Neural Computing and Applications (2023) 35:21–33
https://doi.org/10.1007/s00521-020-05459-5

S.I. : INFORMATION, INTELLIGENCE, SYSTEMS AND APPLICATIONS

Dense neural networks in knee osteoarthritis classification: a study on accuracy and fairness

Serafeim Moustakidis[1] · Nikolaos I. Papandrianos[2] · Eirini Christodolou[2] · Elpiniki Papageorgiou[2,3] · Dimitrios Tsaopoulos[3]

Received: 23 January 2020 / Accepted: 26 October 2020 / Published online: 13 November 2020

Abstract
Dense neural networks (DNNs) are a powerful class of learning algorithms that uses multiple layers to progressively extract higher level features from raw input. Either deep or shallow, their outstanding capabilities made a very significant impact on improving the diagnostic potential in multiple applications including medical data classification. In this research work, DNN and Machine Learning (ML) models are explored to address the diagnosis problem of knee osteoarthritis classification which is a common complex problem in older adults. Knee OA diagnosis is a highly complex problem being related to a large number of medical risk factors including advanced age, gender, hormonal status, body weight or size, family history of disease, etc. The main research objective of this study is to apply DNN in knee osteoarthritis classification and validate it for the first time with respect to both accuracy and fairness. To accomplish this, a hybrid criterion including accuracies, confusion matrix and two fairness metrics (demographic parity (DP) and balanced equalized odds (BEO)) were employed to validate the performance of the proposed methodology. Different subgroups of control participants from self-reported clinical data were considered to prove the performance of the proposed methodology. The best performing DNN method is compared with some popular and well-known machine learning techniques for classification with respect to accuracy and fairness. The results of the conducted experimental analysis show the efficacy of the proposed DNN approach improving the classification accuracy (up to 79.6%) and fairness (BEO: $\sim$ 92% and DP: 98.5%) in the OA case study.

Keywords Data classification · Machine learning · Dense neural networks · Knee osteoarthritis · Fairness

1 Introduction

Osteoarthritis (OA) is a degenerative disease of the articular cartilage and the most common form of arthritis that causes joint pain and mobility limitation and, thus, reduces independence and overall quality of life. Being a complex disease in which biochemical and biomechanical factors are involved, OA is commonly localized in the weight-bearing joints and mainly occurs in the knee [1]. Knee OA, as the most widespread type of OA [2], emerges most often in older adults, over 55 years old [3] with the prevalence of the disease rising in people aged over 65 years [4]. It is also diagnosed in young people and athletes following older injuries [5].

Currently, there is no reliable screening test to identify early-stage OA and to quantify knee joint health sensitively and objectively. This is a major issue as any actions related to proper diagnosis and treatment at this early stage that lifestyle changes are the most effective at averting the disease. Identification of knee OA category (i.e. symptomatic or no) seems to be a step of high importance before applying any treatment of knee OA in athletes and older patients.

The prevalence of OA is certain to increase with the progressive increase in life expectancy of the population. Moreover, OA is the musculoskeletal disease with the highest number of known and modifiable risk factors. A

✉ Serafeim Moustakidis
s.moustakidis@aideas.eu

[1] AIDEAS OÜ, Narva mnt 5, Tallinn, Harju maakond, Estonia

[2] Energy Systems Department, University of Thessaly, Geopolis Campus, 41500 Larissa, Greece

[3] Institute of Bio-Economy & Agri-Technology, Center for Research and Technology Hellas, Papanastasiou 51, Larissa, Greece

number of risk factors for OA have been identified in studies over the last three decades using a variety of definitions. These include genetics, metabolism, obesity, bone density, muscle weakness and joint laxity, sex hormones and gender factors, occupational factors, physical activity, sports and joint injuries. Due to the fact that a relatively large number of risk factors are typically considered in knee OA research, it is a challenge to propose an effective tool/method for early diagnosis of knee OA [6]. Following the literature, there is a significant need for clinical tools that will be able to diagnose and potentially predict KOA with respect to the recognized clinical and biological heterogeneity of knee OA; on top of this, no previous works exist for assessing fairness in making decisions across different data groups in this research domain.

1.1 Literature review

Currently, the diagnosis of knee OA is accomplished by considering patient-reported symptoms and X-ray imaging. Specifically, various approaches have been proposed in the literature [7–9] for the detection and analysis of OA using different knee datasets and images. Despite their known limitations in detecting early disease and subtle changes over time, conventional radiographic images remain the "gold standard" for the imaging diagnosis of knee OA [10, 11]. Despite the growing pool of information, there is little uniformity in the diagnostic application of the various measurement techniques and a lack of their confirmed diagnostic utility, as noted in the "Evidence Based Recommendations for the Diagnosis of Knee OA" published by EULAR in 2009 [12].

As it is reported in the literature, various classification approaches were investigated and deployed to discriminate osteoarthritic (OA) and normal (NL) knee function including Dempster–Shafer theory of evidence, linear discriminant analysis [13] and nearest neighbor classifiers [14]. The discrimination of NL from OA knee gait patterns, based on 3-D ground reaction force (GRF) measurements over 36 subjects, has been also investigated by applying ML algorithms, such as support vector machines (SVM) and fuzzy decision tree-based SVM [15] with an overall accuracy of 93.44%. Another research study in [16] examined the automatic diagnosis and classification of knee OA considering generic subject attributes (like age, sex, assessment of the Knee Injury, Osteoarthritic Outcome Score (KOOS)) and kinematic data derived during a gait cycle. A framework was designed to compute the likelihood and degree to which a subject may have knee OA focusing on the aforementioned attributes. Despite the high classification accuracies achieved, the sample was quite small to provide reliable and robust conclusions.

The classification of knee OA has been further investigated by commonly used and popular multi-layer perceptrons (MLP). Şen Köktaş and his colleagues [17] employed ensemble methods of MLPs for gait classification to discriminate OA and NP subjects. The achieved classification accuracy was 98.5% by using MLPs with features of the knee joint angle. Later, a decision tree-based method with MLP at the leaves was investigated by the same team of researchers [18], to correctly identify four OA-severity categories, formed in accordance with the Kellgren–Lawrence scale: "Normal," "Mild," "Moderate" and "Severe". A moderate classification accuracy was achieved (80%) highlighting the need for further improvements.

Body kinetics have been also investigated as an alternative source of information for the automatic detection of knee OA in [19]. The proposed system was validated on a dataset of 94 subjects (47 subjects with OA and 47 healthy subjects) achieving a fivefold cross-validated mean accuracy of 72.61% $\pm$ 4.24%. EMG signals have been further examined by de Dieu Uwisengeyimana and Ibrikci [20] for the same purpose using artificial neural networks and deep learning. The main outcome of that research work was that knee pathology could be diagnosed more efficiently using surface electromyography signals and ANNs that outperformed deep learning. Moreover, classic ML algorithms have been implemented in a computer-aided diagnosis (CAD) system for early knee OA detection using knee X-ray imaging [21]. The proposed system has presented a good predictive classification rate for OA detection (82.98% for accuracy, 87.15% for sensitivity and up to 80.65% for specificity).

Deep learning, which gains significant popularity nowadays, is dramatically improving the state-of-the-art in many different sectors and industries including healthcare. In knee OA research, the recent literature is focused on automatic osteoarthritis detection and classification through image-based deep learning algorithms. Antony et al. [22] proposed a methodology in which a linear SVM was trained on deep features extracted on X-ray images of the knee using a pre-trained Convolutional Neural Network (CNN) model. The methodology achieved a 94.2% classification accuracy on a multi-class KOA problem (five-point Kellgren and Lawrence (KL) scale). The same team [23] also presented a novel pipeline to automatically quantify knee OA severity including a Fully Convolutional Network (FCN) for localizing knee joints and a CNN jointly trained for classification and regression of knee joints. The pretrained CNN ResNet-34 network was also employed by Tiulpin et al. [24] on radiographic images with the objective to automatically score KOA severity where a 66.71% accuracy was achieved. A DL methodology employing DenseNet and MRI, T2 relaxation time

maps was finally proposed by Pedoia et al. [25] concluding that T2 maps have the potential to reveal information that will enhance the diagnosis of KOA.

1.2 Scope and contribution

Similarly to human beings, learning algorithms are vulnerable to biases thus providing "unfair" decisions. In the context of decision-making, fairness can be defined as the absence of any prejudice toward any individual or group based on its inherent or acquired characteristics. Several mathematical definitions of fairness have been recently proposed in the literature. Fairness definitions from political philosophy have been transferred to the area of machine learning in [26], whereas a 50-year history of fairness definitions in the areas of education and machine-learning has been provided in [27]. Fairness in classification problems has been also studied in [28] and the general public's perception of some of these fairness definitions in computer science literature has been presented in [29]. An epidemic spread of AI and machine learning has been reported in different applications including healthcare. Given that ML can be used in many sensitive environments making life-changing decisions, it is crucial to take fairness issues into account and ensure that discriminatory behaviors toward certain groups or populations are omitted.

This paper makes a contribution toward KOA diagnosis through the application of DNN models on self-reported clinical data (such as symptoms, disability, function and general health) from the osteoarthritis initiative study (https://oai.epi-ucsf.org/) building on the knowledge gained from the author's recent research works ([30, 31]). To the best of our knowledge, this paper contains original content in the first-ever validation of machine and deep learning models with respect to fairness in the KOA classification research. Through this study, different DNN architectures were tested with respect to their ability to recognize participants with symptomatic KOA or being at high risk of developing KOA in one knee at least. Different subgroups were investigated defined by gender, age and obesity. The subgroups considered are (i) participants older than 70 years, (ii) participants under 70 years old, (iii) male participants, (iv) female participants, (v) non-obese and (vi) obese participants. The performance of the proposed DL methodology was validated in terms of both accuracy and fairness calculated using the aforementioned subgroups. Finally, a comparative analysis was conducted with various benchmark machine learning algorithms aiming to show the superiority of the proposed DNN structure for the knee OA classification task.

The structure of the paper is as follows. In Sect. 2 a description of the medical dataset is given including the main characteristics used in this paper. Section 3 presents the proposed DNN methodology along with the preprocessing and validation steps. Section 4 gathers the results and holds a discussion on them. Section 5 concludes this research work drawing the main advantages and future work.

2 Medical data description

2.1 Osteoarthritis dataset

In this research study, the dataset was selected from the osteoarthritis initiative (OAI) database that is designed to identify risk factors associated with the incidence and progression of knee OA [32]. Osteoarthritis initiative study (https://oai.epi-ucsf.org/datarelease/) was launched in 2002, enrolling people, aged 45–79 years, with symptomatic knee OA or being at high risk of developing KOA in at least one knee in four US medical centers. 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 related to joint symptoms, disability, function and general health from all individuals with or without KOA from the baseline visit.

2.2 Dataset characteristics

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

A short description of the 141 risk factors is given in [33, 34]. The 68 out of 141 features describe any type of symptoms over the past 7 days, such as any back pain, symptoms' frequency, limited activities due to back pain, number of days stayed in bed due to back pain, etc. Moreover, 10 out of 141 features describe any type of the same symptoms over the past 30 days, and 13 out of 141 features describe any type of the same symptoms over the past 12 months. Next, 64 out of 141 features are related to pain in various activities for both knees, hips and joints in all time intervals, 27 out of 141 features are related to stiffness in all the time intervals, 37 out of 141 features are related to the knee difficulty on either right or left leg on various activities in all time intervals, 12 out of 141 are symptoms such as swelling, grinding sensation, knee catch or hang up in all time intervals, 15 out of 141 features are related to health, emotional problems, lifestyle and psychology, 8 are indexes which consist a score of questions

about pain, symptoms and quality of life for both of knees and 5 are indexes which consist a score of questions about pain, stiffness and disability for both of knees.

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

- *Class 1*: Incidence: This class comprises 3284 participants who do not have symptomatic knee OA, but who do meet the risk factor eligibility criteria for their age group.
- *Class 2*: Progression: This class involves 1390 participants with frequent knee symptoms, which are defined as "pain, aching or stiffness in or around the knee on most days".

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

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

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

The dataset characteristics (including the description of features and the number of samples per subgroup per class) are presented in Table 1.

3 Methodology

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

3.1 Preprocessing

For handling missing values, mean imputation was performed [35]. Specifically for numerical features, missing values were replaced by the mean feature value. In the case of categorical features, the most frequent category was used to replace NaNs. Since activation functions of DNNs do not generally map into the full spectrum of real numbers, we first standardized our data to be drawn from N(0; 1). Normalization also allowed us to compute more precise errors in this standardized space, rather than in the raw feature space.

Data resampling was employed to cope with the class imbalance problem. Specifically, a variant of SMOTE (SMOTE-SVM [36, 37]) was utilized providing borderline over-sampling especially designed for imbalanced data classification problems. In SMOTE-SVM, a borderline area is approximated by the support vectors obtained after training a standard SVMs classifier on the original training set. New instances are then randomly created along the lines joining each minority class support vector with a number of its nearest neighbors using interpolation.

3.2 Dense neural networks

A DNN is actually a fully connected ANN. With respect to the learning process, DNNs use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. They can learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners. A DNN consists of a series of fully connected layers. A fully connected layer is a function from $\mathcal{R}^m$ to $\mathcal{R}^n$. Let $x \in \mathcal{R}^m$ represent the input to a fully connected layer. Let $y_j \in \mathcal{R}$ be the j-th output from the fully connected layer. Then y_j is computed as follows: $y_j = g\left(\sum_{i=1,..m} w_{ij}x_i\right)$ where g is a predefined function known as the activation function and w_{ij} are learnable parameters in the network. This transformation is iterated from layer to layer until we reach the final layer where a Softmax function is applied. For the purpose of this paper we used H2O [38] that is an open-source library widely used for constructing and learning DNNs in prediction and classification tasks. A more detailed description of H2O's learning features, parameter configurations, and computational implementation can be found in [38]. The design space of a DNN is practically infinite severely depending on the number of layers of the DNN and the number of neurons in each of those layers.

Due to the limited available computational power, the size of a DNN needs to be adjusted according to each problem's characteristics [39]. In this study, we used fully connected, dense neural layers where the output of one layer serves as the input for the next layer. We investigated a number of different DNN architectures with varying: (i) number of hidden layers, (ii) number of nodes per hidden layer. The rectified linear activation was selected given that it has demonstrated high performance on a variety of recognition tasks and is a more biologically accurate model of neuron activations [40]. The final neural layer reduces the dimensionality to two nodes using

Table 1 Dataset characteristics

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

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

"Softmax" as an activation function. The adaptive learning rate was employed with ADADELTA [41] that automatically combines the benefits of learning rate annealing and momentum training to avoid slow convergence. Weight initialization was performed by using uniform distribution. Early stopping was implemented based on the convergence of the log-loss metric.

3.3 Validation

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

Definition 1 (*Demographic Parity*) also known as statistical parity [42]. A predictor satisfies demographic parity if the likelihood of a positive outcome is the same regardless of whether the person is in the protected (e.g., female) group.

$$\mathrm{DP}(\%) = 100 - \mathrm{std}(\mathrm{ACU}_i), \forall i = 1\ldots.6 \quad (1)$$

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

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

$$\mathrm{BEO}(\%) = 100 - \mathrm{std}([\mathrm{TP}_1, \mathrm{TN}_1, \ldots, \mathrm{TP}_K, \mathrm{TN}_K]) \quad (2)$$

where K the number of subgroups ($K = 6$ in our paper). BEO receives 100 in the ideal case in which $\mathrm{TP}_i = \mathrm{TN}_i, \forall i$.

3.4 Validation using benchmark machine learning algorithms

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

4 Results and discussion

4.1 Accuracy performance on the full dataset

This section reports the results of the conducted experiments with different DNN architectures on the full dataset. The proposed DNN models were applied on the 2–-class problem, and the obtained classification accuracies along with associated confusion matrixes and class accuracies are given in Tables 2 and 3 with the without data resampling, respectively.

Best accuracies in the majority of the DNN architectures were received without the application of data resampling, whereas the best overall performance (79.6%) was achieved by the DNN model with 1 hidden layer and 50 nodes per layer (see Table 1).

With respect to the effectiveness of the SMOTE-SVM resampling mechanism, the following remarks can be extracted:

(i) The reported confusion matrixes (gray area in Table 2) reveal the inability of the proposed methodology (without data resampling) to recognize participants in the progression class that receives moderate class accuracies (from 62.63% to 69.19%).

(ii) The application of data resampling on the training sets leads to increased class accuracies for the progression class (from 68.18 to 76.52%) and consequently more balanced confusion matrixes (Table 3). Nevertheless, this increase in the class accuracies comes with a small reduction in the overall accuracies of the models in Table 3 (best accuracy observed: 78.81%).

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

Table 2 Overall testing performance of the proposed DNN methodology for different network architectures

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

Table 3 Overall Figutesting performance of the proposed DNN methodology with SMOTE for different network architectures

Hidden layers	Num. of nodes		*progression*	*Incidence*	Class accuracy	Overall accuracy
1	50	*progression*	303	93	76.52	77.95
		Incidence	215	786	78.52	
1	100	*progression*	281	115	70.96	77.59
		Incidence	198	803	80.22	
2	50	*progression*	270	126	68.18	78.10
		Incidence	180	821	82.02	
2	100	*progression*	283	113	71.46	77.88
		Incidence	196	805	80.42	
3	50	*progression*	290	106	73.23	**78.81**
		Incidence	190	811	81.02	
3	100	*progression*	288	108	72.73	78.31
		Incidence	195	806	80.52	

Table 4 Best performance achieved on subgroups

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

4.2 Results on subgroups

Next, the proposed DNN architectures were trained on data from six subgroups of participants: (i) participants older than 70 years, (iii) participants under 70 years, (iii) male participants, (iv) female participants, (v) non-obese and (vi) obese participants. Table 4 cites classification accuracies obtained by the proposed methodology (without data resampling) trained on the aforementioned data subgroups with the full feature set. Significant differences were observed between these subgroups and the entire dataset. In the following subsections, the results of each subgroup are analyzed and explained.

4.2.1 Results from gender effect in diagnosis

Overall accuracies of $\sim$ 78.6% and a negligible difference of approximately 0.1% were received for the male and female subgroups suggesting that gender is not a factor that could considerably differentiate the diagnosis capacity of the DNN models.

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

4.2.2 Results from age subgroups

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

4.2.3 Results from obesity subgroups

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

Figure 1 summarizes the overall and per-class accuracies obtained from the models built on participants' data from separate subgroups. The variability on the obtained accuracies can be attributed to the fact that any learning methodology strongly depends on the dataset in which is trained on. In our case, the proposed DNN methodology has provided higher accuracies for the majority class (incidence) in 5 out of the 6 cases with the overall accuracy in between the two class accuracies. The most balanced distribution of accuracies (for progression, incidence and overall) was achieved in the female subgroup.

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

4.3 Accuracy versus fairness

This subsection provides a more detailed representation of the obtained performance of the proposed methodology

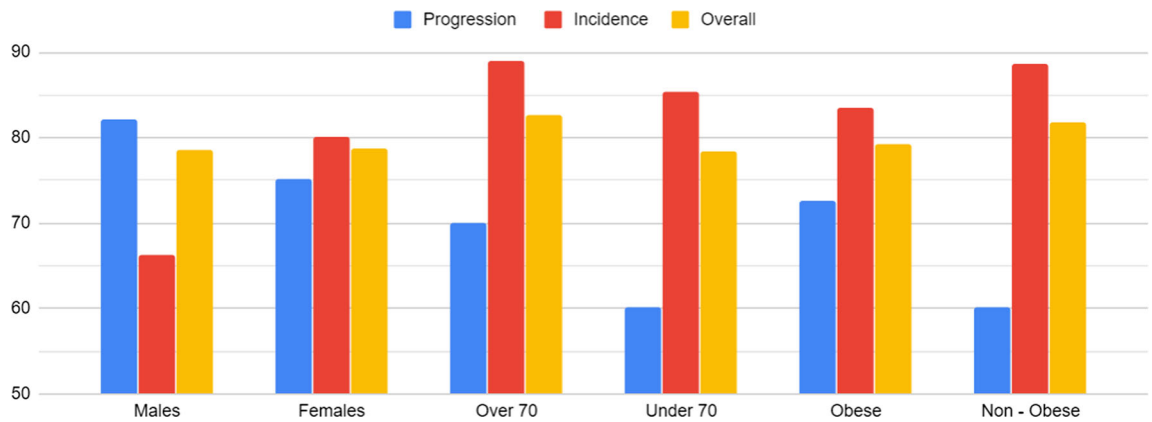

Fig. 1 Best class- and overall accuracy obtained on subgroups

Table 5 Performance achieved by the proposed DNN methodology with and without SMOTE

		Best accuracy on Full set (DNN architecture: 1 hidden layer of 50 nodes)				Best accuracy on full Set with SMOTE (DNN architecture: 3 hidden layers of 50 nodes)			
		Progression	Incidence	Class accuracy	Overall subgroup accuracy	Progression	Incidence	Class accuracy	Overall subgroup accuracy
obese	*progression*	133	52	71.89		139	46	75.14	
	Incidence	51	271	84.16	79.68	57	265	82.30	79.68
non-obese	*progression*	141	70	66.82		151	60	71.56	
	Incidence	112	567	83.51	79.55	133	546	80.41	78.31
over70	*progression*	61	29	67.78		66	24	73.33	
	Incidence	26	172	86.87	80.90	25	173	87.37	82.99
under70	*progression*	213	93	69.61		*224*	*82*	73.20	
	Incidence	137	666	82.94	79.26	165	638	79.45	77.73
male	*progression*	127	41	75.60		126	42	75.00	
	Incidence	60	349	85.33	82.50	72	337	82.40	80.24
female	*progression*	147	81	64.47		164	64	71.93	
	Incidence	103	489	82.60	77.56	118	474	80.07	77.80

with respect to both accuracy and fairness with and without the application of data resampling through SMOTE. Specifically, the DNN methodology was trained on the entire training dataset and the performance is presented separately for each one of the six subgroups on the testing set. Table 5 presents the performances accomplished by the most accurate DNN architectures with and without SMOTE-SVM (on the right and left side on the table, respectively).

Comparable subgroup accuracies were received for both approaches (with and without data sampling), whereas a significant difference was observed in the class accuracies. Specifically, the class accuracies of the SMOTE-enabled models obtained on the 6 subgroups received values in the range of 71.76%–87.37%, whereas the respective class accuracies of the non-SMOTE models were in the range of 64.47%–86.87%. These findings are verified in Fig. 2a that presents the fairness performance (as measured by BEO) with respect to overall accuracy for all the different DNN architectures that were investigated in this paper. It is concluded that SMOTE has a positive effect on the fairness performance (BEO) but at the same time it leads to slightly less accurate models. In terms of demographic parity (Fig. 2b), both approaches had comparable performance with negligible differences in DP values (with the range of $< 1\%$).

Figure 3 shows the fairness performance of the proposed DNN methodology as trained on participants of each one of the 6 subgroups and the full set (with and without data sampling). The best BEO performance was achieved by the SMOTE-enabled model trained on the full set. Training the proposed DNN methodology on the full set (without SMOTE) led to the highest DP performance. Overall, the following remarks can be extracted from the results of this subsection:

(i) Data sampling has a positive effect on the fairness performance of the DNN methodology leading at the same time to more balanced rates for TP and TN throughout all data subgroups.
(ii) The increase in fairness performance comes with a small decrease ($< 1\%$ on average) on the overall predictive accuracy of the models.
(iii) Training on the full dataset increases fairness (both BEO and DP). Thus, special attention should be given in the selection of the training sets that need to represent the whole data variability comprising participants from all sensitive subgroups.

4.4 Comparative analysis with benchmark classifiers

One of the aims of this work was to compare DNN with a variety of well-known machine learning algorithms on the 2-class classification problem using the entire feature sets. To further validate the proposed DNN, the following machine learning algorithms were evaluated for the KOA

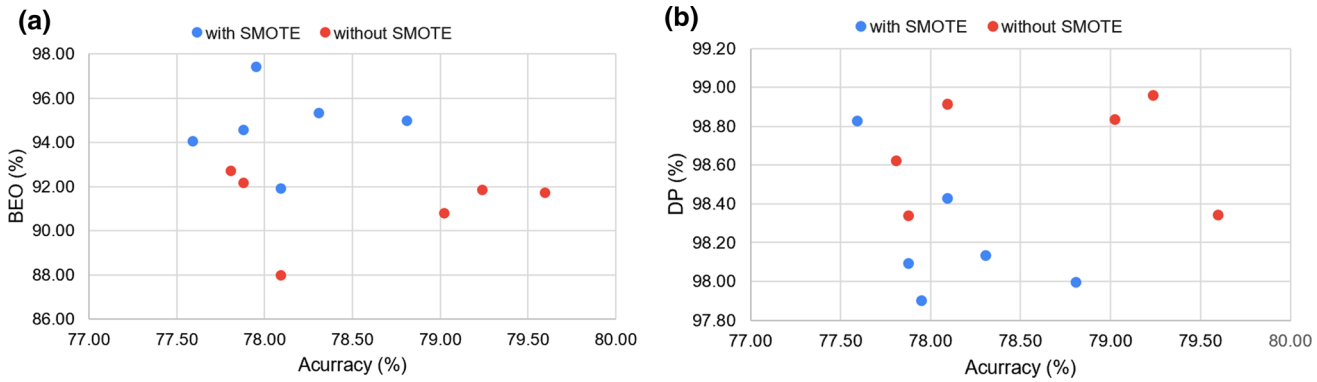

Fig. 2 Fairness with respect to accuracy with and without SMOTE for the proposed DNN methodology: **a** BEO versus accuracy and **b** DP versus accuracy

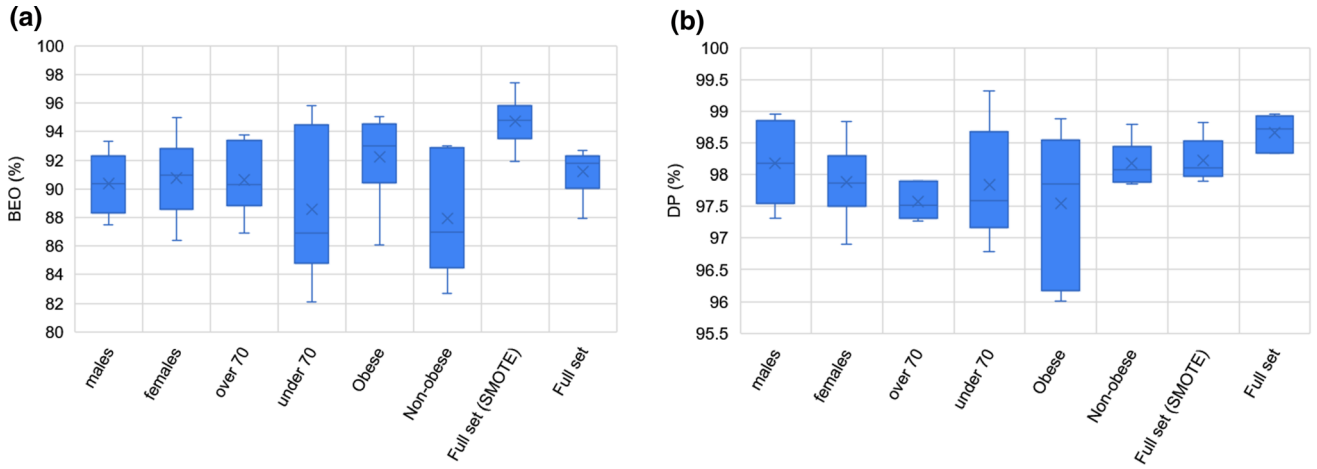

Fig. 3 Fairness achieved by the proposed DL methodology in subgroups and the full set: **a** BEO and **b** DP

classification problem: Decision trees (DTs), KNN [43] with k = 1 and 5, support vector machines (SVM) algorithms with RBF kernel [44, 45], and two ensemble techniques, AdaBoost [46] and Random Forest [47]. Figure 4 compares the performance of the proposed DNN methodology with benchmarks with respect to both accuracy and fairness. DNN accomplished the optimum overall performance with the best accuracy (79.6%) and high fairness values (BEO: $\sim$ 92% and DP:98.5%). The second-best performance was received by AdaBoost that was slightly less accurate ($\sim$ 79%) and less fair (BEO: $\sim$ 92% and DP: 97.9%). High BEO values ($>$ 96%) were achieved by Random Forest that received lower DP values compared to DNN and was less accurate (78.6%). The highest BEO values were achieved by KNN1 without SMOTE. However, this model was less accurate with ACU $<$ 77%. The rest of the ML models had moderate performances in terms of accuracy and/or fairness. Consequently, the proposed DNN outperforms the above well-known machine learning techniques in the knee OA diagnosis task.

4.5 Comparison with existing non-invasive techniques

This subsection focuses on a comparison between the predictive accuracy of the proposed methodology and existing non-invasive AI-based techniques of the recent literature. A deep neural network for detecting the occurrence of osteoarthritis has been presented in [48] using the patient's statistical data of medical utilization and health behavior information. The study was based on 5749 subjects and resulted in 76.8% of area under the curve (AUC). Similarly to the previous study, a DNN-based methodology was proposed in [30] utilizing risk factors from self-reported clinical data about joint symptoms, disability, function and general health. The proposed methodology was demonstrated in the entire OAI population (with an accuracy of 80.74%) as well as in subgroups defined by gender and age where higher accuracies were reported. History and clinical characteristics of the subjects such as age, body mass index and pain level have been also considered for decision-making in OA diagnosis [18]. A

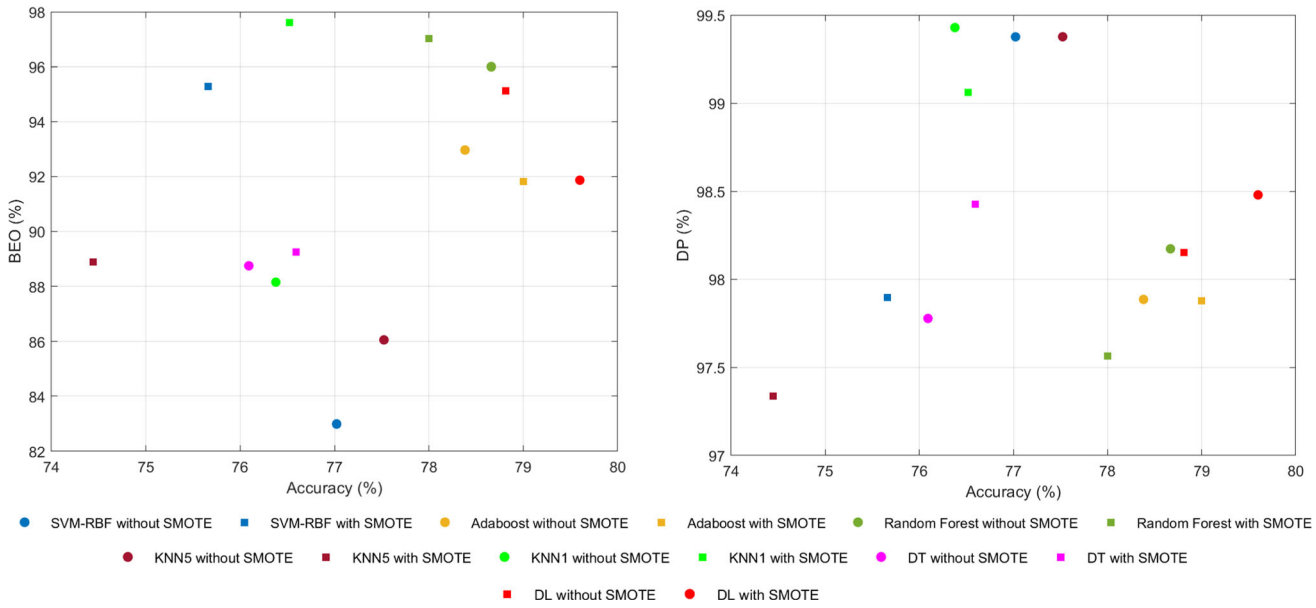

Fig. 4 Comparison of the proposed DNN methodology with benchmarks with respect to accuracy and fairness

success rate of about 80% was achieved using a decision tree equipped with multilayer perceptrons at its leaves. Alternatively, biomechanical data from human body motion analysis has been also explored as risk factors that could contribute to OA diagnosis [15] resulting in detection accuracies up to 93% (demonstrated in datasets of moderate size). The predictive capacity of physical activity measures as contributing factors in the progression of KOA has been also investigated in [49] leading to accuracies up to 74.5%. More information about the recent literature in OA classification studies can be found in [50]. Overall in terms of accuracy, the proposed in this paper methodology provided comparative results with studies employing similar features (non-imaging history and/or clinical data). However, unlike all the aforementioned papers, the main novelty of this paper lies on the inclusion of fairness metrics for the performance evaluation of the classification results.

5 Conclusions

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

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

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Prieto-Alhambra D, Judge A, Javaid M et al (2013) Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis

affecting other joints. Ann Rheum Dis 73:1659–1664. https://doi.org/10.1136/annrheumdis-2013-203355
2. Martin D (1994) Pathomechanics of knee osteoarthritis. Med Sci Sports Exerc 26(12):1429–1434. https://doi.org/10.1249/00005768-199412000-00003
3. 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. https://doi.org/10.1136/ard.60.2.91
4. Dieppe P (1993) Management of osteoarthritis of the hip and knee joints. Curr Opin Rheumatol 5:487–493. https://doi.org/10.1097/00002281-199305040-00014
5. Ackerman I, Kemp J, Crossley K et al (2017) Hip and knee Osteoarthritis affects younger people, too. J Orthop Sports Phys Ther 47:67–79. https://doi.org/10.2519/jospt.2017.7286
6. Wang T, Wen C, Yan C et al (2013) Spatial and temporal changes of subchondral bone proceed to microscopic articular cartilage degeneration in guinea pigs with spontaneous osteoarthritis. Osteoarthr Cartil 21:574–581. https://doi.org/10.1016/j.joca.2013.01.002
7. Janvier T, Jennane R, Toumi H, Lespessailles E (2017) Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the Osteoarthritis Initiative. Osteoarthr Cartil 25:2047–2054. https://doi.org/10.1016/j.joca.2017.09.004
8. Shamir L, Ling S, Scott W et al (2009) Early detection of radiographic knee osteoarthritis using computer-aided analysis. Osteoarthr Cartil 17:1307–1312. https://doi.org/10.1016/j.joca.2009.04.010
9. Antony J, McGuinness K, O'Connor N, Moran K, (2016) Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: 23rd International Conference on Pattern Recognition (ICPR), 2016, pp 1195–1200.
10. Hayashi D, Roemer F, Jarraya M, Guermazi A (2017) Imaging in Osteoarthritis. Radiol Clin North Am 55:1085–1102. https://doi.org/10.1016/j.rcl.2017.04.012
11. Guermazi A, Hayashi D, Roemer F, Felson D (2013) Osteoarthritis: a review of strengths and weaknesses of different imaging options. Rheum Dis Clin North Am 39:567–591. https://doi.org/10.1016/j.rdc.2013.02.001
12. Zhang W, Doherty M, Peat G et al (2009) EULAR evidence-based recommendations for the diagnosis of knee osteoarthritis. Ann Rheum Dis 69:483–489. https://doi.org/10.1136/ard.2009.113100
13. Beynon M, Jones L, Holt C (2006) Classification of osteoarthritic and normal knee function using three-dimensional motion analysis and the Dempster-Shafer theory of evidence. IEEE Trans Syst Man CybernPart A Syst Hum 36:173–186. https://doi.org/10.1109/tsmca.2006.859098
14. Mezghani N, Boiven K, Turcot K, Aissaoui R (2008) Hagmeister N and De Guise J A (2008) Hierarchical analysis and classification of asymptomatic and knee osteoarthritis gait patterns using a wavelet representation of kinetic data and the nearest neighbor classifier. J Mech Med Biol 8(1):45–54
15. Moustakidis S, Theocharis J, Giakas G (2010) A fuzzy decision tree-based SVM classifier for assessing osteoarthritis severity using ground reaction force measurements. Med Eng Phys 32:1145–1160. https://doi.org/10.1016/j.medengphy.2010.08.006
16. Kotti M, Duffell L, Faisal A, McGregor A (2013) Towards automatically assessing osteoarthritis severity by regression trees & SVMs. In: XXIV Congress of the International Society of Biomechanics.
17. Şen Köktaş N, Yalabik N and Yavuzer G (2006) Ensemble classifiers for medical diagnosis of knee osteoarthritis using gait data. In: Proceedings - 5th International Conference on Machine Learning and Applications, ICMLA 2006.
18. Şen Köktaş N, Yalabik N, Yavuzer G, Duin R (2010) A multi-classifier for grading knee osteoarthritis using gait analysis. Pattern Recogn Lett 31:898–904. https://doi.org/10.1016/j.patrec.2010.01.003
19. Kotti M, Duffell L, Faisal A, McGregor A (2017) Detecting knee osteoarthritis and its discriminating parameters using random forests. Med Eng Phys 43:19–29. https://doi.org/10.1016/j.medengphy.2017.02.004
20. de Dieu Uwisengeyimana J, Ibrikci T (2017) Diagnosing knee Osteoarthritis using artificial neural networks and deep learning. Biomed Stat Inf 2(3):95
21. Brahim A, Jennane R, Riad R et al (2019) A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from the OsteoArthritis Initiative. Comput Med Imaging Graph 73:11–18. https://doi.org/10.1016/j.compmedimag.2019.01.007
22. Antony, J, et al. (2017) Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: Proceedings - International Conference on Pattern Recognition.
23. Antony J, et al (2017) Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 376–390.
24. Tiulpin A et al (2018) Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 8(1):1727
25. Pedoia V, Lee J, Norman B et al (2019) Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort. Osteoarthr Cartil 27:1002–1010. https://doi.org/10.1016/j.joca.2019.02.800
26. Binns R (2018) Fairness in machine learning: lessons from political philosophy. In: Proceedings of the 1st conference on fairness, accountability and transparency, PMLR, vol 81, pp 149–159
27. Hutchinson B, Mitchell M (2019) 50 Years of Test (Un) fairness: Lessons for Machine Learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, pp 49–58.
28. Verma S, Rubin J (2018) Fairness definitions explained. In: 2018 IEEE/ACM International Workshop onSoftware Fairness (FairWare). IEEE, pp 1–7
29. Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y (2019) How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, pp 99–106. https://doi.org/10.1145/3306618.3314248
30. Moustakidis S, Christodoulou E, Papageorgiou E et al (2019) Application of machine intelligence for osteoarthritis classification: a classical implementation and a quantum perspective. Quantum Mach Intell 1:73–86. https://doi.org/10.1007/s42484-019-00008-3
31. Christodoulou E, Moustakidis S, Papandrianos N, Tsaopoulos D, Papageorgiou E (2019) Exploring deep learning capabilities in knee osteoarthritis case study for classification. In: 10th International Conference on Information, Intelligence, Systems and Applications (IISA), PATRAS, Greece, 2019, pp 1–6. https://doi.org/10.1109/IISA.2019.8900714
32. Eckstein F, Wirth W, Nevitt M (2012) Recent advances in osteoarthritis imaging—the Osteoarthritis Initiative. Nat Rev Rheumatol 8:622–630. https://doi.org/10.1038/nrrheum.2012.113
33. https://oai.epi-ucsf.org/datarelease/docs/presentations/oarsi092009/MN_OARSI2009WS.pdf

34. https://oai.epi-ucsf.org/datarelease/docs/presentations/acr102008/MNACR2008.pdf
35. Malley B, Ramazzotti D, Wu JT (2016) Data pre-processing. In: Secondary analysis of electronic health records. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_12
36. Nguyen H, Cooper E, Kamei K (2011) Borderline over-sampling for imbalanced data classification. Int J Knowl Eng Soft Data Paradig 3:4. https://doi.org/10.1504/ijkesdp.2011.039875
37. Wang Q, Luo Z, Huang J, Feng Y, Liu Z (2017) A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM. Comput Intell Neurosci. https://doi.org/10.1155/2017/1827016
38. https://h2o-release.s3.amazonaws.com/h2o/rel-turan/4/docs-website/h2o-py/docs/modeling.html#h2odeeplearningestimator
39. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
40. Lecun Y, Chopra S, Hadsell R, Ranzato MA, Huang FJ (2006) A tutorial on energy-based learning. In: Bakir G, Hofman T, Scholkopt B, Smola A, Taskar B (eds) Predicting structured data. MIT Press
41. Zeiler M D (2012) ADADELTA: an adaptive learning rate method," arXiv preprint arXiv:1212.5701.
42. Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30. Curran Associates Inc, New York, pp 4066–4076
43. Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artif Intell Rev 11:11–73
44. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
45. Scholkopf B, (1997) Support vector learning, Ph. D. thesis, Technische Universitat Berlin.
46. Freund Y, Schapire R (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
47. Breiman L (2001) Random forests. Mach learn 45(1):5–32
48. Lim J, Kim J, Cheon S (2019) A deep neural network-based method for early detection of osteoarthritis using statistical data. Int J Environ Res Public Health 16(7):1281. https://doi.org/10.3390/ijerph16071281
49. Alexos A, Moustakidis S, Kokkotis C, Tsaopoulos D (2020) Physical activity as a risk factor in the progression of osteoarthritis: a machine learning perspective. Lect Notes Comput Sci. https://doi.org/10.1007/978-3-030-53552-0_3
50. Kokkotis C, Moustakidis S, Papageorgiou E, Giakas G, Tsaopoulos D (2020) Machine learning in knee osteoarthritis: a review. Osteoarthr Cartil Open 2(3):100069. https://doi.org/10.1016/j.ocarto.2020.100069