Int. J. Appl. Math. Comput. Sci., 2019, Vol. 29, No. 1, 51–67
DOI: 10.2478/amcs-2019-0004

amcs

FUSION OF CLINICAL DATA: A CASE STUDY TO PREDICT THE TYPE OF TREATMENT OF BONE FRACTURES

ANAM HAQ [a,*], SZYMON WILK [a], ALBERTO ABELLÓ [b]

[a] Institute of Computing Science
Poznan University of Technology, Pl. M. Skłodowskiej-Curie 5, 60-965 Poznan, Poland
e-mail: anam.haq@put.poznan.pl, szymon.wilk@cs.put.poznan.pl

[b] Department of Informatics
Polytechnic University of Catalonia, Campus Nord, Carrer de Jordi Girona, 1–3, 08034 Barcelona, Spain
e-mail: aabello@essi.upc.edu

A prominent characteristic of clinical data is their heterogeneity—such data include structured examination records and laboratory results, unstructured clinical notes, raw and tagged images, and genomic data. This heterogeneity poses a formidable challenge while constructing diagnostic and therapeutic decision models that are currently based on single modalities and are not able to use data in different formats and structures. This limitation may be addressed using data fusion methods. In this paper, we describe a case study where we aimed at developing data fusion models that resulted in various therapeutic decision models for predicting the type of treatment (surgical vs. non-surgical) for patients with bone fractures. We considered six different approaches to integrate clinical data: one fusion model based on combination of data (COD) and five models based on combination of interpretation (COI). Experimental results showed that the decision model constructed following COI fusion models is more accurate than decision models employing COD. Moreover, statistical analysis using the one-way ANOVA test revealed that there were two groups of constructed decision models, each containing the set of three different models. The results highlighted that the behavior of models within a group can be similar, although it may vary between different groups.

Keywords: clinical data, data fusion, combination of data, combination of interpretation, prediction models, decision support.

1. Introduction

Intuitively, data fusion can be illustrated by explaining the working of the human brain system. To understand and perceive the surrounding conditions, the human brain initially gathers content or relevant information from all senses such as sight, hearing, smell, taste, and touch. It then performs integration (fusion) by bringing together results that correspond to a conclusive output. In other words, the conclusion extracted from the senses is integrated together and with past data and experience, and thus the brain generates actions accordingly. Another biological example is the human visual perception system. The field of view of each eye is limited, but their combination provides us with an extended field view of 210°. From these examples, we can understand the importance of the data fusion concept. According to a more technical definition, data fusion is described as combination of information or data acquired from various sources of diversified formats, structures and incremental learning experiences (Mitchell, 2014).

There have been numerous successful applications of data fusion in geospatial systems, intelligent services, and surveillance systems (see the work of Castanedo (2013) for a review). This increases the acceptability and confidence in employing data fusion methods in health care (Lahat *et al.*, 2015), where they can provide substantial help in constructing clinical decision support systems (CDSSs) and smart patient monitoring systems.

Currently, decision models in most CDSSs make use of a single kind of clinical data. However, they can prove to be more efficient if we are able to combine clinical data obtained from different sources and in

*Corresponding author

different formats, i.e., patient demographics, images, lab results or genomic data. Therefore, in order to develop such an adequate decision model, information from every heterogeneous source has to be transformed into a common homogeneous space, which is one of the major challenges that is associated with the process of data fusion. Some of the other challenges linked with clinical data fusion are the following:

1. extraction of relevant features from heterogeneous data sources,
2. transformation of heterogeneous information into a homogeneous format,
3. selection of an appropriate data fusion method, i.e., combination of data (COD) or combination of interpretation (COI).

The most prevalent data fusion methods are *combination of data* (COD) and *combination of interpretation* (COI). Both the techniques are explained below.

COD assumes that in the first stage (also known as the "aggregation phase") all the extracted features from given data sources are initially aggregated into a uniform data space. A single classifier (base classifier) is constructed from this space.

In COI, for every data source, a separate classifier is constructed. All individual outcomes are then subject to the aggregation phase carried out by a "combiner." The latter can be regarded as a base classifier that generates a final decision or outcome. COI resembles an ensemble of classifiers (Ponti, 2011) (in particular the stacking scheme).

Unfortunately, none of these two techniques has given a complete solution to the challenges indicated above, and they both have their own inherent drawbacks. The biggest drawback associated with COD is the curse of dimensionality. In contrast, COI is subject to sub-optimality as it cannot preserve the dependencies between data from different sources. To address these shortcomings, Lee *et al*. (2009) developed a *general fusion framework* (GFF) where COI and COD are considered two extremes of a continuous spectrum.

The GFF is illustrated in Fig. 1. It employs multiple transformations which are applied to selected data sources to bring the data into a common space. In general, there are two types of transformations: simple—aimed at data pre-processing, and complex—aimed at constructing classifiers. In the latter case classification outcomes become part of the common space. Transformations are effectively guided by data formats and characteristics—for example, pre-processing transformations for clinical images employ various feature extraction schemes, and transformations for genetic data usually rely on dimensionality reduction schemes like principal component analysis (PCA) (Lee *et al*., 2009). If separate classifiers are constructed for all sources, then we have COI. On the other hand, if all transformations applied only pre-processes data, then the GFF boils down to COD.

In this paper, we present our clinical case study aimed at building a therapeutic decision model (classifier) to suggest an appropriate treatment (surgical or non-surgical) for patients with bone fractures. Such decisions should be based on general patient characteristics, the result of the physical examination and laboratory results (i.e., non-image data) as well as X-ray images. Given this, we apply data fusion—in particular, we consider various data fusion models derived systematically using the GFF and evaluate their impact on the accuracy of resulting classifiers.

In this work, we significantly extend our previous analysis (Haq and Wilk, 2017) by

1. increasing the amount of data used in experiments from 103 to 210 patients;
2. considering five COI data fusion models in addition to a COD model;
3. performing statistical analysis of performance demonstrated by decision models constructed according to specific data fusion models;
4. comparing the performance of the obtained decision models with that of ensemble classifiers constructed using bagging and boosting schemes.

Our goal was to examine not only the basic COD and COI models, as suggested in previous studies (Rohlfing *et al*., 2005), but also how varying complexity of COI

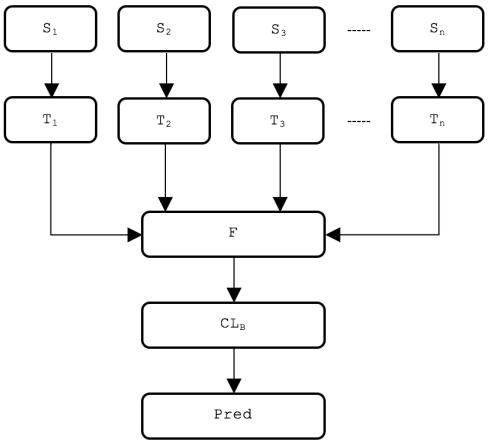


Fig. 1. Schema of GFF (S_i: i-th data source, T_i: transformation applied to S_i, F: common data space, CL_B: base classifier, $pred$: decision outcome of CL_B).

fusion models affects the performance of derived decision models. Therefore, we considered three variants of simple COI models and two variants of complex COI models. According to the experimental results, the COI fusion models proved to be more accurate decision models than COD. This observation was further supported by statistical tests with ANOVA. Moreover, decision models created following COI were better in performance than ensemble classifiers based on bagging and boosting schemes. The proposed COI fusion models can be applied to other problems where image and non-image data need to be considered, in particular to clinical ones. However, among the possible five versions of COI models, those which have separate classifiers for each data source (image and non-image) provide more balance performance across classes. Therefore, in certain circumstances they may be preferred by decision makers (clinicians).

The rest of the paper is organized as follows. In the next section, we present an overview of related work on data fusion and its clinical applications. Then, in Section 3, we introduce our case study and describe its goals, available data sources, and the customized data fusion process. Next, in Section 4, we report the results of our analysis. Finally, in Section 5, we provide conclusions and discuss future work.

2. Related work

CDSSs are mostly based on a single modality, i.e., they process data coming from a single source and represented in a single format (which can be an image, demographics combined with examination and lab results, clinical notes or genomic information). Image processing techniques prove to be very effective for detecting abnormalities in clinical images, and thus they are often employed in image-based CDSSs. In such systems, various relevant features describing the structure and texture of the abnormality are extracted from images, and then learning algorithms are used to build a decision model. Recent developments in deep learning (in particular deep neural networks) allow constructing accurate decision models without explicit feature extraction phase (see the work of de Bruijne (2016) for a review). However, deep neural networks require a significant amount of learning images, which may be not always available. Apart from image-based CDSSs, there exist those that rely solely on non-image clinical information. This group may be represented by systems for diagnosing diabetes (Sim *et al*., 2017). Finally, there is also a growing group of CDSSs built using genomic data, thus providing support for personalized medicine (Douali and Jaulent, 2012).

Specific individual modalities have proven to be useful when building decision models for CDSSs. Their further integration should allow obtaining a more comprehensive description of patients, and thus result in more accurate CDSSs (Viswanath *et al*., 2017).

Applications of the COD method are discussed in detail by Lanckriet *et al*. (2004) and Kourou *et al*. (2015). The former applied COD to develop a support vector machine (SVM) for generating predictions of the yeast proteins function. The proposed technique made use of kernel sets to combine complex protein data, gene expressions and, amino acids. The latter presented the results obtained by applying COD to construct decision models for detection and prognosis of various types of cancers.

Selected applications of COI are presented in detail by Ponti (2011), Jesneck *et al*. (2006) or Zorluoglu and Agaoglu (2015). Jesneck *et al*. (2006) used this method to combine objective findings obtained from mammograms, radiologist-interpreted findings and patient history for breast cancer diagnosis. The authors employed detection theory to construct classifiers. Specifically, a distinct binary classifier exploiting the concept of the likelihood ratio was formulated for each set of sources. The outputs of these classifiers were then combined using a base classifier.

As already mentioned, COI is similar to constructing ensembles of classifiers, more so in the case of the mixture of experts (ME) (a brief overview of ME models is presented by Yuksel *et al*. (2012)) and stacking schemes. A model for breast cancer diagnosis was developed by Zorluoglu and Agaoglu (2015) by making use of the decision tree, support vector machines and neural networks, along with the COI model of these three techniques. The conclusion presented was based on comparisons of individual classifiers' performance with the ensemble of these classifiers.

A functional comparison of COI and COD is provided by Rohlfing *et al*. (2005). The aim of that study was to demonstrate the importance of data fusion in image processing. Since no ground truth was available, performance evaluation of both techniques was informal and based on subjective observations (visual assessment). The experiment involved the development of models for four different kinds of biomedical image processing functions. These included segmentation of atlas-based images, multi-spectral classification, average image tissue based segmentation and deformation based group morphometry. Performance of decision models obtained with COI and COD was compared by the authors based on the capacity of producing reliable and consistent results along with versatility.

Viswanath *et al*. (2017) made a comparison of various dimensionality reduction (DR) schemes when applied in the context of data fusion. They considered multiple data fusion approaches elaborated by Lee *et al*. (2009) and Tiwari *et al*. (2011), applied to genomic data and images, and concluded that the choice of the

fusion scheme was dependent upon the type of data under consideration. Also, diagnostic systems based on a single modality were consistently less accurate as they ignore other aspects of relevant patient information.

Except for our earlier research (Haq and Wilk, 2017), there are no other papers discussing exactly the same decision problem. However, several similar issues have been considered. For example, Edward and Hepzibah (2015) developed a system to identify the type of bone fracture by first locating the fracture and then extracting shape features (e.g., area, perimeter) from the identified location. These features were then used to train a neural network that achieved an overall accuracy of 90%. In the work of Al-Ayyoub and Al-Zghool (2014), fractures in long bones along with the type of fracture were detected using image processing algorithms, and performance of classifiers learned on the basis of extracted features was evaluated. The best accuracy of 85% was demonstrated by an SVM classifier.

3. Case study

3.1. Problem statement. From the clinical perspective, it is important to distinguish between these patients with bone fractures who require surgery and those who can be managed non-surgically. Surgical treatment is not only more expensive than a non-invasive one, but it is also more painful. Several studies have shown that surgery is not needed in every case (Hossain *et al*., 2008). Moreover, there is a group of patients who may be not sufficiently clinically fit for the surgery. Thus, the decision about the type of treatment should be based not only on the characteristics of a fracture captured on X-ray images, but also on the patient's "fitness", and in order to develop an appropriate therapeutic decision, model image and non-image data should be fused together.

In this study, we respond to the challenge mentioned above and use data fusion techniques to build a decision model to support therapeutic decisions for the patient with fractures. This decision model takes into account complete characteristics of a patient involving both image and non-image information (see the next section for a detailed description of data sources) and suggests one of two possible types of treatment: surgical and non-surgical. Such a model could be integrated into computer-aided diagnostic (CAD) tools that automatically identify the presence and type (severity) of bone fracture (Khatik, 2017) in order to offer a comprehensive decision support during subsequent stages of the management process.

3.2. Data sources. In this case study, we used an educational registry of cases provided by the Wielkopolska Center of Telemedicine (Brzezinski *et al*., 2013)—a teleconsultation platform for patients with multiple injuries. This data set is part of the vast teaching resources provided by this platform (other resources include video lectures and clinical algorithms), and it includes 2030 patients with bone fractures: 1593 (78.5%) underwent a surgery and the remaining 437 (21.5%) were treated non-surgically. Patients are described using 301 features capturing demographics (e.g., age and gender), results of physical examinations and basic laboratory tests (e.g., blood work), and detailed structured characterization of injuries. For the sake of simplicity, we will refer to these data and features as clinical data and clinical features, respectively. Moreover, for nearly every patient there is a collection of X-ray images (usually between 2 and 4) of fractured bones.

From the available data set, we randomly selected 210 patients: 76 (36.2%) non-surgical and 134 (63.8%) surgical cases. Such a distribution was established following the suggestions by Dittman *et al*. (2014), who advocate the 35:65 sampling ratio based on their experiments with biomedical data. We changed the distribution of classes in comparison with the entire set to make the resulting classifiers less biased towards the surgical class. While there are other more sophisticated techniques to make the distribution of classes more balanced than random resampling (e.g., Koziarski and Woźniak, 2017), they often introduce synthetic examples and such "artificial patients" may be questionable from a clinical perspective; therefore, we did not consider such methods in this study. Finally, for each patient considered, we selected a single X-ray image showing the initial state of the fractured bone—we will refer to these selected images as image data.

3.3. Data fusion process. The overall goal of our study was to build a therapeutic decision model to predict the type of treatment for patients with fractures. When achieving this goal, we wanted to demonstrate the benefits of data fusion, in particular of combining both image and clinical data. Our secondary specific goal was to apply, evaluate and compare different variants of COD and COI derived systematically from the GFF. Specifically, we considered six different data fusion models (one COD model and five COI models) presented in Fig. 2. They are described in the latter part of this section.

All the fusion models considered rely on two data sources, S_C and S_I, corresponding to clinical and image data, respectively. Moreover, in all six models, the same transformations T_C and T_I were applied to S_C and S_I, respectively. These transformations employed diversified techniques of feature extraction, construction, and selection. The transformation T_I was aimed at extracting from an X-ray image two numerical features (see the description below) indicating the "severity" of a fracture (such information was not recorded explicitly in the clinical data), and it included the following steps:

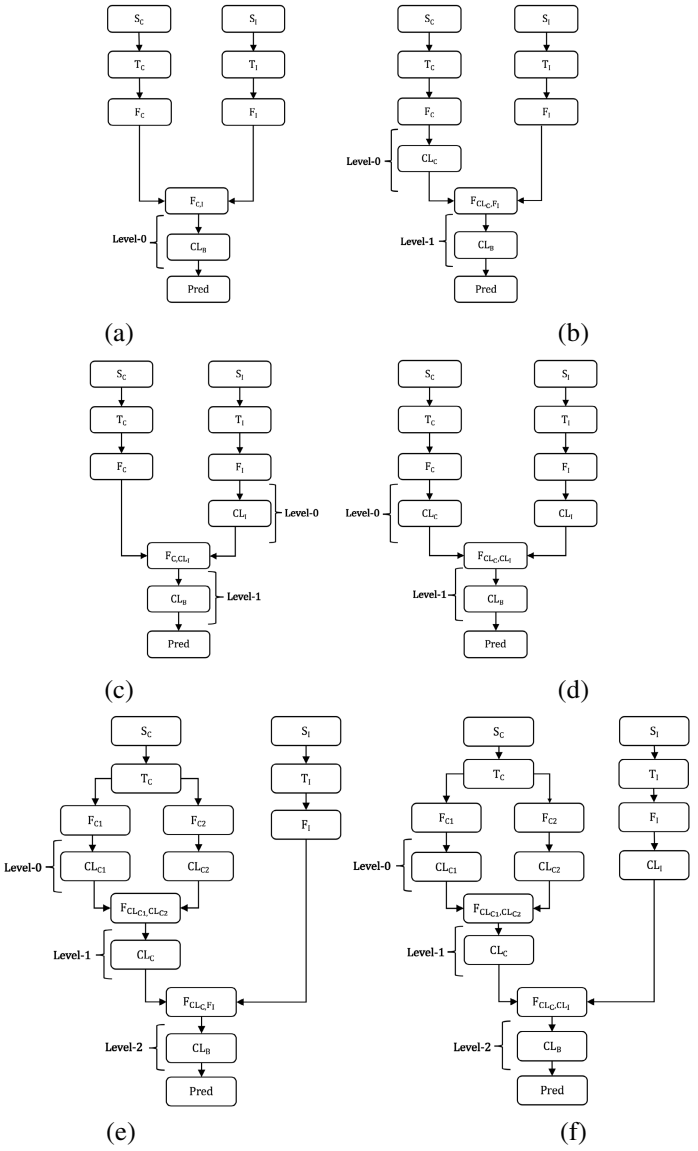


Fig. 2. Data fusion models considered in our study: COD (a), S-COI-C (b), S-COI-I (c), S-COI-CI (d), C-COI-C (e), C-COI-CI (f).

1. noise removal with a median filter and contrast adjustment;
2. bone edge detection with the Canny operator and removal of disconnected components;
3. application of the Hough transform to detect the bone fracture (see the work of Haq and Wilk (2017) for details). The parameter values were set in such a way that the transform produced two peak values for minor fractures (see Fig. 3(a)) and multiple peak values for significant fracture bones (see Fig. 3(b))

Application of T_I to S_I results in the data space F_I, where each patient was described with values of two image-related features—the mean value and standard deviation of peak points from the Hough transform.

The transformation T_C applied to S_C was less complex and it involved the following steps:

1. discretization of numerical features capturing results of laboratory tests using norms defined by clinical experts;
2. introduction of additional features capturing information about injuries at a lower granularity level;
3. removal of "useless" features, e.g., those with the majority of missing values (more than 90%), or extremely low or high variance.

Some of the above steps may require additional explanation. First, in preliminary experiments, we

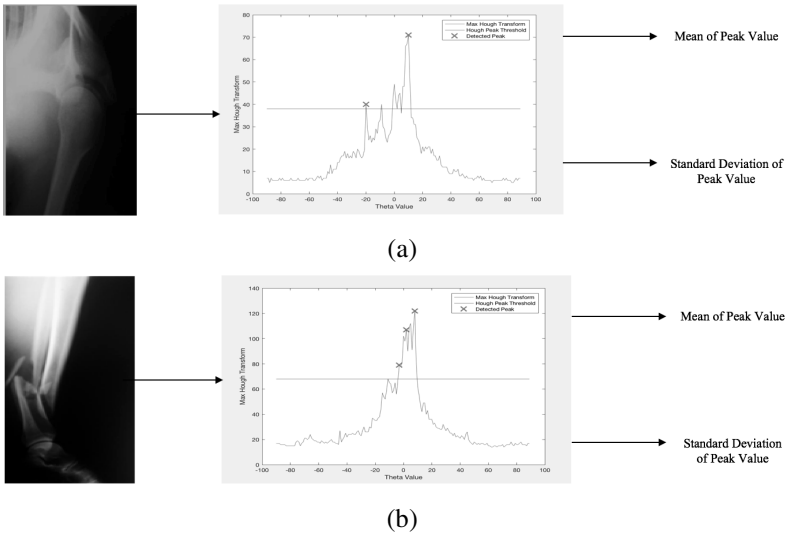


Fig. 3. Application of the Hough transform to an X-ray image and extraction of peak values: two peak values for a minor fracture (a) and three peak values for a major fracture (b).

considered both continuous (original) and discretized values of laboratory results. Since the latter improved the classification performance of the tested classifiers, we decided to apply expert-based norms as part of T_C. Second, the originally recorded information about injuries was very detailed (e.g., it indicated their very precise location, like a specific toe or finger), and thus it was impossible to identify any strong patterns in the clinical data. To address this issue, we introduced additional variables that captured more general information (i.e., they indicated the number of injuries within a given body section, like the upper limb). Finally, in Step 3, in preliminary experiments we also considered other feature selection techniques, such as Relief (see the work of Kuhn and Johnson (2013) for a more detailed description). However, we did not observe any additional benefits in terms of performance and therefore these techniques were not included in T_C. Here we should also note that a common dimensionality reduction technique employed within data fusion is PCA (Viswanath *et al.*, 2017). While it was developed for numerical data, it can be also applied to binary features. We used the one-hot encoding with the data and then applied PCA to further reduce the dimensionality, but similarly to feature selection, we did not observe any benefits (more detailed results from our attempts are reported in the next section).

Application of T_C to S_C results in the F_C data space, where each patient was described with values of 96 features (as compared to 301 features from S_C). These are briefly summarized in Table 1.

Further processing (fusion) of F_I and F_C data spaces was specific for each data fusion model and it is described below:

1. COD model (see Fig. 2(a)): F_C and F_I were combined (join operation) into the single $F_{C,I}$ data space. This representation was then used to build the base CL_B classifier responsible for providing the final therapeutic prediction.

2. Simple COI models—S-COI-C, S-COI-I, and S-COI-CI (see Figs. 2(b), (c) and (d), respectively): F_C and F_I were combined using transformations that involved building intermediary classifiers at level 0. For S-COI-C and S-COI-I, such classifiers were constructed from a single data space, and their outcomes were combined with the other space. For S-COI-CI, two level-0 classifiers were built from both data spaces and their outcomes were joined. Here we need to explain that we combined numerical "votes" provided by classifiers for specific classes rather than binary outcomes, as the former allowed us to capture "richer" information. The data spaces obtained at level 0 (F_{CL_C,F_I}, F_{F_C,CL_I}, F_{CL_C,CL_I}) were finally used to construct the CL_B classifier at level 1.

3. Complex COI models—C-COI-C and C-COI-CI) (see Figs. 2(e) and (f), respectively): F_C was split into two sub-spaces F_{C1} and F_{C2} that captured transformed clinical data corresponding to results of physical examinations (including injuries) and laboratory tests, respectively. Values of basic demographic features (age and gender) were included in both spaces. F_{C1} and F_{C2} were used to derive CL_{C1} and CL_{C2} classifiers at level 0. Then, the outcomes of these two classifiers were fused into the $F_{CL_{C1},CL_{C2}}$ space, further employed to

Table 1. Clinical features considered in the study.

Group of features	Number of features
Demographic information	2
Laboratory test results	22
Physical examination results	72

construct the CL_C, classifier at level 1. In C-COI-C the outcome (votes) from CL_C was fused together with F_I to form a feature space of F_{CL_C,F_I} that was finally used to build CL_B. In C-COI-CI, we also introduced the CL_I classifier at level 0 and its outcome was then combined with the outcome of CL_C. This resulted in F_{CL_C,CL_I} used to build CL_B.

3.4. Experimental design. In the experiments, we applied the six data fusion models described in the previous section to construct and combine different types of classifiers into specific therapeutic decision models. Classifiers employed in our study are described in Table 2; the selection was based on our past experience with analysis of clinical data (Wilk *et al*., 2016) and on results of other studies related to data fusion (Tiwari *et al*., 2011). All classifiers were implemented in WEKA (Hall *et al*., 2009), and for most of the parameters, we used default values (as they were performing well on default parameter settings). The parameters that were modified for our study are presented in Table 2. Such customization was performed during a preliminary analysis limited to individual data spaces and their combination according to the COD model. Specifically, we used a grid search over the possible values of the parameters considered and selected those that resulted in the best performance. A similar search was performed to analyze the values of the parameters C (cost) and γ parameters for the SVM. For the KNN classifier, we checked values of k ranging from 3 to 19, and the best performance was observed for 7 neighbors. This range of the parameter k was defined as arbitrary based on the literature and our previous experience with other data sets (Wilk *et al*., 2016).

When evaluating the performance of classifiers, we considered the following measures: classification accuracy (or true-positive rate) for non-surgical and surgical classes and the average accuracy over these two classes (i.e., micro-average of accuracy). The latter measure is often used in the context of imbalanced data sets as a better alternative to overall accuracy (Ferri *et al*., 2009). In order to obtain more reliable and statistically sound results, the evaluation was performed according to the 10-fold cross-validation scheme repeated 10 times.

The experimental design covered two phases. In the first one, we evaluated decision models constructed from a single data space—either F_I or F_C (i.e., without data fusion). This allowed us to establish the baseline performance for the second phase, where we applied COD, S-COI-C, S-COI-I, S-COI-CI, C-COI-C, and C-COI-CI data fusion models. To limit the number of possible combinations of classifiers in C-COI-C and C-COI-CI, we first evaluated possible combinations for the CL_{C1}, CL_{C2}, CL_I, and CL_C classifiers and selected the best performing (in terms of the average accuracy) set of classifiers. Then, we considered different possible classifiers for CL_B and evaluated the performance of the entire decision model.

4. Results

The results of the first phase of our experiment are given in Table 3 and visualized in Fig. 4 (evaluation of the obtained results from a practical perspective would depend upon the type of fracture under consideration). In this stage we employed all types of classifiers listed in Table 2 and constructed 14 decision models either from the F_I image data space or F_C clinical data space.

The most important observations from the first stage are the following:

- Decision models constructed from the space F_I turned out to be more accurate than models derived from F_C (this is clearly visible in Fig. 4, where the former models are concentrated in the top right corner and dominate the latter). This confirms the importance of image information for deciding about the type of treatment.
- The highest average accuracy of 80.0% was achieved by the FI-5 decision model (DT classifier), and it can be characterized as a good result compared with the decision model derived from clinical features F_C (KNN classifier)—its average accuracy was 72.6%.
- For the surgical class, the highest accuracy of 95.0% was obtained by the FC-1 decision model (KNN classifier). At the same time, the accuracy for the non-surgical class demonstrated by this model was one of the lowest. The best accuracy for the non-surgical class was equal to 71.0% and it was achieved by the FI-4 decision model (NB-D classifier).

The non-surgical class is more diversified and scattered. Thus it is more difficult to identify areas in data space with a majority of examples from this class and to induce

Table 2. Classifiers considered in the study and their parameters.

Symbol	Description	Parameters
KNN	A k-nearest neighbor classifier with Euclidean distance.	$k = 7$
NB, NB-K, NB-D	A naïve Bayes classifier where normal distribution, kernel density estimator and internal discretization were used for numerical values, respectively.	default
DT	A decision tree classifier constructed using the C4.5 algorithm.	default
RF	A random forest classifier.	default
SVM	An SVM classifier with a radial basis kernel function.	$C = 1000$, $\gamma = 0.0001$ for F_I (SVM as CL_I), 0.01 for F_C (SVM as CL_C), and 0.001 otherwise

Table 3. Performance of decision models constructed from a single data space (standard deviation given in brackets, ID = identifier of a decision model).

			Accuracy [%]		
ID	CL_B	Data space	Average	Non-surgical	Surgical
FI-1	KNN	F_I	78.0 (12.5)	68.0 (16.0)	88.0 (9.0)
FC-1	KNN	F_C	72.6 (9.5)	50.2 (13.0)	95.0 (6.0)
FI-2	NB	F_I	78.0 (11.5)	67.0 (15.0)	89.0 (8.0)
FC-2	NB	F_C	65.5 (14.0)	43.0 (19.0)	88.0 (9.0)
FI-3	NB-K	F_I	78.5 (11.5)	68.0 (15.0)	89.0 (8.0)
FC-3	NB-K	F_C	70.5 (13.0)	53.0 (17.0)	88.0 (9.0)
FI-4	NB-D	F_I	79.6 (11.5)	71.0 (15.0)	88.0 (8.0)
FC-4	NB-D	F_C	60.0 (11.5)	28.0 (15.0)	92.0 (8.0)
FI-5	DT	F_I	80.0 (12.0)	70.0 (17.0)	90.0 (7.0)
FC-5	DT	F_C	65.5 (14.5)	52.0 (18.0)	80.0 (11.0)
FI-6	RF	F_I	76.0 (13.5)	68.0 (18.0)	84.0 (9.0)
FC-6	RF	F_C	72.0 (13.5)	57.0 (19.0)	87.0 (8.0)
FI-7	SVM	F_I	78.5 (10.5)	65.0 (14.0)	92.0 (7.0)
FC-7	SVM	F_C	65.5 (14.0)	56.0 (16.0)	75.0 (12.0)

stronger patterns capturing this class. Having said this, the FC-1 model relies on KNN—such a classifier is sensitive to noisy or rare cases and it may be biased towards the class that is more uniform (i.e., forms more uniform clusters). Specifically, when predicting the class for a non-surgical case, the classifier may easily focus on the neighborhood with the majority of surgical patients and thus establish an incorrect outcome, which results in a higher accuracy value for the majority class (i.e., 95%).

The second phase started with developing decision models following the COD scheme. The results obtained for this fusion model are reported in Table 4 and summarized visually in Fig. 5. The most important observations from this analysis are as follows:

- For decision models with DT, RF and SVM classifiers (COD-3, COD-4, and COD-5, respectively), fusion of data spaces resulted in improved average accuracy in comparison with the decision models derived from single spaces. This confirms the benefits of data fusion techniques, even when using a simple scheme such as COD.
- In the case of the COD-4 decision model, fusion improved also accuracy in each class in comparison to the baseline results. Moreover, for the COD-3 and COD-5 models we observed increased accuracy for the non-surgical class (with gains up to 5%); however, there was a decrease (2%) in the accuracy for the surgical class.
- Although the COD fusion model turned out to be beneficial, we still need to emphasize high accuracy of decision models constructed in the first phase from the F_I image data space and importance of features from that space.

At this point, we also examined the impact of dimensionality reduction on the performance of decision

Table 4. Performance of decision models constructed following the COD models (standard deviation given in brackets, ID = identifier of a decision model).

ID	CL_B	Accuracy [%] Average	Non-surgical	Surgical
COD-1	KNN	56.5 (9.0)	17.0 (12.0)	96.0 (6.0)
COD-2	NB-K	76.5 (12.0)	60.0 (17.0)	93.0 (7.0)
COD-3	DT	80.0 (12.0)	72.0 (16.0)	88.0 (8.0)
COD-4	RF	81.5 (11.5)	73.0 (15.0)	90.0 (8.0)
COD-5	SVM	80.5 (11.0)	71.0 (15.0)	90.0 (7.0)

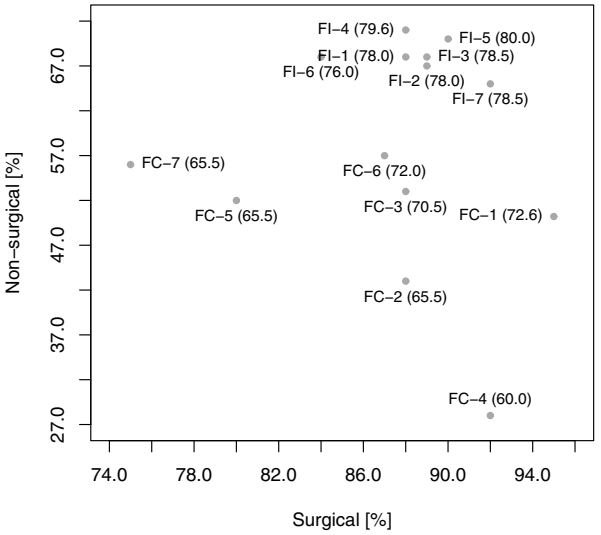


Fig. 4. Decision models constructed from a single data space (average accuracy given in brackets).

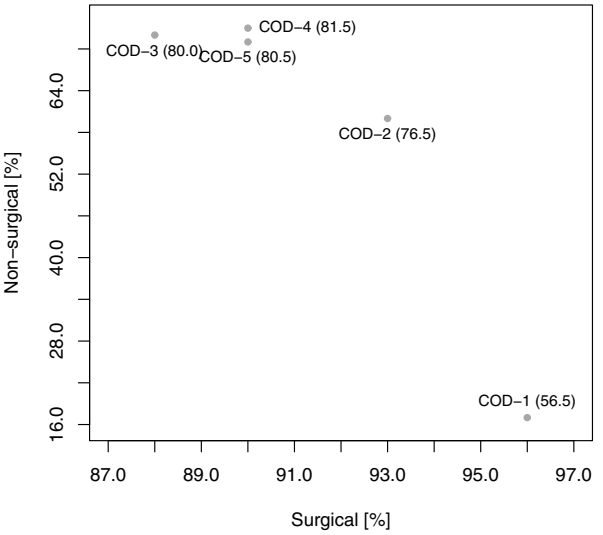


Fig. 5. Decision models constructed following the COD model (average accuracy given in brackets).

models. Specifically, we applied PCA to the $F_{C,I}$ data space (with categorical features encoded using the one-hot technique) and preserved top 50 principal components capturing 95% of the variance. Then, the reduced data space was used to build possible classifiers. The obtained results are listed in Table 5. Since accuracy deteriorated in comparison with the decision models constructed following COD, we decided not to use PCA.

In the third phase, we explored the COI fusion models. We started with the simple models—S-COI-C, S-COI-I, and S-COI-CI—and then proceeded with the complex ones—C-COI-C and C-COI.

In S-COI-C we applied various classifiers to F_C. Specifically, we focused on RF and KNN as CL_C as these classifiers performed best in the first stage when applied to clinical data space. As previously, we tried multiple classifiers as CL_B and the results are reported in Table 6 and Fig. 6. The key observations are the following:

- The S-COI-C-11 decision model with KNN at both levels, i.e., CL_C and CL_B, provided the highest average accuracy. Similar performance was achieved using the combination with KNN at the CL_C level and with RF at the CL_B level (S-COI-C-9).
- Decision models with RF as CL_C demonstrated lower average accuracy then models with KNN. For the former models, the average accuracy did not exceed 77% and accuracy for the non-surgical class was low.
- Overall, S-COI-C proved to be a less accurate model as compared to COD.

In S-COI-I we applied different classifiers to F_I. In the first the stage we tried DT and SVM as CL_I. However, DT demonstrated the highest average accuracy when applied to image data. Moreover, we tried multiple classifiers as CL_B and only the best ones are reported in Table 7 and Fig. 7. The most important observations are the following:

- The highest accuracy was achieved for the S-COI-I-2 decision model (DT as CL_I and NB-D as CL_B). Similar performance was also demonstrated by other decision models with DT as CL_I, except for these where SVM, RF, and KNN were used as CL_B.

Table 5. Performance of decision model using feature transformation with PCA scheme using top 50 features capturing 95% of variance (standard deviation given in brackets, CL = applied classifier).

CL	Accuracy [%] Average	Non-Surgical	Surgical
KNN	65.5 (13.0)	41.0 (18.0)	90.0 (8.0)
NB	65.5 (13.0)	42.0 (17.0)	89.0 (9.0)
NB-K	63.0 (13.0)	36.0 (18.0)	90.0 (8.0)
NB-D	69.0 (16.5)	55.0 (21.0)	83.0 (12.0)
DT	67.0 (16.0)	57.0 (20.0)	77.0 (12.0)
RF	73.0 (14.0)	56.0 (19.0)	90.0 (9.0)
SVM	78.0 (12.5)	67.0 (17.0)	89.0 (8.0)

Table 6. Performance of decision models constructed following the S-COI-C model (standard deviation given in brackets, ID = identifier of a decision model).

ID	CL_C	CL_B	Accuracy [%] Average	Non-surgical	Surgical
S-COI-C-1	RF	DT	72.1 (10.1)	59.9 (18.4)	84.3 (9.1)
S-COI-C-2	RF	RF	71.9 (9.8)	57.4 (18.2)	86.4 (8.1)
S-COI-C-3	RF	SVM	75.7 (9.3)	63.9 (17.0)	87.9 (8.6)
S-COI-C-4	RF	KNN	76.6 (7.9)	69.2 (15.2)	84.0 (11.3)
S-COI-C-5	RF	NB	74.0 (9.9)	64.7 (17.7)	83.3 (9.5)
S-COI-C-6	RF	NB-K	73.0 (9.9)	59.4 (18.1)	86.5 (8.2)
S-COI-C-7	RF	NB-D	72.0 (9.9)	57.4 (18.2)	86.4 (8.1)
S-COI-C-8	KNN	DT	71.5 (1.3)	87.3 (9.4)	79.4 (8.1)
S-COI-C-9	KNN	RF	81.4 (8.1)	77.1 (13.5)	85.7 (9.7)
S-COI-C-10	KNN	SVM	75.7 (9.3)	68.5 (17.2)	83.3 (12.3)
S-COI-C-11	KNN	KNN	83.9 (6.7)	78.7 (12.9)	89.2 (8.3)
S-COI-C-12	KNN	NB	79.9 (8.8)	76.2 (13.9)	83.8 (10.5)
S-COI-C-13	KNN	NB-K	77.9 (9.6)	67.0 (15.4)	88.9 (9.5)
S-COI-C-14	KNN	NB-D	77.8 (9.1)	68.2 (15.8)	87.3 (9.6)

- In the case of SVM as CL_I, the average accuracy did not exceed 80%. However, the highest accuracy for the surgical class was reported using SVM as CL_I and NB as CL_B.

For the S-COI-CI fusion model, we constructed a series of decision models where we applied different classifiers to F_I and F_C data spaces. The results obtained by applying the S-COI-CI fusion model are given in Table 8 and Fig. 8. They can be summarized as follows:

- For several decision models, there was a large increase in the accuracy for the non-surgical class in comparison to the COD approach. It was especially evident for the S-COI-CI-8 model, which attained 85.0%. However, it was associated with the decrease in accuracy for the surgical class, and thus the average accuracy for the best decision models was comparable to the performance of the best models from the previous step (84.2% for S-COI-CI-8 vs. 81.5% for COD-4). The S-COI-CI-8 model was also the one that demonstrated the most even performance across both decision classes.
- The trade-off between both decision classes was also clearly visible for decision models that demonstrated the best accuracy for the surgical class, e.g., S-COI-CI-3 or S-COI-CI-8. These models were relatively weaker in recognizing the non-surgical class. Nevertheless, their performance for this class was usually better than that of decision models following the COD approach.

As already mentioned, we divided the construction of decision models following C-COI-I and C-COI-CI into two steps in order to minimize the number of possible combinations of classifiers to consider. In the first step, we focused on the CL_{C1} and CL_{C2} classifiers derived from data sub-spaces capturing various aspects of a patient's clinical image (physical examinations vs. laboratory results) and on the CL_C classifier derived from their fused outcomes (see Figs. 2(e) and (f)). The performance of the discussed combinations of classifiers is given in Table 9. In comparison with the baseline results for F_C given in Table 3, one could notice the benefits of splitting the F_C space into two sub-spaces—the performance for

Table 7. Performance of decision models constructed following the S-COI-I model (standard deviation given in brackets, ID = identifier of a decision model).

ID	CL_I	CL_B	Accuracy [%] Average	Non-surgical	Surgical
S-COI-I-1	DT	NB	80.1 (8.7)	70.6 (16.3)	89.7 (7.2)
S-COI-I-2	DT	NB-D	80.3 (8.8)	70.6 (16.3)	90.1 (7.1)
S-COI-I-3	DT	NB-K	80.2 (9.1)	70.3 (17.0)	90.1 (7.1)
S-COI-I-4	DT	SVM	78.6 (8.4)	65.6 (15.1)	91.7 (7.9)
S-COI-I-5	DT	RF	75.9 (8.9)	67.3 (17.5)	84.7 (9.2)
S-COI-I-6	DT	DT	80.2 (9.1)	70.3 (17.0)	90.1 (7.1)
S-COI-I-7	DT	KNN	78.6 (8.9)	71.7 (17.4)	85.5 (8.5)
S-COI-I-8	SVM	NB	78.9 (8.6)	65.5 (15.9)	92.3 (6.7)
S-COI-I-9	SVM	NB-D	78.7 (8.3)	66.5 (15.2)	90.9 (7.2)
S-COI-I-10	SVM	NB-K	78.7 (8.3)	65.1 (15.5)	92.3 (6.7)
S-COI-I-11	SVM	SVM	78.8 (8.3)	65.2 (14.8)	92.3 (7.8)
S-COI-I-12	SVM	RF	75.6 (8.8)	66.7 (16.9)	84.4 (8.6)
S-COI-I-13	SVM	DT	77.9 (8.9)	65.2 (15.9)	90.7 (7.0)
S-COI-I-14	SVM	KNN	77.1 (8.4)	66.5 (16.2)	87.7 (8.9)

Table 8. Performance of decision models constructed following the S-COI-CI model (standard deviation given in brackets, ID = identifier of a decision model).

ID	CL_C	CL_I	CL_B	Accuracy [%] Average	Non-surgical	Surgical
S-COI-CI-1	NB-K	NB-D	KNN	81.0 (12.5)	75.0 (15.0)	87.0 (10.0)
S-COI-CI-2	NB-K	NB-D	NB	82.8 (9.7)	84.0 (12.0)	81.7 (7.4)
S-COI-CI-3	NB-K	RF	KNN	80.5 (11.5)	71.0 (15.0)	90.0 (8.0)
S-COI-CI-4	NB-K	RF	NB	80.7 (10.5)	80.0 (13.0)	81.4 (7.9)
S-COI-CI-5	NB-K	RF	NB-K	81.0 (10.5)	75.0 (15.0)	87.0 (6.0)
S-COI-CI-6	NB-K	RF	RF	79.5 (13.0)	72.0 (16.0)	87.0 (10.0)
S-COI-CI-7	NB-K	SVM	KNN	83.0 (11.0)	77.0 (14.0)	89.0 (8.0)
S-COI-CI-8	NB-K	SVM	NB	84.2 (9.2)	85.0 (11.0)	83.4 (7.3)
S-COI-CI-9	NB-K	SVM	NB-K	83.0 (12.0)	82.0 (13.0)	84.0 (11.0)
S-COI-CI-10	RF	NB	SVM	83.3 (10.0)	83.0 (13.0)	83.6 (6.9)
S-COI-CI-11	RF	NB-D	NB-K	80.0 (13.5)	73.0 (17.0)	87.0 (10.0)
S-COI-CI-12	RF	RF	KNN	80.5 (13.5)	73.0 (18.0)	88.0 (9.0)
S-COI-CI-13	RF	RF	NB	80.5 (11.5)	79.0 (15.0)	82.1 (7.9)
S-COI-CI-14	RF	RF	NB-K	81.5 (12.5)	77.0 (16.0)	86.0 (9.0)
S-COI-CI-15	RF	SVM	NB-K	82.5 (13.0)	81.0 (15.0)	84.0 (11.0)

the non-surgical class was improved and the accuracy for the surgical class in most cases was preserved at the same level. To further proceed with the C-COI-CI model, we selected the classifiers resulting in the best average accuracy of 74.0%, i.e., NB-K as CL_{C1} and CL_C and RF as CL_{C2}.

Performance of decision models constructed following the C-COI-C fusion model is reported in Table 10 and Fig. 9. The most important observations from this part of the analysis are as follows:

- The highest accuracy of 78.6% was achieved by the C-COI-C-5 model with NB-K as CL_C and SVM at CL_B.
- average accuracy reported by the C-COI-C model is quite similar to the ones reported in the S-COI-I and S-COI-C models.
- performance of the C-COI-C is relatively low as compared to the performance of decision models following S-COI-CI (see Table 8). Moreover, the complexity of the C-COI-C model is much higher than that of S-COI-CI.

The results obtained for the C-COI-CI fusion model are reported in Table 11 and Fig.10. The most important observations from this analysis are as follows:

1. Similarly to the COD approach, also C-COI-CI

Table 9. Internal performance at level 2 in decision models constructed following the C-COI-C and C-COI-CI models (standard deviation given in brackets).

			Accuracy [%]		
CL_{C1}	CL_{C2}	CL_C	Average	Non-surgical	Surgical
NB-K	RF	NB	72.0 (14.0)	56.0 (18.0)	88.0 (10.0)
NB-K	RF	NB-K	74.0 (13.0)	63.0 (16.0)	85.0 (10.0)
NB-K	RF	NB-D	73.0 (14.5)	63.0 (17.0)	83.0 (12.0)
NB-K	RF	RF	66.0 (14.5)	53.0 (17.0)	79.0 (12.0)
RF	NB-K	NB	72.0 (13.0)	64.0 (16.0)	80.0 (10.0)
RF	NB-K	NB-K	73.0 (13.5)	67.0 (16.0)	79.0 (11.0)
RF	NB-K	NB-D	71.0 (16.5)	66.0 (19.0)	76.0 (14.0)

Table 10. Performance of decision models constructed following the C-COI-C model (standard deviation given in brackets, ID = identifier of a decision model).

			Accuracy [%]		
ID	CL_C	CL_B	Average	Non-surgical	Surgical
C-COI-C-1	NB-K	RF	70.2 (9.7)	57.8 (17.8)	82.6 (9.9)
C-COI-C-2	NB-K	NB-K	71.1 (9.8)	61.2 (17.5)	80.9 (10.5)
C-COI-C-3	NB-K	DT	69.9 (10.1)	59.1 (18.7)	80.8 (10.7)
C-COI-C-4	NB-K	7NN	71.6 (9.9)	60.8 (17.5)	82.4 (10.7)
C-COI-C-5	NB-K	SVM	78.6 (7.5)	72.8 (14.3)	84.4 (10.1)

Table 11. Performance of decision models constructed following the C-COI-CI model (standard deviation given in brackets, ID = identifier of a decision model).

				Accuracy [%]		
ID	CL_C	CL_I	CL_B	Average	Non-surgical	Surgical
C-COI-CI-1	NB-K	NB-D	KNN	82.0 (10.5)	76.0 (13.0)	88.0 (8.0)
C-COI-CI-2	NB-K	NB-D	NB-K	83.0 (11.0)	78.0 (14.0)	88.0 (8.0)
C-COI-CI-3	NB-K	DT	KNN	82.0 (12.0)	75.0 (15.0)	89.0 (9.0)
C-COI-CI-4	NB-K	DT	NB-K	83.0 (11.5)	78.0 (15.0)	88.0 (8.0)
C-COI-CI-5	NB-K	RF	KNN	83.5 (11.5)	76.0 (15.0)	91.0 (8.0)
C-COI-CI-6	NB-K	RF	NB-K	84.5 (11.0)	79.0 (14.0)	90.0 (8.0)
C-COI-CI-7	NB-K	RF	RF	81.0 (11.5)	74.0 (14.0)	88.0 (9.0)
C-COI-CI-8	NB-K	SVM	KNN	81.5 (11.0)	73.0 (14.0)	90.0 (8.0)
C-COI-CI-9	NB-K	SVM	NB-K	84.0 (11.5)	80.0 (14.0)	88.0 (9.0)

results in decision models biased toward the surgical class. The best accuracy for this class was equal to 91.0% and it was achieved for the C-COI-CI-5 model.

2. The attained accuracy for the non-surgical class was in the range from 73.0% to 80.0%, thus it was better than for models constructed according to COD.

3. The best average accuracy of 84.5% was reported by the C-COI-CI-6 model. It is also the best value attained by any of the decision models considered in the experiment approach.

To provide a concise summary of our experiments, we selected the most promising decision models (both baseline ones from the first phase and the ones established according to various fusion approaches in the second phase) and presented them in Fig. 11. However, evaluation of the obtained results from a practical perspective would depend upon the type of fracture. The most important observations are the following:

- While decision models based on the F_I data space performed surprisingly well, the analysis revealed benefits of fusing images with clinical data—the FI-5 model was dominated by several other models based on fused data spaces (e.g., COD-4 and C-COI-CI-5).

- Decision models constructed according to the COI approach turned out to be more accurate than those derived using COD—this emphasizes the benefits of applying different classifiers to specific data spaces.

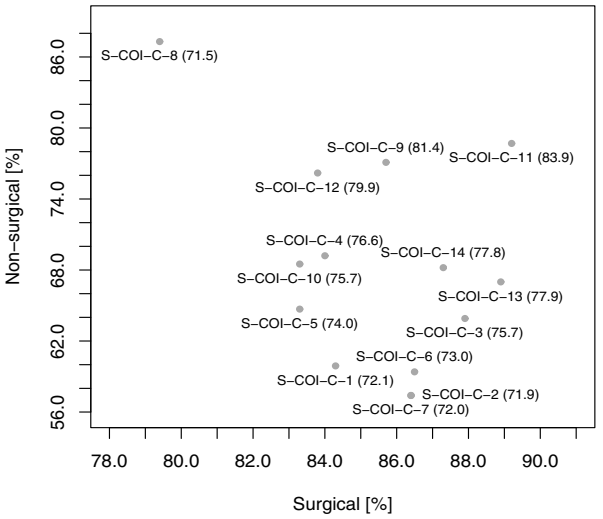


Fig. 6. Decision models constructed following the S-COI-C model (average accuracy given in brackets).

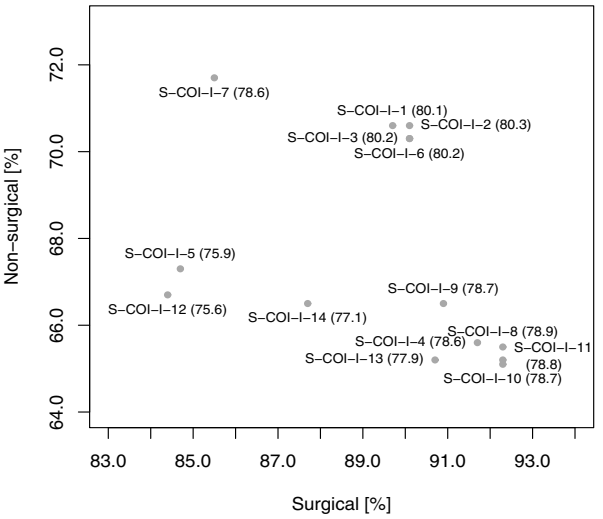


Fig. 7. Decision models constructed following the S-COI-I model (average accuracy given in brackets).

- Decision models based on S-COI and C-COI demonstrated different "classification biases." The S-COI models were more accurate for the non-surgical class, while the C-COI ones were focused on the surgical class. Thus they could be seen as complementary in terms of captured expertise.

- Results reported for decision models constructed with S-COI-CI model (and also C-COI-CI) provide better accuracy for the non-surgical class by using various variants of the naïve Bayes (NB) classifier as the base classifier (CLB). In such a situation, also accuracies across classes are more balanced. This indicates that NB is better able to estimate probabilities when it is built from a data space combing outcomes of classifiers operating on individual data sources.

To perform statistical analysis of decision models constructed according to different fusion models, we applied a one-way ANOVA test (Salzberg and Fayyad, 1997). For each fusion model, we selected the best performing (in terms of the average accuracy) decision model, and the purpose of this test was to report whether there were statistically significant differences between these models. The findings are reported in Table 12 and their graphical representation is shown in Fig. 12. It indicates that there are two groups of decision models (and thus underlying fusion models), with statistical differences between these groups and no differences inside groups. However, if you focus on each group (in Figs. 12(a) and (b)) individually, you will observe that group (a) contains those COI models which have balanced class accuracies, whereas the models mentioned in group (b) are more biased towards the majority class.

As we have already mentioned, there is a similarity between data fusion and ensemble classifiers. To further explore it, we applied two popular approaches to building such classifiers—boosting and bagging (see Kuhn and Johnson, 2013). We combined these schemes with classifiers considered in our study and applied them to the combined $F_{C,I}$ data space, also used by the COD fusion model.

Results obtained for the bagging scheme are reported in Table 13. We briefly summarize them as follows:

- The highest accuracy obtained is 82.0% and it was attained for a DT classifier.

- Bagging provided similar performance to the COD and S-COI-C models for the average class accuracy and also for individual class accuracies. The highest accuracy obtained using bagging for the surgical class is 92.0%.

Performance achieved by boosted classifiers is given in Table 14. It is close to that of decision models based on the S-COI-I, S-COI-C and C-COI-C fusion models. The summary of the obtained results in is as follows:

- The highest average accuracy using boosting was 82.0% obtained using the RF classifier.

- Similarly to bagging, boosting was also more biased towards the surgical class which may have been caused by many factors such as class imbalance, non-surgical class capturing more rare cases and thus being possibly more diversified (it is more difficult to find stronger patterns), with the surgical class being more uniform and thus easier to learn.

Table 12. Result of statistical testing performed using a one-way ANOVA test (p-value $< \alpha = 0.05$ is highlighted in boldface).

	S-COI-I	S-COI-C	S-COI-CI	C-COI-C	C-COI-CI
COD	**9.00E-06**	**4.60E-02**	1.81E-01	**5.00E-03**	7.13E-01
S-COI-I	-	4.50E-01	**2.50E-08**	1.39E-01	**3.70E-06**
S-COI-C	-	-	**6.24E-04**	4.63E-01	**2.20E-02**
S-COI-CI	-	-	-	**1.65E-05**	3.67E-01
C-COI-C	-	-	-	-	**2.20E-02**

Table 13. Performance of decision models constructed following bagging scheme (standard deviation given in brackets, CL = applied classifier).

	Accuracy [%]		
CL	Average	Non-surgical	Surgical
KNN	57.5 (10.0)	19.0 (15.0)	96.0 (5.0)
NB	76.0 (13.5)	61.0 (16.0)	91.0 (8.0)
NB-K	80.0 (11.5)	69.0 (15.0)	91.0 (8.0)
NB-D	76.0 (13.5)	61.0 (18.0)	91.0 (9.0)
DT	82.0 (11.0)	75.0 (15.0)	89.0 (7.0)
RF	81.0 (12.0)	70.0 (16.0)	92.0 (8.0)
SVM	81.0 (13.0)	73.0 (17.0)	89.0 (9.0)

Table 14. Performance of decision models constructed following the boosting scheme (standard deviation given in brackets, CL = applied classifier).

	Accuracy [%]		
CL	Average	Non-surgical	Surgical
KNN	64.0 (13.5)	51.0 (16.0)	77.0 (11.0)
NB	77.0 (13.0)	68.0 (17.0)	86.0 (9.0)
NB-K	77.0 (12.5)	66.0 (16.0)	88.0 (9.0)
NB-D	75.5 (14.5)	65.0 (19.0)	86.0 (10.0)
DT	79.5 (12.0)	71.0 (16.0)	88.0 (8.0)
RF	82.0 (11.0)	73.0 (14.0)	91.0 (8.0)
SVM	80.0 (12.5)	71.0 (17.0)	89.0 (8.0)

5. Conclusions

In this paper, we presented the concept of data fusion and described its application in a case study aimed at developing a therapeutic decision model to predict the type of treatment (surgical vs. non-surgical) for patients with bone fractures. While such decision should be based on the characteristics of a fracture, it should also consider the overall state or "fitness" of the patient. Thus, a comprehensive decision model should be based on both image and non-image patient data.

In our case study, we considered six models of data fusion: a simple COD and five variants of COI with increasing degree of complexity. The first model fused various data sources into a single space, while the latter combined outcomes of multiple classifiers. To obtain a baseline, we also checked decision models constructed from single data sources. The results clearly demonstrated the benefits of data fusion, in particular of COI models. Moreover, we performed an ANOVA test and the results revealed that COI models offer complementary performance for specific decision classes. In particular, decision models following the most complex variant of COI (S-COI-CI and C-COI-CI) demonstrated the best performance for the non-surgical class and the most balanced accuracy across classes, while other decision models were biased towards the surgical class. Since the acceptable decision bias may be dependent on a specific problem (e.g., in the case of hip fracture a delayed surgery may serious consequences (Cha *et al.*, 2017), while in the case of many hand fractures surgeries do not improve the outcomes (Giddins, 2015)), in clinical practice it may be reasonable to use decision models from both groups and let the clinician make the final decision.

For a more comprehensive evaluation, we also compared the performance of the obtained decision models with ensemble classifiers constructed according to bagging and boosting schemes. This analysis further confirmed the benefits of more complex data fusion

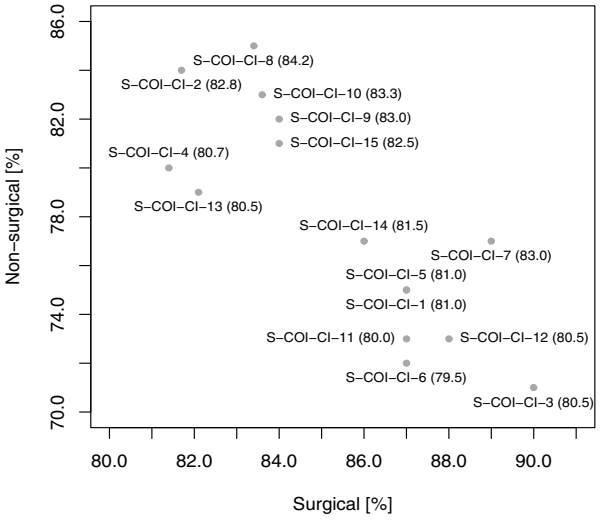


Fig. 8. Decision models constructed following the S-COI-CI model (average accuracy given in brackets).

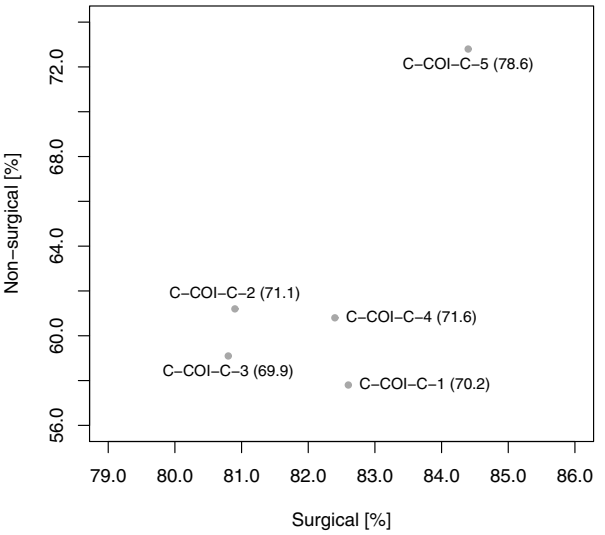


Fig. 9. Decision models constructed following the C-COI-C model (average accuracy given in brackets).

models, which resulted in better accuracy of the oncoming decision models than the ensembles considered. In a comparative analysis of complex COI and PCA (used in combination with COD), COI tends to perform better and is higher accuracy.

6. Future work

Our study has several limitations. The size of the analyzed data (in terms of the number of patients and their images) was limited. We also focused only on image-based features extracted through custom transformation and did not apply deep learning techniques that may have been able to discover relevant features automatically (giving up deep learning was in fact related to the limited amount of data). All of these shortcomings will be addressed in the future work. We also plan to develop a control mechanism for data fusion that would recommend transformations appropriate for specific data sources and their sequence. This involves constructing transformations to deal with data difficulty factors such as the overlapping of decision boundaries, rare cases, outliers, and noise (Wilk *et al.*, 2016). Our ultimate goal is to construct a comprehensive data fusion framework for clinical data, and we believe this study constitutes a relevant step towards this end.

Acknowledgment

This research has been funded by the European Commission through the Erasmus Mundus Joint Doctorate *Information Technologies for Business Intelligence—Doctoral College* (IT4BI-DC).

References

Al-Ayyoub, M. and Al-Zghool, D. (2014). Determining the type of long bone fractures in X-ray images, *WSEAS Transactions on Information Science and Applications* **10**(8): 261–270.

Brzezinski, J., Kosiedowski, M., Mazurek, C., Slowinski, K., Slowinski, R., Stroinski, M. and Weglarz, J. (2013). Towards telemedical centers: Digitization of inter-professional communication in healthcare, *in* M. Cruz-Cunha *et al.* (Eds.), *Handbook of Research on ICTs and Management Systems for Improving Efficiency in Healthcare and Social Care*, IGI Global, Hershey, PA, pp. 805–829.

Castanedo, F. (2013). A review of data fusion techniques, *The Scientific World Journal* **2013**: 704504, DOI: 10.1155/2013/704504.

Cha, Y.-H., Ha, Y.-C., Yoo, J.-I., Min, Y.-S., Lee, Y.-K. and Koo, K.-H. (2017). Effect of causes of surgical delay on early and late mortality in patients with proximal hip fracture, *Archives of Orthopaedic and Trauma Surgery* **137**(5): 625–630.

de Bruijne, M. (2016). Machine learning approaches in medical image analysis: From detection to diagnosis, *Medical Image Analysis* **33**: 94–97, DOI: 10.106/j.media.2016.06.032.

Dittman, D.J., Khoshgoftaar, T.M. and Napolitano, A. (2014). Selecting the appropriate data sampling approach for imbalanced and high-dimensional bioinformatics datasets, *IEEE 14th International Conference on Bioinformatics and Bioengineering, BIBE 2014, Boca Raton, FL, USA,* pp. 304–310.

Douali, N. and Jaulent, M. (2012). Genomic and personalized medicine decision support system, *2012 IEEE Interna-*

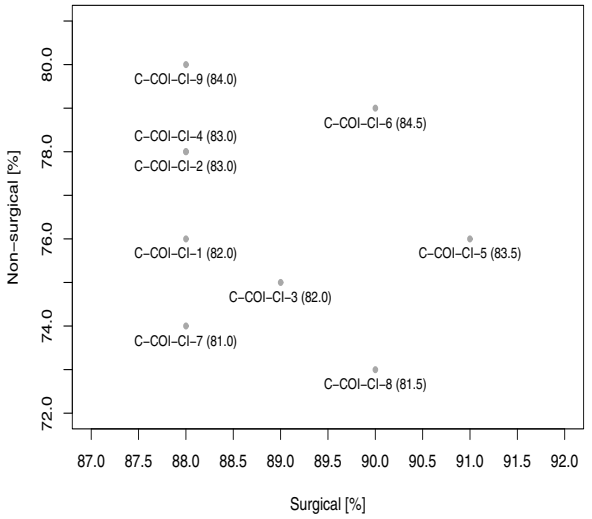


Fig. 10. Decision models constructed following the C-COI-CI model (average accuracy given in brackets).

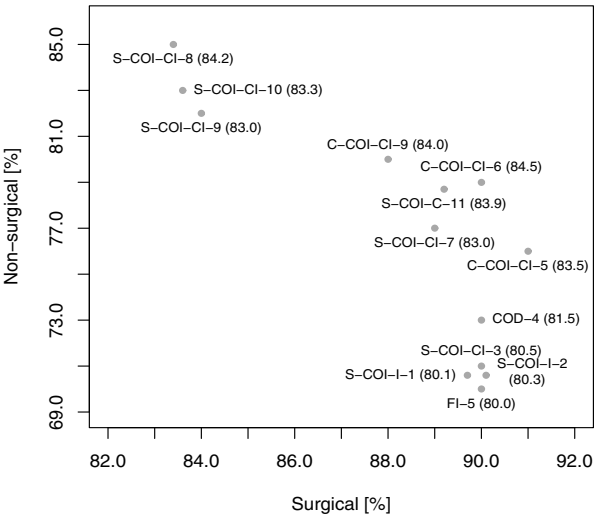


Fig. 11. Selected decision models (average accuracy given in brackets).

tional Conference on Complex Systems (ICCS), Agadir, Morocco, pp. 1–4.

Edward, C.P. and Hepzibah, H. (2015). A robust approach for detection of the type of fracture from X-ray images, *International Journal of Advanced Research in Computer and Communication Engineering* **4**(3): 479–482.

Ferri, C., Hernndez-Orallo, J. and Modroiu, R. (2009). An experimental comparison of performance measures for classification, *Pattern Recognition Letters* **30**(1): 27–38.

Giddins, G.E.B. (2015). The non-operative management of hand fractures, *Journal of Hand Surgery (European Volume)* **40**(1): 33–41.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009). The WEKA data mining software: An update, *ACM SIGKDD Explorations Newsletter* **11**(1): 10–18.

Haq, A. and Wilk, S. (2017). Fusion of clinical data: A case study to predict the type of treatment of bone fractures, *in* M. Kirikova *et al.* (Eds.), *New Trends in Databases and Information Systems*, Springer, Cham, pp. 294–301.

Hossain, M., Neelapala, V. and Andrew, J.G. (2008). Results of non-operative treatment following hip fracture compared to surgical intervention, *Injury* **40**(4): 418–421.

Jesneck, J., Nolte, L., Baker, J., Floyd, C. and Lo, J. (2006). Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis, *Medical Physics* **33**(8): 2945–2954, DOI: 10.1118/1.2208934.

Khatik, I. (2017). A study of various bone fracture detection techniques, *International Journal of Engineering and Computer Science* **6**(5): 21418–21423.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. (2015). Machine learning applications in cancer prognosis and prediction, *Computational and Structural Biotechnology Journal* **13**: 8–17.

Koziarski, M. and Woźniak, M. (2017). CCR: A combined cleaning and resampling algorithm for imbalanced data classification, *International Journal of Applied Mathematics and Computing Sciences* **27**(4): 727–736, DOI: 10.1515/amcs-2017-0050.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, Springer, New York, NY.

Lahat, D., Adali, T. and Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects, *Proceedings of the IEEE* **103**(9): 1449–1477.

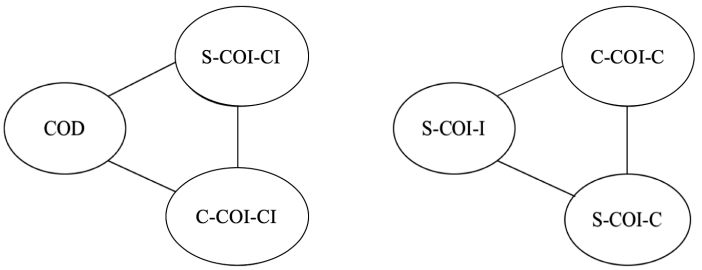
Lanckriet, G., Deng, M., Cristianini, N., Jordan, M. and Noble, W. (2004). Kernel-based data fusion and its application to protein function prediction in yeast, *Pacific Symposium on Biocomputing (PSB 2004), Big Island, HI, USA*, pp. 300–311.

Lee, G., Doyle, S., Monaco, J., Madabhushi, A., Feldman, M.D., Master, S.R. and Tomaszewski, J.E. (2009). A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: Preliminary results in predicting prostate cancer recurrence by fusing mass spectrometry and histology, *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA*, pp. 77–80.

Mitchell, H.B. (2014). *Data Fusion: Concepts and Ideas*, Springer, Berlin/Heidelberg.

Ponti, M. (2011). Combining classifiers: From the creation of ensembles to the decision fusion, *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials, Maceio, Alagoas, Brazil*, pp. 1–10.

Rohlfing, T., Pfefferbaum, A., Sullivan, E. and Maurer, C. (2005). Information fusion in biomedical image analysis: Combination of data vs combination of interpretations, *19th International Conference on Information Processing in Medical Imaging (IPMI'05), Glenwood Springs, CO, USA*, pp. 150–161.



(a) (b)

Fig. 12. Graphical illustration of statistical testing (edges in the graph correspond to non-significant differences in performance).

Salzberg, S.L. and Fayyad, U. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach, *Data Mining and Knowledge Discovery* **1**(3): 317–328, DOI: 10.1023/A:1009752403260.

Sim, L.L.W., Ban, K.H.K., Tan, T.W., Sethi, S.K. and Loh, T.P. (2017). Development of a clinical decision support system for diabetes care: A pilot study, *PLOS ONE* **12**(2): 1–15, DOI:10.1371/journal.pone.0173021.

Tiwari, P., Viswanath, S., Lee, G. and Madabhushi, A. (2011). Multi-modal data fusion schemes for integrated classification of imaging and non-imaging biomedical data, *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Chicago, IL, USA*, pp. 165–168.

Viswanath, S.E., Tiwari, P., Lee, G. and Madabhushi, A. (2017). Dimensionality reduction-based fusion approaches for imaging and non-imaging biomedical data: Concepts, workflow, and use-cases, *BMC Medical Imaging* **17**(1): 2.

Wilk, S., Stefanowski, J., Wojciechowski, S., Farion, K.J. and Michalowski, W. (2016). Application of preprocessing methods to imbalanced clinical data: An experimental study, *in* E. Pietka *et al.* (Eds.), *Information Techmologies in Medicine*, Springer, Berlin/Heidelberg, pp. 503–515.

Yuksel, S.E., Wilson, J.N. and Gader, P.D. (2012). Twenty years of mixture of experts, *IEEE Transactions on Neural Networks and Learning Systems* **23**(8): 1177–1193.

Zorluoglu, G. and Agaoglu, M. (2015). Diagnosis of breast cancer using ensemble of data mining classification methods, *International Journal of Bioinformatics and Biomedical Engineering* **1**(3): 318–322.

Anam Haq received her MSc in computer engineering from the National University of Sciences and Technology, Rawalpindi, Pakistan. Currently, she is pursing her PhD in business intelligence at the Poznan University of Technology under the IT4BI-DC Erasmus Mundus Joint Doctorate program. Her research interests include medical image processing, data pre-processing, machine learning and clinical data fusion.

Szymon Wilk received his MSc, PhD and DSc degrees in computer science from the Poznan University of Technology, where he is now an assistant professor at the Faculty of Computing. He is also an adjunct professor at the Telfer School of Management, University of Ottawa. His research interests include medical informatics and clinical decision support, with special focus on construction of decision models from clinical data using machine learning techniques, formal representation and reasoning over clinical practice guidelines, modeling the behavior of healthcare teams and providing patient-oriented support for improved therapy adherence.

Alberto Abelló obtained his doctorate at the Polytechnic University of Catalonia, Barcelona, in 2002 and has been working as a lecturer and researcher at that university since then. Currently he is an associate professor. He has held research stays at the University of Granada (Spain), Technische Universität Darmstadt (Germany), Claude Bernard Lyon 1 University (France) and the University of the Republic (Uruguay). He has participated in more than 10 national research projects or networks of excellence, and has signed R&D agreements with companies such as Hewlett Packard, Zurich Insurance, SAP or the World Health Organization. Currently he is a local coordinator of the IT4BI-DC Erasmus Mundus PhD programme.

Received: 19 March 2018
Revised: 10 October 2018
Accepted: 6 December 2018