Int. J. Man-Machine Studies (1986) **24**, 475–499

A comparative analysis of methods for expert systems

CONNIE LOGGIA RAMSEY†, JAMES A. REGGIA†‡, DANA S. NAU† AND ANDREW FERRENTINO§

Departments of † Computer Science and ‡ Neurology, University of Maryland, College Park, Maryland 20742, *U.S.A. and § Software Architecture and Engineering*

(*Received* 28 *March* 1985 *and in revised form* 24 *January* 1986)

Given the current widespread interest in expert systems, it is important to examine the relative advantages and disadvantages of the various methods used to build them. In this paper we compare three important approaches to building decision aids implemented as expert systems: Bayesian classification, rule-based deduction, and frame-based abduction. Our critical analysis is based on a survey of previous studies comparing different methods used to build expert systems as well as our own collective experience over the last five years. The relative strengths and weaknesses of the different approaches are analysed, and situations in which each method is easy or difficult to use are identified.

1. Introduction

The importance of expert systems is growing in industrial, medical, scientific, and other fields. Several major reasons for this are: (1) the necessity of handling an overwhelming amount of knowledge in these areas; (2) the potential of expert systems to train new experts; (3) cost reductions sometimes provided by expert systems and (4) the desire to capture corporate knowledge so it is not lost as personnel changes (Waterman, 1986). In medicine, for example, expert systems are beginning to actually be used in practice (Reggia, 1982). An expert system named HELP is currently being used at LDS Hospital in Salt Lake City and has been shown to reduce healthcare costs (Nathanson, 1984; Pryor, Gardner, Clayton & Warner, 1984). HELP analyses patient data whenever a test order or result is entered into the system, and it warns physicians about such things as drug–drug interactions and drug contraindications. Studies showed that 80% of HELP's drug–drug interaction alerts were used by physicians to change prescriptions, test orders, or other forms of treatment, and this led to shortened hospital stays as well as improved quality of care (Pryor *et al.*, 1984). Examples of expert systems developed for industrial tasks include PROSPECTOR, which has successfully been used to locate mineral deposits (Campbell, Hollister, Duda & Hart, 1982), Dipmeter Advisor, which is used for oil exploration (Davis, Austin, Carlbom, Frawley, Pruchnik, Sneiderman & Gilreath, 1981), and R1, which is used to configure computers (McDermott & Steele, 1981). For further information and examples, see (Nau, 1983).

In spite of the widespread interest in expert systems, very little has been written comparing the relative advantages and disadvantages of the intrinsically different approaches available for building them. In fact, a number of authors, particularly in

Please send correspondence to James A. Reggia, Department of Computer Science, University of Maryland, College Park, Maryland 20742.

0020-7373/86/050475+25$03.00/0

Artificial Intelligence (AI), have even suggested that this is a closed issue because of the clear superiority of deductive approaches. For example, consider the following statements:

> "As a result, pattern-directed inference systems based on antecedent–consequent rules make a strong claim to being the best available scheme for knowledge representation . . ." (Hayes-Roth, Waterman & Lenat, 1978).
>
> "There is only one language suitable for representing information, whether declarative or procedural, and that is first-order predicate logic. There is only one intelligent way to process information and that is by applying deductive inference methods" (Kowalski, 1980).

Although rule-based deduction may seem synonymous with expert systems to some, it is actually only one of several widely used methods. Two other important approaches are statistical pattern classification and frame-based abduction.

In contrast to the opinions expressed by some, the authors believe that none of the available methods is obviously "the best" method. This paper presents an analysis of the relative limitations and trade-offs among the three methods listed above to provide guidelines for those who are considering the development of expert systems, and to identify areas where further research is needed. Our critical analysis is based on the following: (1) a review of previous studies comparing different methods used to build expert systems; (2) our own collective experience of over five years with developing and analysing methods for expert systems and (3) several studies which we undertook to evaluate different methods applied to the same application. In short, this paper is intended to provide a fairly comprehensive overview of the advantages and disadvantages of several major methods for building expert systems as they are understood today.

The rest of this paper is organized as follows. Section 2 provides a brief review of the three expert-system methods discussed in the paper: Bayesian classification (a statistical pattern classification method), rule-based deduction, and frame-based abduction. Section 3 summarizes past studies which have compared expert systems and section 4 discusses the expert-system projects with which we have been involved to give the reader an understanding of the basis for our conclusions. Section 5 then discusses the comparative advantages and disadvantages of the three approaches, and section 6 suggests considerations in selecting an appropriate method for building a particular knowledge-based system. Finally, section 7 discusses current and future research needs.

2. Background: three methods for implementing expert systems

In general, an expert system consists of two basic components, a domain-specific knowledge base and a domain-independent inference mechanism. The knowledge base consists of data structures which represent general problem-solving information for some application area. The inference mechanism is a computer program which uses the information in the knowledge base along with problem-specific input data to generate useful information about a specific case.

Various methods have been used to build expert systems in the past, and this section briefly discusses three of them: one form of statistical pattern classification, rule-based deduction, and frame-based abduction. The basic properties of these three approaches are shown in Table 1, and the details are discussed below.

TABLE 1
Three methods for constructing expert systems

Method	Representation	Inference method	Examples
Statistical pattern classification (mainly Bayesian)	*a priori* and conditional probabilities, discriminant functions, etc.	Calculation of posterior probabilities, calculation of discriminant score, etc.	Ben-Bassat *et al.* (1980) deDombal (1975) Gustafson *et al.* (1977) Knapp, Levi, Lurie & Westphal (1977) Matthys, Fischer, Ulrichs & Ruhle (1979) Templeton *et al.* (1967) Warner *et al.* (1964) Zagoria, Reggia, Price & Banko (1981) Zagoria & Reggia (1983)
Rule-based deduction	Conditional rules	Deduction	Campbell *et al.* (1982) Davis *et al.* (1977) Fagan (1979) Futo, Darvas & Szeredi (1978) Kunz *et al.* (1978) Nathanson (1984) Reggia (1978) Reggia (1980) Reggia & Perricone (1981) Reggia, Tabb, Price, Banko & Hebel (1984) Shortliffe (1976) VanMelle *et al.* (1981) Weiss, Kulikowski & Safir (1978)
Frame-based abduction	Frames, semantic networks	Hypothesize-and-test	Aikins (1980) Catanzarite & Greenburg (1979) Miller *et al.* (1982) Mittal, Chandrasekaran & Smith (1979) Pauker *et al.* (1976) Pople, Myers & Miller (1975) Reggia (1981) Reggia, Nau & Wang (1983) Shubin & Ulrich (1982)

2.1. STATISTICAL PATTERN CLASSIFICATION

For our discussion of statistical pattern classification, we will focus on Bayesian classification as the most widely used statistical approach. A more general discussion of statistical pattern classification can be found in any of several existing texts on this topic (e.g. Duda & Hart, 1973). Domain-specific problem-solving knowledge in

Bayesian systems is represented as tables of probabilities, including (1) prior probabilities of outcomes and (2) conditional probabilities of problem features given each possible outcome. The inference mechanism applies Bayes' Theorem to this information to calculate the probability of each possible outcome when given a particular case.

More specifically, suppose we are given the problem features M for a specific case and want to estimate the probability of each of n mutually exclusive and exhaustive outcomes $C_1, C_2, \ldots, C_n$ which could occur. Let:

$P(C_i)$ be the prior probability of C_i (i.e. how commonly C_i occurs in the general population);
$P(M|C_i)$ be the conditional probability of M, given the presence of C_i (i.e. how often C_i would be associated with the set of problem features represented by M);
$P(C_i|M)$ be the conditional probability of C_i given the presence of M (i.e. how often C_i would occur given the particular set of problem features M, also called the posterior probability).

Then Bayes' Theorem can be invoked to generate $P(C_i|M)$ for each C_i given M, and is written as:

$$P(C_i|M) = \frac{P(C_i)P(M|C_i)}{\sum_{j=1}^{n} P(C_j)P(M|C_j)}$$

Formal derivations for Bayes' Theorem can be found in Duda & Hart (1973) and an example of a simple Bayesian knowledge base with a demonstration of how the inference mechanism would work is found in Reggia (1982).

It is generally impractical in real-world domains to derive conditional probabilities for $P(M|C_i)$ for every possible combination of problem features. In practice, therefore, it is usually assumed that the m individual problem features in $M(M_1, M_2, \ldots, M_m)$ are independent (i.e. that the presence of one problem feature does not influence the probability that any other problem feature will be present). Using this independence assumption, it follows that:

$$P(M|C_i) = P(M_1|C_i) * P(M_2|C_i) * \cdots * P(M_m|C_i).$$

Therefore, assuming binary features, only m probabilities $P(M_k|C_i)$ (one for each M_k) are needed in the knowledge base for each outcome C_i rather than the potentially huge set of 2^m probabilities $P(M|C_i)$ which would be needed for each C_i if this assumption was not made.

There are numerous examples of Bayesian classification systems, especially in medicine. In one well-known example, a Bayesian expert system outperformed physicians in diagnosing the cause of acute abdominal pain (about 90% vs. 80% correct diagnoses respectively) (deDombal, 1975). Other expert systems have been developed to diagnose congenital heart disease (Warner, Toronto & Veasy, 1964), identify potential suicide victims (Gustafson, Griest, Stauss, Erdman & Laughren, 1977), classify stroke patients (Zagoria & Reggia, 1983), and diagnose solitary pulmonary nodules (Templeton, Jansen, Lehr & Hufft, 1967).

2.2. RULE-BASED DEDUCTION

A second widely-used method for expert systems, and essentially the "standard" in AI today, is rule-based deduction. In this approach, domain-specific problem-solving knowledge is represented in rules which are basically of the form:

"IF ⟨antecedents⟩ THEN ⟨consequents⟩",

although the exact syntax used may be quite different (e.g. PROLOG). If the antecedents of such a rule are determined to be true, then it logically follows that the consequents are also true. Note that these rules are not branching points in a program, but are non-procedural statements of fact.

The inference mechanism consists of a rule interpreter which, when given a specific set of problem features, determines applicable rules and applies them in some specified order to reach conclusions about the case at hand. Rule-based deduction can be performed in a variety of ways, and rules can be chained together to make multiple-step deductions. (For a fuller description, see Hayes-Roth *et al.*, 1978). In addition, in many systems one can attach "certainty factors" to rules to capture probabilistic information, and a variety of mechanisms can be used to propagate certainty measures during problem solving. MYCIN (Shortliffe, 1976) and PROSPECTOR (Campbell *et al.*, 1982) are two well-known examples of expert systems which incorporate rule-based deduction, and PROLOG successfully uses the fundamental ideas of this method (implemented as a restricted form of resolution wih Horn clauses).

2.3. FRAME-BASED ABDUCTION

A third important method will be referred to in this paper as "frame-based abduction." Here, the domain-specific problem-solving knowledge is represented in descriptive "frames"† of information (Minsky, 1975), and inference is typically based on hypothesize-and-test cycles which model human reasoning as follows. Given one or more initial problem features, the expert system generates a set of potential hypotheses or "causes" which can explain the problem features. These hypotheses are then tested by (1) the use of various procedures which measure their ability to account for the known features, and (2) the generation of new questions which will help to discriminate among the most likely hypotheses. This cycle is then repeated with the additional information acquired. Reasoning from observed facts to the "best explanation" is sometimes referred to as abduction (Reggia, 1985).

As an example, various studies have concluded that diagnostic reasoning is a sequential hypothesize-and-test process (see Reggia, 1982 for a review), so it is not surprising that many of the expert systems built with this approach are directed towards diagnostic problem solving. INTERNIST (Miller, Pople & Myers, 1982), KMS.HT (Reggia & Perricone, 1982; Reggia, Nau & Wang, 1983), PIP (Pauker, Gorry, Kassirer & Schwartz, 1976), and IDT (Shubin & Ulrich, 1982) are typical systems using this approach. For example, KMS.HT is a domain-independent expert-system generator for diagnostic problem solving. In order to simulate hypothesize-and-test reasoning, this system employs a generalized-set-covering model in which there is a universe of all possible manifestations(symptoms) and a universe which contains all possible causes (disorders). For each possible cause, there is a set of manifestations which that cause can explain. Likewise, for each possible manifestation, there is a set of causes which could explain the manifestation. Given a diagnostic problem with a specific set of manifestations which are present, the inference mechanism finds all sets of causes with

† Here, we are not referring to those expert systems in which each rule is a separate "frame." These are considered to be rule-based systems.

minimum cardinality† which could "explain" (cover) all of the manifestations. For a more detailed explanation of the theory underlying this approach and the problem-solving algorithms, see Reggia *et al.* (1983); Reggia, Nau, Wang & Peng (1985b); Nau & Reggia (1984); Peng (1986).

3. Previous studies

Several previous empirical studies have compared expert systems built using different methods to solve the same problem. The expert systems in each study were compared primarily by measuring the accuracy of each system on the same set of cases. We review here those comparative studies which deal with medical problem solving (these studies are summarized in Table 2). We feel these studies are also fairly representative of other application areas. In addition, we describe another study which compares the ease of implementation and the run-time efficiency of several expert systems.

One study compared three statistical models by applying them to screen for thyroid disorders (Nordyke, Kulikowski & Kulikowski, 1971). The three models used were a simple Bayesian model, a linear discriminant model, and a "pattern-recognition method".‡ Five systems were built for each of the three models based on the level of completeness of available data. Each model performed best for one or more of the five completeness levels, with the Bayesian method performing best when all of the data were included (having no higher than a 13·7% misclassification rate). Thus, the methods had comparable performance rates in which the differences depended on the amount and the type of information given to the systems.

Ten discriminant function models of varying complexity, which had been suggested for diagnostic problem solving, were compared in another study (Croft & Machol, 1974). Of the nine models which were applicable to the particular data set tested, all produced relatively similar results. The authors concluded that the derivation of increasingly complex mathematical models similar to the ten used in this study may not be a worthwhile venture.

In another study, three expert systems which generated psychiatric diagnoses were compared by measuring the agreement between the computer systems and clinical diagnoses made on actual cases (Fleiss, Spitzer, Cohen & Endicott, 1972). The three methods used were Bayesian classification, discriminant function classification, and logical decision-tree analysis (branching logic). All three of the methods performed about equally well on test problems derived from the same population that provided the probabilities for the statistical systems. When compared with diagnoses made by clinicians, agreement between programs was about as close as clinicians have been found to agree among themselves. However, when a test sample drawn from a different patient population was used, the decision tree method was found to be more accurate. The authors thus recommend the decision-tree method because of its superior performance on the latter sample and the fact that statistics from a large sample are not needed to develop an expert system using this method.

† Ockham's razor, which states that the simplest explanation is usually the correct one, together with the assumption of independence among causes motivate the requirement of minimum cardinality. For other notions of parsimony, see Peng (1986).

‡ The "pattern recognition method" extracts the most characteristic features of each diagnostic category and uses this as the general problem-solving information. A particular input case is then classified into the category with which its data shares the most features.

TABLE 2
Previous studies comparing different methods for building expert systems

Compared methods	Topic area	Results	Reference
Three statistical pattern classification techniques (Bayesian classification, linear discriminant model, pattern recognition method)	Thyroid dysfunction diagnosis	Comparable performance rates	Nordyke *et al.* (1971)
Ten discriminant function models	Liver disease diagnosis	All produced similar results	Croft & Machol (1974)
Bayesian classification, discriminant function classification, branching logic	Psychiatric diagnosis	Similar performance on one sample; branching logic performed better on another sample	Fleiss *et al.* (1972)
Bayesian classification, rule-based deduction	Dyspepsia diagnosis	Both produced similar results; rule-based system needed less information to determine diagnosis	Fox *et al.* (1980)
Bayesian classification, rule-based deduction	Emergency medical diagnosis	Similar results when normal operating cut-off for each system was applied	Solomon *et al.* (1984)
Rule-based deduction, frame-based abduction	Pulmonary function diagnosis	Frame-based abduction performed better and generated fewer, focused questions	Aikins (1980)

In another study, two different expert systems which diagnosed dyspepsia (gastrointestinal pain) were compared (Fox, Barber & Bardhan, 1980). A successful Bayesian system had previously been developed. A rule-based system was then developed with the purpose of comparing the two systems and determining if this was a viable alternative to the Bayesian method. The performance of the rule-based system was found to be comparable to that of the Bayesian system, with 78% vs 76% correct diagnoses respectively using an initial sample of 50 patient records, and 66% vs 68% correct diagnoses using a somewhat more difficult, second sample of 50 patient records. The use of well-known patterns to form the rules seemed to be able to compensate for the loss of the quantitative precision of the statistics. In addition, the rule-based system asked fewer, focused questions and therefore needed less information in order to determine its diagnosis.

Another experiment also compared rule-based and Bayesian approaches for medical diagnosis (Solomon, Kenevan, Evens, Koschmann & Weil, 1984). A subset of the MEDAS system (Ben-Bassat, Carlson, Puri, Davenport, Schriver, Latif, Smith, Portigal, Lipnick & Weil, 1980) for emergency medical diagnosis was used as the Bayesian system. A knowledge base in the form of rules with certainty factors was then constructed with EMYCIN (VanMelle, Scott, Bennett & Peairs, 1981) to be equivalent to the knowledge in the Bayesian system. Fifty-three test cases were selected from a set of simulated cases prepared by physicians, and the results of test runs were compared. Each system gave its best performance when normal operating thresholds (0.2 for final certainty factors of outcomes in the rule-based system and 0.5 for final probabilities of outcomes in the Bayesian system) were applied, and the systems gave results of similar accuracy.

In another study, rule-based deduction used in PUFF was compared against a second-generation system called CENTAUR which used a frame-based hypothesize-and-test method (Aikins, 1980). Both systems provided interpretations of pulmonary function tests. CENTAUR's diagnoses agreed more frequently with physicians than did PUFF's on 100 test cases. CENTAUR's agreement with the physican who helped create the systems was 91% while PUFF's agreement was 85%. CENTAUR's agreement with a second physician was 84%, and PUFF's was 74%. Furthermore, CENTAUR's consultations were more focused with fewer questions being asked of the user, and this allowed the physicians to understand better how the system was reasoning. The descriptive organization of CENTAUR's knowledge also allowed improved answer justification and easier knowledge acquisition.

In the one study we are aware of in which criteria other than system accuracy were used, four pilot expert systems were built to solve the problem of risk management of a large construction project (Niwa, Sasaki & Ihara, 1984). The four types of models were a simple rule-based deduction system, a structured rule-based deduction system,† a frame-based system (which used a generalization hierarchy supporting inheritance of attributes, Winston & Horn, 1981), and a resolution-based logic system. The four systems were compared subjectively in terms of the difficulties of implementing the knowledge bases and inference mechanisms. They were also compared on the basis of run-time efficiency. It was found that the use of structured knowledge representations (structured rule-based deduction and frame systems) increased run-time efficiency, but the implementation of these systems was more difficult.

In summary, most past studies involve the comparative evaluation of the accuracy of multiple expert systems for the same application domain. In some studies, the various expert systems being compared were built using very different techniques, while in several other studies, the expert systems were built using similar techniques. The major results from these studies indicate that none of the methods is significantly superior to the others. Even where minor differences do exist, one can generally attribute them to differences in the problem-specific information in the knowledge bases as opposed to fundamental differences in the methods being used.

4. Our results

During the last five years, the authors have been involved personally in the development and evaluation of a wide variety of expert systems. These systems are outlined in Table 3, and have involved problems ranging from "toy examples" to real-world systems

† The production rules were divided into several units, based on the temporal order in the model of domain knowledge relationships. Control functions were included in the inference mechanism which allowed the proper unit to be accessed, thereby limiting the number of rules to be searched during usage.

TABLE 3

Some expert systems constructed by our group during the last five years. (SPC stands for Statistical pattern classification, RBD stands for Rule-based deduction, and FBA stands for Frame-based abduction).

Expert-system task	Methods used	Reference
(i) *Medical*		
Stroke diagnosis	SPC	Zagoria & Reggia (1983)
Treatment of transient ischemic attacks	RBD	Reggia *et al.* (1984)
Prediction of arteriorgraphy risk	SPC	Reggia, Pula, Price & Taylor (1981)
Prognosis following subarachnoid hemorrhage	SPC, RBD	Unpublished
Neurological localization in coma	RBD, FBA	Reggia (1978)
Dizziness diagnosis	FBA	Reggia *et al.* (1983)
Coma prognosis	RBD	Unpublished
Acute quadraparesis diagnosis	FBA	Unpublished
Peroneal muscular atrophy diagnosis	FBA	Reggia *et al.* (1983)
Cancer screening	RBD	Reggia & Perricone (1981)
Dementia diagnosis	RBD	Unpublished
(ii) *Non-medical*		
Software acquisition advisor	RBD	Ferrentino (1983)
Software engineering advisor	RBD, FBA	Basili & Ramsey (1985)
Statistical test selection	FBA	Reggia (1981)
Plumbing disorder diagnosis	FBA	Reggia & Perricone (1982)
Chemical-spill diagnosis	SPC, RBD, FBA	Described in this paper
Process planning in automated manufacturing	RBD, FBA	Unpublished
Nuclear reactor monitor	RBD	KES (1984)

intended for eventual use. As an example of the latter, a large rule-based system (about 400 rules) for the difficult medical problem of classifying and treating transient ischemic attacks was implemented and evaluated in a clinical study involving 103 patients (Reggia, Tabb, Price, Banko & Hebel, 1984). Development of other systems has involved statistical pattern classification techniques, rule-based deduction, and frame-based abduction. In addition, for a number of years we have developed and used a domain-independent expert-system generator, and we have recently completed a study of the use of this same software by 70 medical students to build small expert systems (many of these students were computer-inexperienced) (Reggia, Tuhrim & Perricone, 1986). While conclusions drawn from this work must necessarily be partially subjective, it has clearly delineated the issues involved and provided a major influence on the conclusions and recommendations in subsequent sections of this paper.

In addition to the work outlined above, we have undertaken a number of studies specifically to compare the three methods for building expert systems discussed in this paper. The goal in each study was not only to evaluate system performances comparatively, but also to evaluate the advantages and disadvantages of each method in terms of ease of knowledge representation and acquisition. As noted earlier, most previous empirical studies have not addressed these issues. Most of the applications which we have used in this fashion were real-world problems including neurological localization, stroke-related problems, estimating prognosis following subarachnoid hemorrhage, process planning in automated manufacturing, and software development management. However, to illustrate clearly the approach to comparative analysis we used in each case, we briefly describe below a collection of three small expert systems constructed for the same "toy problem" of diagnosing the cause(s) of a chemical spill contaminating a creek. This simple application is used to illustrate the basic approach we employed to do comparative analyses because the domain-specific concepts involved should be relatively easy for most readers to understand.

In the chemical-spill problem, an expert system receives as input a set of manifestations such as high acidity of creek water or changed water color. The goal of the expert system is to produce the name(s) of the chemical(s) responsible for the spill (more than one chemical could be present). The knowledge for the expert systems was originally given in a descriptive, natural-language format. Each type of spill which could occur was listed along with the manifestations it might produce (see Appendix).

KMS (Reggia & Perricone, 1982), an experimental domain-independent system which can be used to build Bayesian, rule-based, and frame-based systems, was then used to construct the three expert systems for this problem. In order to keep the systems consistent, the causes and manifestations used were identical in all three cases. Figure 1 shows a sample section of each knowledge base. The certainty factors used for rules, the probabilities used for statistical pattern classification, and the measures of likelihood used in frames were also kept fairly consistent. For example, given a specific chemical spill, a manifestation with a "high" likelihood of being present in the frame-based system was given a 0·75 conditional probability in the Bayesian system. The rule-based system presented some problems because the number of times this manifestation could appear as a result of other chemical spills also had to be taken into account. While many other formulations of the rules could have been used, we elected for this toy problem simply to use rules of the form shown in Fig. 1(b). (This was not the case in most of the other comparative studies where the antecedents were composed of more complex logical relationships.)

For the chemical-spill problem, it was observed that the abductive frame-based system was, by far, the easiest and most natural to develop. This was not surprising in that a diagnostic problem was involved, and the knowledge was originally presented in a descriptive, frame-like fashion (see Appendix). The three expert systems were each tested on a small set of simulated problems. All three systems worked about equally well for simple test cases involving only one chemical spill, consistent with the previous studies done by others that were described above. However, for test cases where multiple chemical contaminants were present, the frame-based abductive expert system clearly outperformed the other two approaches.

Unlike the chemical-spill study, our other comparative studies involved real-world applications. In most cases, little effort was made to make the expert systems of these

FIG. 1. A small section of each of the three knowledge bases used in the chemical spill expert systems. (a), Frame-based abduction—the letters enclosed in angular brackets indicate a likelihood factor. In the first frame for example, the "⟨H⟩" next to "Acidic" indicates that there is a high likelihood that the pH is acidic when there is a sulfuric acid spill. The other likelihood factors are A (always), N (never), M (medium), and L (low). (b), Rule-based deduction—the numbers enclosed in angular brackets are certainty factors. (c), Bayesian classification—the number next to "Sulfuric Acid" is the prior probability of a sulfuric acid spill occurring. The third row of conditional probabilities under "Sulfuric Acid" corresponds to the acidic/alkaline manifestation. For example, the first probability of 0·75 indicates the probability that the water is acidic given that there is a sulfuric acid spill.

(a)

Sulfuric Acid
[Description:
Spill Alarm = On ⟨A⟩;
Month of Year = May ⟨H⟩, June ⟨H⟩;
PH = Acidic ⟨H⟩;
Spectrometry Results = Sulfur ⟨A⟩],
Petroleum
[Description:
Spill Alarm = On ⟨A⟩;
Month of Year = July ⟨H⟩, August ⟨H⟩, September ⟨H⟩;
Water Color = Black;
Photometry Results = Oily;
Spectrometry Results = Carbon ⟨H⟩;
Specific Gravity of Water = Decreased],

(b)

PH1 IF PH = Acidic
THEN Type of Spill = Sulfuric Acid ⟨0·37⟩.
COL2 IF Water Color = Black
THEN Type of Spill = Petroleum ⟨1·0⟩.
PHOTO1 IF Photometry Results = Oily
THEN Type of Spill = Petroleum ⟨0·33⟩.
SPECT1 IF Spectrometry Results = Carbon
THEN Type of Spill = Petroleum ⟨0·22⟩.
SPECT2 IF Spectrometry Results = Sulfur
THEN Type of Spill = Sulfuric Acid ⟨0·55⟩.
GRAV1 IF Specific Gravity of Water = Decreased
THEN Type of Spill = Petroleum ⟨0·34⟩.

(c)

Attributes:
\INPUT ATTRIBUTES
Spill Alarm (SGL): on, off.
Month of Year (SGL):
April, May, June, July, August, September.
⋮
\INFERRED ATTRIBUTE
Type of Spill [DETERMINANTS:*] (SGL):
Sulfuric Acid ⟨0·08⟩
1·00 0·00;
0·11 0·28 0·28 0·11 0·11 0·11;
0·75 0·24 0·01;
0·98 0·01 0·01;
0·99 0·01;
0·99 0·01
0·99 0·01;
0·00 1·00;
0·99 0·01;
0·01 0·98 0·01,
Petroleum ⟨0·07⟩
1·00 0·00;
0·09 0·08 0·08 0·25 0·25 0·25;
0·01 0·98 0·01;
0·49 0·01 0·50;
0·50 0·50;
0·99 0·01;
0·25 0·75;
0·99 0·01;
0·99 0·01;
0·50 0·49 0·01,

studies maximally consistent with each other as in the "toy" chemical-spill expert system. Instead, the goal was maximum performance with each method, although in fairness, an effort was made to incorporate the same information and scope in each system for the same application. For example, two expert systems for predicting outcomes following subarachnoid hemorrhage were implemented, one using Bayesian classification and one using rule-based deduction. Both were based on the same published data for this clinical problem (Yoshimoto, Uchida, Kaneko, Kayama & Suzuki, 1979), and both systems performed well when tested using data from 40 patients.

As another example, an ongoing comparative study involves the development of two prototype expert systems to aid in software engineering management (Basili & Ramsey, 1985). These systems work as follows. First, it is determined whether or not a software project is following normal development patterns by comparing measures such as programmer hours per line of code against historical, environment-specific baselines of such measures. The "manifestations" detected by this comparison, such as an abnormally high rate of programmer hours per line of code, serve as input to each expert system, and each system then provides the causes, such as low productivity, for any abnormal software development patterns. The knowledge bases consist of relationships between various potential causes (such as poor testing or unstable specifications) and abnormal values of measures.

The two methods used to build these expert systems were rule-based deduction and frame-based abduction. Furthermore, these two systems were intentionally built to be as consistent with one another as possible; the causes and manifestations used were identical in both cases, as were the relationships between them. The initial knowledge was derived from empirical software engineering research and organized in a table format, so the first sets of simple rules (which contained only one antecedent for each rule) and frames were straightforward to develop. However, the situation became more complex as additional knowledge was added, reflecting the fact that the science of software engineering is still ill-defined. Also, an attempt was made to develop rules with complex antecedents, but the more involved patterns necessary to develop these rules are not known yet.

A preliminary evaluation of the two expert systems was performed (Basili & Ramsey 1985). The method used to do the evaluation was to compare the interpretations (causes) provided by the expert systems for particular projects against what actually happened during the development of those projects, thereby obtaining a measure of agreement. It was found that the rule-based system performed better, agreeing with 45% of the actual interpretations; the frame-based system agreed with 33% of the actual interpretations. The expert systems were viewed as performing moderately well given that (1) so much of the knowledge and relationships are unclear in the field of software engineering and (2) only nine metrics (manifestations) were used to determine the interpretations. Both expert systems provided the exact same interpretations for seven out of nine projects. In the two cases where the expert systems differed, the frame-based system provided very few interpretations which covered the entire set of manifestations, while the rule-based system provided more interpretations and agreed with more of the actual interpretations. Also, these differences resulted in 31% fewer extra interpretations for the frame-based system. However, when dealing with uncertain relationships, it is better to provide extra interpretations (and then to let the user decide which are applicable) than to aim for the "one" correct explanation and thereby miss correct explanations. Therefore, the authors concluded that a rule-based deduction system using simple rules seems more applicable to the field of software engineering than does a frame-based abduction system which determines explanations of minimum cardinality. This conclusion can probably be extended to include other ill-defined application areas as well.

5. Comparative advantages and disadvantages

As a result of the work summarized in sections 3 and 4, we can state several conclusions about the relative strengths and weaknesses of methods for building expert systems. While we have attempted to be as comprehensive as possible, we make no claim to completeness. Our intent is to provide a first attempt to organize the relative merits of the different methods involved as a guide to the practitioner, and to provide a focus for further research to confirm and extend the points made here.

5.1. GENERAL COMMENTS

There are common strengths and weaknesses for all three of the methods discussed in this paper. All three methods can generally be adapted to any problem involving the selection of different alternatives. Furthermore, they have all shown reasonable performances in various expert systems in a wide range of application domains.

A second strength which all the methods share is a strong theoretical foundation. Statistical pattern classification is based on probability theory, rule-based deduction is based on deductive logic (e.g. first-order predicate calculus), and frame-based abduction can be based on the theory of set covering which provides a formal theory of diagnostic inference (Reggia *et al.*, 1983; Reggia *et al.*, 1985*b*).

Third, all three expert-system methods can support answer justification, the ability to explain to the user how or why a certain result was derived. This ability is generally perceived to be of major importance by potential expert system users who may be reluctant to trust a machine-generated recommendation unless it is clear why that recommendation has been made (Teach & Shortliffe, 1981). The conventional wisdom in AI has been that rule-based systems, with their ability to support a limited form of answer justification (Davis, Buchanan & Shortliffe, 1977), have a major advantage over other methods for expert systems. However, it has recently been shown that Bayesian classification systems can also support answer justification by analysing and presenting the problem features most responsible for the relative ranking of outcomes (Reggia & Perricone, 1985). Answer justification for frame-based abductive expert systems based on an analysis of the causal associations in the knowledge base has also been described recently (Reggia, Perricone, Nau & Peng, 1985*a*). Thus, it appears that many different expert-system methods involving associative knowledge can support answer justification, and that previous development of answer justification primarily in rule-based systems simply reflected different research priorities. It might even be argued that among the three methods discussed here, rule-based systems are the weakest with respect to answer justification: for many domains, citing artificial program-specific rules is far less educational than organizing and citing naturally occurring associations between domain-specific concepts in Bayesian classification and frame-based abduction.

Just as all three methods have common strengths, all three are limited in their ability to represent conveniently certain types of information such as spatial and temporal knowledge. In addition, all of the methods have a "shallow" nature, using simple associative mappings between problem features and outcomes rather than a "deep" reasoning mechanism which would involve many levels of mappings. None of these models is an adequate model of human cognition, and none is especially robust in handling "noisy data".

Although the three methods are similar in some respects, they are quite different in other respects. We now turn our attention to the relative strengths and weaknesses of each individual method.

5.2. STATISTICAL PATTERN CLASSIFICATION

A major advantage to statistical pattern classification is that it has been repeatedly used with proven success. If one has the needed probabilities and a domain-independent expert-system generator, these systems are very easy to organize and implement. For example, one fully operational real-world example took a total of only 8h to create! (Zagoria & Reggia 1983). In addition, these systems are based on real statistics gathered from actual cases in the problem domain, so the knowledge base many contain knowledge which the experts in the field are not aware of or cannot verbalize (Solomon *et al.*, 1984).

However, statistical pattern classification systems face several major disadvantages. Perhaps the biggest disadvantage to this method is that it requires the availability of exact probabilities. While these probabilities can often be measured, this is usually a time-consuming and very costly task, and estimations of probabilities by domain experts has repeatedly been shown to be unsatisfactory for such systems (Shapiro, 1977; Tversky, 1974; Leaper, Horrocks, Staniland & deDombal, 1972). Another problem is that certain unrealistic assumptions must be made to use some statistical techniques. For example, in Bayesian classification, the outcomes involved must be mutually exclusive. Unfortunately, in many problems, multiple simultaneous outcomes may be present (e.g. a patient may have multiple diseases). The Bayesian chemical-spill system did not perform well for multiple chemical spills, an observation often made with real-world expert systems (e.g. Zagoria & Reggia, 1983). Another assumption, not required by Bayes' Theorem *per se* but generally required in practice, is that the problem features are independent. In very many application domains, problem features are not independent, and making this assumption incorrectly can result in a degraded performance and compromise theoretical claims of optimality (Fryback, 1978; Norusis & Jacquez, 1975*a*). A modification of Bayesian classification to handle multiple simultaneous outcomes has been suggested in Ben-Bassat *et al.* (1980), and a variety of solutions has been proposed for handling the problem of assuming that the problem features are independent (Davies, 1972; Fryback, 1978; Norusis & Jacquez, 1975*b*).

Another disadvantage of most traditional Bayesian and statistical systems is that one must have ALL of the relevant information about a case before one can use the system. This can be unrealistic if one is working on such things as a diagnostic problem where one would need certain (perhaps expensive) diagnostic tests only in certain situations. Some solutions to this are that unknown attributes can simply be dropped from the calculation (Reggia, Pula, Price & Perricone, 1980) or one can use an elaborate sequential application of Bayes' Theorem (Gorry & Barnett, 1968).

Finally, geographic location may be an important factor when using statistical pattern classification. The probabilities may be dependent on the specific environment from which they were gathered and may not be transportable to other locations. However, some evidence has suggested that some statistical systems may be more transferable than previously suspected (Zagoria & Reggia, 1983; Zoltie, Horrocks & deDombal, 1977).

5.3. RULE-BASED DEDUCTION

Rule-based deduction has also exhibited proven successes. Examples include PROSPECTOR (Campbell *et al.*, 1982) and HELP (Nathanson, 1984; Pryor *et al.*, 1984) which were discussed earlier. This method provides the ability to chain associative information to make deductions, a capability that can be very useful. Some say rules provide a good model of human reasoning (Larkin, McDermott, Simon & Simon, 1980), but this is a controversial issue (Davis *et al.*, 1977; Aikins, 1980). The use of rules allows for non-numeric, judgemental knowledge because one does not need exact probabilities.

One of the disadvantages of rule-based deduction is that it is often difficult to represent knowledge in terms of rules, especially if one already has available descriptive information, such as knowledge from texts and experts. This was true of the chemical-spill system, and this type of problem has been encountered during many other attempts to write rule-based systems in the past (Reggia, 1978; Shortliffe, 1976; Buchanan, Sutherland & Feigenbaum, 1970). Rules are not a convenient way to organize knowledge in many domains, and one often needs to introduce non-intuitive intermediate problem features for bookkeeping purposes.

Part of this problem is due to the fact that the "directionality" of production rules can present problems.† For example, the rules used in diagnostic rule-based systems are typically of the form "IF ⟨manifestations⟩ THEN ⟨cause⟩". However, much of the knowledge used to create such rules as it is familiar to domain experts is descriptive and goes in the opposite direction: if some cause is present, then certain manifestations will typically occur. As a result, many people claim that one is "thinking backward" when using rules (Davis *et al.*, 1977). For example, as part of a research project in electronic diagnosis (J. Miller, pers. comm.), a group of electronic technicians were given a brief introduction to the concept of production rules. They were then asked to write some production rules describing an electronic circuit for use in fault diagnosis. The rules they produced were almost all of the form

"IF ⟨cause⟩ THEN ⟨manifestations⟩".

Presumably the technicians were specifying their knowledge in an intuitively familar form, more along the lines of describing each cause rather ~~than providing fixed~~ rules for diagnosis.

Another source of difficulty with rule development is that one must include all of the necessary context for a rule's application in its antecedent clauses (Davis *et al.*, 1977). Since all of the relevant factors for applying a rule are not always obvious, this can result in errors of omission in the antecedents of rules. In addition, the interpretation of some problem features can be very context-dependent. As an example taken from neurology, the "meaning" of a unilateral dilated and fixed pupil depends on whether

† By "directionality", we are not referring to the way in which the rules are used by the inference mechanism (i.e. forward vs backward chaining).

long-tract findings are present, whether a patient is awake or unconscious, whether the neck is supple or rigid, whether topical eye medications are being used, etc. (Reggia, 1978). To encode adequately all of the necessary contextual information into antecedent clauses could require a large number of rules which interact in complicated ways. This problem is magnified in the presence of multiple simultaneous disorders. (For a further discussion of these difficulties, see Nau & Reggia, 1984). Furthermore, there are generally many ways to organize rules, and it is usually not clear *a priori* which way is best. Often, one ends up rewriting the entire rule system (Reggia, 1978).

The "directionality" of certainty factors can also present problems. As a simple example, in the chemical-spill problem one may know that if a hydrochloric or other acid spill occurs there is a high likelihood that the water will become acidic. If one were asked to choose a conditional probability for this association, one might choose 0·95. However, the knowledge base would generally contain rules which go in the opposite direction such as:

"IF ⟨creek water is acidic⟩ THEN ⟨chemical-spill is sulfuric acid⟩ ⟨CF1⟩."
"IF ⟨creek water is acidic⟩ THEN ⟨chemical-spill is hydrochloric acid⟩ ⟨CF2⟩."
"IF ⟨creek water is acidic⟩ THEN ⟨chemical-spill is carbonic acid⟩ ⟨CF3⟩."

Here, CF1, CF2 and CF3 stand for the certainty factors of drawing the consequents if the antecedents are true. Now, if the creek water is acidic, how certain is one that the spill is hydrochloric acid? If the prior probabilities of outcomes are available, one could derive certainty factors from the given information, but without that knowledge, it is often difficult to determine them, and subsequently the certainty factors used in real-world expert systems are, at best, rather arbitrary.

5.4. FRAME-BASED ABDUCTION

One advantage of frame-based abduction is that for many applications, frames are easy and natural to write. They can often be taken almost directly from descriptive information in a textbook or paper. For example, the chemical-spill system was derived from object-oriented descriptive knowledge, and the frames were therefore easily written. Another advantage is that all of the information about each outcome is placed in one frame so context-dependent information can be handled more readily. Further, in diagnostic and other problems involving selection, this method can work very well even when multiple causes/disorders/selections are involved. This was apparent with the chemical-spill system, and has been observed for a number of real-world expert systems. Finally, the hypothesize-and-test algorithms often used in frame-based abduction focus on the most likely outcome, thereby typically generating fewer questions. An obvious advantage to this is that less time is consumed in a question-answering session with a user. In addition, people sometimes feel more comfortable using such systems because they can understand the reasoning of the system (Aikins, 1980).

The major disadvantages of this method can be summarized by observing that it is the most experimental of the three methodologies. Frame-based abductive inference has not been studied or used sufficiently in real-world applications that have objectively assessed accuracy, making it difficult to understand fully the strengths and weaknesses of this approach. Even the relatively small empirical trials with real-world systems that have been done have exhibited somewhat limited performance (e.g. Miller *et al.*, 1982).

Many technical issues concerning this method also remain to be resolved. The optimal point at which to terminate the question generation and decision-making process remains to be determined. While hypothesize-and-test algorithms focus on the most likely hypotheses/outcomes and thereby limit the questions being asked, such a strategy may at times leave some important but unrelated information undiscovered. For example, a diagnostic expert system may move along one direction and determine a set of disorders which accounts for all of the symptoms it knows about. However, it is very possible that a "pure" abductive system might never ask about some unrelated symptom which actually exists, and therefore, its final set of diagnostic explanations would be incomplete.

Secondly, how to select optimally the next question to ask the user remains to be determined. A variety of heuristic approaches have been adopted for question generation (Miller *et al.*, 1982; Reggia *et al.*, 1983), but all of these approaches appear to be limited in their completeness, naturalness, and theoretical basis.

Finally, the abductive inference mechanisms used with this approach involve relatively complex algorithms and problem-solving techniques which potentially have exponential time complexity (although real-world cases fall in the low, flat portion of the exponential curve) (Reggia *et al.*, 1985*b*). Even so, they are weak models of the actual abductive inference methods used by people. It is not clear, for example, in abductive diagnostic problem solving, exactly how a person decides that one set of disorders is a plausible explanation for a given set of manifestations and another set is not (see Peng, 1986 for further details).

6. Recommendations for method selection

Given that all existing methods for building expert systems face limitations to their general applicability, an important question for the individual planning to develop expert systems for various applications is which method(s) to select. There is surprisingly little information on this topic in the literature, so in this section we describe a set of criteria for method selection that we have found useful based on both our experiences and that of others as outlined above. These criteria are based on three main factors: the pre-existing format of the application knowledge, the type of classification that is desired, and the amount of context-dependence inherently present in a problem. These factors are discussed below.

One important factor in method selection is the pre-existing format of the application knowledge. Since this format is often very natural for the particular knowledge involved, keeping the knowledge in that same format has a certain intuitive appeal. Furthermore, it requires much less work to keep the knowledge in the original form, rather than to transform it into some other representation. This was true for the chemical-spill system where the frames for the abductive system were easily derived from descriptive information. The software engineering system referred to earlier was developed from information which existed in a table format, and rules were easily derived from this information.

A second important consideration is the type of classification desired. Some problems inherently involve probabilistic inferences. For example, predictions of outcome (e.g. prognosis) can often only be done in a probabilistic manner. On the other hand, for some problems one may desire to make predominantly categorical inferences, such as for part selection in automated manufacturing or assignment of a patient to an artificial disease "stage". Still other problems involve a significant mixture of probabilistic and categorical inferences. For example, diagnostic problem solving and problems like determining the type of chemical spill often require a combination of these two types of classification. Clearly, one wants to select an approach to knowledge management for a specific problem whose inference mechanism embodies the type(s) of classification involved.

A third factor is how context-dependent the inference process needs to be. If inferences depend on just a few input features, then rule-based deduction may be very appropriate. On the other hand, if inferences depend on numerous input features, then frame-based abduction might be a preferable approach. Writing a set of rules would be difficult in this latter case because all of the context for using each rule would have to be included in the antecedents of that rule. Even assuming that one could identify *a priori* all of this relevant context, the resulting knowledge base would be a potentially huge set of rules. In contrast, by encoding the knowledge in a descriptive fashion, one largely defers the problem of context until problem-solving time.

Given the above factors and the comparative advantages and disadvantages we have cited, guidelines can be proposed for selecting an approach to building an expert

TABLE 4

Method selection criteria (+ means positive influence, and − means negative influence. SPC stands for Statistical pattern classification, RBD stands for Rule-based deduction, and FBA stands for Frame-based abduction).

Factor	SPC	RBD	FBA
Predominant pre-encoding organization of knowledge			
Branching Logic	−	+++	+
Rules	−	+++	−−
Description/tables	+	++	+++
Uncertain/poorly formed	−	++	+
Probabilities			
Available/easily collected	+++	+	+
Otherwise	−−−	+	+
Predominant type of classification			
Mutually exclusive outcomes			
Probabilistic	+++	+	+
Categorical	−	++	+
Mixed		+	+
Multiple simultaneous outcomes			
Probabilistic		+	+
Categorical	−−−	++	++
Mixed	−−	+	+
Context-dependence			
Small	+	+	+
Large	−	−	+++
Some example application areas			
Small diagnostic problems	+	+	+
Large diagnostic problem	−−	+	+++
Course of action to take		++	
Predicting outcome	++	+	−

system for a specific problem. Table 4 summarizes these guidelines for method selection, and they are briefly discussed in the following paragraphs.

Statistical pattern classification in the form of Bayesian classification would be appropriate to consider when: (1) relevant probabilities, or the data needed to derive them, are readily available; (2) the outcomes where probabilities are being estimated are mutually exclusive; (3) input features are relatively independent of one another and (4) predominantly probabilistic inferences are involved. In general, this method seems to work best for limited size problems involving selection of one outcome from a fixed set of alternatives (e.g. in medicine, small diagnostic or prognostic problems). If non-Bayesian statistical methods are used, restrictions (2) or (3) might be relaxed.

Similarly, rule-based deduction may be most appropriate when (1) the underlying knowledge is already organized as rules or in a table format, (2) the type of classification involved is predominantly categorical and (3) there is not a large amount of context dependence. The types of problems for which this method is suited thus include "screening" for particular situations, categorizing a situation into well-defined artificial categories (e.g. staging a disease), and selecting a course of action to be taken for an established situation.

Finally, the use of frame-based abduction should be considered when: (1) the underlying knowledge pre-exists in a descriptive format such as that often found in a textbook or review article; (2) there is a mixture of probabilistic and categorical classification involved, particularly in situations requiring diagnostic problem solving; (3) there is a large amount of context dependence and (4) there are potentially multiple simultaneous outcomes to be selected (e.g. multiple machining operations or diagnoses). In medicine, for example, this approach to knowledge processing appears especially suited for large general diagnostic problems.

7. Discussion

In contrast to some opinions expressed in the literature, this paper has argued that there is not a single "best method" for building expert systems. The real issue for consideration is which methods are best suited for a specific application problem. We therefore have tried to delineate the relative strengths and weaknesses of three important methods in this paper, and to provide guidelines for the selection of an appropriate method. Obviously, there are problems for which none of these methods is ideal and for which no existing methods are ideal (see the discussion of common weaknesses above).

While the observations and recommendations we have presented should be viewed as preliminary, they provide at least initial guidelines for the potential expert system developer, and they can serve as a starting point for further discussion and research on this important topic. In addition to a need for fundamental research on methods for knowledge representation and automated inference (such as methods for dealing with spatial and temporal knowledge), further comparative studies of the various methods used for expert systems need to be undertaken in an objective, controlled, quantitative fashion. Most previous research has focused primarily on accuracy of different methods for constructing expert systems. The few existing comparative studies which have addressed the issues of knowledge acquisition and representational adequacy, including the work described in this paper, are largely exploratory and therefore qualitative and subjective in nature. One could therefore take virtually any conclusion reached in this or previous papers as a *hypothesis* to be studied quantitatively in a controlled setting. Such a study would involve a group of individuals each building the same expert system (e.g. in a classroom setting). Only through such future research will the match between knowledge-processing method and application become clearly defined and founded on a strong empirical basis.

The work described in this paper has involved the efforts of many individuals over the last several years. Funding for the several projects described has been provided by NASA (NSG-5123), NIH (P50-NS-16332) and NSF (DCR-8451430). Dr Nau is the recipient of a Presidential Young Investigator award from NSF. Industrial sponsorship of the Department of Computer Science is provided by Software Architecture and Engineering, AT&T Information Services, General Motors, IBM and Martin Marietta.

References

AIKINS, J. (1980). *Prototypes and Production Rules—A Knowledge Representation for Computer Consultations, Memo HPP*-80-17. Stanford Heuristic Programming Project. Department of Computer Science, Stanford University.

BASILI, V. & RAMSEY, C. (1985). ARROWSMITH-P—A prototype expert system for software engineering management. In KARNA, K., Ed., *Proceedings of Expert Systems in Government Symposium*, pp. 252–264, Washington, D.C.: IEEE Press.

BEN-BASSAT, M., CARLSON, R. W., PURI, V. K., DAVENPORT, M. D., SCHRIVER, J. A., LATIF, M., SMITH, R., PORTIGAL, L. D., LIPNICK, E. H. & WEIL, M. H. (1980). Pattern-based interactive diagnosis of multiple disorders: the MEDAS system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 148–160.

BUCHANAN, B., SUTHERLAND, G. & FEIGENBAUM, E. (1970). Rediscovering some problems of artificial intelligence in the context of organic chemistry. In MELTZER & MICHIE, Eds, *Machine Intelligence*, **5**, pp. 253–280. Edinburgh: Edinburgh University Press.

CAMPBELL, A. N., HOLLISTER, V. F., DUDA, R. O. & HART, P. E. (1982). Recognition of a hidden mineral deposit by an artificial intelligence program. *Science*, **217**, 927–928.

CATANZARITE, V. & GREENBURG, A. (1979). Neurologist—a computer program for diagnosis in neurology. *Proceedings of the Third Symposium on Computer Applications in Medical Care*, pp. 64–72. Washington, D.C.: IEEE Press.

CROFT, D. & MACHOL, R. (1974). Mathematical methods in medical diagnosis. *Annals of Biomedical Engineering*, **2**, 69–89.

DAVIES, P. (1972). Symptom diagnosis using Bahadur's distribution. *Bio-Medical Computing*, **3**, 307–312.

DAVIS, R., BUCHANAN, B. & SHORTLIFFE, E. (1977). Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence*, **8**, 15–45.

DAVIS, R., AUSTIN, H., CARLBOM, I., FRAWLEY, B., PRUCHNIK, P., SNEIDERMAN, R. & GILREATH, J. A. (1981). The dipmeter advisor: interpretation of geological signals. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 846–849. Wm. Kaufmann, Inc.

DEDOMBAL, F. (1975). Computer assisted diagnosis of abdominal pain. In ROSE, J. & MITCHELL, J., Eds., *Advances in Medical Computing*, pp. 10–19. New York: Churchill-Livingston.

DUDA, R. & HART, P. (1973). *Pattern Classification and Scene Analysis*, pp. 10–39. New York: Wiley-Interscience.

FAGAN, L. (1979). A symbolic processing approach to measurement interpretation in the intensive care unit. In *Proceedings of the Third Symposium on Computer Applications in Medical Care*, pp. 30–33. Washington, D.C.: IEEE Press.

FERRENTINO, A. (1983). Software manager's workstation. In *Proceedings of Electronics and Aerospace Systems Conference*, Washington, D.C.

FLEISS, J., SPITZER, R. L., COHEN, J. & ENDICOTT, J. (1972). Three computer diagnosis methods compared. *Archives of General Psychiatry*, **27**, 643–649.

FOX, J., BARBER, D. & BARDHAN, K. (1980). Alternative to Bayes? A quantitative comparison with rule-based diagnostic inference. *Methods of Information in Medicine*, **19**, 210–215.

FRYBACK, D. (1978). Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research*, **11**, 423–434.

FUTO, I., DARVAS, F. & SZEREDI, P. (1978). The application of PROLOG to the development of QA and DBM Systems. In GALLAIRE, H. & MINKER, J., Ed., *Logic and Databases*, pp. 347–376. New York: Plenum Press.

GORRY, G. & BARNETT, G. (1968). Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, **1**, 490–507.

GUSTAFSON, D., GRIEST, J., STAUSS, F., ERDMAN, H. & LAUGHREN, T. (1977). A probabilistic system for identifying suicide attemptors. *Computers and Biomedical Research*, **10**, 83–89.

HAYES-ROTH, F., WATERMAN, D. & LENAT, D. (1978). Principles of pattern-directed inference systems. In WATERMAN, D. A. & HAYES-ROTH, F., Eds, *Pattern-Directed Inference Systems*, pp. 577–601. New York: Academic Press.

KES (1984). *Knowledge Engineering System, Knowledge Base Author's Reference Manual.* Software Architecture & Engineering.

KNAPP, R., LEVI, S., LURIE, D. & WESTPHAL, M. (1977). A computer-generated diagnostic decision guide—a comparison of statistical diagnosis and clinical diagnosis. *Computers in Biology and Medicine*, **7**, 223–230.

KOWALSKI, R. (1980). In *SIGART Newsletter* (special issue on Knowledge Representation), February 1980.

KUNZ, J., FALLAT, R., MCCLUNG, D., OSBORN, J., VOFTERI, B., NII, H., AIKINS, J., FAGAN, L. & FEIGENBAUM, E. (1978). *A Physiological Rule-based System for Interpreting Pulmonary Function Test Results*, HPP-78-19. Department of Computer Science, Stanford University.

LARKIN, J. H., MCDERMOTT, J., SIMON, D. P. & SIMON, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science*, **4**, 317–345.

LEAPER, D., HORROCKS, J., STANILAND, J. & DEDOMBAL, F. (1972). Computer-assisted diagnosis of abdominal pain using estimates provided by clinicians. *British Medical Journal*, **4**, 350–354.

MATTHYS, H., FISCHER, J., ULRICHS, H. & RUHLE, K. (1979). Functional patterns of different lung diseases for computer-assisted diagnostic procedures. *Progress in Respiration Research*, **11**, 188–201.

MCDERMOTT, J. & STEELE, B. (1981). Extending a knowledge-based system to deal with ad hoc constraints. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 824–828. Wm Kaufmann Inc.

MILLER, R., POPLE, H. & MYERS, J. (1982). Internist-1: an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, **307**, 468–476.

MINSKY, M. (1975). A framework for representing knowledge. In WINSTON, P., Ed., *The Psychology of Computer Vision*, pp. 211–277. New York: McGraw-Hill, Inc.

MITTAL, S., CHANDRASEKARAN, B. & SMITH, J. (1979). Overview of MDX—a system for medical diagnosis. *Proceedings of the Third Symposium on Computer Applications in Medical Care*, pp. 34–46. Washington, D.C.: IEEE Press.

NATHANSON, M. (1984). Physicians get HELP in cutting patient stays. *Modern Healthcare*, April, 140–145.

NAU, D. S. (1983). Expert computer systems. *Computer*, February, 63–85.

NAU, D. S. & REGGIA, J. A. (1984). Relationships between deductive and abductive inference in knowledge-based diagnostic expert systems. In KERSCHBERG, L., Ed., *Proceedings of the First International Workshop on Expert Database Systems*, pp. 500–509. Menlo Park, CA: Benjamin/Cummings Publishing Co.

NIWA, K., SASAKI, K. & IHARA, H. (1984). An experimental comparison of knowledge representation schemes. *AI Magazine*, Summer, 29–36.

NORDYKE, R. A., KULIKOWSKI, C. A. & KULIKOWSKI, C. W. (1971). A comparison of methods for the automated diagnosis of thyroid dysfunction. *Computers and Biomedical Research*, **4**, 374–389.

NORUSIS, M. & JACQUEZ, J. (1975*a*). Symptom nonindependence in mathematical models for diagnosis. *Computers and Biomedical Research*, **8**, 156–172.

NORUSIS, M. & JACQUEZ, J. (1975*b*). Diagnostic models based on attribute clusters—a proposal and comparisons. *Computers and Biomedical Research*, **8**, 173–188.

PAUKER, S., GORRY, G., KASSIRER, J. & SCHWARTZ, W., (1976). Towards the simulation of clinical cognition. *American Journal of Medicine*, **60**, 981–996.

PENG, Y. (1986). A formalization of parsimonious covering and probabilistic reasoning in abductive diagnostic inference. *Ph.D. Thesis*, Computer Science Department, University of Maryland.

POPLE, H., MYERS, J. & MILLER, R. (1975). DIALOG: a model of diagnostic logic for internal medicine. *Proceedings of The Fourth International Joint Conference on Artificial Intelligence*, pp. 848–855. Wm. Kaufmann, Inc.

PRYOR, T., GARDNER, R., CLAYTON, P. & WARNER, H. (1984). The HELP system. In BLUM, B., Ed., *Information Systems for Patient Care*, pp. 109–128. New York: Springer-Verlag.

REGGIA, J. (1978). A production rule system for neurological localization. *Proceedings of Second Annual Symposium Computer Applications in Medical Care*, pp. 254–260. Washington, D.C.: IEEE Press.

REGGIA, J., (1980). A domain-independent system for developing knowledge bases. *Proceedings of the Third Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, University of Victoria, Victoria, B.C. Canada, May 14–16, 1980, pp. 289–295.

REGGIA, J. (1981). Knowledge-based decision support systems: development through KMS. *Ph.D. Thesis*, Tech. Report TR-1121, Computer Science Department, University of Maryland.

REGGIA, J. (1982). Computer-assisted medical decision making. In SCHWARTZ, M., *Applications of Computers in Medicine*, pp. 198–213. Washington, D.C.: IEEE Press.

REGGIA, J. (1985). Abductive inference. In KARNA, K., Ed., *Proceedings of the Expert Systems in Government Symposium*, pp. 484–489. Washington, D.C.: IEEE Press.

REGGIA, J., NAU, D. & WANG, P. (1983). Diagnostic expert systems based on a set covering model. *International Journal of Man-Machine Studies*, **19**, 437–460.

REGGIA, J., NAU, D., WANG, P. & PENG, Y. (1985*b*). A formal model of diagnostic inference. *Information Sciences*. In press.

REGGIA, J. & PERRICONE, B. (1982). *KMS Reference Manual, Technical Report TR*-1136. Computer Science Department, University of Maryland.

REGGIA, J. & PERRICONE, B. (1981). Knowledge-based decision support systems—development through high-level languages. *Proceedings of the 20th Annual Technical Symposium of Washington, D.C.*, pp. 75–82. New York: Association of Computing Machinery.

REGGIA, J. & PERRICONE, B. (1985). Answer justification in medical decision support systems based on Bayesian classification. *Computers in Biology and Medicine*, **15**, 161–167.

REGGIA, J., PERRICONE, B., NAU, D. & PENG, Y. (1985*a*). Answer justification in abductive inference systems. *IEEE Transactions on Biomedical Engineering*, **32**, 263–272.

REGGIA, J., PULA, T., PRICE, T. & PERRICONE, B. (1980). Towards an intelligent textbook of neurology. *Proceedings Fourth Annual Symposium on Computer Applications in Medical Care*, Washington, D.C., November 1980, pp. 190–199. Washington, D.C.: IEEE Press.

REGGIA, J., PULA, T., PRICE, T. & TAYLOR, R. (1981). A computer repository of neurological decision making knowledge. In DUVOISIN, R., Ed. *Transactions of the American Neurological Association*, **106**, 129–132.

REGGIA, J., TABB, R., PRICE, T., BANKO, M. & HEBEL, R. (1984) Computer-aided assessment of transient ischemic attacks. *Archives of Neurology*, **41**, 1248–1254.

REGGIA, J., TUHRIM, S. & PERRICONE, B. (1986). Feasibility of physician-developed computer decision aids. *Medical Decision Making*, **6**. In press.

SHAPIRO, A. (1977). The evaluation of clinical predictions. *New England Journal of Medicine*, **296**, 1509–1514.

SHORTLIFFE, E. (1976). *Computer-Based Medical Consultations: MYCIN*. American Elsevier.

SHUBIN, H. & ULRICH, J. (1982). IDT: an intelligent diagnostic tool. *Proceedings of the National Conference on Artificial Intelligence*, pp. 290–295. AAAI.

SOLOMON, D., KENEVAN, J., EVENS, M., KOSCHMANN, T. & WEIL, M. (1984). An experimental comparison of rule-based and Bayesian approaches to medical diagnosis using MEDAS and EMYCIN. *Proceedings of the Conference on Intelligent Systems and Machines*.

TEACH, R. & SHORTLIFFE, E. (1981). An analysis of physican attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, **14**, 542–558.

TEMPLETON, A., JANSEN, C., LEHR, J. & HUFFT, R. (1967). Solitary pulmonary lesions. *Radiology*, **89**, 605–613.

TVERSKY, A., (1974). Assessing uncertainty. *Journal of the Royal Statistical Society*, **36(B)**, 148–159.

VANMELLE, W., SCOTT, A., BENNETT, J. & PEAIRS, M. (1981). The EMYCIN manual, STAN-CS-81-885. Stanford University: Department of Compuer Science.

WARNER, H., TORONTO, A. & VEASY, L. (1964). Experience with Bayes' Theorem for computer diagnosis of congenital heart disease. *Annals of the New York Academy of Science*, **115**, 558–567.

WATERMAN, D. (1986). *A Guide to Expert Systems*. Reading, MA: Addison-Wesley.

WEISS, S., KULIKOWSKI, C. & SAFIR (1978). Glaucoma consultation by computer. *Computers in Biology and Medicine*, **8**, 25–40.

WINSTON, P. H. & HORN, B. K. P. (1981). *LISP*, pp. 291–301. Reading, MA. Addison-Wesley.

YOSHIMOTO, T., UCHIDA, K., KANEKO, U., KAYAMA, T. & SUZUKI, J. (1979). An analysis of follow-up results of 1000 intracranial saccular aneurysms. *Journal of Neurosurgery*, **50**, 152–157.

ZAGORIA, R., REGGIA, J., PRICE, T. & BANKO, M. (1981). Bayesian classification in medicine—the transferability question. *Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care*, pp. 250–252. IEEE Press.

ZAGORIA, R. & REGGIA, J. (1983). Transferability of medical decision support systems based on Bayesian classification. *Medical Decision Making*, **3**, 501–510.

ZOLTIE, N., HORROCKS, J. & DEDOMBAL, F. (1977). Computer-assisted diagnosis of dyspepsia—report on transferability of a system, with emphasis on early diagnosis of gastric cancer. *Methods of Information in Medicine*, **16**, 89–92.

Appendix

CHEMICAL-SPILL EXPERT SYSTEM SPECIFICATIONS (FICTIONAL PROBLEM)

The AJAX Company maintains a number of manufacturing facilities along Willow Creek. To protect the environment, the company has just set up an automated monitoring station downstream from the manufacturing site to detect chemical spills. It now wants to increase the sophistication of this monitoring station by embedding an expert system in it. Should a chemical spill occur, one or more monitoring instruments at the station will detect an abnormal condition and a "spill alarm" will go off, activating the expert system. The expert system should then analyse the situation, determine what chemical or chemicals have inadvertently entered the creek, decide which manufacturing facilities might be at fault, notify appropriate company personnel, make recommendations concerning emergency clean-up procedures, etc.

When a spill occurs and is detected, the expert system should analyse measurements (water pH, spectrometry, etc.) of the creek water at the monitoring station to decide what chemicals are likely to be contaminating the water. The company's chemical expert has given the following report which will form the basis for the knowledge base.

Expert's summary of relevant knowledge

The type of spill which has occurred depends on monitoring station measurements. The monitoring station is only operational from April to September. One can assume that unless noted otherwise, a particular chemical might contaminate the creek during any of these months with equal likelihood. The other information that is relevant to determining the type of spill is the following data from the monitoring station:

whether the pH of the creek water is acidic, normal or alkaline;
whether the water color is its normal green or brown, or whether it is discolored either red or black;
photometry results, which indicate whether the water has its normal clear appearance or whether it appears oily;
whether or not radioactivity is present in the water;
spectrometry results, which are only capable of determining whether carbon, sulfur, or a metal is present in the water; and
the specific gravity of the water.

The following are possible contaminants of the creek in that they are chemicals used in the manufacturing facilities. Note that it is possible for more than one chemical to contaminate the creek simultaneously. Also, only some of the possible manifestations which may be caused by a chemical may be present when it contaminates the water (e.g. whether the water becomes acidic when an acid is spilled in the creek depends on how much acid is involved).

Sulfuric acid (H_2SO_4), also called "oil or vitriol". This can contaminate the river at any time, but is especially likely in May and June (i.e. months of heavy use in manufacturing). This is a very strong acid, so it can be expected to usually make the water acidic. Spectrometry will always detect sulfur.

Hydrochloric acid (HCl). This is used all the time at the manufacturing facility, but only a little is used in April while large amounts are used in August and September (so spills are more likely during these latter months). It is also a strong acid.

Carbonic acid. This is heavily used in manufacturing during April to June, occasionally in July, and never during other months. It is a relatively weak acid. Spectrometry may detect carbon when this substance is a contaminant.

Benzene. This gives water an oily appearance that may be detected by photometry. Spectrometry may detect carbon.

Petroleum. This is used constantly, but most heavily in the months of July, August and September. It may turn the water black and give it an oily appearance and may decrease the specific gravity of the water. Spectrometry usually detects carbon.

Benzenesulfonic acid. This is a weak acid that may give the water an oily appearance. Spectrometry may detect carbon and/or sulfur. It decreases the specific gravity of the water.

Thioacetamide. This may occasionally turn the water red. Spectrometry may detect carbon and/or sulfur.

*Chromogen R*23 (red dye number three). This may make the water alkaline, and usually colors it red. It may decrease the specific gravity of the water. Spectrometry may detect carbon.

Hydroxyaluminum. This may make the water red or alkaline. Spectrometry may detect metal, and the specific gravity may be increased.

Sulfur isotope. This may cause radioactivity to be detected and sulfur to be found by spectrometry.

Carbon isotope. This is only used during July and August at the manufacturing facilities. It may cause radioactivity to be detected and carbon to be found by spectrometry.

Cesium. This is used heavily in manufacturing during May, rarely during April and June, and never during any other months. It may make the water sufficiently radioactive that it can be detected. The specific gravity may be increased, and spectrometry may detect metal.

Rubidium. This is used in all months, but most heavily in April. It is a very weak base, but if massive quantities were spilled it might make the water alkaline. Spectrometry may detect metal, radioactivity may be present, and the specific gravity of the water might be increased.

Radium. This is used in all months, but most heavily in August. It is extremely radioactive, so when it is a contaminant, radioactivity will always be detected. Spectrometry may detect metal, and the specific gravity of the water may be increased.