Computational Models for Estimating Liver Iron Overload with the Magnetic Iron Detector

Barbara Gianesin
Istituto Nazionale di Fisica Nucleare
(INFN)
Genova, Italy
Email: gianesin@ge.infn.it

Luca Baldassarre
Computer Science Department
Università di Genova
Genova, Italy
Email: baldassarre@disi.unige.it

Abstract—An accurate measurement of the liver iron overload is essential for the management of diseases such as thalassemia and hemocromathosis. The Magnetic Iron Detector is a susceptometer, which measures the total iron overload in the liver and has been used on more than 800 patients of Galliera Hospital (Genoa, Italy) since February 2005. The iron overload is obtained by calculating the difference between the measured magnetization signal and the patient's background signal, which is the magnetization signal that would be measured for that patient with a normal iron content. This study describes two models for calculating the background signal using the measurements and the anthropometric features of 84 healthy volunteers. The first model introduces a statistical correction to the signal computed from the body shape of the subject assuming it to be made of water. The second model is based on statistical learning and learns from the volunteers' data a mapping from the anthropometric features to the background signal. We present two approaches to combine the models. The assessment of the models on the 84 volunteers show that the performances of the models are comparable and that we can confidently estimate the background signals of patients. The model sensitivity (0.9 g) allows the physicians to monitor the iron overload variations due to the therapy. These models are currently in use at the Galliera Hospital.

Keywords-Medical Computation; Liver Iron Overload; Magnetic Iron Detector; Thalassemia; Statistical Learning; Kernel Methods; Forward Problem.

I. INTRODUCTION

An accurate assessment of body iron accumulation is essential for the diagnosis and therapy of iron overload in diseases such as hereditary hemochromatosis, thalassemia and other forms of severe congenital or acquired anaemia. For example, in hereditary hemochromatosis, the subject adsorbs an excess of iron from the diet every day, while in thalassemia major the iron overload is caused by the frequent blood transfusions administrated to the patient to contrast its anaemia. Being toxic, the iron in excess must be removed by a tuned therapy: for this reason hemochromatosis patients are subjected to phlebotomy therapy while a chelation therapy is administrated to transfusion dependent patients.

The liver is the target organ for evaluation of the iron overload. A normal liver contains about 0.5 g of iron [1], whereas the overload can be higher than 10 g in severe iron-overload states. The liver biopsy is considered the gold standard to evaluate liver iron overload [1], [2]: this invasive measure evaluates the iron concentration in a small sample of hepatic tissue. Validated non-invasive techniques are MRI [3] and SQUID-susceptometer [4].

The Magnetic Iron Detector (MID) susceptometer [5], [6], [7], [8] quantifies the amount of iron in the liver by measuring the susceptibility of the human body. A magnetic field B, applied to the body, induces a magnetization of its tissues. Because the iron deposits in the living biological tissues exhibit paramagnetic behavior [7], [9], the change of the field we are interested in is very small ($\sim 10^{-6}B$ for a normal iron level). Since the MID generates a low frequency magnetic field, the measurement is performed with a pickup coil using the synchronous detection. The apparatus of the MID susceptometer is sketched in Figure 1A. Two pickup coils are assembled symmetrically with respect to the magnet in order to cancel the common mode signal induced with no patient is placed between the magnet and the lower pickup. The signal becomes non-zero when the patient is placed in the measurement area.

Since February 2005, the MID is in use at the Galliera Hospital of Genoa and more than 1300 iron overload evaluations have been performed [7]. MID obtains the iron overload by computing the difference between the measured magnetization signal of the patient and its *background*
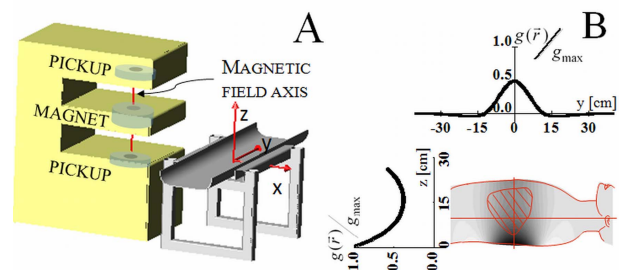

Figure 1. MID Instrumentation. (A) Two pickups are placed symmetrically with respect to the magnet: the sum of their signals is zero in the absence of a patient. (B) Weight function $g(\vec{r})$ of the MID (see Section III-A).

signal, defined as the magnetization signal that the patient would generate with a normal iron level. The evaluation of the background signal is performed under the hypothesis that the magnetization signal of a patient, without iron overload, is the same as the one of a volunteer with the same anthropometric characteristics.

Until January 2010, the background signal was calculated by a statistical model [7], that was developed exploiting the correlation coefficient between the measured magnetization signal of volunteers (i.e., subjects with a normal iron level) and the anthropometric variables. Moreover, a linear dependence of the magnetization signal from the input variables of the model was assumed. The sensitivity of this model is about 1 g, which must be compared with the reproducibility of the iron overload measurement of the same patients (less than 0.5 g) and with the iron contents of a healthy liver (about 0.5 g).

Two models have been developed in order to improve the MID sensitivity. Each one were the object of two PhD theses [8], [10]. The first model was developed following [11], in which the direct calculation of the magnetization signal was proposed assuming for the body a uniform magnetic susceptibility distribution equal to that of water. The second model is based on statistical learning theory [10], [12].

This paper is devoted to the description and the comparison between the two models. Moreover, we propose two different methods for combining their results. The performances of the models, evaluated on a common set of 84 healthy volunteers, are quite equivalent, each one introducing a benefit in the background signal calculation. From January 2010 both this models are being used at the Galliera Hospital (Genoa, Italy) for assessing the iron overload with MID.

The paper is organized as follows: in Section II we describe the measurement of the magnetic signal and the anthropometric features recorded for each subject. In Section III we first detail the two models developed to estimate the background signal and then two approaches adopted to combine their predictions; secondly we describe the experimental protocol adopted to compare them. In Section IV we present the experimental results and conclude the paper with the discussion in Section V.

II. THE MEASUREMENTS

Here we describe the measurement of the magnetization signal and the anthropometric features recorded for each subject.

A. The Magnetization Signal

During the measurement, the patient is placed inside the measurement area and the patient's position along the stretcher is such that the Y axis, positively oriented towards the head, lays along the longitudinal symmetry axis of the body (Figure 1A). The center of the patient's trunk falls
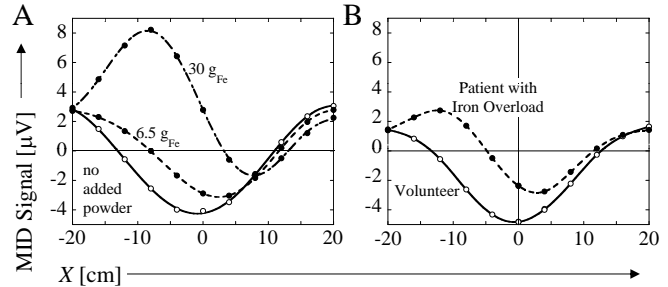

Figure 2.   The iron-overload contribution to the magnetic signal. The abscissa X (cm) is the position of the magnetic field axis relative to the center of the trunk. (A) Anthropomorphic plastic phantom with different doses of paramagnetic powder. (B) Magnetization signal of a patient with liver iron overload and of a volunteer with similar anthropometric data.

on the X axis origin and the abscissa of the liver center of mass is negative. As the stretcher moves along the rails, the vertical axis of the magnetic field slides along the X axis, allowing us to scan the whole liver region. Figure 2A reports the magnetization signals of an anthropomorphic plastic phantom dosed with paramagnetic powder, equivalent to 6.5 g and 30 g of iron. Figure 2B shows the comparison between the signal from a patient with an iron overload of about 9 g in the liver and that of a healthy volunteer with similar anthropometric data.

The contribution to the magnetic susceptibility of the iron overload is obtained by calculating the difference between the magnetization signal and the background signal attributed to the patient on the basis of their anthropometric characteristics. This difference is maximum when the magnetic field axis crosses the liver center of mass. The iron overload in grams is obtained by dividing this signal by the contribution to the signal of 1 g of iron uniformly distributed in the liver of the subject [7].

B. Anthropometric Features

The anthropometric features measured for each subject are: age, body weight and height, body mass index, body surface area, torso cross section areas at the level of the liver, the shoulders and the hips respectively and torso mean thicknesses at the level of the liver, the shoulders and the hips respectively. The last 6 features are calculated from the *3D-shape* of the body which is measured with a system of 6 lasers located on the ceiling. An example of acquisition is reported at the top of Figure 3. Note that the system is not able to detect the empty regions under the body.

The liver iron overload is common to several diseases [7]. As a consequence, no special requirements had been imposed on volunteers enrolling. We applied the Student's t-tests for comparing the means of the anthropometric features of the volunteers and patients populations. We obtained that, with a confidence level of 0.01, the means are the same, with the exception of the age and the torso mean thicknesses.

III. COMPUTATIONAL MODELS FOR ESTIMATING THE BACKGROUND SIGNAL

In this section we first present the two models developed to compute the background signal. Secondly, we propose two approaches to combine their predictions. Finally, we discuss the experimental protocol adopted to compare their performances.

A. Hybrid Waterman Model

Knowing the geometry of the body and the distribution of its magnetic susceptibility, $\chi(\vec{r})$, the magnetization signal of the body, at position $\vec{X} = (X, 0, 0)$, is obtained by solving the forward problem [13]

$$\phi(\vec{X}) = \int_V g(\vec{r} + \vec{X})\,\chi(\vec{r})\,d\vec{r}, \qquad (1)$$

where V is the volume of the body and $g(\vec{r})$ is the weight function (Figure 1B) that gives the contribution to the magnetization signal of a unitary volume of matter with unitary magnetic susceptibility. The function $g(\vec{r})$ was obtained measuring the magnetization signal of a ferromagnetic probe. The quality of this weight function was verified [8] with cylindrical samples of water whose magnetization signals were measured by MID and calculated via (1).

The signal calculated assuming a uniform distribution of susceptibility, equal to that of water ($-9 \cdot 10^{-6}$, SI units), is called the *waterman* [11]. Figure 3A reports the comparison between the measured magnetization signal and the *waterman* signal for an average-built volunteer. Note that in the position around $X = 0$ cm, the measured signal is larger in absolute value than the *waterman*. The opposite observation can be made for the border positions. This is true quite in general as it is shown in Figure 3C, which reports the spatial dependence of the differences between the measured signal and the *waterman* for each volunteer.

The most likely explanation for this discrepancy is that the *waterman* is computed on a volume bigger than the actual body volume, since the *3D-shape* includes the empty regions under it. It was verified that a volume equivalent to that of these regions and filled with water generates a signal comparable to the error of the *waterman* [8]. In addition to this contribution, the error also depends on the difference between the magnetic susceptibility of the water and that of the body tissues. As a first approximation, we used the mean of the errors to correct the *waterman* signal and named this method the *hybrid waterman* model. To compute the background signal, first the *waterman* is calculated from the *3D-shape*, then for each measuring position, the mean of the errors reported in Figure 3C is added.

B. Statistical Learning Model

The statistical learning approach [14] aims to build an estimator of the background signal from a set of given input-output examples, called the *training set*. Another approach
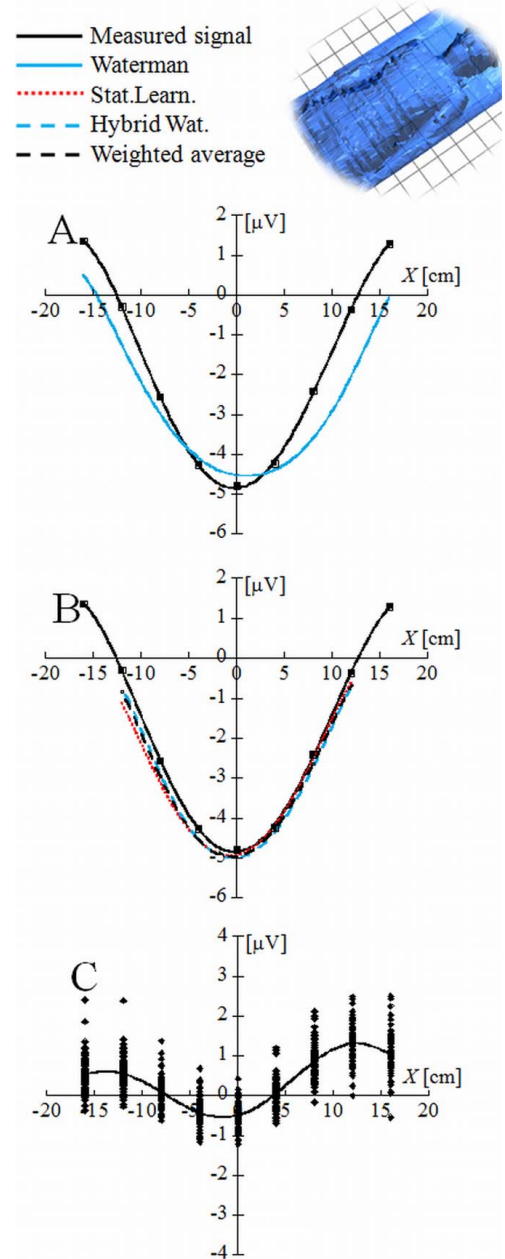

Figure 3. The measured magnetization signal compared to the *waterman* signal (A) and to the estimates of the other models (B) for a volunteer (53 kg). (C) Differences (dots), and their average (solid line), between the measured signal and the *waterman* for the 84 volunteers.

would be to directly model the probabilistic relationship between input and output, for instance with Bayesian techniques [15]. The input examples represent the anthropometric features of the volunteers and the output examples are the corresponding background signals. Since we are interested in estimating the background signal only at the fixed measuring positions, it is a vector-valued learning problem, where each output component corresponds to the measure in a specific

position. It is important that the estimator we learn has good generalization properties on new unseen data and does not predict correctly only the training examples. We begin by formulating the problem in mathematical terms, continue presenting a well-known learning method and conclude the section discussing some choices we made.

We consider a training set of input-output pairs $\{(x_i, y_i) : x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^d\}_{i=1}^n$, where $\mathbb{R}^p$ is called the *input space* (p is the number of anthropometric features) and $\mathbb{R}^d$ is the *output space* (d is the number of measuring positions). We want to estimate a function $f : \mathbb{R}^p \to \mathbb{R}^d$ able to generalize on unseen examples. An estimator, f, can be found minimizing the *empirical error*

$$\mathcal{E}_n(f) = \frac{1}{n} \sum_{i=1}^n ||y_i - f(x_i)||_d^2 , \qquad (2)$$

which is the average prediction error on the examples of the training set. We search for the estimator in a Reproducing Kernel Hilbert Space (RKHS) [16] defined by a matrix-valued kernel, $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathcal{B}(\mathbb{R}^d)$, where $\mathcal{B}(\mathbb{R}^d)$ is the space of $d \times d$ matrices. The function that minimizes the empirical error $\mathcal{E}_n$ in a RKHS can be written as

$$f(x) = \sum_{i=1}^n K(x, x_i)c_i \qquad \text{with } c_i \in \mathbb{R}^d \ \forall i = 1, \cdots, n . \qquad (3)$$

The coefficients c_i must satisfy

$$\mathbf{K}\mathbf{c} = \mathbf{y} ,$$

where $\mathbf{c} = (c_1, \ldots, c_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are nd-dimensional vectors where we concatenated the coefficients c_i and the outputs y_i, respectively. $\mathbf{K}$ is called the Gram matrix and is a $n \times n$ block matrix, where each block (i, j) is the $d \times d$ scalar matrix $K(x_i, x_j)$.

To recover the coefficients $\mathbf{c}$ we should invert the matrix $\mathbf{K}$, but $\mathbf{K}$ is not guaranteed to be invertible, nor stable under inversion. Tikhonov regularization [17], [18] solves these issues by adding a regularization term to the empirical risk, thus minimizing the following functional

$$\frac{1}{n} \sum_{i=1}^n ||y_i - f(x_i)||_d^2 + \lambda ||f||_K^2 .$$

The norm defined by the kernel K controls the smoothness of the candidate vector-valued estimator f and the relationships between its components; λ is called the *regularization parameter* that balances the trade-off between fitting the training data and choosing simpler estimators. The estimator obtained with Tikhonov regularization can still be written as in (3), but the coefficients are now given by

$$\mathbf{c} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{y} ,$$

where $\mathbf{I}$ is $nd \times nd$ identity matrix. We note that the penalty term has the effect of stabilizing the inversion of the matrix $\mathbf{K}$ by increasing all its eigenvalues by $n\lambda$.

Several interesting kernels for vector-valued functions have been recently introduced in the literature [19], [20]. Unfortunately, often these kernels require to fine tune many parameters in order to properly leverage the relationships among the output components.

After some preliminary assessment, we opted for a simple matrix-valued kernel of the form $K(x, x') = k(x, x')A$, where $k(x, x')$ is a scalar kernel and A is a positive semi-definite $d \times d$ matrix. We chose a linear kernel $k(x, x') = x \cdot x'$ that proved to produce results as accurate as the non linear gaussian kernel, which requires the tuning of the width and is more expensive to compute. The matrix A was chosen to be the d-dimensional identity matrix. This choice renders the vector-valued learning problem equivalent to solving d scalar learning problems and the relationships that exist among the output components are not exploited. The only coupling among them is the common regularization parameter λ, that imposes the same level of regularization on each output. To choose the proper value of λ we followed the procedure described in Section III-E.

C. Weighted Average Combined Model

The first combined model consists in computing the weighted average of the predictions of the statistical learning model (y_{sl}) and of the *hybrid waterman* model (y_{hw}). For each measuring position, the weighted average is computed according to the formula

$$\overline{y} = \frac{\frac{y_{sl}}{\sigma_{sl}^2} + \frac{y_{hw}}{\sigma_{hw}^2}}{\frac{1}{\sigma_{sl}^2} + \frac{1}{\sigma_{hw}^2}}, \qquad (4)$$

where σ_{sl}^2 and σ_{hw}^2 are the variances of the differences between the measurements and the predictions.

D. Learning the Error of the Waterman Model

The second combined model uses the statistical learning framework to estimate the difference between the measured signal and the *waterman*, using as input variables the anthropometric features. If there exists a relationship between the anthropometric features and the error of the *waterman* model, we should be able to learn it and use the estimates to correct the predictions of the *waterman* model. In the following, we refer to this model as *learning waterman*.

E. Assessment of the Models

In order to assess the performance of a model we need to separate the data from which the model is trained from the data on which it is evaluated. Usually these are called the training and the test sets. All model parameters must be chosen using only the training set, otherwise we obtain a biased estimate of the performance of the model. Since our data is scarce we cannot split it into two sets. Therefore

we resort to the Leave-One-Out Cross Validation (LOO-CV), which consists in holding out one example for testing and using the remaining $N-1$ for training. The procedure is repeated until all examples have been used for testing. We thus obtain the predictions for each volunteer and the corresponding errors. The distribution of these errors can be used to assess and compare different models.

For the *statistical learning* and the *learning waterman* models, we used two loops of LOO-CV. The inner loop is used to assess the estimators obtained with different values of the regularization parameter, λ, while the outer loop is used to estimate the performance of the model trained with the optimal value of λ. More precisely, one volunteer at a time is held out for testing. The model is trained on the remaining $N-1$ volunteers with a regularizing parameter λ_1. This parameter is chosen via the inner LOO-CV loop. That is, one of the $N-1$ volunteers is held out to assess the predictions error of the models trained on $N-2$ volunteers with regularizing parameters λ_j, $j = 1, \ldots, 150$. Therefore, 150 models will be trained and tested. Then another volunteer (of the $N-1$) is held out, and other 150 models are trained on the remaining $N-2$ volunteers. The prediction errors associated to the same regularizing parameter value λ_j are averaged. The parameter value associated to the smallest average error is used to train the model on all $N-1$ examples. We repeat this procedure until all volunteers are used once for the external loop, yielding N models with their respective regularization parameters and predictions.

IV. RESULTS

Figure 4 reports the distributions (as box-plot representation) of the differences between the measurements and the background signal computed with the different models for the position $X = 0$ cm. The LOO-CV procedure described in the previous section was used to obtain the predictions. Table I reports the standard deviations of the LOO-CV error distributions for positions between $X = -8$ cm and $X = +8$ cm, which is the most significant range for the background signal calculation.

We observe no significant differences between the models in all positions, but for the position X = -8 cm, for which the F-test yields $p < 0.01$ for the differences between the variances of their error distributions. However, combining the two models we do not achieve a significant increase in prediction accuracy. This indicates that most of the background signal is explained by the *3D-shape* and that the other features do not carry additional information.

Furthermore, we found a positive linear correlation (angular coefficient 0.7, R=0.77) between the errors of the *hybrid waterman* and the statistical learning model, indicating that the two models are not independent. In fact, we recall that the statistical learning model uses 6 features that are directly computed from the *3D-shape*.
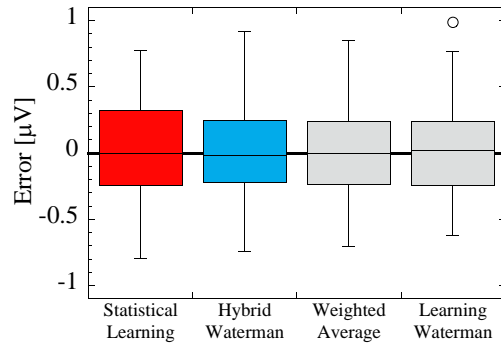

Figure 4.   Box-plots of the distributions of the errors

Table I
STANDARD DEVIATION OF THE DISTRIBUTION OF THE DIFFERENCES
BETWEEN THE MEASURED AND CALCULATED SIGNAL $[\mu V]$

Model	Measurement Position [cm]				
	−8	−4	0	+4	+8
Statistical Learning	0.51	0.41	0.36	0.42	0.50
Hybrid Waterman	0.36	0.37	0.35	0.39	0.44
Weighted Average	0.39	0.37	0.33	0.38	0.43
Learning Waterman	0.39	0.38	0.34	0.38	0.41

Since the center of mass of the patient's liver falls between $X = -8$ cm and 0 cm, the accuracy of the predictions in the positions $X = -8, -4$ and 0 cm is paramount. The average error in these position for the statistical learning model is 0.43 μV, while for the other models is 0.36 μV, which corresponds to an error of about 0.9 g of iron. This error is 20% better than the error (1.1 g) of the first model developed for calculating the background signal, that was evaluated on a dataset consisting of 142 volunteers [8].

V. CONCLUSION AND FUTURE WORK

In this paper, we presented two models for calculating the background signal of patients measured with the MID susceptometer. Since 2005 this apparatus is in use at the Galliera Hospital of Genoa and more than 1300 iron over-load evaluations have been performed. Both models have been developed on the measurements and the anthropometric features of 84 healthy volunteers. The first model, named *hybrid waterman*, calculates the magnetization signal generated by a water volume with the same external geometry of the subject (*waterman*) and corrects it by adding the mean error of the *watermen* evaluated in a population of healthy volunteers. The second model is based on statistical learning and learns from the volunteer data a mapping from the anthropometric features to the background signal. Finally, two methods to combine these models were proposed. Their performances are very similar and the combination of the two does not introduce significant accuracy gains. The evaluated model error (about 0.36 μV, equivalent to 0.9 g of iron) allows the physicians to monitor the iron overload

variations due to the therapy of patients. In order to detect mild overloaded states (between 0.5 and 1 g of iron) this error should be reduced. The limit of 0.5 g corresponds to the reproducibility error of the instrument [6], [7]. To improve the MID error, the number and distribution of healthy volunteers must be increased, while also improving the quality and the number of measured anthropometric parameters. Furthermore, we believe that classifying the volunteers with respect to their anthropometric features and developing of a different model for each category (e.g., babies, adults, oversize, etc) would reduce the error. Regarding the statistical learning model, current work is focused on developing a matrix-valued kernel that is able to exploit the correlations among the measurements in the different positions. A new generation Magnetic Iron Detector is now under construction, all the techniques presented here will be the basis for the development of new models for the estimation of the background signal of the patients that will be examined with the new susceptometer.

From January 2010 the *statistical learning*, the *hybrid waterman* and the *weighted average* models are in use at the Galliera Hospital in Genoa, Italy, for assessing the iron overload with MID.

REFERENCES

[1] A. Maggio and A. Hoffbrand, Eds., *Clinical Aspect and Theraphy of Thalassemia*. SEE-Firenze, 2004.

[2] G. Brittenham and D. Badman, "Noninvasive measurement of iron: report of an niddk workshop," *Blood*, vol. 101, pp. 15–19, 2003.

[3] T. G. St. Pierre, P. R. Clark, W. Chua-anusorn, A. Fleming, G. Jeffrey *et al.*, "Noninvasive measurement and imaging of liver iron concentrations using proton magnetic resonance," *Blood*, vol. 105, pp. 855–861, 2005.

[4] G. Brittenham, D. Farrell, J. Harris, E. Feldman, E. Danish *et al.*, "Magnetic susceptibility measurement of human iron stores," *N. Engl. J. Med.*, vol. 307, pp. 1671–1675, 1982.

[5] M. Marinelli, S. Cuneo, B. Gianesin, A. Lavagetto, M. Lamagna *et al.*, "Non-invasive measurement of iron overload in the human body," *IEEE Trans. Applied Superconductivity*, vol. 16, no. 2, 2006.

[6] M. Marinelli, B. Gianesin, M. Lamagna, A. Lavagetto, E. Oliveri *et al.*, "Whole liver iron overload measurement by a non cryogenic magnetic susceptometer," *International Congress Series 1300*, pp. 299–302, 2007.

[7] M. Marinelli, B. Gianesin, M. Balocco, P. Beruto, C. Bruzzone *et al.*, "Total iron overload measurement in the human liver region by the magnetic iron detector (mid)," *IEEE Trans. Biom. Eng.*, 2010, (to appear).

[8] B. Gianesin, "Total iron overload measurement in the human liver region by the susceptometer magnetic iron detector (mid)," Ph.D. dissertation, Dept. of Physics - University of Genoa, 2010.

[9] R. Brooks, J. Vymazal, R. Goldfarb, J. Bulte, and P. Aisen, "Relaxometry and magnetometry of ferritin," *Magnetic Resonance in Medicine*, vol. 40, no. 2, pp. 227–235, 1998.

[10] L. Baldassarre, "Multi-output learning with spectral filters," Ph.D. dissertation, Dept. of Physics - University of Genoa, 2010.

[11] INFN, DIFI, Ospedale Galliera, DISI, MedService.com Srl, and Genova Robot Srl., "Sistema integrato intelligente di supporto alla decisione per il trattamento delle patologie caratterizzate dal sovraccarico di ferro," Parco Scientifico e Tecnologico della Liguria, Report di attivita', 2008, (Pos. N. 26 Avv. 2/2006).

[12] L. Baldassarre, A. Barla, B. Gianesin, and M. Marinelli, "Vector valued regression for iron overload estimation," in *19th International Conference on Pattern Recognition*, 2008.

[13] G. Tripp, "Physical concept and mathematical models." in *Biomagnetism: an Interdisciplinary Approach: Proc. NATO Advanced Study Institute on Biomagnetism*, ser. Series A: Life Science, N. A. S. Institutes, Ed., 1983.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[15] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[16] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[17] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*. John Wiley, 1977.

[18] C. A. Micchelli and M. Pontil, "On learning vector–valued functions," *Neural Computation*, vol. 17, 2005.

[19] ——, "Kernels for multi-task learning," in *NIPS*, 2004.

[20] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying, "Universal kernels for multi-task learning," *JMLR*, vol. 9, 2008.