# Intelligent data analysis for medical diagnosis: Using machine learning and temporal abstraction

Nada Lavrač*, Igor Kononenko**,
Elpida Keravnou***, Matjaž Kukar** and
Blaž Zupan*
* *J. Stefan Institute, Jamova 39, 1001 Ljubljana, Slovenia*

** *Faculty of Computer and Information Science, Tržaška 25, 1000 Ljubljana, Slovenia*

*** *Department of Computer Science, University of Cyprus, Kallipolenos 75, CY-1678 Nicosia, Cyprus*

Extensive amounts of knowledge and data stored in medical databases request the development of specialized tools for storing and accessing of data, data analysis, and effective use of stored knowledge and data. This paper focuses on methods and tools for intelligent data analysis, aimed at narrowing the increasing gap between data gathering and data comprehension. The paper sketches the history of research that led to the development of current intelligent data analysis techniques, discusses the need for intelligent data analysis in medicine, and proposes a classification of intelligent data analysis methods. The main scope of the paper are machine learning and temporal abstraction methods and their application in medical diagnosis. A selection of methods and diagnostic domains is presented, and the performance or viability of approaches discussed. The paper concludes with the evaluation of selected intelligent data analysis methods and their applicability in medical diagnosis.

Keywords: intelligent data analysis, machine learning, temporal abstraction, medical applications, medical diagnosis

## 1. Introduction

"Now that we have gathered so much data, what do we do with it?" This is the opening statement of the editorial by Usama Fayyad and Ramasamy Uthurusamy in the Communications of the ACM, Special issue on Data Mining [17]. Recently, many statements of this kind appeared in journals, conference proceedings, and other materials that deal with data analysis, knowledge discovery, and machine learning. They all express a concern about how to "make sense" from the large volumes of data being generated and stored in almost all fields of human activity.

Especially in the last few years, the digital revolution provided relatively inexpensive and available means to collect and store the data. For example in medicine, still in the mid-nineties one of the fathers of Artificial Intelligence in Medicine, Edward H. Shortliffe, partially blamed the underdeveloped hospital infrastructure for the failure to fulfill the initial promise of the field [81]. Recently, however, the situation is changing rapidly: modern hospitals are well equipped with monitoring and other data collection devices, and data is gathered and shared in inter- and intra-hospital information systems. In fact, medical informatics has become a must and an integral part of every successful medical institution [83].

The increase in data volume causes greater difficulties in extracting useful information for decision support. The traditional manual data analysis has become insufficient, and methods for efficient computer-based analysis indispensable. From this need, a new interdisciplinary field of *Knowledge discovery in databases* (KDD) was born [21]. KDD encompasses statistical, pattern recognition, machine learning, and visualization tools to support the analysis of data and discovery of principles that are encoded within the data.

KDD is frequently defined as a *process* [18] consisting of the following steps: understanding the domain, forming the dataset and cleaning the data, extracting of regularities hidden in the data thus formulating knowledge in the form of patterns, rules, etc. (this step in the overall KDD process is usually referred to as *data mining* (DM)), postprocessing of discovered knowledge, and exploitation of the results.


ISSN 0921-7126 / $8.00 © , IOS Press

In this paper we use the term *intelligent data analysis* (IDA) rather than KDD, despite the fact that it is hard to make the distinction between the two. IDA and KDD have in common the topic of investigation, which is interactive and iterative process of data analysis, and they share many common methods. A possible distinguishing feature is that the methodologies and techniques used in IDA are mostly (but not exclusively) knowledge-based (and therefore "intelligent" in the sense used in Artificial Intelligence): they either use the knowledge about the problem domain, of the underlying principles or the knowledge about the data analysis process itself. Another aspect involves the size of data: KDD is typically concerned with the extraction of knowledge from very large datasets, whereas in IDA this is not necessarily the case. This also affects the type of data mining tools used: in KDD the data mining tools are executed mostly in batch mode (despite the fact that the entire KDD process is interactive), whereas in IDA the tools can either be batch or applied as interactive assistants.

Specific goals of applying IDA in medicine are:

- the extraction (discovery) of medical knowledge for diagnostic, screening, prognostic, monitoring, therapy support or overall patient management tasks (i.e., *data mining*), and
- the intelligent interpretation of patient data in a context-sensitive manner and the presentation of such interpretations in a visual or symbolic form (i.e., *data abstraction*); the temporal dimension in the representation and intelligent interpretation of patient data is of primary importance.

As any other research in medicine is aimed at directly or indirectly enhancing the provision of health care, IDA research in medicine is no exception. As such, the benchmark tests for these methods and techniques can only be real world problems. Viable IDA proposals for medicine must be accompanied by detailed requirements that delineate the spectrum of real applications addressed by such proposals; in-depth evaluation of resulting systems thus constitutes a critical aspect.

Another consideration is the role of IDA systems in a clinical setting. Their role is clearly that of an intelligent assistant that tries to bridge the gap between data gathering and data comprehension, in order to enable the physician to perform his task more efficiently and effectively. If the physician has at his disposal the right information at the right time, doubtless he will be in a better position to reach correct decisions or instigate correct actions within the given time constraints. The information revolution made it possible to collect and store large volumes of data from diverse sources on electronic media. These data can be on a single case (e.g., one patient) or on multiple cases. Raw data as such are of little value since their sheer volume and/or the very specific level at which they are expressed make its utilization (operationalization) in the context of problem solving impossible. However such data can be converted to a mine of information wealth if the real gems of information are gleaned out by computationally intelligent means. The useful, operational information/knowledge, which is expressed at the right level of abstraction, is then readily available to support the decision making of the physician in managing a patient.

Important issues that arise from the rapidly emerging globality of data and information are:

- the provision of standards in terminology, vocabularies and formats to support multi-linguality and sharing of data,
- standards for the abstraction and visualization of data,
- standards for interfaces between different sources of data,
- seamless integration of heterogeneous data; images and signals are important types of data,
- standards for electronic patient records, and
- reusability of data, knowledge, and tools.

Clinical data constitute an invaluable resource, the proper utilization of which impinges directly on the essential aim of health care which is "correct patient management". Investing in the development of appropriate IDA methods, techniques and tools for the analysis of clinical data is thoroughly justified and this research ought to form a main thrust of activity by the relevant research communities.

Numerous intelligent data analysis methods have already been applied for supporting decision making in medicine (e.g., see [51]). These methods can be classified into two main categories: *data mining* and *data abstraction*. The majority of data mining IDA methods belong to *machine learning* and the majority of data abstraction methods perform *temporal abstraction*. This is the reason for machine learning and temporal abstraction to be the focus of investigation in this paper.

The structure of this paper is as follows. In Section 2, the paper first sketches the history of research (Section 2.1) that led to the development of current intelligent data analysis techniques. It then discusses the need for intelligent data analysis in medicine (Section 2.2), proposes a classification of intelligent data analysis methods (Section 2.3), and outlines some related characteristics of medical diagnosis problems (Section 2.4). The main scope of the paper are machine learning and temporal abstraction methods, and their application in medical diagnosis. A selection of methods and diagnostic domains is presented, and the performance or viability of such approaches is discussed in Sections 3 and 4. In Section 6 the paper concludes with the evaluation of selected intelligent data analysis methods and their applicability in medical diagnosis.

## 2. Intelligent data analysis in medicine

### *2.1. Knowledge versus data: A historical sketch*

In late seventies and early eighties, AI in medicine was mainly concerned with the development of medical expert systems aimed at supporting diagnostic decision making in specialized medical domains. Shortliffe's MYCIN [80], representing pioneering work in this area, was followed by numerous other efforts leading to specialized diagnostic and prognostic expert systems, e.g., HODGKINS [75], PIP [62, 84], CASNET [88], HEADMED [23], PUFF [46], CENTAUR [1], VM [16], ONCOCIN [82], ABEL [61], GALEN [85] MDX [9], and many others. The most general and elaborate systems were developed for supporting diagnosis in internal medicine [64, 58, 57]: INTERNIST-1 and its follower CADUCEUS, which, in addition to expert-defined rules as used in INTERNIST-1, included also a network of patophysiological states representing "deep" causal knowledge about the problem. The main problems addressed at this early stage of expert system research concerned knowledge acquisition [14, 15], knowledge representation, reasoning and explanation [86]. A typical early expert system schema is shown in Fig. 1.

Rules were proposed from the early days of knowledge-based systems, and expert systems in particular, as a prime formalism for expressing knowledge in a symbolic way. Rules have the undisputed advantages of simplicity, uniformity, transparency, and ease of inference, that over the years have made them one of the most widely adopted approaches for representing real world knowledge. Rules elicited directly from domain experts are expressed at the right level of abstraction from the perspective of the expert, and are indeed comprehensible to the expert since they are formulations of his rules of thumb. However, human-defined rules risk capturing the biases of one expert, and although each rule individually may appear to form a coherent, modular chunk of knowledge, the analysis of rules as an integral whole can reveal inconsistencies, gaps, and various other deficiencies due to their largely flat organization (i.e., the lack of a comprehensive, global, hierarchical organization of the rules).

It soon became clear that knowledge acquisition is the hardest part of the expert system development task. This was identified as the so-called "Feigenbaum bottleneck" [19, 20] in the construction of a knowledge base. The knowledge base is the heart of an expert system. For the effective use of expert system technology a knowledge base needs to be consistent and as complete as possible, throughout its deployment; to attain these desirable characteristics, both manual knowledge maintenance should be facilitated and the system should be able to evolve on the basis of its problem solving experience. The limitations of the first generation of expert systems [36, 49] coupled with the relatively high costs (in human and other terms) involved in acquiring knowledge directly from the experts, as well as the fact that databases of example cases started becoming readily available, made the learning of rules from such data especially appealing as a more efficient, less biased, and more cost-effective approach. On the one hand, this led to the developments in the area of machine learning (as described below), and on the other hand, to the investigations of the use of deep causal knowledge that could potentially overcome the difficulties encountered when using unstructured shallow-level sets of rules [37, 28]. An early approach to combining the use of deep knowledge and machine learning was used in the development of KARDIO, a system for ECG diagnosis of cardiac arrhythmias [5].

In late eighties and early nineties it thus became apparent that knowledge acquired from experts alone

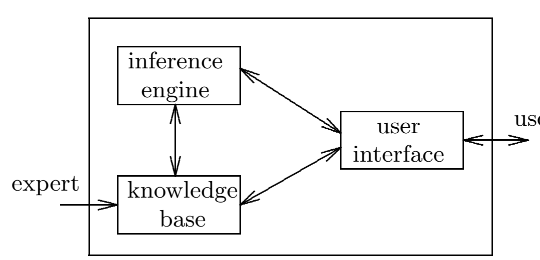


Fig. 1. An expert system schema of early '80s.

is unsuitable for solving difficult problems and that, when developing decision support systems, the analysis of data gathered in the daily practice of experts and stored systematically in databases can play an important role for the decision making support. This led to the development of early machine learning algorithms [52, 65] aimed at the automatic extraction of rules or decision trees from data. Early machine learning systems, aimed at dealing with real-world data which may be erroneous (noisy) and incomplete, include CART [6], Quinlan's extensions to ID3 [66], ASSISTANT [4, 8], AQ [53], and CN2 [12, 11]; C4.5 [68] is an efficient and probably the most popular machine learning system of the nineties.

Machine learning approaches do not advocate the bypassing of experts. Far from so. Experts are actively involved, but in a different and more constructive way than in the development of early expert systems. The example cases come from the experts and the resulting rules are validated by the experts for comprehensibility and other desired qualities. The learning approaches ensure that the derived rules are consistent, hierarchically organized (for example in terms of a decision tree), and, assuming that the collection of case examples used provides an adequate coverage of the particular domain, the resulting set of rules will be of sufficient accuracy and adequate coverage (i.e., without significant gaps of knowledge). Furthermore, the expert provides important background knowledge for focusing and guiding the learning of rules. Irrespective of whether rules are learned or directly acquired from experts, their format should be simple, intuitive, and adequately expressive for the purposes of the particular application.

The nineties are characterized by the increasing gap between the excessive storage of uninterpreted data and the understanding of the data, and the need to overcome this gap by the effective use of data analysis techniques. The main emphasis of current research is thus on data analysis. This led to the challenging new research areas of knowledge discovery in databases [21], data mining, and intelligent data analysis, in which machine learning techniques have a major role when the goal of data analysis is knowledge extraction. Current machine learning research is characterized by a shift of emphasis towards relational learning (ILP, [59, 50]) and more elaborate statistics applied in learning and evaluation methodologies. In data analysis, another trend is towards data abstraction and, in particular, towards temporal data abstraction [30] that can be viewed as a form of preprocessing for further data analysis. In the late nineties, data analysis has an increased role also due to the fact that data gathering is becoming distributed (e.g., telemedicine [2]), and that the analysis of such data is even more demanding. Fig. 2 shows a possible schema of a decision support system of the nineties, where decision support needs to deal also with large volumes of data, as well as data gathering and analysis via the Internet and an intranet (see also the account by [3]).

### *2.2. The need for IDA in medicine*

The gap between data generation and data comprehension is widening in all fields of human activity. In medicine, overcoming this gap is particularly crucial since medical decision making needs to be supported by arguments based on basic medical knowledge as well as knowledge, regularities and trends extracted from data.

There are two main aspects that define the significance of and the need for intelligent data analysis in medicine:

- The first important aspect concerns the discovery of new medical knowledge that can be extracted through data mining of representative collections of example cases, described by symbolic or numeric descriptors. The available datasets are often incomplete (missing data) and noisy (erroneous). The methods for extracting meaningful and understandable symbolic knowledge will be referred to as *data mining methods*. The quality assessment of these methods is based both on the performance (classification and prediction accuracy, misclassification cost, sensitivity, specificity, etc.), as well as the understandability and significance of the discovered knowledge.
- The second aspect concerns the support of specific knowledge-based problem solving activities (diagnosis, prognosis, monitoring, etc.) through the intelligent analysis of individual patients' raw data, e.g., a time series of data collected in monitoring. Data are mostly numeric and often quite noisy and incomplete. The aim is to glean out, in a dynamic fashion, useful abstractions (e.g., summaries) on the patient's (past, current, and hypothesized future) situation which can be matched against the relevant (diagnostic, prognostic, monitoring, etc.) knowledge for the purposes of the particular problem solving activity. Such data analysis methods are referred to as *data abstrac-*

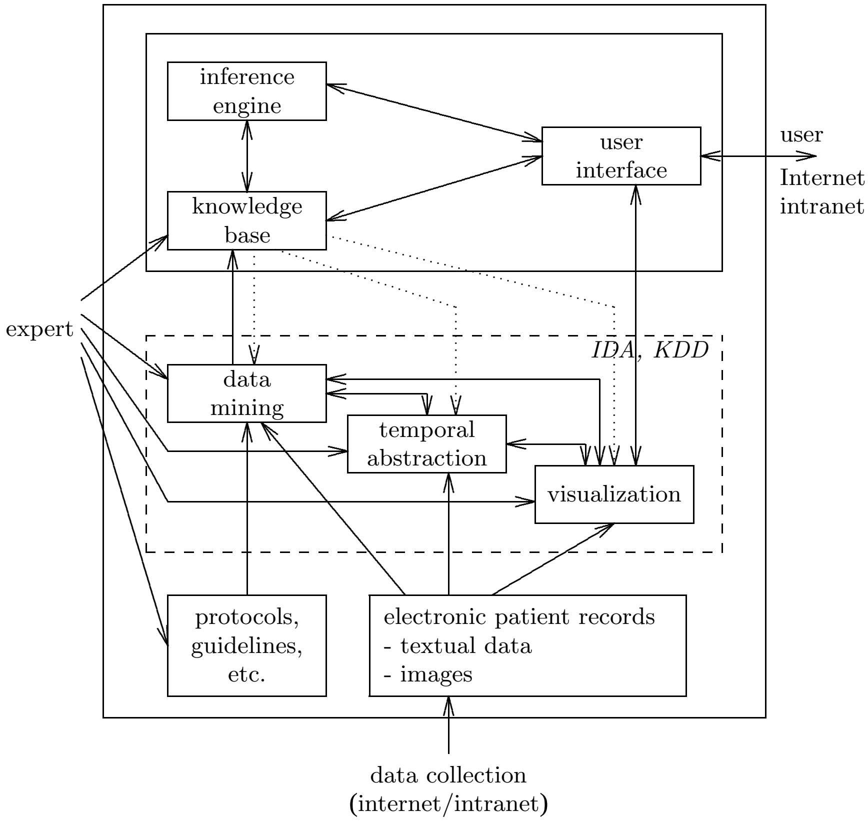


Fig. 2. A decision support system schema of late '90s.

*tion methods*, a term originally coined by Clancey in his now classical proposal on heuristic classification [10], where these methods form an integral part of the reasoning process. Recently, data abstraction methods have been concerned with the interpretation of temporal data (*temporal data abstraction*), where temporal trends and more complex temporal patterns constitute main types of such abstractions. Since the primary goal of (temporal) data abstraction methods is on-line decision support, their quality assessment is performance-based: for instance, does a method provide adequate support for diagnostic and prognostic reasoning, does it predict well a trend or a value to be expected at the next point in time? In this respect, *visualization of data* is extremely important for supporting decision making and even invaluable for successfully performing a problem solving task.

Since the goal of data abstraction is to describe the data in more abstract terms, it can also be used in the preprocessing of data for further analysis by data mining techniques and tools.

### 2.3. A classification of IDA methods

Based on the main aspects of the use of IDA methods in medicine discussed in Section 2.2, we propose the following classification of IDA methods:

– *Data mining methods* are intended to extract knowledge preferably in a meaningful and understandable symbolic form. Most frequently applied methods in this context are supervised symbolic learning methods. For example, effective tools for inductive learning exist that can be used to generate understandable diagnostic and prognostic rules. Symbolic clustering, discovery of concept hierarchies, qualitative model discovery, and learning of probabilistic causal networks fit in this framework as well. Sub-symbolic learning and case-based reasoning methods can also be classified in the data mining category. Other frequently applied sub-symbolic methods are nearest neighbor, Bayesian classifier, and (non-symbolic) clustering.

– *Data abstraction methods* are intended to support specific knowledge-based problem solving activities (data interpretation, diagnosis, prognosis, monitoring, etc.) by gleaning out the useful abstractions from the raw, mostly numeric data. *Temporal data abstraction methods* represent an important subgroup where the processed data are temporal. The derivation of abstractions is often done in a context sensitive and/or distributed manner and it applies to discrete and continuous supplies of data. Useful types of temporal abstractions are trends, periodic happenings, and other forms of temporal patterns. Temporal abstractions can also be discovered by visualization. The abstraction can be performed over a single case (e.g., a single patient) or over a collection of cases.

The scope of this paper regarding data mining is *machine learning methods*, with the emphasis on symbolic concept learning and Bayesian classification. In data abstraction, the scope of the paper is limited to *temporal abstraction methods*.

### *2.4. IDA for medical diagnosis*

A typical diagnostic process is the following. In an interview the patient's anamnestic data is obtained and after the preliminary examination of the patient the physician records the status data. Depending on the anamnestic and the status data, the patient takes additional laboratory examinations. The diagnosis is then determined by the physician who takes into account the whole available description of the patient's state of health. Depending on the diagnosis the treatment is prescribed and after the treatment the whole process may be repeated. In each iteration the diagnosis may be confirmed, refined, or rejected. The definition of the final diagnosis depends on the medical problem. In some problems the first diagnosis is also the final, in some others the final diagnosis is determined after the results of the treatment are available, and in some problems there is no way to obtain a completely reliable final diagnosis. For example, in the problem of the localization of the primary tumor the final diagnosis can always be obtained with an operation where the location of the primary tumor is verified, although this "examination" is avoided and replaced with other laboratory tests unless it is really necessary to obtain the verified diagnosis. And in urology, in the problem of diagnosing the type of incontinence, in practice the final diagnosis is never obtained as there is no practical way to verify it.

Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician, his intuition and biases, and even on the psycho-physiological condition of the physician. Several studies have shown that the diagnosis of a patient can differ significantly if the patient is examined by different physicians or even by the same physician at different times (different day of the week or different hour of the day).

#### *2.4.1. Machine learning*

*Machine learning methods* can be used to automatically derive diagnostic rules from the descriptions of the patients treated in the past for which the final diagnoses were verified. Automatically derived diagnostic knowledge may assist physicians to make the diagnostic process more objective and more reliable.

Typically, automatically generated diagnostic rules slightly outperform the diagnostic accuracy of physician specialists when physicians have available exactly the same information as the machine. Table 1 provides a comparison of the performance of two machine learning algorithms, the naive Bayesian classifier and a decision tree induction algorithm Assistant [8], with the average performance of four physician specialists in three different medical diagnostic problems: the localization of the primary tumor (PRIM), the diagnosis of thyroid diseases (THYR), and rheumatology (RHEU).

Table 1
The comparison of performance of different classifiers in three medical domains.

| Classifier | PRIM | THYR | RHEU |
|---|---|---|---|
| naive Bayes | 49% 1.60bit | 70% 0.79bit | 67% 0.52bit |
| Assistant | 44% 1.38bit | 73% 0.87bit | 61% 0.46bit |
| physicians | 42% 1.22bit | 64% 0.59bit | 56% 0.26bit |

The following are the brief descriptions of the diagnostic problems (see also [45]).

– **Localization of primary tumor:** The medical treatment of patients with metastases is much more successful if the location of the primary tumor in the body of the patient is known. The diagnostic task is to determine one of 22 possible locations of the primary tumor on the basis of age, sex,

histological type of carcinoma, the degree of differentiation and 13 possible locations of discovered metastases. The data set of 339 patients with known locations of the primary tumor was provided for our experiments by the Institute of Oncology in Ljubljana.
- **Thyroid diseases:** The diagnostic problem is to determine one of the four possible diagnoses from age, sex, histological data, and results of laboratory tests. However, in everyday practice physicians use much more additional information for diagnostics, which was not available for computer processing. The data set of 884 patients with known final diagnoses was provided for our experiments by the Clinic for Nuclear Medicine of the University Medical Center, Ljubljana.
- **Rheumatology:** The diagnostic problem is to select one of the six groups of possible diagnoses from anamnestic data and status data. There is over two hundred diagnoses used by physicians specialists in rheumatology. However, general practitioners have to decide among rheumatological and orthopedical diseases for patients to be further investigated and treated by specialists. Such decisions are unreliable and, by the opinion of the physician specialist in rheumatology, in more than 30% of cases wrong. The data set of 355 patients with known final diagnoses was provided for our experiments by the Clinic for Rheumatology of the University Medical Center, Ljubljana. All diagnoses were verified with additional observations, laboratory tests, and X-raying.

Detailed characteristics of the data sets used in this experiment are summarized in Table 2. The entropy ($Ent$, measured in bits) together with the number of classes ($Cl$) shows the difficulty of the diagnostic problem. The number of attributes ($Att$) approximately tells how well the patients are described. The majority class ($MC$, given by the percentage of cases belonging to this class) approximates the prior probability of the most probable diagnosis. This is in fact the classification accuracy of the *default* classifier which, regardless of the patient, always selects the same most probable diagnosis.

In our experiments one run was performed by randomly selecting 70% of instances for learning and 30% for testing. The results in Table 1 are averages of 10 runs. The average accuracy is given along with the average *information score per answer* [43]. Information score is a performance measure that eliminates the influence of prior probabilities of classes and can be applied to various kinds of incomplete and probabilistic answers. This measure is necessary as in each domain the default classifier would achieve high classification accuracy.

Table 2
Basic description of medical data sets.

| Domain | Cl | Att | Val/Att | Ins | MC | Ent |
|---|---|---|---|---|---|---|
| PRIM | 22 | 17 | 2.2 | 339 | 25% | 3.89 bit |
| THYR | 4 | 15 | 9.1 | 884 | 56% | 1.59 bit |
| RHEU | 6 | 32 | 9.1 | 355 | 66% | 1.73 bit |

Four physician specialists in each domain were tested to estimate their diagnostic accuracy. From a set of training data, a subset of patients was randomly selected and their description printed on paper without the diagnosis. The physicians were asked to select the most probable diagnosis for each patient. The performances of physicians in Table 1 are the averages of four physician specialists in each domain. The physicians were tested at the University Medical Center in Ljubljana. While in rheumatology, diagnosing a patient on paper is somewhat unnatural, for the other two domains it often occurs in practice.

Both algorithms significantly outperform the diagnostic performance of the physicians in terms of the classification accuracy and the average information score of the classifier. However, these results need a qualification. It should be emphasized that in these experiments both the physicians and the computer had available exactly the same information. This is often unrealistic in medical practice. During the examination of the patient the physician often observes the patient's condition in terms of intuitive impressions which cannot be formally described and therefore cannot be typed in the computer. The lack of such information may be in some cases of crucial importance for the (in)ability to obtain more reliable diagnosis. The accuracy results in Table 1 should therefore be understood as an estimate of how well the algorithms perform, and not necessarily how badly the physicians diagnose. Although machine learning may induce more reliable diagnostic algorithms from the limited description of the patient, such diagnostic tools definitely cannot, and also are not intended to, replace the physicians, but should be rather considered as helpful tools that can improve the physicians' performance. The results in this section and from other experiments convincingly demonstrate that physicians' diagnostic accuracy could be improved with the aid of machine learning.

#### 2.4.2. *Temporal data abstraction*

Time is intrinsically relevant to many medical diagnostic domains. Disease processes evolve in time, patient records give the history of patients, and therapeutic actions, like all actions, are indescribable void of time. In diagnostic systems for such domains, time should be explicitly represented in an integral fashion and reasoned with. The modeling of time enables a more accurate formation of potential diagnostic solutions (e.g., the presence of an abnormality may not be diagnostically significant as such, but its specific pattern of appearance is) and a more accurate evaluation of the entertained solutions (e.g., the expected picture of a disease is different depending on the state of its evolution). Below we give a concrete example of the significance of time in medical diagnostic reasoning, from the domain of skeletal dysplasias and malformation syndromes. These are developmental disorders which affect the skeletal system to varying degrees. A simplified description of the skeletal dysplasia Spondylo-epiphyseal Dysplasia Congenita (SEDC) reads as follows:

SEDC *presents from birth* and can be lethal. It *persists throughout the lifetime* of the patient. People suffering from SEDC can exhibit the following: short stature, due to short limbs, *from birth*; mild platyspondyly *from birth*; absence of the ossification of knee epiphyses *at birth*; bilateral severe coxa-vara *from birth, worsening with age*; scoliosis, *worsening with age*; wide triradiate cartilage *up to about the age of 11 years*; pear-shaped vertebral-bodies *under the age of 15 years*; variable-size vertebral-bodies *up to the age of 1 year*; and *retarded ossification* of the cervical spine, epiphyses, and pubic bones.

The text given in italic font refers to time, directly or indirectly. The references to time are absolute, where occurrences are specified with respect to some (generic) fixed time-point, which here is birth, and absolute durations are explicitly or implicitly specified. For example property “SEDC present”, in the context of some patient, persists throughout the span of that person’s lifetime, whatever that might be. Since SEDC can be lethal this duration could be zero (events birth and death coincide). The occurrences (and hence durations) of properties “wide triradiate cartilage” and “pear-shaped vertebral-bodies”, at the granularity of years, are approximated through the qualitative expressions “up to about the age of ..” and “under the age of ..” respectively. We refer to this characteristic as *absolute vagueness* [29, 30]. Some of the SEDC manifestations express abnormalities, namely *retarded*, with respect to some ossification process. Other manifestations express temporal trends, namely the *worsening* of properties “scoliosis” and “bilateral, severe, coxa-vara”.

The above description of SEDC gives the overall model for this disorder. Such a model need to be (temporally) adapted to the case under consideration. For example SEDC presents a different picture for an one year old, a twelve year old, or a seventeen year old.

The diagnostic task for this domain is [35]: Given a set of patient data determine which skeletal dysplasia or malformation syndrome is the best explanation of the patient situation. Patient data are largely obtained from radiographs that give discrete snapshots of the development of the patient’s skeleton. For example consider the following data on some patient for whom the available radiographs were for the pelvis and the lateral spine at the ages of 2 and 7 years old and for the hands and the lateral skull at the age of 10 years:

Carpal-bones small *at the age of 10 years*; femoral-capital-epiphyses abnormal *at the age of 2 years*; femoral-capital-epiphyses flat and irregular *at the age of 7 years*; vertebral-end-plates irregular *at the age of 7 years*.

The patient information is point-based in contrast to the medical knowledge which is largely interval-based. In this domain patient information tends to be grossly temporally incomplete. A competent, knowledge-based, diagnostic system must be able to process the available patient data in an intelligent way. This usually entails an ability to derive abstractions from the given information, which fill in the gaps and can be directly matched against the model of a disorder, for a patient of that age.

Abstractions for which time plays a central role are called *temporal abstractions*. For example temporal reasoning is central in establishing the existence of some delay or prematurity in the unfolding of some ossification process, or the existence of some trend. Temporal data abstraction is presently attracting considerable research interest [22, 27, 33, 48, 56, 60, 74, 76–79], as a fundamental intermediate reasoning process for the intelligent interpretation of temporal data in support of tasks such as diagnosis, monitoring, etc. Background domain knowledge [34] can be effectively utilized in the context of temporal data abstraction. In the domain of skeletal dysplasias, knowledge on the normal ‘behavior’ of ossification processes constitutes significant background domain knowledge. A piece of this knowledge is given below:

The ossification process of the cervical-spine begins at the eighth *gestation week* and terminates at the 25th *year* of age.

This gives a high level description of the particular ossification process. It is interesting to note that the initiation of this process is given at the granularity of weeks, with respect to fetal period, while its termination is given at the granularity of years and with respect to maturity. The particular ossification process spans a chain of temporal contexts (developmental periods); this becomes apparent once the process is decomposed into subprocesses at finer levels of description.

Knowledge on *normality* (such as descriptions of normal ossification processes) serves different purposes in a diagnostic process. Firstly such knowledge can be used for establishing whether some observation describes an abnormal situation, and therefore warrants an explanation. For example, the earliest, the primary centers of ossification of the anterior center for arch of the first cervical vertebra, are expected to appear is the age of 12 months. Thus the non appearance of these centres for a child of 10 months does not represent an observation of abnormality. Secondly knowledge on normality can be used to further abstract observations of abnormality. For example if the above observation was made at the age of 2 years this would be an abnormality, more specifically a delay in the ossification of the vertebra since the latest age that such centers are expected to appear is 15 months.

In general knowledge-based temporal data abstraction is a necessary process for diagnostic, monitoring, therapy planning, and other medical tasks dealing with dynamic situations. The data which is processed with the aim of deriving intelligent abstractions (maximal persistences, identification of distinct (compound) occurrences, temporal trends, periodic occurrences, other complex temporal patterns, etc.), i.e, abstractions that are directly matchable against diagnostic knowledge, can have vastly different characteristics (grossly incomplete or excessive in volume, numeric or qualitative, vague, noisy, etc.). The temporal abstractions try to fill in the gaps or explicate the significant information from a large volume of very specific data, eliminate the noise or vagueness, bring out potential dependencies or interactions, etc. The derived abstractions can be fallible since everything is dynamic and changeable; moreover current abstractions may need to be modified on the basis of old data that has now become available (view updating [77]), or past abstractions are revoked by new data (hindsight [73]). Deriving temporal abstractions on the basis of what is currently available or believed is one aspect of the problem; truth maintenance of the derived abstractions is the other aspect. The overall problem of temporal data abstraction is open-ended and as already illustrated its solution depends critically on knowledge. Thus straightforward algorithmic solutions are not appropriate and this is fairly evident in the various approaches proposed, a selection of which is presented in the sequel.

## 3. Machine learning for medical diagnosis

In this section we give a description of specific requirements that any machine learning system has to satisfy in order to be used in the development of applications in medical diagnosis. Several learning algorithms are then briefly described. We compare the performance of all the algorithms on several medical diagnostic problems and their appropriateness for medical diagnostic applications.

### *3.1. Requirements for machine learning systems*

For a machine learning system to be useful in solving medical diagnostic problems the following features are desired: good performance, ability to appropriately deal with missing and noisy data (errors in data), transparency of discovered diagnostic knowledge, ability to explain the proposed diagnosis for a new patient, and a reduction in the number of tests necessary to obtain a reliable diagnosis.

In this section we first discuss these requirements. Then we present a comparison study of seven representative machine learning algorithms to illustrate more concretely the points made.

– **Good performance:** The algorithm must be able to extract significant information from the available data. The diagnostic accuracy on new cases has to be as high as possible. Typically, most of the algorithms perform at least as well as the physicians and often the classification accuracy of machine classifiers is better than that of physicians when using the same patient descriptions. Therefore, if it is possible to measure the accuracy of physicians, their performance can be used as a lower bound on the required accuracy of the machine learning system in the given problem.
  In the majority of learning problems, various approaches typically achieve similar performance in terms of classification accuracy although in some cases some algorithms may perform significantly

better than the others [55]. Therefore, almost none of the algorithms can be excluded apriori with respect to the performance criterion. Rather, several learning approaches should be tested on the available data and one or few with best estimated performance should be considered for the development of the application.

- **Dealing with missing data:** In medical diagnosis very often patient descriptions lack certain data. Machine learning algorithms have to be able to appropriately deal with such incomplete descriptions.
- **Dealing with noisy data:** Medical data typically suffer from uncertainty and errors. Therefore machine learning algorithms, appropriate for medical applications, have to have effective means for handling noisy data.
- **Transparency of diagnostic knowledge:** The generated knowledge and the explanation of decisions accruing from the application of this knowledge should be transparent to the physician. He should be able to analyze and understand the generated knowledge. Ideally, the automatically generated knowledge will provide to the physician a novel point of view on the given problem, and may reveal new interrelations and regularities that the physician did not see before in an explicit form.
- **Explanation ability:** The system (applying the discovered diagnostic knowledge) must be able to explain decisions when diagnosing new patients. When faced with a curious solution of a new problem the physician shall require further explanation, otherwise he will not seriously consider the system's suggestions. The only possibility that physicians would accept a "black box" classifier is where such a classifier would outperform by a very large margin all other classifiers including the physicians themselves in terms of classification accuracy. However, such a situation is highly improbable and the authors of this paper are not aware of any.
- **Reduction of the number of tests:** In medical practice the collection of patient data is often expensive, time consuming, and harmful for the patients. Therefore, it is desirable to have the classifier that is able to reliably diagnose with a small amount of data about the patients. This can be verified by providing all candidate algorithms with the limited amount of data. However, the process of determining the right subset of data may be time consuming as it is essentially a combinatorial problem. Some of the machine learning systems are themselves able to select the appropriate subset of attributes, i.e., the selection is done during the learning process and may be more appropriate than others that lack this facility.

### 3.2. *Description of the tested algorithms*

In this subsection we briefly describe seven algorithms that were used in our experiments: Assistant-R, Assistant-I, LFC, the naive and semi-naive Bayesian classifier, backpropagation with weight elimination, and the $k$-nearest neighbors algorithm.

**Assistant-R:** This is a reimplementation of the Assistant learning system for top down induction of decision trees [8]. The basic algorithm goes back to CLS (Concept Learning System) developed by Hunt et al. [25] and was reimplemented and improved by several authors (see [66] for an overview). The main features of Assistant are binarization of attributes, decision tree prepruning and postpruning, incomplete data handling, and the use of the naive Bayesian classifier to calculate the classification in "null leaves".

The main difference between Assistant and its reimplementation Assistant-R is that ReliefF is used for attribute selection [42]. ReliefF is an extended version of Relief, developed by Kira and Rendell [38, 39], which is a non-myopic heuristic measure that is able to estimate the quality of attributes even if there are strong conditional dependencies between attributes. For example, Relief can efficiently estimate the quality of attributes in parity problems. In addition, wherever appropriate, instead of the relative frequency, Assistant-R uses the $m$-estimate of probabilities, which was shown to often improve the performance of machine learning algorithms [7].

**Assistant-I:** A variant of Assistant-R that instead of ReliefF uses information gain for the selection criterion, as does the original Assistant. However, the other differences to Assistant remain ($m$-estimate of probabilities).

**LFC:** Ragavan and Rendell [69] use limited lookahead in their LFC (Lookahead Feature Construction) algorithm for top down induction of decision trees to detect significant conditional dependencies between attributes for constructive induction. They show interesting results on some data sets. Robnik [70] developed a reimplementation of their algorithm which was then used in our experiments. LFC generates binary decision trees. At each node, the algorithm constructs new binary attributes from the original attributes, using logical operators (conjunction, disjunction, and negation). From the constructed binary attributes, the best attribute is selected and the process is recursively repeated on two subsets of training instances, corresponding to two values of the selected attribute. For constructive induction a limited lookahead is used. The space of possible useful constructs is restricted, due to the geometrical representation of the conditional entropy which is the estimator of the attributes' quality. To further reduce the search space, the algorithm also limits the breadth and the depth of search.

As LFC uses lookahead it is less myopic than the greedy algorithm of Assistant. The comparison of experimental results of LFC and Assistant-R contrasts the performance of the greedy search in combination with ReliefF versus the lookahead strategy. To make results comparable to Assistant-R we equipped LFC with the same pruning and probability estimation facilities. All tests were performed with a default set of parameters (depth of the lookahead 3, beam size 20), although in some domains better results may be obtained by parameter tuning. However, higher values of the parameters may combinatorially increase the search space of LFC, which makes the algorithm impractical.

**Naive Bayesian Classifier:** A classifier that uses the naive Bayesian formula to calculate the probability of each class $C$ given the values $V_i$ of all the attributes for an instance to be classified, assuming the conditional independence of the attributes given the class:

$$P(C|V_1..V_n) = P(C)\prod_i \frac{P(C|V_i)}{P(C)}$$

A new instance is classified into the class with maximal calculated probability. We used the $m$-estimate of probabilities [7]. For prior probabilities the Laplace's law of succession was used. In our experiments, the parameter $m$ was set to 2.0 (this setting is usually used as default and, empirically, gives satisfactory results [7]).

The relative performance of the naive Bayesian classifier can serve as an estimate of the conditional independence of attributes.

**Semi-naive Bayesian Classifier:** Kononenko [40] developed an extension of the naive Bayesian classifier that explicitly searches for dependencies between the values of different attributes. If such dependency is discovered between two values $V_i$ and $V_j$ of two different attributes then they are not considered as conditionally independent. Accordingly the term

$$\frac{P(C|V_i)}{P(C)} \times \frac{P(C|V_j)}{P(C)}$$

in the naive Bayesian formula is replaced with

$$\frac{P(C|V_i, V_j)}{P(C)}$$

For such a replacement a reliable approximation of the conditional probability $P(C|V_i, V_j)$ is required. Therefore, the algorithm trades-off between the non-naivety and the reliability of approximations of probabilities.

**Backpropagation with weight elimination:** The multilayered feedforward artificial neural network is a hierarchical network consisting of two or more fully interconnected layers of processing units - neurons. The task of the learning algorithm is to determine the appropriate weights on the interconnections between neurons. Backpropagation of error in multilayered feedforward neural network [72] is a well known learning algorithm and also the most popular among algorithms for training artificial neural networks. Well known problems with backpropagation are the selection of the appropriate topology of the network and overfitting the training data. An extension of the basic algorithm that uses the weight elimination technique [87] addresses both problems. The idea is to start with too many hidden neurons and to introduce into the criterion function a term that penalizes large weights on the connections between neurons. With such criterion function the algorithm, during training, eliminates an appropriate number of weights and neurons in order to obtain the appropriate generalization on the training data.

**$k$-NN:** The $k$-nearest neighbor algorithm. For a given new instance the algorithm searches for $k$ nearest training instances and classifies the instance into the most frequent class of these $k$ instances. The presented results in the next section were obtained with Manhattan-distance. The results using Euclidian distance are practically the same. The best results with respect to parameter $k$ are presented, although for fair comparison such parameter tuning should be allowed only on the training and not the testing sets of data.

### 3.3. Performance of algorithms on medical diagnostic problems

We compared the performance of the algorithms on several medical data sets.

- Data sets obtained from the University Medical Center in Ljubljana, Slovenia: the problem of locating the primary tumour in patients with metastases (PRIM), the problem of determining the type of the cancer in lymphography (LYMP), diagnosis in rheumatology (RHEU), and diagnosis of sport injuries (SPORT). The problem domains PRIM, LYMP and RHEU are briefly described in Section 2.4.1 (see also [45]). The problem domain SPORT is described in detail in [89].
- Data sets obtained from the StatLog database [55]: diagnosis of diabetes (DIAB) and diagnosis of heart diseases (HEART).

The characteristics of these data sets are given in Table 3.

Table 3
Basic description of medical data sets.

| Domain | Cl | Att | Val/Att | Ins | MC | Ent |
|---|---|---|---|---|---|---|
| PRIM | 22 | 17 | 2.2 | 339 | 25% | 3.89 bit |
| THYR | 4 | 15 | 9.1 | 884 | 56% | 1.59 bit |
| RHEU | 6 | 32 | 9.1 | 355 | 66% | 1.73 bit |
| LYMP | 4 | 18 | 3.3 | 148 | 55% | 1.28 bit |
| DIAB | 2 | 8 | 8.8 | 768 | 65% | 0.93 bit |
| HEART | 2 | 13 | 5.0 | 270 | 56% | 0.99 bit |
| SPORT | 30 | 49 | 3.7 | 118 | 15% | 4.49 bit |

Results of the experiments on these data sets are given in Fig. 3. These are averages over 10 runs for each domain. In each run, the dataset was randomly partitioned into 70% of data for learning, and 30% for testing.

### 3.4. Appropriateness for medical diagnosis

In this section we discuss how various algorithms fit the requirements described in Section 3.1. Table 4 summarizes the comparison of algorithms with respect to the appropriateness for developing applications in medical diagnostic problems.

Among the compared algorithms only decision tree builders are able to select the appropriate subset of attributes. With respect to the criterion of reduction of the number of tests, these algorithms have clear advantage over other algorithms.

With respect to the performance criterion the algorithms are more similar. The best performance was achieved by naive and semi-naive Bayesian classifiers. In medical data sets, attributes are typically relatively conditionally independent given the class. Physicians try to define conditionally independent attributes. Humans tend to think linearly and independent attributes make the diagnostic process easier. Therefore, it is not surprising that the Bayesian classifiers show clear advantage on medical data sets. It is interesting that the performance of the $k$-NN algorithm is also good in these domains.

In our experiments, on the DIAB dataset, all classifiers perform equally well, with the exception of the Bayesian classifiers which are significantly better. LFC achieved significantly better results than the other two inductive algorithms in the LYMP domain, where constructive induction seems to be useful. However, LFC performed significantly worse in the RHEU domain while in the other domains the three inductive algorithms performed equally well.

With respect to the transparency and the explanation ability criteria there are great differences between the algorithms:

- **$k$-nearest neighbors:** As $k$-NN does no generalization, the transparency of knowledge representation is poor. However, to explain the decision of the algorithm, a predefined number ($k$) of nearest neighbors from the training set is shown. This approach is analogous to the approach used by domain experts who make decisions on the basis of previously known similar cases. Such explanation ability is assessed by physicians as acceptable.
- **Naive and semi-naive Bayes:** Here, knowledge representation consists of a table of conditional probabilities which seems to be of interest to physicians. Therefore such knowledge representation is assessed as good. On the other hand, the decisions of Bayesian classifiers can be naturally interpreted as the sum of information gains [41]. The amount of information necessary to find out that an instance belongs to class $C$, is given by:

$$-\log_2 P(C|V_1,...,V_n) = -\ \log_2 P(C) - \\ - \sum_i(-\log_2 P(C) + \log_2 P(C|V_i))$$

Therefore, the decisions of the Bayesian classifiers can be explained with the sum of information gains from all attributes in favor or against the given class. In the case of the semi-naive Bayesian classifier, the process is exactly the same, except when the tuples of joined attribute/value pairs oc-

Fig. 3. Classification accuracy of learning systems on medical data sets.

Table 4
The appropriateness of various algorithms for medical diagnosis.

| classifier | performance | transparency | explanations | reduction | miss. data handling |
|---|---|---|---|---|---|
| Assistant-R | good | very good | good | good | acceptable |
| Assistant-I | good | very good | good | good | acceptable |
| LFC | good | good | good | good | acceptable |
| naive Bayes | very good | good | very good | no | very good |
| semi-naive Bayes | very good | good | very good | no | very good |
| backpropagation | very good | poor | poor | no | acceptable |
| $k$-NN | very good | poor | acceptable | no | acceptable |

cur. In this case, instead of simple attribute values, the joined values are used.
Such information gains can be listed in a table to sum up the evidence for/against the decision. Figures 4 and 5 provide a typical explanation of one decision. Each attribute has an associated strength, which is interpreted as the amount of information in bits provided by that attribute. It can be in favor or against the classifier's decision. One of the main advantages of such explanation is that it uses all available attributes. Such explanation was found by physicians as very good and they feel that Bayesian classifiers perform the task in a way similar to how they diagnose. Namely, they also sum up the evidence for/against a given diagnosis.

– **Backpropagation neural networks** have non-transparent knowledge representation and in gen-

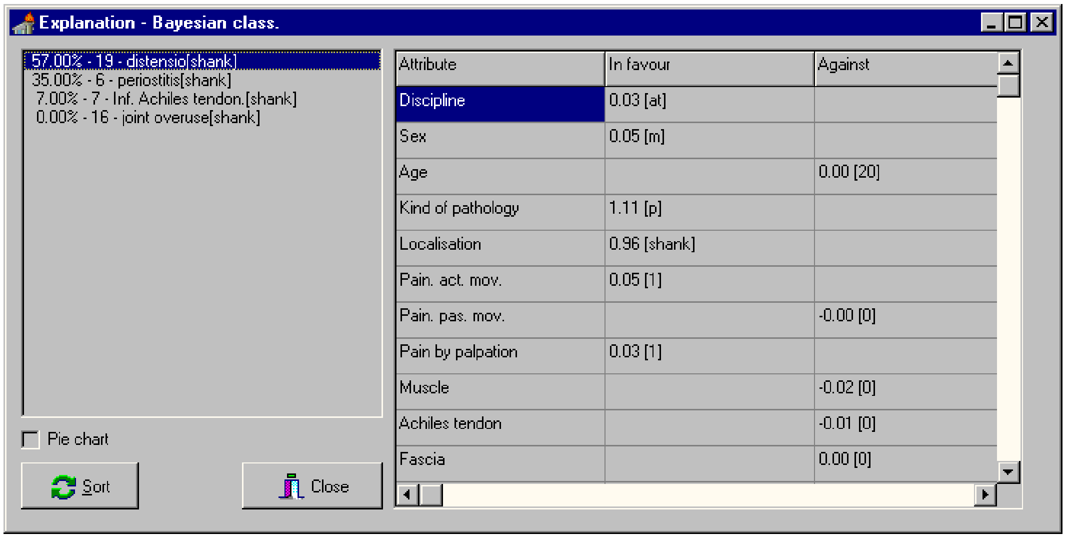


Fig. 4. Naive Bayes: An explanation of the decision in the diagnosis of sport injuries.

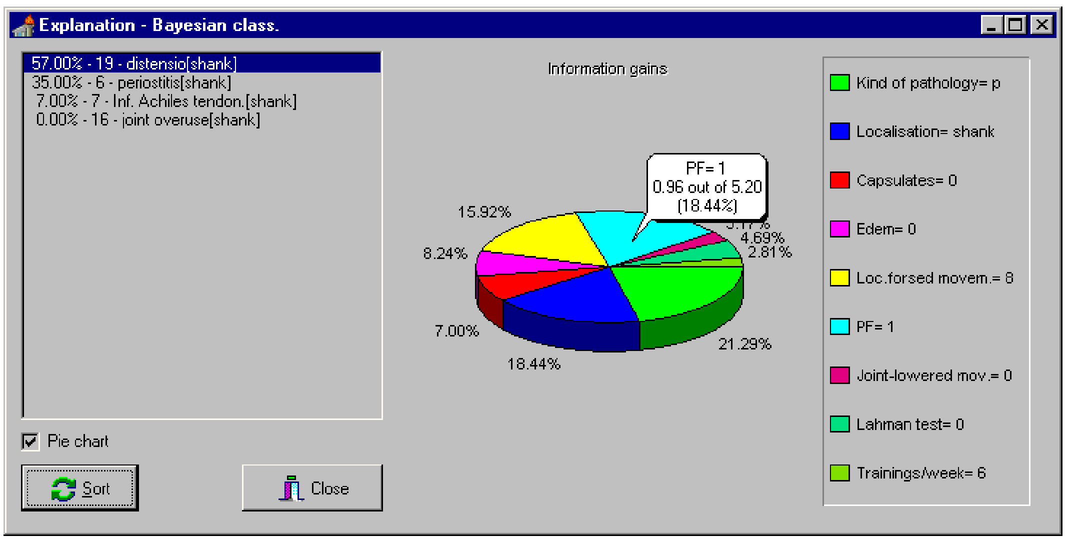


Fig. 5. Graphical explanation of the decision in the diagnosis of the sport injuries by the naive Bayesian classifier.

eral cannot easily explain their decisions. This is due to the large number of real-valued weights which all influence the result. In some cases it is possible to extract symbolic rules from the trained neural network. However, the rules tend to be large and relatively complex. Craven and Shavlik [13] compare rules extracted from a neural network with rules produced by Quinlan's [68] C4.5 system. The rules for a *NetTalk* data set extracted from a neural network have on average over 30 antecedents per rule compared to 2 antecedents for C4.5. Such rules are too complicated and hardly offer a useful explanation to a non-technically oriented domain expert.

– **Decision trees** (**Assistant-I and Assistant-R**) can be used without the computer and are fairly easy to understand. Positions of attributes in the tree, especially the top ones, often directly correspond to the domain expert's knowledge. However, in order to produce general rules, these methods use pruning which drastically reduces the tree sizes. Correspondingly, the paths from the root to the leaves are shorter, containing only few, although most informative attributes. In many cases the physicians feel that such a tree describes the diagnoses too poorly and is therefore not sufficiently informative [63]. In several problems the physicians have preferred the decision trees generated by Assistant-R. It seems that the estimates of ReliefF correspond to the way the physicians estimate the importance of attributes. In fact, the structure of trees generated by Assistant-I are often considered by physicians as strange and unnatural.

– **Lookahead feature construction** (**LFC**) also generates decision trees. However, in each node a potentially complex logical expression is used instead of a simple attribute value. The generated trees can therefore be smaller. The expressions may represent valid concepts from the domain. However, on the lower levels of the tree the expressions are often very specific and typically meaningless. Due to complex logical expressions in nodes, the number of attributes used to classify an instance can be higher than in usual decision trees.

### *3.5. Multistrategy learning*

Multistrategy learning approaches, as proposed by [54], construct several classifiers using different machine learning approaches and then use *all* the classifiers on new problems by combining their decisions. We used this idea in the prediction of the femoral neck fracture recovery problem [47]. In this study the results of different classifiers were combined (using the naive Bayesian formula) to make the final decision which can be explained as a weighted sum of single decisions. The physicians' evaluation of this approach indicates that with a multistrategy approach the reliability and the interpretability of the results is much better than when using a single learning strategy (e.g., decision tree learning only).

## 4. Temporal abstraction for medical diagnosis

Abstraction is a central concept in many disciplines especially those that deal with modeling and problem solving, such as informatics and engineering. Any process that hides the detail (in some situation), which can be erroneous and misleading at places, and brings out the essence (from the perspective of some goal) can be termed as an abstraction process. Such a process enables one to view the given situation from a more global and in some sense detached perspective and avoids getting bogged down to local, unimportant, detail. Data abstraction in the context of knowledge-based problem solving was introduced by Clancey in the mid eighties [10]. Temporal data abstraction research is not more than ten years old, but in spite of its young age, this emerging technology has managed to break substantial ground with respect to real life medical problems. Atemporal data abstraction is substantially simpler than temporal data abstraction; time brings a whole new dimension and complexity.

The proposed approaches to temporal data abstraction are to a large extent domain independent although the motivation for most of these lies in specific, real, medical problems. So unlike machine learning approaches that have been applied to many different problems and thus concrete evaluation results on their performance are available, this is not so for temporal data abstraction methods. The majority of these have only been applied to a single medical problem and the evaluation results, although very promising, are at best of a preliminary stage. Furthermore a number of these approaches have been developed primarily for monitoring tasks, either for the sort of acute monitoring that takes place in intensive care units, or for discrete monitoring over long periods of time, perhaps for the rest of the patient's life, as in the case of chronic disorders. In continuous monitoring of acute problems one is faced with large volumes of data that need to be interpreted, while in discrete monitoring one is faced with largely incomplete and often vague recollections of what has happened. Noise can be a characteristic of either type. The scope of this paper is diagnosis and not monitoring as such. Monitoring functions to detect as timely as possible (potentially) alarming situations, such as when a therapeutic regime is not working for the patient, or when the status of a person diverges from normality or some steady state. Thus a monitoring system is required to answer diagnostic questions such as "Are things steady or are they improving or worsening?", or "Is the situation normal or some misbehavior is detected?". Monitoring therefore involves repetitive application of diagnostic reasoning.

Temporal data abstraction can in fact be used as a stand alone method for the intelligent interpretation, and possibly visual presentation, of a patient's relevant history as dictated by a moving time window underlying the higher level decision making process. So a temporal data abstraction engine can be viewed as an intelligent assistant to a physician. This engine is (continuously) fed with new data on the patient (both referring to the present or the past, and possibly revoking previously specified data), and presents, to the physician, in a transparent and comprehensive fashion the information abstracted from the raw data. The physician can then use this information to draw diagnostic, therapeutic, or whatever other conclusions of relevance to the decision making task at hand.

In the remaining of this section first we list the general requirements for temporal data abstraction methods and then we overview some of the recent temporal data abstraction proposals. We start with Shahar and Musen's proposal, that has been implemented in the R'ESUM'E system [76–79]; this aims to provide a very generic and reusable approach to temporal data abstraction. Next we overview two proposals that are more specific in scope, namely the proposals by Haimowitz and Kohane [22], and Miksch et al. [56] respectively implemented in the systems TrenDx and VIE-VENT. Both proposals focus on the derivation of trends. We conclude the section by presenting an approach by Keravnou regarding the derivation of periodic occurrences. Periodicity, or more generally repetition, is not a specific focus in any of the other approaches.

All these approaches are knowledge-based and thus for their application in a specific domain (a) the types of knowledge underlying their operation must be available, and (b) the assumed (generic) characteristics and properties of the processed data must be satisfied by the data of the given domain.

### 4.1. *Requirements for temporal data abstraction*

Temporal data abstraction can have the following uses:

- For the support of a decision making task (diagnosis, monitoring, therapy planning, etc.) with respect to an individual patient, by presenting the relevant (moving) history of the patient at a level of abstraction appropriate to the particular task. The latter does not need to be automated in the form of a computer-based system, but it can be directly performed by the human problem solver. In this mode of operation, the temporal data abstraction process would be required to present its output in a visual form.
- For the preprocessing of the (frozen) histories of a number of patients. The abstracted histories are then fed to a data mining algorithm for the induction of knowledge (such as diagnostic rules) for the particular task. Current machine learning approaches do not attempt to first abstract, on an individual basis, the example cases that constitute their training sets, and then to apply whatever learning technique they employ for the induction of further generalizations. Strictly speaking every machine learning algorithm performs a kind of abstraction over the entire collection of cases; however it does not perform any abstraction on the individual cases. Cases tend to be atemporal, or at best they model time (implicitly) as just another attribute. Data abstractions on the selected cases are often mannually performed by the domain experts as a preprocessing step. Such manual processing is prone to non uniformity and inconsistency, while the automatic extraction of abstractions is uniform and objective. The integration of temporal data abstraction methods with machine learning algorithms will give a new perspective to machine learning with the aim of inducing "deeper" knowledge.

The essential requirement for temporal data abstraction, under either of its uses, is derivation of all appropriate abstractions, of any degree of complexity, and no derivation of erroneous or misleading abstractions. Under its first use discussed above, where everything is dynamic, truth maintenance is also an essential requirement. Completeness, relevance, and correctness of derivations and truth maintenance are the overall requirements. Within these requirements, the more specific requirements are (these are quite similar to the corresponding requirements for machine learning):

- Dealing with noisy or vague data.
- Dealing with missing data.
- Dealing with a variety of data.
- Deriving transparent and comprehensible abstractions.
- Visually presenting (if required) the derived abstractions in a highly explanatory manner.
- Deriving and revising abstractions in a time efficient manner (for dynamic decision support).

Since temporal data abstraction is knowledge driven, the means for achieving, at least the conceptual requirements, are knowledge-based.

### 4.2. *Selected approaches*

In this section, for illustration purposes, we briefly overview four approaches to temporal data abstraction. Detailed accounts on the selected approaches and their so far evaluation are available in the literature and the reader is referred there.

*4.2.1. Shahar and Musen's approach*

Shahar and Musen [76–79] have developed a knowledge-based framework for the creation of abstract, interval concepts from time-stamped clinical data. The framework has been implemented in the R'ESUM'E system under the CLIPS environment. The principles underlying this framework are genericity and reusability and the use of knowledge is emphasized. More specifically the proposers define the types of knowledge required (structural, classification, temporal semantic, and temporal dynamic, knowledge) for the identified temporal abstraction functionalities (context formation, contemporaneous abstraction, temporal inference, temporal interpolation, and temporal pattern-matching). In a specific application of the framework the actual knowledge is organized under various ontologies for parameter-properties, events, contexts, and dynamic induction relations of context intervals.

The framework supports four types of abstractions: state, gradient, rate and pattern. Given a historic database, R'ESUM'E aims to infer, in a non directed fashion, all derivable abstractions of any degree of complexity. The process of derivation is repeatedly applied since by its very nature a historic database is never fixed, and also truth maintenance is supported.

A significant novelty of this approach, is the dynamic derivation of interpretation contexts; these could be contemporaneous, prospective and retrospective. Interpretation contexts are induced by events, such as therapeutic actions. Two or more interpretation contexts could define generalized interpretation contexts; moreover contexts could be nonconvex, if they are induced on the basis of repetitive events. Abstractions are generated on the basis of interpretation contexts, thus the interpretation of the patient data is context sensitive. Several concurrent interpretation contexts can be induced, maintained and queried, thus creating different interpretations for the same set of data points.

In summary, the underlying ontologies, required knowledge, and supported functionalities have been specified in great detail, and the soundness of the proposal has been demonstrated through its application to a number of medical domains (therapy for insulin-dependent diabetes, protocol-based care of AIDS and of chronic GVHD, and monitoring of children's growth) with promising results.

*4.2.2. Haimowitz and Kohane's approach*

Haimowitz and Kohane [22] have developed a system, TrenDx, with the specific focus of medical trend diagnosis. Generic trends are defined through the notion of a trend template that gives great power of expression. This is both the strength and the limitation of this approach. Strength because of the higher power of expression supported. Limitation because this expressiveness is required if one wishes to define dynamic processes (e.g. disorder processes) in terms of the different phases comprising them, the uncertainty governing the transitions from one phase to the next, the significant events marking these transitions and various constraints on parameter-values associated with the different phases. Thus in using this approach one is forced to intermix data abstraction knowledge with diagnostic knowledge per se; there is no clear separation between the two, and no diagnostic independent specification of temporal abstraction knowledge (of the types advocated by Shahar). In other words a trend template is a fairly sophisticated mechanism for the specification of temporal models for dynamic processes, both normal and abnormal processes. There is no decoupling between an intermediate level of data interpretation (derivation of abstractions) and a higher level of decision making. Data interpretation involves the selection of the trend template instantiation that matches best the raw temporal data (this covers noise detection and positioning of transitions). The selected trend template instantiation is the final solution; thus temporal data abstraction and diagnostic (or other) reasoning per se are tangled up into a single process. This makes the overall reasoning more efficient, but it limits the genericity of the approach; the derivation of the abstractions is very much directed (trend template driven) and hence the potentiality of this approach as a preprocessing tool for machine learning is somewhat limited; for the discovery of new knowledge (i.e., new diagnostic rules) the abstractions used should be derived in a nondirected, i.e., in a non-biased fashion.

TrenDx has been applied, with promising results, to the diagnosis of pediatric growth disorders and the detection of significant trends in hemodynamics and blood gas in intensive care unit patients.

*4.2.3. Miksch et al. approach*

The third approach to be discussed is by Miksch et al. [56]. This approach, like the one by Haimowitz and Kohane, is aimed at a specific type of applications, and thus unlike the approach by Shahar and Musen, the aim is not to formulate in generic terms a knowledge-based temporal abstraction task. This proposal has been realized in VIE-VENT, a system for data validation and therapy planning for artificially ventilated newborn infants. Like R'ESUM'E, VIE-VENT is implemented in the CLIPS environment.

The overall aim is the context-based validation and interpretation of temporal data, where data can be of different types (continuously assessed quantitative data, discontinuously assessed quantitative data, and qualitative data). The interpretation contexts are not dynamically derived, but they are defined through schemata with thresholds that can be dynamically tailored to the patient under examination. The context schemata correspond to potential treatment regimes; which context is actually active depends on the current regime of the patient. If the interpretation of data points to an alarming situation, the higher level reasoning task of therapy assessment and (re)planning is invoked which may result in changing the patient's regime thus switching to a new context. Context switching should be done in a smooth way and again relevant thresholds are dynamically adapted to take care of this. The data abstraction process per se is fairly decoupled from the therapy planning process. Hence this approach differs from the previous one where the selection and instantiation of an interpretation context (trend template) represents the overall reasoning task. In VIE-VENT the data abstraction process does not need to select the interpretation context, as this is given to it by the therapy planning process.

The types of knowledge required are classification knowledge and temporal dynamic knowledge (e.g., default persistences, expected qualitative trend descriptions, etc.). Everything is expressed declaratively in terms of schemata that can be dynamically adjusted depending on the state of the patient. First quantitative point-based data are translated into qualitative values, depending on the operative context. Smoothing of data oscillating near thresholds then takes place. Interval data are then transformed to qualitative descriptions resulting in a verbal categorization of the change of a parameter over time, using schemata for trend-curve fitting. The system deals with four types of trends: very short-term, short-term, medium-term and long-term.

#### *4.2.4. Keravnou's periodicity approach*

The three approaches discussed do not explicitly address the derivation of periodic happenings. So we conclude this section by overviewing an approach proposed by Keravnou that focuses on the derivation of periodicity [33]. This approach, that has not yet been evaluated in a real medical domain, is part of a bigger effort that aims to develop a generic and reusable temporal kernel for medical knowledge-based problem solvers; temporal data abstraction features as one of the derivation functionalities of this kernel and the derivation of periodicity is a subfunctionality of temporal data abstraction [31, 32].

Periodicity is very relevant to medical reasoning. As Kahn [26] puts it "Most medical phenomena recur. Illnesses reappear, symptoms return, treatments start, stop, and resume. Frequently, events from one clinical episode provide key patient-specific insights about what might transpire during a later episode. Thus, the ability to reason about recurring events is an essential aspect of temporal problem-solving.".

The principle underlying the time-ontology that constitutes the foundations of the proposed periodicity approach is that for time to be properly integrated in a knowledge-based system, it should be an integral aspect of the entities that form the processing elements of the system. The central primitive of the ontology is the *time-object* which is a dynamic entity, viewed as a tight coupling between a property and an existence; its existence can be expressed with respect to different temporal contexts (time-axes), and thus depending on the context of reference (and associated time granularity) the time-object can be treated as a point-object (and thus indivisible) or as an interval-object whose existence can also be governed with uncertainty. Time-objects can be compound and can be involved in causal interactions. This way the notion of a time-object unifies three essential types of knowledge, temporal, structural, and causal.

Periodic occurrences are modeled as compound time-objects, subsuming a number of other time-objects. A generic periodic time-object is specified through a *repetition element*, a *repetition pattern*, and a *progression pattern* (over the sequence of instantiations of the repetition element). The repetition element could itself be a periodic occurrence (thus having nested periodicity), or a trend, etc.

The aim of the proposed approach is to derive, in a nondirected fashion, all periodic occurrences, of any order of complexity, which are derivable from some patient history, where the patient history is a collection of concrete time-objects. The types of knowledge required include temporal-semantic knowledge of properties, regularity patterns, knowledge on dominant/subordinate relations between property subjects, and knowledge relating to the justification of exclusions.

There are two basic algorithms: (a) an algorithm that derives periodic occurrences within a sequence of time-objects sharing the same property subject (order-

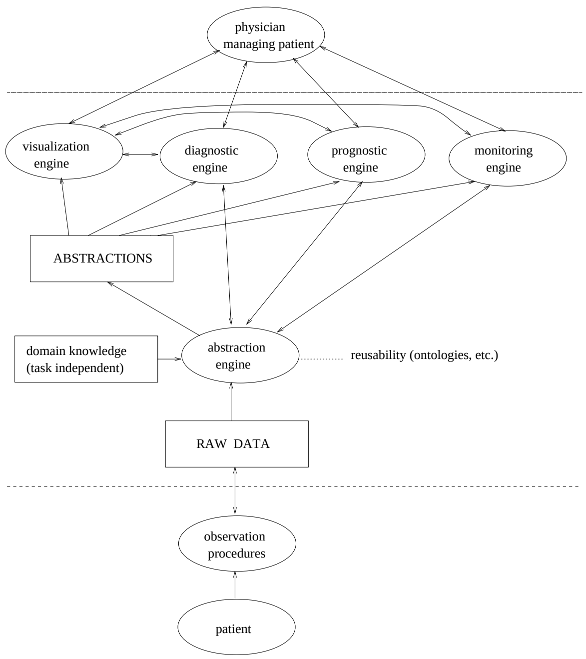


Fig. 6. Data abstraction as a decoupled process.

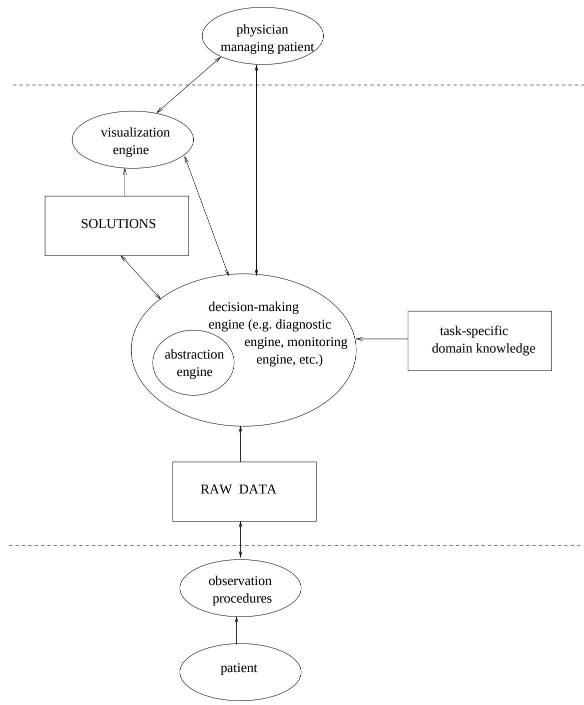


Fig. 7. Data abstraction as a task-dependent process.

1 periodicity); and (b) an algorithm that derives periodic occurrences across two sequences of time-objects with different property subjects (periodicity across two subjects). These algorithms are recursively used in the context of the algorithm for the derivation of order-$n$, $n > 1$, periodicity (periodic occurrences involving n distinct property subjects). The acceptable regularity patterns as passed as parameters to these algorithms. The selection of the best periodic occurrence out of a set of competing plausible periodic occurrences can be based on domain specific heuristics and the justification of the exclusion of time-objects (whose existences overlap with the existence of the conjectured periodic occurrence) is knowledge-based. This effort is ongoing.

In summary, a data abstraction engine is more usefully deployed as a process decoupled from a higher level reasoning engine (Fig. 6), as this achieves reusability and enables its utilization by a data mining engine. An abstraction engine that is embedded within a decision-making engine cannot be deployed by a data mining engine since the generation of abstractions (or more accurately solutions) is biased by the needs of the relevant decision-making engine (Fig. 7).

## 5. Temporal data abstraction for machine learning

This section discusses the potential of temporal data abstraction for the discovery of medical knowledge by machine learning.

Data is patient specific, while medical knowledge is patient independent and consists of generalizations that apply across patients. Machine learning for medical domains aims to discover medical knowledge by inducing generalizations from records of representative samples of patients. Trying to induce such generalizations directly from the raw patient data is particularly hard when generalizing from time stamped monitoring data. Consider a patient record stating that “the blood pressure reading was 9 at 10 am on March 26th 1966”. Making generalizations from patient data recorded at this level means, for instance, trying to find the same datum in more than one patient’s record; this is highly unlikely. This example shows that generalizations can be more effectively discovered by comparing patient profiles at a higher level of abstraction, in terms of derived data abstractions such as periodic occurrences, trends and other temporal patterns.

Different raw data can yield the same abstractions, even if they differ substantially in volume. The number of derived abstractions is relatively constant across patients with the same medical situation, and of course this number is considerably smaller than the number of raw data. Temporal data abstractions reveal the essence of the profile of a patient, hide superfluous detail, and last but not least eliminate noisy information. Furthermore, the temporal scope of abstractions like trends and periodic occurrences are far more meaningful and prone to adequate comparison than the time-points corresponding to raw data. If the same complex abstraction, such as a nested periodic occurrence, is associated with a significant number of patients from a representative sample, it makes a strong candidate for being a significant piece of knowledge. Sharing a complex abstraction is a strong similarity while sharing a concrete datum is a weak similarity, if at all.

Roughly, medical knowledge discovery aims to either refine the model of a known disorder, or to discover a new disorder/syndrome from undiagnosed cases. Temporal data abstraction can support both these types of knowledge discovery.

In the first case, tools for inductive concept learning can be used to synthesize or refine the model of a known disorder. A sample of patients to be used for learning is in theory easily identifiable; it consists of patients that have been correctly diagnosed to have suffered from the particular disorder. However, if the knowledge of this disorder is still vague and incomplete, it is possible that the sample includes also patients that were wrongly diagnosed to have suffered from that disorder. Temporal data abstraction can reveal such mistakes. To avoid such mistakes, the patients used in a sample should be carefully monitored over relevant periods of time to ensure that the raw data on each of them is as complete as possible. Complex abstractions shared by an adequate proportion of the monitored patients may well be promoted to features of the given disorder.

In the second case, tools for clustering can be used to discover new disorders. In this case the sample of patients is not immediately determined. This leads to the difficulty that it may not be possible or justifiable to closely monitor every undiagnosed case, and hence the patient records can be severely incomplete. However, a data abstraction process can reveal patients warranting closer inspection and monitoring. For example, if a number of undiagnosed patients appear to share some (complex) temporal abstraction, this could be a strong similarity and closer comparison between these patients can reveal a common picture at a high level of abstraction.

The above examples indicate the way how to effectively combine temporal data abstraction and machine learning. In the future, temporal data abstraction may have an important role in data preprocessing for machine learning and much research in expected in this area.

## 6. Conclusion

Although the results of applying various machine learning algorithms in medical diagnosis, reported in Section 3, seem excellent, this technology has not been widely accepted in medical practice. Reasons usually given by physicians themselves are diverse:

– Inflexibility of the knowledge representation. The set of attributes that describe the patients must be fixed. The information that is used by the rules to derive the final diagnosis is limited to strictly defined parameters while subjective, informal, and fuzzy notions (like intuition, impression, etc.) can not be represented in a formal and symbolic way.
– Physicians often claim that if they are not sure about the final diagnosis, usually further examinations (e.g. laboratory tests) may be performed to verify the diagnosis. In situations where further examinations are easy, the physicians do not feel the need for assistance in the diagnostic process. In prognosis there is no possibility for further examination that would confirm the prediction. For that reason the prognostic problems are even more attractive for machine learning than diagnostic problems [90].
– Physicians often claim that they are too busy to use any additional tool for decision making. In everyday practice it is too time and/or energy consuming to type in the data into the computer in order to use the computer support in the diagnostic process.
– In their regular education, physicians do not encounter intelligent data analysis techniques, which are seldomly included even in more specialized courses of medical informatics. In their research, they therefore prefer to use classical statistical analysis, which they are much more familiar with.
– Non-negligible is also subjective resistance of physicians to new diagnostic technology. It is often felt that the diagnosis problem, considered as perhaps the most critical and sensitive task, will be left to machines, thereby leaving the physicians without power to control and without responsibility.

– We have also rather frequently encountered quite irrational reasons for resisting computer diagnosis. These have actually been described by some physicians as follows. Diagnosis is regarded by some physicians as the premium intellectual task of their profession. As such, this task requires in-depth knowledge, unexpected ideas, and in particular intuition. Therefore diagnosis is a bit of an art that is impossible to explain and formalize. How can it then be done by computers? And if computers could do it, that would destroy all the magic and professional pride.

In the past we have developed several applications for medical diagnostic problems using decision tree technology [44, 71, 24]. Besides the above mentioned problems, decision trees suffer also from the following deficiencies:

– Learning and classification is sensitive to missing data [67] which is often the case in medical data.
– The generated decision rules typically include too few attributes [63]. The explanation of decisions is therefore poor and does not typically support exact decisions of generated diagnostic rules.

To solve these two problems and to increase the reliability and the transparency of automatically generated classifiers a multistrategy learning approach has recently gained much attention [54]. The approach of combining decisions of several classifiers when solving new problems is analogous to decision making in hospitals where the decisions for harder cases are solved by a group of physicians rather than by one physician alone. In [47] the results of different classifiers were combined to make the final decision which can be explained as a weighted sum of single decisions. Physicians felt that with a multistrategy approach the reliability and the comprehensibility of the results of learning were much better than when using decision trees only.

Temporal data abstraction represents a very young technology, and concrete evaluation results of its performance are just beginning to emerge. Assuming that adequate levels of performance are attained, there is every reason to expect that this technology will be fully accepted by physicians, as it can be of great assistance to them. Bringing the patient's relevant history to a form, and visually presenting it in an immediately discernible fashion, that enables the direct application of the higher level decision making performed by the physician, represents invaluable assistance.

In our view, the awareness of the challenging new fields of intelligent data analysis, data mining and knowledge discovery in databases, and emerging new technologies has been much larger in industry, finance, and economy than in medicine. Hence, the purpose of this paper is to increase the awareness of various techniques and methods that are available for intelligent data analysis in medicine, and to present some case studies of their application. The paper presents the state-of-the-art of selected intelligent data analysis techniques (machine learning and temporal data abstraction) and evaluates their applicability in medical diagnosis.

Despite the technical advantages of these technologies that may lead to more reliable diagnosis, the reviewed intelligent data analysis methods have not yet been widely accepted in practice. Among the reasons for such a slow acceptance of these new technologies is the fact that the data analysis tools are not yet integrated into the existing instrumentation, which would make its use simpler and more natural. In addition, the tools should be made more intuitive and equipped with functional and visually attractive user interfaces. Since for many algorithms there is a need to set certain numeric parameters in order to achieve best performance, a method for automatic parameter setting would be highly desirable. To make the preparation of data easier for users, the intelligent data analysis systems should also be able to provide an interface to standard database software.

*Acknowledgement*

Our thanks goes to physicians Matjaž Zwitter (PRIM, LYMP), Sergej Hojker (THYR), Vlado Pirnat (RHEUM) in Vasja Vuga (SPORT), for their collaboration and expertise in the selected diagnostic domains.

## References

[1] Aikins, J. S. (1979). Prototypes and production rules: An approach to knowledge representation for hypothesis formation. In *Proc. Sixth International Joint Conference on Artificial Intelligence*, pages 1–3.
[2] Barahona, P., and Christensen, J.P., editors (1994). *Knowledge and Decisions in Health Telematics*, IOS Press.
[3] van Bemmel, J. H. (1996). Medical informatics, art or science? *Meth. Inform. Med.*, 35:157–172.
[4] Bratko, I. and Kononenko, I. (1987). Learning diagnostic rules from incoplete and noisy data. In Phelps, B., editor, *AI Methods in Statistics*. Gower Technical Press, London.

[5] Bratko, I., Mozetič, I., and Lavrač, N. (1989). *KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems*. MIT Press, Cambridge, MA.

[6] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.

[7] Cestnik B., Estimating Probabilities: A Crucial Task in Machine Learning, *Proc. European Conf. on Artificial Intelligence*, Stockholm, August, 1990, pp. 147-149.

[8] Cestnik B., Kononenko I.& Bratko I., ASSISTANT 86: A knowledge elicitation tool for sophisticated users, in: I. Bratko, N. Lavrač (eds.): *Progress in Machine learning*, Wilmslow: Sigma Press, 1987.

[9] Chandrasekaran, B., and Mittal, S. (1983). Conceptual representation of medical knowledge for diagnosis by computer: MDX and related systems. *Advances in Computers*, 22:217–293.

[10] Clancey, W.J. (1985). Heuristic classification. *Artificial Intelligence*, 27:289–350.

[11] Clark, P. and Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Proc. Fifth European Working Session on Learning*, pp. 151–163. Springer, Berlin.

[12] Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4):261–283.

[13] Craven M.W. and Shavlik J.W., Learning symbolic rules using artificial neural networks, *Proc.* $10^{th}$ *Intern. Conf. on Machine Learning*, Amherst, MA, Morgan Kaufmann, 1993, pp.73-80.

[14] Davis, R. (1978). Knowledge acquisition in rule-based systems: knowledge representation as a basis for system construction and maintainance. In Shine, M. E. and Coombs, M. J., editors, *Designing for Human-Computer Communication*, pages 87–137. Academic Press.

[15] Davis, R. (1983). TEIRESIAS: experiments in communicating with a knowledge-based systems. In Waterman, D.A., and Hayes-Roth, F., editors, *Pattern Directed Inference Systems*, pages 99–134. Academic Press.

[16] Fagan, L. M., Shortliffe, E. H., and Buchanan, B. G. (1980). Computer-based medical decision making: From MYCIN to VM. *Automedica*.

[17] Fayyad, U. M., and Uthurusamy, R. (1996). Data minning and knowledge discovery in databases (editorial). *Communications of the ACM*, 39(11):24–26.

[18] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knwoledge from volumes of data. *Communications of the ACM*, 39(11):27–41.

[19] Feigenbaum, E. A. (1977). The art of artificial intelligence 1: Themes and case studies of knowledge engineering. Technical report, Pub. no. STAN-SC-77-621, Stanford University, Department of Computer Science.

[20] Feigenbaum, E. A. and McCurduck, P. (1984). *The fifth generation*. Pan Books, London.

[21] Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1991). Knowledge discovery in databases: An overview. In Piatetsky-Shapiro, G. and Frawley, W., editors, *Knowledge discovery in databases*. The AAAI Press, Menlo Park, CA.

[22] Haimowitz, I. J., and Kohane, I. S. (1996). Managing temporal worlds for medical trend diagnosis. *Artificial Intelligence in Medicine*, Vol.8, No.3, pp.299–321.

[23] Haiser, J. F., Brooks, R. E., and Ballard, J. P. (1978). Progress report: A computerized psychopharmacology advisor. In *Proc. 11th Collegium Internationale Neuro-Psychopharmatologicum*, Vienna.

[24] Hojker S., Kononenko I., Jauk A., Fidler V. & Porenta M., Expert System's Development in the Management of Thyroid Diseases, *Proc. European Congress for Nuclear Medicine*, Milano, Sept., 1988.

[25] Hunt E., Martin J & Stone P., *Experiments in Induction*, New York, Academic Press, 1966.

[26] Kahn, M. G. (1993) In pursuit of time's arrow: temporal reasoning in medical decision support. *Proc. of the 4th Conference on Artificial Intelligence in Medicine Europe*, Munich, IOS Press, pp.3–6.

[27] Kahn, G. M., Fagan, L. M., and Sheiner, L. B. (1991). Combining physiologic models and symbolic methods to interpret time-varying patient data. *Meth Inform Med*, Vol.30, pp.167–178.

[28] Keravnou, E. T., editor (1992). *Deep models for medical knowledge engineering*. Elsevier.

[29] Keravnou, E. T. (1995). Temporal vagueness in medical reasoning. *Int. J. Systems Research and Info Science*, Vol.7, pp.3–28.

[30] Keravnou, E. T., editor (1996a). Special issue on temporal reasoning in medicine. *Artificial Intelligence in Medicine*, 8(3):187–326.

[31] Keravnou, E. T. (1996b). Engineering time in medical knowledge-based systems through time-axes and time-objects. *Proc. TIME-96*, IEEE Computer Society Press, pp.160–167.

[32] Keravnou, E. T. (1996c). An ontology of time using time-axes and time-objects as primitives. Technical Report TR-96-9, Department of Computer Science, University of Cyprus.

[33] Keravnou, E. T. (1997). Temporal abstraction of medical data: deriving periodicity. In Lavrač, N., Keravnou, E. T., and Zupan, B. (Eds.) *Intelligent Data Analysis in Medicine and Pharmacology*. Kluwer Academic Publishers, 1997.

[34] Keravnou, E. T., Dams, F., Washbrook, J., Hall, C. M., Dawood, R. M., and Shaw, D. (1992). Background knowledge in diagnosis. *Artificial Intelligence in Medicine*, Vol.4, pp.263–279.

[35] Keravnou, E. T., Dams, F., Washbrook, J., Hall, C. M., Dawood, R. M., and Shaw, D. (1994). Modelling diagnostic skills in the domain of skeletal dysplasias. *Computer Methods and Programs in Biomedicine*, Vol.45, pp.239–260.

[36] Keravnou, E. T., and Washbrook, J. (1989). What is a deep expert system: an analysis of the architectural requirements of second generation expert systems. *Knowledge Engineering Review*, 4(3):205–233.

[37] Keravnou, E. T., and Washbrook, J. (1989). Deep and shallow models in medical expert systems. *Artificial Intelligence in Medicine*, 1(1):11–28.

[38] Kira K. & Rendell L., A practical approach to feature selection, *Proc. Intern. Conf. on Machine Learning* (Aberdeen, July 1992) D.Sleeman & P.Edwards (eds.), Morgan Kaufmann, 1992, pp.249-256.

[39] Kira K. & Rendell L., The feature selection problem: traditional methods and new algorithm. *Proc. AAAI'92*, San Jose, CA, July 1992.

[40] Kononenko I., Semi-naive Bayesian classifier, *Proc. European Working Session on Learning-91* (Y.Kodratoff (ed.), Springer-Verlag), Porto, March 4-6 1991, pp.206-219.

[41] Kononenko I., Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7:317-337, 1993.

[42] Kononenko I., Estimating attributes: Analysis and extensions of Relief. *Proc. European Conf. on Machine Learning* (Catania, April 1994), L. De Raedt & F.Bergadano (eds.), Springer Verlag, 1994, pp. 171-182.

[43] Kononenko I. & Bratko I., Information based evaluation cri-

terion for classifier's performance, *Machine Learning*, Vol.6, No.1, 1991, pp.67-80.

[44] Kononenko I., Bratko I., Roškar E.: Experiments in automatic learning of medical diagnostic rules, International School for the Synthesis of Expert's Knowledge Workshop, Bled, Slovenia, August, 1984.

[45] Kononenko I., Bratko I., Kukar, M.: Application of machine learing to medical diagnosis. In *Machine Learing and Data Mining: Methods and Applications*, R. S. Michalski, I. Bratko, and M. Kubat (eds.), John Willey & Sons, in press.

[46] Kunz, J. C., et al. (1978). A physiological rule-based system for interpreting pulmonary function test results. Technical report, Stanford HPP Memo HPP-78-19.

[47] Kukar M., Kononenko I., Silvester T., Machine learning in prognosis of the femoral neck fracture recovery, *Artificial Intelligence in Medicine*, Vol. 8, 1996, pp. 431-451.

[48] Larizza, C., Bellazzi, R., and Riva, A. (1997). Temporal abstractions for diabetic patients management. *Proc. AIME-97*, Lecture Notes in Artificial Intelligence, Vol.1211, Springer, pp.319–330.

[49] Lavrač, N. Mozetič, I. (1992) Second generation knowledge acqusition methods and their application to medicine. In Keravnou, E. T., editor, *Deep models for medical knowledge engineering*, pages 177–199, Elsevier.

[50] Lavrač, N. and Džeroski, S. (1994). *Inductive logic programming: Techniques and applications*. Ellis Horwood.

[51] Lavrač, N., Keravnou, E. T. and Zupan, B. (Eds.) *Intelligent Data Analysis in Medicine and Pharmacology*. Kluwer Academic Publishers, 1997.

[52] Michalski, R. (1983). A theory and methodology of inductive learning. In Michalski, R., Carbonell, J., and Mitchell, T., editors, *Machine Learning: An Artificial Intelligence Approach*, volume I, pages 83–134, Palo Alto, CA. Tioga.

[53] Michalski, R., Mozetič, I., Hong, J., and Lavrač, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In *Proc. Fifth National Conference on Artificial Intelligence*, pages 1041–1045, San Mateo, CA. Morgan Kaufmann.

[54] Michalski R., Tecuci G., *Machine learning: A multistrategy approach, Vol. IV*, Morgan Kaufmann, 1994.

[55] Michie D., Spiegelhalter D.J., Taylor C.C (eds.) *Machine learning, neural and statistical classification*, Ellis Horwood, 1994.

[56] Miksch, S., Horn, W., Popow, C., and Paky, F. (1996). Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. *Artificial Intelligence in Medicine*, Vol.8, pp.543–576.

[57] Miller, R. A. (1984). Internist-I/CADUCEUS: Problems facing expert consultant programs. *Meth. Inform. Med.*, 23:9–14.

[58] Miller, R. A., Pople, H. E., and Myers, J. D. (1982). Internist-I, An experimental computer-based diagnostic consultant for general internal medicine. *The New England Journal of Medicine*, 307(8):468–476.

[59] Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, 8(4):295–318.

[60] Nejdl, W., and Gamper, J. (1994). Harnessing the power of temporal abstraction in model-based diagnosis of dynamic systems. *Proc. ECAI-94*, pp.667–671.

[61] Patil, R. S., Szolovits, P., and Schwartz, W. B. (1982). Modelling knowledge of the patient in acid-base and electrolyte disorders. In Szolovits, P., editor, *Artificial Intelligence in Medicine*, AAAS Selected Symposium Series, pages 345–348. West View Press.

[62] Pauker, S. G., Gorry, G. A., Kassirer, J. P., and Schwartz, W. B. (1976). Towards the simulation of clinical cognition: Taking a present illness by computer. *The American Journal of Medicine*, 60:981–995.

[63] Pirnat, V., Kononenko, I., Janc, T., Bratko, I., Medical Estimation of Automatically Induced Decision Rules, *Proc. of 2nd Europ. Conf. on Artificial Intelligence in Medicine*, City University, London, August 29-31, 1989, pp. 24–36.

[64] Pople, H. E. (1982). Heuristic methods for imposing structure on ill structured problems: The structuring of medical diagnosis. In Szolovits, P., editor, *Artificial Intelligence in Medicine*, AAAS Selected Symposium Series, pages 119–185. West View Press.

[65] Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end-games. In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, *Machine Learning: An artificial intelligence approach*. Tioga Publishing Company, Paolo Alto.

[66] Quinlan J.R., Induction of Decision Trees. *Machine Learning*. Vol. 1, No. 1, 1986, pp. 81–106.

[67] Quinlan J.R., Unknown attribute values in induction, *Proc. 6th Int.Workshop on Machine Learning*, Cornell University, Ithaca, June 26-27, 1989, pp.164-168.

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

[69] Ragavan H. & Rendell L., Lookahead feature construction for learning hard concepts. *Proc. 10th Intern. Conf. on Machine Learning*. (Amherst, MA, June 1993), Morgan Kaufmann, 1993, pp.252-259.

[70] Robnik M., Constructive induction with decision trees. B.Sc. Thesis, University of Ljubljana, Faculty of Electrical Eng. & Computer Science, Ljubljana, Slovenia, (In Slovene), 1993.

[71] Roškar E., Abrams P., Bratko I., Kononenko I., Varšek A., MCUDS - An expert system for the diagnostics of lower urinary tract disorders, *Journal of Biomedical Measurements, Informatics and Control*, Vol. 1, No. 4, 1986, pp. 201 - 204.

[72] Rumelhart D.E. & McClelland J.L. (eds.), *Parallel Distributed Processing, Vol. 1: Foundations*. Cambridge: MIT Press, 1986.

[73] Russ, T. A. (1989). Using hindsight in medical decision making. *Proc. Symposium on Computer Applications in Medical Care*, New York NY: IEEE Computer Society Press, pp.38–44.

[74] Russ, T. A. (1995). Use of data abstraction methods to simplify monitoring. *Artificial Intelligence in Medicine*, Vol.7, pp.497–514.

[75] Safrans, C., Desforges, J., and Tsichlis, P. (1976). Diagnostic planning and cancer management. Rep. No. TR-169. Technical report, Laboratory for Computer Science, M. I. T., Massachusetts.

[76] Shahar, Y. (1994). A knowledge-based method for temporal abstraction of clinical data. *PhD thesis*, Department of Computer Science, Stanford University (Report No. KSL-94-64 or STAN-CS-TR-94-1529).

[77] Shahar, Y., and Musen M. A. (1992). A temporal-abstraction system for patient monitoring. *Proc. Symposium on Computer Applications in Medical Care*, New York NY: McGraw-Hill Inc., pp.121–127.

[78] Shahar, Y., and Musen, M.A. (1996). Knowledge-based temporal abstraction in clinical domains. *Artificial Intelligence in Medicine*, Vol.8, No.3, pp.267–298.

[79] Shahar, Y., Tu, S.W., Das, A.K., and Musen, M.A. (1992). A problem-solving architecture for managing temporal data and their abstractions. *Proc. AAAI-92 Workshop on Implementing Temporal Reasoning*

[80] Shortliffe, E. H. (1976). *Computer-Based Medical Consultations: MYCIN*. Elsevier.

[81] Shortliffe, E. H. (1993). The adolescence of AI in medicine: Will field come of age in the '90s? *Artificial Intelligence in Medicine*, 5(2):93–106.

[82] Shortliffe, E. H., Scott, C. A., and Bischoff, M. B. (1981). ONCOCIN: An expert system for oncology protocol management. In *Proc. Seventh International Joint Conference on Artificial Intelligence*, pages 876–881.

[83] Spackman, K., Elert, J. D., and Beck, J. R. (1993). The CIO and the medical informaticist: alliance for progress. In *Proc. Annual Symposium on Computer Applications in Medical Care*, pages 525–528.

[84] Szolovits, P. and Pauker, S. G. (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11.

[85] Thompson, W. B., Johnson, P. E., and Moen, J. B. (1983). Recognition-based diagnostic reasoning. In *Proc. Eight International Joint Conference on Artificial Intelligence*, pages 236–238.

[86] Wallis, J. W., and Shortliffe, E. H. (1982). Explanatory power of medical expert systems: studies in the representation of causal relationships for clinical consultatins. *Meth. Inform. Med.*, 21:127–136.

[87] Weigand S., Huberman A., & Rumelhart D.E., Predicting the future: a connectionist approach, *International Journal of Neural Systems*, Vol. 1(3), 1990.

[88] Weiss, S. M., Kulikowski, C. A., Amarel, S., and Safir, A. (1978). A model-based method for computer-aided medical decision making. *Artificial Intelligence*, 11:145–172.

[89] Zelič, I., Kononenko, I., Lavrač, N., Vuga, V. (1997) Diagnosis of sport injuries with machine learning: Explanation of induced decisions. In *Proc. Tenth IEEE Syposium on Computer-Based Medical Systems*, Maribor, Slovenia, June 11-13.

[90] Zwitter, M., Bratko, I., and Kononenko, I. (1983) Rational and irrational reservations against the use of computer in medical diagnosis and prognosis, *Proc. 3th Mediterranean Conf. on Medical and Biological Engineering*, Portorož, Slovenia, Sept. 5-9.