Quantum Machine Intelligence (2019) 1:73–86
https://doi.org/10.1007/s42484-019-00008-3

RESEARCH ARTICLE

Application of machine intelligence for osteoarthritis classification: a classical implementation and a quantum perspective

Serafeim Moustakidis[1,2] · Eirini Christodoulou[3] · Elpiniki Papageorgiou[1,4] · Christos Kokkotis[4,5] · Nikos Papandrianos[6] · Dimitrios Tsaopoulos[4]

Received: 27 December 2018 / Accepted: 29 September 2019 / Published online: 31 October 2019

Abstract

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

Keywords Deep learning · Osteoarthritis · Diagnosis · Clinical data · Symptoms · Osteoarthritis initiative · Quantum computing perspective

1 Introduction

Osteoarthritis (OA) is the most common chronic condition of the joints. Compared with other types of OA, knee osteoarthritis (Martin 1994) is the most widespread having direct correlation with quality of life. It is a degenerative form of arthritis and specifically is called "wear-and-tear" type, because the cartilage in the knee joint gradually wears away. The progressive loss of articular cartilage with concomitant changes in underlying bone leads to the development of abnormal bony growths, which are called osteophytes or bone spurs. Knee osteoarthritis occurs most often in people over 55 years old (Peat et al., 2001) with the prevalence of the disease rising in people over 65 (Dieppe 1993). OA is also diagnosed in young and athletes following older injuries (Ackerman et al., 2017). The particularity of this disease is that the knee osteoarthritic process is gradual with a variation in symptoms intensity, frequency and pattern. The complexity of the disease combined with the lack of longitudinal data, as well as an absence of reproducible, non-invasive

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42484-019-00008-3) contains supplementary material, which is available to authorized users.

✉ Serafeim Moustakidis
s.moustakidis@aideas.eu

1 Faculty of Technology, University of Thessal, Gaiopolis, 41500 Larisa, Greece

2 AIDEAS OÜ, Narva mnt 5, Tallinn, Harju maakond, Estonia

3 Computer Science Department, University of Thessaly, Lamia, Greece

4 Center for Research and Technology Hellas, Institute of Bio-Economy & Agri-Technology, Volos, Greece

5 Department of Physical Education & Sport Science, University of Thessaly, Trikala, Greece

6 General Dept. (prior Nursing Department), University of Thessaly, Lamia, Greece

methods to measure changes in joint morphology and function limit our understanding of the processes governing osteoarthritis progression. Factors directly related with knee OA are medical risks factors including advanced age, gender, hormonal status, body weight or size, usually quantified using body mass index (BMI), family history of disease along with joint loading during occupational or physical activities. There is also evidence supporting genetic association with OA (Cicuttini and Spector 1996), (Valdes et al., 2011).

Knee OA is not easy to define, predict or treat. The proliferation of large observational studies and the availability of big heterogeneous clinical databases bring new challenges as well as opportunities for enhanced diagnosis of OA through advanced data-driven approaches. Several techniques have been reported in literature in which machine learning models were employed for knee OA diagnosis. Parameters extracted from 3D ground reaction forces have been investigated for their capability to discriminate osteoarthritic (OA) and normal (NL) knee function in a number of studies. In this context, various classification models have been proposed in the literature including Dempster–Shafer theory of evidence, linear discriminant analysis (Beynon et al., 2006) and nearest neighbor classifiers (Mezghani et al., 2008a, Mezghani et al., 2008b). More advanced SVM-based decision trees were also proposed by Moustakidis et al. (2010) to distinguish NL from OA knee gait patterns as well as investigate OA severity over thirty-six subjects with an overall accuracy of 93.44%. The combined use of GRF and 3D kinematic data was also investigated to predict the level of joint deterioration and pain in participants suffering from knee OA (McBride et al., 2011). Generic subject attributes, osteoarthritic outcome scores and kinematics were further combined by Kotti et al. (2013) for automatic knee OA recognition on a small sample achieving a perfect class separation. Genetic parameters along with demographic characteristics were finally employed for knee OA classification using support vector machines and probabilistic neural networks achieving accuracy rates at 76.77% and 90.55%, respectively.

Biomechanical data forms another critical source of information for knee OA diagnosis. Şen Köktaş et al. (2006) presented a method with high accuracy at 98.5% by using MLPs with features of the knee joint angle. Deluzio and Astephen (2007) investigated the association of biomechanical features of gait waveform with the knee OA by using principal component analysis. A hybrid approach to the analysis of motion analysis data using principal component analysis, Dempster-Shafer theory of evidence and simplex plots was proposed by Jones et al. (2008) to characterize the differences between OA and NL knee function data and to produce a hierarchy of those variables that are most discriminatory in the classification process. Mechanical measurements of human walking patterns and clinical characteristics have been also validated for grading knee OA using multilayer perceptron with a moderate accuracy of 80% (Şen Köktaş et al., 2010). Kotti et al. (2017) applied random forest on joint kinematics from 94 subjects (47 subjects with OA and 47 healthy subjects) for automatic knee OA detection achieving a 5-fold cross-validated accuracy of 72.61%. EMG signals have been finally examined by de Dieu Uwisengeyimana and Ibrikci (2017) for the same purpose using artificial neural networks and deep learning. As it was concluded in this paper, knee pathology could be diagnosed more efficiently using surface electromyography signals and artificial neural network algorithms that outperformed deep learning.

Identification of risk factors for developing osteoarthritis has been limited by an absence of reproducible, non-invasive methods to inform clinical decision making and enable early detection of people who are most likely to progress to severe OA. Given the recognized clinical and biological heterogeneity of knee OA, there is an urgent need for clinical tools that will be able to diagnose and potentially predict KOA. This paper makes a contribution towards KOA diagnosis through the application of various machine intelligence models on self-reported clinical data (such as symptoms, disability, function and general health) from the osteoarthritis initiative study (http://www.oai.ucsf.edu/). Different machine learning models as well as deep learning architectures were tested with respect to their ability to recognize participants with symptomatic KOA or being at high risk of developing KOA in at least one knee. The effect of various feature subsets was also investigated. These feature categories are related to (i) the temporal occurrence of symptoms, (ii) symptoms' type and (ii) participants' quality of life status. WOMAC and KOOS features were also evaluated for their capacity to diagnose KOA. Finally, the best performing approach (deep learning) was demonstrated in subgroups defined by gender and age. A quantum perspective of the application of deep learning techniques for the task of OA diagnosis is also given in the discussions.

The paper is organized as follows. Section 2 gives a description of the dataset that was used in our paper including origin and main characteristics. In Section 3, the proposed deep learning methodology along with the necessary data preparation and validation mechanisms is presented. Results are given in Section 4. Discussion of the results is provided in Section 5, whereas Section 6 gives a quantum perspective of the proposed methodology. Conclusions and future work are finally drawn in Section 7.

2 Data description

Data was obtained from the osteoarthritis initiative (OAI) database which is a multi-centre prospective longitudinal cohort study designed to identify risk factors associated with the incidence and progression of KOA (Eckstein et al., 2012). Launched in 2002, OAI began enrolling people, aged 45–79 years, with symptomatic KOA or being at high risk of

developing KOA in at least one knee in four US medical centres. In total 4796 participants were recruited and followed over an 8-year period with a follow-up rate of more than 90% over the first 48 months.

The current study only includes self-reported data about joint symptoms, disability, function and general health from all individuals with or without KOA from the baseline visit. The selected dataset, that comprises of 141 risk factors from 4796 participants, was further separated into 10 overlapping feature subsets with different characteristics. Three subsets are relevant with the temporal occurrence of symptoms, four subsets refer to different types of self-reported symptoms and one involves features related to health, emotional problems, lifestyle and psychology. Hybrid metrics related to WOMAC and KOOS have been also considered as separate sets. The effect of each feature subset on the KOA diagnosis was investigated in the following sections of the paper, providing insights about their clinical significance. Table 1 cites the main characteristics of the 10 feature subsets considered in our paper.

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

- Class 1: Incidence: This class comprises of 3284 participants who do not have symptomatic tibiofermoral knee OA at the screening clinic visit in at least one knee, but who do meet the risk factor eligibility criteria for their age group.
- Class 2: Progression: This class involves 1390 participants with frequent knee symptoms, which are defined as "pain, aching or stiffness in or around the knee on most days". These participants had knee symptoms on most days of 1 month of the preceding year and radiographic tibiofemoral knee OA (Osteoarthritis Research Society International (OARSI) atlas grades 1–3) on a fixed-flexion radiograph at recruitment in at least one knee.
- Class 3: Non-exposed control group: 122 participants have been assigned in this class without any knee symptoms in either knee, who do not have any of the eligibility risk factors and who have OARSI grade 0 in both tibiofermoral compartments for osteophytes.

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

3 Methodology

The proposed in this paper machine intelligence methodology for OA classification includes three processing steps: data pre-processing to handle missing values and normalize the collected clinical data, a learning process for training, and evaluation of the classification results, as illustrated in Fig. 1. The proposed methodology is thoroughly presented in the following sections.

3.1 Pre-processing

Mean imputation was performed to handle missing values. Specifically, for numerical features missing values were replaced by the mean feature value. In case of categorical features, the most frequent category was used to replace NaNs.

Table 1 Main characteristics of the feature subsets considered in this paper

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

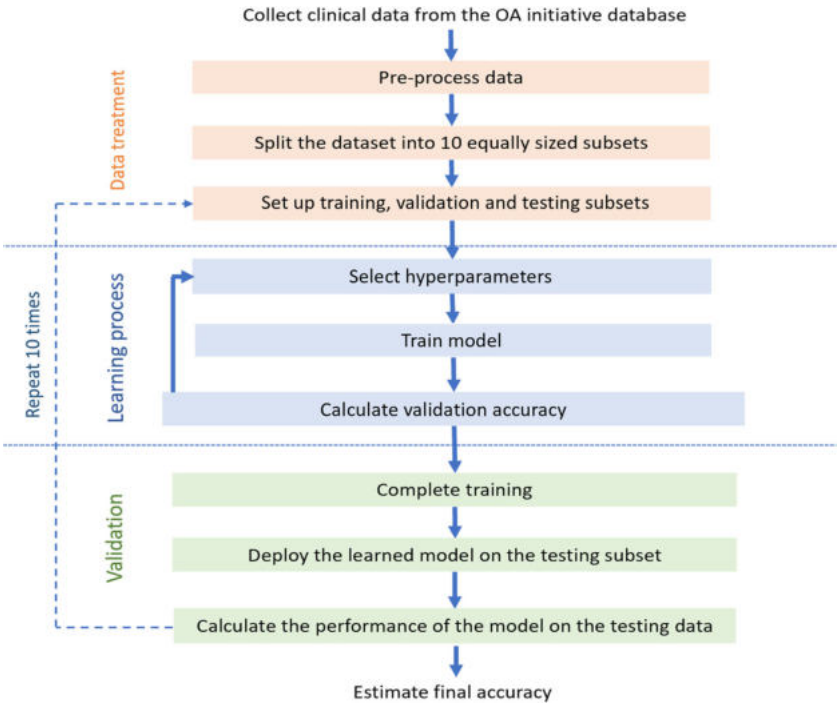

Fig. 1 Flowchart of the proposed machine intelligence methodology

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.

3.2 Learning process

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

Machine learning models We tested linear discriminant analysis (LDA) (Duda et al., 2012) to provide a baseline for comparisons with more advanced models. We also evaluated decision trees (Belson 1959, Witten et al., 2016) driven by Gini's diversity index, KNN and weighted KNN (Atkeson et al., 1997), as well as non-linear support vector machines (SVM) algorithms with Gaussian kernel (Cortes and Vapnik 1995, Scholkopf 1997), which can deal with the overfitting problems that appear in high-dimensional spaces. The ensemble techniques AdaBoost (Freund and Schapire 1997) and Random Forest (Breiman 2001) were also evaluated using DT models as weak learners. Three fuzzy based algorithms were also tested including Fuzzy K-Nearest Neighbors (FKNN) and Fuzzy Nearest Prototype classifier (Fuzzy NPC) by Keller et al., 1985) as well as Condensed Fuzzy K-Nearest Neighbors (CFKNN by Zhai et al., 2014).

Deep learning models Deep learning (LeCun et al., 2015) holds great promise to fulfill the challenging needs of various industries including data-driven healthcare. It performs human-like reasoning and extracts compact features which embody the semantics of input data. Deep neural networks are stacked layer models in which a series of layers is connected together including an input layer, an output layer and a few hidden layers placed between them. The number of nodes in the input and output layers correspond to the dimensionality of the input and the target data, respectively. The nonlinear relationship between the DNN layers is indicated by the following equations:

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

$$h_{W,b}(x) = f\left(z_j^l\right) = f\left(\sum_i w_{i,j}^l x_i^{l-1} + b_j^l\right) \tag{2}$$

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

Our DNN models use fully connected, dense neural layers where the output of one layer serves as the input for the next layer. A number of different DNN structures were investigated

in this paper with varying: (i) input dimensionality (as described in Table 1), (ii) number of hidden layers and (iii) number of nodes per hidden layer. Rectified linear activation was selected given that it has demonstrated high performance on a variety of recognition tasks, and is a more biologically accurate model of neuron activations (LeCun et al., 2006). The final neural layer reduces the dimensionality to either 2 or 3 nodes using Softmax as activation function. Adaptive learning rate was employed with ADADELTA (Zeiler 2012) that automatically combines the benefits of learning rate annealing and momentum training to avoid slow convergence. Weight initialisation was performed using uniform distribution. Early stopping was implemented based on the convergence of the *logloss* metric.

3.3 Validation

Ten-fold cross validation (10FCV) was used to evaluate the effectiveness of the learned classification models. The dataset was split into 10 subsets, called folds. The train-test method was applied iteratively by using each of the 10 folds for testing, while the learning model was trained with the remaining nine. The performance was calculated by averaging the individual ten test scores. To achieve a fair comparison between the different approaches, hyperparameter selection was performed for each one of the investigated machine and deep learning algorithms. A validation subset was held out from the training sets (a randomly selected 10%) as a criterion for selecting the optimum hyperparameters by means of a grid search process.

4 Results and comparisons

4.1 Comparative analysis

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

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

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

Table 3 demonstrates the results of DNN, SVM, RF and Adaboost on the 2-class problem. Apart from being the best model in terms of overall accuracy, DNN achieved the highest sensitivity (92.54%) as well as the highest precision for the class ‘progression’ (84.96%) maintaining a 78.96% precision for class ‘incidence’. The highest specificity was achieved by RF that although did not perform well on the progression class (having a class precision of 73.95% that was the lowest among the four models). On the 3-class problem (Table 4), DNN accomplished the highest accuracies for two of the three classes (incidence and non-exposed) having also the second highest accuracy for the progression class. SVM achieved the highest class accuracy for ‘progression’ samples and it failed in recognizing the non-exposed class with only 9.83% correct assignments. Overall, DNN was proved to be the most effective model in terms of the overall accuracy and the rest of the validation metrics. Despite being the most computationally intensive, DNN was selected for the rest of the experimentation on this paper given that its execution time was not prohibitive for performing multiple runs. The role of quantum computing is also discussed in Section 6 towards a more efficient implementation of such complex models that will alleviate the computational burden of the existing models nowadays.

4.2 Effect of feature categories on the classification performance

DNN is utilized in this subsection as a criterion for evaluating the discriminating capability of different feature categories. Results are demonstrated in different DNN architectures to assess the effect of the DNN structure on the classification

Table 2 Comparative analysis between the best DNN models and state-of-the-art ML models

Model type	10 fold cross validation accuracy (%)			
	2 classes		3 classes	
	Overall (std)	Time (s)	Overall (std)	time
Decision trees (minimum leaf size: 5, Split criterion: Gini's index, Maximal number of decision splits: 7)	79.3* (2.1)	0.22	77.7* (2.0)	0.26
Linear Discriminant	80.1* (2.3)	0.07	76.2* (2.8)	0.08
SVM Gaussian (C = 1, sigma =0.15)	80.2 (1.05)	2.8	79.1 (1.34)	3.2
KNN (k = 9)	79.1* (1.8)	**0.016**	76.9* (2.2)	**0.03**
Fuzzy KNN (k = 11)	79.2* (1.33)	0.034	77.39* (1.45)	0.06
Fuzzy NPC (k = 5)	77.8* (1.24)	0.09	72.4* (1.9)	0.11
CFKNN (K = 9)	78.6* (1.06)	0.1	73.6* (2.05)	0.14
Adaboost (number of weak learners: 130, Maximal number of decision splits: 1024, weak learner: DT)	80.6 (1.33)	25.6	78.6* (1.2)	28.7
Random Forest (number of weak learners: 130, Maximal number of decision splits: 4, weak learner: DT)	80.1 (1.1)	5.1	77.7* (1.86)	5.5
Deep Learning (Adam optimization, ReLU functions, adaptive learning rate, 3 hidden layers, 50 nodes per layer)	**80.7 (1.1)**	31.5	**79.5 (1.2)**	36.4

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

Bold refers to maximum performance achieved per category

performance. Figure 2 shows the performance of different DNN architectures on the 2-class problem using feature subsets that correspond to symptoms occurred at different time periods before the visit. The best accuracy (79.35%) was obtained for the feature subset '*past month*' using an architecture of 3 hidden layers (with 100 nodes at each layer) applied after data resampling. The feature subsets 'past week' and 'past year' were proved to be slightly less informative achieving accuracies marginally higher than 78%.

The effect of symptoms' type on the diagnosis of KOA was also investigated. Figure 3 depicts the performance of DNN using features that correspond to symptoms related to pain, stiffness, knee difficulty and other symptoms such as swelling and grinding. Stiffness was proved to be the most informative symptom with the maximum accuracy of 80.3% achieved by the best DNN using only features related to pain. It is worth to notice that this accuracy is very close to the best accuracy achieved using the entire feature set. Pain-related features were the second best that led to accuracies of 78.2%–79.2% using the deepest DNN models. The rest of symptom types achieved lower performances in the range of 73%. Figure 4 shows the performance obtained using WOMAC-based, KOOS-based features and risk factors related to health and quality of life. A 10FCV performance of approximately 80% was received using DNN models trained on WOMAC features, whereas KOOS and QoL features led to accuracies up to 75.33% and 73.68%, respectively.

Figure 5 summarizes all the classification accuracies obtained from the best performing DNN architectures trained on the 2-class problem (blue line) using the proposed 10 feature subsets. The same analysis was performed on the 3-class problem and the best accuracies per feature category are shown in same figure in orange. It was concluded that the addition of the third class led to a small decay in all the performances received over all the feature subsets. As far as the class accuracy of the non-exposed participants, the following remarks could be drawn from Fig. 6: (i) WOMAC features

Table 3 Confusion matrix for the best DNN architecture on the 2-class problem using the entire feature set

Model		Incidence	Progression	Precision	Sensitivity	Specificity	Overall accuracy
DNN	**Incidence**	2593	691	78.96%	**92.54%**	63.08%	**80.74%**
	Progression	209	1181	**84.96%**			
SVM	**Incidence**	2633	651	80.17%	90.54%	63.13%	80.19%
	Progression	275	1115	80.21%			
RF	**Incidence**	2720	564	**82.82%**	88.25%	**64.57%**	80.1%
	Progression	362	1028	73.95%			
ADA	**Incidence**	2617	667	79.68%	91.59%	63.29%	80.59%
	Progression	240	1150	82.73%			

Table 4 Confusion matrix for the best DNN architecture on the 3-class problem using the entire feature set

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

Bold refers to maximum performance achieved per category

provided an almost perfect (99.19%) identification of class 3, (ii) the feature subsets 'stiffness' and 'pain' accomplished a moderate performance, classifying correctly only 51.64% and 47.55% of the control participants, respectively and (iii) from the rest of the feature subsets, only QoL features contributed with a 12.3% per-class accuracy for class 3. The remaining feature subsets, that do not appear in Fig. 6, did not contribute at all in the identification of class 3 participants.

4.3 KOA diagnosis with respect to gender and age

Table 5 cites classification accuracies obtained by the proposed methodology trained on data subgroups with the full feature set. The following four subgroups were considered: (i) participants older than 70 years, (iii) participants under 70 years, (iii) male participants and (iv) female participants. Significant difference was observed between the two age subgroups. Specifically, a performance of 86.95% was achieved on the

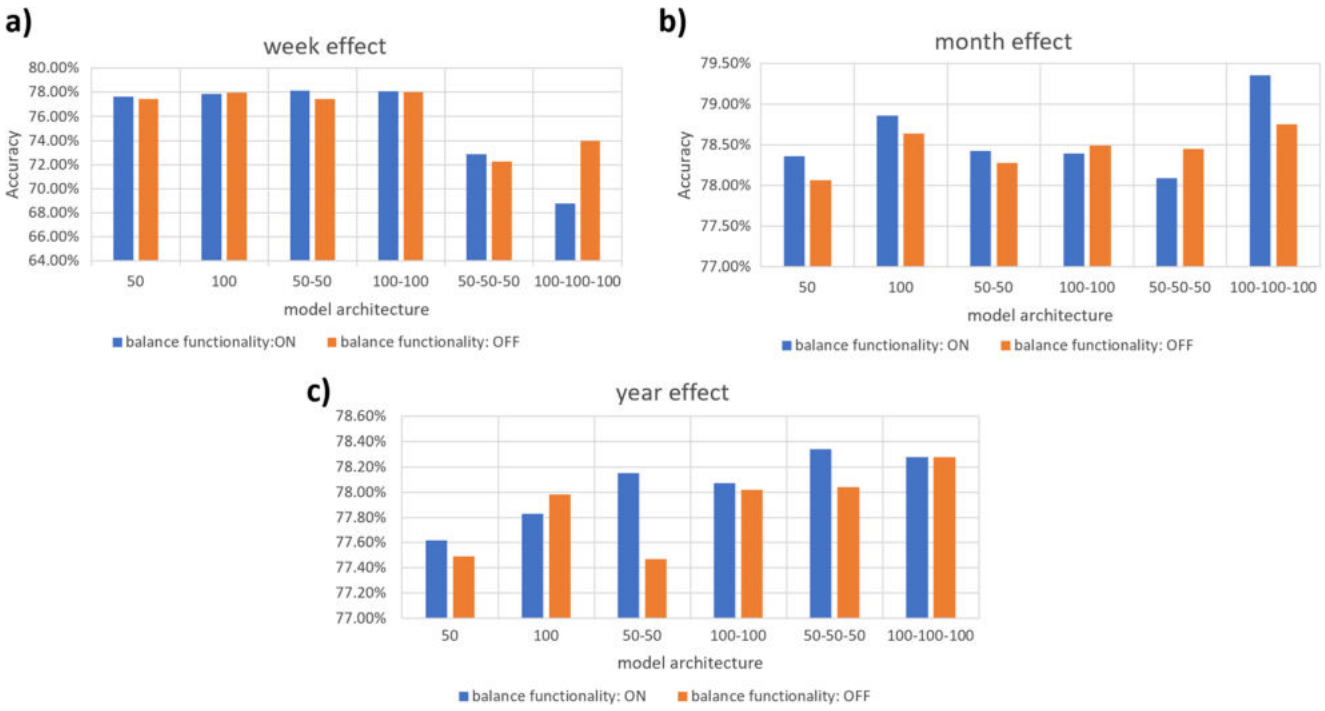

Fig. 2 Results for different DNN architectures for the 2-class classification problem using features that corresponds to symptoms occurred over the: (a) last week, (b) last month and (c) last year

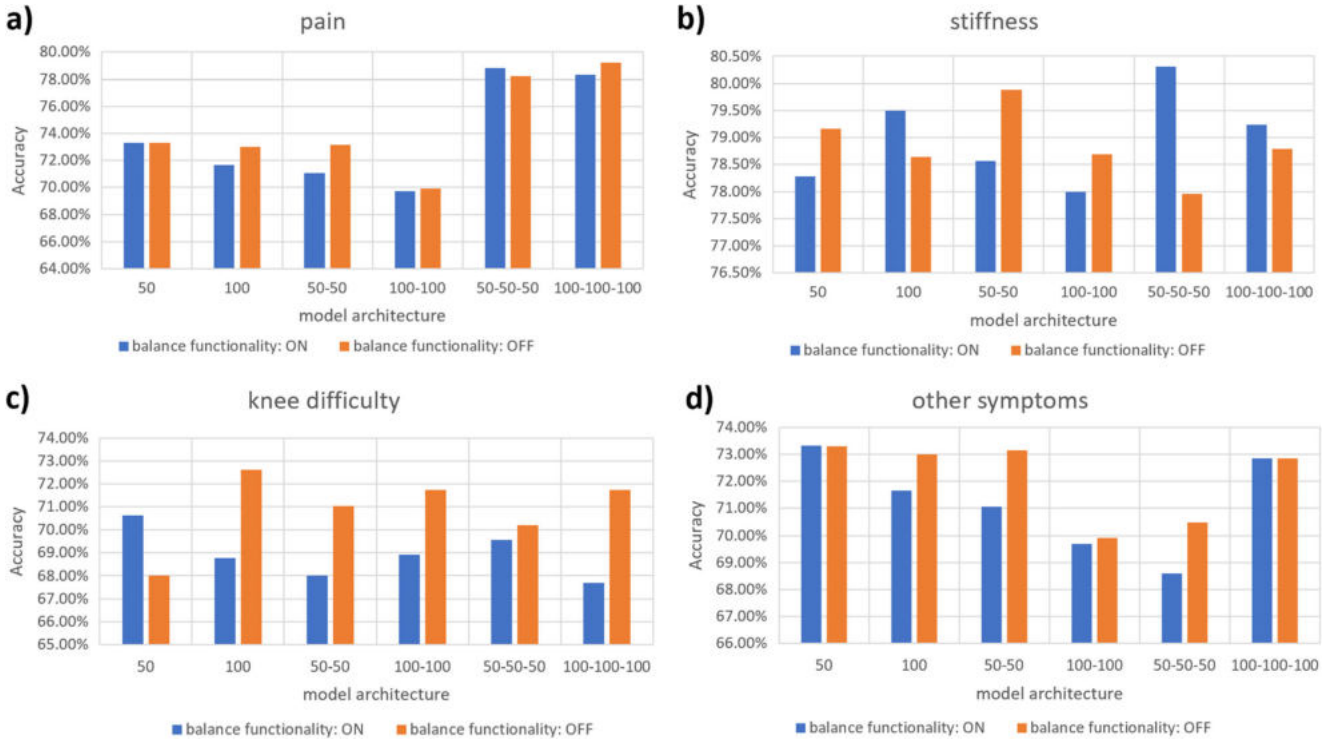

Fig. 3 Results for different DNN architectures for the 2-class classification problem using features that corresponds to symptoms related to: (a) pain, (b) stiffness, (c) knee difficulty and (d) other symptoms

KOA recognition for older participants, whereas the KOA diagnosis accuracy of the 70- age subgroup (80.81%) was closer to the overall accuracy taken on the entire dataset. Accuracies of ~81% and a negligible difference of approximately 0.5% were received for the male and female

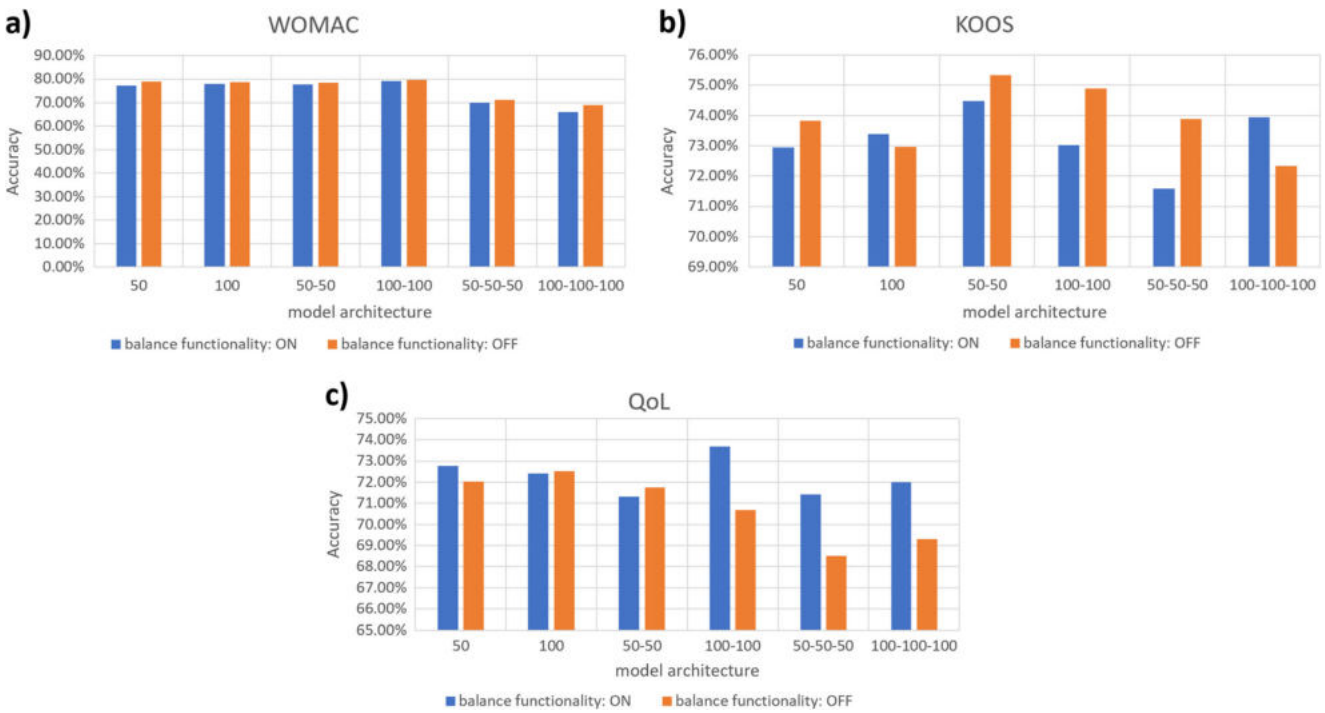

Fig. 4 Results for different DNN architectures for the 2-class classification problem using: (a) WOMAC features, (b) KOOS features and (c) features related with participants' quality of life

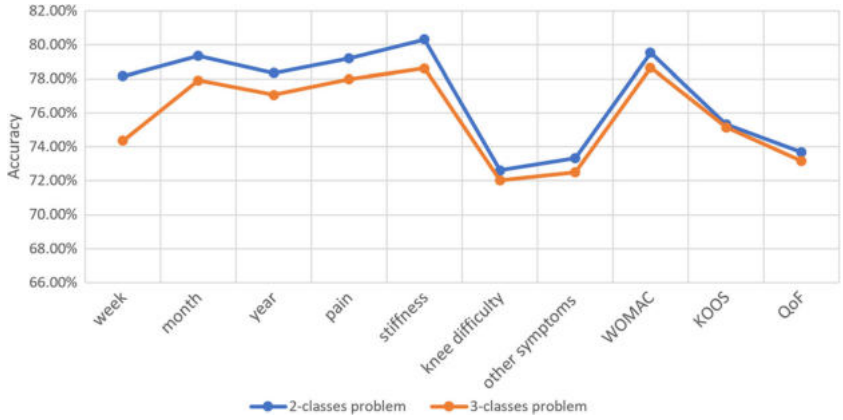

Fig. 5 Results of the best performing DNN architectures for the 2-class (blue line) and 3-class (orange line) classification problem using different feature subsets

subgroups suggesting that gender is not a factor that could considerably differentiate the diagnosis capacity of the DNN models.

5 Discussion of results

An overall performance of 80.74% was achieved in the 2-class problem by the best DNN model trained on the entire feature set, whereas a small accuracy decay was observed when the third class of control participants was added. This decay can be attributed to the inability of the model to cope with the data imbalance issue where class 3 is much smaller in size than classes 1 and 2. Specifically, only half of the control participants were correctly classified indicating the difficulty to differentiate them from participants of high risk to develop KOA. The inclusion of data resampling contributed to better accuracies for class 3 participants outperforming the performance of all the DNN models trained on the original datasets (without data resampling). Finally, the proposed DNN was compared with well-known machine learning techniques and the results

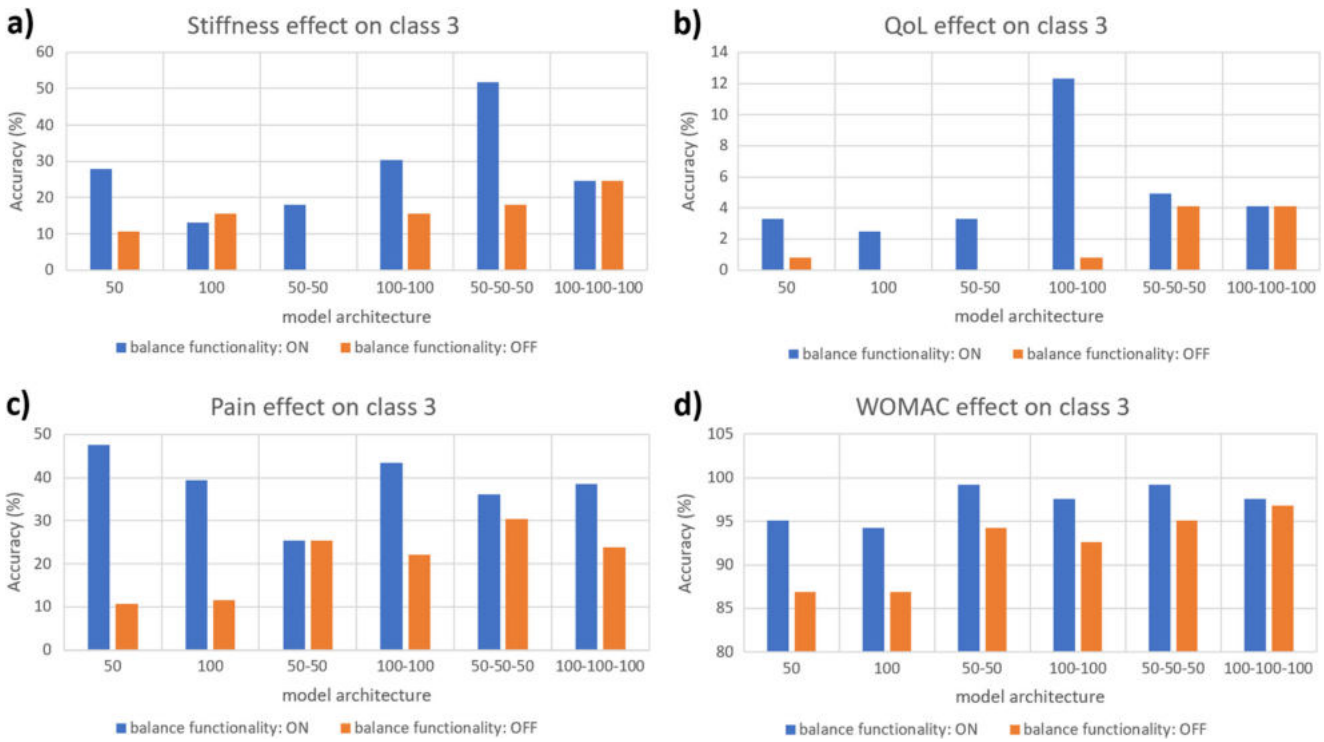

Fig. 6 Accuracy rates for the participants belonging to class 3 (control) for different DNN architectures using features related to (a) stiffness, (b) quality of life, (c) pain and (d) WOMAN features

Table 5 Classification accuracies on different data subgroups

subgroups		problem	accuracy	Best model	
				DNN architecture	Data sampling
age	70+	3-class	86.95%	3 hidden layers (100–100-100)	on
	70-	3-class	80.81%	3 hidden layers (50–50-50)	off
gender	Male	3-class	81.37%	2 hidden layers (100–100)	on
	Female	3-class	81.81%	2 hidden layers (100–100)	off

verified the superiority of deep learning in the KOA diagnosis task in terms of accuracy while being more computationally intensive. As far as the architecture of the selected DNN model, it was concluded that adding more layers (apart from increasing computational complexity to the training and testing phases), allowed for more easy representation of the interactions within the input data and therefore led to the highest accuracy in the case of using the full feature set (Tables 2, 3 and 4). Acting as a universal approximator, DNN architectures with 2 hidden layers also gave high accuracy during the evaluation of the different feature subsets.

As far as the effect of the symptoms' type on the diagnosis of KOA, stiffness was proved to be the most informative symptom leading to an accuracy of 80.3%. Accuracies in the range of 78.2% - 79.2% were received by the deepest DNN models (with 3 hidden layers) trained on pain-related features revealing the importance of pain as a critical risk factor in KOA diagnosis. The rest of symptom types achieved lower performances at the level of 73%. WOMAC also had a significant effect on the KOA diagnosis as demonstrated by the approximately 80% accuracy of the DNN models trained only with WOMAC features. KOOS and features related with quality of life led to lower accuracies (up to 75.33% and 73.68%, respectively). Small difference in accuracy was observed between the three feature categories that were defined by the temporal occurrence of symptoms (last week, month and year). In the challenging task of discriminating control participants from those in high risk, WOMAC features provided an almost perfect (99.19%) identification of class 3, whereas the DNN models trained on the feature categories 'stiffness' and 'pain' classified correctly only 51.64% and 47.55% of the control participants, respectively. The rest of the feature subsets had a minor or no effect on the identification of class 3 participants.

The application of the proposed method in subgroups revealed that it is possible to build even more accurate diagnostic models that work for specific populations. The model built on the aged subgroup (70+) accomplished an 86.95% accuracy that was the highest reported in this paper. This finding implies that local models trained on more focused populations could outperform the global one. The model trained on the 70- subgroup provided an accuracy (80.81%) closer to the performance of the global model. No significant difference was received in the accuracies from male and female subgroups except to a slight increase for both in the range of approximately 2% compared to the global model.

6 Quantum classification perspective for osteoarthritis classification

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

Recent findings by Havlíček et al., 2019set new horizons on the effective combination of machine and deep learning with quantum computing altering how computations are performed to address previously untenable problems without requiring fundamentally new algorithms. Quantum computing is expected to give AI such a boost that it would be able to discover hidden patterns within huge datasets alleviating the computational burden of the existing deep learning algorithms. Significant progress has been recently made in this area towards a better understanding of quantum computers' power for learning tasks. Quantum Neural Networks (QNN) have been proposed by Farhi and Neven 2018 investigating how a popular classification task might be carried out on quantum processors. Despite being primarily theoretical, this study envisions the practical implementation of QNN in the near future. Issues related with the robust training of such networks have been also discussed by McClean et al., 2018with the aim of guiding future strategies for initializing and training QNNs.

The results of this paper on the task of OA classification revealed that DL offers the best solution which unfortunately

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

7 Conclusions

The proposed methodology shows potential for non-invasive OA diagnosis. Here we demonstrated its potential to reliably identify informative risk factors from self-reported clinical data and recognize at a certain level participants with symptomatic KOA or being at high risk of developing KOA in at least one knee. A quantum computing perspective of the future application of the proposed methodology is also discussed highlighting the potential to massively speed up certain types of classification problems. Our method may promote future development and clinical implementation of non-invasive tools for KOA diagnosis and prediction. Future work includes the development of machine learning and deep learning models that could predict the progression of the disease using selected risk factors. More emphasis will be given to local prediction models that will be trained on data subgroups defined by parameters such as body mass index combined with demographics and social indicators. The methodology will be finally extended to include parameters from more disciplines including nutrition, medical history, biomarkers and physical measurements of participants performed in the clinic. Research at the intersection of machine learning and clinical research offers great promise for improving OA related research, advancing clinical decision-making and accelerating intervention programs. To enhance appropriate use of machine/deep learning techniques and stay abreast of new developments in advance analytical techniques, open data and scientific tools must be dynamically encouraged within the OA research community.

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

Appendix

Hyperparameter selection over the validation sets (average) for different classification methods on the 3-class problem

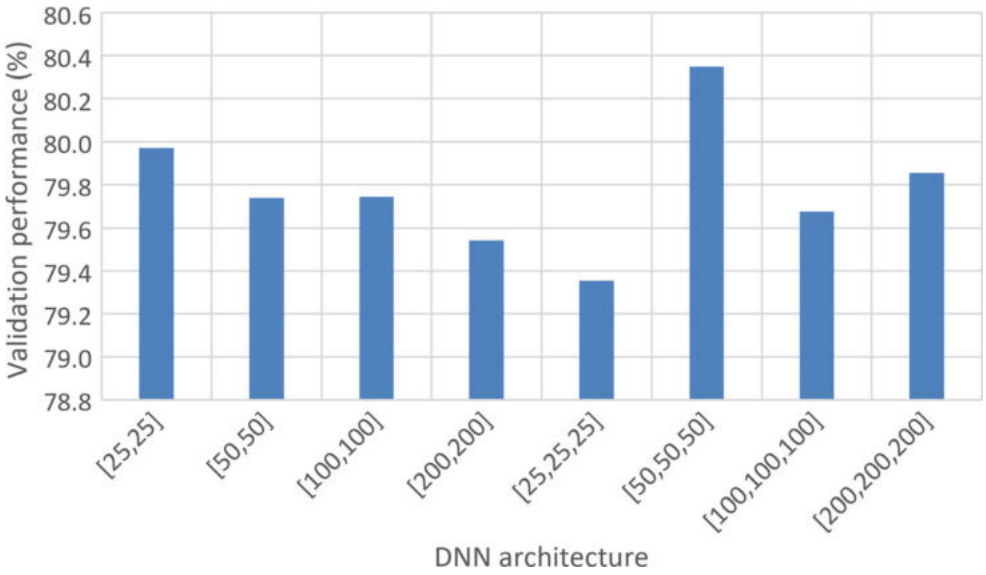

Fig. 7 Average validation performance of various DNN architectures

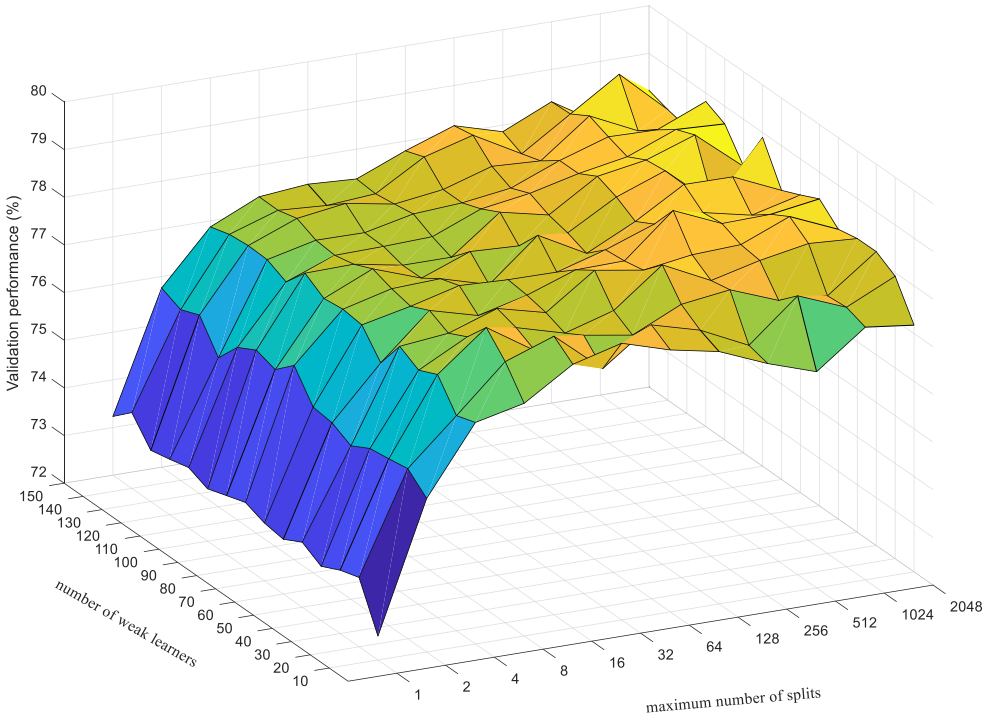

Fig. 8 Average validation performance for Adaboost with respect to the number of weak learners and the maximum number of splits (DT as weak learner)

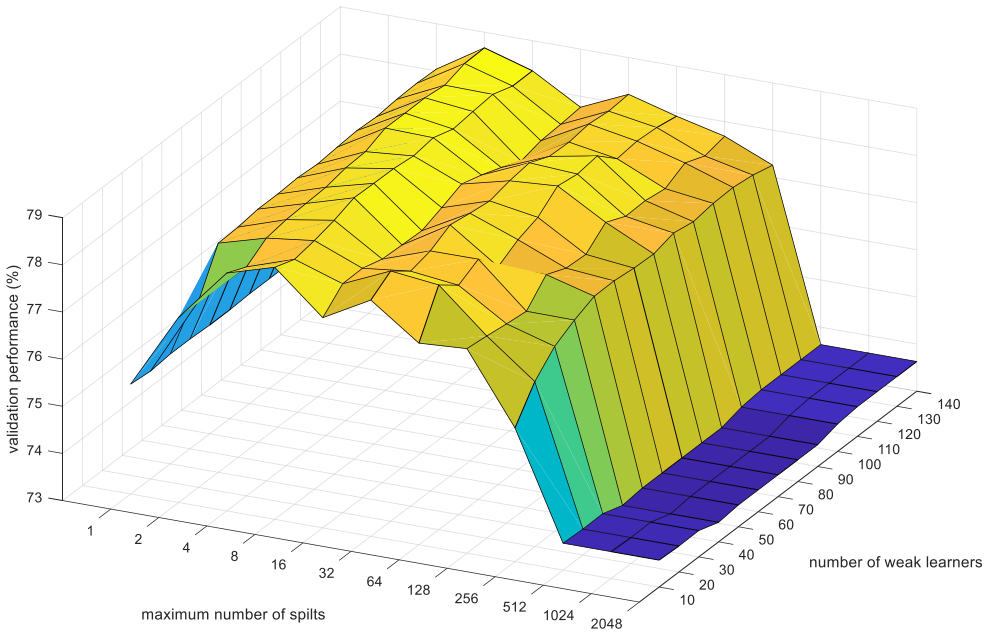

Fig. 9 Average validation performance for Random Forest with respect to the number of weak learners and the maximum number of splits (DT as weak learner)

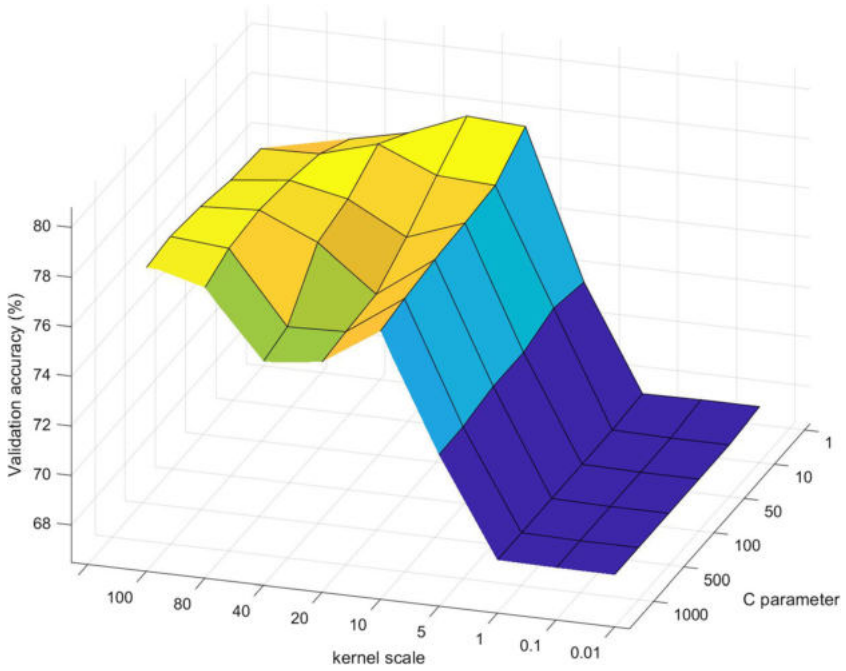

Fig. 10 Average validation performance for SVM with respect to the kernel scale and the C parameters

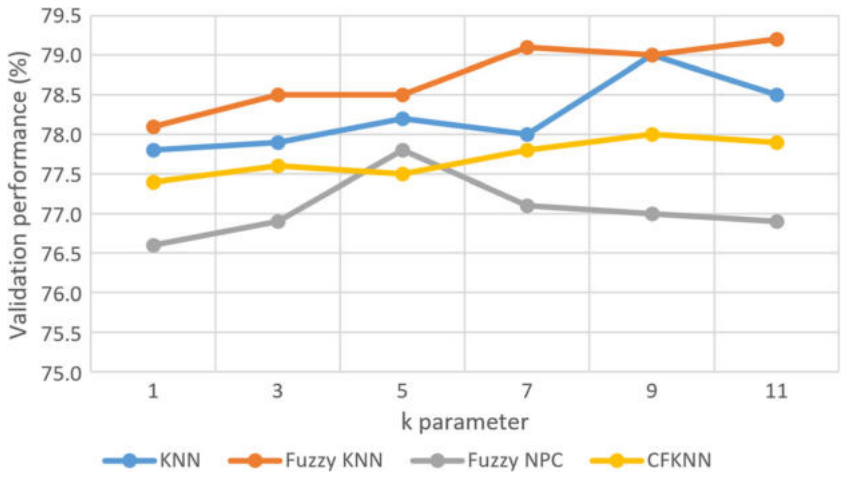

Fig. 11 Average validation performance for KNN, Fuzzy KNN, Fuzzy NPC and CFKNN with respect to k parameter

References

Ackerman IN, Kemp JL, Crossley KM, Culvenor AG, Hinman RS (2017) Hip and Knee Osteoarthritis Affects Younger People, Too. J Orthop Sports Phys Ther 47(2):67–79

Atkeson CG, Moore AW, Schaal S (1997) Locally Weighted Learning. Artif Intell Rev 11(1):11–73

Belson WA (1959) "Matching and Prediction on the Principle of Biological Classification." Journal of the Royal Statistical Society. Series C (Applied Statistics) 8(2):65–75

Beynon MJ, Jones L, Holt CA (2006) Classification of osteoarthritic and normal knee function using three-dimensional motion analysis and the Dempster-Shafer theory of evidence. IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans 36(1):173–186

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Cicuttini FM, Spector TD (1996) Genetics of osteoarthritis. Ann Rheum Dis 55(9):665–667

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3): 273–297

de Dieu Uwisengeyimana J, Ibrikci T (2017) Diagnosing Knee Osteoarthritis Using Artificial Neural Networks and Deep Learning. Biomedical Statistics and Informatics 2(3):95

Deluzio KJ, Astephen JL (2007) Biomechanical features of gait waveform data associated with knee osteoarthritis: An application of principal component analysis. Gait Posture 25:86–93

Dieppe P (1993) Management of osteoarthritis of the hip and knee joints. Curr Opin Rheumatol 5(4):487–493

Duda RO, Hart PE, Stork DG (2012) Pattern classification, John Wiley & Sons

Eckstein F, Wirth W, Nevitt MC (2012) Recent advances in osteoarthritis imaging–the osteoarthritis initiative. Nat Rev Rheumatol 8(10):622–630

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

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

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

Jones L, Holt CA, Beynon MJ (2008) Reduction, classification and ranking of motion analysis data: an application to osteoarthritic and normal knee function data. Comput Methods Biomech Biomed Engin 11(1):31–40

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

Kotti M, Duffell L, Faisal A, McGregor A (2013) Towards automatically assessing osteoarthritis severity by regression trees & SVMs

Kotti M, Duffell LD, Faisal AA, McGregor AH (2017) Detecting knee osteoarthritis and its discriminating parameters using random forests. Med Eng Phys 43:19–29

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553): 436–444

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

Martin DF (1994) Pathomechanics of knee osteoarthritis. Med Sci Sports Exerc 26(12):1429–1434

McBride J, Zhang S, Wortley M, Paquette M, Klipple G, Byrd E, Baumgartner L, Zhao X (2011) Neural network analysis of gait biomechanical data for classification of knee osteoarthritis. Proceedings of the 2011 Biomedical Sciences and Engineering Conference: Image Informatics and Analytics in Biomedicine, BSEC 2011

McClean J, Boixo S, Smelyanskiy V, Babbush R, Neven H (2018) Barren plateaus in quantum neural network training landscapes. Nat Commun 9(1). https://doi.org/10.1038/s41467-018-07090-4

Mezghani N, Boiven K, Turcot K, Aissaoui R, Hagmeister N, De Guise JA (2008a) Hierarchical analysis and classification of asymptomatic and knee osteoarthritis gait patterns using a wavelet representation of kinetic data and the nearest neighbor classifier. Journal of Mechanics in Medicine and Biology 8(1):45–54

Mezghani N, Husse S, Boivin K, Turcot K, Aissaoui R, Hagemeister N, de Guise JA (2008b) Automatic classification of asymptomatic and osteoarthritis knee gait patterns using kinematic data features and the nearest neighbor classifier. IEEE Trans Biomed Eng 55(3):1230–1232

Moustakidis SP, Theocharis JB, Giakas G (2010) A fuzzy decision tree-based SVM classifier for assessing osteoarthritis severity using ground reaction force measurements. Med Eng Phys 32(10):1145–1160

Peat G, McCarney R, Croft P (2001) Knee pain and osteoarthritis in older adults: a review of community burden and current use of primary health care. Ann Rheum Dis 60(2):91–97

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

Şen Köktaş N, Yalabik N, Yavuzer G (2006) Ensemble classifiers for medical diagnosis of knee osteoarthritis using gait data. Proceedings - 5th International Conference on Machine Learning and Applications, ICMLA 2006

Şen Köktaş N, Yalabik N, Yavuzer G, Duin RPW (2010) A multi-classifier for grading knee osteoarthritis using gait analysis. Pattern Recogn Lett 31(9):898–904

Valdes AM, Arden NK, Vaughn FL, Doherty SA, Leaverton PE, Zhang W, Muir KR, Rampersaud E, Dennison EM, Edwards MH, Jameson KA, Javaid MK, Spector TD, Cooper C, Maciewicz RA, Doherty M (2011) Role of the Nav1.7 R1150W amino acid change in susceptibility to symptomatic knee osteoarthritis and multiple regional pain. Arthritis Care Res 63(3):440–444

Witten IH, Frank E, Hall MA, Pal CJ (2016) Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann

Zeiler MD (2012) ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701

Zhai J, Zhai M, Kang X (2014) Condensed fuzzy nearest neighbor methods based on fuzzy rough set technique. Intelligent Data Analysis 18(3):429–447. https://doi.org/10.3233/ida-140649