Semi-Automated Detection of Anterior Cruciate Ligament Injury from MRI

Ivan Štajduhar[a,d,*], Mihaela Mamula[b], Damir Miletić[b], Gözde Ünal[c]

[a]*Faculty of Engineering, University of Rijeka, Vukovarska 58, Rijeka, Croatia*
[b]*Clinical Hospital Centre Rijeka, University of Rijeka, Krešimirova 42, Rijeka, Croatia*
[c]*Istanbul Technical University, Department of Computer Engineering, Maslak, Sarıyer, Istanbul, Turkey*
[d]*Faculty of Engineering and Natural Sciences, Sabanci University, Üniversite Cd. No:27, Tuzla, Istanbul, Turkey*

Abstract

Background and Objectives: A radiologist's work in detecting various injuries or pathologies from radiological scans can be tiresome, time consuming and prone to errors. The field of computer-aided diagnosis aims to reduce these factors by introducing a level of automation in the process. In this paper, we deal with the problem of detecting the presence of anterior cruciate ligament (ACL) injury in a human knee. We examine the possibility of aiding the diagnosis process by building a decision-support model for detecting the presence of milder ACL injuries (not requiring operative treatment) and complete ACL ruptures (requiring operative treatment) from sagittal plane magnetic resonance (MR) volumes of human knees.
Methods: Histogram of oriented gradient (HOG) descriptors and *gist* descriptors are extracted from manually selected rectangular regions of interest enveloping the wider cruciate ligament area. Performance of two machine-learning models is explored, coupled with both feature extraction methods: support vector machine (SVM) and random forests model. Model generalisation properties were determined by performing multiple iterations of stratified 10-fold cross validation whilst observing the area under the curve (AUC) score.
Results: Sagittal plane knee joint MR data was retrospectively gathered at the Clinical Hospital Centre Rijeka, Croatia, from 2007 until 2014. Type of ACL injury was established in a double-blind fashion by comparing the retrospectively set diagnosis against the prospective opinion of another radiologist. After clean up, the resulting dataset consisted of 917 usable labelled exam sequences of left or right knees. Experimental results suggest that a linear-kernel SVM learned from HOG descriptors has the best generalisation properties among the experimental models compared, having an area under the curve of 0.894 for the injury-detection problem and 0.943 for the complete-rupture-detection problem.
Conclusions: Although the problem of performing semi-automated ACL-injury diagnosis by observing knee-joint MR volumes alone is a difficult one, experimental results suggest potential clinical application of computer-aided decision making, both for detecting milder injuries and detecting complete ruptures.

Keywords: Anterior cruciate ligament (ACL) injury, Knee joint MRI, Feature extraction, Machine learning, Computer-aided diagnosis

1. Introduction

Anterior cruciate ligament (ACL) is the most commonly injured ligament in a human body, for which surgery is frequently performed [1]. Although this type of knee injury is typical for athletes, it can happen to anyone. Presence of an ACL injury is usually determined by performing a magnetic resonance (MR) scan of a knee joint and then visually inspecting the scan. This analysis is usually performed by a radiologist who determines the level of injury, i.e. whether the rupture is complete, partial (or strained) or the ACL is not injured at all [2]. Posterior cruciate ligament (PCL) injury is also possible, but less frequent, because this ligament is wider and stronger than the ACL. Cruciate ligament locations in a human knee are illustrated in Fig. 1. Representative sagittal-plane MR slices of several human knees, each with a different condition of the ACL, are shown in Fig. 2.

Related to the above-mentioned problem, but also applicable to various other problem types, physicians' work in diagnosing various diseases and pathologies from medical imaging can be time consuming, tiresome, expensive and prone to errors. Computer-aided diagnoses (CAD) aims to reduce these factors by assisting physicians in the interpretation of these images [3], mostly in the form of decision-support systems (i.e. a computer intermediary that focuses on a suspicious region in an image, prepares it for inspection, perhaps suggests some of the more probable outcomes, and then lets the human expert make a decision). Some of the more interesting problems encountered in CAD implementations are: 1) segmentation of a region of interest, e.g. the exact space in a 2D or 3D image occupied by a human organ, and 2) detection of anomalies in a region of interest, e.g. detecting lesion presence or traces of a pathology in that organ. Traditionally, for fully-automated or

*Corresponding Author: Tel. +385-51-651448, e-mail: istajduh@riteh.hr. Ivan Štajduhar was a visiting scholar at Sabanci University, Istanbul, where parts of this work were conducted.

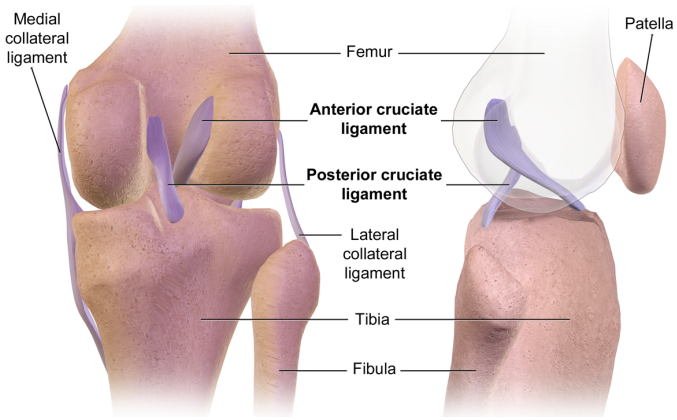


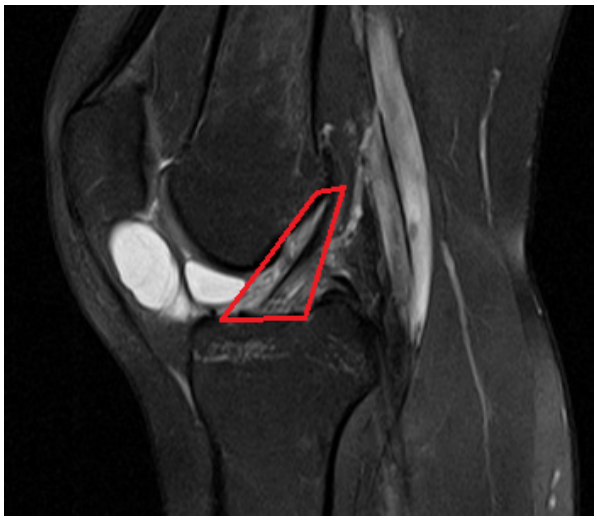
Figure 1: An illustration of a human right-knee joint area (left: posterior view, right: side view), emphasising the cruciate ligament positions.

semi-automated segmentation, algorithms that perform well on specific problem domains were developed. Nowadays, with the development of good evaluation metrics for segmentation performance, machine learning methods are more often used [4] because of their ability to automatically search for new and better models that optimise the chosen metric [5, 6].

Knee joint area was the focus of several CAD research efforts, some of which are mentioned here. An automated method for detection of knee meniscus tears from MR images was introduced in [7]. An automated method for quantification of knee osteoarthritis severity from CR (computed radiography) images was described in [8]. A method of detection of osteoarthritis from X-ray images was suggested in [9]. There is also mention of successful knee joint cartilage segmentation implementations ([10, 11, 12], just to name a few). There have been reports of other efforts too, but further from the proposed research area or with less impact.

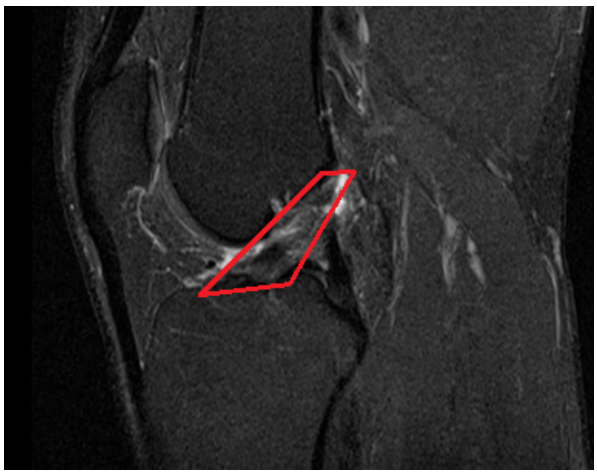
Although the scientific literature regarding the ACL injury problem is abundant, none of the reported work deals with the problem of (semi-)automated detection of this injury (from radiological scans), regardless of methods and procedures used. One probable reason for this is the obvious problem of acquiring the data needed to perform this kind of research, i.e. a sufficient quantity of image collections of adequate resolution (both injured and healthy cases).

In this paper, we tackle the problem of building a predictive model capable of automatically establishing whether an injury of the ACL is present or not, simply by observing MRI data as a radiologist would. We must stress out that the methods described in this paper can be considered only semi automated, because they require the use of extracted regions of interest from MRI volumes. We examine two scenarios: (1) detecting if any kind of ACL injury is observed and (2) detecting if only a complete rupture is observed. The first scenario is useful because it should enable building a decision-support model for alerting a radiologist to a probable case of injury, if it were observed in the MR scan given. If alerted, a radiologist would then dedicate more of his/her time to examining the ACL area in the scan, thus reducing the possibility of establishing an er-



(a)

(b)



(c)

Figure 2: Sagittal plane slices showing the scans of three example knees, each having a different ACL diagnosis: (a) not injured, (b) partially injured, and (c) completely ruptured. The area depicting the ACL is roughly bounded with a red quadrilateral shape.

roneous diagnosis. In spite of the fact that complete ACL ruptures (in contrast to partial injuries) are quite easily detected using visual observation [2], one would also benefit from the second scenario. Because complete ruptures frequently require operative treatment, if such an injury were observed in the scan, this could be used to notify the patient (and the hospital) of the probable impending operative treatment, immediately. On the other hand, if such an injury were not observed, then the patient could be immediately dismissed.

We were interested in determining whether a clinically-useful predictive model could be trained directly from gathered labelled data using standard machine-learning algorithms. Many machine-learning algorithms and models require performing some kind of feature extraction when dealing with images. This is usually performed as a preprocessing step, prior to learning, with a sole purpose of capturing the expressiveness of the visual content by reducing problem dimensionality and removing possible unimportant variation or noise. We explored two popular feature-extraction methods: histogram of oriented gradient (HOG) descriptors [13] and *gist* descriptors [14]. Feature extraction was performed on manually extracted ligament-enveloping regions of interest volumes (ROIs) in the original MR volumes. Transformed datasets were then paired with two powerful machine-learning models: support vector machine (SVM) [15] and random forests (RF) [16]. Detailed tests were performed using different hyperparameter values. Experimental results on a large clinical dataset (917 cases) suggest possible clinical application, both for detection of partial injuries and complete ruptures of the ACL.

This paper is organised as follows. In section II, data acquisition, parsing and the problems encountered are described in detail, following the descriptions of the feature-engineering step and the learning and testing steps. In section III, experimental results are presented and interpreted and, finally, in section IV, they are summarised.

2. Materials and methods

First, we describe in detail the data used for our research - its origin, label extraction and volume transformation procedures.

2.1. Data

A total of 969 knee sagittal plane DICOM MR volumes in 12-bit grayscale were acquired from Clinical Hospital Centre Rijeka picture archiving and communication system (PACS), along with their respective assigned diagnoses. The volumes were recorded between the year 2007 and 2014 using a Siemens Avanto 1.5T MR scanner, and obtained by proton density weighted fat suppression technique, having 0.56 mm in-plane spacing (X and Y axes), 3 mm slice thickness and 3.6 mm spacing between slices (Z axis). X and Y axes constitute a high-resolution plane, whereas the view along the Z axis will be relatively blurry and hold less information, depending on the parameters. The Hospital has been using the described setup as a standard for morphological assessment of a knee, along with axial- and coronal-plane volumes (in some cases even using additional sequences, having different parameters). Similar setups are mentioned as a standard for morphological and compositional assessment of knee cartilage [17]. At the Hospital, a patient's cruciate ligaments conditions are normally established by observing only the sagittal plane volumes in proton density weighted fat suppression technique, because they are the most informative concerning this problem. Therefore, we decided to concentrate our efforts only on those volumes.

From the acquired dataset, three volumes were discarded due to data corruption, such as missing DICOM slices. Additional 22 volumes were discarded for containing abnormal physiological characteristics, either portraying knees after ACL reconstruction or knees exhibiting severe stages of osteoarthritis, leaving a total of 944 sequence volumes. Gathered valid volumes varied in size from $290 \times 300 \times 21$ to $320 \times 320 \times 60$ with median dimensions $320 \times 320 \times 32$ (slice height $\times$ slice width $\times$ number of slices) voxels. Voxel intensities of each distinct volume were, therefore, represented by integer 3D matrices of varying sizes. For the purpose of reducing the unwanted intra-class variation, all MR volumes portraying right knees were mirrored to resemble the volumes portraying left knees.

2.1.1. Labelling data

Each of the volumes in our dataset included a lengthy diagnosis from the PACS, concerning the physical state of the entire knee joint area under inspection, e.g. in what condition were the ligaments, menisci, bones, cartilage, and so on. This kind of diagnosis is normally recorded by radiologists when performing an MR exam of the knee joint area. Each diagnosis was established by one out of four different radiologists involved in this study. Three of those were experienced radiologists - consultants (their initials are D.M., D.V. and S.B.) and one was a senior resident with experience in musculoskeletal radiology (D.J.).

Recorded diagnoses concerning the ACL condition usually differentiate the following states: 1) ACL is not injured, 2) ACL is partially injured (partially ruptured or strained), and 3) ACL is completely ruptured. We manually inspected all of the diagnoses linked to distinct volumes and assigned them appropriate labels. This was done under the supervision of a skilled radiologist (M.M.). Several examples of relevant diagnosis excerpts, along with assigned labels, are displayed in Table 1. Most of the diagnoses were additionally amended with a summarised conclusion (e.g. "ACL partially ruptured") from which we have drawn our final labelling decisions, but in their absence we have drawn our conclusions from the exhaustive descriptions (Table 1). The following label distribution was established: 717 cases ($\approx$ 76%) were originally labelled as non-injured, 182 ($\approx$ 19%) as partially injured and 45 ($\approx$ 5%) as completely ruptured.

After establishing labels from diagnoses, during visual inspection of given volumes we discovered a serious flaw in our data extraction design, one that is inherent to radiology exams of this type - uncertainty due to the lack of visually distinguishable characteristics for differentiating between some partially injured cases (representing smaller partial lesions) and some of those that are not injured at all (anatomical variations of normal conditions). The data we had at our disposal were abun-

Table 1: Examples of some of the extracted diagnoses and our simple labelling principle. Quoted relevant diagnosis description excerpts were translated from Croatian language. The diversity of the problems being analysed and the diagnosis variations depict a complex problem.

Diagnosis excerpt	ACL condition
"Cruciate ligaments are followed in their continuity" "... is of proper tone, characteristic signal" "... is of proper tone and followed in its continuity" "... somewhat thinned, but of proper tone and characteristic direction, without loss of continuity" "... adequately low signal"	Not injured
"... not of characteristically low signal nor thick enough, although certain fasciculi are followed in continuity, but is somewhat thinned on its proximal junction" "... of a more heterogeneous signal, which can point to a partial distortion" "... of a more irregular contour and of a more heterogeneous signal in its middle segment, which can point to a distortion" "... heightened signal ... in its distal junction and partially of irregular contour which fits a strain" "... lesioned at its proximal segment" "... shows chronic lesion in the middle third part, partial rupture" "... thicker and for the most part of altered signal, in the sense of partial interstitial lesion" "... more oedematous, which points to its strain" "... of heightened signal, but kept continuity, which points to its strain" "Part of the ligament laying on the tibial plateau, and only a part of the fasciculus is being directed towards the femoral junction"	Partially injured
"... loss of continuity ..." "Rupture of the ACL" "ACL is of heightened signal and obscured in its middle part – rupture" "ACL continuity is interrupted at the half of its length and is not followed on its proximal segment... complete rupture of the ACL" "... middle segment thicker and towards the proximal segment is not visualised, probably a complete rupture" "... interrupted continuity at the proximal junction" "... distal half grounded on the tibia floor, and thread rupture visible in its proximal part"	Completely ruptured

dant with such cases. For the most part, this refers to such cases where the ACL: 1) is of physiological shape, but its entire region is saturated with higher pixel intensities; 2) looks thinned at one point (either proximal or distal), exhibiting a wider higher pixel intensity region close to it; 3) exhibits a thinning at any of its smaller portions, usually represented by a lack of a low intensity region, and; 4) is of physiological shape, but is not followed as a completely straight line – rather it looks a bit curved. These visual characteristics are often attributed to a strained or oedematous ligament, meaning partially injured. On the other hand, they are also quite often completely ignored by radiologists in establishing a diagnosis and stating that the ligament is completely healthy [18]. The reason for this discrepancy in establishing a diagnosis is rather simple. When interpreting the MR scans, radiologists are often given additional input regarding the condition of the patient at hand and the reason why they are being examined (e.g. sports activity injury or car crash injury), thus making their findings potentially biased, based on additional information. For example, if a patient came to an emergency room from a basketball field, then the radiologist would probably be biased towards concluding that the ligament injury is present, even if the evidence is otherwise inconclusive. Finally, there also exists a possibility of a radiologist making an error diagnosing an injury. The severity of this problem is illustrated by a couple of examples shown in Fig. 3. These examples are much harder to differentiate, as opposed to those shown in Fig. 2 where intergroup differences are obvious to the naked eye. Introducing additional labels for dealing with this problem was not an option because it would make the problem of learning even harder. Same holds for introducing a continuous scoring scheme in spite of the fact that designated labels are inherently ordinal.

Ground truth regarding ligament injury type can only be determined with invasive interventions, such as operation or autopsy. Patients who have not been diagnosed with complete ruptures are seldom sent to the operating room, thus rendering it impossible to confirm the majority of cases postoperatively. We managed to obtain a confirmation that 25 out of 28 patients who underwent a surgical operation of the knee at the University Orthopaedic Clinic Lovran were confirmed to have a completely ruptured ACL. The remaining 3 cases were postoperatively diagnosed as severe cases of partial ACL rupture. Postoperative confirmations of morphologically established diagnoses were unfortunately not available for other cases. Therefore, we had to ensure data labels were accurate enough only using visual inspection of the MR volumes. To accomplish this, we decided to assign another experienced radiologist (M.M.) with a task of diagnosing injuries from the observed 944 MR volumes, assigning a label to each one (normal, partially injured or completely ruptured), blind to the original annotations and extracted labels. This round of labelling was obviously somewhat biased by the fact that the radiologist was familiar with the approximate distribution of the original labels. For each exam case, we then compared the labels reflecting both diagnoses, and retained only those cases where both radiologists agreed on the diagnosis. This led to the exclusion of another 27 cases. Inconsistencies between the original labels and the newly assigned ones are presented in Table 2. Majority of the inconsistencies pertains to non injured and partially injured cases. Only one instance, originally labelled as partially injured, was assigned

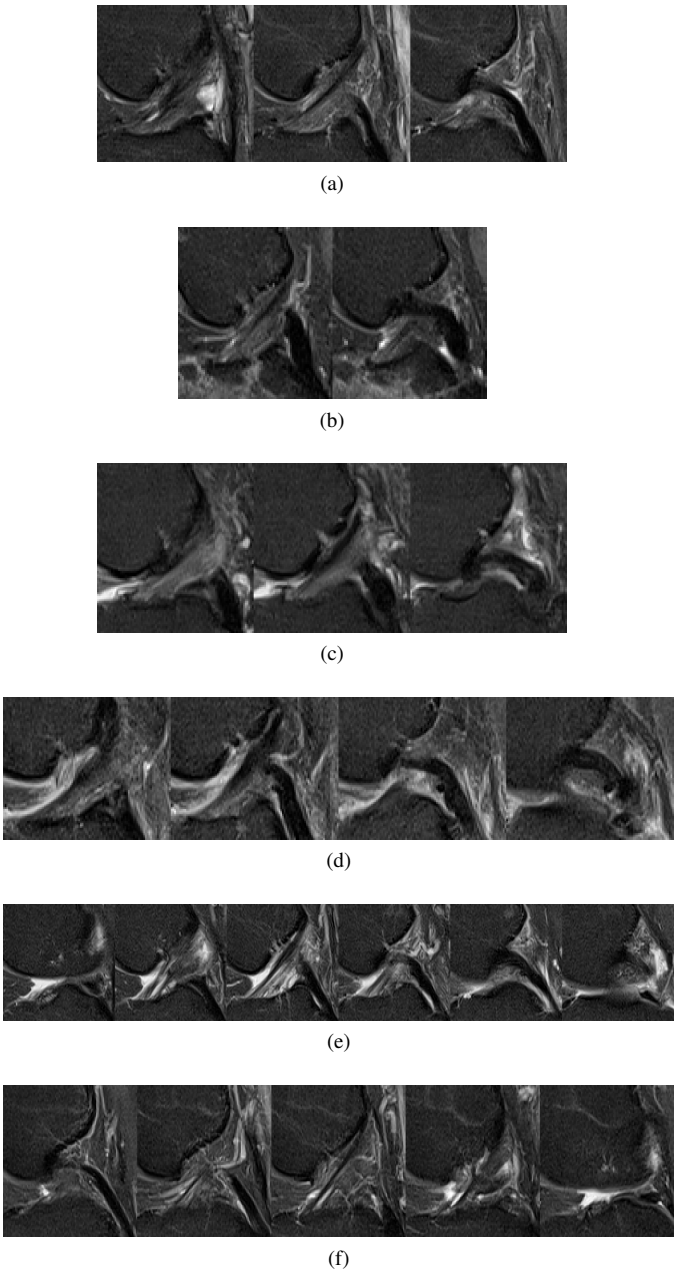


Figure 3: Several region of interest (ROI) sequences depicting non-injured (a,b,c), and partially injured (d,e,f) ACL exam cases, labelled according to their respective diagnoses extracted from the PACS. Notice the lack of sound distinguishable differences between the two groups. Ligament line shape is present in both groups, pixel intensities vary in both groups (suggesting possible strains) and texture differences are practically indistinguishable.

a fully-ruptured label and only one, originally labelled as fully ruptured, was assigned a partially-injured label. After the exclusions, final dataset used in our experiments consisted of 690 non-injured ($\approx$ 75%), 172 partially injured ($\approx$ 20%) and 55 completely ruptured ($\approx$ 5%) cases.

We were interested in observing the potential of building a model capable of detecting injured ACL cases (partially and completely ruptured), differentiating them from normal (healthy) cases. Presuming it were reasonably accurate, such a model could be utilised to put forward an early warning, alerting a radiologist that an injury is probably present in the volume under observation. Furthermore, both patients and hospitals would also benefit from a model capable of automatically detecting completely ruptured cases because this would give them an immediate notice of the impending operative treatment. For these reasons, we examine both scenarios in section 3.

Table 2: Distribution of inconsistencies between the originally extracted labels and the newly assigned ones, presented in a form of a confusion matrix.

		Original extracted labels		
		NI[1]	PI[2]	CR[3]
Newly assigned labels	NI[1]	690	4	0
	PI[2]	21	172	1
	CR[3]	0	1	55

[1] NI = Not injured
[2] PI = Partially injured
[3] CR = Completely ruptured

2.1.2. Focusing on a region of interest

When a radiologist evaluates ACL condition from a recorded volume, he/she first concentrates his/her efforts on locating a smaller region of the entire volume using his/her prior knowledge of sound morphological properties of the knee and then focusing his/her vision on the details of this smaller region, disregarding the rest. This rather intuitive procedure is mimicked here, as follows: a rectangular region of interest (ROI) was manually extracted from each MR volume by a radiologist (M.M.) using visual inspection. This region was to envelop a wider ACL area, as can be seen in a couple of examples in Fig. 4. Although the ROIs were manually extracted here, we speculate that equally good results could have been obtained by using a reference (template) knee MRI volume, which can be reasonably accurate at pinpointing the wider ACL area.

ROI selection can be automated by building a reference knee MRI volume, from a selected subset of the training set [19]. One can rigidly register the subset of volumes and compute an average volume afterwards, which can be further refined in a second iteration of registration [20]. The bounding box of the ROI on the template space can then be transformed to the space of a given patient volume. However, the described procedure has possible sources of error due to data registration process. Location offsets in registration could result in two stages: both in creation of the reference volume and in the second registration step, while estimating the parameters for transforming the ROI from the reference volume to the given data space. Due to potential errors in registration as well as its added computational costs, a semi-automated detection approach is adopted in this work. Hence, the main focus of this paper is on automatically detecting the ACL condition from a given ROI, which implies a semi-automated detection system.

Extracted ROIs varied in size from 54×46×2 to 124×136×6, having median dimensions 92×91×3 (slice height × slice width × number of slices) voxels. All the ROIs were then rescaled using linear interpolation to fit one standard size, $90 \times 90 \times 3$, giving a total of 24300 intensity features. This rescaling led to an obvious loss of distinguishing visual features in some cases.

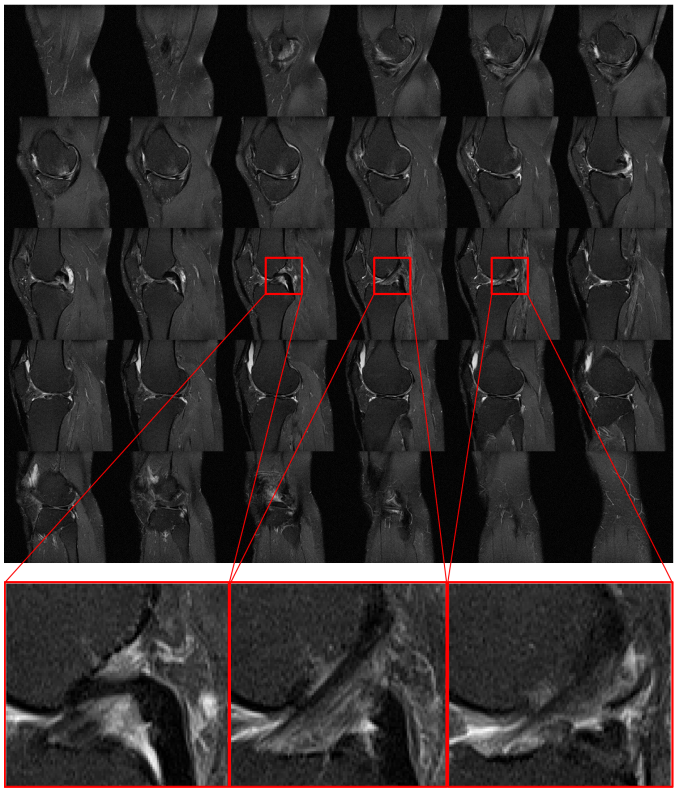


Figure 4: Manual extraction of a rectangular region of interest

It was empirically determined in later steps that this approach was more efficient than the alternative lossless rescaling (cropping/expanding the ROI). Next, we describe feature extraction mechanisms used in our experiments.

2.2. Feature extraction techniques

Image volumes are represented as 3D arrays containing voxel intensities. In this form, they cannot be directly handled by most machine-learning algorithms because of the overwhelming number of features, as opposed to the number of observations. Instead, we preprocess the volumes to extract smaller numbers of potentially useful features (descriptors) per observed volume. We examine two popular feature extraction techniques, namely histogram of oriented gradients and scene spatial envelope descriptors. Both are described next.

2.2.1. Histogram of oriented gradient descriptor

Histogram of oriented gradient (HOG) feature descriptors are nowadays commonly used for object detection in image processing. Initially developed for improving human detection in images [13], soon they were found to be equally convenient in solving various other problems. Some of the more recent uses of HOG descriptors in medical image analysis involve vertebrae detection and labelling in lumbar (spine) MR images [21], vocal folds detection on video laryngostroboscopy images [22], breast cancer diagnosis from mammographic images [23], prostate MR segmentation for prostate cancer diagnosis [24], lung tissue classification from chest CT images [25], and diagnosis of tuberculosis from chest X-ray images [26]. HOG descriptors are built by taking a non-linear function of image edge orientations (gradient) in a dense grid and pooling into smaller spatial regions with local contrast normalisation. The combined image histograms from the patches form the new representation. A visualisation of the calculated HOG descriptors from several randomly chosen cases is portrayed in Fig. 5. HOG descriptor representations are commonly used for learning linear-kernel support vector machine models for object detection. These models are described in section 2.3.1.

2.2.2. Scene spatial envelope descriptor

Gist descriptor [14] represents holistic spatial scene properties (spatial envelope) of an image. It summarises gradient information on different spatial scales and orientations by splitting the image into a grid of cells on several scales, and convolving each cell using a Gabor filter bank from different perspectives. Calculated responses are then concatenated to form the descriptor. It was first developed for the purpose of differentiating between several types of environmental scenes [14], but has since been applied to other problem domains also. Some of the more notable and recent uses of *gist* in medical image analysis involve automatic annotation of medical images concerning modality, body orientation, body region and biological system axes [27] and diagnosis of tuberculosis [26] and chest pathology detection [28] from chest X-ray images. Unlike HOG, *gist* is a global image descriptor. Therefore, we were interested in observing how well this holistic semantic approach would fare in the ACL detection problem domain. A visualisation of *gist* descriptors is depicted in Fig. 5. Although originally used with linear discriminant analysis for the purpose of categorising scenes [14], *gist* descriptors can also be used for learning more complex representations, if the underlying domain requires it.

2.3. Machine-learning techniques

Next, we describe two popular machine-learning classification models for dealing with the calculated descriptors extracted from image volumes: support vector machine and random forests model.

2.3.1. Support vector machine

Support vector machines (SVMs) [15] are one of the most popular supervised learning techniques. SVM models are learned from data by searching for a hyperplane in a high dimensional feature space, which separates the classes, optimising the generalisation bounds. A hyperplane optimising this measure is calculated using sequential minimal optimisation (SMO) with L^1 soft margin. If the classes are not linearly separable, non-linear kernels, such as polynomial kernels (quadratic, cubic, and so on) or radial basis function (RBF) kernels, can be used to implicitly transform the feature space. This expansion simplifies hyperplane separation at the cost of overfitting the data. Regardless of the kernel used, if the data in the transformed feature space, $\mathbf{x}$, is not linearly separable, target hyperplane parameters, $(\mathbf{w}, b)$, are estimated by minimising the cost function

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i, \tag{1}$$

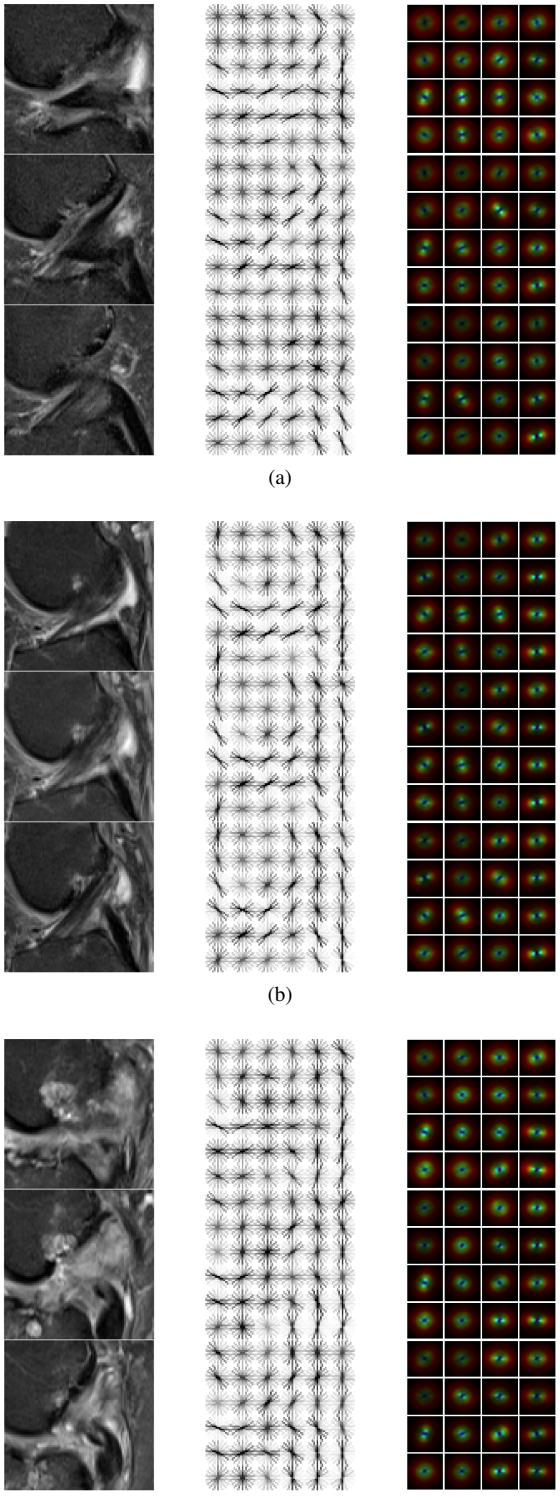


Figure 5: Visualisation of calculated feature descriptors for a randomly chosen case of each class: (a) not injured, (b) partially injured, and (c) completely ruptured. Left column depicts scaled 3-slice ROIs. Middle column depicts a visualisation of calculated HOG descriptors for patch size 15×15. Right column depicts a visualisation of calculated *gist* descriptors for a 4×4 grid of blocks.

subject to the following soft constraints

$$y_i[\mathbf{x}_i^\top \mathbf{w} + b] \geq 1 - \xi_i, \tag{2}$$

$$\xi_j \geq 0, \tag{3}$$

where $y_i \in \{-1, 1\}$ represents the label of the i-th instance, ξ_i represents its respective slack variable (allowing misclassification of the i-th instance) and C represents the box constraint. Larger value of C incurs a larger penalty regarding the distance from the separating hyperplane for misclassified instances in the cost function, thus forcing stricter separation between labels (also leading to model overfitting). On the other hand, smaller value of C, closer to 0, produces models which allow more misclassification by preferring simpler models. We refer the reader to [29, 30] for additional information regarding SVMs. There exist other boundary-optimisation algorithms, like twin SVMs [31], which have been successfully adapted and applied recently at solving brain MRI classification and pathology detection problems [32, 33]. It is important to note that SVMs are non-probabilistic binary classifiers. For the purpose of calculating the evaluation metrics, described in one of the following sections, posterior probability transform function was estimated from the scores and the labels, and then applied to the scores [34].

In this paper, we explore using linear-kernel SVMs and RBF-kernel SVMs. Time complexity of the SMO solver in both cases is roughly equal to $O(n^3)$, where n is the number of instances [35].

2.3.2. Random forests

Random forests (RF) model [16] is an ensemble of decision trees that can be used for modelling both classification and regression problems. Each decision tree forming an ensemble is learned separately from a subset of instances, randomly sampled from the entire dataset with replacement. When growing a tree, each node split is determined by observing only a randomly chosen subset of available features and selecting the one giving the best split. This combination of bagging and random feature subset selection ensures excellent generalisation properties of an RF model as the number of weak learners (unpruned individual trees) becomes large. In our experiments, the number of features forming a subset equalled square root of the total number of features, and the number of instances forming a data subset equalled the size of the dataset used for learning. Trees were grown to their full sizes (i.e., no depth limit). Time complexity of the algorithm for building an m-trees RF model learned from n instances by observing d features roughly equals $O(mnd \log n)$ [36].

Model performance evaluation metrics are described next.

2.4. Evaluation metrics

Ordinary scalar performance metrics, such as classification accuracy, sensitivity and specificity, can often be misleading when dealing with class-imbalanced data [37]. This is often solved by carefully setting suitable cost-function hyperparameters for cost-sensitive learning, tuning the function in such way that it penalises wrong classification of minority class instances more than it penalises wrong classification of majority class instances. Because finding suitable cost-function hyperparameters by hand can be quite exhausting, and bearing in mind that

the above mentioned performance metrics are not as good in describing the model strength, we decided to use their more powerful graphical counterpart - the receiver operating characteristic (ROC) curve. ROC curve is a graphical plot, commonly used for illustrating a model's predictive power under various discriminative threshold values [38]. It represents a relationship between the true-positive rate (sensitivity), against the false-positive rate (one minus specificity), at a given threshold. Following the curve, a proper decision threshold can be chosen by aligning the sensitivity/specificity trade-off, to reflect the needs of the problem at hand [6]. In addition, we used the area under the curve (AUC or AUROC) to get a quantitative measurement of the robustness of the models learned.

Another important metric, often used for evaluating predictive properties of models involving biomedical applications, is the F_1 score. It is calculated as a harmonic mean of precision and sensitivity [36]. Bearing in mind that the data used in this research was class imbalanced, we decided to observe those values at different probabilistic thresholds, ranging from 0.05 to 0.95, using step size 0.05.

In our experiments, evaluation metrics were calculated using stratified 10-fold cross-validation. Empirical performance was calculated by averaging the results obtained using the learning folds, whereas expected performance was calculated by averaging the results obtained using the test folds. Both types of performance were under inspection in order to keep track of possible data underfitting or overfitting (bias-variance tradeoff) in regard to the model used.

3. Results

Both feature extraction methods were paired with an appropriate-kernel SVM model-learning algorithm and an RF algorithm, thus producing a total of 4 independent experimental models whose names are self explanatory: HOG+linSVM, HOG+RF, GIST+rbfSVM, GIST+RF. Other combinations of described methods were investigated as well, but the results were not reported here in detail due to the fact that they had worse generalisation properties. E.g., using HOG descriptors coupled with polynomial (HOG+polSVM) or RBF (HOG+rbfSVM) kernels was prone to overfitting the SVM model. Similarly, using *gist* descriptors coupled with linear (GIST+linSVM) or polynomial (GIST+polSVM) kernels gave results that were far worse than the ones reported here. This did not come as a complete surprise because these combinations were shown to perform somewhat worse in certain applications [13, 39, 26]. Because the performance of a predictive model built using an algorithm for learning from extracted features was highly dependent on the choice of some hyperparameters, several most important ones were varied, while others were left as they were (default values used by certain tools or programming libraries). Hyperparameter ranges and step sizes were estimated from experience in order to cover the most promising scenarios, while retaining computational feasibility.

Prior to feature extraction, images were convolved by an isotropic Gaussian kernel of $\sigma = 1$ to reduce noise. We observed that better classifier performance was obtained for all of the experimental models when using this filter, as opposed to not using it. Convolving ROI slices with the Gaussian kernel has a smoothing effect on the images, which can be observed as a preprocessing step towards eliminating some of the unneeded variation in the data, such that would be otherwise embodied in the extracted descriptors, and possibly lead to overfitting. This smoothing effect was proven to be beneficial for use with the MRI volumes processed in this research. Model accuracy under different values of σ was not inspected. To summarize the volume preprocessing steps, after the manual extraction of a ROI, the extracted volume is first rescaled using linear interpolation to size $90 \times 90 \times 3$, then each volume slice is convolved separately with an isotropic Gaussian kernel of $\sigma = 1$, following the feature extraction phase which is described next.

HOG descriptors were extracted from ROI volume data using VLFeat [40], an open source library. For each of the volume slices, a separate descriptor vector was calculated using a quadratic patch of a certain size, which was varied in the experiments. Slice descriptor vectors were then concatenated to form a HOG feature vector for the ROI volume. Length of the resulting feature vector depended on the input patch size.

Gist descriptors were calculated using a set of tools provided by [14]. Number of spatial scales and orientations was used as suggested by [14], that is 4 spatial scales, each having 8 orientations, thus resulting in using 32 Gabor filters without any boundary extension. The only parameter that was varied was the number of blocks used to determine the coarseness of the descriptor (grid size). *Gist* descriptor was calculated for each slice in a ROI volume separately, later concatenating them to form ROI volume descriptors. Length of the resulting feature vector depended on the number of blocks used.

As can be seen in Fig. 5, morphological properties of a ROI under inspection are to some degree observable in certain segments of visualisations of the extracted features, both for HOG and *gist* descriptors. In this example, inter-class differences are the most visible when observing the middle slice only. When observing a healthy ACL, the gradient in its nearest region forms a feature visualisation in which one can clearly follow the shape of the ligament. This shape is more obfuscated in the partially-injured case, and is practically impossible to follow in the completely-ruptured case. These characteristics were exploited by the ML algorithms for differentiating between different clinical conditions. Although the differences between visualisations for the example depicted in Fig. 5 are easily distinguishable with the naked eye, this did not hold for a larger portion of the used data. Therefore, ML algorithms were utilised for finding a connection between the features and the observed clinical outcomes.

SVM training and testing functions and the score-to-posterior-probability transform function used for performing the experiments were ready-made commercial-software-package-native functions[1]. RBF kernel was calculated using

[1]MATLAB 2015a, The MathWorks, Inc., Natick, Massachusetts, United States

the expression

$$K(x, x') = \exp\Big(- \frac{\|x - x'\|^2}{2\sigma^2} \Big), \quad (4)$$

where $\sigma = 1$. Box constraint term, indirectly regulating the maximum number of allowed support vectors was varied in the experiments. RF training and testing functions used for performing the experiments were also commercial-software-package native. The seeds used were random. The number of trees in an RF was varied. Other details concerning both algorithm implementations are presented in section 2.3.1 and section 2.3.2.

3.1. Evaluation of detector accuracy

For each distinct hyperparameter setup concerning each of the 4 experimental models (HOG+linSVM, HOG+RF, GIST+rbfSVM, GIST+RF) applied to both problems (detecting either injured or completely ruptured cases), a full evaluation was performed using stratified 10-fold cross-validation. The results regarding the expected model performance were then summarised and are presented in Fig. 6. Hyperparameter ranges and values used can be interpreted from the plots. Basically, each experimental model was observed under 24 different hyperparameter value pairs. It took around 100 hours to perform these experiments on a computer having an I5 quad-core processor, running at 3.2GHz clock frequency, and having 32GB of DDR3 RAM. Consequently, finer graining of hyperparameter values or increasing them in range was not feasible. Nevertheless, reported results on expected AUC scores can be used as reference for conducting further experiments (e.g., the concavity and gradient of AUC surfaces depicted in Fig. 6). Details regarding specific execution times for feature extraction, learning and inferring is presented in Table 3. Descriptor extraction and model training times can get quite large if feature vectors are lengthy, especially when using RFs. On the other hand, time for parsing and classifying a single instance is under one second for all experimental scenarios.

After observing the estimated AUC values for different combinations of hyperparameter values in Fig. 6, most promising choices were selected and were then used with their respective experimental models, again by performing stratified 10-fold cross-validation, but this time over 10 iterations of equal fold random splits for all experimental models. Hyperparameter values used for performing each test, along with characteristics of extracted feature vectors, calculated AUC score iteration mean and standard deviation, are presented in Table 4. Related ROC curves describing both empirical and expected performance are shown in Fig. 7. For every experimental setup, each distinct ROC curve plotted in Fig. 7 represents performance of 10 distinct models, one for each iteration. Relative differences in standard deviations reported in Table 4 can also be observed in this plot. Finally, predictive properties of these models, observed under different probabilistic thresholds, are reported in Fig. 8, using F_1 score and sensitivity.

Best peak generalisation performance in terms of the AUC score was obtained using a linear-kernel SVM model trained from extracted HOG descriptors. For the problem of discriminating between injured-ACL cases and healthy-ACL cases (left column in Fig. 6), the model achieved an expected AUC of 0.894 using HOG patch size 10×10 and box constraint 0.01. Models for both smaller and larger patch sizes performed worse, regardless of the box constraint value used. Using smaller HOG patch sizes can be beneficial for describing local morphological characteristics of the observed area in more detail, but at the cost of overfitting the model. This is because smaller patches also produce larger feature vectors, which lead to the necessity for training more complex models, consisting of larger numbers of parameters. Given an equal number of input points (data instances), this can easily lead to overfitting. Comparable results were obtained using RF, where the highest performance was recorded using an equally sized HOG patch. Slightly worse results were obtained using RBF-kernel SVMs for learning from *gist* descriptors extracted using 3 blocks. Although the number of features obtained this way is rather small (only 864 features), it achieves good generalisation performance regardless of the box constraint value used (Fig. 6). When using a larger number of blocks, which results in many more extracted features, generalisation performance deteriorates rapidly due to overfitting. This is not the case when using RF. In this scenario, best performance is achieved using 15 blocks. A multitude of weak learners in an RF model is, therefore, obviously capable of generalising well when using a lengthier feature vector (21600 extracted features).

For the problem of detecting completely ruptured ACL cases only (right column in Fig. 6), best peak generalisation performance was obtained using again a linear-kernel SVM model trained from extracted HOG descriptors. Best model achieved an expected AUC of 0.943 using HOG patch size 5×5 (resulting in a larger number of features, compared to the previous scenario), having box constraint 1. Better generalisation performance when using lengthier feature vectors (30132 extracted features) can be attributed to the smaller intra-class variations of fully-ruptured cases, compared to the larger intra-class variations of normal, non-injured, cases. These differences in variations can be easily observed when manually inspecting the ROIs. Another factor, leading us to conclude that intra-class variations present in the fully-ruptured cases are smaller, is the fact that the role of the box constraint value used for training a discriminative model is relatively insignificant. Finally, reducing HOG patch size leads to a smaller deterioration in AUC score. Slightly worse generalisation performance is obtained using the RF model, using HOG descriptors having patch size 15×15 (3348 extracted features). For models learned from *gist* descriptors, peak generalisation AUC scores were both well below the ones obtained using HOG features. When observing the columns in Fig. 6, comparable performance characteristics can be observed for both, in regard to the hyperparameter values chosen.

Experimental results on the F_1 score and sensitivity, which can be observed in Fig. 8, are consistent with the conclusions drawn so far. A sensitivity of 90% or more can only be achieved by using a rather small decision threshold (between ≈ 0.05 and ≈ 0.15), greatly favouring the minority class, but at the ob-

Table 3: Code execution times for descriptor extraction (entire dataset), unit descriptor extraction (only for one instance), model learning from the entire dataset and unit inference (only for one instance), measured in seconds. Hyperparameter values used are presented in Table 4. Computer configuration: I5-4460 quad core processor, running at clock frequency 3.2GHz; 32GB of DDR3 RAM.

Experimental model	Descriptor extraction	Unit descriptor extraction	Model learning	Unit inference	Detection problem
HOG+linSVM	1.89925	0.00207	0.63236	0.00066	Injured
	2.35348	0.00256	1.89528	0.00262	Completely ruptured
HOG+RF	1.89650	0.00207	158.36788	0.04427	Injured
	1.85817	0.00203	72.39095	0.02894	Completely ruptured
GIST+rbfSVM	95.83711	0.10451	0.04398	0.00004	Injured
	96.44974	0.10517	0.02592	0.00002	Completely ruptured
GIST+RF	806.42644	0.87941	380.25262	0.09604	Injured
	809.67757	0.88296	376.60347	0.10781	Completely ruptured

Table 4: Comparison of all distinct experimental models performing under most promising hyperparameter values for both detection problems. HOG and *gist* hyperparameter values are supplemented with their respective numbers of resulting extracted features. AUC score means and standard deviations calculated from multiple iterations are presented. Best performing results (comparing experimental models) for each detection problem are emphasised.

Experimental model	Hyperparameter values				Score		Detection problem
	HOG patch size / #features	Gist #blocks / #features	SVM box constraint	RF #trees	AUC	σ(AUC)	
HOG+linSVM	10 / 7533	-	0.01	-	**0.894**	0.002	Injured
	5 / 30132	-	1	-	**0.943**	0.004	Completely ruptured
HOG+RF	10 / 7533	-	-	2000	0.884	0.002	Injured
	15 / 3348	-	-	2000	0.937	0.003	Completely ruptured
GIST+rbfSVM	-	3 / 864	1	-	0.889	0.001	Injured
	-	3 / 864	0.1	-	0.913	0.008	Completely ruptured
GIST+RF	-	15 /21600	-	2000	0.880	0.001	Injured
	-	15 / 21600	-	2000	0.895	0.003	Completely ruptured

vious expense of a greatly diminished specificity. Sensitivity-specificity tradeoffs can be observed by inspecting the F_1 score in the graphs. Again, the best results for both prediction problems (highest average and particular F_1 scores) are achieved using linear-kernel SVMs learned from HOG descriptors.

As was expected, for both detection problems, RF models having best performance were ensembles consisting of the largest number of trees used (2000). They were also the most time demanding for learning.

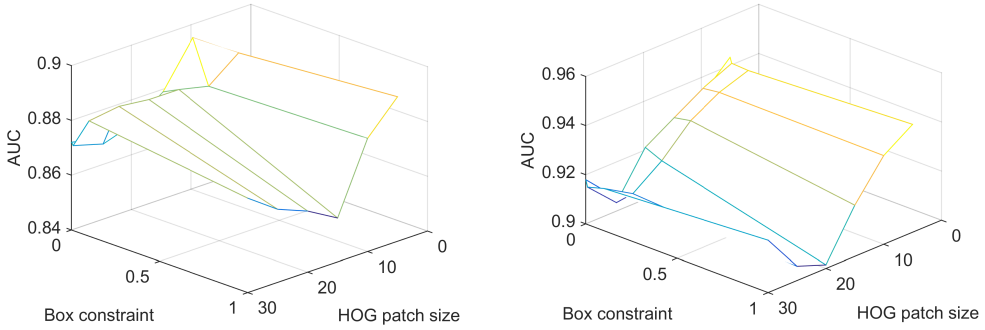
3.2. Influence of ROI selection on overall detector accuracy

Next, we were interested in observing the influence of the ROI selection phase on the ACL-condition detection phase. This is important for two reasons. First, this analysis gives us some idea on the possibility of presence of involuntarily introduced bias during the ROI extraction phase. Seeing that ROIs were extracted by a skilled radiologist, there exists a possibility that ROI position or shape is somewhat affected by the observed ACL condition. Second, this analysis gives us a rough estimate on the needed properties of an algorithm for determining the exact ROI position. This is an important step towards constructing a fully-automated CAD detection system.

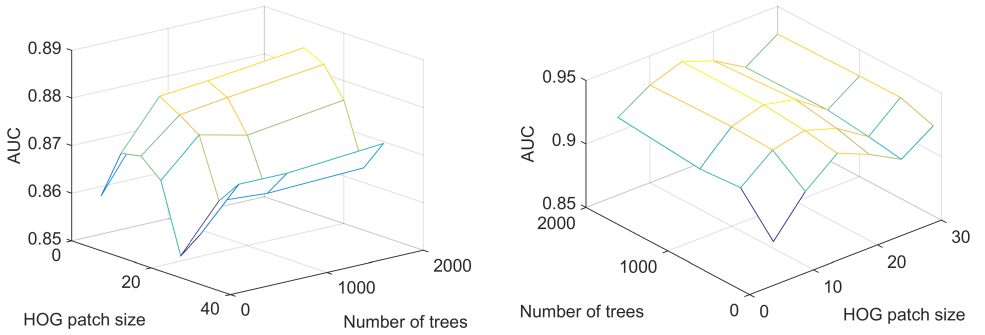
The influence of ROI selection on detector accuracy is measured by introducing a certain amount of error in the original ROI selection, and then calculating the AUC score. The data is corrupted by randomly expanding or contracting and shifting the ROI area along the sagittal plane. Specifically, in our implementation, starting and ending coordinates along the axes X and Y were modified by a percentage of their respective distances, multiplied by a number sampled from uniform distribution $\mathcal{U}$, e.g., $x_1 = x_1 + (x_2 - x_1) \cdot p \cdot \mathcal{U}(-1, 1)$, where p equals 0.03, 0.05 or 0.1 (3%, 5%, or 10%). The experiment is conducted by generating multiple instances of partially corrupted input datasets, learning and evaluating detector models from each dataset independently, using 10-fold cross validation, and then comparing the calculated means against the results reported in Table 4. Hyperparameter values used for performing this experiment are, again, equivalent to those presented in Table 4. The results are presented in Table 5. Our experiment did not include introducing noise along the third axis (Z) because the resolution along that axis is rather small (median value is 3 slices). To introduce an error along this axis, e.g. shifting the ROI by only one slice in either direction, would render the learned detector model almost useless.

Although a performance drop in terms of detector AUC is evident, it is rather small. Therefore, we can assume that slightly worse performance can be expected when ROIs are not optimally selected. For the problem of discriminating between injured-ACL cases and healthy-ACL cases, experimental models involving the use of SVMs have the smallest performance deterioration at 10% corruption (from $\approx$ 1.7% to $\approx$ 2.1% drop in AUC). For the problem of detecting completely ruptured cases, smallest performance deterioration at 10% corruption was observed for the RBF-kernel SVM, learned from *gist* descriptors ($\approx$ 3.3% drop in AUC). Regardless of relatively larger performance deterioration, linear-kernel SVMs learned from HOG descriptors retain the highest AUC scores for both classification scenarios.

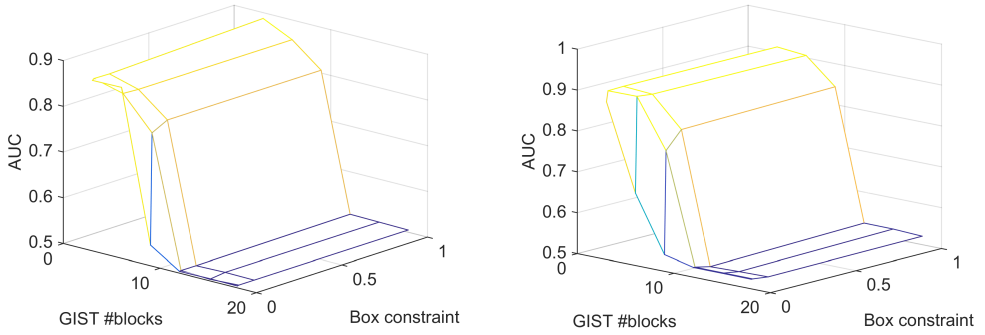
We believe that these results are influenced largely by the fact that suboptimal choice of a ROI introduces more unnecessary variation into the data, causing a negative change in classifier AUC. Furthermore, seeing that the change in AUC is rather small, we can assume that the ROI selection phase was not bi-



(a) HOG+linSVM



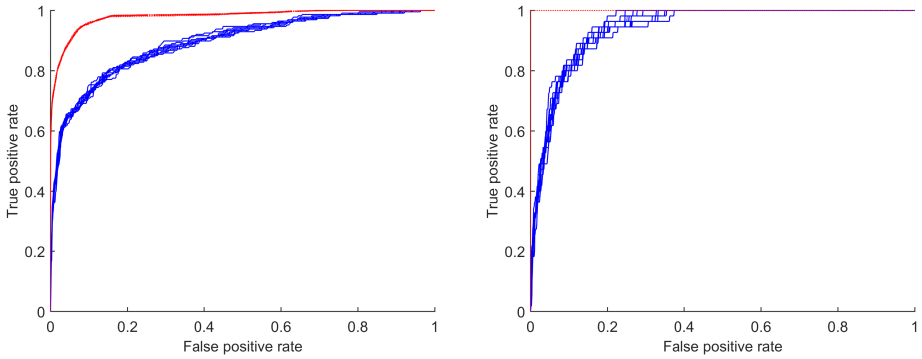
(b) HOG+RF



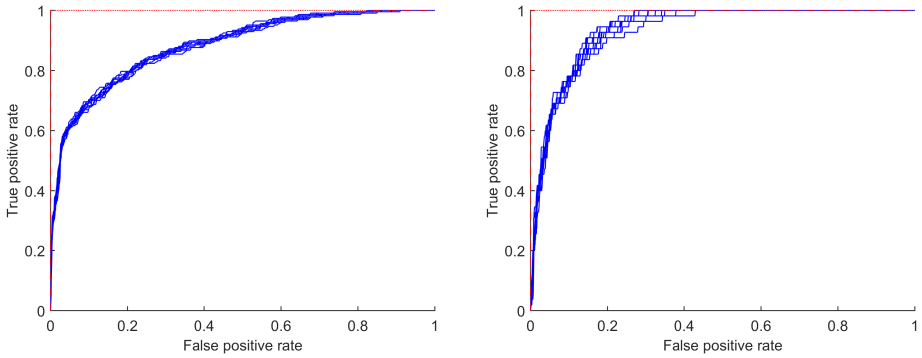
(c) GIST+rbfSVM

(d) GIST+RF

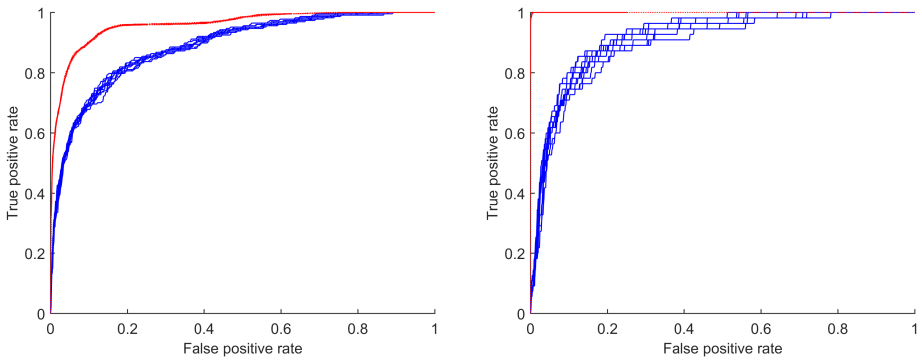
Figure 6: A linear interpolation of the expected AUC scores for the following experimental models: HOG+linSVM, HOG+RF, GIST+rbfSVM, and GIST+RF. Line intersections in the grids represent calculated AUC values. Column on the left represents the problem of detecting injured cases (partially injured or completely ruptured). Column on the right represents the problem of detecting completely ruptured cases only. Hyperparameter values that were under consideration can be interpreted from the plots.



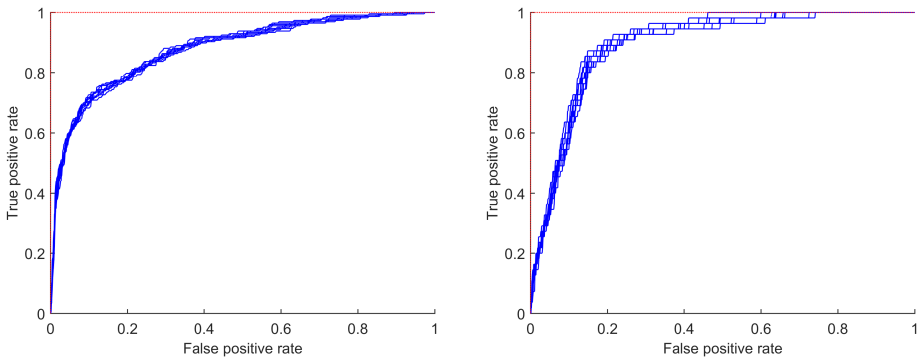
(a) HOG+linSVM



(b) HOG+RF

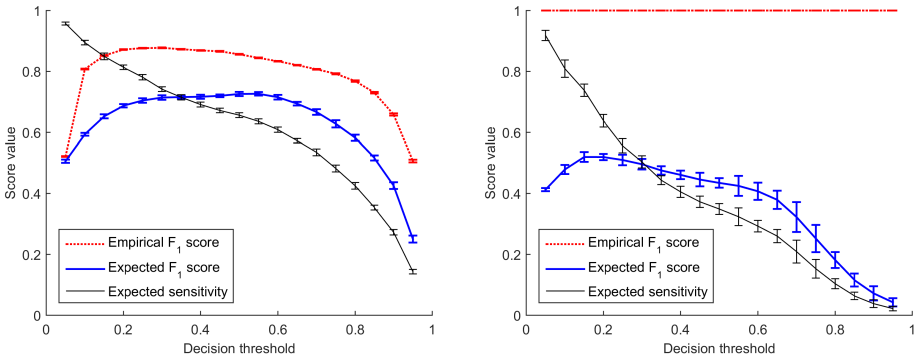


(c) GIST+rbfSVM

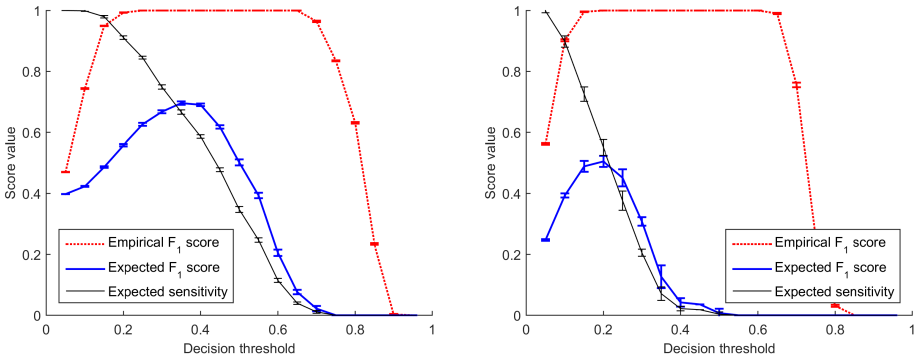


(d) GIST+RF

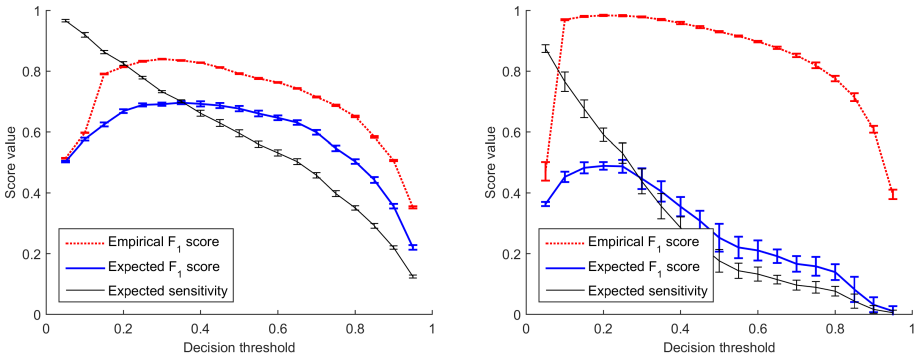
Figure 7: Expected ROC curves for the following experimental models: HOG+linSVM, HOG+RF, GIST+rbfSVM, and GIST+RF. Column on the left represents the problem of detecting injured cases (partially injured or completely ruptured). Column on the right represents the problem of detecting completely ruptured cases only. Expected ROC curves are presented using blue solid lines, whereas empirical ROC curves are presented using red dotted lines. Hyperparameter values used for each setup are stated in Table 4. Results were obtained in 10 iterations of stratified 10-fold cross-validation.



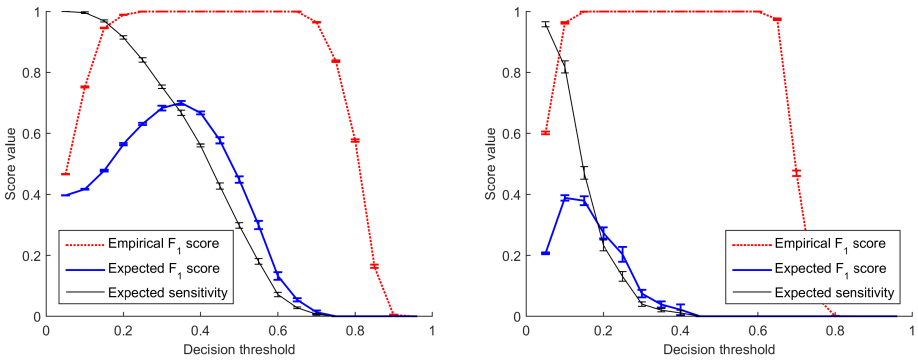
(a) HOG+linSVM



(b) HOG+RF



(c) GIST+rbfSVM



(d) GIST+RF

Figure 8: Calculated F_1 score and sensitivity values for different probability boundaries (decision thresholds), corresponding to the results presented in Fig. 7, based on the experimental setup described in Table 4: HOG+linSVM, HOG+RF, GIST+rbfSVM, and GIST+RF. Column on the left represents the problem of detecting injured cases (partially injured or completely ruptured). Column on the right represents the problem of detecting completely ruptured cases only.

Table 5: Influence of ROI selection on overall detector accuracy. Sagittal plane ROI coordinates are varied by 3%, 5% or 10%, relative to the ROI size of a distinct axis. The third axis is not varied, due to the small number of slices used. All distinct experimental models performing under the most promising hyperparameter values (Table 4) for both detection problems are compared. AUC score means and standard deviations calculated from multiple iterations are presented. Best performing results (comparing experimental models) for each detection problem are emphasised.

Experimental model	Origin	Variation level introduced						Relative AUC change			Detection problem
		3%		5%		10%		3%	5%	10%	
	AUC	AUC	σ(AUC)	AUC	σ(AUC)	AUC	σ(AUC)				
HOG+linSVM	0.894	**0.891**	0.002	**0.885**	0.003	**0.875**	0.006	-0.298	-0.969	-2.088	Injured
	0.943	**0.938**	0.005	**0.927**	0.006	**0.903**	0.017	-0.495	-1.661	-4.242	Completely ruptured
HOG+RF	0.884	0.878	0.002	0.879	0.001	0.858	0.004	-0.641	**-0.566**	-2.941	Injured
	0.937	0.933	0.002	0.922	0.006	0.892	0.009	-0.427	-1.636	-4.838	Completely ruptured
GIST+rbfSVM	0.889	0.887	0.001	0.882	0.004	0.874	0.002	**-0.262**	-0.750	**-1.687**	Injured
	0.913	0.910	0.005	0.911	0.005	0.888	0.009	**-0.292**	**-0.219**	**-2.738**	Completely ruptured
GIST+RF	0.880	0.876	0.002	0.871	0.006	0.849	0.009	-0.492	-1.023	-3.561	Injured
	0.895	0.883	0.004	0.862	0.002	0.823	0.014	-1.378	-3.724	-8.045	Completely ruptured

ased with the observed ACL condition, thus the data used in these experiments is plausible. If ROI selection was even more corrupted, classifier performance would surely degrade even further. It would be interesting to observe whether a different choice of hyperparameters, e.g. using a smaller box constraint, would improve the experimental results.

4. Discussion and conclusion

Computer-aided diagnosis, with its ability to advise medical specialists in their decision-making process, plays an important role in today's world. Decision-support models have often in the past been created by manual assembly of prior specialist knowledge, but are nowadays more often constructed or learned directly from existing data. Help with decision making is of paramount importance especially when the amount of information that needs to be considered for establishing a diagnosis is large, e.g. recognising objects in an obfuscated image. This is often the case when analysing radiology images.

One of such cases is the thorough analysis of a human knee from MRI, which takes into account the condition of ligaments, menisci, cartilage, bones, and so on. Anterior cruciate ligament is the most commonly injured ligament in the human body. A decision-support system that is able to differentiate between normal and injured ACLs would aid in establishing diagnosis and preventing human errors. Same goes for the problem of detecting completely ruptured ACLs, which could be used as an early warning system for both patients and hospitals, notifying them of an impending operative treatment, thus allowing them to immediately plan ahead.

In this paper, we present in detail the possibilities and the difficulties encountered regarding constructing an ACL-injury classification model for semi-automated diagnosis from knee MRI data. We also discuss the feasibility of building a fully-automated system through an automated selection of the ROI boundaries, based on a priorly constructed reference volume. We study the problem of differentiating between healthy and injured knees as well as the problem of detecting only completely ruptured cases (regarding the ACL condition). We treat these problems as separate binary classification problems instead of a single 3-class classification problem, because of the inherent imbalanced distribution present in the data, following the need for utilising binary-class performance measures. We compare two feature extraction techniques, paired with two popular machine-learning models. Linear-kernel SVM models used for performing supervised learning from HOG descriptors achieved the best performance in terms of the AUC score. Experimental results suggest that this method has clinical potential for differentiating complete ACL tears (AUC=0.943) from other cases. All of the experimental models used in this work can be considered suitable for real-time CAD, because the inference algorithm execution times are just under one second, using a standard desktop computer. The performance decreases when differentiating between the non-injured and the remaining cases (AUC=0.894). This is largely due to the fact that many of the partially-injured cases are rather hard to distinguish from the non-injured ones in the available data, even by an expert radiologist [18]. Although we demonstrate that both HOG and scene-spatial-envelope descriptors have excellent properties in this application, the question is whether they were able to fully delineate the representation needed for such discrimination. However, the results reported in this work constitute a good starting point for the challenging computer-aided ACL injury detection problem.

Acknowledgment

This research was funded by The Scientific & Technological Research Council Of Turkey (TÜBİTAK 2221 Programme), Croatian Science Foundation's funding of the project UIP-2014-09-7945 and by the University of Rijeka Research Grant 13.09.2.2.16.

References

[1] K. P. Spindler, R. W. Wright, Anterior cruciate ligament tear, New England Journal of Medicine 359 (20) (2008) 2135–2142.

[2] D. Nenezić, I. Kocijancic, The value of the sagittal-oblique MRI technique for injuries of the anterior cruciate ligament in the knee, Radiology and oncology 47 (1) (2013) 19–25.

[3] K. Doi, Computer-aided diagnosis in medical imaging: Historical review, current status and future potential, Computerized Medical Imaging and Graphics 31 (4-5) (2007) 198–211. doi:10.1016/j.compmedimag.2007.02.002.

[4] A. P. Kansagra, J.-P. J. Yu, A. R. Chatterjee, L. Lenchik, D. S. Chow, A. B. Prater, J. Yeh, A. M. Doshi, C. M. Hawkins, M. E. Heilbrun, S. E.

Smith, M. Oselkin, P. Gupta, S. Ali, Big Data and the Future of Radiology Informatics, Academic Radiology 23 (1) (2016) 30–42.

[5] V. Jain, H. S. Seung, S. C. Turaga, Machines that learn to segment images: A crucial technology for connectomics, Current Opinion in Neurobiology 20 (5) (2010) 653–666. doi:10.1016/j.conb.2010.07.004.

[6] S. Wang, R. M. Summers, Machine learning and radiology, Medical Image Analysis 16 (5) (2012) 933–951. doi:10.1016/j.media.2012.02.005.

[7] C. Köse, O. Gençalioğlu, U. Şevik, An automatic diagnosis method for the knee meniscus tears in MR images, Expert Systems with Applications 36 (2) (2009) 1208–1216. doi:10.1016/j.eswa.2007.11.036.

[8] H. Oka, S. Muraki, T. Akune, A. Mabuchi, T. Suzuki, H. Yoshida, S. Yamamoto, K. Nakamura, N. Yoshimura, H. Kawaguchi, Fully automatic quantification of knee osteoarthritis severity on plain radiographs, Osteoarthritis and Cartilage 16 (11) (2008) 1300–1306.

[9] L. Shamir, S. M. S. Ling, W. W. Scott, A. Bos, N. Orlov, T. J. MacUra, D. M. Eckley, L. Ferrucci, I. G. Goldberg, Knee X-ray image analysis method for automated detection of osteoarthritis, IEEE Transactions on Biomedical Engineering 56 (2) (2009) 407–415. doi:10.1109/TBME.2008.2006025.Knee.

[10] J. Fripp, S. Crozier, S. K. Warfield, S. Ourselin, Automatic segmentation and quantitative analysis of the articular cartilages from magnetic resonance images of the knee., IEEE transactions on medical imaging 29 (1) (2010) 55–64. doi:10.1109/TMI.2009.2024743.

[11] G. Vincent, C. Wolstenholme, I. Scott, M. Bowes, Fully automatic segmentation of the knee joint using active appearance models, in: Medical Image Analysis for the Clinic: A Grand Challenge, CreateSpace, 2010, pp. 224–230.

[12] Y. Yin, X. Zhang, R. Williams, X. Wu, D. D. Anderson, M. Sonka, LOGISMOS—layered optimal graph image segmentation of multiple objects and surfaces: cartilage segmentation in the knee joint, Medical Imaging, IEEE Transactions on 29 (12) (2010) 2023–2037.

[13] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, 2005, pp. 886–893. arXiv:9411012, doi:10.1109/CVPR.2005.177.

[14] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International journal of computer vision.

[15] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[16] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[17] F. Roemer, M. Crema, S. Trattnig, A. Guermazi, Advances in imaging of osteoarthritis and cartilage, Radiology 260 (2) (2011) 332–354.

[18] T. W. Hash, Magnetic resonance imaging of the knee., Sports health 5 (1) (2013) 78–107. doi:10.1177/1941738112468416.

[19] B. Zitová, J. Flusser, Image registration methods: a survey, Image and Vision Computing 21 (11) (2003) 977–1000. doi:10.1016/S0262-8856(03)00137-9.

[20] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, NeuroImage 17 (2) (2002) 825–841. arXiv:arXiv:1011.1669v3, doi:10.1016/S1053-8119(02)91132-8.

[21] M. Lootus, T. Kadir, A. Zisserman, Vertebrae Detection and Labelling in Lumbar MR Images, 16th International Conference on Medical Image Computing and Computer Assisted Intervention, Computational Methods and Clinical Applications for Spine Imaging, Lecture Notes in Computational Vision and Biomechanics 17 (2014) 219–230.

[22] H. Irem Turkmen, M. Elif Karsligil, I. Kocak, Classification of laryngeal disorders based on shape and vascular defects of vocal folds., Computers in biology and medicine 62 (2015) 76–85. doi:10.1016/j.compbiomed.2015.02.001.

[23] D. Moura, M. López, An evaluation of image descriptors combined with clinical data for breast cancer diagnosis, International Journal of Computer Assisted Radiology and Surgery 8 (4) (2013) 561–574.

[24] S. Liao, Y. Gao, A. Oto, D. Shen, Representation learning: A unified deep learning framework for automatic prostate MR segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Vol. 16, 2013, pp. 254–261.

[25] Y. Song, W. Cai, Y. Zhou, D. D. Feng, Feature-based image patch approximation for lung tissue classification., Medical Imaging, IEEE Transactions on 32 (4) (2013) 797–808. doi:10.1109/TMI.2013.2241448.

[26] A. Chauhan, D. Chauhan, C. Rout, Role of Gist and PHOG Features in Computer-Aided Diagnosis of Tuberculosis without Segmentation, PLoS ONE 9 (11) (2014) e112980. doi:10.1371/journal.pone.0112980.

[27] J. Kalpathy-Cramer, W. Hersh, Medical image retrieval and automatic annotation: OHSU at ImageCLEF 2007, Advances in Multilingual and Multimodal Information Retrieval (2008) 623–630.

[28] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, H. Greenspan, Chest pathology detection using deep learning with non-medical training, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 294–297. doi:10.1109/ISBI.2015.7163871.

[29] N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, 1st Edition, Cambridge University Press, 2000.

[30] C. Bishop, Pattern Recognition and Machine Learning, 1st Edition, Springer-Verlag New York, 2006.

[31] Jayadeva, R. Khemchandani, S. Chandra, Twin support vector machines for pattern classification, Pattern Analysis and Machine Intelligence, IEEE Transactions on 29 (5) (2007) 905–910. doi:10.1109/TPAMI.2007.1068.

[32] Y. Zhang, Z. Dong, A. Liu, S. Wang, G. Ji, Magnetic resonance brain image classification via stationary wavelet transform and generalized eigenvalue proximal support vector machine, Journal of Medical Imaging and Health Informatics 5 (7) (2015) 1395–1403.

[33] S. Wang, S. Lu, Z. Dong, J. Yang, M. Yang, Y. Zhang, Dual-Tree Complex Wavelet Transform and Twin Support Vector Machine for Pathological Brain Detection, Applied Sciences 6 (6).

[34] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers 10 (3) (1999) 61–74.

[35] J. Platt, Sequential minimal optimization: A fast algorithm for training support vector machines, Tech. rep., Microsoft Research (1998).

[36] I. Witten, E. Frank, M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann, 2011.

[37] Y. Sun, A. K. C. Wong, M. S. Kame, Classification of Imbalanced Data: a Review, International Journal of Pattern Recognition & Artificial Intelligence 23 (4) (2009) 687–719.

[38] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve., Radiology 143 (1) (1982) 29–36.

[39] H. Bristow, S. Lucey, Why do linear SVMs trained on HOG features perform so well?, arXiv preprint arXiv:1406.2419arXiv:arXiv:1406.2419v1.

[40] A. Vedaldi, B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1469–1472.