Prediction of pain in knee osteoarthritis patients using machine learning: Data from Osteoarthritis Initiative

1st Antonios Alexos
Department of Electrical and Computer Engineering
University of Thessaly
Volos, Greece
aalexos@uth.gr

2nd Christos Kokkotis
Institute of Bio-Economy and Agri-Technology
Center for Research and Technology Hellas
Volos, Greece
c.kokkotis@certh.gr

3rd Serafeim Moustakidis
AIDEAS OÜ
Narva mnt 5, Tallinn, Harju maakond, Estonia
s.moustakidis@aideas.eu

4th Elpiniki Papageorgiou
Faculty of Technology
University of Thessaly
Larissa, Greece
elpinikipapageorgiou@uth.gr

5th Dimitrios Tsaopoulos
Institute of Bio-Economy and Agri-Technology
Center for Research and Technology Hellas
Volos, Greece
d.tsaopoulos@certh.gr

***Abstract*—Knee Osteoarthritis(KOA) is a serious disease that causes a variety of symptoms, such as severe pain and it is mostly observed in the elder people. The main goal of this study is to build a prognostic tool that will predict the progression of pain in KOA patients using data collected at baseline. In order to do that we leverage a feature importance voting system for identifying the most important risk factors and various machine learning algorithms to classify, whether a patient's pain with KOA, will stabilize, increase or decrease. These models have been implemented on different combinations of feature subsets, 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.**

***Index Terms*—machine learning, knee osteoarthritis, pain prediction, feature selection, physical function, knee joint**

I. Introduction

Knee OA is the most common form of osteoarthritis [21]. KOA results from the degeneration of cartilage in the knee, which can happen due to aging, weight and injuries. Furthermore, it results from genetic predisposition, biochemical processes and mechanical forces. The certain disease process begins before pain, symptoms and motion restriction are noticed. Because of this and the fact that KOA is a progressive disease, getting a tool for early detection and prediction as early as possible is key [11]. Predicting this disease is a difficult task because it entails multifactorial causation, which is under investigation by the scientific community.

Increasingly data collection is a challenge for the scientific community and leads to the use of machine learning techniques to develop reliable tools for predicting KOA. According to the literature review several studies have shown that machine learning models are used to predict KOA [14]. In 2017, Lazzarini et. al. [15] investigated the contribution of different variables (including biomarkers) within the predictive models in overweight and obese women. The main aim of this work was to discover and analyse the role of novel biomarkers in KOA [15]. Furthermore, Halilaj et. al. [10] focused to characterize different clusters of OA progression and build models for early prediction of them by using self-reported knee pain and radiographic assessments of joint space narrowing. In another study, Pedoia et. al. [17] used MRI and biomechanics multidimensional data to fill the gap in multidimensional data analysis for prediction of KOA. This study was the first, which provided large-scale integration of compositional imaging and skeletal biomechanics. In 2019, Abedin et. al. [1] investigated 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. They presented comparable accuracies for these two approaches and suggested as future work a model based on both patient's questionnaire data and X-ray images. In another study the same year, Widera et. al. [23] used clinical data and X-ray image assessment metrics in a multi-classifier problem for prediction of KOA progression by using different algorithms and learning process configurations. Therefore, 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.

The purpose of this study is: (i) to identify different clusters of KOA pain progression, (ii) to identify informative parameters that are relevant with pain progression from a big pool of risk factors that are available in osteoarthritis initiative (OAI) database and (iii) to build ML models that can predict long-term pain progression using baseline data. To accomplish the aforementioned targets, we built a ML-

Identify applicable funding agency here. If none, delete this.

empowered methodology capable of achieving state-of-the-art accuracy results with the minimum possible number of features. By using a relatively small number of features, and at the same time not sacrificing test set performance, we can run the algorithm faster at inference time, and implement it in portable devices (e.g. a smartphone). The dataset, as described in Section II has 726 features. By reducing this number to a relatively small number features, for instance 25 we could create more possibilities for the implementation of an algorithm in a small mobile device; or test the algorithm faster in the subjects by requiring less computational power. In order to do this we have developed a hybrid technique, in which we derive the feature importance from different Feature Selection (FS) algorithms via a common voting system. Afterwards, we explored the suitability of different ML algorithms in an extensive comparative experimentation, to distinguish the one that produces the best results for our prediction.

The rest of the paper is organised as follows. Section II gives a short description of the data used in our study, whereas Section III presents the proposed methodology that comprises of the following parts: (i) Grouping/Labeling of the data, (ii) Feature Selection, (iii) Machine Learning models, (iv) Validation. Results and discussions are given in section IV and conclusions are drawn in the final section of the paper.

II. Dataset Description

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

III. Methodology

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

A. Grouping/Labeling

The available data was grouped into three clusters, each one representing a different pain progression condition:1) cluster 1: pain decline, 2) cluster 2: no significant pain change and 3) cluster 3: pain increase. To accomplish this, the following methodology was applied. WOMAC pain (as represented by the variables VxxWOMKPR and VxxWOMKRL for the right and left leg, respectively where xx the number of the visit) was collected for the first four (4) visits from each patient of the OAI dataset. We selected only the first four visits because a significant number of patients did not follow up after visit 4. Linear fitting([20]) was applied on the WOMAC pain progression data of each patient (Figure 1a) and the grouping was implemented based on the slope of the regression line. The thresholds applied on the calculated slopes were carefully selected towards the generation of equally sized, well-represented and non-overlapping clusters. Indicative pain progression data from each cluster are given in Figure 1 b) and c). The proposed clustering methodology was applied separately on each leg.

The rest of the ML components focus on the identification of the most important risk factors from the OAI dataset, described in Subsections III-B and III-C, and as well as the development of the ML models that could discriminate patients belonging to the three aforementioned clusters, described in Subsection III-D.

B. Data Pre-Processing

For the preprocessing of the data we followed two procedures. The dataset has a lot of missing of values so we used mode imputation to handle them([12]). We used this specific method because it is able to handle both numerical and non-numerical variables. We also standarized the features by subtracting the mean and scaling the values with respect to variance. This is a common requirement, because some ML algorithms behave badly, if the features are not normally distributed([7], [8]).

C. Feature Selection

A hybrid feature selection methodology was employed consisting of filter, wrapper and embedded techniques, whereas feature ranking was decided on the basis of a majority voting system. Applying each technique separately, the order of the feature importance emerged from the frequency of feature appearance in the selection criteria. The features were ranked with respect to the votes received.

The proposed feature selection proceeds along the following steps:

Step1. All features were normalized as described in Subsection III-B

Step2. We performed each one of the six FS techniques separately resulting to the creation of the following six feature subsets FSS_i, $\forall i = 1, \ldots, 6$

Step3. Main loop

- **Step**3.1. For each feature j, we set $\forall j = 1, \ldots, M$, where M the total number of features

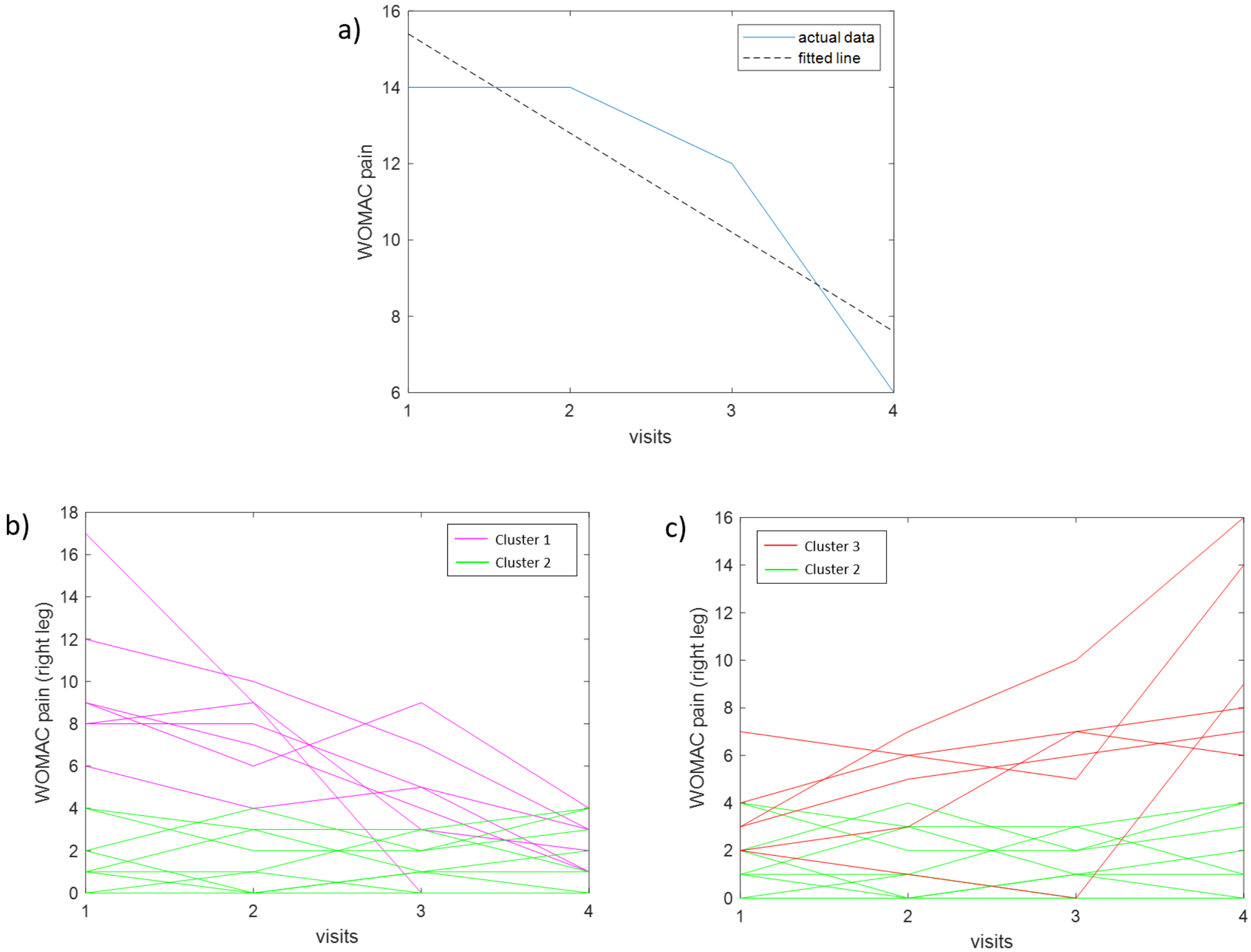

Fig. 1. a) Example of actual pain progression data and the associated fitted line, b) representative examples from pain progression clusters 1 (pain decline) and 2 (no significant pain change), c) representative examples from pain progression clusters 2 (no significant pain change) and 3 (pain increase)

Step3.2. Set $j = 1$

Step3.3. If a feature j is selected in FFS_i, then $V_j = V_j + 1$;

Step3.4. We repeat step 3.2 for each one of the six FS techniques

Step3.5. Set $j = j + 1$ and return to step 3.2
Terminate main loop when $j > m$

Step4. Rank features to descending order with respect to V_j

end

The table III-C gives a brief background description of employed Feature Selection techniques.

D. Machine Learning Algorithms

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

1) Decision Trees: Decision Tree [18] is one of the most famous algorithms for supervised learning for classification problems. It uses a lot of if-then-else decision rule statements in order to come to a decision. Its structure is a branch structure which breaks the data into data subsets, and then it produces decision and leaf nodes. Every node has a minimum of two branches, and every leaf node is for classification or a decision prediction.

2) k Nearest Neighbors: k Nearest Neighbors (kNN) [6] is a non-parametric, lazy learning algorithm. The classification prediction of a sample datapoint, is achieved with the use of data, which are class-separated. The algorithm presumes that similar datapoints are close to each other. More specifically, this algorithm loops over every datapoint in the data and calculates the distance between every datapoint and the chosen datapoint. The distances are sorted in an ascending order and then the algorithm chooses the first k entries.

3) Support Vector Machines: Support Vector Machines (SVM) [5] is an algorithm which finds a line that separates

Category	Technique	Description	Termination criterion
Filter	Pearson Correlation [2]	Pearson correlation examines the correlation between two features. For linear dependence the correlation coefficient is ± 1 and 0 for no dependence. If there is a high correlation between the two variables, each variable is examined with the dependent (target). The variable with the highest correlation remains while the other is rejected. In this way, the other features also emerge.	The selection process was terminated when 30 features have been selected.
	Chi-2 [22]	Chi-2 is a test method for various range of data. It compares two or more sample rates (or composed ratios) and the correlation of two categorical variables. It also works in a hand-wavy way with non-negative numerical and quantitative features.	
Wrapper	Recursive Feature Elimination (RFE) [24] Logistic Regression classifier	Given an external estimator that assigns weights to features (Logistic Regression), recursive feature elimination is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.	
Embedded	Logistics Regression (L2 penalty) [25]	This embedded model returns a subset of selected features relying on regularized logistic regression models. It incorporates feature selection as a part of the model fitting/training process, and features for optimizing the objective function of the learning model.	Each one of the three FS techniques has a different termination criterion leading to a different number of selected features.
	Random Forest [16]	It measures the importance of the features in the prediction. The method rearranges stochastically all values of the features for each tree and uses the RF model to predict this permuted feature.	
	LightGBM [13]	It is a novel Gradient Boosting Decision Tree (GBDT) algorithm. A greedy algorithm can achieve quite good approximation ratio. So, this approach can effectively reduce the number of features without hurting the accuracy of split point determination by much and speeds up the training process of conventional GBDT.	

TABLE I
FEATURE SELECTION TECHNIQUES

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

4) Random Forest: Random Forest is an algorithm consisted of many decision trees algorithms [3]. Its characteristics are the randomness in the sampling of datapoints when building the trees; and the randomness in the features subsets, when splitting nodes. Every tree in the algorithm learns from a random sample of data. These samples of data are being used several times by the trees, which means that the trees take them with replacement. So every tree has high variance because of this fact, but the random forest has lower variance in overall. It is worth noting that the decisions are the average of the predictions of all the trees in the random forest.

5) XGBoost: XGBoost or eXtreme Gradient Boosting [4], is a parallel tree boosting that solves data science problems in a fast and accurate way. After constructing the boosted trees the algorithm calculates the importance score of every feature of the dataset. This score is an indicator of how useful is its feature to the construction of the trees inside the algorithm. The calculation of this score is achieved by the amount that each feature point split improves the performance for the model for the data that the node is responsible for. A popular measure of performance is the Gini index which selects the split points([9]). More specifically the Gini coefficient is a statistic which quantifies the amount of inequality that exists in a population. It is a number between 0 and 1, with 0 representing perfect equality and 1 perfect inequality. XGBoost in fact ranks the features of the data by comparing them to each other.

6) Naive Bayes: Naive Bayes is a probabilistic classifier that uses the Maximum A Posteriori decision rule in a Bayesian setting and is included in supervised learning [19]. The main idea behind this method is the Bayes Theorem. Bayes theorem approximates the probability of an event given the probability of a past event. The Naive Bayes predicts membership of probabilities for every class, such as the probability that the given data point belongs to a particular class. The data point belongs to the class with the highest probability score.

E. Validation

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

IV. RESULTS

Tables II and V present the feature ranking exploration of the first 10 Features of the whole dataset for the left and the right knee, respectively. The feature ranking was decided on the basis of a majority vote scheme by using the proposed feature selection methodology, as discussed in Section III-C. We also note, that features related to symptoms were selected.

A. Results on Left Leg

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

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

TABLE II
MOST IMPORTANT FEATURES FOR THE LEFT LEG

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

B. Results on Right Leg

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

	DT	KNN	NB	RF	SVM	XGB
5	64.46%	57.02%	70.25%	71.07%	66.12%	63.64%
10	71.9%	71.07%	73.55%	79.34%	73.55%	74.38%
15	71.9%	76.86%	76.03%	77.69%	75.21%	80.99%
20	71.9%	76.86%	73.55%	81.82%	72.73%	83.47%
25	66.94%	79.34%	69.42%	**84.3%**	78.51%	80.99%
50	68.6%	69.42%	71.07%	75.21%	78.51%	74.38%
100	73.55%	73.55%	69.42%	76.86%	78.51%	83.47%
150	67.77%	67.77%	68.6%	78.51%	76.86%	79.34%
200	58.68%	67.77%	71.9%	76.03%	78.51%	77.69%
250	61.16%	71.9%	65.29%	74.38%	76.86%	76.86%
300	67.77%	66.12%	61.16%	75.21%	81.82%	79.34%
350	64.46%	61.16%	63.64%	76.03%	76.86%	76.03%
400	60.33%	61.98%	62.81%	74.38%	75.21%	76.03%
450	56.2%	66.12%	61.98%	79.34%	78.51%	77.69%
500	58.68%	58.68%	62.81%	69.42%	77.69%	76.86%
550	61.98%	64.46%	62.81%	77.69%	76.86%	75.21%
600	55.37%	63.64%	60.33%	76.86%	71.07%	75.21%
650	61.98%	61.98%	58.68%	72.73%	74.38%	73.55%
700	50.41%	56.2%	58.68%	70.25%	72.73%	76.03%
750	74.38%	60.33%	58.68%	71.9%	71.9%	75.21%

TABLE III
LEFT LEG: FEATURES AND MODEL ACCURACY SCORES

	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%

TABLE IV
RANDOM FOREST: CONFUSION MATRIX

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

TABLE V
MOST IMPORTANT FEATURES FOR THE RIGHT LEG

2) Performance: For the right leg, Table VI shows the results of the machine learning algorithms that we have applied on different combinations of feature subsets, created by the FS methodology. The best performing algorithm for the right leg is Random Forest with an accuracy score of 84.3%, for 20 features; and as you can see the addition of additional extra features has produced inferior results for our prediction. Table VII shows the confusion matrix of the Random Forest for the best prediction score that it has produced. It is observed that the other algorithms have achieved inferior results

As observed from the Tables V, II and the Figure 2, similar results are obtained on both legs; indicating the repeatability and robustness of the proposed methodology.

V. DISCUSSION AND CONCLUSIONS

In this work we have proposed a methodology in which we identified three different clusters of KOA pain progression

	DT	KNN	NB	RF	SVM	XGB
5	63.33%	68.33%	73.33%	69.17%	70.0%	68.33%
10	75.83%	75.83%	75.83%	76.67%	75.0%	75.0%
15	73.33%	75.0%	75.0%	76.67%	79.17%	70.83%
20	65.0%	70.83%	76.67%	**82.5%**	80.0%	74.17%
25	69.17%	65.0%	77.5%	75.0%	77.5%	75.0%%
50	72.5%	66.67%	73.33%	78.33%	79.17%	75.83%
100	65.0%	64.17%	70.83%	78.33%	78.33%	78.33%
150	67.5%	58.33%	66.67%	77.5%	80.83%	77.5%
200	60.0%	60.83%	62.5%	77.5%	78.33%	79.17%
250	75.0%	58.33%	63.33%	77.5%	78.33%	73.33%
300	68.33%	58.33%	64.17%	78.33%	80.0%	78.33%
350	60.0%	45.83%	63.33%	73.33%	76.67%	78.33%
400	60.83%	53.33%	63.33%	74.17%	75.83%	75.0%
450	59.17%	60.0%	60.83%	75.0%	77.5%	75.0%
500	55.83%	58.33%	59.17%	74.17%	77.5%	73.33%
550	52.5%	53.33%	55.0%	71.67%	77.5%	75.0%
600	65.0%	50.83%	55.0%	69.17%	74.17%	76.67%
650	65.83%	50.83%	54.17%	69.17%	74.17%	74.17%
700	60.83%	51.67%	52.5%	75.83%	77.5%	74.17%
750	65.0%	50.83%	50.83%	71.67%	75.83%	73.33%

TABLE VI
LEFT LEG: FEATURES AND MODEL ACCURACY SCORES

	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%

TABLE VII
RANDOM FOREST: CONFUSION MATRIX FOR THE LEFT LEG

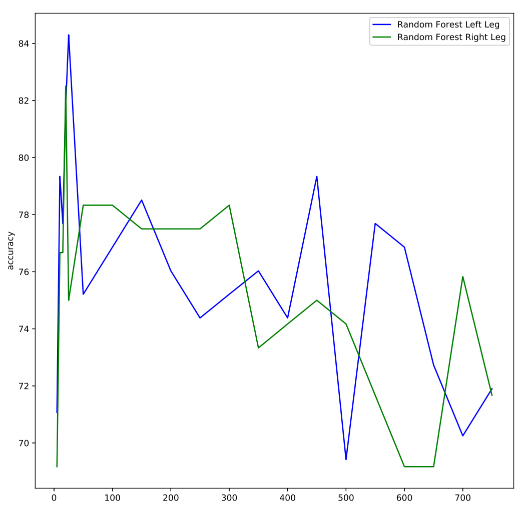

Fig. 2. The comparison of the performance of Random Forest on both legs in accordance to the number of features

along with the most informative parameters towards the development of prognostic ML models that can predict long-term pain progression. In order to achieve this we have developed a voting system for feature importance, in which 6 different methods are used to show the most important features in the dataset. Then we applied 6 different models in various subsets of data, which procedure has proved that XGB achieves a state-of-the-art accuracy score by using only a small number of features. As you can see on Tables VI and III we present the results of our analysis on various combinations of models and numbers of features. Tables V and II present the 10 most important features for KOA pain progression on the right and the left leg respectively.

Summing up we have used for this work only data from the baseline and not from future visits for our prediction. Moreover, we detect the basic trends in pain progression so that we can construct the 3 classes of patients, as we mentioned in Subsection III-A. More specifically we have achieved an **84.3%** for the prediction of pain on the left leg, and an **82.5%** on the right leg. An important observation here is that these high accuracy scores were achieved by using a relatively small subset of features(25 features for the left leg, and 20 for the right leg) that share similar characteristics. It was also observed from the Tables V and II that the most important features for the pain progression prediction are related directly with the pain on each leg respectively. These accuracy scores, with the combination of a small number of features, can set the foundation, for the development of robust tools capable of identifying pain progression at an early stage therefore improving future KOA prevention efforts. Our ultimate goal is to improve the quality of life for people with KOA. For our future work, we are planning to also consider imaging data and associated image-based biomarkers that are expected to further improve the predictive capacity of the proposed methodology.

VI. ACKNOWLEDGMENT

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

REFERENCES

[1] 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.

[2] E. C. Blessie and E. Karthikeyan, "Sigmis: A feature selection algorithm using correlation based method," *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385–394, 2012.

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theor.*, vol. 13, no. 1, pp. 21–27, Sep. 2006.

[7] Y. Dodge and D. Commenges, *The Oxford dictionary of statistical terms*. Oxford University Press on Demand, 2006.

[8] D. Freedman, R. Pisani, and R. Purves, "Statistics edition 2007, 4th," 2007.

[9] J. L. Gastwirth, "The estimation of the lorenz curve and gini index," *The review of economics and statistics*, pp. 306–316, 1972.

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

[11] A. Jamshidi, J.-P. Pelletier, and J. Martel-Pelletier, "Machine-learning-based patient-specific prediction models for knee osteoarthritis," *Nature Reviews Rheumatology*, vol. 15, no. 1, pp. 49–60, 2019.

[12] P. Juszczak, D. Tax, and R. P. Duin, "Feature scaling in support vector data description," in *Proc. asci*. Citeseer, 2002, pp. 95–102.
[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, 2017, pp. 3146–3154.
[14] 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.
[15] 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.
[16] T.-T. Nguyen, J. Z. Huang, and T. T. Nguyen, "Unbiased feature selection in learning random forests for high-dimensional data," *The Scientific World Journal*, vol. 2015, 2015.
[17] 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.
[18] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
[19] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.
[20] S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2018.
[21] I. J. Wallace, S. Worthington, D. T. Felson, R. D. Jurmain, K. T. Wren, H. Maijanen, R. J. Woods, and D. E. Lieberman, "Knee osteoarthritis has doubled in prevalence since the mid-20th century," *Proceedings of the National Academy of Sciences*, vol. 114, no. 35, pp. 9332–9336, 2017.
[22] J. Wang, J. Xu, C. Zhao, Y. Peng, and H. Wang, "An ensemble feature selection method for high-dimensional data based on sort aggregation," *Systems Science & Control Engineering*, vol. 7, no. 2, pp. 32–39, 2019.
[23] 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," *arXiv preprint arXiv:1909.13408*, 2019.
[24] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers," *Genome Research*, vol. 11, no. 11, pp. 1878–1887, 2001.
[25] R. Zakharov and P. Dupont, "Ensemble logistic regression for feature selection," in *IAPR International Conference on Pattern Recognition in Bioinformatics*. Springer, 2011, pp. 133–144.