciden	137	666	82.94	79.26	165	638	79.45	77.73
male	*progres*	127	41	75.60		126	42	75.00	
	Inciden	60	349	85.33	82.50	72	337	82.40	80.24
female	*progres*	147	81	64.47		164	64	71.93	
	Inciden	103	489	82.60	77.56	118	474	80.07	77.80

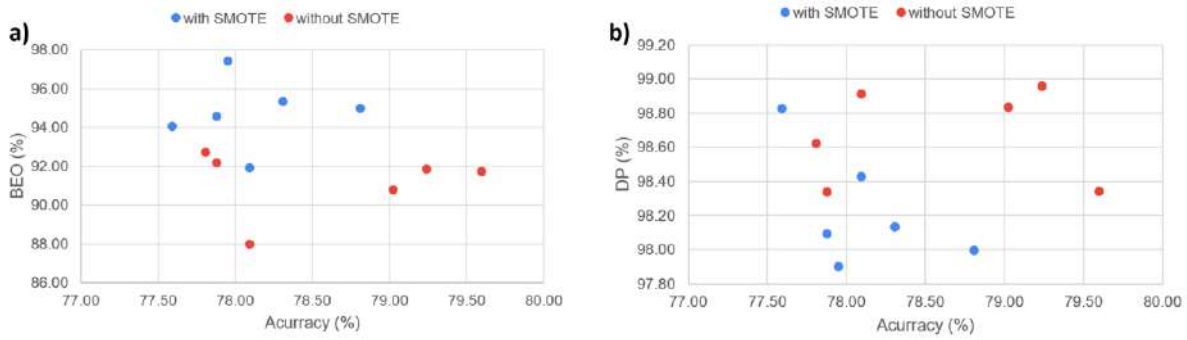


Figure 22. *Fairness with respect to the accuracy with and without SMOTE for the proposed DNN methodology: a) BEO versus accuracy and b) DP versus accuracy.*

A significant difference was observed between the two age subgroups. Specifically, a performance of 82.74% was achieved on the knee OA recognition for older participants, whereas the knee OA diagnosis accuracy of the 70- age subgroup (78.34%) was closer to the overall accuracy taken on the entire dataset. The accuracy obtained by the DNN model built on the aged subgroup (70+) was the highest reported in this work. This finding implies that local models trained on more focused populations could provide better decisions focusing on the specific characteristics of the subgroup population, thus outperforming global models trained on the entire dataset.

Results from obesity subgroups

Examining the results of the two-weight subgroups, a moderate difference of approximately 2.5% was observed. Specifically, a performance of 81.82% was achieved on the knee OA recognition for participants on the non-obese subgroup, whereas the knee OA diagnosis accuracy of the obese subgroup (79.21%) was closer to the overall accuracy taken on the entire dataset.

Figure 21 summarizes the overall and per-class accuracies obtained from the models built on participants' data from separate subgroups. The variability in the obtained accuracies can be attributed to the fact that any learning methodology strongly depends on the dataset in which is trained on. In our case, the proposed DNN methodology has provided higher accuracies for the majority class (incidence) in 5 out of the 6 cases with the overall accuracy in between the two class accuracies. The most balanced distribution of accuracies (for progression, incidence, and overall) was achieved in the female subgroup.

The results above indicate the need for further analysis with respect to the predictive capacity of any learning methodology not only on entire datasets but also on (sensitive or not) data subgroups. To address this challenge, the following subsections focus on a more extended validation of the proposed DNN methodology and benchmarks for both accuracy and fairness.

Accuracy versus fairness

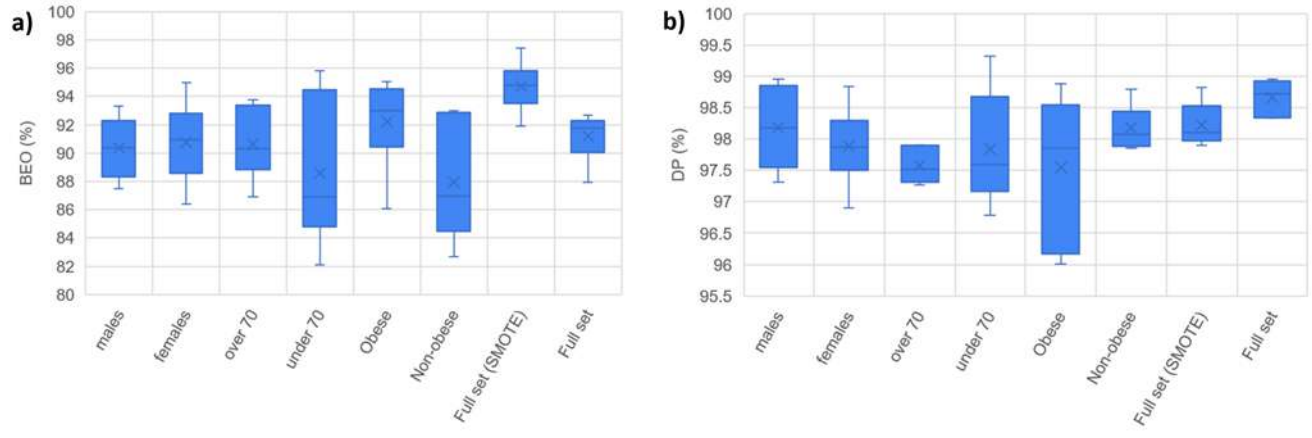


Figure 23. *Fairness achieved by the proposed DL methodology in subgroups and the full set: a) BEO and b) DP*

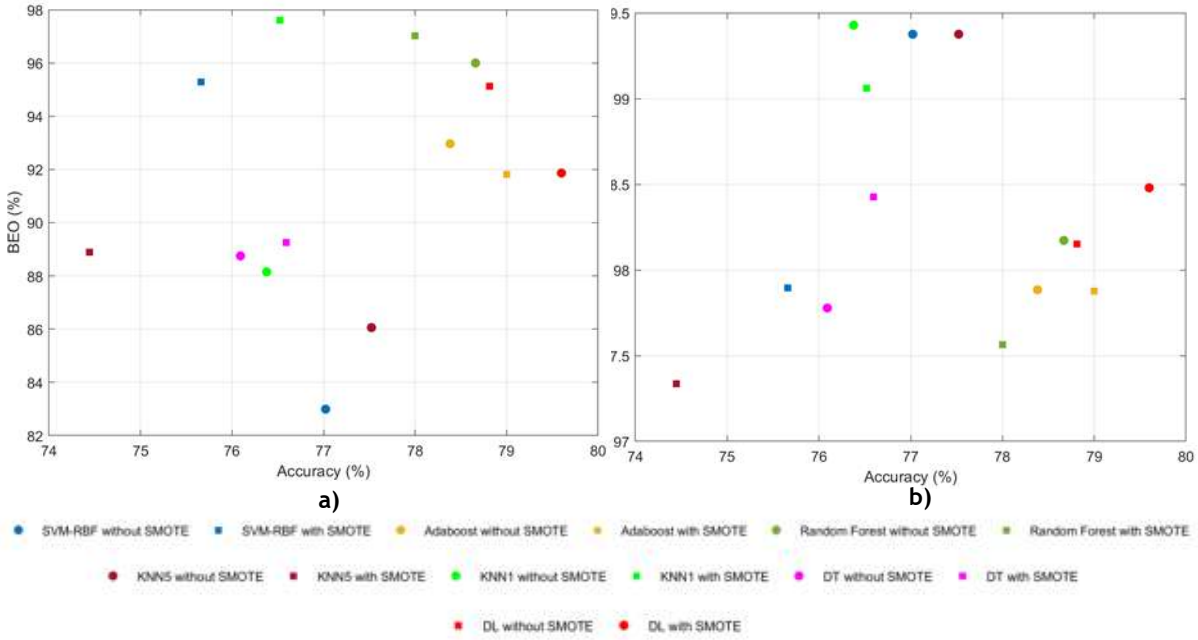


Figure 24. *Comparison of the proposed DNN methodology with benchmarks for accuracy and fairness*

This subsection provides a more detailed representation of the obtained performance of the proposed methodology with respect to both accuracy and fairness with and without the application of data resampling through SMOTE. Specifically, the DNN methodology was trained on the entire training dataset and the performance is presented separately for each one of the six subgroups on the testing set. Table 19 presents the performances accomplished by the most accurate DNN architectures with and without SMOTE-SVM (on the right and left side on the table, respectively).

Comparable subgroup accuracies were received for both approaches (with and without data sampling), whereas a significant difference was observed in the class accuracies. Specifically, the class accuracies of the SMOTE-enabled models obtained on the 6 subgroups received values in the range of 71.76% – 87.37%, whereas the respective class accuracies of the non-SMOTE models were in the range of 64.47% – 86.87%. These findings are verified in Figure 22a that presents the fairness performance (as measured by BEO) with respect to overall accuracy for all the different DNN architectures that were investigated in this work. It is concluded that SMOTE has a positive effect on fairness performance (BEO) but at the same time it leads to slightly less accurate models. In terms of demographic parity (Figure 22b), both approaches had a comparable performance with negligible differences in DP values (with the range of <1%).

Figure 23 shows the fairness performance of the proposed DNN methodology as trained on participants of each one of the 6 subgroups and the full set (with and without data sampling). The best BEO performance was achieved by the SMOTE-enabled model trained on the full set. Training the proposed DNN methodology on the full set (without SMOTE) led to the highest DP performance. Overall, the following remarks can be extracted from the results of this subsection:

Data sampling has a positive effect on the fairness performance of the DNN methodology leading at the same time to more balanced rates for TP and TN throughout all data subgroups.

The increase in fairness performance comes with a small decrease (<1% on average) on the overall predictive accuracy of the models.

Training on the full dataset increases fairness (both BEO and DP). Thus, special attention should be given in the selection of the training sets that need to represent the whole data variability comprising participants from all sensitive subgroups.

Comparative analysis with benchmark classifiers

One of the aims of this work was to compare DNN with a variety of well-known machine learning algorithms on the 2-class classification problem using the entire feature sets. To further validate the proposed DNN, the following machine learning algorithms were evaluated for the KOA classification problem: Decision trees (DTs), KNN with k=1 and 5, support vector machines (SVM) algorithms with RBF kernel and two ensemble techniques, AdaBoost and Random Forest. Figure 4 compares the performance of the proposed DNN methodology with benchmarks with respect to both accuracy and fairness. DNN accomplished the optimum overall performance with the best accuracy (79.6%) and high fairness values (BEO: ~92% and DP:98.5%). The second-best performance was received by AdaBoost which was slightly less accurate (~79%) and less fair (BEO: ~92% and DP: 97.9%). High BEO values (>96%) were achieved by Random Forest that received lower DP values compared to DNN and was less accurate (78.6%). The highest BEO values were achieved by KNN1 without SMOTE. However, this model was less accurate with ACU<77%. The rest of the ML models had moderate performances in terms of accuracy and/or fairness. Consequently, the proposed DNN outperforms the above well-known machine learning techniques in the knee OA diagnosis task.

Comparison with existing non-invasive techniques

This subsection focuses on a comparison between the predictive accuracy of the proposed methodology and existing non-invasive AI-based techniques of the recent literature. A deep neural network for detecting the occurrence of osteoarthritis has been presented in using patient's statistical data of medical utilization and health behavior information. The study was based on 5749 subjects and resulted in 76.8% of the area under the curve (AUC). Similar to the previous study, a DNN-based methodology was proposed in utilizing risk factors from self-reported clinical data about joint symptoms, disability, function, and general health. The proposed methodology was demonstrated in the entire OAI population (with an accuracy of 80.74%) as well as in subgroups defined by gender and age where higher accuracies were reported. History and clinical characteristics of the subjects such as age, body mass index, and pain level have been also considered for decision-making in OA diagnosis. A success rate of about 80% was achieved using a decision tree equipped with multilayer perceptrons at its leaves. Alternatively, biomechanical data from human body motion analysis has been also explored as risk factors that could contribute to OA diagnosis resulting to detection accuracies up to 93% (demonstrated in datasets of moderate size). The predictive capacity of physical activity measures as contributing factors in the progression of KOA has been also investigated in leading to accuracies up to 74.5%. Overall, in terms of accuracy, the proposed in this work methodology provided comparative results with studies employing similar features (non-imaging history and/or clinical data). However, unlike all the aforementioned works, the main novelty of this work lies in the inclusion of fairness metrics for the performance evaluation of the classification results.

6. Quantifying of MRI impact

The entire OAI MRI Data set consists of 2002 variables, one response variable, and 679 observations. Two variables of the data were removed because as data indexes. The rest of the variables are separated into two major categories, femoral and tibial consisting of 1000 variables each. In each of those categories of the variables the variables further separate in two categories, one category for the mean values and one category for the standard deviation.

The response variable is a two classes categorical variable. The first class, represented with "0", is the class of a person without osteoarthritis which will not present osteoarthritis in the near future. The second class, represented with "1", is the class with a person which in the near future will present osteoarthritis in one or both of his knees. From now on the second class represented with "1" will be considered as the positive class.

A corelation analysis of the dataset was applied for all the possible combinations of the above data categories [1]. For brevity, only the corelation, of the variables' categories combination which provide the best results will be provided. The correlation analysis of the selected variables saw that there are highly correlated variables inside the data set in the figure below we can see how many variable pairs are correlated (positively or negatively) and in what degree.

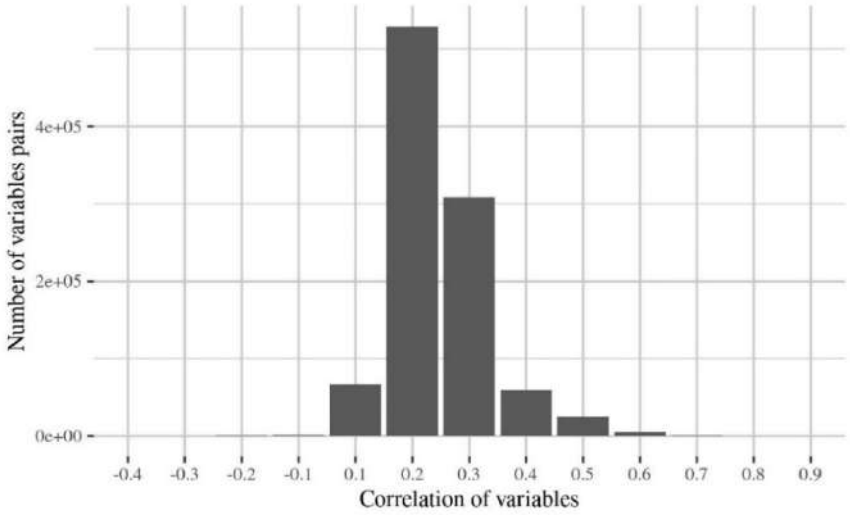


Figure 25. *Corelation analysis.*

To produce the Figure 25, a correlation matrix was created with the Spearman's rank correlation coefficient measure. From this matrix, we removed the values of the major diagonal because it represents pairs of the same variables which they have correlation 1.

Highly corelated variables in the dataset are considered variables with corelation values over 0.7. The number of which will be presented in the table below for all the possible combinations of the OAI MRI variables' categories.

Table 20. *OAI MRI variables' categories.*

Category combination name	*Number of corelated variables*
all	2000 / 2000
femoral	1000 / 1000
femoral mean	305 / 500
femoral std	500 / 500
mean	595 / 1000
std	985 / 1000
tibial	1000 / 1000
tibial mean	290 / 500
tibial std	489 / 500

In the above figure we can observe that there a big number of highly corelated variable in all the variable categories, and in some cases the entire dataset is highly corelated. The smallest proportion of highly corelated variables can be found in the tibial mean variables category while the mean variables category also, contains among the smallest proportions of highly corelated variables.

Data pre-processing

Feature selection (Variable Importance)

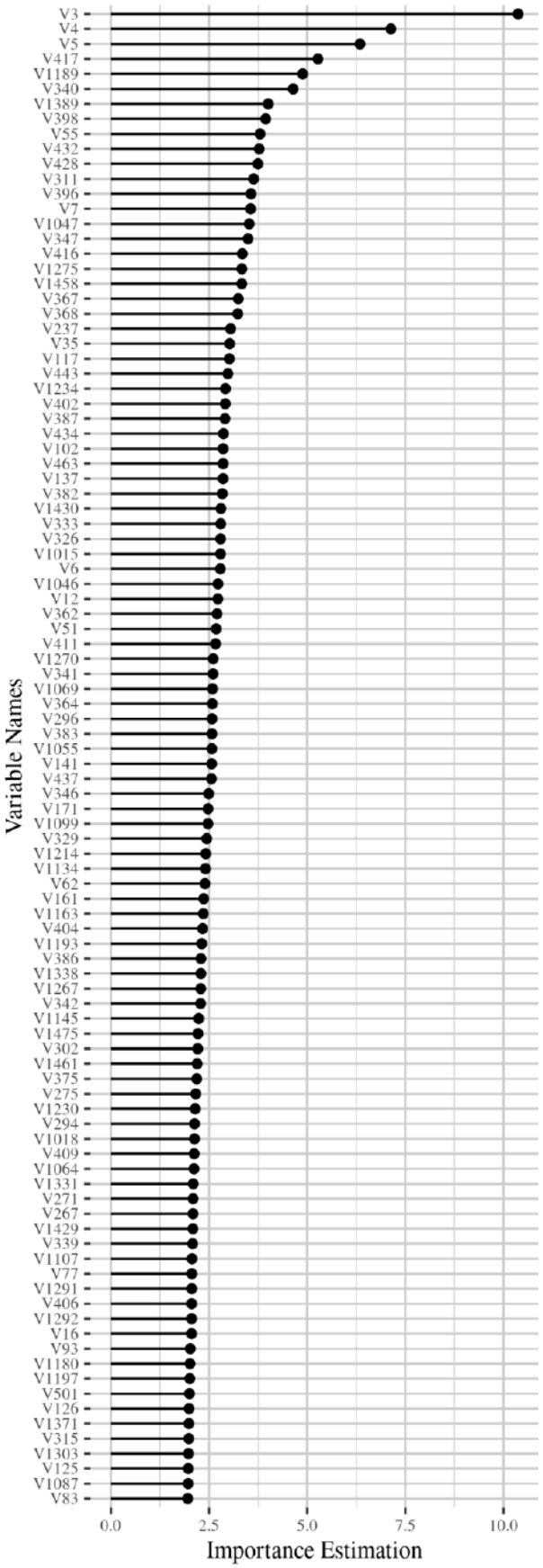
The first pre-processing step executed for the data set is the Variable Importance Analysis with the use of the Random Forest Algorithm (RF). By using the Mean Decrease in Accuracy (MDA) measure. The resulted variables' selection was NOT used in the analysis. Its purpose is a better understanding of the data set.

In order to calculate the MDA with the RF algorithm the permuted out-of-bag (OOB) data where used. Specifically, by recording the prediction error on the OOB portion of data, for each tree. The same process is repeated after permuting each predictor variable. The difference between the two (Decreases in Accuracy of Trees) is then averaged over all trees, and normalized by the standard deviation of the differences.

$$MDA = \frac{Mean(Decreases\ in\ Accuracy\ of\ Trees)}{StandardDeviation(Decreases\ in\ Accuracy\ of\ Trees)}$$

For the calculation of the importance for the OAI MRI dataset variables, a RF model was created with the use of 1000 variables of the mean features. This decision is based on the classification results of all the categories combinations. This model creation was possible because RF is a decision tree-based algorithm.

In the following figures 200 of the 1000 variables contained in then mean dataset, the 100 variables with the highest MDA score on the left figure. On the right figure there are the 100 variables with the lowest MDA score. In both figures the variables are presented in descending order, from the most important to the less important. It is noteworthy that all the variables of the lowest MDA score figure have scores bellow zero.


V3
V4
V5
V417
V1189
V340
V1389
V398
V55
V432
V428
V311
V396
V7
V1047
V347
V416
V1275
V1458
V367
V368
V237
V35
V117
V443
V1234
V402
V387
V434
V102
V463
V137
V382
V1430
V333
V326
V1015
V6
V1046
V12
V362
V51
V411
V1270
V341
V1069
V364
V296
V383
V1055
V141
V437
V346
V171
V1099
V329
V1214
V1134
V62
V161
V1163
V404
V1193
V386
V1338
V1267
V342
V1145
V1475
V302
V1461
V375
V275
V1230
V294
V1018
V409
V1064
V1331
V271
V267
V1429
V339
V1107
V77
V1291
V406
V1292
V16
V93
V1180
V1197
V501
V126
V1371
V315
V1303
V125
V1087
V83
Variable Names
0.0
2.5
5.0
7.5
10.0
Importance Estimation

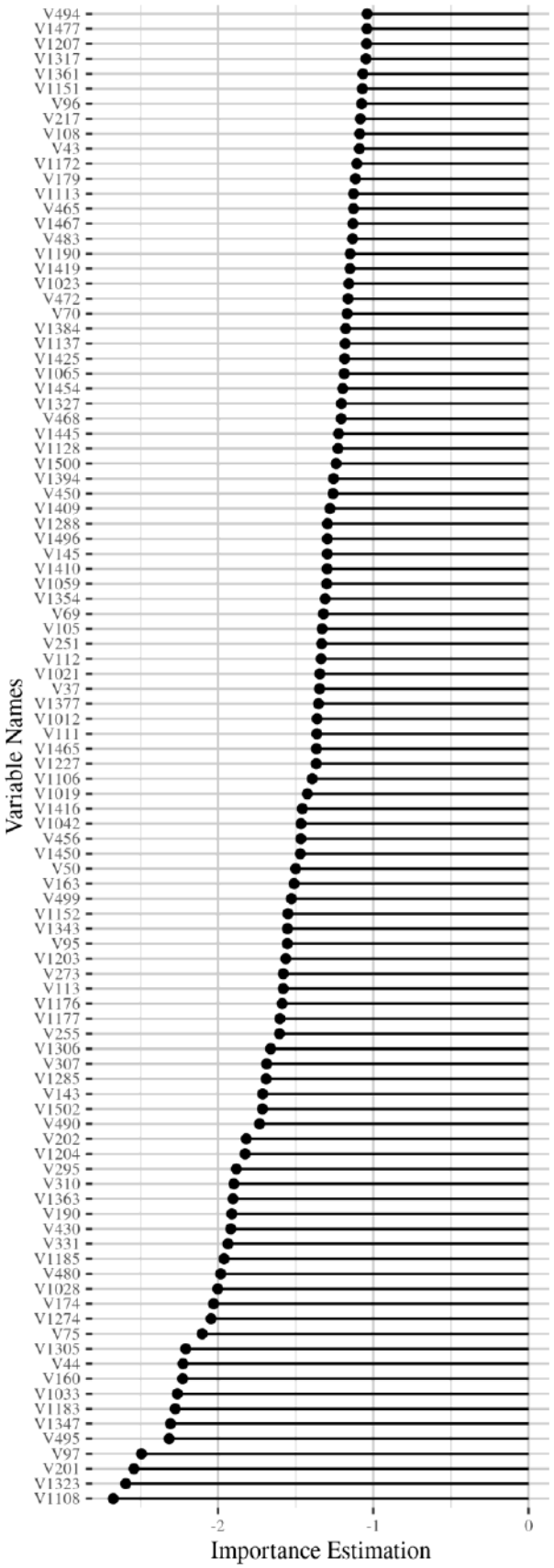


Figure 26. *Variables with the highest MDA score.*

Data normalization

For the normalization of the data the technique of min-max scaling or min-max normalization. This technique is rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)},$$

where x is an original value, x ' is the normalized value. After the scaling of the data the ranges of the 100 first variables are sawn in the figure bellow.

Dimensionality reduction

Because of the data's high dimensionality, it was dimmed necessary to implement dimensionality reduction techniques. The chosen examined techniques are the Principal Components Analysis (PCA) and the Random Projections (RP) [2, 3]. For the PCA technique, we calculate the principal components matrix and the projections of the data on the desired number of dimensions. The number of principal components was selected in two ways, first we arbitrary selected 2, 3, 4, and 5 principal components. The second way is the from the cumulative amount of variance explained by each principal component, from a literature review this number is 0.7. The number of principal components, that explain 0.7 of the total variance in the OAI MRI dataset, is 6.

For the RP technique, a projection matrix was created with use of Gaussian Distribution. The dimensions of the resulted vector space are selected in two ways, similar to the PCA technique, arbitrary and by a literature recommendation. Arbitrary we select 2, 3, and 4 dimensions. After a literature review, we implemented the Johnson-Lindenstrauss lemma with an error tolerance of 0.5. This methodology gave us a result of 6 dimensions.

Table 21. *Classification results with colour coding to aid visual identification of the performance of the different algorithms. Green: indicates specificity or sensitivity value of 1, and the respective sensitivity or specificity value is 0 in those cases. Those values cannot be taken into account; Yellow: maximum values of each of the algorithms; Red: maximum values on each of the metrics.*

	kNN			*LDA*			*mLR*			*RF*			*XGB*		
	acc	sens	spec	acc	sens	spec	acc	sens	spec	acc	sens	spec	acc	sens	spec
PCA 2dims	0.526	0.508	0.544	0.571	0.646	0.500	0.571	0.646	0.500	0.556	0.554	0.559	0.564	0.585	0.544
PCA 3dims	0.489	0.446	0.529	0.474	0.508	0.441	0.474	0.508	0.441	0.519	0.508	0.529	0.571	0.569	0.574
PCA 4dims	0.496	0.492	0.500	0.451	0.508	0.397	0.466	0.508	0.426	0.564	0.600	0.529	0.617	0.692	0.544
PCA 5dims	0.519	0.523	0.515	0.474	0.508	0.441	0.481	0.508	0.456	0.489	1.000	0.000	0.549	0.585	0.515
PCA 0.7var 6dims	0.511	0.523	0.500	0.489	0.477	0.500	0.489	0.492	0.485	0.429	0.431	0.426	0.541	0.523	0.559
RP 2dims	0.459	0.400	0.515	0.444	0.431	0.456	0.444	0.431	0.456	0.511	0.000	1.000	0.571	0.538	0.603
RP 3dims	0.526	0.538	0.515	0.564	0.631	0.500	0.564	0.631	0.500	0.511	0.000	1.000	0.541	0.508	0.574
RP 4dims	0.481	0.477	0.485	0.504	0.585	0.426	0.504	0.585	0.426	0.511	0.000	1.000	0.511	0.492	0.529
RP JL lemma 6dims	0.507	0.532	0.487	0.500	0.500	0.500	0.500	0.500	0.500	0.551	0.000	1.000	0.507	0.468	0.539

As it is consistent with the results of the classification without the application of dimensionality reduction, when we apply dimensionality reduction the results of the classification accuracy relations between the classification algorithms remain almost identical, with the best results to be provided by the XGBoost algorithm. Specifically, the best classification accuracy is achieved from the PCA technique at 4 dimensions, with a drop in accuracy of 0.033 from the original 0.650.

Visualization of the data

For the visualization of the data the Principal Components Analysis was utilized. The visualization was applied on for all the data, the femoral variables, the tibial variables, the mean variables, and the std variables. The projection of the data on the first 2 principal components is given in the Figures 27-31 bellow.

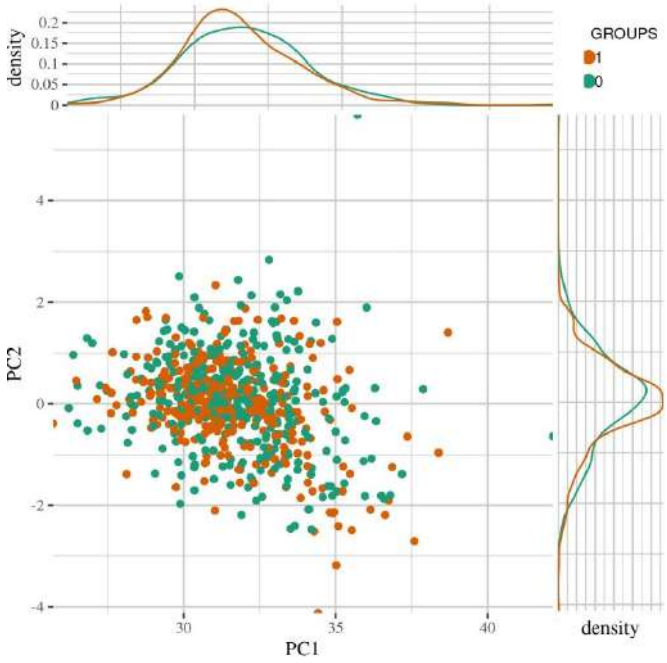


Figure 27. *Principal Components Analysis for all variables.*

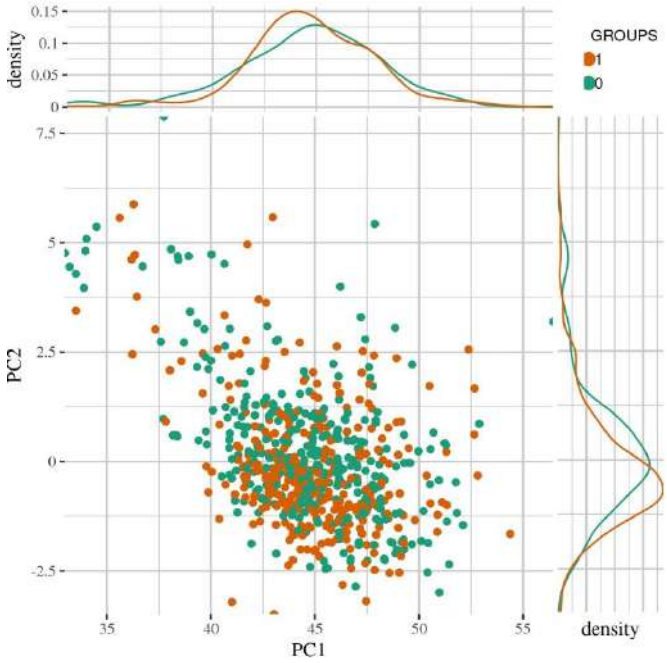


Figure 28. *Principal Components Analysis for femoral variables.*

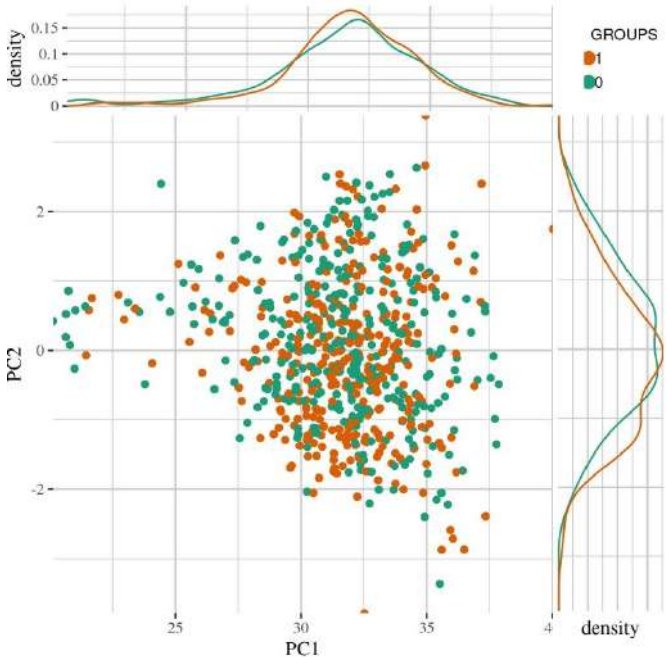


Figure 29. *Principal Components Analysis for tibial variables.*

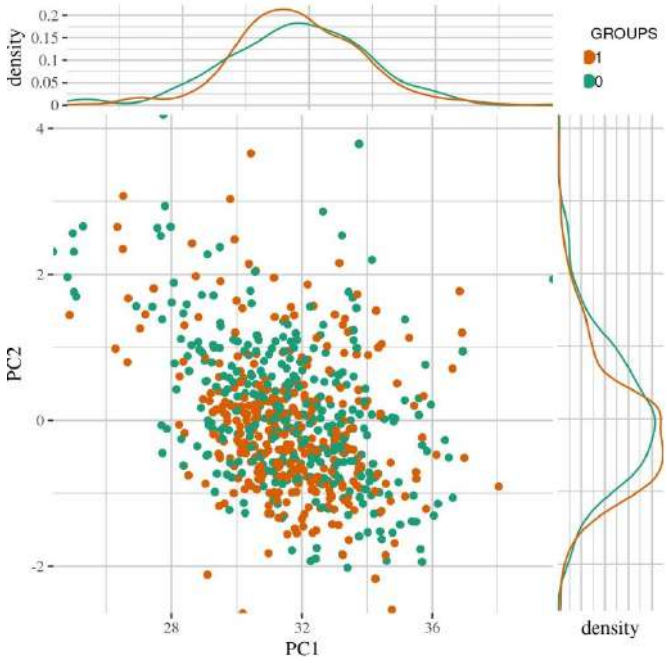


Figure 30. *Principal Components Analysis for mean variables.*

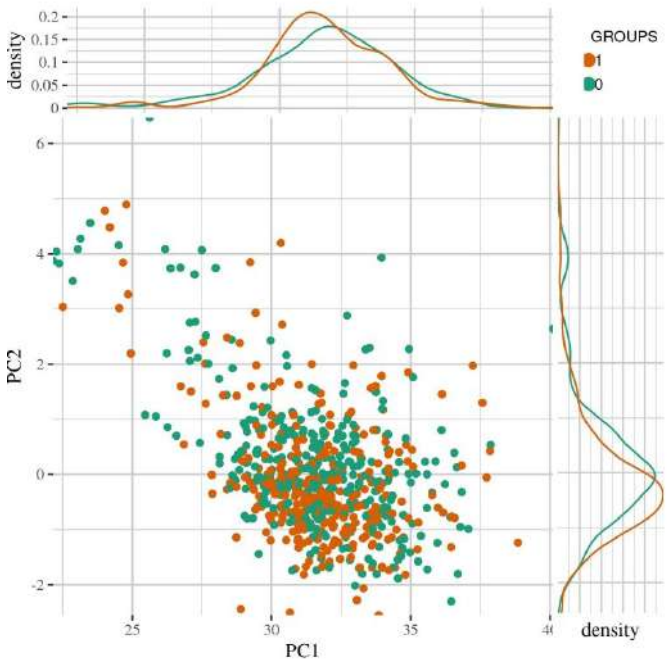


Figure 31. *Principal Components Analysis for std variables.*

As it is immediately apparent from the representation of the data, the Classes "0" and "1" aren't segregated at all, in all the cases. As the density plots of each principal component represent the classes have almost identical distribution in the space. Form these visualizations an initial conclusion is that we have an almost impossible classification problem to tackle.

Classification tasks

The classification algorithms used are the Multinomial Logistic Regression (MLR), the Linear Discriminant Analysis (LDA), the k-Nearest Neighbours (kNN), the Random Forest (RF) and finally the XGBoost. These algorithms are applied on almost all the variable categories of the OAI MRI dataset. Specifically, on all the variables, on the mean variables, on the std variables, on the femoral variables, on the femoral mean variables, on the femoral std variables, on the tibial variables, on the tibial mean variables, and on the tibial std variables.

The results of the classifications can be seen on the table below. The measures used are accuracy, sensitivity and specificity.

Table 22. *Classification results with colour coding to aid visual identification of the performance of the different algorithms. Green: indicates specificity or sensitivity value of 1, and the respective sensitivity or specificity value is 0 in those cases. Those values cannot be taken into account; Yellow: maximum values of each of the algorithms; Red: maximum values on each of the metrics.*

	kNN			*LDA*			*mLR*			*RF*			*XGB*		
	acc	sens	spec	acc	sens	spec	acc	sens	spec	acc	sens	spec	acc	sens	Spec
all	0.577	0.901	0.186	0.562	0.549	0.576	0.538	0.507	0.576	0.330	0.479	0.254	0.538	0.479	0.610
femoral	0.523	0.901	0.086	0.536	0.543	0.529	0.556	0.606	0.504	0.464	0.000	1.000	0.536	0.479	0.594
femoral mean	0.486	0.938	0.105	0.479	0.479	0.479	0.500	0.500	0.494	0.479	0.479	0.479	0.606	0.606	0.606
femoral std	0.532	0.912	0.169	0.479	0.479	0.479	0.506	0.606	0.507	0.479	1.000	0.000	0.538	0.538	0.538
mean	0.480	0.852	0.211	0.538	0.479	0.538	0.538	0.538	0.479	0.479	1.000	0.000	0.606	0.606	0.606
std	0.538	0.829	0.214	0.538	0.606	0.479	0.538	0.538	0.479	0.538	0.000	1.000	0.538	0.538	0.538
tibial	0.538	0.705	0.375	0.538	0.538	0.538	0.538	0.538	0.538	0.479	0.000	1.000	0.538	0.538	0.538
tibial mean	0.479	0.538	0.479	0.538	0.538	0.479	0.479	0.479	0.479	0.479	0.479	0.538	0.479	0.479	0.479
tibial std	0.538	0.901	0.186	0.479	0.538	0.479	0.479	0.479	0.479	0.538	0.538	0.538	0.538	0.538	0.538

In the results we can observe that there is a diversity of the classification accuracy, sensitivity, and specificity based on the variable categories that are used. Even more there is a significant diversity between the used classification algorithms. Between the classification categories it is expected to exist diversity on the results. The fact that proves that claim is shown on the results correlation analysis and further established visually on the PCA visualization of the data in combination with the unique characteristics of each classification algorithm.

From the above results we can conclude that the results of the XGBoost algorithm are among the most satisfactory, between the results of all the classification algorithms. Except the largest classification accuracy in most cases, we can observe a balance between the prediction of the positive and negative class of the

response variable. The only case which one of the other algorithms, LDA specifically, archives those criteria better is the case of the category for all the variables of the OAI MRI dataset.

The above results in the conclusion that the best classification accuracy can be provided from the mean variable category. Bellow, (Figures 32-36) are presented the Ro Curves which illustrate the results, only, of the mean variable category:

MLR

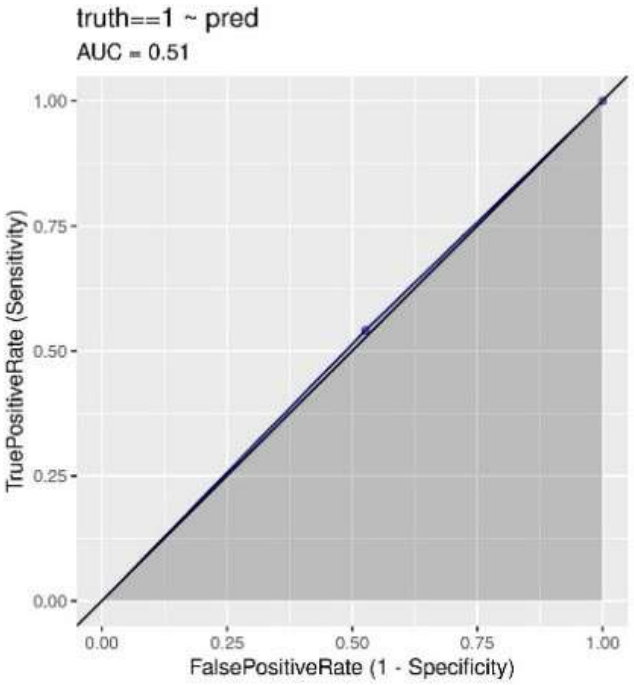


Figure 32. *AUC for MLR model.*

LDA

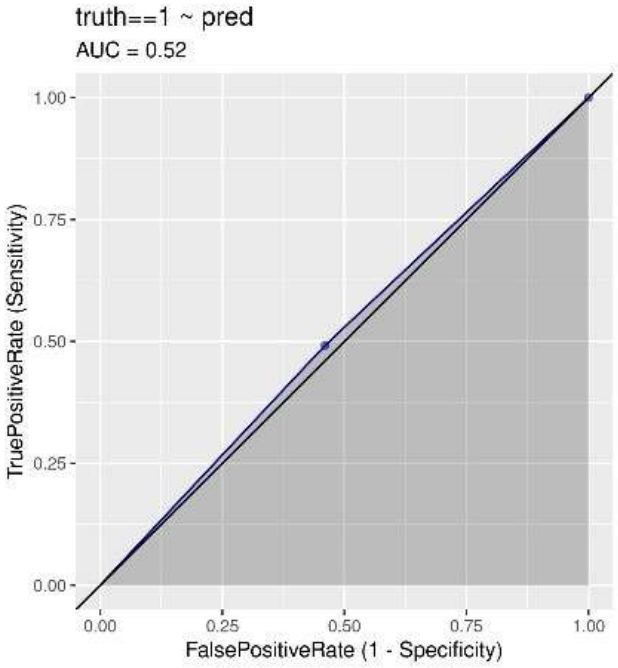


Figure 33. *AUC for LDA model.*

kNN

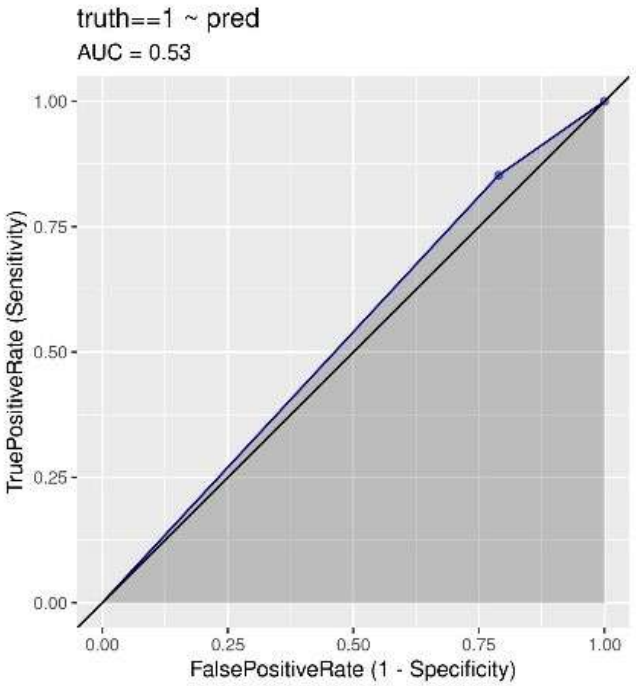


Figure 34. *AUC for kNN model.*

RF

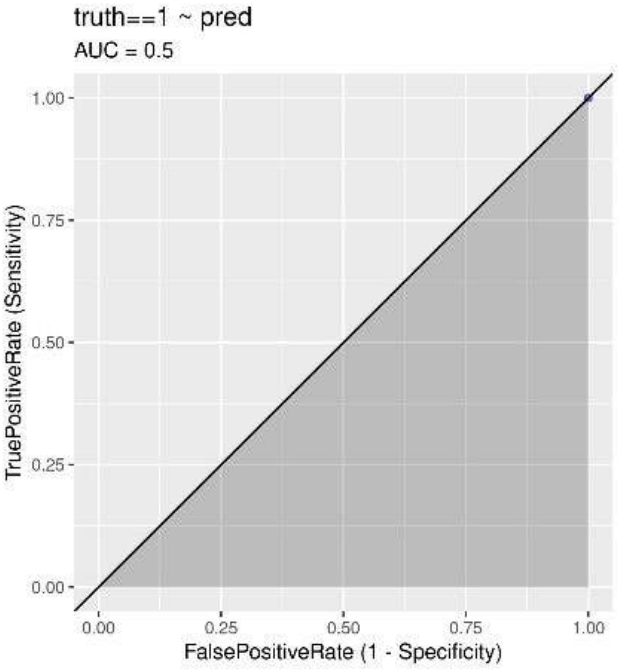


Figure 35. *AUC for RF model.*

XGBoost

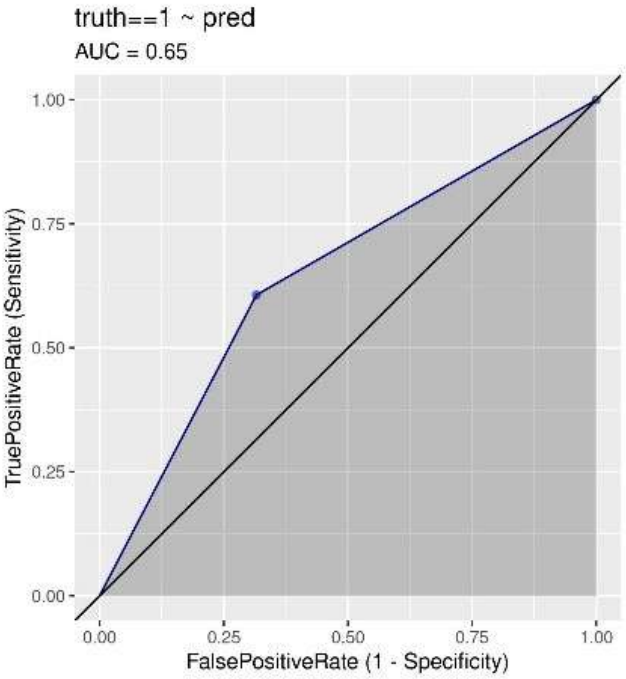


Figure 36. *AUC for XGBoost model.*

The extensive analysis presented in this report shows us the challenges presented in the OAI MRI dataset. The findings of this analysis gave us valuable insight into this database. We found that it is a dataset that

contains a large number of correlated variables. This results in a not so separable dataset, as shown even in all classification tasks. Finally, form the dimensionality reduction analysis we found that we can reduce the dimensions on the dataset with a relatively small cost in the classification accuracy.

7. Interpretable models

Diagnostic Model

The model used in the diagnostic analysis was logistic regression. The logistic regression approach was chosen after considering alternative analysis methods [10]. This is due to the method being preferred by clinicians as it reflects well their decision making process. The goal for the logistic model is to determine, based on 8 features relating to a subject, whether they are likely to have KOA and therefore require further investigations into their symptoms. The presence of clinical KOA, KL grade 2 or above is the outcome. The model was trained and tested using the OAI data with 1353 and 1354 subjects respectively.

Prognostic Model

The prognostic analysis uses Cox regression to model how the covariates jointly influences the probability of the subject developing KOA. After modelling with Cox we created cohorts by risk stratification, to highlight the criteria for being low and high risk at developing KOA in 5 years from the baseline assessment. To stratify the group into cohorts the first step is to establish a cut point that gives the biggest separation in the subjects. This is done with a model containing 5 variables from the subjects initial assessment. The groups used to model this analysis is taken from the original OAI data and removes subjects with KOA at their initial assessment. The model was then trained and tested on 1002 and 1003 subjects from the OAI dataset respectively.

External Validation

To validate and ensure that these models were not overfitting to the OAI data used to train them, we used the MOST data set to validate the results. The MOST data was collected from different centres than those used in the OAI study so this helps to determine if the model is able to stand up to institutional bias. This helps to assess if the model can be used outside of the bounds from which the data was collected, and also contributes to showing that a prediction model is more suitable for use in clinical practice [11]. The validation set for the diagnostic model is 2006 subjects, whilst the validation set is 1155 subjects.

Results

The OAI data test set has a prevalence of KOA at 39%. The MOST validation set prevalence is at 60%. The probabilistic cut-off for binary classification was taken to be 0.5. The AUROC and CI for the diagnostic model, for test and validation are 0.7475 (0.7209-0.7742) and 0.6697 (0.6311-0.7082) respectively.

As the data for the prognostic model differs from that in the diagnostic model the training and test sets are also different. The training set and test set are made up of 1002 and 1003 subjects respectively. The external validation set contains n = 1155 subjects.

Figure 37 shows the stratification curves on the OAI training data for the raw data and the predictions produced on the MOST validation. The last event recorded in cohort 2 on the training data within the 5-year span is at day 1642. The stratification curves produced are well separated with no crossover on the confidence intervals, which indicates that on unseen data the well-separated groups hold true.

In conclusion, the validation with external data of the diagnostic and prognostic models developed in OActive shows that these models have potential clinical validity. The associated web-apps are currently available only within the consortium, while a journal paper is being submitted. Pending the reviews of the

paper, the apps have potential for widespread use by clinicians to stratify patients and to communicate to them the importance of preventive measures.

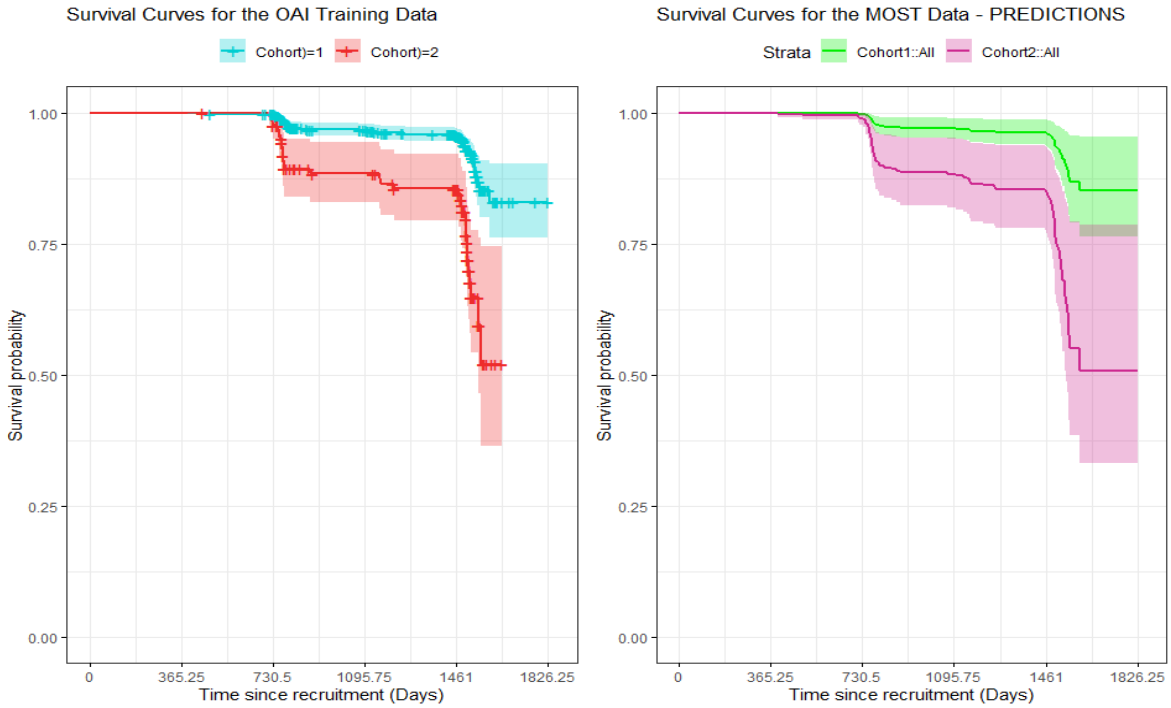


Figure 37. *Stratification curves on the left showing OAI training data showing the high and low risk cohorts. Note the last event recorded in cohort 2 within the 5-year span is at day 1642. The stratification curves on the right are the validation data showing the high and low risk cohorts fitted to the models developed using the OAI data.*

8. Conclusions

This deliverable (Deliverable D9.3) describes the results and outcomes of the long-term evaluation of OACTIVE using data from big data registries in OACTIVE. Deliverable (D9.3) presents the validation of the personalised predictive OACTIVE models, which are used either for prevention, diagnosis, or even during the intervention stage. Specifically, Section 3 presents a baseline performance with PCA on OAI database and in Section 4 the validation of the personalized prediction and diagnosis models is presented. In Section 5 Machine Learning and Deep Learning Diagnosis models with a focus on accuracy and fairness are presented. Section 6 presents a approach for the quantification of MRI impact and finally, in Section 7 Interpretable models validated on MOST data are shown.

Initially, an extensive analysis was presented in this report. This work shows the high quality of OACTIVE's database. The findings of this analysis gave us valuable insight into this database. We found that it is a dataset that contains a large number of correlated variables. This results in a not so separable dataset, as shown even in all classification tasks. In the first look, the resulted accuracy is over 0.7 in all cases. But if we count the unbalanced nature of the data in combination with the sensitivity, which is below 0.16 in all cases, we can conclude that the positive class is not predicted, by the classification models. To cope with this limitation of the data we proposed several technics which have described in Deliverable 6.3 and Deliverable 6.5 and validated in this report.

Then, we worked on the evaluation of the personalised prediction and diagnosis models. 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 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 extensive experimentation, a 74.07% classification accuracy was achieved by SVM on a group of fifty-five selected risk factors (in dataset D). Understanding the contribution of risk factors is a valuable tool for creating more powerful, reliable, and non-invasive prognostic tools in the hands of physicians. For our future work, we are planning to also consider image-based biomarkers and areas with valuable information derived from biomechanical data that are expected to further improve the predictive capacity of the proposed methodology. ML explainability analysis will also be considered to capture the effect of the selected features on the models' outcomes.

Furthermore, the purpose of the second work was: (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 to 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.

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 was initially performed on the JSN progression based on the JSM outcomes of patients over the first five visits. Subsequently, 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). This work was not only focused on the development of prediction models, but also aimed to reveal significant insights regarding the nature of the predictive risk factors that were identified as important. Through this analysis, we concluded that a blend of heterogeneous features from almost all feature categories is necessary to maximize the performance and prediction accuracy of the models. The nature of the selected features along with their impact on the prediction outcome (via SHAP) was also discussed to increase our understanding of their effect on JSN progression. Future work should focus on incorporating morphological knee features as an additional feature category that could potentially increase the performance of the predictive models. These features can be extracted by employing deep learning algorithms for image processing. Alternative data clustering

algorithms, such as self-organizing maps (SOM) could also be explored to improve the clustering performance of the proposed methodology, leading to more informative and distinct data classes.

The fourth work had 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. The nature of the selected features along with their impact on the prediction outcome (via SHAP) was also discussed to increase our understanding of their effect on KOA progression. Identifying and understanding the contribution of risk factors on KOA progression may enable the implementation of better prevention strategies prioritizing non-surgical treatments, essentially preventing an epidemic of KOA.

In addition, we worked in diagnosis of KOA. In this task a machine learning workflow for diagnosis of KOA with a focus on post-hoc explainability is provided. Overall, understanding the inner workings of ML algorithms is of utmost importance. Explainability refers to being able to trace and follow the logic ML algorithms use to form their conclusions. Thus, explainability provides certainty and eliminates prejudices about the correctness of the ML models. The proposed methodology is based on a hybrid approach that combines a robust feature selection technique with a post-hoc explainability analysis via SHAP that enhances our understanding of the methodology which is applied in the diagnosis of KOA. Understanding the contribution of risk factors is a valuable tool for creating more powerful, reliable, and non-invasive diagnostic tools in the hands of physicians.

In section 5, the proposed DNN methodology shows potential for non-invasive OA diagnosis and demonstrates its potential to provide both accurate and fair decisions. In this respect, this work contains original content in the first-ever validation of DNN and machine learning models with respect to fairness in the KOA classification research. Comparative analysis verified the superiority of the proposed methodology for both accuracy and fairness over other common classification methods given similar inputs. This shows that DNNs are a viable tool to be used for medical classification tasks. Future studies should be focused on a wider application of fairness metrics for the assessment of machine and deep learning models applied in medicine. Our plans include 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 evaluate bias and fairness of the generated prediction models that will be trained on data subgroups defined by parameters such as body mass index combined with demographics and social indicators. Open data and scientific tools using unbiased and fair machine/deep learning techniques for OA diagnosis are promising and must be dynamically encouraged within the OA research community.

The extensive analysis in Section 6 shows us the challenges presented in the OAI MRI dataset. The findings of this analysis gave us valuable insight into this database. We found that it is a dataset that contains a large number of correlated variables. This results in a not so separable dataset, as shown even in all classification tasks. Finally, form the dimensionality reduction analysis we found that we can reduce the dimensions on the dataset with a relatively small cost in the classification accuracy.

Statistical models for diagnosis and prognosis were developed with data from the OAI and Multicenter Osteoarthritis Study (MOST) data set. This provided the opportunity for external validation on a substantial data set acquired with a very different protocol. The protocol differences are especially with respect to follow-up. This resulted in apparent differences in the prognostic models. However, the separation of the Kaplan-Meier curves for the predicted high- and low-risk cohorts remained valid. This shows a level of validation that is currently beyond the sate-of-the-art for KOA.

8. References

[1] Sedgwick, P. (2014). Spearman's rank correlation coefficient. Bmj, 349.

[2] Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical methods, 6(9), 2812-2831.

[3] Song, M., Yang, H., Siadat, S. H., & Pechenizkiy, M. (2013). A comparative study of dimensionality reduction techniques to enhance trace clustering performances. Expert Systems with Applications, 40(9), 3722-3737.

[4] 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.

[5] Alexos, A., Kokkotis, C., Moustakidis, S., Papageorgiou, E., & Tsaopoulos, D. (2020, July). Prediction of pain in knee osteoarthritis patients using machine learning: Data from Osteoarthritis Initiative. In 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA (pp. 1-7). IEEE.

[6] Ntakolia, C., Kokkotis, C., Moustakidis, S., & Tsaopoulos, D. (2021). Prediction of Joint Space Narrowing Progression in Knee Osteoarthritis Patients. Diagnostics, 11(2), 285.

[7] Kokkotis, C., Moustakidis, S., Baltzopoulos, V., Giakas, G., & Tsaopoulos, D. (2021). Identifying Robust Risk Factors for Knee Osteoarthritis Progression: An Evolutionary Machine Learning Approach. Healthcare 2021, 9, 260.

[8] Kokkotis, C., Moustakidis, S., Papageorgiou, E., Giakas, G., & Tsaopoulos, D. (2020, July). A Machine Learning workflow for Diagnosis of Knee Osteoarthritis with a focus on post-hoc explainability. In 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA (pp. 1-7). IEEE.

[9] Moustakidis, S., Papandrianos, N. I., Christodolou, E., Papageorgiou, E., & Tsaopoulos, D. (2020). Dense neural networks in knee osteoarthritis classification: a study on accuracy and fairness. Neural Computing and Applications, 1-13.

[10] McCabe, P. G., Olier, I., Ortega-Martorell, S., Jarman, I., Baltzopoulos, V., & Lisboa, P. (2019). 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.

[11] Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. E., & Moons, K. G. M. (2003). External validation is necessary in prediction research:: A clinical example. Journal of clinical epidemiology, 56(9), 826-832.