A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients

Charis Ntakolia
Department of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
cntakolia@uth.gr

Christos Kokkotis
Institute for Bio-Economy & Agri-Technology
Center for Research and Technology Hellas
Volos, Greece
c.kokkotis@certh.gr
&
Department of Physical Education & Sport Science
University of Thessaly
Trikala, Greece
chkokkotis@gmail.com

Serafeim Moustakidis
AIDEAS OÜ
Narva mnt 5,
Harju maakond, Estonia
s.moustakidis@aideas.eu

Dimitris Tsaopoulos
Institute for Bio-Economy & Agri-Technology
Center for Research and Technology Hellas
Volos, Greece
d.tsaopoulos@certh.gr

***Abstract*—Osteoarthritis is the common form of arthritis in the knee (KOA). It is identified as one of the main causes of pain leading even to disability. To exploit the continuous increase in medical data concerning KOA, various studies employ big data and Artificial Intelligence analytics for KOA prognosis or treatment. However, most of the studies are limited to either specific groups of patients or specific groups of features, such as MRI, X-ray images or questionnaires. In this study, a machine learning pipeline is proposed to predict knee joint space narrowing (JSN) in KOA patients. The proposed methodology, that is based on multidisciplinary data from the osteoarthritis initiative (OAI) database, employs: (i) a clustering process to identify groups of people with progressing and non-progressing JSN; (ii) a robust feature selection process consisting of filter, wrapper and embedded techniques that identifies the most informative risk factors that contribute to JSN prediction; and (iii) a decision making process based on the evaluation and comparison of various classification algorithms towards the selection and development of the final prediction model for JSN. The evaluation was conducted with respect to model's overall performance, robustness and highest achieved accuracy. A 78.3% and 77.7% accuracy were achieved in left and right leg by Logistic Regression on the group of the 164 risk factors and SVM on the group of the 88 and 90 risk factors, respectively.**

Keywords—machine learning, knee osteoarthritis, joint space narrowing prediction, feature selection, interpretation

I. Introduction

Knee Osteoarthritis (KOA) has a higher prevalence rate compared with other types of OA. KOA results from a multifactorial, complex interplay of mechanical and constitutional factors, including mechanical forces, local inflammation, joint integrity, biochemical processes and genetic predisposition. Furthermore, KOA is closely associated with obesity, age and injuries [1]. The main consequences of the specific disease are low quality of life, low levels of psychology and social exhaustion due to low public participation [2]. Due to the multifactorial nature of KOA, disease pathophysiology is still poorly understood, and prediction and diagnosis tools are under current investigation.

KOA disease is a challenge for the scientific community, either in prognosis or in treatment. Increasing data collection has led to an increasing number of studies employing big data and Artificial Intelligence analytics applied in the KOA research. According to the literature review, several techniques have been reported in the literature in which Machine learning (ML) models were used to predict the development of KOA [3].

Lazzarini et. al. [4] examined the contribution of different variables (including biomarkers) within the predictive models in overweight and obese women by using five different outcome measures of incident knee OA, including medial joint space narrowing (JSN). The goal of this paper was to discover and analyse the role of novel biomarkers in KOA. In another study, Halilaj et. al. [5] used self-reported knee pain and radiographic assessments of joint space narrowing to characterize different clusters of OA progression and build models for early prediction of them by using. Furthermore, Pedoia et. al. [6] used MRI and biomechanics multidimensional with aim to set up a multidimensional platform for improving OA outcome prediction and patient sub stratification. This approach was the first, which provided large-scale integration of compositional imaging and skeletal biomechanics. Abedin et. al. [7] used Kellgren and Lawrence grading scheme to investigate if the prediction accuracy of a statistical model based on patient's questionnaire data is comparable to the prediction accuracy based on X-ray image. The accuracies were comparable for these two approaches and suggested as future work a model based on both patient's questionnaire data and X-ray images. In addition, Widera et. al. [8] investigated a multi-classifier problem for the prediction of KOA progression by using clinical data and X-ray image assessment metrics, where different algorithms and learning process configurations were investigated. The proposed method reduces by 20–25% the number of patients who show no progression. Hence, there is a need for further studies and development of techniques for determining risk factors that lead to the development of reliable tools for predicting KOA.

Risk factors identification for the prediction of KOA progression has been limited by an absence of non-invasive methods to inform clinical decision making and enable early detection of people who are most likely to progress to severe KOA.

2471-7819/20/$31.00 ©2020 IEEE
DOI 10.1109/BIBE50027.2020.00158

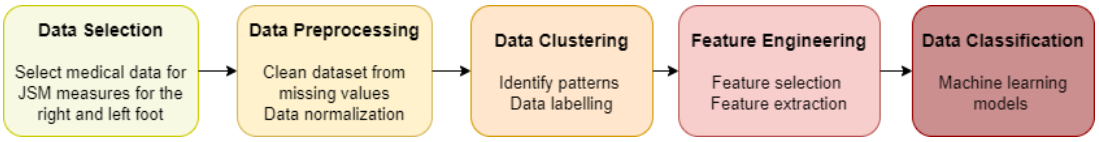


Figure 1. Methodology Flowchart.

However, most of the studies are limited to: (i) specific group of patients, such as overweight and obese women; (ii) single data modalities and most frequently on image-based algorithms utlizing MRI or X-ray images; (iii) questionnaires leading to biased results due to objectivity of patients, such as pain grade and (iv) a small number of factors that are linked with joint space narrowing (JSN) in KOA patients. This paper contributes to the prediction of joint space narrowing on Medial compartment (JSM) progression through the application of a novel machine learning pipeline that is capable of handling a large number of heterogeneous features coming from different feature categories. Clustering was initially applied to identify groups of patients with and without JSM progression, whereas identification of risk factors was based on a robust feature selection technique via a common voting system. The final prediction task was implemented using well-known ML models in an extensive comparative experimentation.

The paper is organized as follows. Section II gives a description of the medical dataset that was used in our paper. In Section III, the proposed methodology along with the necessary data pre-processing, feature selection and validation mechanisms, are presented. Results are given in Section IV. Conclusions and future work are finally drawn in Section V.

II. Medical Data

Data were obtained from the osteoarthritis initiative (OAI) database (available upon request at https://nda.nih.gov/oai/). Specifically, the current study only includes clinical data from baseline from all individuals without or being at high risk to develop KOA in at least one knee. Data from the baseline (625 features in total) from nine feature categories were considered as possible risk factors for the prediction of JSN as shown in Table 1. Clustering was performed on the JSM progression (especially using the variables V00XRJSM, V01XRJSM, V03XRJSM, V05XRJSM, V06XRJSM of the OAI from the first five visits) to group patients into two clusters (non-progressing patients and those whose JSN changes over time).

TABLE I. Maim categories of the feature subsets considered in the proposed methodology.

Category	Description
Anthropometrics	Variables that describe measurements of participants including height, weight, BMI, abdominal circumference etc
Behavioral	Variables which describe the participants' social behaviour
Quality of life	Questionnaire results regarding the quality level of daily routine
Medical history	Questionnaire data regarding a participant's arthritis-related and general health histories and medications
Medical imaging outcome	Variables which contain medical imaging outcomes (e.g. osteophytes and joint space narrowing)
Nutrition	Variables which collected using the modified Block Food Frequency questionnaire
Physical activity	Questionnaire results regarding leisure activities etc
Physical exam	Variables which contain physical measurements of participants, including isometric strength, knee and hand exams, walking tests, and other performance measures
Symptoms	Questionnaire results regarding arthritis symptoms and general arthritis or health-related function and disability

III. Methodology

To discover knowledge from medical data a common approach is to target the correct data, transform them via a preprocessing stage and recognize patterns in order to extract knowledge. Data mining techniques that are commonly used in medical data and healthcare industry are association, clustering and classification [9].

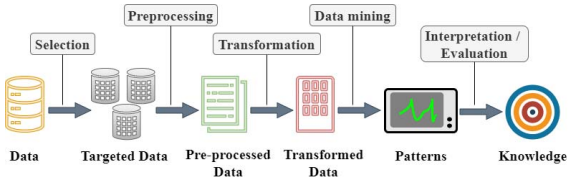


Figure 2. Stages of knowledge discovery process.

A similar approach (Figure 2) was developed in our study by taking advantage of the combination of descriptive and predictive techniques, such as clustering, feature selection and classification. The proposed machine learning methodology for predicting JSN consists of 4 main steps: (i) Data pre-processing; (ii) Data clustering; (iii) Feature Selection; and (iv) Data Classification. In the first step, data cleaning and normalization are performed to remove noise and bring all the variables to the same range. The normalized data are then clustered based on the JSM measures for the left and the right leg, respectively. Thus, the selection and extraction of features are realized based on the identified clusters (that are considered as classes in our case). Consecutively, the selected features are used to develop prediction models for KOA progression of patients (Figure 1).

A. Data pre-processing

In this step, data cleaning was performed by excluding the columns with more than 20% missing values compared to the total numbers of samples. For the rest data, data imputation was implemented to replace missing values of the categorical or numerical variables by the mode (most frequent value) of the non-missing variables. Furthermore, standardization of a dataset is a common requirement for many ML estimators. In our paper, data was normalized to [0, 1] to build a common basis for the feature selection algorithms that follow.

B. Data Clustering

Clustering divides a large dataset into a small number of groups called clusters. The elements of a cluster present similar characteristics called features. Hence, clustering algorithms collects the data so that the elements within a cluster to be more identical than with elements on the other clusters [10].

The most commonly used types of clustering algorithms are the centroid-based, the connectivity-based and the distribution-based clustering methods. Centroid-based clustering is a widely used technique within unsupervised

learning algorithms in many research fields. The success of any centroid-based clustering relies on the choice of the similarity measure under use [11]. Traditional approaches include K-Means [12] and K-Medoids [13]. The connectivity-based or hierarchical clustering is based on a recursive partitioning of data into clusters. If the hierarchy of the clusters follows a bottom-up approach then the strategy is called agglomerative while the divisive strategy follows a top-down approach [14]. The distribution-based clustering methods use the distribution of sequences across data to cluster the dataset [15]. Commonly used approaches are also the Gaussian mixture models [16].

In the proposed methodology, we followed for each leg the clustering process that is illustrated in Figure 3 and described with the pseudo-algorithm in Algorithm 1. Initially, for each patient $p \in \mathcal{P}$, where $\mathcal{P}$ is the set of the patients included in the cleaned medical data, we calculated the differences between the consecutive measurements: $d_j^p = m_j^p - m_i^p, \forall i \in [1, \dots, n-1],\ j \in [2, \dots, n]$, where m_i is the i^{th} JSM measurement and n the number of the measurements performed for the leg under examination, such as left or right. In our case study we had 5 measurements for each leg that correspond to the JSM progression during the first 5 visits. The absolute sum of the differences was calculated: $\sum_{k=2}^{n} \left| d_k^p \right|$ forming an indicator of the overall JSM progression within the first 5 visits. These values were used for performing the clustering.

For the clustering, the centroid-based, the connectivity-based and the distribution-based clustering methods, such as K-Means [12], K-Medoids [13], Hierarchical clustering [17] and Gaussian mixture models [16], were employed, whereas the Davies Bouldin index [18] was used to evaluate the optimal number of clusters. However, further investigation should be done according to the clustered data in order to determine the final number of clusters that will be adopted (2 cluster in our case as described in section IV-B). Data resampling was finally employed to cope with the class imbalance problem. Specifically, the size of majority cluster was reduced in order to have the same number sample as the minority one.

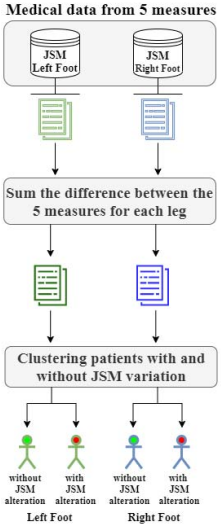


Figure 3. Clustering process of the proposed methodology.

Algorithm 1. Pseudoalgorithm of the clustering process

Input: medical data with JSM measurements
Output: labeled data

1. **For each** patient $p \in \mathcal{P}$:
   Calculate the differences between the consecutive JSM measurements: $d_j^p = m_j^p - m_i^p, \forall i \in [1, \dots, n-1], j \in [2, \dots, n]$
   Calculate the sum of the absolute differences: $\sum_{k=2}^{n} \left| d_k^p \right|$
End
2. For each clustering method m examined
   Perform clustering evaluation with Davies Boulding index and calculate the optimal number of clusters C_m.
   Perform clustering with C_m clusters
End
3. Return labels and evaluate the clustered data

C. *Feature Engineering*

To avoid bias, a robust feature selection methodology was employed that combined the outcomes of six FS techniques: two filter algorithms (Pearson correlation and Chi-2), one wrapper (with Logistic Regression) and three embedded ones (Logistic regression L2, Random Forest and LightGBM). Feature ranking was decided on the basis of a majority vote scheme. Specifically, we performed all six FS techniques separately, each one resulting into a selected FS. A feature receives a vote every time it has been selected by one of the FS algorithms. We finally ranked all features with respect to the votes received.

A short overview of the feature selection algorithms investigated in this research work, is given in what follows.

1) *Filter algorithms*:

a) *Pearson Correlation* is the most important correlation factor being based on the concept of linear relationship. If there is a linear dependence between two features, then their correlation coefficient is ± 1. If there is no dependence, the correlation coefficient is 0. However, if two variables are highly correlated among themselves, they provide redundant information regarding the target. Consequently, the second variable doesn't add additional information, so removing it can help to reduce the dimensionality. In this approach, we set the maximum number of the selected features to be 30 [19].

b) *Chi-squared* independence test [20] was applied to examine the relationship between two quality variables. The Chi-squared statistical test also works manually with non-negative numerical and quantitative characteristics. The specific test compares the degree of agreement (or correlation) between the theoretical frequency and the actual frequency. The algorithm was decided to terminate at 30 selected features. The termination criterion was manually selected after a trial-and-error exploration process.

2) *Wrapper*:

Recursive Feature Elimination (RFE) [21] is a greedy optimization algorithm which aims to find the best performing feature subset. In each iteration, it creates models and keeps aside the best or the worst performing feature. Each next model includes reduced number of features until all the features are exhausted. At the end, it ranks the features based on the order of their elimination. In this approach, the logistic regression classifier was selected to drive the elimination process whereas the termination criterion was also set to 30 features.

3) *Embedded*:

a) *Logistic regression (L2 penalty)* is an embedded method relying on regularized logistic regression models. Furthermore, this approach is based on small subsets of the full feature space by sampling at random this space. The sampling probability depends on the estimated feature relevance. In addition, the initial relevance of each feature is estimated according to a t-test ranking [22].

b) *Random forests* are a popular method for feature ranking, due to the fact that they require very little feature engineering and parameter tuning. But they come with their own limitations, especially when data interpretation is concerned. In high correlated data, strong features can end up with low scores and the method can be biased towards variables with many categories. In addition, this method rearranges stochastically all values of the features for each tree and uses the RF model to predict this permuted feature [23].

c) *Light GBM* is a gradient boosting framework that uses tree-based learning algorithm. Specifically, Gradient-based One-Side Sampling and Exclusive Feature Bundling are used to deal with large number of data instances and large number of features. Light GBM can handle the large size of data, it takes lower memory to run and is faster than Gradient Boosting Decision Tree [24].

D. Data Classification

Classification consist of two steps: training and testing. During training a classification model is built based on the collected training data for generating classification rules. The accuracy of the model is estimated with the test data [25].

Various classification algorithms were tested to identify the optimum model that achieves the higher accuracy on the test data. Gradient Boosting model (GBM) belongs in the family of the decision trees. GBM identifies and uses weak learners to produce strong learners through an additive, gradual and sequential process. A modified version of the initial training data set is fitted to develop a new tree. Let $\{(x_i, y_i)\}_{i=1}^n$ be the training set, $L(y, F(x))$ be the loss function and M be the number of iterations. The Algorithm 2 presents the pseudoalgorithm of the GBM method [26], [27].

Algorithm 2. GBM Algorithm

1. Model's initialization: $F_0(\boldsymbol{x}) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$
2. **For** $m = 1$ *to* M:
   Compute the pseudo-residuals:
   $$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial x}\right]_{F(x)=F_{m-1}(x)}, for\ i = 1, \dots, n$$
   Fit a base learner (or wake learner) $h_m(\boldsymbol{x})$ to pseudo-residuals
   Compute multiplier γ_m by solving the 1-dimensional optimization problem: $\gamma_m = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$
   Update the model: $F_m(\boldsymbol{x}) = F_m(\boldsymbol{x}) + \gamma h_m(\boldsymbol{x})$
   End
3. Retum $F_m(\boldsymbol{x})$

Logistic Regression (LR) is a mathematical model that describes the relationship of data to a dichotomous dependent variable. The model is based on the logistic function, $f(x) = \frac{1}{1+e^{-x}}, where\ x \in (-\infty, +\infty)$ and $0 \le f(x) \le 1$. Thus, regardless the value of x the model is designed to describe the data with a probability in the range of 0 and 1 in a A-shaped graph [28].

Multilayer Perceptron (MLP), belongs in the category of Artificial Neural Networks (ANN) and it is the most common neural network. MLP is based on a supervised training procedure to generate a nonlinear model for prediction. MLP consists of layers, such as the input layer, output layer and hidden layers. Thus, MLP is a layered feedforward neural network where the information is transferred unidirectionally from the input layer to output layer through the hidden layers (Figure 4). In Figure 4a, a simple neuron perceptron is presented with single layer where all inputs connect with only one output. Let $\{x_i\}_{i=0}^n$ be the input, such as features or variables, and $\{w_i\}_{i=0}^n$ be the weights of the neuron. The weighting step consists of three steps: (i) the multiplication of features with the corresponding weight, $\{x_i w_i\}_{i=0}^n$; (ii) their sum, $z = \sum_{i=0}^n x_i w_i$, and (iii) the transfer step where the output y is produced by the application of an activation function f to the sum, $y = f(z)$. Commonly used activation or transfer functions are the unit step (Heaviside), linear or logistic (sigmoid) [29].

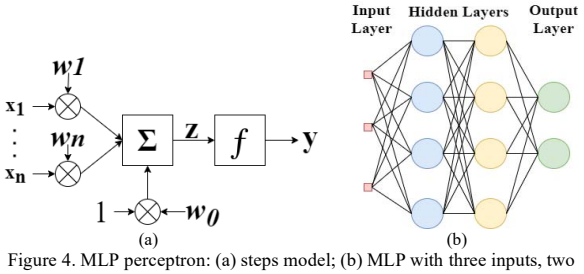


Figure 4. MLP perceptron: (a) steps model; (b) MLP with three inputs, two hidden layers, and two outputs.

Naïve Bayes Gaussian (NBG), is a probabilistic classifier employing the Bayes theorem with strong independence assumptions between the variables / features given the class (Figure 5). NBG adopts the assumption that the data follows the Gaussian (normal) distribution. The process of classifying an instance $\{x_i\}_{i=1}^n$ consists of selecting the class with the highest a posteriori probability, $P(c|\boldsymbol{x})$. The factorization of the joint probability is given by: $P(c|\boldsymbol{x}) \propto P(c) \prod_{i=1}^n p(x_i|c)$, where $p(x_i|c) \sim \mathcal{N}(\mu_i^c, \sigma_i^c)$ is the real joint distribution [30], [31].

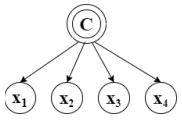


Figure 5. Example Naive Bayes classifier

Random Forest (RF) is an ensemble learning method based on decision trees. RF constructs a large number of decision trees. Each decision tree denotes a class prediction and the class with the most votes represents the model's prediction. Algorithm 3 shows the pseudoalgorithm of RF method [32], [33].

Algorithm 3. Breiman's Random Forest

Input: Training set $\mathcal{D}_n$, the number of trees $M > 0$, $a_n \in \{1, \dots, n\}$, `mtry`$\in \{1, \dots, p\}$, `nodelize`$\in \{1, \dots, a_n\}$ and $\boldsymbol{x} \in \mathcal{X}$
Output: Prediction of the random forest at $\boldsymbol{x}$

1. **For** $j = 1$*to* M:
   Select a_n points, with (or without) replacement, uniformly in $\mathcal{D}_n$ to be used in the following steps.
   Set $\mathcal{P} = (\mathcal{X})$ the list containing the cell associated with the root of the tree
   Set $\mathcal{P}_{final} = \emptyset$ an empty list
   While $\mathcal{P} \neq \emptyset$:
      Let A be the first element of $\mathcal{P}$
      If A contains less than `nodesize` points or if all $\boldsymbol{X}_i \in A$ are equal
         Remove the cell A from the list $\mathcal{P}$
         $\mathcal{P}_{final} \leftarrow Concatenate(\mathcal{P}_{final}, A)$
      Else
         Select uniformly, without replacement, a subset $\mathcal{M}_{try} \subset \{1, \dots, p\}$ of cardinality `mtry`
         Select the best split in A by optimizing the CART-split criterion along the coordinates in $\mathcal{M}_{try}$.
         Cut the cell A according to the best split. Call A_L and A_R the two resulting cells
         Remove the cell A from the list $\mathcal{P}$
         $\mathcal{P} \leftarrow Concatenate(\mathcal{P}, A_L, A_R)$
      End
   End
   Compute the predicted value $m_n(\boldsymbol{x}; \Theta_j, \mathcal{D}_n)$ at $\boldsymbol{x}$ equal to the average of the Y_i falling in the cell of $\boldsymbol{x}$ in partition $\mathcal{P}_{final}$.
End

2. Compute the random forest estimate $m_M(\boldsymbol{x}; \theta_1 \dots \theta_M, \mathcal{D}_n)$ at the query point $\boldsymbol{x}$ according to: $m_M(\boldsymbol{x}; \theta_1 \dots \theta_M, \mathcal{D}_n) = \frac{1}{M}\sum_{j=1}^{M} m_n(\boldsymbol{x}; \theta_j, \mathcal{D}_n)$
3. Return the prediction of the random forest at $\boldsymbol{x}$

Support vector machine (SVM) is another supervised learning model. SVM is based on the statistical learning framework, called VC theory, proposed by Vapnik and Chervonekis [34], [35]. SVM targets to create a decision boundary, the hyperplane, between two classes that enables the prediction of labels from one or more feature vectors, such that the distance between the closest points of each class, called support vectors, and the hyperplane to be maximized (Figure 6). In Figure 6, the labeled training set consists of two classes the blue and the red one. Let $\{(x_i, y_i)\}_{i=1}^{n}$ with $x_i \in \mathbb{R}^d$ and $y_i \in (-1,1)$ be the feature vector representations and the class label, respectively. The optimal hyperplane can be defined as: $\boldsymbol{w}\boldsymbol{x}^T + \boldsymbol{b} = 0$, where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x}$ the input feature vector and b the bias. The $\boldsymbol{w}$ and $\boldsymbol{b}$ satisfy the following inequalities: $wx_i^T + b \geq \begin{cases} 1, if\ y_i = 1 \\ -1, if\ y_i = -1 \end{cases}$. The objective is to find the values of $\boldsymbol{w}$ and $\boldsymbol{b}$ so that the hyperplane separates the data and maximizes the margin $1/\|\boldsymbol{w}\|^2$ [36].

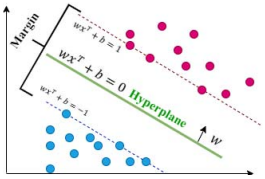


Figure 6. Example of linear SVM with two classes (blue and red).

IV. Evaluation

A. Evaluation Methodology

The proposed methodology was applied on the case study for the prediction of JSN in patients with KOA and it was implemented using the dataset presented in Section II.

The preprocessed data for the right and left leg was clustered and the groups of patients with/without JSM progression within the dataset were identified. To accomplish this task, K-Means, K-Medoids [13], Hierarchical clustering and Gaussian mixture models were employed in a comparative analysis. Davies Bouldin index [18] was used to evaluate the optimal number of clusters in order to discriminate patients into groups but also to identify the magnitude of the variation in the patients' JSM measures. The parameters of the proposed clustering methods are presented In TABLE II.

TABLE II. Parameter settings for clustering methods.

Clustering Method	Parameters
K-Means	City block distance, 5 replicates
K-Medoids	City block distance, 5 replicates
Hierarchical	Agglomerative cluster tree, Chebychev distance, farthest distance between clusters, 3 maximum number of clusters
Gaussian mixture models	using the Expectation-Maximization algorithm

To predict the JSN in KOA patients, six different prediction models were employed and compared separately at each leg, including Gradient Boosting, Logistic Regression, MLP, Naïve Bayes Gaussian, Random Forest and SVM. The data set was split to 70% for training set and 30% for testing set. The evaluation of the models was performed on the presented medical dataset (Section II). For most of the models, hyper parameter tuning (TABLE III) was realized with grid search and 3-fold cross validation. The test size was set to 30% with normalization upon the features. The models were evaluated in feature subsets of increasing dimensionality from 5 to 155 features with a step size of 5, from 155 to 325 features with a step size of 10, from 325 to 415 features with a step size of 15, from 415 to 475 features with a step size of 20 and from 475 to 625 features with a step size of 25.

TABLE III. Hyper parameter settings for tuning.

Classification Model	Hyper parameters tuning
Gradient Boosting	The number of boosting stages to perform from 10 to 500 with 10 step size The maximum depth of the individual regression estimators from 1 to 10 with 1 step size The minimum number of samples required to split an internal node: 2, 5 and 10 The minimum number of samples required to be at a leaf node: 1, 2 and 4 The number of features to consider when looking for the best split: $\sqrt{n_{features}}$ or $\log_2(n_{features})$
Logistic Regression	The inverse of regularization strength was tested on 0.001, 0.01, 0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 Algorithm to use in the optimization problem was set to 4 different solvers that handle L2 or no penalty, such as 'newton-cg', 'lbfgs', 'sag' and 'saga' A binary problem is fit for each label or the loss minimized is the multinomial loss fit across the entire probability distribution, even when the data is binary With and without reusing the solution of the previous call to fit as initialization
MLP	Hidden layers: $(10, 50, 100)$, $(50, 100, 150)$ and $(100, 200, 400)$ Activator function: Relu and $tanh$ Solver for weight optimization: stochastic gradient descent, stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba and an optimizer in the family of quasi-Newton methods L2 penalty (regularization term) parameter: 0.0001 and 0.05 The learning rate schedule for weight updates was set as a constant learning rate given by the given number and as adaptive by keeping the learning rate constant to the given number as long as training loss keeps decreasing.
Naïve Bayes Gaussian	-
Random Forest	The number of trees in the forest from 10 to 500 with 10 step size The maximum depth of the tree from 1 to 10 with 1 step size The minimum number of samples required to split an internal node: 2, 5 and 10 The minimum number of samples required to be at a leaf node: 1, 2 and 4 The number of features to consider when looking for the best split: $\sqrt{n_{features}}$ or $\log_2(n_{features})$ With and without bootstrap
SVM	The regularization parameter was tested on 0.001, 0.01, 0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 Kernel type was set to linear, polynomial, sigmoid and radial basis functions

B. Results

1) *Clustering results*:

Regarding the clustering process, 4 clusters (TABLE IV) were identified in most of the cases representing patients with zero, low, medium and high alterations in JSM measures, respectively. A class size imbalance problem was observed with Cluster 1 being significantly bigger than the rest three clusters for both left and right legs (Figure 7a,b). We should notice that for the right leg some patients with low JSM alterations were erroneously grouped in the cluster with the stable or non-infected patients. To overcome these problems, we decided to perform clustering with only 2 clusters. Among the four potential clustering approaches, we adopted the results from the K-Means method since a better discrimination among the patient groups was achieved. From the resulted clusters, the large one represents the patients with stable JSN or the patients that did not present KOA in their left and/or right leg while the second one represents patients

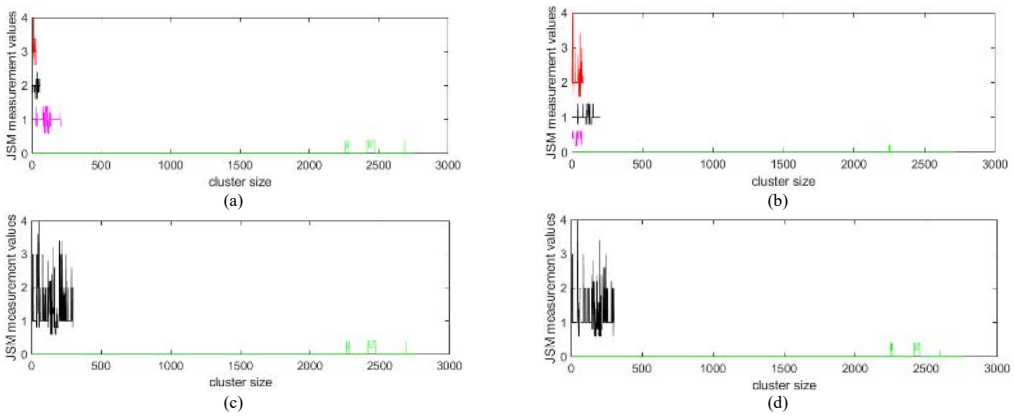


Figure 7. Clustering with K-Means of (a) left leg with Davies Bouldin index; (b) right leg with Davies Bouldin index; (c) left leg with 2 clusters; (d) right leg with 2 clusters.

with JSM alterations (Figure 7c,d). TABLE V cites the results of the finally employed clustering approach.

TABLE IV. Clustering of left and right leg with Davies Bouldin index.

Clustering Method	Number of Clusters		Cluster Elements	
	Left	*Right*	*Left*	*Right*
K-Means	4	4	[2763,209,62,28]	[2733,199,84,50]
K-Medoids	4	4	[2763,209,62,28]	[2733,199,84,50]
Hierarchical	4	3	[2960,74,24,4]	[2989,68,9]
Gaussian mixture models	4	4	[2960,74,24,4]	[2783,206,68,9]

TABLE V. Clusters identified in our study

Clustering Method	Number of Clusters		Cluster Elements	
	Left	*Right*	*Left*	*Right*
K-Means	2	2	**[2763,299]**	**[2764,302]**
K-Medoids	2	2	[2763,299]	[2764,302]
Hierarchical	2	2	[3034,28]	[2989,77]
Gaussian mixture models	2	2	[3034,28]	[2989,77]

2) *Feature selection results*:

Figure 8 shows the first 100 features selected by the proposed FS approach for the left and the right knee. The following conclusions could be drawn from the analysis of Figure 8: (i) Symptoms and medical imaging outcomes seem to be the most informative feature categories. A feature from the symptoms' category was selected first followed by three imaging outcomes in both left and right knees. In total, almost half of the first 40 selected features in both legs come from either the symptoms or the imaging outcomes category; (ii) Nutrition and physical exam outcomes were also proved to be contributing risk factors since approximately 20 out of the 100 features were selected for each one of these two categories; (iii) Anthropometrics and medical history features were also among the feature categories that contribute in the JSM prediction (~5 anthropometrics and 4 medical history indexes were selected among the first 50 features in both knees); (iv) Overall, it was concluded that a combination of heterogeneous features coming from almost all feature categories is needed to predict JSM highlighting the necessity of adopting a multi-parametric approach that could handle the complexity of the available data.

3) *Classification results:*

The proposed classification process was applied to the left and right leg, separately. For the left leg, the Logistic Regression model outperformed the others by achieving the maximum accuracy (≅ 77.7%) for 165 features (TABLE VI, Figure 9). However, SVM and MLP presented a comparative performance. To identify the optimal number of features that maximizes the accuracy, we tested the 2 models that have the best performance in the range of 155 and 175 features (Figure 10), namely Logistic Regression and SVM. The Logistic

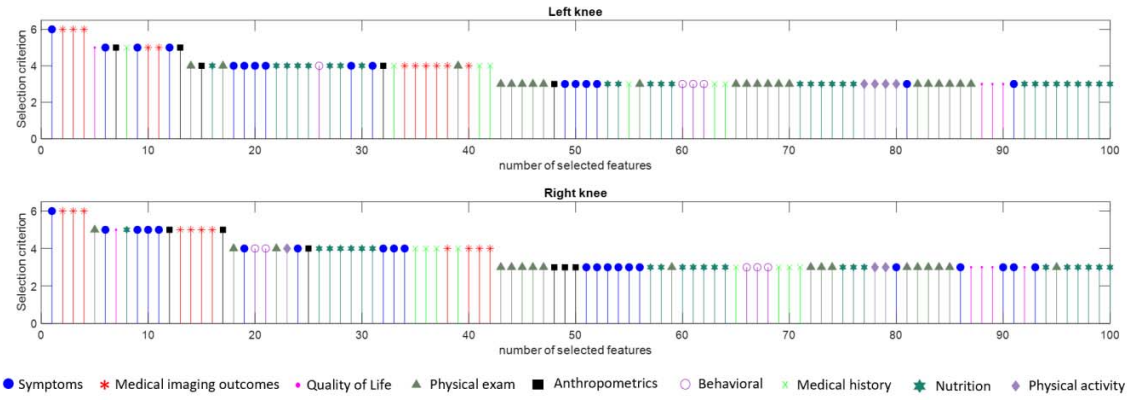


Figure 8. The first 100 features selected for the left (top) and the right knee (down).

Regression model was proved to have the best performance (≅ 78.3%) for 164 features with inverse of regularization strength to 1, without class weight, dual, nor L1 ratio. The maximum number of iterations was 100 and the intercept scale was 1 while the penalty was set to L2. The Newton-cg solver was chosen with reuse of the previous solution as initial one and tolerance at 0.0001.

TABLE VI. Maximum, minimum and mean accuracy of prediction models over the tested set for the left leg.

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
MLP	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

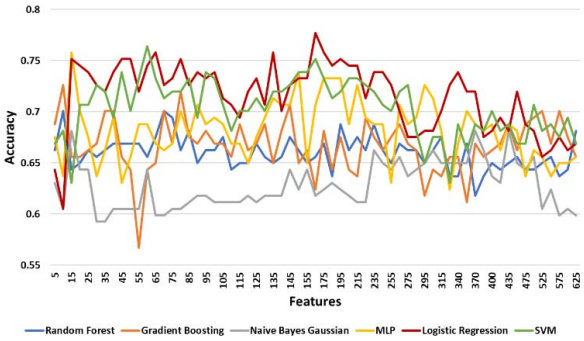


Figure 9. The accuracy of prediction models over test set for various number of features for the left leg.

Figure 10. The accuracy of Logistic Regression and SVM for 155 to 175 features over the test set for left leg.

For the right leg a similar approach was adopted. The SVM model outperformed the rest achieving the maximum accuracy (≅ 77.7%) for 90 features (TABLE VII). However, the MLP and Logistic Regression models presented a comparative performance (Figure 11). The Logistic Regression model achieved the higher mean accuracy (≅ $70.7\% \pm 0.036$) with lower standard deviation from the SVM (≅ $68.6\% \pm 0.039$) (TABLE VII). To this end, these two models were re-evaluated in a neighborhood, $\mathcal{U}_{SVM}(90,5)$ and $\mathcal{U}_{LR}(185,10)$, where they reached their maximum accuracy. Logistic Regression reached its best performance at 185 and 188 features with ≅ 77.1% while the SVM model accomplished a maximum 77.7% accuracy with 88 and 90 features (Figure 12). The hyperparameters of the SVM model with the best performance are the following: the regularization parameter was set at 0.1, the cache size at 200 and the tolerance at 0.001. Also, a linear kernel was chosen.

TABLE VII. Maximum, minimum and mean accuracy of prediction models over the tested set for the right leg.

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
MLP	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

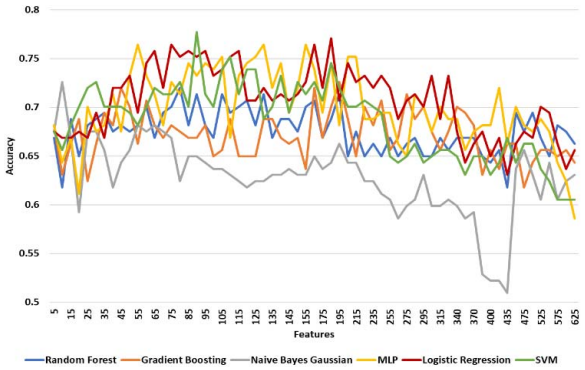
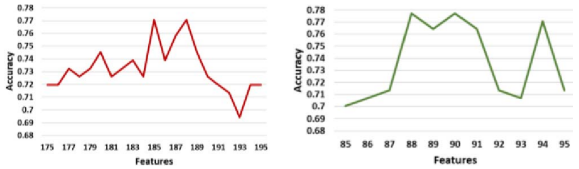


Figure 11. The accuracy of prediction models over test set for various number of features for the right leg.



(a) (b)

Figure 12. The performance evaluation of (a) Logistic Regression in the range of 175-195 features and (b) SVM in the range of 85-95 features.

V. Conclusion and Future Work

In this study the problem of JSN prediction in knee osteoarthritis patients was investigated. To this end, a machine learning pipeline was proposed for the accurate prediction of JSM progression trained on a multidisciplinary pool of heterogenous factors from the OAI database (625 features in total). To facilitate the leaning process, clustering was initially performed on the JSM measures of patients over the first five visits in order to identify and group patients with and without JSN progression. Then, a hybrid feature selection technique was employed to identify those features (only from the baseline) that contribute significantly to the discrimination of patients belonging to the identified clusters (progressing versus non-progressing patients). The selected features were finally used to train various ML models for predicting JSM in KOA patients. The comparison and evaluation of the models showed that for the left leg the LR model achieved the best performance with 78.3% accuracy for 164 features, while for the right leg the SVM model dominated with 77.7% accuracy for 88 and 90 features.

Apart from developing the prediction models, this study also revealed insights with respect to the nature of most important predictive risk factors. The main finding was that a combination of heterogeneous features coming from almost all feature categories is needed to maximize the performance of the predictive models.

Future work includes the incorporation of image-based deep learning algorithms to extract morphological knee features as an additional source of information that could further increase the accuracy of the proposed predictive ML models.

ACKNOWLEDGMENT

This work has received funding from the European Community's H2020 Programme, under grant agreement Nr. 777159 (OACTIVE).

REFERENCES

[1] V. Silverwood, M. Blagojevic-Bucknall, C. Jinks, J. Jordan, J. Protheroe, and K. Jordan, "Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis," *Osteoarthritis and cartilage*, vol. 23, no. 4, pp. 507–515, 2015.

[2] M. J. Lespasio, N. S. Piuzzi, M. E. Husni, G. F. Muschler, A. Guarino, and M. A. Mont, "Knee osteoarthritis: a primer," *The Permanente Journal*, vol. 21, 2017.

[3] C. Kokkotis, S. Moustakidis, E. Papageorgiou, G. Giakas, and D. Tsaopoulos, "Machine Learning in Knee Osteoarthritis: A Review," *Osteoarthritis and Cartilage Open*, p. 100069, 2020.

[4] N. Lazzarini, J. Runhaar, A. Bay-Jensen, C. Thudium, S. Bierma-Zeinstra, Y. Henrotin, and J. Bacardit, "A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women," *Osteoarthritis and Cartilage*, vol. 25, no. 12, pp. 2014–2021, 2017.

[5] E. Halilaj, Y. Le, J. L. Hicks, T. J. Hastie, and S. L. Delp, "Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative," *Osteoarthritis and cartilage*, vol. 26, no. 12, pp. 1643–1650, 2018.

[6] V. Pedoia, J. Haefeli, K. Morioka, H.-L. Teng, L. Nardo, R. B. Souza, A. R. Ferguson, and S. Majumdar, "MRI and biomechanics multidimensional data analysis reveals R2-R1ρ as an early predictor of cartilage lesion progression in knee osteoarthritis," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 1, pp. 78–90, 2018.

[7] J. Abedin, J. Antony, K. McGuinness, K. Moran, N. E. O'Connor, D. Rebholz-Schuhmann, and J. Newell, "Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[8] P. Widera, P. M. Welsing, C. Ladel, J. Loughlin, F. P. Lafeber, F. P. Dop, J. Larkin, H. Weinans, A. Mobasheri, and J. Bacardit, "Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data," *Scientific Reports*, vol. 10, no. 1, pp. 1–15, 2020.

[9] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," *International Journal of Information*, vol. 6, no. 1/2, pp. 53–60, 2016.

[10] K. Sharmila and S. Vethamanickam, "Survey on data mining algorithm and its application in healthcare sector using Hadoop platform," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, no. 1, pp. 567–571, 2015.

[11] A. Sarmiento, I. Fondón, I. Durán-Dıaz, and S. Cruces, "Centroid-based clustering with α$β-divergences," *Entropy*, vol. 21, no. 2, p. 196, 2019.

[12] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," 1997.

[13] L. K. P. J. RDUSSEEUN and P. KAUFMAN, "Clustering by means of medoids," 1987.

[14] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu, "Hierarchical clustering: Objective functions and algorithms," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2018, pp. 378–397.

[15] Z. Yu, X. Zhu, H.-S. Wong, J. You, J. Zhang, and G. Han, "Distribution-based cluster structure selection," *IEEE transactions on cybernetics*, vol. 47, no. 11, pp. 3554–3567, 2016.

[16] D. A. Reynolds, "Gaussian Mixture Models.," *Encyclopedia of biometrics*, vol. 741, 2009.

[17] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[18] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 3, pp. 301–315, 1998.

[19] J. Biesiada and W. Duch, "Feature selection for high-dimensional data—a Pearson redundancy based filter," in *Computer recognition systems 2*, Springer, 2007, pp. 242–249.

[20] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, 2017.

[21] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers," *Genome Research*, vol. 11, no. 11, pp. 1878–1887, 2001.

[22] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2, 1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[23] Q. Zhou, H. Zhou, and T. Li, "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features," *Knowledge-based systems*, vol. 95, pp. 1–11, 2016.

[24] E. Al Daoud, "Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset," *International Journal of Computer and Information Engineering*, vol. 13, no. 1, pp. 6–10, 2019.

[25] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[26] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[27] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," in *The elements of statistical learning*, Springer, 2009, pp. 337–387.

[28] D. G. Kleinbaum and M. Klein, "Logistic regression, statistics for biology and health," *Retrieved from DOI*, vol. 10, pp. 978–1, 2010.

[29] H. Taud and J. Mas, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*, Springer, 2018, pp. 451–455.

[30] A. Perez, P. Larranaga, and I. Inza, "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes," *International Journal of Approximate Reasoning*, vol. 43, no. 1, pp. 1–25, 2006.

[31] A. H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," in *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, 2017, pp. 209–212.

[32] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[36] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics-Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.