A framework for modelling the biomechanical behaviour of the human liver during breathing in real time using machine learning

D. Lorente[a,*], F. Martínez-Martínez[a], M. J. Rupérez[b], M. A. Lago[c], M. Martínez-Sober[a], P. Escandell-Montero[a], J. M. Martínez-Martínez[a], S. Martínez-Sanchis[b], A. J. Serrano-López[a], C. Monserrat[c], J. D. Martín-Guerrero[a]

[a]*Intelligent Data Analysis Laboratory (IDAL), Electronic Engineering Department, Universitat de València, Avda. Universitat s/n, 46100 Burjassot, Valencia, Spain*

[b]*Centro de Investigación en Ingeniería Mecánica (CIMM), Departamento de Ingeniería Mecánica y de Materiales, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain*

[c]*Departamento de Sistemas Informáticos y Computación (DSIC), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain*

Abstract

Progress in biomechanical modelling of human soft tissue is the basis for the development of new clinical applications capable of improving the diagnosis and treatment of some diseases (e.g. cancer), as well as the surgical planning and guidance of some interventions. The finite element method (FEM) is one of the most popular techniques used to predict the deformation of the human soft tissue due to its high accuracy. However, FEM has an associated high computational cost, which makes it difficult its integration in real-time computer-aided surgery systems. An alternative for simulating the mechanical behaviour of human organs in real time comes from the use of machine learning (ML) techniques, which are much faster than FEM. This paper assesses the feasibility of ML methods for modelling the biomechanical behaviour of the human liver during the breathing process, which is crucial for guiding surgeons during interventions where it is critical to track this deformation (e.g. some specific kind of biopsies) or for the accurate application of radiotherapy dose to liver tumours. For this purpose, different ML regression models were investigated, including three tree-based methods (decision trees, random forests and extremely randomised trees) and other two simpler regression techniques (dummy model and linear regression). In order to build and validate the ML models, a labelled data set was constructed from modelling the deformation of eight *ex-vivo* human livers using FEM. The best prediction performance was obtained using extremely randomised trees, with a mean error of 0.07 mm and all the samples with an error under 1 mm. The achieved results lay the foundation for the future development of some real-time software capable of simulating the human liver deformation during the breathing process during clinical interventions.

Keywords: Soft tissue deformation, Biomechanical behaviour, Liver, Machine learning, Tree-based regression

1. Introduction

The use of computer technology for surgical planning and also for guiding surgical interventions, commonly referred to as computer-aided surgery (CAS), has spread rapidly in the last decades. CAS can be conducted by means of a number of medical imaging technologies, such as X-ray computed tomography (CT), magnetic resonance imaging (MRI), X-ray radiography, and medical ultrasound. Such technological progress in medicine has enabled the introduction of minimally-invasive surgical techniques, which limit the size of incisions needed compared to traditional open surgeries. Thus, the recovery time for the patient, as well as the associated pain and risk of acquiring infections, is reduced. On the other hand, the main drawbacks of these surgical techniques are the limited mobility for the surgeon and the loss of direct contact with the operation site. In particular, CAS systems have allowed to assist surgeons during surgical interventions in real time, minimising the problem of visibility and, therefore, maximising their precision, while reducing invasion into human bodies. Therefore, in the literature, different methods have been applied to assist surgeons especially for liver segmentation (Göçeri, 2013, 2016).

When performing surgical interventions on internal organs, such as the liver or the breast, the inclusion of biomechanical models that simulate their mechanical response during the intervention is becoming increasingly important in CAS. An accurate biomechanical model of an organ can significantly improve the performance of the surgical technique, as well as predict the outcome of the intervention. Therefore, progress in biomechanical modelling of human organs is the basis for the development of new clinical applications capable of improving the diagnosis and treatment of some diseases (e.g. cancer), as well as the

*Corresponding author. Tel.: +34 963543421

Email addresses: delia.lorente@uv.es (D. Lorente), francisco.martinez-martinez@uv.es (F. Martínez-Martínez), mjrupere@upvnet.upv.es (M. J. Rupérez), Miguel.Lago@psych.ucsb.edu (M. A. Lago), marcelino.martinez@uv.es (M. Martínez-Sober), pablo.escandell@uv.es (P. Escandell-Montero), jose.maria.martinez@uv.es (J. M. Martínez-Martínez), sanmars1@upv.es (S. Martínez-Sanchis), antonio.j.serrano@uv.es (A. J. Serrano-López), cmonserr@dsic.upv.es (C. Monserrat), jose.d.martin@uv.es (J. D. Martín-Guerrero)

surgical planning and guidance of some interventions. In this sense, a considerable research effort has been made in order to simulate the biomechanical behaviour of the soft tissue (Meier et al., 2005).

Many research works have been focused on using spring-mass methods to simulate the organ deformation in real time due to their simplicity of implementation and their low computational complexity (Kenedi et al., 1975; Waters, 1992; Delingette et al., 1994; Duysak et al., 2003). However, spring-mass models do not allow to reproduce the existing non-linear behaviour of the soft tissue and, therefore, they lead to inaccurate modelling of the mechanical response of the organs.

Instead, the finite element method (FEM) can provide a more physically-realistic and accurate solution by using knowledge about the soft tissue or organ (e.g. organ geometry, elastic constants and boundary conditions of the problem). In fact, FEM is one of the most popular methods used to predict the deformation of the human soft tissue in medical applications (Brock et al., 2006; Ruiter et al., 2006; Brock et al., 2008).

FEM is a well-known numerical method for the simulation of the mechanical behaviour of a continuum body (Zienkiewicz & Taylor, 1989). In FEM, an approximate discrete representation of the organ under study can be obtained by dividing the organ in a high number of elementary building components called finite elements, which are interconnected at points called nodes that define the element size. Finite elements can use physical properties, such as elastic properties, thus integrating tissue characteristics into the organ model. The entire set of these components is called mesh, which is defined through nodes and elements. A mesh is usually built from volumetric images (e.g. CT or MRI) of the organ. The individual equations that govern the mechanical behaviour of the finite elements under external loads are assembled into a larger system of equations that models the entire organ. An approximate solution of these equations can be found through computations on the nodes. Since the number of the equations to solve is proportional to the number of nodes, the larger the number of nodes, the more accurate the solution.

Despite its high accuracy, the use of FEM is limited due to the high computational cost involved. Hence, FEM has been typically used to perform off-line simulations of non-linear/complex behaviours and, therefore, its integration in CAS systems has been difficult due to real-time requirements (i.e. the organ models used in CAS must be deformed in the same way as the real organs, and at the same time). Several techniques have been proposed to reduce the computational time of conventional FEM in clinical applications, such as parallel processing algorithms (Székely et al., 2000; Inoue et al., 2006), the use of graphics processing units (GPU; Courtecuisse et al., 2010) or model reduction techniques (Niroomandi et al., 2008, 2013; González et al., 2016). For example, Székely et al. (2000) used parallel processing algorithms to speed up FEM when simulating the deformation of the uterus due to the interaction with the surgical instruments. Inoue et al. (2006) employed FEM in real time for the development of a liver surgical simulator by means of the use of parallel processing, coupled with volume rendering (i.e. using a relatively coarse volumetric mesh with less than 400 nodes). In other research (Courtecuisse et al., 2010), a GPU implementation was used to significantly improve the computational cost of FEM in the simulation of the interactions between the medical devices and the liver. With regard to the use of model reduction techniques, Niroomandi et al. (2008) applied a new strategy based on proper orthogonal decomposition (POD) techniques to the real-time simulation of the cornea deformation due to the palpation with the surgical tool. Later on, Niroomandi et al. (2013) presented a novel approach based on the use of the so-called proper generalised decomposition (PGD) techniques (i.e. a generalisation of POD techniques), which was also applied to the simulation of the deformation of the liver due to its interaction with the scalpel. Although the commented techniques led to good performance, they were computed only for a organ (i.e. unique geometry and unique elastic constants). Therefore, they could be applied only to predict deformations of that particular organ. This was due to the difficulty in introducing the geometry and the elastic constants as input parameters (i.e. external selectable parameters) into the models based on FEM, since they were fixed parameters required for the construction of the explicit biomechanical model of the organ. Nevertheless, some successful research works have been recently carried out in this regard. For instance, González et al. (2016) presented an approach combining PGD and kernel principal component analysis (kPCA) that was applied to simulate the liver deformation, in which the shape of the liver was also considered to be an external selectable parameter by using several livers with different geometries when building the algorithm, thus making the proposed modelling framework more general. In spite of some promising recent efforts, the development of FEM-based models that can accurately predict the deformation of soft tissue in real time is still a challenge in the field of CAS. In this sense, as computational capacity is increasing rapidly from year to year, it is expected that, in the future, FEM-based simulations will be run in real time and, therefore, they could be used for clinical applications. However, to date, it is not possible to perform real-time simulations based on FEM and, consequently, new techniques that require less computational cost than FEM are arising.

A possible alternative to FEM-based models for simulating the mechanical behaviour of human organs in real time could come from the use of data-driven modelling. In this modelling approach, data are used to feed a supervised machine learning (ML) model in order to find a function of the input variables (e.g. external load applied to the tissue, biomechanical parameters or elastic constants, and the corresponding geometry of the soft tissue) that can approximate the known outputs (e.g. deformation of the soft tissue), with this function being capable of generating an output for future unseen inputs (Izenman, 2008). Hence, the performance of a ML model depends mainly on the collected data and the chosen learning algorithm. Therefore, the ML model does not require an explicit biomechanical model of the organ, as happens in FEM, but it performs simulations based only and exclusively on data. Within this framework, FEM can be used to generate data that the ML model uses to estimate the mapping function (i.e. training samples).

Thus, ML models can extract the underlying properties of training samples, for which the deformations are known, and, then, predict soft tissue deformation when exerting a new load. The main advantage of data-driven modelling compared to FEM is that, although the estimation of the mapping function might be very time-consuming, once this training process is done off-line, ML models are able to provide solutions in real time for complex biomechanical behaviours of organs.

Most of the research works using data-driven modelling in this field have been focused on the real-time simulation of haptic interactions with organs (i.e. soft tissue deformation with cutting and haptic force feedback), such as the liver (Zhong et al., 2006; Morooka et al., 2008; Abdelrahman et al., 2011), the stomach (Deo & De, 2009; De et al., 2011) and the prostate (Jahya et al., 2013). In particular, the cited studies used an artificial neural network (ANN), which exploited FEM-generated data to predict the deformation of organs. FEM was applied several times using different load states (i.e. external forces) on a particular organ (i.e. unique geometry/mesh and unique elastic constants) and the deformed states were stored. After training the ANN with these data, the ANN was able to predict new deformed states from external loads that were not used during training for this particular organ. As a result, the ML models were able to predict deformation when a force/displacement was applied to a contact point of a specific organ with a rigid tool. It should be particularly emphasised that the commented research works were able to perform simulations in real time only for a specific organ, particularly that used for the training process of the ML model.

The research works using a ML model were mainly oriented towards the development of real-time surgical simulators, capable of predicting the deformation of a human organ due to the interaction with the surgical instruments, thus giving surgeons the chance of training surgical techniques. However, as far as the authors know, no research has been reported to perform accurate simulations of other organ deformations in real time using data-driven modelling, such as the liver deformation during breathing, which can be crucial for diagnosis or treatment of some diseases. The simulation of the deformation of the human liver during the breathing process in real time is of vital importance for guiding surgeons during interventions where it is critical to track hepatic lesions suspicious to be a tumour (e.g. some kind of biopsies) or for tracking liver tumours for the accurate application of the radiotherapy dose. The need to predict the liver motion during breathing mainly comes from the fact that changes in the liver geometry, mainly driven by physiological changes (e.g. respiratory and digestive motions), may confound the ability to accurately plan a hepatic biopsy when the tumour is very small. In addition, variations in geometry can also confuse the capacity to accurately deliver the radiation dose and measure the response of the tumour and surrounding normal tissue to radiotherapy treatment (Brock et al., 2006). In fact, due to the immense importance of simulating the liver deformation during breathing in the medical field, some studies have attempted to tackle this problem, to a certain extent, but using FEM-based models (Brock et al., 2006, 2008) with an associated high computational cost. Therefore, the inclusion of a fast and accurate model of the liver deformation during breathing in CAS systems could allow to know the exact location of internal lesions in the liver at every moment.

Regarding the accuracy required for models of soft tissue deformation, some related research works (Brock et al., 2006; Ruiter et al., 2006; Brock et al., 2008) achieved displacement errors of about 3-5 mm in the simulation of the deformation of human organs, such as the breast and the liver, using accurate FEM-generated models. However, the specification of precision requirements when using medical imaging, for instance, in cancer diagnosis and treatment or surgical guidance is still an open question (Brock et al., 2006). For example, Ruiter et al. (2006) stated that a displacement error within 5 mm could be regarded as an acceptable error for these medical systems. On the other hand, Brock et al. (2008) declared that, in stereotactic body radiotherapy (SBRT), a systematic geometric error higher or equal to 3 mm may have clinically-relevant dosimetric consequences, since the delivered radiation dose is really high in this kind of radiation therapy.

In this regard, the main goal of the present research work was to evaluate the feasibility of ML techniques for modelling the biomechanical behaviour of the human liver during the breathing process. As commented above, most of the previous research studies related to the modelling of human organ deformation were able to perform simulations only for a particular organ (i.e. unique geometry and unique elastic constants). Therefore, other important goal was to present a data-driven modelling scheme more general than those in other studies, not only capable of predicting the soft tissue deformation when applying a new load, but also for a new liver. For these purposes, instead of using an ANN as ML model, this research used, among others, several tree-based methods, since they were proved to outperform ANN models in a thorough comparative study of ML techniques (Fernández-Delgado et al., 2014).

The rest of this paper is organised as follows. In Section 2, the FEM-generated data set used to train and evaluate the ML models is described in detail and, in addition, a brief description of the ML models and details about how to use them in the problem of predicting the deformation of the human liver during breathing are also given. The results achieved in this research are shown and discussed in Section 3, including the results obtained from the hyperparameter optimisation for the different ML models and the results corresponding to the performance assessment of the models. Section 4 presents the conclusions drawn from this research work. Finally, possible lines for future research are identified in Section 5.

2. Material and methods

2.1. Data collection

2.1.1. Biomechanical modelling of the liver using FEM

Eight *ex-vivo* human livers from anonymous donors (discarded for transplantation) were used in this research. Each liver was placed in a device called artificial human torso (AHT), which was originally designed and built in order to simulate the deformation of the liver caused by human breathing, as further

described in Martínez-Martínez (2014). The AHT mainly consisted of the artificial diaphragm, the foam and the liver cavity. The diaphragm could be moved in the superior-inferior direction, thus pushing the liver towards the foam, which emulated the behaviour of the remaining abdominal organs. All the materials used for building the AHT were chosen in order to avoid the generation of artefacts during the image acquisition with the CT scanner. The AHT size was approximately $180 \times 400 \times 600$ mm. The AHT was in turn positioned into a 256-slice computer tomography (CT) scanner (Brilliance iCT, Philips Healthcare, Best, The Netherlands), as shown in Figure 1. A single CT image was acquired from each of the livers, with the artificial diaphragm of the AHT being always placed at the position corresponding to complete exhalation (i.e. the diaphragm did not touch the liver at all). The acquired three-dimensional (3D) images had a size of $512 \times 512 \times 258$ voxels with a voxel size of $0.64 \times 0.64 \times 1.5$ mm. The commercial software Simpleware (version 6.0; Synopsys, Inc., Mountain View, California, USA) was used to segment each liver and generate a 3D finite element (FE) mesh.

Figure 1: Image acquisition of an *ex-vivo* human liver with a CT scanner.

The Ogden model was chosen in this work to represent the biomechanical behaviour of each liver, since it provided better results when modelling livers than other models (Hu & Desai, 2004; Martínez-Martínez et al., 2013a). This model is defined through the strain energy potential (W), as shown in Equation (1):

$$W = \sum_{i=1}^{N} \frac{\mu_i}{\alpha_i}(\overline{\lambda}_1^{\alpha_i} + \overline{\lambda}_2^{\alpha_i} + \overline{\lambda}_3^{\alpha_i} - 3) + \frac{K_0}{2}(J-1)^2 \qquad (1)$$

where N indicates the order of the model; α_i and μ_i refer to the elastic material constants; $\overline{\lambda}_1$, $\overline{\lambda}_2$ and $\overline{\lambda}_3$ denote the deviatoric stretches; K is the Bulk modulus, related to the liver compressibility; and J is the determinant of the elastic deformation gradient. The model used in this work was of order $N = 2$, thus requiring four elastic material constants: μ_1, α_1, μ_2 and α_2. For each of the eight livers, three different biomechanical behaviours were modelled using the combinations of elastic material constants shown in Table 1. This strategy allowed to create a richer data set from the eight *ex-vivo* human livers. Therefore, each liver had a different geometry (i.e. FE mesh) and the same three biomechanical behaviours (i.e. combinations of elastic material constants). The elastic constants used in this research were obtained from three *ex-vivo* human livers, different from those used for obtaining the liver FE meshes, following the procedure described in Martínez-Martínez (2014) and their values were comparable in order of magnitude with those reported for the human liver in the literature (Lu et al., 2013). The value of the Bulk modulus (K_0) was set to 10 kPa, based on the results reported in Hostettler et al. (2010) for the Bulk modulus of the human liver *in vivo*.

Table 1: Combinations of elastic material constants used for the Ogden model to represent the liver tissue biomechanical behaviour.

Combination of elastic material constants	μ_1 (kPa)	α_1 (-)	μ_2 (kPa)	α_2 (-)
Combination 1	66.17	61.17	22.03	98.75
Combination 2	59.34	-50.00	66.91	21.50
Combination 3	11.90	-36.46	57.67	99.70

FEBio (version 1.5), freely available software for non-linear FE analysis (Maas et al., 2012), was used to recreate the deformation that the liver suffers during breathing. Breathing process can be divided into two different stages: inhalation and exhalation. During inhalation, the diaphragm is contracted and moved in the superior-inferior direction around 15 mm, this value being approximately in the middle of the range of the measured diaphragm displacements during breathing corresponding to several patients reported in the study by Balter et al. (2001). Thus, the volume of the thoracic cavity is enlarged and the liver is pushed towards the remainder of abdominal organs. The boundary conditions from Martínez-Martínez et al. (2013b) were replicated in order to simulate the liver compression during inhalation process. Ten different external displacements were applied to each liver in the z-axis direction, from an initial displacement of 1.5 mm to a final displacement of 15 mm in steps of 1.5 mm, thus obtaining ten deformed states for each liver. As shown in Figure 2, each external displacement was applied to the 10% of top nodes from the surface of each liver mesh, oriented in the frontal plane (i.e. plane x-z), thus emulating the movement of the upper surface of the liver during inhalation. The 10% of bottom nodes from the surface of each liver mesh were restricted in all directions, thus attempting to imitate the direct contact of the human liver with the rest of abdominal organs during inhalation process. Each liver deformed state was defined through the nodal displacements in the 3D Euclidean space provided by FEBio after the simulation.

An aspect that should be commented is that the boundary conditions used in FEM were chosen just as a simplistic way to simulate the real boundary conditions during breathing. This simplification was performed due to the difficulty in measuring the real loads that the liver undergoes inside the body during breathing or the real displacements to which some parts of the liver are subjected because of the compression produced by the diaphragm. However, since the main objective of this work was to prove that accurate ML models could be constructed using data from FEM-based simulations of the breathing process, the simplistic boundary conditions used in this work provided enough information to do this. In this sense, once fulfilled the main goal, a further step in the framework of this research would be to use more realistic boundary conditions when generating data with FEM, capable of reproducing more faithfully the real breathing process. For example, this could be done by performing two CT scans of each liver (in complete exhalation and in complete inhalation) and then obtaining the real displacement of the nodes of the liver surface by a registration algorithm.

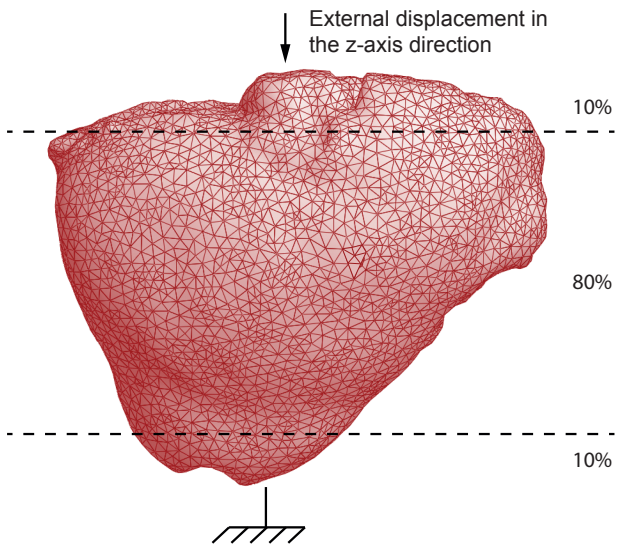


Figure 2: Example of a FE mesh of a human liver with the boundary conditions used by the FE analysis software.

2.1.2. *Labelled data set*

The main goal of this work was to predict the behaviour of the human liver under different external displacements that recreated the liver compression during breathing. This problem was tackled using ML regression models. The supervised nature of regression models required the use of a set of n labelled samples, $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1,\dots,n}$, where $\mathbf{x}_i$ was the m-dimensional input vector for the i-th sample with an associated k-dimensional target output vector $\mathbf{y}_i$, which we desired to predict. Particularly, once the biomechanical behaviour of the human livers was modelled using FEM, it was constructed a labelled data set containing information about the eight human livers. Each organ was described by a mesh of points (or nodes) distributed throughout the volume of the liver. The geometry (i.e. size and shape), and consequently the FE mesh, varied from one liver to another, and three different biomechanical behaviours were modelled for each liver. The following variables were available for each liver node:

- $disp$: External displacement applied to the liver to which the node belonged. This variable could take ten different values, from an initial displacement of 1.5 mm to a final displacement of 15 mm in steps of 1.5 mm.
- V: Volume of the liver to which the node belonged.
- x, y, z: Initial coordinates of the node in the 3D space before applying any displacement.
- $\mu_1, \alpha_1, \mu_2, \alpha_2$: Elastic material constants characterising the biomechanical behaviour of the liver tissue.
- dx, dy, dz: Displacement of the node in the 3D space, obtained using FEM.

The nodes used in the boundary conditions in FEM were not included in the labelled data set, since these nodes were considered as parameters of the FEM model (i.e. they were fixed beforehand). Each of the eight livers was composed of a slightly different number of nodes, as shown in Table 2, including the ten liver deformed states for each of the three biomechanical behaviours, thus resulting in a labelled data set with a total of $n = 3,154,980$ samples. For each sample, $m = 9$ variables were used as inputs, including $disp$, V, x, y, z, α_1, α_2, μ_1 and μ_2, and the remaining $k = 3$ variables were employed as target outputs (i.e. labels), including dx, dy and dz (i.e. the displacement of the corresponding node of the liver mesh).

Table 2: Number of nodes (samples) for each liver after the removal of the nodes used in the boundary conditions in FEM from the labelled data set.

Liver	Number of nodes
Liver 1	379,800
Liver 2	326,520
Liver 3	318,960
Liver 4	494,310
Liver 5	331,830
Liver 6	390,390
Liver 7	492,480
Liver 8	420,690

2.2. *Regression models*

In this work, five different regression models were used to predict the human liver biomechanical behaviour: dummy model (DM), linear regression (LR), decision tree (DT), random forest (RF) and extremely randomised trees (ET). The regression models, as well as all the experiments conducted in this work, were implemented in Python programming language (version 2.7) using mainly the scikit-learn module (version 0.17; Pedregosa et al., 2011), which integrates a wide range of ML algorithms. A brief description of the regression models is given below.

2.2.1. Dummy model

A DM always returns the same output value regardless of the input data. Although this model has not practical usefulness for making predictions, it is commonly used as a simple baseline to be compared with other real regressors. In particular, the output value of a DM is computed by averaging the labels (i.e. targets) belonging to the training set. Therefore, the prediction error associated with this model is coincident with the standard deviation of the data distribution.

2.2.2. Linear regression

In a LR model, the output is computed as a linear combination of the input variables, with the relationship between the inputs being established through the model parameters (i.e. regression coefficients; Bishop, 2006). The regression coefficients are commonly obtained from the training data using the ordinary least squares approach, which minimises the sum of the squares of the differences between the target outputs and the outputs predicted by the linear approximation. Therefore, LR is also known as ordinary least squares linear regression. This model assumes a linear relationship between the input variables and the output. Obviously, this strong assumption is not always fulfilled in real problems, which usually deal with more complex data involving non-linear relationships. However, LR models are commonly used because of their advantages (e.g. simplicity, interpretability and computational cost) compared with more sophisticated methods.

2.2.3. Decision tree

A DT is an efficient non-parametric method widely used for classification and regression problems (Breiman et al., 1984). A DT model is basically built or grown from the training set according to a top-down procedure by computing recursive binary partitions of the input feature space so that the samples with the same label are grouped together. At each step of the procedure, a division rule is specified by choosing the split (i.e. the combination of an input variable and the corresponding split-point for that variable) that maximises a homogeneity measure of the target variable within each of the obtained groups. In regression problems, a commonly used criterion to choose the best split at each node of a DT model is to select the split that minimises the mean squared error. This recursive process continues until some stopping rule specified by the user is satisfied. A common stopping rule is that a tree node can be split if it contains more than a certain number of samples (n_{min}). Therefore, the minimum number of samples required to split a tree node should be adjusted by the user in order to control the size of the DT model, thus preventing overfitting.

Once a DT is grown, the output for a given unseen sample is computed in a straightforward way by passing down the tree through all the nodes at which a decision is made as to which direction to proceed based on the value of each input variable. Finally, a terminal tree node is reached and a predicted output is given, computed as the mean of training samples in that tree node in the case of regression problems.

In addition to the fact that a DT model is simple to understand and interpret, it is a powerful model, thus providing a suitable compromise between accuracy and interpretability. On the other hand, one major problem with DT models is that they can be extremely unstable, since small variations in the training data can lead to a very different tree model. Other drawback is that, if a DT is too complex, the model might not generalise the data well (i.e. it may suffer from overfitting).

2.2.4. Random forest

A RF regressor is a tree-based ensemble learning method, which builds several DT models independently and then averages their individual predictions to compute a final prediction (Breiman, 2001). In a RF model, each single tree belonging to the ensemble is built from samples drawn randomly with replacement (i.e. bootstrap sampling) from the training set. Furthermore, when splitting a tree node during the building of a tree, the split that is chosen is the best split among a random subset of the input variables, selected without replacement, instead of picking the best split among all the input variables as that done in a single DT model.

As a direct consequence of the randomisation of the tree growing method combined with ensemble averaging, the variance of a RF model is reduced compared to that associated with a single non-random DT. Thus, several advantages are achieved, such as the improvement of the prediction performance, the reduction in the sensitivity to small changes in the training data and the control of overfitting, at the expense of a small increase in the bias, the loss of interpretability and a higher computational cost.

The main hyperparameters that need to be specified when building a RF model are the minimum number of samples required to split a tree node (n_{min}), the number of trees in the forest (M), and the number of input variables randomly selected to consider when looking for the best split (K). These parameters should be optimised, since they have considerable influence on the prediction performance and the computational complexity.

2.2.5. Extremely randomised trees

ET is another tree-based ensemble learning method (Geurts et al., 2006). In ET models, one further step of randomisation is added when splitting a tree node by randomising the choice of both input variable and cut-point. As in RF models, ET models use a random subset of the input variables, but instead of searching for the best split-points, for each candidate input variable, its split-point is chosen fully at random (i.e. independently of the target variable), and then the best split among all the randomly-generated splits is picked. Thus, the variance and the computational complexity of a ET model are reduced a little more compared to those of a RF model, at the expense of a slightly greater bias. Another key difference between RF and ET is that ET models use the entire original training set to build each tree, instead of using bootstrap sampling, thus trying to minimise the bias.

The main hyperparameters to adjust when building an ET model are exactly the same as those for a RF model, including the minimum number of samples to split a node (n_{min}), the number of trees in the ensemble (M), and the number of input variables randomly selected at each node (K).

2.3. Experimental setup

In this section, details are given about how the different regression models were used in the problem of predicting the displacement of the nodes of the liver mesh employing the labelled data set described in Section 2.1.2.

2.3.1. Single-output modelling

The problem in this work was multi-output, also known in the literature as multi-target, since it was a supervised learning problem that aimed to predict three output variables simultaneously, namely, the displacement of each liver mesh node in the 3D space: *dx*, *dy* and *dz*. In order to tackle the problem, it was decided to build an independent regression model for each of the three outputs (i.e. a single-output model) and then to use these models to predict each output independently, instead of building a single regression model capable of predicting the three outputs simultaneously (i.e. a multi-output model). This choice was based on the fact that the outputs were weakly correlated, with a Pearson correlation coefficient (Pearson, 1895) close to 0 between each possible pair of output variables (0.038 for *dx*-*dy*, -0.036 for *dx*-*dz* and 0.079 for *dy*-*dz*). Under these circumstances, the use of several single-output models was expected to result in better prediction performance than using a multi-output regression model, especially in the case of tree-based learning algorithms (Struyf & Džeroski, 2006; Appice & Džeroski, 2007; Ikonomovska et al., 2011), as those used in this work. In addition, the use of several single-output models was supported by several preliminary experiments, which were conducted using both single-output and multi-output regression models. In accordance with our expectations, the results (data not shown in this paper) revealed that the prediction performance obtained by using three single-output models surpassed that obtained by only one multi-output model.

2.3.2. Data splitting

Following the common practice when dealing with ML algorithms, the labelled data set was split into two subsets in order to obtain a reliable performance evaluation of the regression models: a training set and a test set. The training set was employed to build the models (i.e. adjust the model parameters by learning the data properties) and optimise their hyperparameters. The test set was used to evaluate the performance of the models, thus checking their generalisation capability.

Two different approaches were used to split the data. In the first approach, the training set was composed of the labelled samples belonging to seven livers, while the test set was composed of the samples belonging to the remaining liver. The experiments using this splitting strategy were repeated leaving one different liver as the test set each time. However, given that the results were very similar in all cases, this article reports only one case for the sake of simplicity. In particular, liver 8, as named in Table 2, was chosen for testing purposes, thus resulting in a training set with 2,734,290 samples (i.e. about 87% of the total) and a test set with 420,690 samples (i.e. about 13% of the total). Since the test set contained samples of a new liver (i.e. a liver different from those used for training), this approach would have been the most suitable in the context of the problem tackled in this work of predicting the displacement of the nodes of an unknown liver, as would happen in a real application. However, the weak point of this splitting strategy was based on the limited number of livers used for the experiments and the subsequent small range of liver geometries that were modelled by FEM. For the implementation of the regression models in a real application, many more livers should be employed to build the models in order to properly predict the displacement of a liver within a wider range of geometries and biomechanical properties.

In this sense, a second splitting approach was used in order to maximise the variety of liver geometries in the training set, which was expected to recreate a more realistic scenario. It should be commented that it was only possible to maximise the range of liver geometries, and not the variety of biomechanical properties as well, since each of the eight livers used in this work had a different geometry, but the same three biomechanical behaviours. In this second approach, 70% of the total samples (i.e. 2,208,486 samples), randomly drawn from the complete labelled data set, were used as the training set, whereas the remaining 30% of samples (i.e. 946,494 samples) were used as the test set.

2.3.3. Hyperparameter optimisation

Hyperparameter optimisation was one of the main tasks performed in this work when developing the regression models. Hyperparameters are parameters that are not directly learnt within the models and, therefore, they should be provided by the user when building the models. The process of hyperparameter optimisation was only required for the DT, the RF and the ET models, since the DM and the LR models are parameterless methods. Therefore, the hyperparameter optimisation procedure was performed for a total of $3 \times 3 \times 2 = 18$ regression models, since the three required single-output models (i.e. one for each target output) were built using each of the tree-based regression methods (i.e. the DT, the RF and the ET techniques), and both data splitting approaches. In addition, it is important to comment that the whole optimisation procedure was carried out using only the training sets.

The main hyperparameters to adjust when constructing the RF and the ET models were exactly the same, including the minimum number of samples to split a tree node (n_{min}), the number of trees in the ensemble (M), and the number of input variables randomly selected at each node (K). In the case of the DT models, only one hyperparameter needed to be optimised: the minimum number of samples to split a tree node (n_{min}). The parameter n_{min} controls the size of the three tree-based models. Small values of n_{min} lead to large trees, high variance and low bias. In consequence, a really small value usually implies that the model memorises the training set, thus resulting in a poor generalisation capability of the model (i.e. overfitting), whereas a large value could prevent the trees from learning from the data (i.e. underfitting). Therefore, making a good choice of this parameter was of critical importance in this work to prevent the overfitting problem and reduce the computational complexity. With respect to the parameter M, it determines the strength of

the variance reduction associated with the ensemble averaging. In principle, the larger the value of M, the better the model from the point of view of prediction performance, but also the more computationally complex the model. Regarding the parameter K, it controls the strength of the randomisation of the tree growing method. Specifically, the smaller the value of K, the stronger the randomisation of the trees in the ensemble, the greater the variance reduction, and the lower the computational complexity, but also the greater the increase in the bias.

The optimisation of all the hyperparameters was performed so that a good compromise between prediction performance and computational complexity was achieved. The methodology used to select the optimal hyperparameters for each regression model was to perform a parameter sweep for each hyperparameter. More specifically, the sweep of a hyperparameter consisted in building different models by varying the value of that parameter and then evaluating their corresponding prediction performance, thus obtaining the evolution of the model performance as a function of the parameter value. For the optimisation of the parameter n_{min} for the DT models, the value of n_{min} was varied from 2 to 100. In the case of the RF and the ET models, each regression model required three different parameter sweeps, one for each hyperparameter. When performing the sweep of a particular hyperparameter of these two models, the value of the hyperparameter to optimise was varied over a range and the remaining two hyperparameters were fixed. In particular, the values of n_{min}, M and K were set to 5, 10 and 9 (i.e. the dimension of the input space), respectively. For optimising the parameter n_{min} for the two ensemble models, the value of the parameter was varied from 2 to 100, as that done for the DT models. In the case of the parameter M, the optimisation of this parameter was performed by varying its value from 2 to 100. For the sweep of the parameter K, the value of this parameter was varied over its possible range between 1 and 9 (i.e. the dimension of the input space).

In order to evaluate and compare more appropriately the performance of the different hyperparameter settings for each regression model, it was used a cross-validation procedure on each of the training sets obtained by both data splitting approaches. Particularly, the experiments were performed with five-fold cross-validation (Hastie et al., 2009), which consisted in splitting the training set randomly into five subsets of equal size and using four subsets as training data to build the regression models and the remaining one as test data to evaluate the prediction performance of the resulting models. This process was repeated five times leaving one different subset for evaluation each time. The five different validation results were subsequently averaged to produce a single performance measure, thus reducing variability due to random partitioning and obtaining a more reliable performance estimation. In this work, the performance measure used for hyperparameter optimisation was the root mean squared error (RMSE), as further explained in Section 2.3.4.

2.3.4. Performance assessment of the regression models

After building the regression models using the training sets associated with both data splitting approaches, the test sets were used to evaluate and compare the performance of the models in the prediction of the displacement of the liver mesh nodes. The first step in the performance assessment procedure was to compute the error associated with the prediction of each individual Cartesian coordinate corresponding to the displacement of each liver node i (i.e. the errors associated with the prediction of dx, dy and dz), defined as the difference between the output estimated by the model $\hat{y}_i$ and the actual output y_i (i.e. the corresponding label). In this work, this kind of error related to each individual displacement coordinate for a particular sample was referred to as the *coordinate error*.

Another more suitable metric used to evaluate the model performance in the prediction of each displacement coordinate was the root mean squared error (RMSE). The RMSE was computed considering all the samples included in the used test set, thus being regarded as a global performance measure. Particularly, the RMSE for each output was computed from all the coordinate errors associated with that particular output, in Equation (2):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n_{test}} (\hat{y}_i - y_i)^2}{n_{test}}} \tag{2}$$

where n_{test} is the number of samples in the used test set, $\hat{y}_i$ is the output estimated by the model for the i-th sample and y_i is the actual output for the i-th sample.

In addition to evaluating the performance of the models in the prediction of each displacement coordinate individually, it was of major importance to use other performance measures capable of taking into account the model performance related to the three displacement coordinates. This type of performance measure was absolutely required due to the fact that good performance when predicting one or two displacement coordinates did not ensure good performance in the prediction of the remaining coordinates, thus possibly leading to a general prediction failure. In this sense, for each liver mesh node i, the Euclidean distance between the displacement of the node in the 3D space predicted by each regression model and the actual displacement of the node was calculated, as shown in Equation (3):

$$d(\hat{\mathbf{y}}_i, \mathbf{y}_i) = \sqrt{(\hat{dx}_i - dx_i)^2 + (\hat{dy}_i - dy_i)^2 + (\hat{dz}_i - dz_i)^2} \tag{3}$$

where $\hat{y}_i$ is the predicted displacement vector in the 3D space for the i-the sample with the Cartesian coordinates $(\hat{dx}_i, \hat{dy}_i, \hat{dz}_i)$ and $\mathbf{y}_i$ is the actual displacement vector for the i-th sample with Cartesian coordinates (dx_i, dy_i, dz_i). An illustrative example of the computation of this Euclidean distance in the 3D space is shown in Figure 3. In this work, this performance measure was referred to as the *Euclidean error*.

Furthermore, as an attempt to consider all the samples in the test set used for assessing the model performance, three different global performance metrics based on the Euclidean errors were computed: the mean Euclidean error, the percentage of samples with a Euclidean error lower or equal to 1 mm and the percentage of samples with a Euclidean error lower or equal to 3 mm. The mean Euclidean error was calculated by averaging the Euclidean errors corresponding to all the samples in the test set. The percentage measures were computed as the number

of samples with a Euclidean error lower or equal to the corresponding upper limits divided by the total number of samples. Regarding the upper limits, the choice of the limit of 3 mm was based on the precision of the results obtained in other similar research works related to the modelling of soft tissue deformation (Brock et al., 2006; Ruiter et al., 2006; Brock et al., 2008), in which displacement errors of about 3-5 mm were achieved in the simulation of the deformation of human organs, such as the breast and the liver, using accurate FEM-generated models. Therefore, the limit of 3 mm was used as a reference to check if this study was comparable with other related works in terms of accuracy. However, the specification of precision requirements when using medical imaging, for instance, in cancer diagnosis and treatment or surgical guidance is still an open question (Brock et al., 2006). For example, Ruiter et al. (2006) stated that a displacement error within 5 mm could be regarded as an acceptable error for these medical systems. On the other hand, Brock et al. (2008) declared that, in stereotactic body radiotherapy (SBRT), a systematic geometric error higher or equal to 3 mm may have clinically-relevant dosimetric consequences, since the delivered radiation dose is really high in this kind of radiation therapy. In addition, in order to check a more ambitious upper limit than that of 3 mm, the limit of 1 mm was also used for the performance assessment of the regression models, thus expecting that this work could go beyond the results achieved in other studies.

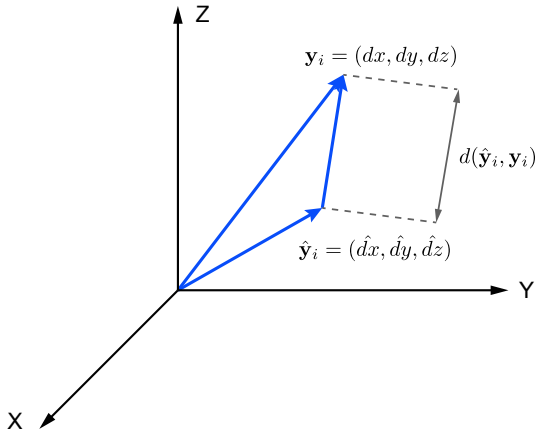


Figure 3: Displacement error in volume.

A peculiarity of the labelled data set used in this work was the great number of samples with an almost negligible displacement (i.e. their associated values of dx, dy and dz were zero or very close to zero). This situation could possibly lead to a misinterpretation of the global performance measures. This could be explained due to the fact that the samples with output values close to zero were therefore expected to be predicted with really low Euclidean errors, thus likely involving also a low mean Euclidean error, even though the predictions were completely wrong (i.e. with a high Euclidean error) for the samples with a displacement different from zero. Hence, it was necessary to study the effect of this issue on the model performance assessment. In particular, the three global performance measures based on the Euclidean errors mentioned above were also computed for the ten different external displacements applied to the liver (i.e. input variable $disp$). The samples with the same value of $disp$ were used for calculating these performance measures, instead of using all the samples in the test set, thus resulting in a total of ten values computed for each global measure. The choice of grouping the samples according to the input variable $disp$ was based on the realistic assumption that the value of the input variable $disp$ was directly related to the values of dx, dy and dz (i.e. the higher the input variable $disp$ was, also the higher dx, dy and dz). Within this framework, the relative Euclidean error was also computed for each liver mesh node i by dividing the corresponding Euclidean error (computed according to Equation (3)) by the length of the actual displacement vector $\mathbf{y}_i$ in the 3D space (computed as shown in Equation (4)). This kind of error also allowed to check if a low Euclidean error did not mean good prediction performance, but it was only due to the insignificant displacement of that particular liver node.

$$\|\mathbf{y}_i\| = \sqrt{dx_i^2 + dy_i^2 + dz_i^2} \tag{4}$$

3. Results and discussion

3.1. Hyperparameter optimisation

In order to select the optimal hyperparameters for the DT, the RF and the ET models, the sweep of each hyperparameter to adjust was performed using the training sets obtained by both data splitting approaches. Figure 4 shows the cross-validated RMSE values of the DT models as a function of the hyperparameter n_{min} for the displacement coordinates dx, dy and dz using the 70%/30% data splitting approach. Similarly, Figures 5 and 6 show the cross-validated performance of the RF and the ET models, respectively, by varying the value of the different hyperparameters (n_{min}, M and K) for the three displacement coordinates using the 70%/30% data splitting approach. For both data splitting approaches, the curves representing the evolution of cross-validated RMSE values of the regression models with the different hyperparameters presented exactly the same trends, although the RMSE values were slightly different. Therefore, for the sake of illustration, figures with the hyperparameter optimisation results are shown only for the 70%/30% data splitting approach. In addition, a glance at Figures 4, 5 and 6 shows that the curves exhibited exactly the same tendencies for the three displacement coordinates, even though there were differences in the RMSE values.

When observing the sweeps of n_{min} for the three tree-based regression models, it can be noticed that the prediction error had a monotonically increasing trend, as expected from the fact that a large value of n_{min} could prevent the trees from learning from the data (i.e. underfitting). Furthermore, the fact that the best performance was obtained using the minimum possible value of n_{min} (i.e. $n_{min} = 2$) suggested that the data set was free (or almost free) of output noise, since the noisier the output, the higher the value of n_{min} for which the best performance was obtained, as reported by Geurts et al. (2006). Considering

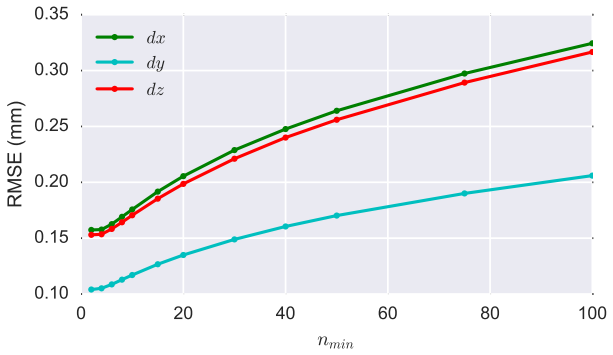


Figure 4: Evolution of the RMSE values of the DT models with the minimum number of samples to split a node (n_{min}) for the three displacement coordinates using the 70%/30% data splitting approach.

the sweeps of the hyperparameter M, as expected, it can be observed that the behaviour of the prediction error was a monotonically decreasing function of M for the RF and the ET models. In addition, it is noticeable that the prediction error converged as the number of trees increased. Smaller values of M implied higher speed of convergence and, therefore, the prediction error stopped decreasing significantly beyond a critical number of trees. With regard to the evolution of the RMSE values with K for the RF and the ET models, it can be noticed that the prediction error was a monotonically decreasing function and it converged when increasing the value of K, as happened for the sweeps of M. However, unlike for the sweeps of M, the slope of the prediction error curve was not so steep for the smallest values of K, but the error decreased more gradually with the value of K. All the commented trends in the sweeps of the different hyperparameters were in agreement with those reported in the research conducted by Geurts et al. (2006), in which the effect of the different hyperparameters on the prediction performance of the ET models was analysed.

From the parameter sweeps, the hyperparameter values were chosen as a trade-off between prediction performance and computational complexity. In particular, for the DT models, the value of the hyperparameter n_{min} was set to 5 for the three displacement coordinates using both splitting approaches. In the case of the RT and the ET models, the hyperparameter values were set to $n_{min} = 5$, $M = 40$ and $K = 7$ for the three displacement coordinates using both splitting approaches. The choice of $n_{min} = 5$ was based on the fact that, although the prediction error obtained using $n_{min} = 5$ was almost identical to the minimum possible error, obtained using $n_{min} = 2$, the setting of $n_{min} = 5$ presented the additional advantage of reducing the computational complexity compared to using fully developed trees (i.e. $n_{min} = 2$). Moreover, the value of n_{min} chosen in this work was in accordance with the value proposed in the study by Geurts et al. (2006), which suggested that the setting of $n_{min} = 5$ was a robust parameter choice for regression problems. Regarding the choice of the parameter M, $M = 40$ was selected as the most suitable because this value was large enough to ensure convergence of the prediction error due to the ensemble averaging effect. Furthermore, due to the enormous influence of parameter M on the computational requirements to build the models, the setting of M was attempted to be as small as possible in order to prevent the training time from bursting, while guaranteeing convergence of the error. Considering the curves for the sweeps of K, it can be seen that the choice of $K = 7$ ensured the convergence of the prediction error. In addition, the setting of K chosen in this work presented the additional benefit of requiring less computational requirements compared to using the highest possible value of K (i.e. $K = 9$), as done by Geurts et al. (2006), since the degree of randomisation of the trees in the ensemble was weaker when using $K = 7$. Hence, the different choice of the parameter K in this work was justified.

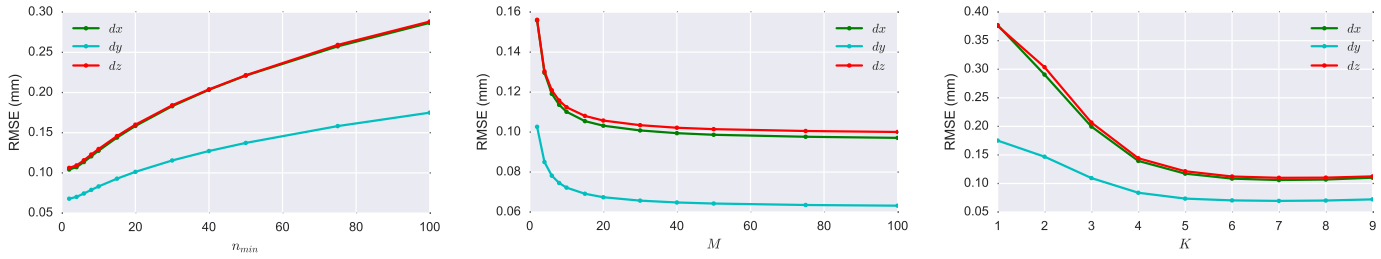
3.2. Performance assessment of the regression models

3.2.1. Overall results

Once the regression models were built using the optimised hyperparameters reported in Section 3.1 and the training sets associated with both data splitting approaches, the test sets were employed to validate the models in the prediction of the displacement of liver mesh nodes. Table 3 summarises the performance assessment results for all the regression models using both data splitting approaches. In particular, the three global performance metrics based on the Euclidean errors are shown, including the mean Euclidean error, the percentage of samples with a Euclidean error lower or equal to 1 mm and the percentage of samples with a Euclidean error lower or equal to 3 mm. An aspect that should be commented is that Table 3 does not show the results from evaluating the performance of the models in the prediction of each displacement coordinate separately, such as the RMSE values, but only the performance metrics that took into account the performance related to the three displacement coordinates as a whole, since the optimisation of this type of performance measure was the main goal of this work. Therefore, the higher the percentages of samples with Euclidean errors under 1 mm and 3 mm and the lower the mean Euclidean error associated with a particular regression model, the better the model from the point of view of prediction performance.

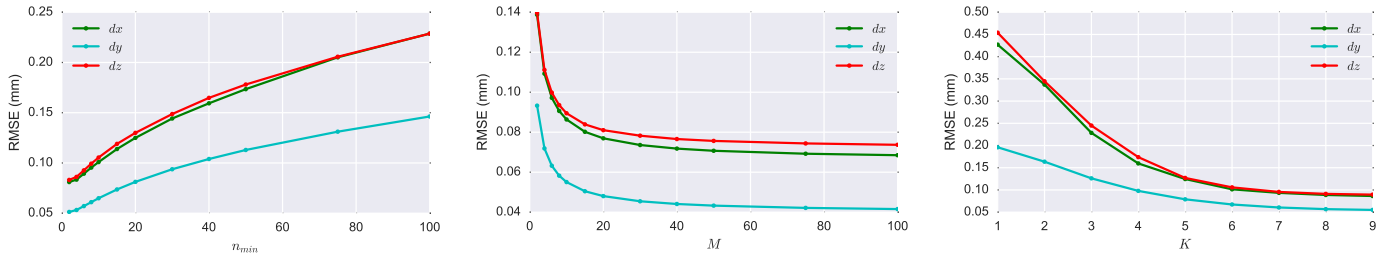
When comparing the performance assessment results for the different regression models, it can be noticed that the three tree-based methods (i.e. the DT, the RF and the ET techniques) clearly outperformed the DM and the LR models for both data splitting approaches. As expected, the DM models provided the worst prediction performance, since these models always returned the mean value of the labels in the training set and, therefore, they were simply taken as a reference to be compared with the other regression models. In contrast, the best results were achieved using the tree-based methods, with the three models yielding similar prediction performance for each data splitting approach. For the LR models, the performance results were in between those for the DM and the tree-based methods, this fact confirming that the problem in this work had non-linear nature.

Furthermore, comparison of the results for both data splitting approaches shows that, in general, the use of the 70%/30% data



(a) (b) (c)

Figure 5: Evolution of the RMSE values of the RF models with the different hyperparameters for the three displacement coordinates using the 70%/30% data splitting approach: (a) the minimum number of samples to split a node (n_{min}), (b) the number of trees in the ensemble (M) and (c) the number of input variables randomly selected at each node (K).



(a) (b) (c)

Figure 6: Evolution of the RMSE values of the ET models with the different hyperparameters for the three displacement coordinates using the 70%/30% data splitting approach: (a) the minimum number of samples to split a node (n_{min}), (b) the number of trees in the ensemble (M) and (c) the number of input variables randomly selected at each node (K).

splitting approach resulted in much better performance results than using the data splitting approach leaving one liver for test. Even though the DM and the LR models gave similar results for both data splitting approaches, the 70%/30% data splitting approach led to remarkably better prediction performance than the data splitting approach leaving one liver for test for the three tree-based methods. In particular, in the case of the three tree-based models, it can be observed that, while the percentage of samples with Euclidean errors under 3 mm was notably high (above 91%) for both data splitting approaches, the other two global performance measures were clearly improved when using the 70%/30% data splitting approach, with the percentage of samples with Euclidean errors under 1 mm increasing from about 55-60% to around 100% and the mean Euclidean error decreasing from a magnitude order of 1 mm to an order of 0.1 mm. In this sense, it should be particularly emphasised that, when using the tree-based methods, almost all the samples were predicted with a Euclidean error lower or equal to 1 mm for the 70%/30% data splitting approach, thus surpassing, to a large extent, the displacement errors of about 3-5 mm mentioned in other similar research studies regarding the modelling of human organ deformation (Brock et al., 2006; Ruiter et al., 2006; Brock et al., 2008). Anyway, even though the best results were obtained for the 70%/30% data splitting, the high percentage of samples with Euclidean errors under 3 mm of above 91% and the mean Euclidean error of the order of 1 mm achieved for the three tree-based models using the data splitting approach leaving one liver for test could be regarded as perfectly comparable with the results of about 3-5 mm obtained in the other related works. In addition, the results achieved for the tree-based models using the one liver for test data splitting could be considered more than acceptable according to the upper precision limit of 3 mm suggested for avoiding clinically-relevant dosimetric consequences in radiotherapy (Brock et al., 2008), and they were also well below the accuracy limit of 5 mm required when using medical imaging on human organs for early cancer diagnosis (Ruiter et al., 2006).

Due to the limited number of livers used for the experiments, the improvement in the results when using the 70%/30% data splitting approach was broadly in line with expectations, since the 70%/30% data splitting maximised the variety of liver geometries in the training set employed to build the models. Nevertheless, if many more livers had been employed to build the models, the results for both data splitting approaches would have been expected to be much closer. In fact, since the data splitting approach leaving one liver for test was more similar to what would happen in a real application, a new research work is currently being planned to be conducted using many more livers, thus attempting to increase the range of geometries and biomechanical properties compared to those used in this work.

Despite the similarity in performance of the three tree-based models, the best prediction performance results for the data splitting approach leaving one liver for test were obtained using the RF regression method, with percentages of samples with

Table 3: Performance assessment results for all the regression models using both data splitting approaches, including the three global performance metrics based on the Euclidean errors. The best results for both data splitting approaches are highlighted in bold.

	One liver for test			70%/30%		
Regression model	Samples with Euclidean errors ≤ 1 mm (%)	Samples with Euclidean errors ≤ 3 mm (%)	Mean Euclidean error (mm)	Samples with Euclidean errors ≤ 1 mm (%)	Samples with Euclidean errors ≤ 3 mm (%)	Mean Euclidean error (mm)
DM	2.20	22.87	5.41	2.31	21.23	5.41
LR	15.40	66.67	3.04	10.94	59.62	3.22
DT	57.77	93.20	1.13	99.66	100.00	0.14
RF	**60.19**	**94.31**	**1.06**	99.99	100.00	0.09
ET	55.45	91.67	1.20	**100.00**	**100.00**	**0.07**

Euclidean errors under 1 mm and 3 mm of 60.19% and 94.31%, respectively, and a mean Euclidean error of 1.06 mm, as can be extracted from Table 3. Similarly, for the 70%/30% data splitting approach, it is noticeable that the lowest mean Euclidean error (0.07 mm) and the highest percentages of samples with Euclidean errors under 1 mm and 3 mm (100.00% for these two performance measures) were achieved using the ET method. For the sake of clarity, the mentioned performance assessment results using the best regression models for both data splitting approaches are highlighted in bold in the table.

Even though the evaluation of ML models in terms of computational cost was beyond the scope of this research work, it is also worth discussing this issue in general terms, without getting into detail. To this end, once performed the training process of the ML models off-line, it was measured the computational time required by the RF regression model to predict the three displacement coordinates of each liver mesh node (i.e. dx, dy and dz) for the liver used for test. Specifically, the experiment consisted in predicting the ten liver deformed states for the biomechanical behaviour modelled using combination 1 of elastic constants (as named in Table 1). In addition, the computational time necessary to perform exactly the same simulation with FEM was also measured using the same computer, thus making it possible to generally compare both modelling approaches in terms of their speed and provide a general idea of how much faster ML models were compared with FEM-generated models. The computer employed in this work to perform the simulations was based on a 3.4 GHz Intel Core i7 processor with 8 GB RAM and OS X El Capitan (version 10.11.6).

The measured computational times were 2.89 s and 51.63 s for the ML and the FEM models, respectively. Therefore, the ML model was approximately 18 times faster than the FEM-based model when performing exactly the same simulation. These results suggested that the computational cost associated to FEM-generated models was much higher than that for ML models, as expected from previous related research works. Furthermore, an aspect that should also be commented is that the most important time for a medical application would be the computational time required to predict only one liver deformed state, instead of ten states, thus resulting in a computational time 10 times lower than that reported in this research of 2.89 s for the ML model. Therefore, the value of around 0.3 s for the computational time to predict one liver deformed state with the ML model would be much closer to fulfilling the real-time requirements of actual CAS systems used in medicine, which work at a high refresh rate. In this sense, in the near future, the performance of the proposed ML-based modelling framework should be thoroughly evaluated in terms of computational cost before its integration in commercial CAS systems.

3.2.2. Results for the best models

This section discusses more thoroughly the performance assessment results obtained using the above-commented regression models leading to the best prediction performance for both data splitting approaches. Even though the measures based on the Euclidean errors were of more importance in this work, it is also worth commenting on the results obtained from evaluating the model performance in the prediction of each displacement coordinate separately. In this regard, Table 4 shows the RMSE values for the three displacement coordinates using the best regression models for both data splitting approaches. As expected, the RMSE values for the 70%/30% data splitting approach were notably lower than those for the data splitting approach leaving one liver for test. Furthermore, it can be observed that, for both data splitting approaches, the RMSE values for the displacement coordinates dx and dz were very close to each other and, in turn, higher than the value for the coordinate dy. This may be explained due to the fact that the displacement of the liver mesh nodes in the whole labelled data set was considerably lower in the y-axis direction than in that of the other two axes (in absolute values), with the target output dx ranging from -22.37 mm to 22.92 mm, dy from -9.99 mm to 8.11 mm and dz from -31.64 mm to 9.24 mm. However, despite the difference in the RMSE values between the coordinates dx and dz and the coordinate dy, the values were of the same order of magnitude for the three displacement coordinates.

Table 4: RMSE values for the three displacement coordinates using the best regression models for both data splitting approaches.

Data splitting approach	dx (mm)	dy (mm)	dz (mm)
One liver for test	0.93	0.66	0.92
70%/30%	0.07	0.04	0.07

For further discussion, Figure 7 shows the boxplots of the coordinate errors for the three displacement coordinates using the best regression models for both data splitting approaches, thus providing information about the error distribution. An aspect that should be commented is that the outliers are not shown in the boxplots in Figure 7 and neither in the remaining boxplots depicted in this work (Figure 8). This is due to the fact that, even though the number of outliers was almost negligible compared with the enormous amount of samples in the labelled data set, the presence of outliers makes it difficult to see the boxplots properly, thus complicating the extraction of meaningful information from the boxplots. Therefore, in general terms, the removal of outliers from the boxplots does not affect the discussion, since the boxplots are shown only for commenting on general trends, without getting into detail. In the case of the 70%/30% data splitting approach, it can be noticed that the error distribution was symmetric for the three displacement coordinates, since the median, represented by the red line in the middle of the box, was equidistant from the minimum and the maximum error values. In addition, taking into account that the coordinate error was defined as the difference between the estimated output and the actual output, the fact that the error distribution was balanced around zero for the three coordinates (i.e. the median had a value of around zero) indicated that the regression models did not tend to overestimate nor underestimate the outputs, since the coordinate error was positive for half of the samples (i.e. overestimated predictions), whereas the error was negative for the other half (i.e. underestimated predictions). Another aspect that can be observed is that the boxplot for the displacement coordinate *dy* was smaller than those for the other two coordinates, thus indicating that there was less variation in the errors for the coordinate *dy* and, therefore, these errors were smaller, this being in agreement with the RMSE values shown in Table 4. For the data splitting approach leaving one liver for test, the corresponding boxplots presented some general similarities to those for the 70%/30% data splitting approach, such as the fact that, in general terms, the error distribution was approximately symmetric and balanced around zero for the three coordinates. However, even though the boxplot for the displacement coordinate *dx* was almost symmetric and the median had a value of around zero, it is noticeable that the boxplots for the displacement coordinates *dy* and *dz* were mildly skewed and, in addition, the corresponding medians were not so close to zero, but slightly above zero. This suggested that the models used to predict these two displacement coordinates tended to slightly overestimate the outputs, since the coordinate error was positive for more than half of the samples.

In order to continue the discussion about the performance assessment results, the boxplots of the Euclidean errors using the best regression models for both data splitting approaches are shown in Figure 8 (outliers not shown). In addition, for convenience when discussing the boxplots, Table 5 shows the three global performance measures based on the Euclidean errors using the best models for both data splitting approaches, directly extracted from Table 3. When observing the boxplots in Figure 8, it is clear that the distribution of the Euclidean errors was positively skewed for both data splitting approaches, since

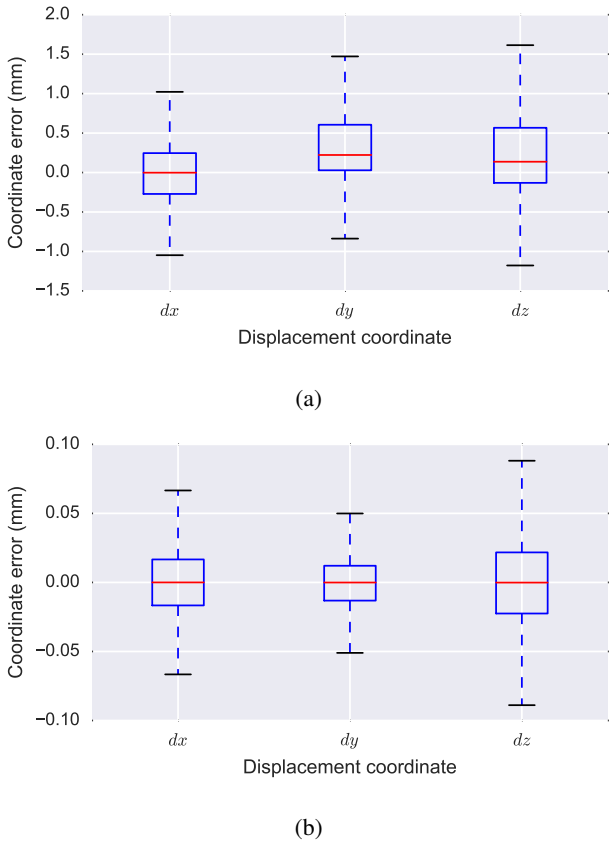


Figure 7: Boxplots of the coordinate errors for the three displacement coordinates using the best regression models for the one liver for test (a) and the 70%/30% (b) data splitting approaches. Outliers are not shown.

the distance from the median to the third quartile was greater than the distance from the median to the first quartile. As a direct consequence of the positive skew of the error distributions, the mean Euclidean error was expected to be higher than the median for both data splitting approaches. In fact, as can be straightforwardly seen from the boxplots, the mean Euclidean errors of 1.06 mm and 0.07 mm for the one liver for test and the 70%/30% data splitting approaches, respectively, were clearly higher than the corresponding medians, thus indicating that the Euclidean error associated with most of the samples was lower than the corresponding mean Euclidean error for both data splitting approaches. In addition, another aspect that can be observed from the boxplots is that, for the data splitting approach leaving one liver for test, the Euclidean error was lower than the upper limit of 3 mm for most of the samples, whereas the error was lower than the limit of 1 mm for more than half of the samples, as expected from the computed percentages of samples with Euclidean errors under 1 mm and 3 mm of 60.19% and 94.31%, respectively. Similarly, for the 70%/30% data splitting approach, it is easy to check that the Euclidean error was lower than the upper limits of 1 mm and 3 mm for all the samples, thus agreeing with the computed percentages of samples with errors under 1 mm and 3 mm of 100.00% for both upper limits.

The last step in this discussion about the performance results for the best models is to check that the samples with a displace-

Table 5: Performance assessment results using the best regression models for both data splitting approaches, including the three global performance metrics based on the Euclidean errors.

Data splitting approach	Samples with Euclidean errors ≤ 1 mm (%)	Samples with Euclidean errors ≤ 3 mm (%)	Mean Euclidean error (mm)
One liver for test	60.19	94.31	1.06
70%/30%	100.00	100.00	0.07

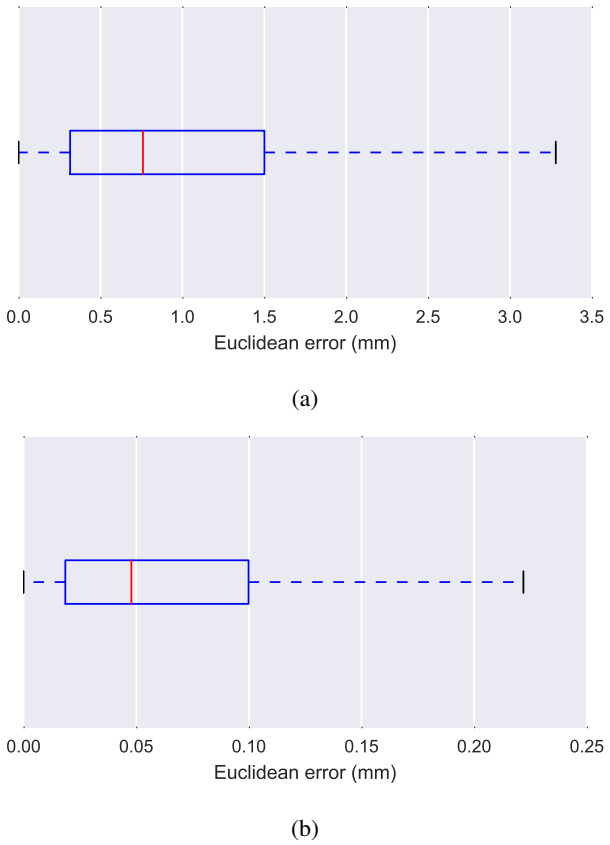


Figure 8: Boxplots of the Euclidean errors using the best regression models for the one liver for test (a) and the 70%/30% (b) data splitting approaches. Outliers are not shown.

ment (i.e. their associated values of dx, dy and dz) very different from zero were predicted with low Euclidean errors and, therefore, the good prediction performance was not only due to the great number of samples in the labelled data set with an almost negligible displacement and, consequently, an expected really low Euclidean error. In this regard, Table 6 shows the performance assessment results for the ten different external displacements applied to the liver (i.e. input variable $disp$) using the best regression models for both data splitting approaches, including the three global performance metrics based on the Euclidean errors. In the case of the 70%/30% data splitting approach, it can be observed that, while the percentages of samples with Euclidean errors under 1 mm and 3 mm remained constant at 100.00% for the different displacement applied to the liver, the mean Euclidean error gradually increased as the value of the input variable $disp$ raised, from 0.01 mm to 0.13 mm. For the data splitting approach leaving one liver for test, the mean Euclidean error increased more steeply with the input variable $disp$, from 0.19 mm to 1.95 mm. Conversely, the percentages of samples with Euclidean errors under 1 mm and 3 mm decreased when increasing the value of $disp$, diminishing from 100.00% to 26.08% and from 100.00% to 80.06%, respectively. The steeper decrease in the percentage of samples with Euclidean errors under 1 mm was expected, since the limit of 1 mm was a much stricter constraint than the limit of 3 mm. For both data splitting approaches, it should be highlighted that, although the mean Euclidean error increased with the displacement applied to the liver $disp$, the mean Euclidean error was quite low even for the maximum value of $disp$ of 15 mm, especially for the 70%/30% data splitting approach, with values of 1.95 mm and 0.13 mm for the one liver for test and the 70%/30% data splitting approaches, respectively. Therefore, it can be said that, for both data splitting approaches, the prediction performance was also good for the samples with a displacement not very close to zero, assuming that the values of dx, dy and dz were directly related to the value of $disp$.

In order to better illustrate and clarify this issue, Figure 9 shows a particular cut of the liver used for test in the frontal plane (i.e. plane x-z) when applying an external displacement of 15 mm, and for the biomechanical behaviour modelled using combination 1 of elastic material constants (as named in Table 1). The colourbar in the figure indicates the value of the actual displacement vector length for each sample. In addition, magenta circles represent those liver nodes with a Euclidean error lower or equal to 3 mm when predicting the displacement using the best model for the data splitting approach consisting in leaving one liver for test; whereas black circles represent the nodes used in the boundary conditions in FEM and, therefore, not included in the labelled data set. Figure 9 clearly shows that a significant part of the samples were predicted with Euclidean errors under 3 mm, thus agreeing with the percentage presented in Table 5 (i.e. 80.06% of the samples had a Euclidean error under 3 mm for the maximum value of $disp$ of 15 mm).

Within this framework, Figures 10 and 11 show the performance results with samples ordered according to the actual displacement vector length using the best regression models for both data splitting approaches. In particular, the estimated and the actual displacement vector lengths for all the samples, the Euclidean error for all the samples and the relative Euclidean error for all the samples are shown, respectively, from top to bottom. It should be commented that, although the samples with a displacement vector length very close to zero led to relative Euclidean errors very much higher than 100%, Figures 10 and 11 only show up to a relative Euclidean error of 100% for the sake of clarity. A quick glance at the figures shows that much better prediction performance was achieved for the 70%/30% data splitting approach compared to the data splitting approach leaving one liver for test, with the estimated displacement vector length being much closer to the actual one and, therefore, the Euclidean error and the relative Euclidean error being much lower, thus agreeing with all the above-commented results. For both data splitting approaches, it can be observed that, in fact,

Table 6: Performance assessment results for the different external displacements applied to the liver using the best regression models for both data splitting approaches, including the three global performance metrics based on the Euclidean errors.

	One liver for test			70%/30%		
disp (mm)	Samples with Euclidean errors $\leq$ 1 mm (%)	Samples with Euclidean errors $\leq$ 3 mm (%)	Mean Euclidean error (mm)	Samples with Euclidean errors $\leq$ 1 mm (%)	Samples with Euclidean errors $\leq$ 3 mm (%)	Mean Euclidean error (mm)
1.5	100.00	100.00	0.19	100.00	100.00	0.01
3.0	96.46	100.00	0.38	100.00	100.00	0.03
4.5	84.60	100.00	0.57	100.00	100.00	0.04
6.0	72.84	99.86	0.76	100.00	100.00	0.06
7.5	61.16	98.74	0.95	100.00	100.00	0.07
9.0	50.56	97.00	1.15	100.00	100.00	0.08
10.5	42.98	93.48	1.35	100.00	100.00	0.09
12.0	36.44	89.17	1.56	100.00	100.00	0.10
13.5	30.78	84.80	1.75	100.00	100.00	0.11
15.0	26.08	80.06	1.95	100.00	100.00	0.13

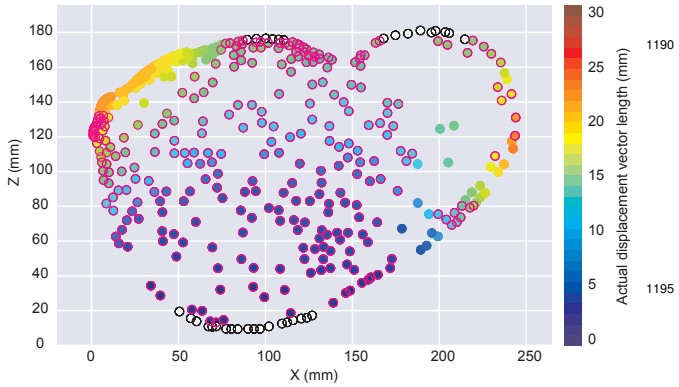


Figure 9: Cut of the liver used for test in the frontal plane (i.e. plane *x*-*z*) when applying an external displacement of 15 mm, and for the biomechanical behaviour modelled using combination 1 of elastic material constants. The colourbar indicates the value of the actual displacement vector length for each sample. The liver nodes with a Euclidean error lower or equal to 3 mm are circled in magenta and the nodes used in the boundary conditions in FEM are circled in black.

most of the samples with a displacement vector length close to zero led to relative Euclidean errors very much higher than 100%. In addition, for the data splitting approach leaving one liver for test, a lot of samples with a displacement vector length not very close to zero had huge relative Euclidean errors due to the generally much poorer prediction performance achieved for this data splitting approach. However, it is noticeable that, despite the huge relative Euclidean errors for samples with a small displacement, the relative Euclidean error tended to decrease with the value of the displacement vector length, this fact being specially significant for the 70%/30% data splitting approach. In this regard, an aspect that should be highlighted is that, for the 70%/30% data splitting approach, almost all the samples with a displacement vector length higher than 1 mm were predicted with relative Euclidean errors much lower than 20%, as can be seen from Figure 11. Therefore, these results also confirmed that the best regression model for the 70%/30% data splitting approach led to particularly good prediction performance over a wide range of displacement vector lengths, with a low relative Euclidean error for all the samples, with the exception of those with a displacement much lower than 1 mm.

4. Conclusions

The potential of ML methods has been proved for modelling the biomechanical behaviour of the human liver during the breathing process. In particular, this research has proposed a ML-based modelling framework capable of predicting the liver deformation (i.e. the displacement of each liver mesh node in the 3D space: *dx*, *dy* and *dz*) when exerting different external displacements that recreated the liver compression during breathing, by means of training the ML models with data previously collected from the FEM-based simulation of this behaviour. Furthermore, unlike other related studies, this modelling scheme was not only able to predict the liver deformation when applying a new external displacement, but also for a new liver. This was due to the use of several livers during the training process of the ML models.

Once the different ML regression models were built using the optimised hyperparameters and the training sets associated with the two data splitting approaches, the test sets were used to evaluate the performance of the models in the prediction of the displacement of the liver mesh nodes. The comparison of the performance assessment results revealed that the three tree-based methods (i.e. the DT, the RF and the ET techniques) clearly outperformed the DM and the LR models for both data splitting approaches, thus confirming that the problem in this work had non-linear nature. In addition, in general, the use of the 70%/30% data splitting approach resulted in much better performance results than using the data splitting approach leaving one liver for test, since the 70%/30% data splitting maximised the variety of liver geometries in the training set employed to build the models. In fact, the best prediction performance results for the data splitting approach leaving one liver for test were obtained using the RF regression method, with percentages of samples with Euclidean errors under 1 mm and 3 mm

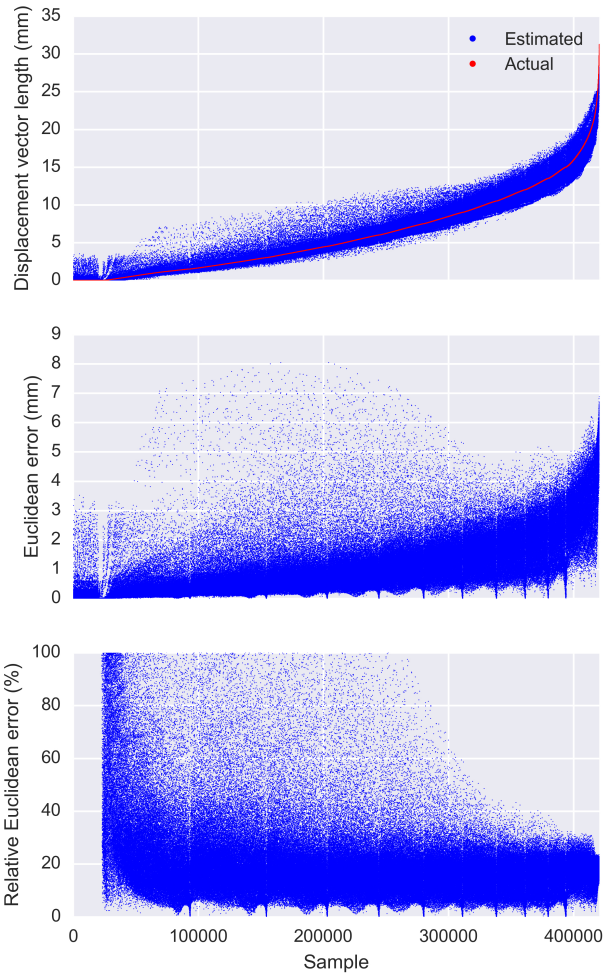


Figure 10: Performance assessment results with samples ordered according to the actual displacement vector length using the best model for the data splitting approach leaving one liver for test, including the estimated and the actual displacement vector lengths for all the samples (*top*), the Euclidean error for all the samples (*middle*) and the relative Euclidean error for all the samples (*bottom*).

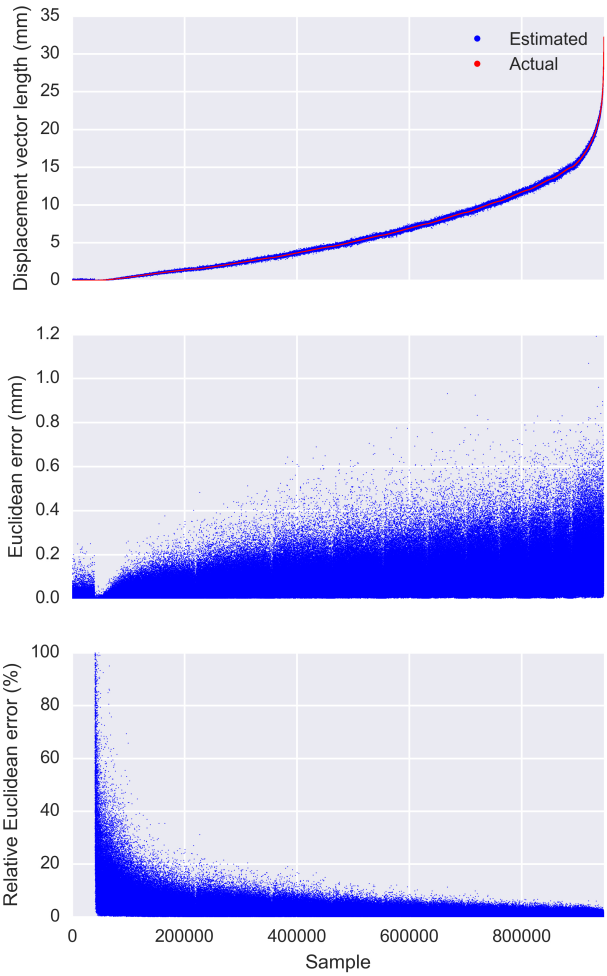


Figure 11: Performance assessment results with samples ordered according to the actual displacement vector length using the best model for the 70%/30% data splitting approach, including the estimated and the actual displacement vector lengths for all the samples (*top*), the Euclidean error for all the samples (*middle*) and the relative Euclidean error for all the samples (*bottom*).

of 60.19% and 94.31%, respectively, and a mean Euclidean error of 1.06 mm. In the case of the 70%/30% data splitting approach, the lowest mean Euclidean error (0.07 mm) and the highest percentages of samples with Euclidean errors under 1 mm and 3 mm (100.00% for these two performance measures) were achieved using the ET method.

It should be particularly emphasised that all the samples were predicted with a Euclidean error lower or equal to 1 mm using the best model for the 70%/30% data splitting approach, thus surpassing, to a large extent, the precision results of about 3-5 mm achieved in other similar research studies regarding the modelling of human organ deformations. Anyway, even though the best results were obtained for the 70%/30% data splitting, the high percentage of samples with Euclidean errors under 3 mm of 94.31% and the mean Euclidean error of 1.06 mm achieved using the best model for the data splitting approach leaving one liver for test could be regarded as perfectly comparable with the results of about 3-5 mm obtained in the other related works. In addition, the results achieved using the best model for the one liver for test data splitting could be considered more than acceptable according to the upper precision limit of 3 mm suggested for avoiding clinically-relevant dosimetric consequences in radiotherapy, and they were also well below the accuracy limit of 5 mm acceptable when using medical imaging on human organs for early cancer diagnosis.

A last remark that is important to comment is that the ML models were constructed using data from FEM-based simulations of the liver deformation during breathing due to the considerable difficulty in obtaining these data directly from real patients. Nowadays, it is really difficult to get liver deformation data from real patients mainly because taking unnecessary CT scans during the breathing process is out of the clinical protocol, since the scanning process implies a lot of radiation for the patients and it is time and resource consuming. However, if acquiring data from tracking the position of some specific points of *in-vivo* livers was possible, ML models could be constructed using only these real data without the need to perform FEM-based simulations of the breathing process off-line. The resultant ML models built from real data would predict the liver deformation during breathing as fast as the ML models constructed from FEM-generated data. As a direct consequence of not requiring FEM-based simulations, an explicit biomechanical model of the liver (e.g. the Ogden model used in this work) would not be necessary to perform predictions about its behaviour. This is an important issue to highlight, since, to date, an explicit model of the human liver able to characterise its mechanical behaviour *in vivo* in a process such as breathing is still a challenge in soft tissue Biomechanics. Therefore, if, as the Biomechanics community hopes in the near future, the acquisition of data from the liver of real patients was feasible, the ML models would provide a powerful modelling framework capable of outperforming the conventional FEM in terms of speed with the additional advantage of not requiring an explicit biomechanical model of the organ, not even for obtaining the training data, thus enabling the real-time simulation of the deformation of *in-vivo* human livers of real patients during the breathing process.

5. Future research work

Further research on some aspects is still required before the clinical application of the presented ML-based modelling scheme. In this sense, a natural way to continue the research presented in this paper would be to use many different livers during the training process of the ML models in order to properly predict the displacement of an unknown liver within a wide range of geometries and biomechanical properties. Thus, the data splitting approach leaving one liver for test, which is more similar to what would happen in a real application, would be expected to lead to much better prediction performance compared to that achieved in the present research. In fact, a new research work is currently being planned to be conducted using many more livers compared to those used in this work. Therefore, it may be possible to check that the results for the two data splitting approaches used in the present work are similar and, in turn, optimal from a clinical point of view when employing many more livers to construct the ML models, which would be in line with current expectations.

A further step in the framework of this research line would be to obtain the elastic properties of the livers, as well as the liver meshes, from CT images of real patients *in vivo*. In fact, the researchers in the present work are currently working in this regard in order to estimate the elastic material constants of soft tissues of each particular patient *in vivo* by means of image analysis. In addition, the boundary conditions used when generating the training data with FEM should reproduce more faithfully the real process of breathing. For instance, this could be done by performing two CT scans of each liver (in complete exhalation and in complete inhalation) and then obtaining the real displacement of the nodes of the liver surface by a registration algorithm.

Once ML models are trained off-line, ML models are expected to be able to perform an accurate simulation of liver deformation during breathing in real-time, thus tracking the exact location of the liver at every moment, since these models generally allow a really fast processing. However, the performance of the proposed ML-based modelling framework should be thoroughly evaluated in terms of computational cost before its integration in commercial CAS systems. Thus, it would be possible to properly check if ML models can fulfil the real-time requirements of actual CAS systems used in medicine, which work at a high refresh rate. In this regard, even though it is common knowledge from previous related research works that the computational cost associated to FEM-generated models is much higher for simulating complex biomechanical behaviours than that for ML models, an exhaustive comparison of the results from both models in terms of their computational time should also be carried out.

Compliance with ethical standards

The research work reported in this paper was approved by the ethics committees of the hospital involved, the *Universitat Politècnica de València* and the *Universitat de València*, and all procedures followed were performed in accordance with the

ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5).

Acknowledgements

This work has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO) through research projects TIN2014-52033-R and DPI2013-40859-R with the support of European FEDER funds. The authors acknowledge the kind collaboration of the personnel from the hospital involved in the research.

Abdelrahman, W., Farag, S., Nahavandi, S., & Creighton, D. (2011). A comparative study of supervised learning techniques for data-driven haptic simulation. In *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2842–2846).

Appice, A., & Džeroski, S. (2007). Stepwise induction of multi-target model trees. In J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Proceedings of the18th European Conference on Machine Learning, ECML 2007* (pp. 502–509). Berlin/Heidelberg, Germany: Springer-Verlag.

Balter, J. M., Dawson, L. A., Kazanjian, S., McGinn, C., Brock, K. K., Lawrence, T., & Haken, R. T. (2001). Determination of ventilatory liver movement via radiographic evaluation of diaphragm position. *International Journal of Radiation Oncology*Biology*Physics*, *51*, 267–270.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. (1st ed.). New York, NY, USA: Springer-Verlag.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, California, USA: Wadsworth Publishing Company.

Brock, K. K., Dawson, L. A., Sharpe, M. B., Moseley, D. J., & Jaffray, D. A. (2006). Feasibility of a novel deformable image registration technique to facilitate classification, targeting, and monitoring of tumor and normal tissue. *International Journal of Radiation Oncology*Biology*Physics*, *64*, 1245–1254.

Brock, K. K., Hawkins, M., Eccles, C., Moseley, J. L., Moseley, D. J., Jaffray, D. A., & Dawson, L. A. (2008). Improving image-guided target localization through deformable registration. *Acta Oncologica*, *47*, 1279–1285.

Courtecuisse, H., Jung, H., Allard, J., Duriez, C., Lee, D. Y., & Cotin, S. (2010). GPU-based real-time soft tissue deformation with cutting and haptic feedback. *Progress in Biophysics and Molecular Biology*, *103*, 159–168.

De, S., Deo, D., Sankaranarayanan, G., & Arikatla, V. S. (2011). A physics-driven neural networks-based simulation system (PhyNNeSS) for multimodal interactive virtual environments involving nonlinear deformable objects. *Presence*, *20*, 289–308.

Delingette, H., Subsol, G., Cotin, S., & Pignon, J. (1994). Craniofacial surgery simulation testbed. In *Visualization in Biomedical Computing '94, Proceedings of SPIE* (pp. 607–618). volume 2359.

Deo, D., & De, S. (2009). PhyNeSS: A physics-driven neural networks-based surgery simulation system with force feedback. In *EuroHaptics conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, World Haptics 2009, Third Joint* (pp. 30–34).

Duysak, A., Zhang, J. J., & Ilankovan, V. (2003). Efficient modelling and simulation of soft tissue deformation using mass-spring systems. In *Proceedings of the 17th International Congress and Exhibition on Computer Assisted Radiology and Surgery, CARS 2003* (pp. 337–342). volume 1256 of *International Congress Series*.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3–42.

Göçeri, E. (2013). *A comparative evaluation for liver segmentation from spir images and a novel level set method using signed pressure force function*. PhD thesis: İzmir Institute of Technology.

Göçeri, E. (2016). Fully automated liver segmentation using Sobolev gradient-based level set evolution. *International Journal for Numerical Methods in Biomedical Engineering*, . doi:10.1002/cnm.2765.

González, D., Aguado, J. V., Cueto, E., Abisset-Chavanne, E., & Chinesta, F. (2016). kPCA-based parametric solutions within the PGD framework. *Archives of Computational Methods in Engineering*, (pp. 1–18).

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. (2nd ed.). New York, NY, USA: Springer-Verlag.

Hostettler, A., George, D., Rémond, Y., Nicolau, S. A., Soler, L., & Marescaux, J. (2010). Bulk modulus and volume variation measurement of the liver and the kidneys in vivo using abdominal kinetics during free breathing. *Computer Methods and Programs in Biomedicine*, *100*, 149–157.

Hu, T., & Desai, J. P. (2004). Modeling large deformation in soft-tissues: experimental results and analysis. In *Proceedings of EuroHaptics 2004*.

Ikonomovska, E., Gama, J., & Džeroski, S. (2011). Incremental multi-target model trees for data streams. In *Proceedings of the 2011 ACM Symposium on Applied Computing, SAC '11* (pp. 988–993). New York, NY, USA: ACM.

Inoue, Y., Masutani, Y., Ishii, K., Kumai, N., Kimura, F., & Sakuma, I. (2006). Development of surgical simulator with high-quality visualization based on finite-element method and deformable volume rendering. *Systems and Computers in Japan*, *37*, 67–76.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. (1st ed.). New York, NY, USA: Springer-Verlag.

Jahya, A., Herink, M., & Misra, S. (2013). A framework for predicting three-dimensional prostate deformation in real time. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *9*, e52–e60.

Kenedi, R., Gibson, T., Evans, J., & Barbenel, J. (1975). Tissue mechanics. *Physics in Medicine and Biology*, *20*, 163–169.

Lu, Y.-C., Kemper, A. R., Gayzik, S., Untaroiu, C. D., & Beillas, P. (2013). Statistical modeling of human liver incorporating the variations in shape, size, and material properties. *Stapp Car Crash Journal*, *57*, 285–311.

Maas, S., Ellis, B., Ateshian, G., & Weiss, J. (2012). FEBio: finite elements for biomechanics. *Journal of Biomechanical Engineering*, *134*, 011005.

Martínez-Martínez, F. (2014). Simulation of the biomechanical behavior of the human liver. In *Determining the biomechanical behavior of the liver using medical image analysis and evolutionary computation* (pp. 133–152). PhD thesis: Universitat Politècnica de València.

Martínez-Martínez, F., Lago, M. A., Rupérez, M. J., & Monserrat, C. (2013a). Analysis of several biomechanical models for the simulation of lamb liver behaviour using similarity coefficients from medical image. *Computer Methods in Biomechanics and Biomedical Engineering*, *16*, 747–757.

Martínez-Martínez, F., Rupérez, M. J., Martín-Guerrero, J. D., Monserrat, C., Lago, M. A., Pareja, E., Brugger, S., & López-Andújar, R. (2013b). Estimation of the elastic parameters of human liver biomechanical models by means of medical images and evolutionary computation. *Computer Methods and Programs in Biomedicine*, *111*, 537–549.

Meier, U., López, O., Monserrat, C., Juan, M. C., & Alcañiz, M. (2005). Real-time deformable models for surgery simulation: a survey. *Computer Methods and Programs in Biomedicine*, *77*, 183–197.

Morooka, K., Chen, X., Kurazume, R., Uchida, S., Hara, K., Iwashita, Y., & Hashizume, M. (2008). Real-time nonlinear FEM with neural network for simulating soft organ model deformation. In D. Metaxas, L. Axel, G. Fichtinger, & G. Székely (Eds.), *Proceedings of the 11th International Conference on Medical image computing and computer-assisted intervention, MICCAI 2008, Part II* (pp. 742–749). Berlin/Heidelberg, Germany: Springer-Verlag. volume 5242 of *Lecture Notes in Computer Science*.

Niroomandi, S., Alfaro, I., Cueto, E., & Chinesta, F. (2008). Real-time deformable models of non-linear tissues by model reduction techniques. *Computer Methods and Programs in Biomedicine*, *91*, 223–231.

Niroomandi, S., González, D., Alfaro, I., Bordeu, F., Leygue, A., Cueto, E., & Chinesta, F. (2013). Real-time simulation of biological soft tissues: a PGD approach. *International Journal for Numerical Methods in Biomedical Engineering*, *29*, 586–600.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Re-*

search, *12*, 2825–2830.

Ruiter, N. V., Stotzka, R., Müller, T.-O., Gemmeke, H., Reichenbach, J. R., & Kaiser, W. A. (2006). Model-based registration of X-ray mammograms and MR images of the female breast. *IEEE Transactions on Nuclear Science*, *53*, 204–211.

Struyf, J., & Džeroski, S. (2006). Constraint based induction of multi-objective regression trees. In F. Bonchi, & J.-F. Boulicaut (Eds.), *Knowledge Discovery in Inductive Databases: 4th International Workshop, KDID 2005* (pp. 222–233). Berlin/Heidelberg, Germany: Springer-Verlag.

Székely, G., Brechbühler, C. H., Hutter, R., Rhomberg, A., Ironmonger, N., & Schmid, P. (2000). Modelling of soft tissue deformation for laparoscopic surgery simulation. *Medical Image Analysis*, *4*, 57–66.

Waters, K. (1992). Physical model of facial tissue and muscle articulation derived from computer tomography data. In *Visualization in Biomedical Computing '92, Proceedings of SPIE* (pp. 574–583). volume 1808.

Zhong, Y., Shirinzadeh, B., Alici, G., & Smith, J. (2006). Cellular neural network based deformation simulation with haptic force feedback. In *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control, AMC '06* (pp. 380–385).

Zienkiewicz, O. C., & Taylor, R. L. (1989). *The finite element method: basic formulation and linear problems, Volume 1*. (4th ed.). London, UK: McGraw-Hill.