Intelligent decision support systems for clinical data patient analysis through data mining techniques

Emanuel Weitschek[1,2*], Paola Bertolazzi[2] and Giovanni Felici[2]

1. Department of Engineering
   Uninettuno International University
   Rome, Italy

2. Institute of Systems Analysis and Computer Science Antonio Ruberti
   National Research Council
   Rome, Italy

* corresponding author

Email addresses:

Emanuel Weitschek	emanuel.weitschek@uninettunouniversity.it
Paola Bertolazzi	paola.bertolazzi@iasi.cnr.it
Giovanni Felici	giovanni.felici@iasi.cnr.it

ABSTRACT

The high growth of clinical data sets originated from the introduction of electronic health records collected in medical environments. Nowadays, the computerization of clinical data is supported by many international projects and frequently national integrated clinical record systems are adopted in both public and private health-care facilities. The study of the diseases and the discovery of effective therapies demands collection, management, integration and analysis of clinical data. The main target is to obtain valuable knowledge from large sets of clinical data. Novel methods for their analysis are therefore required in order to extract relevant information and compact models. A candidate discipline is data mining, an interdisciplinary field that comprises computer science, statistics and artificial intelligence, and is dedicated to automate the knowledge discovery process. Data mining methods are able to consider different types of data - structured and unstructured - from disparate sources. The objectives of clinical data mining are to: recognize the clinical and laboratory variables that characterize a particular disease; deal with often incomplete (missing values) and noisy data sets (e.g. different measure scales); integrate patient samples from different sources; recognize variables name with the same meaning; integrate diverse patient data collection procedures. The management of clinical data, the discovery of patients interactions, and the integration of the disparate data sources are the hardest problems to solve. Finally, after an adequate handling of these issues, a compact and human understandable data model has to be extracted. In this work, consolidated data mining methods (e.g., artificial neural networks, decision trees, rule based classifiers) able to manage and analyze clinical data sets are introduced and applied to a real case study. Demented patient samples collected in different Italian health care facilities are investigated, providing a practical example of clinical data mining. Classification through artificial neural networks, logic rules, decision trees are considered and compared. It is shown that supervised classification is a promising technique to identify the disease of the patients and to extract significant models for biomedical knowledge discovery.

I. INTRODUCTION

Automatic knowledge extraction methods and tools are sound candidates to support medical doctors and scientist in the analysis of currently growing clinical patient data sets. The field of data mining, a discipline that comprises computer science, statistics, and artificial intelligence, is the right choice for dealing with this complex task. When analyzing clinical data sets, a supervised data mining methodology, i.e. classification, is suggested, as the considered patient samples are often previously assigned to a specific class or type by medical doctors [1], e.g. healthy and diseased. The focus in this work is therefore on supervised data mining (classification). Classification is the action of assigning an unknown object into a predefined class after examining its characteristics [2]. A great advantage is also to compute a clear human understandable classification model which fits the data, e.g. "if-then rules" (for example *if HachinskiScore > 4 then the patient sample is healthy*). The classification model can be a valid aid for the medical doctors and investigators to extract the clinical and laboratory variables associated to a disease and to formulate a diagnosis for new patients.
The field of clinical data mining spread in late nineties with the beginning of electronic patient health records collection [3]. The analysis of growing clinical patient data required efficient and effective methods. The authors in [1] provide a comprehensive review of clinical data mining by summarizing 84 papers in this domain. They highlight following main goals of this discipline:

- data understanding;
- healthcare professionals decision support, and
- definition of a methodology for medical data analysis.

In [4], an extended review of clinical data mining is provided, and supervised data mining (classification) is cited as the main technique for analyzing clinical patient sets and a framework for constructing, assessing and exploiting data mining models in clinical medicine is described. The authors of [5] illustrate the high growth of clinical data and the major issues in clinical data sets: inaccuracy, complexity and frequent missing values. Other common problems in clinical data are:

- diverse variables names in different data sets with the same meaning;
- different adopted scales of the variables (also in the same data set);
- outliers coming from wrong measures;
- mixed (numeric - textual) variables types, e.g. some medical doctor rate the value of Blood Urea Nitrogen (BUN) with numbers and some others with words (low, normal, high); when integrating the variable, mixed type variables, like BUN = {12, 16, high, 9, 11, …}, may be found.

In this work, methods to pre-process and classify clinical data are reported and a complete knowledge discovery process with different data mining algorithms on Italian demented clinical patients samples collected in diverse health care facilities is provided as an example to the reader. The paper is structured in the following way. The section *Methods* describes a complete automatic knowledge discovery process on clinical data, composed of following steps:

- data collection;
- data integration;
- data analysis.

The data analysis steps comprises preprocessing, missing values treatments, noise reduction, discretization, feature selection, classification, and model extraction. The section *Results and Discussion* provides to the reader a real world example of a knowledge discovery process (data set

collection, preprocessing, feature selection, classification and model extraction) applied to a clinical data set of demented patient collected in different Italian health care structures. The work ends with the *Conclusions*, where a reasoning towards a computer aided diagnosis with an intelligent decision support system for clinical data patient variables and sample analysis is performed.

II. METHODS

This section provides to the reader a clinical data analysis knowledge discovery process. The flow chart in figure 1 draws the clinical data mining process.

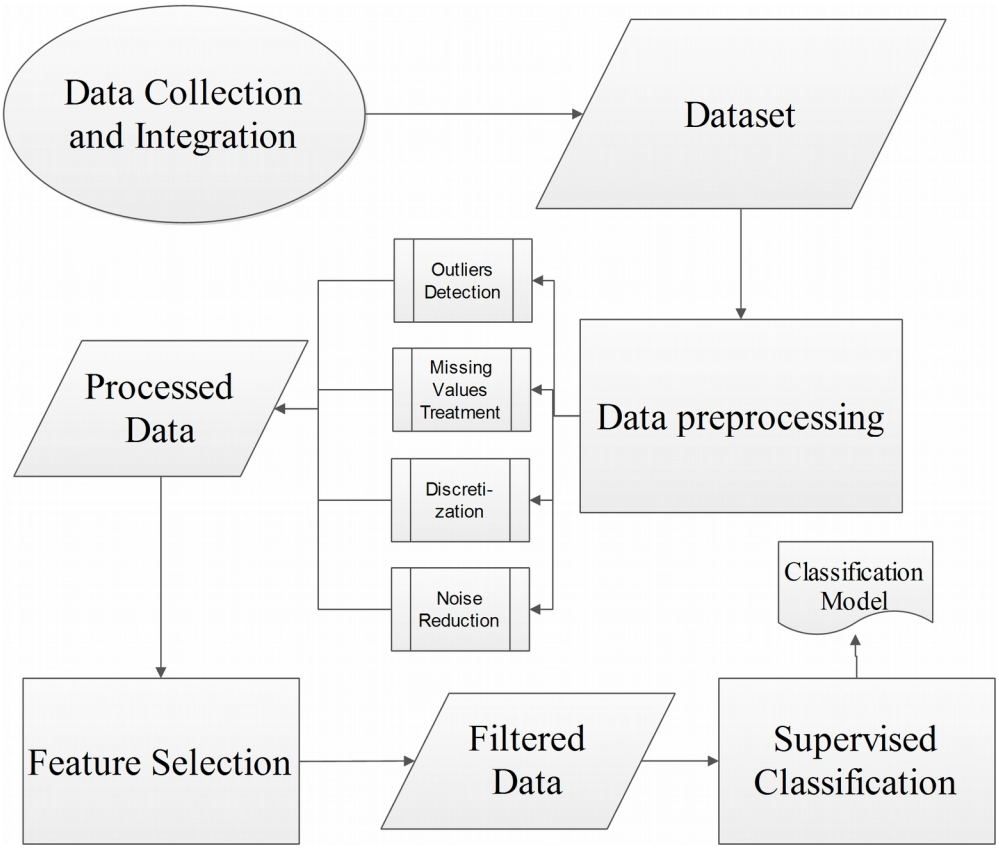

Fig. 1 Clinical data mining flow chart

A clinical data set is composed of a number of patient samples (records) each one associated with its clinical and laboratory variables. These data can be represented as in a matrix depicted in table I.

TABLE I CLINICAL DATA REPRESENTATION

Sample	*Class*	*variable*$_1$	*variable*$_2$	...	*variable*$_n$
Patient$_1$	Healthy	value$_{(1,1)}$	value$_{(1,2)}$	…	value$_{(1,n)}$
Patient$_2$	Diseased	value$_{(2,1)}$	value$_{(2,2)}$	…	value$_{(2,n)}$
...	…	…	…	…	…
Patient$_n$	Helathy	value$_{(n,1)}$	value$_{(n,2)}$	…	value$_{(n,n)}$

A. Data collection and integration

The first step in a clinical data mining study is the collection of the different electronic health patient records from the considered sources, e.g. hospitals, ambulatories, labs, etc. The data set is frequently composed from disparate health care facilities by non IT staff. The variable values and names are inserted manually in spreadsheets or copied from hard copy archives. Often the IT systems (if available) are not integrated and every health care facility uses his own data format. Clinical data sets are therefore very noisy, missing values and outliers are frequent due to wrong measures and mixed variables types (textual - numeric). The integration of this data is very challenging, but is often necessary for the analysis of a representative patient set associated to a disease in order to perform a valuable knowledge extraction. A usual integration technique is to join all the patient records on common variables names, considering all the non-common variables as complementary. A problem may occur when the variables names are different in the diverse sets. This problem may be solved with the creation of ontologies or clinical and laboratory variables names standardization, see for example the "Logical Observation Identifiers Names and Codes (LOINC): a universal code system for identifying laboratory and clinical observations" (http://loinc.org) and the Health Level Seven International organization (http://www.hl7.org). Last but not least in the clinical data domain the privacy issues are very important to respect. The patient confidential information, as names, has not to be shared among the analysts. So a private numeric identifier has to be assigned to the patient by the source structure and be stored in a non-accessible encrypted data base. The quality of data collection is a necessary, but not sufficient condition, for good analysis results. It is therefore suggested to plan and invest heavily on these steps, e.g. with integrated IT systems and adequate training for medical staff.

B. Data analysis: Preprocessing and missing values detection

Cleaning the clinical data with an adequate preprocessing step is always necessary when dealing with multiple sources. The knowledge extraction is aided by this important step. The investigator could start to apply the classical statistical measures on every variable present in the data set, e.g. mean, standard deviation, minimum, maximum, number of missing values, modal value and attribute type. According to these measures the analyst is able to filter the variables that present high noise.
It is recommended:

- to exclude the variables with a high number of missing values, e.g. more than 20%;
- to fill the remaining missing values with the class mean [7];
- to detect the outliers by mean - standard deviation - minimum - maximum values analysis and records removal;
- to correct the mixed attributes types.

Alternative methods for processing the missing values are reported in [6]. The methods are frequently tailored in replacing them with statistical artifacts:

- most common attribute value;
- assignment of all possible attribute values;
- removal of records with unknown attribute values;
- event-covering method;
- insertion of special values.

The authors in [7] suggest to fill the missing values of numeric variables with the class mean and of categorical with the modal value. The investigator needs to implement ad-hoc preprocessing programs and procedures fitted on the analyzed clinical data.

C. Data analysis: Discretization

The discretization is suggested by the presence of data of mixed nature (numerical, categorical, logical) or by the presence of numerical variables, when using tree based or logic based classification algorithms. Discretization is a procedure that converts numeric variables in categorical by the definition of intervals. The discretization step determines a number of cutpoints in the range of values of the considered variable and defines a set of intervals in which the variable can be. Clinical data sets are often large matrices of real and integer values, which represent the patients variables measures. These data is often not optimally accepted as input by some classification algorithms, like logic data mining and trees, and a conversion in the discrete domain is necessary to improve their performance. A binarization is suggested, i.e. transform the data set composed of real numbers in binary. Define cutpoints for each variable by intervals identification and convert it into a new discrete one. Apply then a binary conversion of the variables in order to let the data be treatable in a logic framework. Following methods are suggested for applying a discretization and binarization of the data: [8] and [9]. Other effective approaches are available on www.keel.es. Also the well-known data mining software Weka [10] provides a large collection of discretization procedures.

D. Data analysis: Noise reduction

After the preprocessing phase (II-B), that provides a major cleaning of the data, advanced noise reduction techniques could be necessary in very noisy clinical data sets. An effective noise reduction technique in logic frameworks is described in [11], where a logic data mining software is enriched with special cleaning procedures. These procedures operate at discretization (II-C), feature selection (II-E), and classification model extraction (II-F) levels. In addition, the authors in [12] provide an extensive list of noise removal techniques.

E. Data analysis: Feature selection

Feature selection aims at selecting the candidate variables that are able to distinguish the different classes in the data set. Another main goal is the reduction of the data by filtering out the irrelevant and redundant variables, and by extracting only high informative ones. These variables are then considered as input to the real knowledge extraction algorithm, that is so able to work better, faster and effectively [13]. The classification problem becomes tractable and only informative variables are considered to extract the model. The authors of [14], [15], [16] and [17] provide a comprehensive review of feature selection algorithms and describe the main computational steps. The importance of feature selection in clinical domains is highlighted, as the data sets are often

composed by a large number of variables (100 - 1000) associated to every patient sample that belongs to a given class.

F. Data analysis: Classification

Classification is the action of assigning an unknown object into a predefined class after examining its characteristics [2]. The classical example from medicine is distinguishing diseased patients from healthy ones. Automatic classification is known as supervised learning: Unknown objects are automatically assigned to a class by analyzing their attributes (variables) using a classification model computed from objects with a known class (training set). Each object is a record of the data set and is composed by a tuple (**x**, c), where x is the attribute set and c is the class. A formulation of the classification problem is reported in [7]. Given t objects $\{x_1, \ldots, x_t\}$ whose class is known, build a function and use it to classify new objects, whose class is unknown. This function is called also classifier and is defined as $c' : R^n \rightarrow \{1, ..., m\}$.. The goal is to have $c'(x) = c(x)$ and for each x, the predicted class should be the correct class. A classification function is learned from (or fitted to) a training data and applied to a test data. This process is called supervised learning. Systematic approaches able to build a classification model by analyzing the training set are called classification methods. Currently many classification methods are present in the literature; the most suited for clinical data mining are the following:

1) *Support Vector Machines*: Support Vector Machines (SVM) [18], [19] have high classification performance when classifying data sets composed of numeric variables. The classification task is considered a two class distinction problem (in case of more classes a one class versus all the other approach is performed) and the two classes are represented in two n dimensional vectors. The Support Vector Machine builds the best separating hyperplane, which maximizes the margin (defined as the minimum distance between the hyperplane and the closest point in each class) between the two vectors. The main feature of this method is its ability to efficiently project the data into a higher dimension with very rich non-linear transformations. SVM generally perform very well on clinical patient data, as the most attributes are numeric. The main disadvantage is that no clear and human understandable classification model is given as output to the analyst and so no real knowledge is transmitted.

2) *Nearest Neighbor classifiers*: Nearest Neighbor classifiers are classifiers that consider the closest similarity between the unknown object in the test set and the training set and perform so the class assignment. To compute the similarity they often use Euclidean distance. Given a training set with known class object, a new object is assigned to a class by computing the closest neighbors of the training set. Also in this case, no real descriptive model of the data is computed and only the classification assignments are provided to the analyst.

3) *Classification Trees*: Classification trees (or decision trees) are mathematical structures composed of nodes and edges. Each node represents a predicate, associated to the objects in the data set, each edge a binary decision. The class labels are on the leaves of the trees. A path from the root to the leave is a set of decision on the objects values that leads to a classification of an object. One of the best performing classification tree method is C4.5 [21], as it is based on the minimization of the class entropy between each node. Classification trees provide a clear human understandable model that is extracted by following the paths from the root to each leaves (classes).

4) *Logic Data Mining*: Logic Data Mining is often referred to as rule based classification or classification with logic formulas. The characteristic of this method is the extraction of logic formulas (”if-then rules”) in disjunctive or conjunctive normal form as classification model. The logic rules are evaluated on the objects to perform classification. The most powerful methods for logic data mining are DMB [14], [24], RIPPER [22], LSQUARE [23], LAD [25], RIDOR [26], and PART [27]. The evident advantage of rule based classification is the data model, in terms of logic formulas, that is compact and human understandable. This last fact is very valuable in clinical data

mining, as a human understandable classification model could aid the medical doctor to identify the variables values that are characteristic for a certain disease.

5) *Artificial Neural Networks:* The Artificial Neural Network is a computational model, which aims to simulate the human brain structure [28]. The biological neural system is composed by neurons, connected with other neurons. The links between neurons are called axons, which transmit nerve impulses form one neuron to another. A neuron is connected with an axon via dendrites. A synapse is the contact point between a dendrite and an axon. The human brain learns by changing the strength of the synaptic connection between neurons stimulated by an impulse. An artificial neural network is engineered following this model: it is composed of interconnected nodes and directed links. The simplest attempt to build an artificial neural network is the perceptron [29]. The perceptron is composed of two types of nodes: input nodes, which represent the input attributes, and an output node, which represents the model output. The nodes are the neurons. Each input node is connected via a weighted link to the output node. The weight is the strength of a synaptic connection between neurons. A perceptron calculates its output value by computing a weighted sum on its input, subtracting a noise factor t from the sum and then evaluating the sign of the result. An input node simply transmits the value to the outgoing link without performing any computation. The output node is a mathematical system that calculates the weighted sum of its inputs, subtracts the noise and then produces the output by applying the sign function to the resulting sum. The multilayer artificial neural networks allow the resolution of more advanced classification problems. The network has a more complex structure. It contains some intermediate layers between the input and output layers. These layers are called hidden layers and its nodes hidden nodes. In feed forward neural networks the nodes are only connected with nodes of the next layer. In recurrent neural networks the nodes may be connected also with nodes from the same layer. The network may also use different activation functions, like linear, sigmoid and hyperbolic tangent. These functions allow the resolution of non-linear classification problems, e.g. the XOR problem.

III. RESULTS AND DISCUSSION

The following section provides to the reader a real case of clinical data mining. A collection of classification algorithms and methods are used to process an integrated data set of patient records containing different medical experiments (variables). These variables are in different formats and type (real, integer, binary, categorical). Final goals of the experimentation is to define a work-flow and to extract a diagnostic model for the identification of dementia. The model should ideally be human understandable and able to distinguish successfully the various classes of patients present in the data set. A new diagnostic model, in terms of logic classification formulas, is found by analyzing the patients clinical variables of following classes: *Demented, Depressed, Mild Cognitive Impairment (MCI), Psicotic, Uncertain and Control*.

A. The data set

More than 4,000 patients' psychometric and blood tests, imaging and other clinical data collected in several Geriatrics and Alzheimer departments in Italy [30] compose the data set. The data was collected in the framework of the ReGAl project [31] in 36 health-care facilities from Italy under supervision of Perugia University. Every facility had its own collection procedures and formats and a prior quality control is performed on a central unity. The data set comprises demographic characteristics, medical history, pharmacological treatments, clinical and neurological examination, psychometric tests, laboratory blood tests, imaging (MRI and CT). The total number of patients is 4,728 each one with 722 variables. Approximately 30% of missing values are present in the variables. 367 numeric, 318 ordinal and 37 of mixed type variables (attributes) are present. The variables represent demographic characteristics, medical history, pharmacological treatments, clinical and neurological examination, psychometric tests, laboratory blood tests, imaging and

various other clinical patient observations. Six different classes are assigned to the patients by domain expert medical doctors: *Demented, Depressed, Mild Cognitive Impairment (MCI), Control, Psicotic and Uncertain*. Table II provides a summary of the data set composition and distribution.

TABLE IIIII THE REGAL CLINICAL DATA

class	*Control*	*Demented*	*Depressed*	*MCI*	*Psicotic*	*Uncertain*	*Total*
# of patients	365	3110	375	406	49	423	4728
# of variables	722	722	722	722	722	722	722
% missing values	30.41	29.45	37.13	25.07	31.77	31.48	30.88

B. *The data mining analysis*

Several data mining methods and systems are considered for analyzing the previously described clinical data set in order to extract disease characteristic variables and ad-hoc classification models. A supervised classification approach is adopted, as the classes of the patients are present in the data set provided by an accurate diagnosis from domain expert medical doctors. This data forms the training set. Considering the training set the data mining methods select candidate variables for patient distinction (feature selection) and, based on these, compute the classification model. This model is then evaluated on the test set that may be composed of object with unknown class or object of known class, the latter is used for verification of the classification performances of the model.

1) *Preprocessing phase*: All the preprocessing procedure described in section II-B are implemented in scripts and software to effectively clean the data set. The variables that present a missing values ratio greater than 20% are discarded, the rest of missing values are filled with the class mean (or modal value in case of categorical variables), and the outliers are detected and removed. Special noise reduction procedures, like [11], are additionally adopted when using the DMB data mining system.

2) *Feature selection phase*: Feature selection present in the DMB system and described in [14] is performed in order to extract the characteristic clinical and laboratory variables for every disease state present in the data set. DMB feature selection detects following representative features for the disease classes drawn in table III. These variables are confirmed by multiple runs of the feature selection step by repetitively changing the random sampling sets of the data (80%).

TABLE IV PATIENTS CLASSES CHARACTERISTIC VARIABLES

Control	**MCI**	**Dementia**	**Depression**
Albumin	Albumin	Albumin	Babcock_1
Anxiety_symptoms	Anxiety_symptoms	Babcock_1	CIRS_cognitive_psych
Azotemia	Azotemia	CIRS_cognitive psych	CIRS_hypertension_artery
Babcock_1	Babcock_1	CIRS_hypertension_artery	CIRS_skeletal_muscle
CIRS_cognitive psych	CIRS_cognitive psych	CIRS_skeletal_muscle	Draw_copying
Copy_drawing_corrected	Copy_drawing_corrected	Draw_copying	Delayed_recall_corr
Cortical_CT	Cortical_CT	Frontal_horn_hypodensity	Diffused_hypodensity
Delayed_Recall_corrected	Delayed_Recall_corrected	Lives_alone	FAS
ECG_patological	ECG_patological	No_drinks_since	Lives_alone
Lives_alone	Lives_alone	IADL_Total	Number_of_drinks
IADL_Total	IADL_Total	MMSE_Total	IADL_Total
Main_job	Main_job	NPI_Depression	Main_job

NPI_Depression Token_test	NPI_Depression Token_test Number_of_drinks	Years_Education	Marital_status MMSE_Total NPI_Depression Son_daughter Token_test

3) *Model extraction phase*: Final goal of the data mining analysis is to compute a reliable classification model of the patients, that is able to distinguish the various disease states present in the considered clinical data set. The model should contain the characteristic variables for each class of the patient samples. In noisy data sets this is a hard task and the model creation is more challenging. A relatively small number of variables is used to build the classification model. The patients belonging to the Demented class are the most easy to classify, Control patients are more difficult and classification is more time consuming and complex with MCI and Depressed subjects. Uncertainty in MCI and Depressed classes is due to different facts: loose criteria to define the MCI disease, compared to Dementia one; overlapping value intervals for similar variables in the MCI and Depressed classes [30]. Ruled based classifiers (RIPPER and DMB), classification trees (C4.5), classification functions (SVMs), nearest neighbor classifiers (KNN), and artificial neural networks (ANN) are considered and applied to classify the previously described clinical data set. Table IV shows the classification accuracy of the different data mining methods tested with 10-fold cross validation [7].

TABLE IV CLASSIFICATION RESULTS IN %

method	DMB	RIPPER	C4.5	SVM	KNN	ANN
settings	no settings	opt. run = 10	unpruned, minobj = 2	Polykernel= 1	k = 1	auto
correct %	87.36	86.86	88.49	90.31	78.27	90.45
discretization	yes	no	no	yes	yes	yes
model	yes	yes	yes	no	no	no

The results show that the systems were able to classify the data with a reasonable accuracy. The preprocessing steps play a key role in this process: without them the classification performances are very low (in the range of 60% of accuracy) and no clear classification model and patient distinction can be performed. When using SVM, KNN, and ANN as classification algorithms a discretization of the data is performed with the method of Fayyad et al. [32], these improves the correct classification rate in the range of 10%. When analyzing the classification accuracy of the methods it can be seen that ANN performs best, but it is not able to provide a human understandable model to the investigator. DMB, RIPPER (logic data mining) and C4.5 (classification tree) have a slightly inferior accuracy, but output a classification model in terms of logic classification formulas. The model is a collection of ”if-then rules”, which consider a subset of clinical variables, that identify each class in the data. Figure 3 shows an example of classification model for the Mild Cognitive Impairment (MCI) state.

```
(MMSE_Totale >= 24) and (IADL_Punteggio_2 >= 16) and (Dilatazione_SottoCorticale_AtrofiaTAC <= 0) and
(Fluidita_Verbale_Categorie_Corretto >= 12.25) and (etaprecisa >= 73.08) OR \\
(MMSE_Totale >= 24) and (IADL_Farmaci_1 >= 1) and (Hachinski_Score >= 4) and
(Fluidita_Verbale_Categorie_Corretto <= 10) and (Corni_Frontali_IpodensitaTAC <= 0) OR \\
(MMSE_Totale_Rettificato >= 22.5) and (Matrici_Corretto >= 34)  and (Assurdita_Test_Verbali <= 13) and
(GB >= 4.9) OR \\
(MMSE_Totale_Rettificato >= 23.200001) and (IADL_Farmaci_1 >= 1) and
(Dati_Test_Base_Valutazione_Disturbi_ID_Utente <= 3) and (regione <= 6) and (Forgetting >= 4) OR \\
(MMSE_Totale_Rettificato >= 22) and (IADL_Spostamenti_2 >= 4) and (Depressione_Psichiatrico <= 0) and
(regione <= 2) and (CIRS_Comorbilita_Complessa >= 1) OR \\
(MMSE_Totale >= 23) and (Token_Test_Corretto >= 30) and (Raven_Test_Corretto >= 215) and
(Familiarita_Demenza <= 0) and (Richiamo_Immediato_Corretto <= 34) and (CIRS_Vascolare_Linfatico <= 1)
```

Fig. 3 Example of MCI classification model

With supervised classification methods, that provide a human understandable model, the clinicians are able to highlight the characteristic variables for each disease present in the data set and to aid the medical doctors in a diagnosis - a potential new faster and cheaper diagnostic work-flow. Conversely, with the other classifiers, as ANN, the clinicians are able to perform a more accurate automatic diagnosis.

IV. CONCLUSIONS

This work delineated the main guidelines for a clinical data analysis, focusing on the particular data mining application of classification. Classifying patient samples collected in different health-care facilities is a challenging task, that may be solved with ad-hoc data processing programs and consolidated supervised machine learning methods. These methods have been extensively described providing a possible work-flow for clinical data analysis. A full and automatic knowledge discovery process was performed on a large data set of real clinical data collected in Italy. With a robust preprocessing and cleaning, reliable classification results and models have been achieved, providing to medical doctors and clinicians new insights of the different diseases present in the data set. As suggested in [33], supervised classification models could finally aid the medical staff to accelerate the diagnosis process (by analyzing only the extracted variables) and to formulate the diagnosis more effectively.

ACKNOWLEDGMENTS

This work is partially supported by the FLAGSHIP ”InterOmics” project (PB.P05) funded by the Italian MIUR and CNR institutions.

REFERENCES

[1] J. Iavindrasana, G. Cohen, A. Depeursinge, H. M¨uller, R. Meyer, and A. Geissbuhler, “Clinical data mining: a review,” Yearb Med Inform, vol. 2009, pp. 121–133, 2009.
[2] S. Dulli, S. Furini, and P. E., Data Mining. Springer, 2009.
[3] H. Koh and G. Tan, “Data mining applications in healthcare,” Journal of Healthcare Information Management, vol. 19, no. 2, pp. 64–72, 2010.
[4] R. Bellazzi and B. Zupan, “Predictive data mining in clinical medicine: current issues and guidelines,” international journal of medical informatics, vol. 77, no. 2, pp. 81–97, 2008.
[5] G. Hripcsak and D. J. Albers, “Next-generation phenotyping of electronic health records,” Journal of the American Medical Informatics Association, vol. 20, no. 1, pp. 117–121, 2013.
[6] J. W. Grzymala-Busse and M. Hu, “A comparison of several approaches to missing attribute values in data mining,” in Rough sets and current trends in computing. Springer, 2001, pp. 378–385.
[7] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison Wesley, 2005.

[8] E. Weitschek, G. Felici, and P. Bertolazzi, “Mala: A microarray clustering and classification software,” in 23rd International Workshop on DEXA, BIOKDD, 2012.
[9] L. A. Kurgan and K. J. Cios, “Caim discretization algorithm,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 145– 153, 2004.
[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278
[11] G. Felici and E. Weitschek, “Mining logic models in the presence of noisy data,” in ISAIM, 2012.
[12] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, “Enhancing data analysis with noise removal,” Knowledge and Data Engineering, IEEE Transactions on, vol. 18, no. 3, pp. 304–319, 2006.
[13] P. Bertolazzi, G. Felici, and G. Lancia, “Application of feature selection and classification to computational molecular biology,” in Biological Data Mining, S. L. e. J.K. Chen, Ed. Chapman & Hall, 2010, pp. 257–294.
[14] P. Bertolazzi, G. Felici, and E. Weitschek, “Learning to classify species with barcodes,” BMC Bioinformatics, vol. 10, no. S-14, p. 7, 2009.
[15] V. De Angelis, G. Felici, and G. Mancinelli, “Feature selection for data mining,” in Data Mining and Knowledge Discovery Apporaches Based on Rule Induction Techniques, T. E. e. Felici G., Ed. Massive Computing Series, Springer, 2006, pp. 227–252.
[16] T. J. Umpai and S. Aitken, “Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes,” BMC Bioinformatics, vol. 6, no. 1, pp. 1–11, 2005.
[17] M. Dash and H. Liu, “Feature selection for classification,” Intelligent Data Analysis, vol. 1, pp. 131–156, 1997.
[18] V. N. Vapnik, Statistical learning theory. Wiley, 1998.
[19] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.
[20] B. V. Dasarathy, Nearest neighbor (NN) norms: NN pattern classification techniques, Dasarathy, B. V., Ed. IEEE Computer Society Press, 1990.
[21] J. R. Quinlan, “Improved use of continuous attributes in C4.5,” Journal of Artificial Intelligence Research, vol. 4, pp. 77–90, 1996. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi= 10.1.1.46.3240
[22] W. W. Cohen, “Fast effective rule induction,” in In Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, 1995, pp. 115–123.
[23] G. Felici and K. Truemper, “A minsat approach for learning in logic domains,” INFORMS Journal on Computing, vol. 13, no. 3, pp. 1–17, 2002.
[24] E. Weitschek, A. Lo Presti, G. Drovandi, G. Felici, M. Ciccozzi, M. Ciotti, and P. Bertolazzi, “Human polyomaviruses identification by logic mining techniques,” BMC Virology Journal, vol. 58, no. 9, 2012.
[25] E. Boros, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik, “An implementation of logical analysis of data,” Rutgers University, NJ, Tech. Rep. 29-96, 1996.
[26] B. R. Gaines and P. Compton, “Induction of ripple-down rules applied to modeling large databases,” Journal of Intelligent Information Systems, 1995.
[27] E. Frank and I. H. Witten, “Generating accurate rule sets without global optimization,” in In: Proc. of the 15th Int. Conference on Machine Learning. Morgan Kaufmann, 1998.
[28] J. E. Dayhoff and J. M. DeLeo, “Artificial neural networks: opening the black box,” Cancer, 91(8):1615–1635, 2001.
[29] B. Widrow and M. A. Lehr, “30 years of adaptive neural networks: Perceptron, madaline, and backpropagation,” Proceedings of the IEEE, 78(9):1415–1442, 1990.

[30] I. Arisi, M. D'Onofrio, R. Brandi, A. Cattaneo, G. Drovandi, G. Felici, E. Weitschek, P. Bertolazzi, S. Brancorsini, S. Ercolani, F. Mangialasche, and P. Mecocci, "New diagnostic model for the early diagnosis of alzheimer's disease and other dementias, based on logic mining of clinical variables," in 10th International Conference on Alzheimer's and Parkinson's Diseases, 2011.
[31] E. Mariani, R. Monastero, S. Ercolani, P. Rinaldi, F. Mangialasche, E. Costanzi, D.F. Vitale, U. Senin, P. Mecocci ; ReGAl Study Group. Influence of comorbidity and cognitive status on instrumental activities of daily living in amnestic mild cognitive impairment: results from the ReGAl project. Int J Geriatr Psychiatry, vol.23(5):523-30, 2008.
[32] U. Fayyad and K. Irani, "Multi-interval discretization of continuous valued attributes for classification learning," in Proceedings of the International Joint Conference on Uncertainty in AI., 1993.
[33] E. Weitschek, G. Felici, and P. Bertolazzi, "Clinical data mining: problems, pitfalls and solutions," in IEEE DEXA Workshop on BIOKDD, 2013.