# Applying Image Mining to Support Gestational Age Determination

Anderson V. Araújo[1], Olga R.P. Bellon[1], Luciano Silva[1], Everton Vieira[1] and Mônica Cat[2]

June 30, 2006

[1]Imago Research Group – Departament of Informatics, Universidade Federal do Paraná
[2]Departament of Pediatrics – University Hospital

**Abstract**

*The gestational age (GA) determination is one of the most used measures to guide diagnostics of specialists in neonatology to guarantee the newborn surveillance. However, the currently available methods have been subject to a number of evaluations and criticism by important medical publications concerning both their low precision and their invasive procedures. Based on this scenario, the FootScanAge method was conceived, seeking to determine the GA through digital imaging of the newborn plantar surface. To support the evaluation and evolution of this new method, it was developed an Open Source and Java based Decision Support System that combines Data Mining and Image Processing techniques to implement the FootScanAge method. The system was developed taking advantage of high degree of interaction between experts in neonatology and computer science.*

## Contents



Current methods for gestational age (GA) determination – Ballard [1,2], Parkin [3], Dubowitz [4], Capurro [5] – contain three main limitations: (1) have low precision, (2) are invasive, requiring intensive handling of the newborns, and (3) involve few premature infants [6]. Based on this scenario, it was conceived the FootScanAge method, which presents a new alternative to current methods, once it is a non-invasive approach based on features automatically extracted from the newborn plantar surface image.

To support evaluation, evolution of this new method it was developed an Open Source and Java based Decision Support System that is composed by two main modules: FootScanAge Web and FootScanAge Desktop. The first one is an Electronic Patient Record tool that manages a database of infant and maternal data, including personal information, exams and diagnostics. The second one combines Image Processing techniques to extract features from infant’s footprint and analyzes them by using Data Mining algorithms. This combination is called Image Mining [7]. The focus of this paper is the Data Mining Tool.

# 1 Concepts

To determinate the GA score, it is necessary to face some challenges: current methods are based on researches involving few premature newborns; best methods involve collecting many features, most of them requiring intensive handling of the newborns; faster methods are more imprecise, once they use less features. The FootScanAge method was developed aiming to obtain the GA score automatically based only on a simpler source of features: the image of newborn plantar surface [6, 8].

The essence of the FootScanAge method is based on learning a model of rules from a database of image features and applying the discovered knowledge to estimate the GA of a new infant footprint. Therefore, there is a need for building an appropriate image database and validate the method as long as the newborn population grows. To corroborate to this strategy, the FootScanAge system was developed to allow database management as well as to improve the method by adding more features and rules as necessary.

In this context, techniques were borrowed from different disciplines creating a new one, called Image Mining [7]: the feature extraction and data cleaning techniques from Data Mining [9] are combined with algorithms for feature extraction from Computer Vision and Image Processing [10], and also some Machine Learning strategies used by Data Mining are improved and applied on the database. In fact, one of the main contributions of the FootScanAge system is the ability to combine multiple algorithms, techniques and paradigms, once all these approaches are merged to help the expert in neonatology to analyze the newborn and infer a final, ultimate gestational age score.

# 2 System architecture

The FootScanAge system is composed by two modules: FootScanAge Web, which serves as basis for cataloging and management of mother/newborn information, including pre-birth and post-birth data, and the FootScanAge Desktop, which implements the Image Mining solution.

This FootScanAge Web module was developed using several Java-based technologies like: an application server and a database management system to store and retrieve data. To run the Web module, the user has to access the website location (see Figure 1). Then, after selecting the appropriate menu, the user is able to do a variety of operations: (a) to manage mother information; (b) to include diagnostics about exams applied to the mother and newborn; (c) to acquire plantar surface images and submit them to the FootScanAge Desktop module (Figure 2).


**Figure 1** FootScanAge Web initial page.

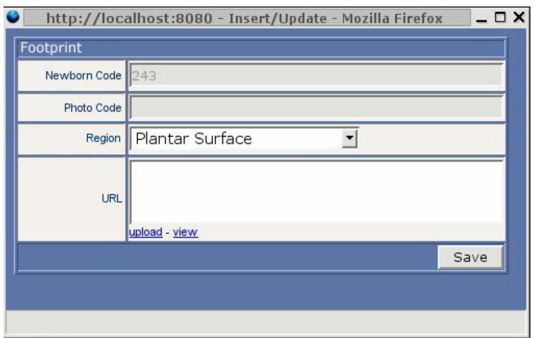


**Figure 2** Acquiring and uploading a footprint.

To allow the implementation of the Desktop module, many Java-based and free libraries were combined. The Image Processing algorithms were developed based on the JAI framework[1] (Java Advanced Imaging). JAI is a library that comprises tools to deal with different image formats and contains built-in operators and filters to allow the execution of various tasks, when dealing with images, and to create new algorithms. The data-mining phase is developed based on Weka API [11] machine learning algorithms. After uploaded, the footprint image is then processed and analyzed, as described in the next sections.

## 3 Image Processing

To generate the set of image features that will be analyzed and mined in the next phase, the following steps are executed by the system:

a) Preprocessing: performs image rotation, noise reduction and selection of the region of interest.
b) Binarization: converts the image to black and white, keeping the footprint quality.
c) Labeling: removes small regions and helps the removal of the fingers.
d) Finger Removal: isolates the plantar surface.
e) Region Detection: helps the feature extraction step.
f) Feature Extraction: extracts the necessary characteristics from the plantar surface.

At the end of this process, the user confirms and saves the results into the database, starting the phase described in the next section. The FootScanAge´s Image Processing Tool is described in details in [12].

## 4 Image Mining

The main purpose of the Image Mining tool is to estimate the GA score. This task is accomplished by combining two different mining approaches, classification and regression, generating two scores. These results, when used together with the knowledge obtained by traditional methods (if available), guide the specialist in neonatology to infer the final score. This process is illustrated in Figure 3 and Figure 4:

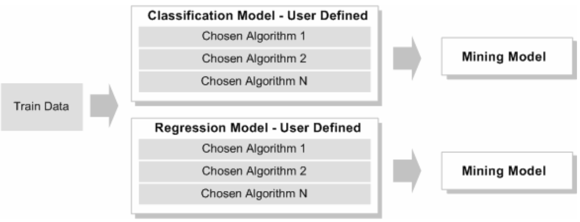


**Figure 3** Creating a mining model.

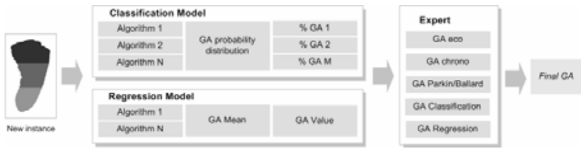


**Figure 4** Determining the final score.

The first step consists in building a knowledge database from the training data, called mining model. The system confronts the image features of each record against its GA, generating a mining model by means of supervised learning. Each model can have one or more algorithms combined, creating an ensemble [13] of algorithms. In fact, each algorithm can be an ensemble of other algorithms. It is possible to save these models for future use, *i.e.*, the user has no need to generate a new model for every new instance.

[1]Java Advanced Imaging – Available at: http://java.sun.com/products/java-media/jai/index.jsp

When a new image is analyzed, its features are extracted and the classification and/or regression models are applied to the current instance, generating two kinds of output: (1) a probability distribution per GA for the classification model, and (2) a suggested score from the regression model. These outputs are presented in textual and graphical form, according to Figure 6 and Figure 7.

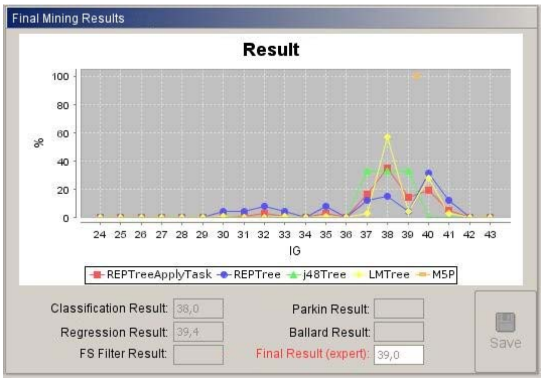


**Figure 6** Graphical output of regression and classification scores.

Classification Result:

| Algorithm | IG | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Final | 38 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 1 | 0 | 3 | 0 | 16 | 35 | 14 | 19 | 5 | 0 | 0 |
| REPTree | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 8 | 4 | 0 | 8 | 0 | 12 | 15 | 4 | 31 | 12 | 0 | 0 |
| j48Tree | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 | 33 | 33 | 0 | 0 | 0 | 0 |
| LMTree | 38 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 3 | 57 | 4 | 28 | 3 | 0 | 0 |

**Figure 7** Example of combined probability distribution for the classification algorithms.

The purpose of combining algorithms and mining strategies is to reflect the uncertainty inherent of the GA determination process. At this level, the FootScanAge system acts as a Decision Support System where the user considers the opinion of different specialists – the FootScanAge method, traditional methods and his/her own expertise – to determine the final result.

## 5 Tests and Results

To assert the feasibility of the proposed method, several evaluations were executed on a test database containing 280 instances. The tests had the following objectives: to explore the best regression and classification algorithms available from the Weka API; to define strategies to clean the data (remove outlier instances); and to assess how the combination of features affects the final results.

The tests were applied using cross-validation [14] and reveals that the algorithms that best suit the test database are: J.48 (C4.5 [15] based implementation), REPTree [16] (Reduced Error Pruning Tree), M5P [15] (combination of regression and decision trees) and LMT [17] (Logistic Model Trees). These algorithms had similar results, although with different probability distribution per gestational age (for classification algorithms) and different precision scores (for regression algorithms).

Diverse combinations of features were also analyzed. To select the attributes, the same classification and regression algorithms were applied, giving a hint of the best attribute grouping. To clean the database, a classification or regression algorithm was applied and the outliers were removed (cleaning filter). Table 1 shows the results obtained from the application of the M5P algorithm with different features/instance sets. The first column indicates how many attributes were used. The second column shows if the cleaning filter was used. The third column specifies if the applied algorithm used an ensemble method (Bagging). The last column demonstrates the final result, the mean absolute error of the algorithm, in weeks.

Other experiments were also performed to evaluate how the classification strategy could behave with different combinations of algorithms and attributes as well as to assess how the probability distribution could influence the GA score. The tests exhibited in Table 2 considered the training of the classification algorithms (REPTree combined with J.48, REPTree standalone and Bagged REPTree) using 67% of the 280 instances and the application of the mining model on the remaining. This table shows four samples of the test instances (A, B, C, D), the corresponding probability distribution (some values were omitted), the expected GA and the predicted GA (FS).

Each test group had a mean absolute error of 1.77, 2.08, 1.57 and 1.4 weeks, respectively, for the entire test set. These results illustrate the importance of to weigh carefully the combination of algorithms and attributes and how the use of Bagging with classification can also improve the precision of the method. Furthermore, it is important to notice that in some cases, even when the mining algorithms picked the incorrect GA, the probability distribution shows that the correct class or a closer class could be chosen. Considering current GA methods, the mean absolute error of 1.14 weeks is a very promising score. The achieved results also reveal that the use of all attributes can reduce the precision of the method. The use of an ensemble strategy and other information regarding the binomial newborn/mother could also enhance the solution and its accuracy.

**Table 1** Regression results**.**

| **Attributes** | **Cleaning Filter** | **Bagged** | **Mean Absolute Error** |
|---|---|---|---|
| All | No | No | 1.4484 |
| All | Yes | No | 1.1991 |
| All | No | Yes | 1.3733 |
| All | Yes | Yes | 1.1480 |
| 10 | No | No | 1.3573 |
| 10 | Yes | No | 1.1638 |
| 10 | No | Yes | 1.3313 |
| 10 | Yes | Yes | 1.1488 |

**Table 2** Experiments with Classification.

**(a) REPTree and J4.8 - All attributes**

| # | **GA** | **FS** | **32** | **33** | **34** | **35** | 36 | **37** | **38** | **38** | **40** | **41** | **42** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 40 | 40 | 0 | 0 | 0 | 0 | 0 | 8 | 9 | 0 | 69 | 13 | 1 |
| B | 40 | 37 | 0 | 0 | 0 | 0 | 25 | 33 | 9 | 0 | 19 | 3 | 1 |
| C | 33 | 33 | 4 | 73 | 0 | 0 | 19 | 4 | 0 | 0 | 0 | 0 | 0 |
| D | 42 | 41 | 0 | 0 | 0 | 0 | 0 | 8 | 9 | 17 | 19 | 46 | 1 |

**(c) REPTree - 7 attributes**

| # | **GA** | **FS** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **38** | **40** | **41** | **42** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 40 | 40 | 0 | 0 | 3 | 1 | 1 | 12 | 21 | 1 | 43 | 16 | 2 |
| B | 40 | 40 | 0 | 0 | 3 | 1 | 1 | 12 | 21 | 1 | 43 | 16 | 2 |
| C | 33 | 33 | 4 | 31 | 15 | 0 | 19 | 4 | 0 | 0 | 4 | 0 | 0 |
| D | 42 | 40 | 0 | 0 | 3 | 1 | 1 | 12 | 21 | 1 | 43 | 16 | 2 |

**(b) REPTree and J4.8 - 7 attributes**

| # | **GA** | **FS** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **38** | **40** | **41** | **42** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 40 | 41 | 0 | 0 | 2 | 1 | 1 | 6 | 11 | 1 | 3 | 5 | 1 |
| B | 40 | 38 | 0 | 0 | 2 | 1 | 1 | 6 | 44 | 1 | 27 | 14 | 7 |
| C | 33 | 36 | 2 | 15 | 8 | 0 | 6 | 2 | 0 | 0 | 2 | 0 | 0 |
| D | 42 | 41 | 0 | 0 | 2 | 1 | 1 | 6 | 11 | 1 | 3 | 50 | 1 |

**(d) Bagged REPTree - 7 attributes**

| # | **GA** | **FS** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **38** | **40** | **41** | **42** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 40 | 40 | 0 | 0 | 1 | 0 | 0 | 2 | 18 | 0 | 31 | 29 | 1 |
| B | 40 | 40 | 0 | 0 | 2 | 3 | 1 | 16 | 2 | 1 | 37 | 3 | 18 |
| C | 33 | 33 | 1 | 26 | 22 | 0 | 23 | 17 | 0 | 0 | 3 | 4 | 0 |
| D | 42 | 40 | 0 | 0 | 0 | 0 | 0 | 19 | 6 | 2 | 47 | 24 | 3 |

Although the use of fewer attributes shows best results, we cannot say yet that those selected attributes are adequate for all kind of newborns. We believe that using a greater population of infants, from different geographic regions, could allow us to have a definitive conclusion regarding the method. However, the presented results show the importance of to build the FootScanAge system as a modular system that can accompany the database evolution with the flexibility to help the user to infer the final result from different sets of features and algorithms.

## 6 Final Remarks

To support the FootScanAge method for GA determination, it was developed a system that helps gathering data about infants, analyzes features obtained from the footprints and helps the expert to infer the final gestational score. To succeed in this task, the application uses an Image Mining process that merges techniques from different disciplines, such as Image Processing and Data Mining. The final result is a system that maintains a knowledge database that evolves as long as new instances of newborns and new types of features are obtained. An important contribution of the FootScanAge System is the employment of state-of-art techniques of Image Processing, Machine Learning and Open Source software, showing that this strategy is reliable to improve research in the medical area.

Preliminary results shows that the obtained FootScanAge score is very promising compared to traditional methods. In order to validate and to enhance the method, the current version of the system is being applied in the University Hospital, aiming to obtain new instances, new features and new mining models.

Another relevant aspect of the system is its ability to combine different machine learning approaches and its flexibility to handle different features, obtained from the image or combined with other data. With a careful analysis of the mining models and image features, the FootScanAge method can be applied in other cities or countries, creating more robust, precise models. Also, this solution can be applied to other images and features, allowing its use in different scenarios regarding Medical Decision Support Systems.

The authors would like to thank FINEP, CNPq and CAPES for the financial support.

## References

[1] J. L. Ballard, J. C. Khouri, K. Wedig, L. Wang, B. Eilers-Walsman, and R. Lipp. *New Ballard Score, expanded to include extremely premature infants.* The Journal of Pediatrics, v. 199, p.417-423, 1991

[2] J. Ballard, K. Novak, and M. Driver. *A simplified score for assessment of fetal maturation of newly born infants.* Journal of Pediatrics, v.95, n.1, p.769-774, 1979.

[3] J. M. Parkin, E. N. Hey, and J. S. Clowes. *Rapid assessment of gestational age at birth.* Archives of Diseases in Childhood, v.51, n.4, p.259-263, 1976.

[4] L. M. Dubowitz, V. Dubowitz, and C. Goldberg. *Clinical assessment of gestational age in the newborn infant.* Journal of Pediatrics, v.77, p.1-10, 1970.

[5] H. Capurro, S. Konichezki, D. Fonseca, and R. Caldeyro-Barcia. *A simplified method for diagnosis of gestational age in the newborn infant.* Journal of Pediatrics, v.93, p.93-120, 1978.

[6] O. R. P. Bellon, M. Cat, L. Silva, and K. L. Boyer. *Using computer vision to help the determination of the gestational age of newborns.* Academic Radiology, v.12, n.5, p.544-553, 2005.

[7] J. Zhang, W. Hsu, and L. M. Lee. *Image mining: issues, frameworks and techniques.* Proceedings of the 2$^{nd}$ International Workshop on Multimedia Data Mining, p.13-20, 2001.

[8] M. N. L. Cat. *The FootScanAge method for gestational age determination*. Ph.D. thesis. Universidade Federal do Paraná, Curitiba (PR), 2003, in Portuguese.

[9] U. M. Fayyad, G. Piatestsky-Shapiro, and P. Smyth. *From Data Mining to Knowledge Discovery: An Overview.* In: Advances in Knowledge Discovery and Data Mining, p.1-34, 1996.

[10] R. C. Gonzales, and R. E Woods, R. E. *Digital Image Processing.* Addison Wesley, 2000.

[11] I. H. Witten, and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2005.

[12] L. Silva, O. R. P. Bellon, R. P. Lemes, J. A. Meira, and M. N. L. Cat. *An image processing tool to support gestational age determination*. Proceedings of the $19^{th}$ IEEE International Symposium on Computer-Based Medical Systems, pp.867-872, 2006.

[13] T. G. Dietterich. *Ensemble Methods in Machine Learning.* Lecture Notes in Computer Science, v.1857, p.1-15, 2000.

[14] C. Schaffer. *Selecting a Classification Method by Cross-Validation, Machine Learning*. v.13, p.135-143, 1993.

[15] R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[16] R. Quinlan. *Learning with continuous classes.* Proceedings of the $5^{th}$ Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore, p. 343-348, 1992.

[17] N. Landwehr, M. Hall M, and E. Frank. *Logistic Model Trees.* Proceedings of the 14th European Conference on Machine Learning, p.241-252, 2003.