

# PREDICTION OF QUALITY FEATURES IN IBERIAN HAM BY APPLYING DATA MINING ON DATA FROM MRI AND COMPUTER VISION TECHNIQUES

Daniel Caballero<sup>1</sup>, Andrés Caro<sup>1</sup>, Trinidad Perez-Palacios<sup>2</sup>, Pablo G. Rodriguez<sup>1</sup>,  
Ramón Palacios<sup>3</sup>

<sup>1</sup>Departamento de Ingeniería de Sistemas Informáticos y Telemáticos, Escuela  
Politécnica, Universidad de Extremadura, Av. Universidad s/n, 10003 Cáceres, Spain

<sup>2</sup>Tecnología de los Alimentos, Facultad Veterinaria, Universidad de Extremadura, Av.  
Universidad s/n, 10003 Cáceres, Spain

<sup>3</sup>Servicio de Radiología, Hospital Universitario Infanta Cristina, Ctra. de Portugal s/n,  
06800 Badajoz, Spain

## **ABSTRACT**

*This paper aims to predict quality features of Iberian hams by using non-destructive methods of analysis and data mining. Iberian hams were analyzed by Magnetic Resonance Imaging (MRI) and Computer Vision Techniques (CVT) throughout their ripening process and physico-chemical parameters from them were also measured. The obtained data were used to create an initial database. Deductive techniques of data mining (multiple linear regression) were used to estimate new data, allowing the insertion of new records in the database. Predictive techniques of data mining were applied (multiple linear regression) on MRI-CVT data, achieving prediction equations of weight, moisture and lipid content. Finally, data from prediction equations were compared to data determined by physical-chemical analysis, obtaining high correlation coefficients in most cases. Therefore, data mining, MRI and CVT are suitable tools to estimate quality traits of Iberian hams. This would improve the control of the ham processing in a non-destructive way.*

## **KEYWORDS**

*Data mining, MRI, Active Contours, Iberian ham, Quality traits prediction*

## **1. INTRODUCTION**

Iberian dry-cured hams are highly appreciated by consumers all over the world. Quality attributes of dry-cured hams depend on the characteristics of the raw material and the processing conditions. The ripening of hams leads to physico-chemical (P-C) changes in the pieces, which influences on the quality of the final product. For example, at the end of the processing weight losses are around 30–32% , mainly due to the desiccation throughout the maturation process (45-50 % of water loss) [1].

Traditional methods for evaluating the P-C characteristics (i.e. moisture and lipid content) of dry-cured hams throughout the whole processing are tedious and time-consuming, and involve the destruction of the pieces for carrying out the analysis. The use of non-destructive techniques, such as Computed Tomography (CT), Near Infra-red Reflectance spectroscopy (NIRs) and Magnetic Resonance Imaging (MRI) have been suggested as an alternative and a complement to conventional methods. CT has been used for prediction of moisture, water activity and salt content in dry-cured hams [2,3], and NIRs has been applied to predict the fatty acid composition of ham subcutaneous fat [4].

MRI is a non-destructive, non-invasive, non-intrusive, non-radiant and innocuous technique [5]. Some parameters of Italian hams have been determined by applying MRI [6-7]. In Iberian ham, several achievements have been allowed by using this technique. The implementation of active contours in MRI allows the identification of the Biceps femoris and Semimembranosus muscles, the estimation of their volume and weight, and other parameters such as moisture [8-9]. The intramuscular fat levels of Iberian hams has also been estimated by using MRI applications [10-11], obtaining good, but not very high (around 0.50 – 0.63), correlation coefficients ( $R^2$ ) [12].

Data mining is a relevant part of a larger process known as Knowledge Discovery in Databases (KDD) [13]. It is associated with big data. Nowadays, these techniques are booming, since they can help to reveal the hidden information in datasets. This technique has been recently used to predict the quality of meat products, e.g. beef [14] and lamb [15]. However, except for the investigations carried out by our research group [16], no other studies have applied these techniques to Iberian ham.

The goal of this work was to predict weight, moisture and lipid content of Iberian hams by applying data mining on data obtained from MRI and CVT. The accuracy of the prediction was also determined.

## **2. MATERIAL & METHODS**

### **2.1. Experimental design**

This work was carried out with two batches of Iberian hams. The first batch was composed by 15 Iberian hams, which were ripened following the traditional processing, as described in [8]. Four stages were considered: raw hams, 0 days (Step 1); end of post-salting, 90 days (Step 2); end of drying, 270 days (Step 3); and dry-cured hams, 660 days (Step 4). At each stage, 6 hams were scanned to obtain MRI images. After scanning, 3 hams were destroyed for carrying out the P-C analysis. The second batch was composed by 38 Iberian hams. Stages of processing for these hams were as described previously for the first batch. MRI images and sampling for P-C analysis of the second batch of hams were acquired only at step 1 (raw hams) and 4 (dry-cured hams). Figure 1 shows the experimental design of first and second batches.



Figure 1. Sampling throughout the Iberian ham processing for the MRI acquisition and P-C analysis for both batches of hams

## 2.2. Physico-Chemical analysis

At the indicated stages for each batch, and after the MRI scanning, ham weight was recorded and then both the Biceps femoris and the Semimembranosus muscles were dissected, weighed and analyzed for measuring the moisture [17] and lipid content [18]. Analyses were carried out in triplicate.

## 2.3. MRI acquisition

Magnetic resonance sequences enable the exploration of the Biceps femoris and the Semimembranosus muscles in Iberian hams through image analysis. MRI images were obtained by using a MRI scanner (Philips Gyroscan NT Intera 1.5 T) at the “Infanta Cristina” University Hospital (Badajoz, Spain). The quadrature whole-body coil was used according to the sequences of T1 with the following parameters: 120 x 85 mm for field-of view (FOV), 20 ms for echo time (TE), 500 ms for repetition time (TR), 2 mm of thickness for slices, 90° for flip angle, i.e. a T1-weighted spin echo (SE), 0.23 x 0.20 mm for pixel resolution, and 60 as the number of slices for each ham. All the images were in DICOM format, with a 512 x 512 resolution.

## 2.4. Computer Vision Techniques

Once the images were obtained from the MRI scanner, our own computer vision algorithms were applied to extract numerical data from these images. Active contours were applied to identify the Biceps femoris and Semimembranosus muscles, as described in [9]. First, a previous image pre-processing stage was carried out. Secondly, the Biceps femoris and the Semimembranosus muscles were distinctly recognized (Figure 2) by using Active Contours [9], which apply a greedy algorithm method [8]. Active Contours are curves that can be moved due to the influence of internal and external forces. These forces are defined by an energy function. By minimizing

this energy function, the contour converges, and the solution is achieved. Volume for the ham, Biceps femoris and Semimembranosus muscles was calculated based on classical methods in analytical geometry.

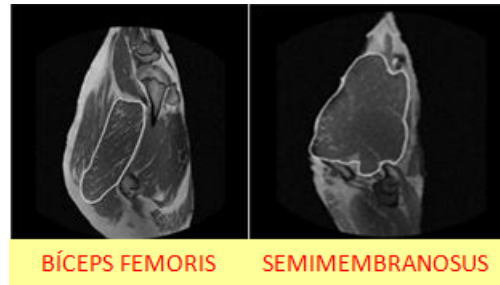


Figure 2. Biceps femoris and the Semimembranosus muscles recognition by Active Contours

## 2.5. Databases

An initial database was built with data obtained from the first batch of hams throughout its processing, including: i) stage of the ham processing, ii) data from P-C analysis (ham weight, and weight, moisture and lipid content of the Biceps femoris and the Semimembranosus muscles) and iii) data from MRI-CVT (volume of ham, Biceps femoris and Semimembranosus muscles).

As previously explained, the first batch initiates with 15 hams, three of which were destroyed to be analyzed at each stage. Thus, the number of samples at raw, end of post-salting, end of drying and dry-cured hams was 15, 12, 9 and 6, respectively. At each stage six hams were scanned to obtain MRI-CVT data. The weight of all hams was recorded, and three hams were physico-chemically analyzed. In this way, the initial database contained a total of 44 records.

Another database was built out with data obtained from the second batch of hams, containing: i) stage of the processing (raw material or dry-cured ham), ii) parameters from P-C analysis (ham weight, and weight, moisture and lipid content of the Biceps femoris and the Semimembranosus muscles), and iii) parameters from MRI-CVT (ham volume). This second database contained 40 records.

## 2.6. Data Mining

The free software WEKA (Waikato Environment for Knowledge Analysis) [19] was used for carrying out the data mining analysis. The primary groups in data mining tasks are descriptive and predictive techniques. The first ones include deductive techniques, which have the ability to infer new values based on actual data. In predictive techniques, future models can be predicted from current data by trend analysis [20,21]. Both, deductive and predictive techniques were applied in this study.

Multiple linear regression was used for the deductive tasks. The dependent variable to be estimated was always unique and numerical and this method enables the removal of collinear attributes. In addition, regression techniques seem to be the most appropriate to forecast values, as it allows inferring numerical data from the available numerical values. The M5 method of attribute selection and a ridge value of  $1 \times 10^{-4}$  were applied. This method steps through the

attributes, and removes the one with the smallest standardised coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion [22].

Again, multiple linear regression was used for the experiments of prediction. This technique obtains a linear regression equation, which can be used to predict future values [22]. The M5 method of attribute selection and a ridge value of  $1 \times 10^{-4}$  were also applied.

Figure 3 summarizes the analyses carried in this study, from ham selection to the prediction equations.

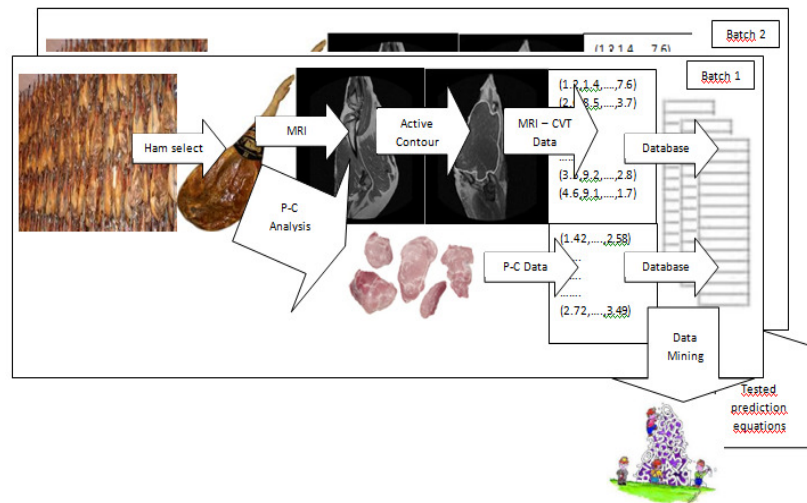


Figure 3. Summary of the analyses carried in this study, from ham selection to the prediction equations

### 3. RESULTS & DISCUSSION

Most of the records of the database from the first batch of hams were incomplete. By applying deductive techniques of data mining, multiple linear regression, the unknown information of the records in the database is estimated. Hence, a database of 44 full records is computed. This process could be seen as a type of data reconstruction: data that did not exist is reconstructed by using various algorithms with some degree of confidence. Correlation index  $R^2$  is used to prove the correctness and precision of the estimated values by using multiple linear regression. Table 1 shows the correlation coefficients between P-C and deduced data for the features analyzed: ham weight; Biceps femoris and Semimembranosus muscles weight, moisture and lipid content; ham, Biceps femoris and Semimembranosus muscles volume. As can be seen, high correlations ( $R^2 > 0.900$ ) have been obtained for all traits, except for lipid content of the Semimembranosus muscle ( $R^2 = 0.665$ ). This lower correlation could be related to the high variability of fat content in Iberian ham. Particularly noteworthy is the high correlation obtained for moisture in the two muscles ( $R^2 > 0.990$ ).

Table 1. Correlation coefficients ( $R^2$ ) between analyzed and estimated data in each features of the complete database from the first batch

|                                 | $R^2$ |
|---------------------------------|-------|
| <i>Biceps femoris</i> Weight    | 0.975 |
| <i>Biceps femoris</i> Moisture  | 0.994 |
| <i>Biceps femoris</i> Lipid     | 0.908 |
| <i>Semimembranosus</i> Weight   | 0.916 |
| <i>Semimembranosus</i> Moisture | 0.993 |
| <i>Semimembranosus</i> Lipid    | 0.665 |
| Ham Volume                      | 0.975 |
| <i>Biceps femoris</i> Volume    | 0.999 |
| <i>Semimembranosus</i> Volume   | 0.993 |

Table 2 shows the range of values of the parameters (ham weight; weight, moisture and lipid content of *Biceps femoris* and *Semimembranosus* muscle; volume of ham, *Biceps femoris* and *Semimembranosus* muscle) of the complete database from the first batch of hams at the different stages of the processing.

Table 2. Range of values for the features on the initial database. HW = Ham Weight. BW = Biceps Weight. BM= Biceps Moisture Content. BL = Biceps Lipid Content. SW = Semimembranosus Weight. SM= Semimembranosus Moisture Content. SL = Semimembranosus Lipid Content. HV = Ham Volume. BV = Biceps Volume. SV = Semimembranosus Volume

|    |        | Stage       |               |             |                |
|----|--------|-------------|---------------|-------------|----------------|
|    |        | Raw hams    | Post- Salting | Drying      | Dry-cured hams |
| HW | g      | 10600-11200 | 9000-10200    | 7800-9000   | 7600-7800      |
| BW | g      | 1235-1475   | 1100-1187     | 880-1110    | 690-756        |
| BM | g/100g | 64.97-71.95 | 59.79-62.31   | 52.27-54.63 | 40.66-45.88    |
| BL | g/100g | 7.61-9.52   | 10.09-12.90   | 9.41-13.43  | 15.21-18.38    |
| SW | g      | 695-760     | 510-608       | 422-585     | 256-423        |
| SM | g/100g | 69.82-73.96 | 53.80-62.38   | 42.39-48.03 | 29.11-40.21    |
| SL | g/100g | 3.46-6.40   | 5.26-6.60     | 5.66-7.70   | 3.88-8.29      |
| HV | Voxel  | 68270-92720 | 57710-77050   | 58940-67210 | 48580-62330    |
| BV | Voxel  | 19020-20850 | 15720-18220   | 12120-15980 | 8120-11340     |
| SV | Voxel  | 22980-31990 | 16950-23190   | 12160-17090 | 9860-14200     |

Once the database was built, weight, moisture and lipid content of Biceps femoris and Semimembranosus muscles were tried to be predicted. For that, predictive techniques of data mining, again multiple linear regression, was applied on different combinations of parameters achieved non destructively: i) ham weight and stage of processing; ii) ham weight, ham volume and stage of processing; iii) ham volume and stage of processing. Table 3 shows prediction equations obtained for weight, moisture and lipid content of Biceps femoris and Semimembranosus muscles for each combination, and Table 4 shows values of the correlation coefficients ( $R^2$ ) for each prediction equation. This statistical parameter allows verifying the quality of the equations as it determines the relation between analyzed data and predicted values of the first ham batch.

Table 3. Prediction equations obtained by applying multiple linear regression on different combination of parameters: A = Ham Weight + Stage of processing. B = Ham Weight + Ham Volume + Stage of processing. C= Ham Volume + Stage of processing. BW = Biceps Weight. BM= Biceps Moisture Content. BL = Biceps Lipid Content. SW = Semimembranosus Weight. SM= Semimembranosus Moisture Content. SL = Semimembranosus Lipid Content

|    | Prediction Equations                        |   |   |
|----|---|---|---|
|    | A   | B   | C   |
| BW | $BW = -158.61 * Stage + 0.05 * HW + 971.04$ | $BW = -159.61 * Stage + 0.05 * HW + 971.04$ | $BW = -221.82 * Stage + 1615.77$          |
| BM | $BM = -8.41 * Stage + 77.65$                | $BM = -8.41 * Stage + 77.65$                | $BM = -8.41 * Stage + 77.65$              |
| BL | $BL = 2.37 * Stage + 5.99$                  | $BL = 2.37 * Stage + 5.99$                  | $BL = 2.37 * Stage + 5.99$                |
| SW | $SW = -123.77 * Stage + 847.19$             | $SW = -123.77 * Stage + 847.19$             | $SW = -123.77 * Stage + 847.19$           |
| SM | $SM = -7.98 * Stage + 0.01 * HW + 27.54$    | $SM = -7.98 * Stage + 0.01 * HW + 27.54$    | $SM = -12.99 * Stage + 0.01 * HV + 79.30$ |
| SL | $SL = 0.29 * Stage + 6.00$                  | $SL = 1.02 * Stage + 0.01 * HW - 1.59$      | $SL = 0.28 * Stage + 5.99$                |

For weight and moisture content of the *Biceps femoris* and *Semimembranosus* muscles and for the lipid content of the *Biceps femoris* muscle, high correlation coefficients ( $R^2$ ) (0.89- 0.99) were obtained. The lipid content correlations related to the *Semimembranosus* muscle were not particularly accurate (0.02). In the three combinations of parameters on which multiple linear regression was applied, similar correlation coefficients were found. The option using ham volume and stage of processing would be selected since its computational costs are slightly lower in comparison to the others.

Table 4. Correlation coefficients ( $R^2$ ) of the prediction equation obtained by applying multiple linear regression techniques on different combination of parameters: A= Ham Weight + Stage of processing. B = Ham Weight + Ham Volume + Stage of processing. C= Ham Volume + Stage of processing. BW = Biceps Weight. BM= Biceps Moisture Content. BL = Biceps Lipid Content. SW = Semimembranosus Weight. SM= Semimembranosus Moisture Content. SL = Semimembranosus Lipid Content

|    | $R^2$ |       |       |
|----|-------|-------|-------|
|    | A     | B     | C     |
| BW | 0.977 | 0.977 | 0.975 |
| BM | 0.989 | 0.989 | 0.990 |
| BL | 0.897 | 0.897 | 0.910 |
| SW | 0.959 | 0.959 | 0.959 |
| SM | 0.990 | 0.989 | 0.984 |
| SL | 0.001 | 0.001 | 0.024 |

Once the prediction equations were obtained, they were tested by comparing their results to the data from its own batch of hams (Table 5). For weight and moisture content in Biceps femoris and Semimembranosus muscles and lipid content in Biceps femoris, high correlation coefficients (0.87- 0.97) were obtained. However, correlation coefficient was low for the lipid content in Semimembranosus (0.03).

Table 5. Correlation coefficient ( $R^2$ ) between chemical and predicted data in the first batch of hams

|                                 | $R^2$ |
|---------------------------------|-------|
| <i>Biceps femoris</i> Moisture  | 0.966 |
| <i>Biceps femoris</i> Lipid     | 0.871 |
| <i>Semimembranosus</i> Moisture | 0.969 |
| <i>Semimembranosus</i> Lipid    | 0.035 |

After then, results obtained from the predicted equations were compared to data from the second batch of hams (Table 6). High correlation coefficients ( $R^2$ ) (0.95- 0.98) were found in moisture content in *Semimembranosus* and *Biceps femoris* muscles, as can be observed graphically in Figure 4 and 5, respectively.

Table 6. Correlation coefficient ( $R^2$ ) between chemical data from the second batch of hams and predicted values

|                                 | $R^2$ |
|---------------------------------|-------|
| <i>Biceps femoris</i> Moisture  | 0.950 |
| <i>Biceps femoris</i> Lipid     | 0.242 |
| <i>Semimembranosus</i> Moisture | 0.980 |
| <i>Semimembranosus</i> Lipid    | 0.196 |

Data related to the lipid content in *Biceps femoris* and *Semimembranosus* muscles showed low correlation coefficients ( $R^2$ ) (0.19– 0.25). This outcome could be explained by the high variability of the chemical values for the lipid content in hams.

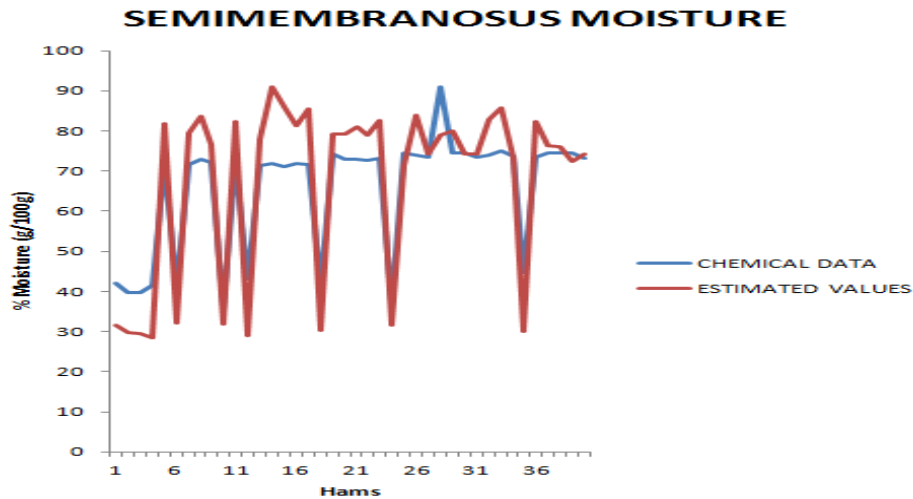


Figure 4. Comparison between the chemical moisture values of the *Semimembranosus* muscle and the estimated values from the prediction equation



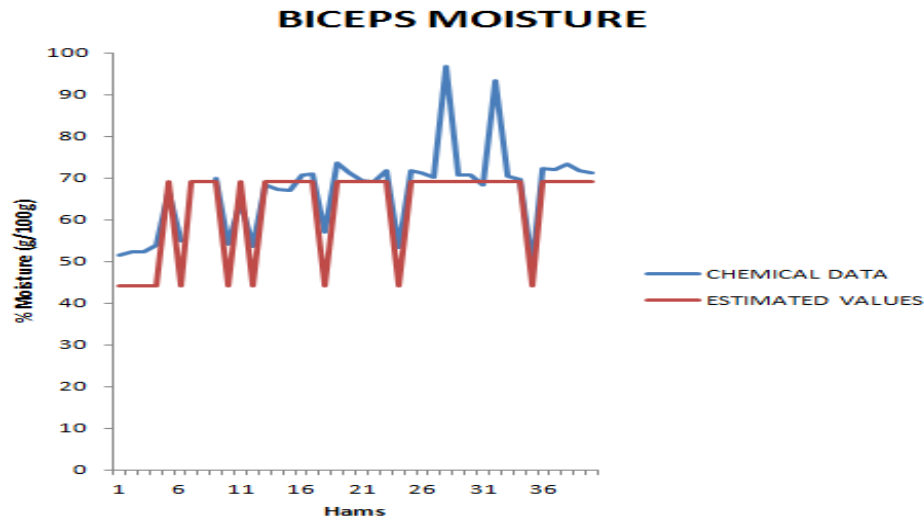


Figure 5. Comparison between chemical moisture values of the Biceps femoris muscle and estimated values from the prediction equation

#### 4. CONCLUSION

The combination of data mining, MRI and CVT is a pioneer approach to analyze ham quality features. These tools can be useful for calculating P-C parameters related to ham quality and improving the control of the processing without destructing the ham pieces.

Prediction equations have been obtained for measuring the moisture content of the Biceps femoris and the Semimembranosus muscle with a high correlation coefficient.

Multiple linear regression is an accurate method of data mining for predicting the weight and moisture values in Iberian ham.

To predict the lipid content of ham muscles (Biceps femoris and Semimembranosus) other techniques of data mining, e.g. isotonic regression, should be tested.

#### ACKNOWLEDGEMENTS

The authors wishes to acknowledge the funds received to carry out this research from both the Junta de Extremadura (Regional Government Board – Research Projects 3PR05B027 and PDT08A021; Consejería de Economía, Comercio e Innovación and FEDER– economic support for research 345 groups: GRU09148 and GRU09025) and from the Spanish Government (National Research Plan) and the European Union (FEDER funds) through the grant reference TIN2008-03063. We also wish to thank the “Hermanos Roa” company from Villar Del Rey (Badajoz).

## REFERENCES

- [1] Martin, L., Antequera, T., Ruiz, J., Cava, R., Tejada, J. F., &Cordoba, J.J., (1998) “Influence of the processing conditions of Iberian ham on proteolysis during ripening”, *Food Science and Technology International*, Vol. 4, pp. 17–22
- [2] Fulladosa, E., Santos-Garcés, E., Picouet, P., &Gou, P., (2010)“Prediction of salt and water content in dry-cured hams by computed tomography”, *Journal of Food Engineering*, Vol. 96, pp. 80–85
- [3] Santos-Garcés, E., Gou, P., Garcia-Gil, N., Arnau, J., &Fulladosa, E., (2010) “Non-destructive analysis of aw, salt and water in dry-cured hams during the drying process through computed tomography”, *Journal of Food Engineering*, Vol. 101, pp. 187–192
- [4] Pérez-Juan, M., Afseth, N.K., González, J., Díaz, I., Gispert, M., Furnols, M.F., Oliver, M.A., &Realini, C.E., (2010) “Prediction of fatty acid composition using a NIRS fibre optics probe at two different locations of ham subcutaneous fat”, *Food Research International*, Vol. 43, pp. 1416–1422
- [5] Bonny, J. M., Laurent, W., Labas, R., Taylor, P., Berge, P., &Renou, J. P., (2000) “Magnetic resonance imaging of connective tissue: A non-destructive method for characterizing muscle structure”, *Journal of the Science of Food and Agriculture*, Vol. 81, pp. 337–341
- [6] Fantazzini, P., Gombia, M., Schembri, P., Simoncini, N., &Virgili, R., (2009) “Use of Magnetic Resonance Imaging for monitoring Parma dry-cured ham processing”, *Meat Science*, Vol. 82, pp. 219-227
- [7] Manzoco, L., Anese, M., Marzona, S., Innocente, N., Lagazio, C., &Nicoli, M. C., (2013) “Monitoring dry-curing of S. Daniele ham by magnetic resonance imaging”, *Food Chemistry*, Vol. 141, pp. 2246-2252
- [8] Antequera, T., Caro, A., Rodriguez, P.G., & Pérez-Palacios, T., (2007) “Monitoring the ripening process of Iberian Ham by computer vision on magnetic resonance imaging”, *Meat Science*, Vol. 76, pp. 561-567
- [9] Caro, A., Rodríguez, P.G., Cernadas, E., Durán, M.L., & Villa, D., (2001)“Applying active contours to muscle recognition in Iberian ham MRI”, In *IASTED International Conference Signal Processing, Pattern Recognition and Applications*, Rhodes, Greece
- [10] Ávila, M., Durán, M.L., Caro, A., Antequera, T., & Gallardo, R., (2005) “Thresholding methods on MRI to evaluate intramuscular fat level on Iberian ham”, *Lectures Notes in Computer Science (LNCS 3523)*, *Pattern Recognition and Image Analysis*, pp. 697–704
- [11] Caro, A., Durán, M.L., Rodríguez, P., Antequera, T., & Palacios, R., (2003) “Mathematical morphology on MRI for the determination of Iberian ham fat content” *Lecture Notes in Computer Science (LNCS 2905)*, *Progress in Pattern Recognition, Speech and Image Analysis*, pp. 359-366
- [12] Colton, T., (1974) “*Statistical in Medicine*”, Ed. Little Brown and Co
- [13] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., (1996) “From data mining to knowledge discovery in databases”, *AI Magazine*, Vol. 17, pp. 37-54
- [14] Song, Y.H., Kim, S.J., & Lee, S.K., (2002) “Evaluation of ultrasound for prediction of carcass meat yield and meat quality in Korean native cattle”, *Asian Journal Animal Science*, Vol. 15, pp. 591-595
- [15] Cortez, P., Portelinha, S., Rodrigues, S., Cadavez, V., & Teixeira, A., (2006) “Lamb Meat Quality Assessment by support vector machines”, *Neural Processing Letters*, Vol. 24, pp. 41-51
- [16] Caballero, D., Pérez-Palacios, T., Ávila, M., Duran, M.L., Antequera, T., (2013) “Estudio de la difusión de sal en el jamón ibérico mediante la aplicación de minería de datos a imágenes de resonancia magnética”, *VII Congreso Mundial del Jamón*, Ourique, Portugal
- [17] Association of Official Analytical Chemist (AOAC), (2000), “*Official Methods of Analysis of the Association of Official Analytical Chemists*”, 17th ed., Gaithersburg, Maryland
- [18] Pérez-Palacios, T., Ruiz, R., Martin, D., Muriel, E., &Antequera, T., (2008) “Comparison of different methods for total lipid quantification”, *Food Chemistry*, Vol. 110, pp. 1025–1029
- [19] <http://www.cs.waikato.ac.nz/ml/weka> WEKA web reference tool
- [20] Witten, I.H., & Frank, E., (2005) “*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*”, Ed. Morgan Kaufmann, San Francisco, California

- [21] Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z-H., Steinbach, M., Hand, D.J., & Steinberg, D., (2008) "Top 10 algorithms in data mining", Knowledge and Information Systems, Vol. 14, pp. 1-37
- [22] Hastie, T., Tibshirani, R., Friedman, J., (2001) "The Elements of Statistical Learning: Data Mining, Inference and Prediction", Springer-Verlag, New York, New York