A Novel Approach for Labelling Health Insurance Data for Cardiovascular Disease Risk Classification

César L. C. Mattos*, Amauri H. Souza Júnior†, Ajalmar R. Rocha Neto‡
Guilherme A. Barreto†, Ronaldo F. Ramos‡, Hélio A. Mazza* and Márcio O. Mota§

*Dyad & Associates, Rua Antônio Augusto, 1271 - Sala 106
Ed. Medical Gênesis, Aldeota, CEP 60110-370, Fortaleza, Ceará, Brazil
Telephone: +55 85 4008–4711, Emails: {cesar,mazza}@dyad.com.br
†Department of Teleinformatics Engineering, Federal University of Ceará
Center of Technology, Campus of Pici, Fortaleza, Ceará, Brazil
Emails: guilherme@deti.ufc.br, amauri01@gmail.com
‡Federal Institute of Education, Science and Technology of Ceará, Fortaleza, Ceará, Brazil
Emails: ajalmar@gmail.com, ronaldo.ramos@gmail.com
§ State University of Ceará, Master Program in Business Administration (CMAAd)
Av. Paranjana, 1700, Bloco CESA, Fortaleza, Ceará, Brazil
Email: marcio@marciomota.com

***Abstract*—Health insurance companies own very large databases built from the history of clinical exams and/or hospital procedures undergone by their beneficiaries. An important challenge faced by these companies is then to mine useful information from those database for the purpose of preventive care and financial costs reduction. Bearing this in mind, in this paper we propose a novel approach for building and labelling feature vectors for the beneficiaries of health insurance companies with the aim of building classifiers capable of predicting the risk level (high or low) of a given beneficiary to undergo serious cardiovascular events within a predefined horizon in the near future. The proposed approach was evaluated in the design of neural network classifiers using real-world health data from a Brazilian insurance company. The obtained results show that the proposed method is rather promising and can be used to aid the management of health insurance plans.**

I. Introduction

Health insurance companies store a huge amount of usage data from their beneficiaries. It is well known that such data may be used for analysis of risk factors for diseases [1]. More interestingly, whole sequences of procedures can be tracked, since the first medical exams to more serious interventions. That sequence of procedures defines what may be called a *disease evolution*. When a serious disease is considered, such as heart disease, early diagnosis can be very useful for both the beneficiary and the insurance company: the first may receive preventive care and the second may avoid high cost procedures, such as surgeries.

Many authors have studied health data mining for useful information about risk of diseases. In [2], statistical risk prediction models were studied for managing patients in the cardiovascular domain, with the focus of evaluating risk equations retrieved from the MEDLINE database. In [3], besides rule based models, other techniques, such as decision trees and neural networks, are surveyed for health data mining.

In [4] several algorithms to estimate the risk of cardiovascular disease in asymptomatic patients are described and the importance of preventive treatments of individuals with high risk are emphasized. A study about risk prediction of cardiovascular events in near future was also done in [5]. In that paper, the main objectives included the definition of risk factors, biomarkers and specific clinical results. However, this information is often not available and difficult to obtain.

When using non-specific attributes, artificial neural networks may be applied. Recently, for example, neural networks were used for severity prediction of several acute pancreatitis from hospital data in [6]. Breast cancer risk estimation was carried out in [7] using both logistic regression and multilayer perceptron network, which proved to be useful to aid in decision making. In [8], grammatical evolution neural networks [9] were used to identify individuals with high risk for aging-related macular degeneration.

In [10], the author lists different types of data for healthcare risk prediction: physician referral/chart, enrollment data, claims, pharmacy data, laboratory values and self-reported information. The present paper focus on medical and hospital claims, as they are widely available in health plan context. Thus, in this paper we propose a novel methodology for designing a classifier able to find patterns in health usage data that indicate high or low risk of a certain beneficiary to undergo serious cardiovascular events in the near future.

The core of the proposed methodology is a new strategy for labelling feature vectors based on a time window into the future within which we search if the beneficiaries underwent any kind of serious cardiovascular problem. If a given beneficiary underwent any serious cardiovascular event within this time window, his/her feature vector is tagged with a high risk label; otherwise, his/her feature vector is labelled as a low risk one. This labelling strategy is then henceforth referred to as *predictive labelling*. Using the proposed predictive labelling strategy, linear and nonlinear neural network classifiers were

trained in order to produce a risk score for an individual based on its past clinical and/or hospital procedures, as registered by the health insurance company.

The remainder of the paper is organized as follows. In Section II we provide a detailed description of the proposed approach for feature vector building and labelling for risk classification of health insurance data. In Section III we report the results achieved by the evaluated classifiers in the task of of predicting the risk level (high or low) of a given beneficiary to undergo serious cardiovascular events within a predefined horizon in the near future using real-world data. Finally, the paper is concluded in Section IV.

II. Proposed Approach

This section will describe the proposed approach used to tackle the problem of risk classification from beneficiaries usage data. The methodology consists of steps before the training of the classifier, such as building the feature vectors, the classifier training itself and the post-training evaluation process. It worth noting that the proposed labelling approach can be used used by any type of pattern classifier, not only neural network based ones.

A. Building the Feature Vectors

Health insurance companys generate large amounts of data due to the several procedures a beneficiary may use. For instance, based on this longitudinal research from 2006 to 2011, there are almost 150 million events of data, more than 900.000 beneficiaries and 6.877 procedures. Therefore, the first limitation of this study was the computer processing so that all data was organized into binary patterns before sending to the training algorithm.

These attributes consist mainly of health procedures (e.g. electrocardiogram, echocardiogram, cardiologist appointment) used by beneficiaries in a prefixed time window (e.g. 6, 12 and 24 months), denominated henceforward *procedure window*. Each evaluated procedure window results in its associated attribute having the value 1 (one). In the case where the procedure was not done, its attribute value is 0 (zero).

Moreover, some personal information is appended to the previous string. More specifically, two more binary attributes are added: one representing the gender of the beneficiary (male or female) and other representing its inclusion in a given age group. For example, as this study analysed cardiovascular diseases, it was also added an age attribute for the ones over or below 40 years old. According to the Brazilian Health Department [11] , there is a higher risk of having arterial hypertension and heart attack the ones over 40 years old.

The result of the feature vector building process is a binary feature vector representing the usage pattern of a beneficiary in a given time window. The Equation 1 formalizes mathematically a pattern $\mathbf{p}$:

$$\begin{aligned}
\mathbf{p} &= [a_1\ a_2\ \cdots\ a_j\ \cdots\ a_J]\,, \quad \text{where} \\
a_j &= 1, \forall j \in [1, J-2], \text{if procedure } i \text{ was} \\
&\qquad \text{done within the time window,} \\
a_j &= 0, \forall j \in [1, J-2], \text{ otherwise,} \\
a_{J-1} &= 1, \text{ if the beneficiary is male,} \\
a_{J-1} &= 0, \text{ otherwise,} \\
a_J &= 1, \text{ if the beneficiary is in a prefixed age group,} \\
a_J &= 0, \text{ otherwise,}
\end{aligned} \tag{1}$$

where J is the total number of attributes and a_j is the j-th attribute.

B. Feature Selection

There are 6.877 different procedures used at least once in our database. This would imply 6.877 attributes for the training algorithm. However, as the target procedures are known in advance, it is possible to do a statistical based feature selection before training.

Let A be the event of occurrence of one target procedure (e.g. myocardial revascularization) and $P(A)$ its probability. Let also B_j be the event of occurrence of the procedure B_j. The conditional probability $P(A|B_j)$ is calculated by:

$$P(A|B_j) = \frac{P(A \cap B_j)}{P(B_j)}. \tag{2}$$

The probabilities in the right side of Equation 2 may be estimated from the examples available in our database as follows:

$$P(A|B_j) = \frac{|A \cap B_j|/N}{|B_j|/N} = \frac{|A \cap B_j|}{|B_j|}, \tag{3}$$

where N is the total number of events and $|\cdot|$ denotes the cardinality operator.

Equation 3 is calculated for every procedure B_j. The result is the estimated probability that a beneficiary went through a target event A, given that he/she also had a procedure B_j. These conditional probabilities are sorted from greatest to lowest and only a fraction is selected. Indeed, due to the quantity of procedures, we arbitrary used only the 400 procedures with the greater conditional probability.

C. Labelling the Feature Vectors: A 'Time Window into the Future' (TWINF) Approach

The risk classifier is going to be trained in a supervised way, i.e., given a set of training patterns built through the instructions described in Section II-A, thus it is necessary to label them for the training process.

In traditional classification tasks, patterns are labeled in a very objective way, usually through simple observation of the sources of the patterns. However, our risk classifier intends to predict, within a given future window, the risk of a beneficiary does some target procedures. Therefore, there is a time component in the analysis.

This paper proposes a novel approach called Time Window into the Future (TWINF). This methodology labels a pattern

based on known future events of the beneficiary from whom the pattern was built. For instance, if the beneficiary would do a target procedure in a prefixed future time window (e.g. 6 months), the pattern is considered to have a high risk and it will be labeled with the value 1. Otherwise, if it is considered to have a low risk, it will be labeled with the value 0. Hereafter, this future time window will be denominated Time Window into the Future (TWINF).

A useful feature of this labeling strategy is its ability to track the changes of a beneficiary along the time, i.e., its risk changes. For example, given several training patterns obtained from the same beneficiary for different time periods, the label of each pattern may vary as varies the risk of the beneficiary over time.

D. Training Process

The training set, i.e., the set of labeled patterns used for training, is obtained through the following steps:

- **Step 1** Extract raw data from the database which contains usage information of the beneficiaries;
- **Step 2** Perform feature selection as described in Section II-B;
- **Step 3** Find every beneficiary that did a target procedure somewhere in the time period analysed;
- **Step 4** For each beneficiary found on Step 3, build a pattern considering a procedure window counted from the first time a target procedure was done. This is intentioned to capture the moment of transition when the beneficiary passes from low risk to high risk. These patterns are called *target patterns*;
- **Step 5** Find every beneficiary that never did a target procedure during the same time period;
- **Step 6** For each beneficiary found on Step 5, build patterns considering a procedure window counted from each available month, which generates multiple patterns per beneficiary. These patterns are called *non-target patterns*;
- **Step 7** Obtain a training set formed by every target pattern and the same quantity of non-target patterns, randomly sampled. Note that the amount of non-target patterns is much greater than the amount of target patterns. In our database, the approximate ratio of target beneficiaries is 1:370. Random undersample of the majority class, i.e., non-target patterns, has been a simple and effective choice for learning from imbalanced datasets, as experimented in [12].

It is important to emphasize that, although the procedure window is relative to a given month, the objective is to learn the pattern of a beneficiary who is close by doing a target procedure, independently the month that it occurs.

Figure 1 illustrates the training set building process from the database to the attribute vectors.

E. Classifier Evaluation

During the evaluation of the classifier, unlabelled patterns are obtained as described in Section II-A. The operational of the risk classifier evaluation is run monthly so that the procedure and prediction windows must be counted from the present month. Note that in this case there will be only one pattern per beneficiary.

Figure 2 shows a diagram of the evaluation process of the risk classifier from unlabelled patterns. The steps below summarize the analysis:

- **Step 1** From the beneficiaries active on month M, a set of patterns is built following the methodology described in Section II-A;
- **Step 2** The risk classifier previously trained receives the set of test patterns;
- **Step 3** The estimations outputted by the classifier are sorted;
- **Step 4** A fraction of the correspondent beneficiaries are selected. The selected beneficiaries are considered to have high risk and are, for example, sent to preventive care.

After Step 4, the *recall* metric may be calculated by the following equation:

$$R = \frac{|S \cap T|}{|T|}, \tag{4}$$

where S is the set of beneficiaries selected by the classifier, $|S \cap T|$ is the number of target beneficiaries selected, i.e., correctly identified, and $|T|$ is the total number of target beneficiaries in the test set. For example, if the test set is known to have 100 high risk beneficiaries and, after the application of the risk classifier, 50 of these were found, the recall would be equal to 0.5, or 50%.

Furthermore, other important metric is the fraction of beneficiaries selected in Step 4:

$$C = \frac{|S|}{|T| + |G|}, \tag{5}$$

where $|G|$ is the number of non-target beneficiaries in the test set. This fraction represents the *sensitivity*[1] of the evaluation, i.e., the ratio of beneficiaries selected after the risk estimation done by the risk classifier. For example, if the test set consists of 1000 beneficiaries (with high or low risk) and the sensitivity is equal to 5%, the 50 beneficiaries with greater risk estimation are selected and labeled as high risk.

It is important to emphasize that the greater the recall and the less the sensitivity the better. Besides, there is a balance between these two metrics, for when the sensitivity is lower, the recall is also lower. In this paper the sensitivity was arbitrary fixed as 5% of the number of beneficiaries in the test.

III. Experimental Results

The real-world data described in Section II-A was used to validate the methodology proposed in Section II. The evaluated neural networks were the linear Perceptron (LP) and a one-hidden-layered MLP network, as described in [13]. The target

[1]Here the term *sensitivity* is applied in a different sense of the one usually found in machine learning literature, where one wishes higher values. In our definition lower values are preferable.

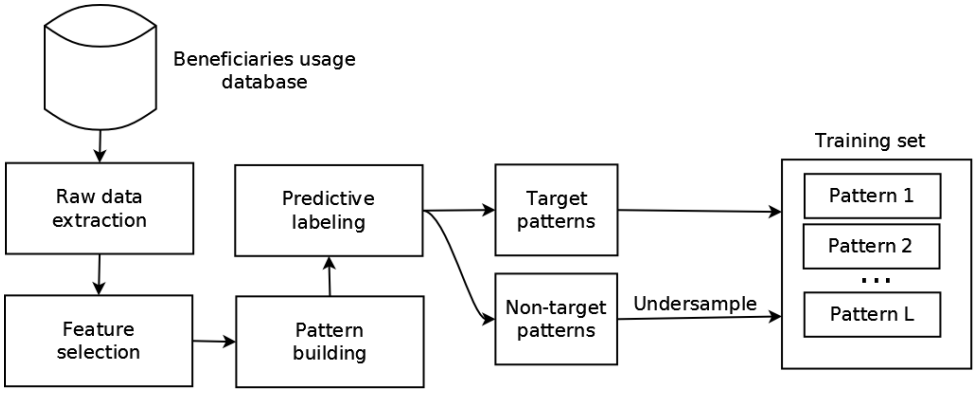


Fig. 1. Diagram of training set building process for the risk classifier from the database.

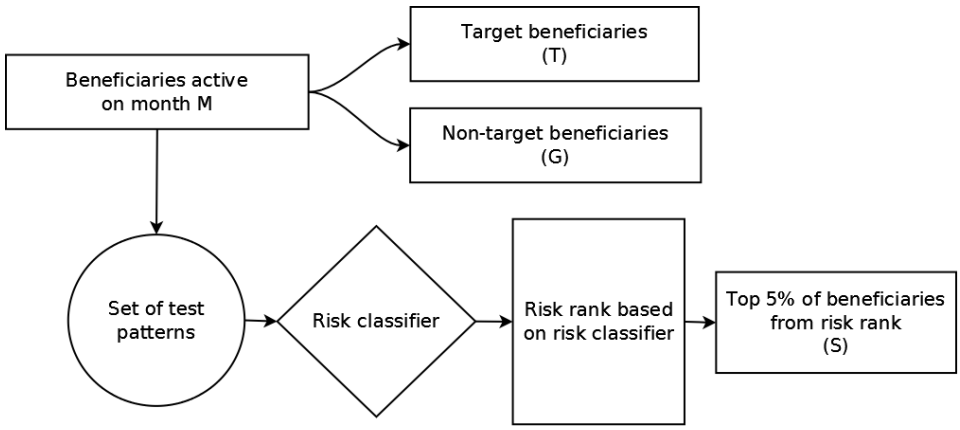


Fig. 2. Diagram of the evaluation process of the risk classifier. The sensitivity was fixed as 5%.

procedures are 116 high cost procedures related to serious heart diseases. The database structure follows the description done in Section II-A.

This section is divided in two parts, one showing the results of the neural networks and the other introducing some empirical financial implications of the use of the proposed methodology for risk classification.

A. Neural Networks Evaluation

The evaluation methodology follows the steps of Sections II-D and II-E. The training set was obtained from the 2007 and 2008 database of a large Brazilian health insurance company. After the undersampling step, described in Section II-D, 4.755 pattern were built for training, half from target beneficiaries, half from non-target ones. Two groups of test sets were built, one from the year 2009 and another from 2010. The evaluation of the classifier was run monthly, as described in Section II-E. Hence, there are 24 test sets, each one with around 695.000 patterns. The procedure and prediction windows were maintained in 12 and 6 months, respectively.

A MLP and a LP were trained using a learning step of 0.01 for 150 epochs. The number of hidden neurons of the MLP was set to 10 and the weights were updated through backpropagation. As mentioned in Section II-A, the number of features applied were 402, including 400 procedures selected through the method explained in Section II-B, the gender of the beneficiary and a boolean indicating if his age is over or below 40 years. It is worth mentioning that the feature selection step was made only in the training set.

Figure III-A shows the results for the test set of 2009. The same trained neural network was used to classify the beneficiaries of 2010, with results shown in Figure III-A. As a general result, one can easily note that the performance of the MLP is better than that of the LP network.

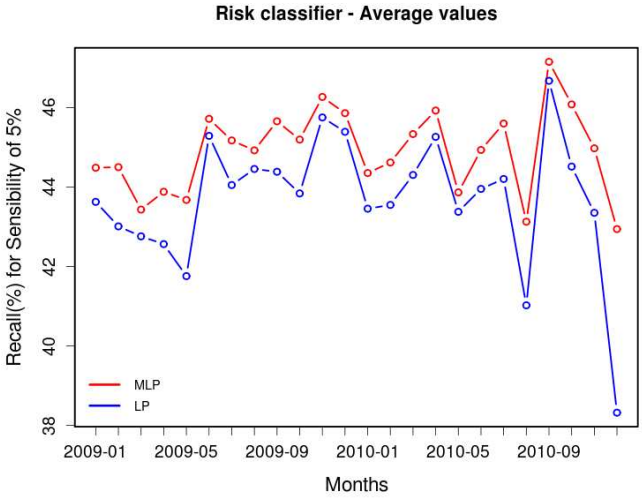
The percentage of high risk beneficiaries in both years is around 0.27% per month. With a fixed sensitivity of 5%, i.e., a selection of 5% of the active beneficiaries, the risk classifier was able to recognize 46.04% of the target beneficiaries on average per month in 2009 and 49.66% in 2010.

One interesting indication is that a trained classifier for a specific year can be successfully used in more than one year afterwards, a property which may indicate that the risk model does not change abruptly, as shown in Figures III-A and III-A.

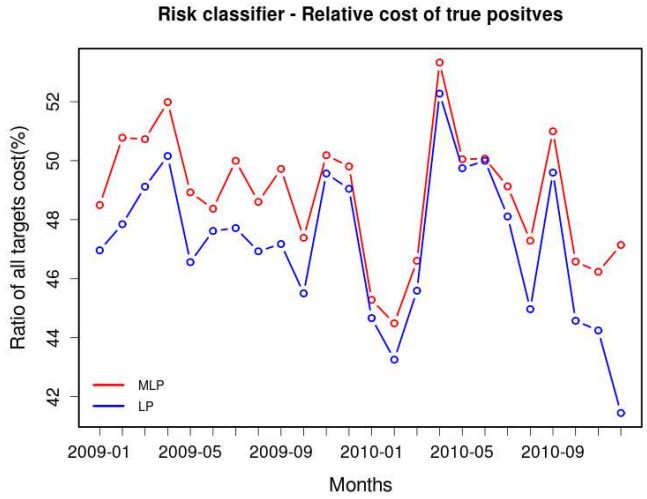
B. Empirical Financial Analysis

Besides, the recall values presented in the previous section is also interesting to estimate the potential financial impact of the application of a risk classifier with the proposed methodology.

Figure 4 presents a comparative analysis of the average costs of the correctly classified high risk beneficiaries (true



(a) Results for year 2009.



(b) Results for year 2010.

Fig. 3. Results of the risk classifier based on the MLP and LP networks.

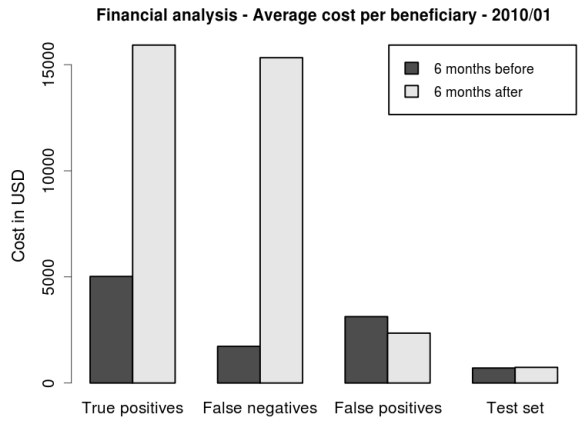


Fig. 4. Average costs per beneficiary before and after the risk analysis.

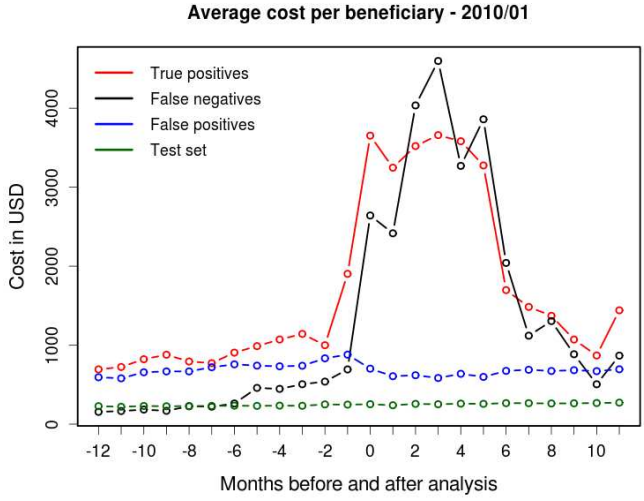


Fig. 5. Time evolution of average costs per beneficiary before and after the risk analysis.

positives), the high risk beneficiaries not found (false negatives) and the false positives 6 months before and 6 months after the risk analysis.

It may be perceived that the detected high risk beneficiaries are more costly than the non-detected. Besides, comparing the false positives and the average cost in the test set, the average cost of the first are greater, which might indicate that, although they did not have any of the target procedures in the predictive window, they possibly had other high cost procedures.

Figure 5 represents the time evolution of the average costs per beneficiary along the months before and after the risk analysis. The average cost of the true positives are greater than the cost of the false negatives during almost all period. The cost of the false positive is greater than the average values of the test set, as observed before. Figure 5 also shows a possible indication of the incorrect classification of the false negatives: their cost before the risk analysis is much lower than the true positives.

The Figure 6 shows the average cost of the beneficiaries labeled by the risk classifier as high risk, i.e., true positives and false positives. These amounts might guide the investments of a health insurance company in preventive programs.

IV. Conclusions and Further Work

In this paper, we proposed a novel approach, called TWINF (*Time window into the Future*), for building and labelling feature vectors for the beneficiaries of health insurance companies. The main goal of the TWINF approach is the design of pattern classifiers capable of predicting the risk level (high or low) of a given beneficiary to undergo serious cardiovascular events within a predefined horizon in the near future. The proposed approach was evaluated in the design of neural network classifiers using real-world data from a Brazilian health insurance company. The obtained results have shown that the proposed method is rather promising and can be used to aid the management of health insurance plans.

Currently we are evaluating several feature selection strategies for reducing even further the number of procedures used to build the feature vectors.

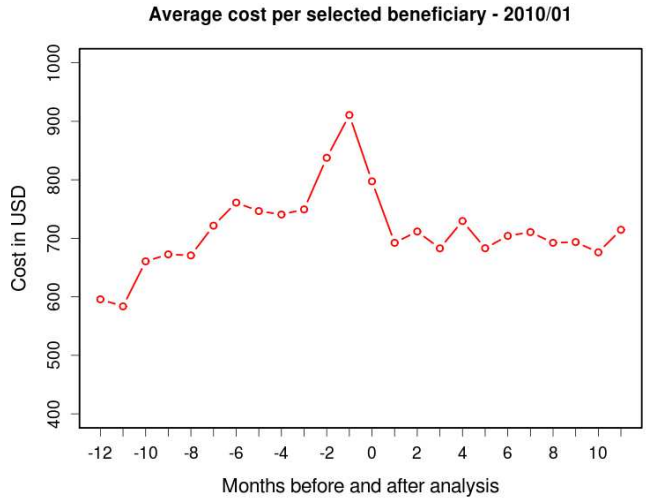


Fig. 6. Time evolution of average costs per beneficiary selected through the application of the risk classifier.

ACKNOWLEDGMENT

The authors would like to thank FUNCAP (Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico) for the financial support.

REFERENCES

[1] J. Li, A. Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks, and C. Kelman, "Mining risk patterns in medical data," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2005, pp. 770–775.

[2] H. Dréau, I. Colombet, P. Degoulet, G. Chatellier *et al.*, "Identification of patients at high cardiovascular risk: a critical appraisal of applicability of statistical risk prediction models," *Methods of Information in Medicine*, vol. 40, no. 1, pp. 6–11, 2001.

[3] H. Kaur and S. Wasan, "Empirical study on applications of data mining techniques in healthcare," *Journal of Computer Science*, vol. 2, no. 2, pp. 194–200, 2006.

[4] J. Berger, C. Jordan, D. Lloyd-Jones, and R. Blumenthal, "Screening for cardiovascular risk in asymptomatic patients," *Journal of the American College of Cardiology*, vol. 55, no. 12, pp. 1169–1177, 2010.

[5] K. Eagle, G. Ginsburg, K. Musunuru, W. Aird, R. Balaban, S. Bennett, R. Blumenthal, S. Coughlin, K. Davidson, E. Frohlich *et al.*, "Identifying patients at high risk of a cardiovascular event in the near future," *Circulation*, vol. 121, no. 12, pp. 1447–1454, 2010.

[6] B. Andersson, R. Andersson, M. Ohlsson, and J. Nilsson, "Prediction of severe acute pancreatitis at admission to hospital using artificial neural networks," *Pancreatology*, vol. 11, no. 3, pp. 328–335, 2011.

[7] T. Ayer, J. Chhatwal, O. Alagoz, C. Kahn Jr, R. Woods, and E. Burnside, "Comparison of logistic regression and artificial neural network models in breast cancer risk estimation1," *Radiographics*, vol. 30, no. 1, pp. 13–22, 2010.

[8] K. Spencer, L. Olson, N. Schnetz-Boutaud, P. Gallins, A. Agarwal, A. Iannaccone, S. Kritchevsky, M. Garcia, M. Nalls, A. Newman *et al.*, "Using genetic variation and environmental risk factor data to identify individuals at high risk for age-related macular degeneration," *PLOS ONE*, vol. 6, no. 3, 2011.

[9] I. G. Tsoulos, D. Gavrilis, and E. Glavas, "Neural network construction using grammatical evolution," in *Proceedings of the 5th IEEE International Symposium on Signal Processing and Information Technology*, 2005, pp. 827–831.

[10] I. Duncan, *Healthcare Risk Adjustment and Predictive Modeling*. ACTEX Publications, Inc., 2011.

[11] Brasil, "Hipertensão arterial sistêmica para o sistema único de saúde," 2006, Ministério da Saúde, Secretaria de Atenção à Saúde, Departamento de Atenção Básica.

[12] J. van Hulse, T. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, 2007, pp. 935–942.

[13] S. Haykin, *Neural networks and learning machines*, 3rd ed. Prentice Hall, 2009.