A PATTERN RECOGNITION APPROACH

TO COMPUTER-AIDED MEDICAL DIAGNOSIS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAII IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN ELECTRICAL ENGINEERING

SEPTEMBER 1970

By

Casimir Alexander Kulikowski

Thesis Committee:

Michael S. Watanabe, Chairman
Norman Abramson
Richard H. Jones
Bharat Kinariwala
W. Wesley Peterson

ACKNOWLEDGEMENT

I wish to thank Dr. R. A. Nordyke of the Straub Medical Research Institute for providing the data and encouraging me in this project. Long discussions with him have been invaluable in contributing to my understanding of some of the problems involved in differential diagnosis.

ABSTRACT

This dissertation describes a pattern recognition model which has been successfully used to simulate a doctor's diagnostic process. A computer program implementing this model can be a valuable aid to the specialist, freeing him from routine screening procedures and making his skills more readily available to those patients who need them.

The process of diagnostic inference is formulated as a pattern recognition task for which each disease category is represented by a characteristic pattern of symptoms and other patient variables. The method of class featuring information compression that was used assumes these characteristic patterns to form a subspace of the variable space. The subspaces are defined in terms of the components of an optimal expansion for the data vectors of a class. The use of this optimal expansion guarantees that the representative samples from which it is calculated will lie closer, on the average, to a subspace spanned by its own principal components than to any other subspace of the same dimension.

The principal results of this dissertation in the pattern recognition field are procedures for selecting the dimensionality of the class subspaces to obtain good discrimination. Two quantities were found to be good predictors of recognition performance. One is a ratio of the inclusions of the paradigms of two classes within the subspace of one of them; the other is the average margin of correct classification for the paradigms of a class. Both, under certain conditions, are highly correlated with recognition performance. Therefore, a procedure for subspace selection is the maximization of the average margin of correct classification for one class subject to a constraint

on the margins for all other classes. A similar procedure can be derived with the ratios. The average inclusions of paradigms within a subspace can be calculated from the autocorrelation and projection matrices of the classes. Thus, the ratios and margins are found and performance predicted without the need of actually performing any classifications.

The model was tested with data obtained for 3291 patients who were examined at the Straub Clinic in Honolulu between 1963 and 1969 for the possibility of thyroid dysfunction. Data from 1963 to 1968 were used as paradigms and 1969 data as a test sample. In the diagnosis of hyperthyroidism the pattern recognition program performed consistently better than a linear discriminant method. In the diagnosis of hypothyroidism, however, the original class featuring information compression program did not perform as well as the other methods with which it was compared. Subspace selection by the method of constrained margin maximization improved the performance of the pattern recognition program considerably. The results indicate that this method can serve as a good representation of a realistic diagnostic situation.

In order that the method be useful clinically a sequential version of the program was developed giving the classification of a patient at every stage of diagnosis. A category of deferred judgment was included in order that more data could be gathered when a diagnosis was uncertain. The sequential program satisfied clinical tolerances of accuracy as determined by a specialist. The on-line performance of this program has been simulated successfully and is to be implemented clinically.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

THE DIAGNOSTIC PROCESS

The Logical Basis of Diagnostic Inference

What logical procedure does a doctor follow in making a diagnosis? The doctor knows the particular set of signs, symptoms, and test results (hereafter referred to as variables) exhibited by the patient. He also knows the corresponding typical sets of variables characterizing certain diseases. If a unique pattern of variables characterized each disease and if there were no overlaps between the patterns of different diseases the problem of diagnostic inference would be a simple one, solved by matching the patient's observed variables to a disease. The logic of diagnosis could then be deductive. Unfortunately, such a simple state of affairs rarely occurs in medicine. The first problem is that the set of diseases from which the doctor has to make a choice is essentially infinite. Secondly, many of these diseases are not yet well-defined. Reflecting this, the pattern of variables that is typical for a disease does not have to be present in an affected patient. There is usually quite a range of quantitative values (or qualitative alternatives) that a variable can take in patients having a certain disease. Thus, a particular distribution of values for each variable, and joint distributions for patterns of variables, would characterize a disease. To complicate matters, the distributions for various diseases usually overlap. Therefore, the physician's reasoning becomes inductive: from the observed pattern of variables he must infer which disease is most likely. The doctor is making a decision in the face

of uncertainty, incurring a certain risk if his decision proves to be incorrect.

The doctor's decision is not usually reached in a single step. It is a process of narrowing down many possibilities to the most likely alternatives. The initial step of going from a practically infinite number of possibilities to a finite set of disease categories is a difficult problem in "pattern recognition" for which adequate models and solutions are lacking. This dissertation will be concerned with a more limited problem of diagnosis--that which occurs when, because of prior knowledge or obvious indications, the doctor initially considers only a finite set of alternatives. The problem of diagnosis and the allied choice of treatment can then be posed in the context of pattern recognition or decision theory.

To be able to apply decision theory certain probabilistic information must be available relating the occurrence of sets of variables to all possible diseases. In what follows it will be assumed that a set of good variables has been selected to characterize a patient. It has been correctly noted elsewhere[1] that the diagnostic process involves the selection of variables (questions and tests) as well as the inference made on their basis. The selection problem will be considered separately after the framework of diagnostic inference has been examined.

Ideally, the doctor should know the conditional probabilities of each possible set of symptoms for every population of diseased (or healthy) patients. The diagnostic problem is then solvable by the calculation of the inverse conditional probabilities of the patient's

falling into one of the disease categories, given a particular set of observed variables. This can be done using Bayes' theorem if the prior probability of each disease is known. Alternatively, the diagnostic problem with full statistical information can be solved by direct inference from the disease likelihoods (the conditional probabilities of the variables). To do this the ratios of the disease likelihoods are computed and compared to thresholds which are functions of the prior probabilities and costs of misclassification. If prior probabilities are not admitted diagnosis can be viewed as the testing of the hypothesis that a patient has a certain disease versus the alternative that he does not. Such a test can be designed to minimize the number of false positives at a fixed tolerance level for false negatives. This test would have to be repeated for every disease under consideration.

The practical problem arising from the application of any of the above theories is that the availability of full probabilistic information can very rarely be assumed. Thus, estimates have to be used based on finite samples of patients with confirmed diagnoses; the recognition schemes incorporating such estimates result in lower percentages of correct diagnoses than are theoretically attainable. The above methods are still reasonable if good estimates are available. However, there are drawbacks to obtaining such estimates. If there were a large enough sample from the population of patients with a given disease D_k it would be possible to estimate the conditional probabilities of the various patterns of variables (represented by the vector $\underline{x} = x_1, x_2, \ldots, x_m$) from their frequencies in this population. Usually this is not at

all practical since for N discrete variables, which can assume M possible values each, there will be M^N possible patterns. Except for only the simplest (binary) variables this approach would require impossibly large samples in order to densely cover the variable space in a way that would ensure good estimates. If all the variables are probabilistically independent, however, any required probability of a pattern is the product of the conditional probabilities of the component variables. These can be estimated by their frequencies in smaller samples because only the univariate spaces must now be covered densely. However, not only is the assumption of independence most often untrue, but, even if it holds, the required sample sizes might still be too large to be obtained practically.

If it is not possible to assume independence, and in cases where estimation of the conditional probabilities is impractical, a different approach to diagnostic inference is needed. Methods which have been used in pattern recognition and discriminatory analysis are well-suited to the purpose. They incorporate a learning procedure by which rules of classification are derived from data on samples of patients with confirmed diagnoses without the explicit use of a probability assignment. The two main choices to be made in any of these methods are how the diseases are to be represented in terms of the variables, and which decision rule will be used for classification. Classical discriminatory analysis has concerned itself mainly with finding those linear combinations of variables that discriminate best between a certain group of classes, whereas pattern recognition methods encompass a wide variety of approaches. The term "feature extraction" denotes

any pattern recognition method that represents classes by "features" (general mathematical combinations of the variables) extracted as being most representative of a class and serving as the basis for classification.

One pattern recognition method, proved effective in both character[2] and speech recognition[3] and being applied now to medical diagnosis,[4] is based on a representation of classes in terms of their optimal coordinate systems (Karhunen-Loève expansion). In this model features are extracted from the variable space to form distinct class subspaces. This dissertation shows that such a model can serve as a useful representation of the inferential process in diagnosis.

How do the above diagnostic approaches (motivated by statistical or pattern recognition methods) compare with the logical process followed by a physician? For most doctors diagnostic reasoning is not easily formulated in a set of rules, whether deterministic or probabilistic. Rules about certain facets of disease may be followed but, in general, diagnosis is an intuitive process. Some of the possible components of such a process, their interrelationships and limitations, will be examined in later chapters to gain insight into the underlying modes of a doctor's reasoning. In this manner they can be accounted for in attempts to automate diagnosis.

The first defining constraint under which a physician operates is the limit of his knowledge: he will not be able to diagnose a disease that he does not know about. A good doctor, however, will form a rather complete picture of what a healthy patient should look like, and will compensate for his ignorance of unknown diseases by a good

ability to detect them as departures from the normal.

Part of the doctor's knowledge might be translatable into probability statements expressing the value of certain variables as indicators of possible disease. It is most likely, however, that such probabilistic information does not contain everything that is useful or crucial in making a diagnosis. For example, the doctor will tend to be alert for specific combinations of symptoms that are only occasionally dangerous, though in an overwhelming number of cases they are not. In a simple probabilistic model the dangerous cases would not stand out because of the averaging effect of statistics. A consideration of the losses incurred by misclassifying such dangerous cases would, then, be most desirable. Decision theory considers precisely such information.

Another problem that the doctor faces is the occurrence of unusual data. If a patient's variables are such that only a few or no other patients have exhibited the same or similar values, inference on the basis of the unusual data is impossible. More data is needed. Therefore, diagnostic inference must always allow for a suspended-judgment class of decisions from which the doctor can proceed to consider additional information.

A physician's reasoning can be represented by models of inference which bypass the underlying probability structure. For example, a patient's variables might be scored in such a way that different diseases are characterized by very different values of the total score. This approach has much in common with discriminant methods which incorporate linear scoring of the variables. Alternatively, a doctor

having difficulty in giving probabilities for suspected diseases might, nevertheless, be able to rank the diseases in order of increasing correlation with the observed variables of a given patient. Pattern recognition methods which employ correlations might thus prove good simulators of the doctor's process.

The Selection of Diagnostic Variables

In this discussion diagnostic inference has been considered as based on a number of patient variables--signs, symptoms, and tests--that the doctor believes to be good and relevant to diagnosis. The process by which these variables are selected will, therefore, strongly determine the outcome of the inference.

Diagnosis can be seen as a process in which the doctor at every step selects a variable of the patient to examine, updates his inference on this basis, and uses the new overall information to guide him in the selection of further variables. This selection is, optimally, such that it leads to a correct diagnosis most quickly, constrained only by the cost and risk involved in acquiring the information. Thus, the problem lies in determining which sequence of the yet unchosen variables carries the most diagnostic information, given the known data of the patient. Unfortunately, the number of possible sequences leading to a correct diagnosis is very large. Though conceptually it is simple to think of searching through them for the shortest sequence, this is, in practice, impossible, except for very small groups of variables. The doctor relies on his intuition to make a rapid choice of the next variable at every step.

The selection of each variable (highly contingent on previous

information) is characteristic of most initial diagnoses, but need not be so important in specialty diagnosis. The patients a specialist sees are already suspected of having one of a certain group of diseases. His responsibility is primarily to diagnose which, if any, of the diseases the patient has. In this context much can be done to make the examination and testing of patients uniform. The problem of selecting good variables is a much simpler one because now they are a finite set, preselected over the years by the experience of many specialists. In a given situation, a good subset of the variables must be chosen to lead rapidly to a correct diagnosis. As a matter of effectiveness, simplicity, and cost the order in which the patient's variables are acquired can be divided into three subgroups that are comparatively homogeneous within themselves: 1) the doctor questions the patient; 2) he performs a physical examination; and 3) if needed, laboratory tests are taken. Within the first subgroup all questions are equally easy to ask and not costly. In many cases items included in a physical examination are all about comparable in patient discomfort and cost, and in time used. The same cannot be said of the third subgroup of variables. Laboratory tests vary in both cost and risk and each must be considered separately.

Within each of the two first subgroups of variables the optimization of diagnosis consists of the selection of those sets of variables which are best indeces of disease, regardless of the order in which they are chosen. The inference based on these variables can be made after the entire subgroup is completed. In addition to simplicity an advantage of this approach is that consistent data will

be obtained from patient to patient, yielding statistical information that can be used as a basis for automated diagnosis. This dissertation will be concerned with the modelling of a specialty diagnostic situation by the above approach.

The Computer as an Aid to Diagnosis

The decision-making capabilities of a computer are very different from those of the clinical diagnostician. The computer is both more resourceful than the doctor and more limited. Its principal advantages are a much larger accessible memory and the availability of many explicit, rapidly executable decision schemes. Its disadvantages are a lack of the unique experience of the human diagnostician and an inability to stray from prescribed decision rules, even when learning functions are incorporated.

The fundamental difference between a doctor and the computer lies in their variable selection and decision-making processes. The doctor's process is relatively unstructured. His final diagnosis is the result of an induction; the value he attaches to specific information varies from patient to patient, based on a particular pattern and correlation between symptoms. The doctor also takes into account more subtle points, such as an overall evaluation of a patient and the observation of slight but unique characteristics. On the other hand, the computer's decisions are completely structured: a unique diagnosis corresponds to a unique set of symptoms. The set of variables from which the computer can choose is fixed and finite, though it can be large. It is crucial, then, that both variable selection and decision rules be made optimal in some manner. It would be reasonable if such

rules gave the best diagnosis for a predetermined acceptable risk and cost. Various optimal decision rules have been mentioned in the first section on diagnostic inference. There is, in addition, another optimality to consider: the most efficient utilization of the machine. A good strategy for diagnosis must also be reasonable in its requirements for computer time and memory. Such factors are properly part of the cost of diagnosis, but for the purposes of this discussion will be treated separately.

Given the differences between man and machine it would be a difficult task, as well as a poor use of resources, to try to directly simulate the doctor's diagnostic process on the computer. The computer has a better role as an efficient complement to the physician. In its large accessible memory can be stored data about many diseases, and programs can be designed to aid the doctor in diagnosis within specific disease groupings. The initial narrowing down of alternatives will still depend very much on the physician's intuition, and the final decision of diagnosis and allied treatment has to be his responsibility entirely. However, between the initial intuition and the final treatment the computer can serve as an effective tool to test out various hypotheses that a doctor might consider for a given patient. Thus, the development of diagnostic computer programs for disease specialities has a high priority for practical clinical usage. This dissertation will be concerned with the development of pattern recognition methods for variable selection and disease discrimination applicable in such programs.

CHAPTER II

MATHEMATICAL MODELS FOR DIAGNOSIS

Literature Review

Attempts to develop automatic aids for differential diagnosis began in the 1950's with various schemes for matching symptom groups to diseases. They were implemented by card sorting systems and other mechanical means.[5,6,7] The introduction of probabilistic models for diagnostic inference coincided with the availability of computers that could process long symptom lists at high speeds. Ledley and Lusted[8,9] began investigations in this field, suggesting a Bayesian inference model with frequency-count estimates for the symptom conditional probabilities. Probabilities of symptom patterns were estimated by the product of the marginal distributions of the constituent symptoms. Because of its simplicity this diagnostic model has become one of the most popular. It was soon applied to the diagnosis of heart disease by Warner and his colleagues[10,11,12] and by other investigators.[13,14] The same model was used for diagnosis of thyroid dysfunction by Overall and Williams[15] who, with Fitzgerald, developed a computer program that has been made available for general use.[16] Overall and Williams compared their method to discriminant function and factor analysis models in the diagnosis of thyrotoxicosis.[17,18,19] The discriminant analysis method had been used earlier in this context by Crooks, Murray, and Wayne.[20] A hypothesis testing approach was taken by Collen, Rubin, et al.[21,22] for diagnosis in a multiphasic screening situation. The first fully sequential diagnostic system (using the

Bayesian model) was developed by Gorry[1] and implemented at the Massachusetts General Hospital.[23,24] A decision theoretic approach to the overall problem of diagnosis and treatment has been taken by Ginsburg and Offensend for cases where the selection of variables (laboratory tests) and treatment involves substantial risk.[25] Pattern recognition methods have been used for diagnosis only in the special-purpose recognition of forms, such as chromosome analysis.[26] The only general pattern recognition method that has been used to model the diagnostic process as such is the perceptron model.[27]

Despite ten years of active research in computer-aided diagnosis there appears to be a lack of systems that have actually been used in clinical situations. Most studies have been retrospective and hospital-based. Some of the characteristics and results of the studies will be examined to gain insight into the reasons for this.

The simplified Bayes' model is most widely used in automated diagnosis. If $\underline{x}$ is the vector of the patient's variables, composed of elements $x_j (j = 1,2,\ldots,m)$, and D_k is any disease category, then the posterior probability that a patient with variables $\underline{x}$ will exhibit disease D_k is given by Bayes' formula:

$$p(D_k/\underline{x}) = \frac{p(\underline{x}/D_k)p(D_k)}{\sum_j p(\underline{x}/D_j)p(D_j)} \qquad (2.1)$$

where $p(\underline{x}/D_k)$ is the conditional probability of observing the vector of variables $\underline{x}$ in a patient with disease D_k (often called the likelihood of D_k), and $p(D_k)$ is the prior probability of D_k. Because estimation of every possible $p(\underline{x}/D_k)$ would require excessively large

samples a zero-order approximation is used:

$$p(\underline{x}/D_k) = \prod_{i=1}^{m} p(x_i/D_k) \qquad (2.2)$$

This estimate is, in general, biased. It will be exact only if all the x_i are independent within the population D_k. The estimate allows great computational simplification, requiring storage of only the marginal distributions $p(x_i/D_k)$. Despite its low order, this estimate has proved surprisingly effective in the diagnostic model, comparing well with the performance of specialist physicians. In reporting the first application of this method to diagnosis (in the area of congenital heart disease) Warner and his colleagues state:

> "...it is apparent from our experience to date with 36 cases that the most probable diagnosis estimated with Equation 10 (Bayes' formula) agrees with the actual diagnosis made by physiologic studies and observations at surgery at least as often as does the most probable diagnosis estimated by 3 experienced cardiologists from the same clinical information."[10]

Three years later, with 83 patients tested by the program, the authors arrived at essentially the same conclusions.[11] In addition to simplicity the zero-order estimation using Bayes' formula has the advantage of being able to represent qualitative data, so common in diagnosis. Care must be taken, however, not to include mutually exclusive or highly dependent variables. Yet another advantage is that patients with incomplete data are easily included in a study by exclusion of the missing variables from the computation of the $p(\underline{x}/D_k)$.

The same Bayes' model, applied to the diagnosis of thyroid dysfunction, gave satisfactory results in the classification of 286

patients with complete data in another study by Overall and Williams;[15] in this study laboratory tests were included, however, and they are themselves the best indicators of disease. The percentage of correct diagnoses was 96.3%, but when signs and symptoms alone were used this became 88%. Applied to an independent sample in Germany by Reichertz,[28] the program diagnosed 93 out of 100 hyperthyroid patients correctly, and 96 out of 105 patients with mixed diseases. This performance was obtained despite differences between the statistics of the new sample and the paradigms, proving the insensitivity of the model to variations of culture and geography. The author noted that

> "...the Bayes theorem led to a high degree of correctness in classification even when the mutual exclusivity and possible independence of some symptoms was not taken into consideration."[28]

The maximum posterior probability rule used in the above research corresponds to a minimization of the average probability of misclassification--if the exact distributions and not estimates are used. A more general approach, motivated by decision theory, minimizes the risk (or cost) involved in making a decision. This was considered by Gorry[1] who developed the first fully sequential diagnostic program. His model can be described in three parts: 1) an inference function based on the simple zero-order Bayes' model, but with facilities for handling prescribed dependent variables; 2) a pattern selection function that allows the program to discard variables irrelevant for diagnosis; and 3) a test selection function employing heuristic methods to select good variables. The depth of search for evaluation of such goodness was restricted to one stage. Though minimization of average

risk proved too complicated to implement, a simpler minimum loss criterion was used in the strategy for variable selection. The program was tested with Warner's data on congenital heart disease reproducing the final diagnoses of the single stage Bayes' programs, but using a significantly smaller number of variables. An interesting diagnostic index was used: it was the product of the average probability of the correct diagnoses and the frequency of cases in which correct diagnosis was given a probability greater than 1%. This quantity reflected both the efficacy of correct classification and the proportion of complete failures. It is questionable whether this index is useful clinically, since a doctor is interested in all errors of diagnosis (regardless of the academic definition of "gross" or "slight" error). If the severity of error is to be considered a criterion of program performance it should not be established absolutely, but, rather, as a function of all the posterior probabilities of classification.

Discriminant methods were first applied to the diagnosis of hyperthyroidism without the use of computers.[20] It was found that a scoring system for the symptoms yielded a diagnostic index that resulted in about 85% correct classification for 171 cases of an independent sample. In this model a measure of relative importance (score) is attached to each variable, reflecting its contribution to the linear discriminant function (LDF) of the class. This function has the general form:

$$Y_k = \sum_{i=1}^{m} c_i^{(k)} x_i + d^{(k)} \qquad (2.3)$$

The best LDF in the sense of maximizing the distance between the

average vectors of two classes for a constant intraclass dispersion is the one such that

$$c_i^{(k)} = \sum_{j=1}^{m} g_{ij} m_j^{(k)} \qquad \text{and} \qquad d^{(k)} = -1/2 \sum_{j=1}^{m} [m_j^{(k)}]^2$$

where g_{ij} are the elements of the inverse of the pooled covariance matrix of the classes and m_j is the j-th component of the average vector of class j. This Y_k, then, differs on the average more from one disease group to another than any other linear combination of variables. A vector $\underline{x}$ is classified into the class for which the corresponding LDF is greatest. From the point of view of decision theory the LDF is optimal (minimizes the probability of misclassification) only if the populations from which the LDF's are calculated are distributed normally in the $\underline{x}$'s with equal covariances for all classes. This method has also been applied to thyroid diagnosis by Overall and Williams.[29] They reached a correct diagnostic rate of 96% for 225 patients by including laboratory tests in the data.

A review of the various methods used to date with the diseases diagnosed, the number of variables considered, and the results of classification are given in Table 1.

Most of the studies carried out have been restricted to a small or medium number of diseases within a specialty, indicating that it is within this field that feasible solutions to the diagnostic problem can be obtained. The exceptions are studies aimed primarily at screening, where large numbers of common diseases have been treated.[30]

The results of most automated diagnostic studies have been good to excellent, but often the performance was measured on the same set

TABLE 1

SOME STUDIES IN AUTOMATED DIAGNOSIS SUMMARIZED

Method	Disease	Reference Number	Number of Variables	Number of Patients	Correct Classification (%)	
					Paradigms	Test Sample
Bayes' theorem with zero-order probability estimates	Heart disease	10	53	36	--[a]	--
		11	53	83	--	--
		14	20	231	78.0	**[b]
	Hematology	12	12	103	**	95.0
	Thyroid function	15	21	450	93.3	**
		28	--	205	**	92.3
Linear discriminant	Thyroid function	29	27	225	96.0	**
	Pulmonary stenosis & patent ductus	27	7	286	65.5	**
Pattern recognition	Thyroid function	4	28	2208	94.3	89.0
	Pulmonary stenosis & patent ductus	27	7	75	77.3	**

a -- indicates not reported
b ** indicates not performed

of data that was used for modelling, thus not giving a fair assessment of future performance in other samples. Good results could also be expected in those studies where sophisticated laboratory tests were included in the data, since such tests in themselves are often the best indicators of a disease. Probably the most limiting factors of such studies are that they have been retrodictive and tested on small numbers of patients under relatively nonhomogeneous circumstances. In many cases, studies consisted of an analysis of records coming from a number of doctors where observer variation could be expected. The principal reason why there has not been clinical application of these methods is probably the lack of an adequate and trusted data base in each of the specialties.

With the single exception of the MIT-Massachusetts General Hospital effort[1,23] all research to date has concentrated on the classification problem, with little attention being paid to the development of formal variable selection procedures. This probably reflects an initial choice (preselection) of good variables by the various specialty physicians. Usually, however, the number of symptoms, signs, and laboratory tests that are in common use within a specialty is quite large. Each doctor will prefer a subset of these and depend on them for his diagnoses. It would be a worthwhile task to seek out from among the large set of generally accepted variables those which prove the best discriminators without going to the more complicated, fully sequential selection procedures. This is an integral part of designing a program adequate for clinical use.

Because this dissertation develops a pattern recognition method for the diagnostic process, the characteristics of some of these methods and their possible advantages will now be considered.

Pattern Recognition Methods in Diagnosis

The Bayesian inference and discriminant models used to date in diagnostic programs are optimal in application only if the data satisfies the assumptions of optimality. This requires conditional independence of the variables in the zero-order Bayesian case, and normal distributions for the data of each class, the data differing only in their mean values for the discriminant case. That these models perform as well as they do is a strong indication of their tolerance to violations of the underlying assumptions. In specific disease areas, however, these diagnostic models may not reach the level of accuracy required for clinical implementation. There is a need to search for other methods that can aid the clinician in diagnosis.

Pattern recognition methods, developed mainly for the recognition of visual images, have proved equally useful in other classification tasks. The basic approach of pattern recognition is the formation of a system that will learn some rule of classification on the basis of samples of objects of known classification. A certain structure for the recognition system is specified and during a "training" period the parameters of the system are adjusted according to some criterion so that a large proportion of future objects of unknown classification can then be correctly recognized. The structure of most pattern recognition systems is chosen heuristically to solve particular problems where statistical classification methods are not applicable.

There are, however, some mathematical models generally applicable to problems of classification.

In medical diagnosis a perceptron model was used by Abraham and Caceres[27] for the diagnosis of pulmonary stenosis and patent ductus with better results (77.3% correct diagnoses) than a linear discriminant model (65.5%) or a probability density estimation procedure (69.6%). Unfortunately, the perceptron analysis was carried out on a subsample of 75 patients out of 286 patients that were diagnosed by the two other methods making comparisons difficult. The first application of a pattern recognition method to the diagnosis of a large number of patients was the work preliminary to this dissertation.[4] It applied the method of class featuring information compression (CLAFIC)[2] to the diagnosis of hyperthyroidism in a sample of over 2000 patients. The main purpose of the research was to find a method that would give the best diagnosis on the basis of symptoms and physical findings without the use of laboratory tests. The CLAFIC method performed consistently better than a linear discriminant method (96% vs. 91.4%) when they were compared in a subsample of 271 cases. The performance of the pattern recognition program was tested on an independent sample of data as well as checked on the set of paradigms. In both sets the correct classification of hyperthyroid patients was about 95% while the rate of false negatives was kept between 10% and 15%. These rates were considered by the clinician as satisfactory in his specialty. A mathematical description of this method will be presented next, serving as the basis for more general pattern recognition diagnostic procedures.

Subspace Models for Class Representation

There are three principal tasks in any classification problem: 1) the weighting and selection of variables; 2) the specification of a measure of nearness between objects; 3) the choice of a decision-making rule.[31] These tasks are interrelated in a manner that depends on the approach taken toward classification.

In the absence of sufficient probabilistic information classificatory procedures can attempt to directly extract from the samples of paradigms those discriminatory features that best distinguish the objects of one class from another. A feature is any combination or function of the observed variables. In practice, the search for such discriminatory features must be restricted to certain categories, such as linear or quadratic combinations of variables, polynomials, etc. It is simplest and most usual to seek best linear discriminants for the classes, but these may be inadequate if most of the distinguishing information lies in quadratic or other functions of the variables.

Another approach to classification is to first find some good representation of a class in the variable space. Then a recognition rule is specified by choosing a criterion of similarity or closeness of an object to the class representation. This approach is fruitful if the class representation is good in the sense of containing a broader class of discriminating features than might be contained in the natural categories of discriminators mentioned above. On the other hand, if the class representation contained only a subset of the features within a class of discriminators it could not be expected to perform any better than these discriminators. In general, however,

pattern recognition methods can be seen as attempts to broaden the set of features from which best discriminators can be extracted. Their success depends on taking advantage of properties of the data that can aid in classification. For example, if there is reason to believe that the objects of different classes have unimodal distributions over different regions of the variable space, a useful class representation might be obtained by using those features that are maximally clustered about each mean; or, equivalently, if mean square distance is taken as a criterion of clustering, a representation can be obtained by finding the Gaussian distributions that best fit the data.[32]

If there is reason to suspect that the data of each class is mostly confined to a subspace of the variable space--as might be suggested by strong correlations between some of the variables--it would be reasonable to characterize each class by its subspace of best fit. The subspace for each class would be unique and contain most of the information of the class. If there were not much overlap between the subspaces of different classes, this representation would contain effectively the discriminating information about the classes. This is the approach taken by the method of class featuring information compression (CLAFIC).

Consider a set of data vectors $\underline{x}^{a^T} = (x_1^a, x_2^a, \ldots, x_m^a)^T$ of m variables that represent the objects of known classification. Let there be n_j objects for class C_j. Then the autocorrelation matrix for the sample of class C_j will be

$$G_j = \frac{1}{n_j} \sum_{a=1}^{n_j} \underline{x}^a \underline{x}^{a^T} \tag{2.4}$$

A reasonable choice for the best-fitting subspace is the one in which the data vectors are reproduced with minimum error. If mean square error is taken as a criterion, the best subspace is the one that is spanned by the principal components of the autocorrelation matrix G_j.[3,31] Thus, vectors of class C_j will be represented on the average with high accuracy (depending on the number of principal components taken) in the subspace of C_j. Vectors of any other class C_k will be represented with much greater average error in the subspace of C_j. A classification rule that proves effective on the average should be the assignment of vectors to the class for which least error of representation is obtained.

An alternative derivation of this subspace model is obtained if one considers all vectors initially normalized, such that $\underline{x}^T\underline{x} = 1$, disregarding vector lengths but preserving the angles between vectors. A probability-like measure of the importance of a coordinate in characterizing a class is given by the mean square value of the corresponding components of the paradigms. For variable x_k it is possible to define the measure ρ_k:

$$\rho_k = \frac{1}{n_j} \sum_{a=1}^{n_j} [x_k^a]^2 \tag{2.5}$$

The degree of uneveness with which these weightings are distributed along the different coordinate axes is given by the entropy function:

$$S(\{\rho_k\}) = -\sum_{k=1}^{m} \rho_k \log \rho_k \tag{2.6}$$

A best subset for minimum error estimation of the paradigms can now be defined by choosing an orthogonal basis spanning the original space but such that the distribution of weights ρ_k is most uneven. Truncation can be used to specify the subspace in which maximum weight is concentrated. The orthogonal basis for which $S(\{\rho_k\})$ is minimized is the set of eigenvectors $\{\underline{u}_k\}$ of the autocorrelation matrix G_j. If the eigenvalues are arranged in descending order of magnitude so that

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$$

the subspace that is spanned by the first m^* eigenvectors is chosen, such that

$$\sum_{j=1}^{m^*} \lambda_j \leq \sigma < \sum_{j=1}^{m^*+1} \lambda_j$$

The quantity σ is the fidelity with which the vectors of the class are represented in the subspace of dimension m^*.

As a criterion of classification it is now possible to use the angular distance of a vector to a subspace--or, conversely, the closeness of a vector to a subspace in terms of the cosine squared of the angle θ with the subspace:

$$\cos^2\theta = \sum_{k=1}^{m^*} (\underline{x}^T \underline{u}_k)^2 \tag{2.7}$$

A projection operator can be introduced to characterize a subspace:

$$P = \sum_{k=1}^{m^*} \underline{u}_k \underline{u}_k^T \tag{2.8}$$

Then the cosine squared can be written in terms of it:

$$\cos^2\theta = \underline{x}^T P \underline{x} \tag{2.9}$$

This is also computationally more convenient.

Thus, a vector $\underline{x}$ is classified by the rule

$$\underline{x} \in C_i \quad \text{if } \underline{x}^T P_i \underline{x} = \max_j \{\underline{x}^T P_j \underline{x}\} \tag{2.10}$$

where P_i is the projection operator of class C_i. If a decision to withhold judgment is desired for application in sequential pattern recognition procedures, the rule can be modified so that

$$\underline{x} \in C_i \quad \text{if} \quad \underline{x}^T P_i \underline{x} = \max_j \{\underline{x}^T P_j \underline{x}\}$$

and

$$\underline{x}^T P_i \underline{x} > f(\{\underline{x}^T P_j \underline{x}\}) \; ; \; j \neq i$$

where $f(\cdot)$ is some function of the closeness of the $\underline{x}$ to the subspace of all classes except the closest one (C_i). This requires that the maximum closeness be high in relation to the closeness of the vector to other subspaces. It would be simplest to choose $f(\cdot)$ as a constant that could be empirically adjusted by performance on a test set. If no C_i satisfies the above conditions $\underline{x}$ is classified into C_{n+1}, a class of deferred judgment.

The CLAFIC method will be effective if 1) the data is mostly concentrated (in the mean square sense) in a subspace of the variable space; 2) the subspaces of different classes do not overlap significantly. If the subspaces were orthogonal, they would indeed be good discriminators because this would imply that the most heavily weighted

features of each class had the smallest weight in all other classes. However, it is more usual that there be some overlap and that the subspaces be at different angles to one another. The worst condition is that in which the only significant discriminating information is contained in the subspaces complementary to the class subspaces defined by the fidelity level σ. This implies that clustering (in the subspaces of smallest weighting) is more important than weighting or dispersion and that the CLAFIC model is not adequate for discrimination. That this method has succeeded in such diverse fields as character and speech recognition and medical diagnosis indicates that it can be a good model of realistic data. However, this is never known until it is tested in specific instances. Thus, it becomes important to ask whether it is possible to find a subspace representation guaranteeing that the classes will contain good discriminating features. One answer to this question has been given for the two-class recognition problem. Fukunaga and Koontz[33] have suggested a distortion of the variable space such that in the new space the principal components of one class are the smallest of the other and vice versa. A transformation T of the variable space is chosen, such that it diagonalizes and scales the sum of the autocorrelation matrix of the two classes into the unit matrix

$$TG_s T^T = I$$

where

$$G_s = G_1 + G_2$$

is the joint autocorrelation matrix. If the transformed autocorre-

lation matrices are denoted by $R_1 = TG_1T^T$ and $R_2 = TG_2T^T$ they satisfy $R_1 + R_2 = I$. If $\{\lambda_i^{(k)}\}$ and $\{\underline{u}_i^{(k)}\}$ are the sets of eigenvalues and corresponding eigenvectors of matrix R_k it is easy to see that

$$R_2\underline{u}_i^{(2)} = (I - R_1)\,\underline{u}_i^{(2)} = \lambda_i^{(2)}\underline{u}_i^{(2)}$$

or

$$R_1\underline{u}_i^{(2)} = (1 - \lambda_i^{(2)})\,\underline{u}_i^{(2)}$$

but since

$$R_1\underline{u}_i^{(1)} = \lambda_i^{(1)}\underline{u}_i^{(1)}$$

it can be seen that

$$\underline{u}_i^{(2)} = \underline{u}_i^{(1)}$$

and that

$$\lambda_i^{(1)} = 1 - \lambda_i^{(2)}$$

Though this method gives a set of discriminating features for distinguishing between two classes, its direct extension to multiclass recognition is difficult because a different distortion for each pairwise comparison would make multiple comparisons inconsistent and meaningless. Thus, one could seek a single distortion of the space so that the subspaces of the different classes in the new space would become orthogonal. Some approaches to this problem of multiclass discrimination will be discussed in the next chapter.

CHAPTER III

DISCRIMINATORY SUBSPACE MODELS FOR MULTICLASS RECOGNITION

Discriminatory Subspaces

The class subspaces of the CLAFIC method were chosen because of their optimality in representing the paradigm vectors with minimum mean square error. Thus, for a given reduction of dimensionality, the best estimates of the paradigms lay in the class subspace. When this method was used for pattern recognition rather than for estimation the first step was to implicitly specify a criterion of closeness between the objects (the angular distance between their vectors). A weighting of variables was carried out next (calculation of the optimal coordinate system), followed by the selection of variables with highest weight defining the class subspaces. The classification rule (maximum inclusion within the subspace) was a natural result of the choice of closeness measure and class definition.

Such an approach differs from the more usual probabilistic and discriminant approaches in which the classification rule is chosen immediately after the specification of the object closeness measure. Thus, a maximum linear discriminant rule can follow the specification of a Euclidean norm, and the maximum posterior probability rule can follow a probabilistic measure defined over the space. The weighting and selection of variables follows from the choice of classification rule. The CLAFIC subspace method, by its earlier specification of variable weighting and selection, leaves open the question of whether the variables selected are, in fact, the best discriminators. In this

chapter an attempt will be made to overcome this indeterminacy.

In general, a discriminatory feature will be one which has very different values for objects of different classes. In terms of the subspace model the mean square value of such a feature should be very different from one class to another. In other words, a good discriminating feature would be contained in the subspace of one class but not in the subspaces of the other classes. Different discriminating features, however, will result from choosing different levels of fidelity σ to define the classes' subspaces. This situation will be further discussed in the section on adjustment of σ.

First a method will be given by which a retrenched subspace of discriminating features can be obtained[34] from an already defined CLAFIC subspace. All subspaces S_k will be characterized by their projection operators P_k. A discriminating subspace S_k^* is the intersection of the subspace S_k with the complementary subspaces $\neg S_\ell$ of all the other classes ℓ. The projection operator of S_k^* is

$$P_k^* = \left(\bigcap_{\substack{\ell=1 \\ \ell \neq k}}^{n} [I - P_\ell] \right) \cap P_k \tag{3.1}$$

The operation $P_i \cap P_j$ can be approximated by an iterated matrix product[31]:

$$P_i \cap P_j = P_i \cdot P_j \cdot P_i \cdot P_j \ldots \tag{3.2}$$

The resulting sub-subspaces S_k^* of all classes become orthogonal, because the multiple intersections with the complementary subspaces guarantee that there will be no common subspace between any two S_k^* and S_ℓ^*. By definition, any two subspaces S_k^* and S_ℓ^*, where $k \neq \ell$,

constructed according to (3.1) have projection operators

$$P_k^* = (I - P_1) \cap \ldots \cap P_k \cap (I - P_\ell) \cap \ldots \cap (I - P_n)$$

$$P_\ell^* = (I - P_1) \cap \ldots \cap (I - P_k) \cap P_\ell \cap \ldots \cap (I - P_n)$$

The projection operator of their intersection $S_k^* \cap S_\ell^*$ is denoted by

$$P_k^* \cap P_\ell^* = (I - P_1) \cap \ldots \cap P_k \cap (I - P_k) \cap P_\ell \cap (I - P_\ell) \cap \ldots \cap (I - P_n)$$

but, by definition of complementary subspaces,

$$P_k \cap (I - P_k) = P_\ell \cap (I - P_\ell) = 0$$

Hence,

$$P_k^* \cap P_\ell^* = 0 \qquad (k \neq \ell) \tag{3.3}$$

The remainder of the space not included in the S_k^* subspaces can be described by the projection operator

$$P_{n+1}^* = I - \sum_{k=1}^{n} P_k^* \tag{3.4}$$

The class $(n + 1)$, defined by P_{n+1}^*, is a reject category consisting of nondiscriminatory features. A unique set of basis vectors $\{\underline{u}_i\}$, $(i = 1,\ldots,m)$, can be defined over the space, such that a subset of these corresponds to each class k. Such a subset is denoted by $\{\underline{u}_j^{(k)}\}$, $(j = 0, 1,\ldots,m_k^*)$. This direct consequence of the orthogonality condition is most useful for recognition, because now each class is characterized by a unique subspace (if it exists), which, by being orthogonal to those of all other classes, is at the maximum possible angular distance from them. According to the classification rule (2.10) this results in the best discriminatory ability.

The main drawback to the method described above is that the subspace S_k^* corresponding to a given class may be degenerate,

consisting only of the origin ($m_k^* = 0$). This makes the method useless for recognizing that class. In this case the rejection subspace S_{n+1}^* is too large, meaning that the method is too strict in its specification of discriminating features. The restrictions on subspace definition (as determined by σ) that the above situation implies can be quite severe. Thus, in two-class discrimination each subspace and complementary subspace must be at least two-dimensional, unless the complementary subspace of one class coincides with the subspace of another. In three-class discrimination at least two of the subspaces that intersect to form a retrenched subspace must be of two dimensions and the other must span three dimensions (barring coincidences between the subspaces). The restriction on two-class discrimination also implies that the retrenched subspace approach cannot be expected to work in spaces of two or three dimensions at all, and that four-dimensional spaces are the very least for which two-class discrimination can be attempted. In general, this method is expected to be practicable only for spaces of large dimensionality. The possibility of reducing the reject category by calculating the retrenched subspaces only for pairs of classes runs into the same objection as Fukunaga and Koontz's method[33] for multiclass recognition: it does not give a consistent representation of the class.

Because of the limited application possible for the above method a different approach to the extraction of discriminating subspaces is needed. One approach that leads to discriminating subspaces in the sense that they contain variables which are not common to any other classes is the subtraction from each class subspace those sub-subspaces

shared with other classes. Thus, a subspace S_k^{**} belonging solely to class k would be defined by its projection operator

$$P_k^{**} = P_k - \sum_{\substack{i=1 \\ i\neq k}}^{n} P_k \cap P_i + \sum_{\substack{i=1 \\ i\neq k}}^{n} \sum_{\substack{j=i+1 \\ j\neq k}}^{n} P_k \cap P_i \cap P_j - \ldots \tag{3.5}$$

Unlike the subspaces S_k^*, the members S_k^{**} of this new set will not necessarily be mutually orthogonal. It was noted earlier that orthogonality of the subspaces is a useful property when classification is accomplished by the angular distance criterion. It will be shown next that orthogonality of class subspaces has a more general and desirable quality, that of allowing the definition of a type of conditional probability on the space.

Criteria of Classification

In Chapter II the closeness of a vector $\underline{x}$ to a subspace S_k was given by

$$d_k(\underline{x}) = \frac{\underline{x}^T P_k \underline{x}}{\underline{x}^T \underline{x}} \tag{3.6}$$

When $\underline{x}^T\underline{x} = 1$ and when (2.8) is substituted for P_k:

$$d_k(\underline{x}) = \sum_i^m \sum_j^m x_i x_j \sum_{s=1}^{m_k^*} u_{si} u_{sj}$$

The decision rule (2.10) for classifying $\underline{x}$ was to assign $\underline{x}$ to class C_k if

$$d_k(\underline{x}) = \max_j \{d_j(\underline{x})\}$$

An important property relating the average of $d_k(\underline{x})$ to subspace fidelity is now derived. The average of $d_k(\underline{x})$ over the paradigms

of class k is

$$E_k\{d_k(\underline{x})\} = \sum_{i=1}^{m} \sum_{j=1}^{m} G_{kij} P_{kij} = \mathrm{tr}(G_k P_k) \tag{3.7}$$

The eigenvalue equation

$$\lambda_s \underline{u}_s = G_k \underline{u}_s$$

is premultiplied by $\underline{u}_s^T$ so that

$$\lambda_s \underline{u}_s^T \underline{u}_s = \underline{u}_s^T G_k \underline{u}_s$$

$$\lambda_s = \underline{u}_s^T G_k \underline{u}_s \; / \; \underline{u}_s^T \underline{u}_s$$

$$\sigma = \sum_{s=1}^{m_k^*} \lambda_s = \sum_{s=1}^{m_k^*} \underline{u}_s^T G_k \underline{u}_s \; / \; \underline{u}_s^T \underline{u}_s$$

The eigenvectors are normalized: $\underline{u}^T \underline{u} = 1$, so that

$$\sigma = \sum_{s=1}^{m_k^*} \underline{u}_s^T G_k \underline{u}_s = \sum_i \sum_j G_{kij} P_{kij} \tag{3.8}$$

The right-hand sides of (3.7) and (3.8) are equal, hence

$$\sigma = E_k \{d_k(\underline{x})\} \tag{3.9}$$

This is the average weight of the paradigms of class k in their own subspace. The average inclusion of the paradigms of class k in the subspace of another class ℓ is

$$E_k\{d_\ell(\underline{x})\} = \sum_{s=1}^{m_\ell^*} \sum_i^m \sum_j^m E_k\{x_i x_j\} v_{si} v_{sj} = \mathrm{tr}(G_k P_\ell) \tag{3.10}$$

where $\underline{v}_s$ is the s-th eigenvector of class ℓ. This average can also be interpreted as the inclusion of a representative vector of class k in the subspace S_ℓ.

For a given vector it is possible to find a quantity that satisfies all the rules of a probability measure by normalizing the weight function $d_k(\underline{x})$:

$$w_k(\underline{x}) = \frac{d_k(\underline{x})}{\sum_{k=1}^{n} d_k(\underline{x})} \tag{3.11}$$

The denominator is a function of $\underline{x}$; hence, $w_k(\underline{x})$ has the form of a posterior probability:

$$p(k/\underline{x}) = \frac{f(k \cap \underline{x})}{f(\underline{x})} \tag{3.12}$$

where f(.) is an arbitrary function satisfying the probability axioms.

A more basic measure is the conditional probability of the vector $\underline{x}$, given that it belongs to class k:

$$p(\underline{x}/k) = \frac{f(k \cap \underline{x})}{f(k)} \tag{3.13}$$

The subspace model defines the structure of a class and can implicitly define a conditional probability measure on the space. From (3.10) it is seen that a measure representing the average inclusion of a vector of class k in the subspace of class ℓ is $\mathrm{tr}(G_k P_\ell)$. Division by $\sum_{\ell} \mathrm{tr}(G_k P_\ell)$ normalizes this quantity to satisfy the probability axioms, but it does not fit the functional form (3.13) of a conditional probability. However, in the special case when the class subspaces S_ℓ^* are orthogonal to each other,

$$\sum_{\ell}^{n+1} \mathrm{tr}(G_k P_\ell^*) = \mathrm{tr}(G_k I) = \mathrm{tr}(G_k) \tag{3.14}$$

where I is the unit matrix which is the projection operator of the whole space onto itself. The average weight of the paradigms of

class k in S_ℓ^* is then

$$E_k\{d_\ell(\underline{x})\} = tr(G_k P_\ell^*)/tr(G_k) \tag{3.15}$$

This quantity has the functional form (3.13) of a conditional probability. By normalization of the vectors $tr(G_k) = 1$ for all k. Hence, the conditional probability of the representative vector of class k falling in the subspace of class ℓ is

$$p(S_\ell^*/k) = tr(G_k P_\ell^*) \tag{3.16}$$

Then,

$$\sum_{\ell=1}^{n+1} p(S_\ell^*/k) = 1$$

and

$$p(S_\square / k) = 1$$

$$p(S_\phi / k) = 0$$

where $S_\square$ is the whole space and S_ϕ is the zero vector. Because $S_k^* \cap S_\ell^* = S_\phi$,

$$p(S_k^* \cup S_\ell^* / j) = p(S_k^* / j) + p(S_\ell^* / j)$$

This definition of a conditional probability on the space allows the introduction of the prior probabilities p(k) in defining a posterior probability of a representative vector of class k belonging to the subspace of class ℓ by Bayes' formula:

$$p(k/S_\ell^*) = \frac{p(S_\ell^*/k)\ p(k)}{\sum_{j=1}^{n+1} p(j)p(S_\ell^*/j)} \tag{3.17}$$

It can be reasoned that a vector $\underline{x}$ partakes of the subspace of class ℓ with weight $w_\ell(\underline{x})$ and that the probability of $\underline{x}$ really being

of class k when it is in the subspace of class ℓ is $p(k/S_\ell^*)$. To find the posterior probability of any vector $\underline{x}$ belonging to class k the inclusions of $\underline{x}$ in all the subspaces must be taken into account, weighted by their respective probabilities $p(k/S_\ell^*)$ or

$$p(k/\underline{x}) = \sum_{\ell=1}^{m+1} d_\ell(\underline{x}) p(k/S_\ell^*) \tag{3.18}$$

The advantage of being able to represent the classes by a set of orthogonal subspaces is now clear. Without this representation it would be impossible to introduce the conditional probability which in turn allows specification of prior probabilities in the analysis. A maximum posterior probability rule can be then used for classification.

Subspace Definition by Adjustment of Threshold σ

In the discussion so far the threshold σ has been fixed. The value of σ can be adjusted to improve discrimination.

The threshold

$$\sigma_k = \sum_{s=1}^{m_k^*} \lambda_s$$

determines the number of eigenvectors entering into the subspace S_k of class k. Conversely, σ_k is a discontinuous function of the number of dimensions m_k^* of the subspace. Typically the first few eigenvalues are much larger than the others, accounting for a predominant part of the mean square variation of the class. Two typical examples of the relationship between m_k^* and σ are illustrated in Fig. 1 for m = 10.

A fixed level of σ for all classes guarantees that the mean square error of representation of all is homogeneous. It results, however, in the dimensionality of the subspace of each class being different.

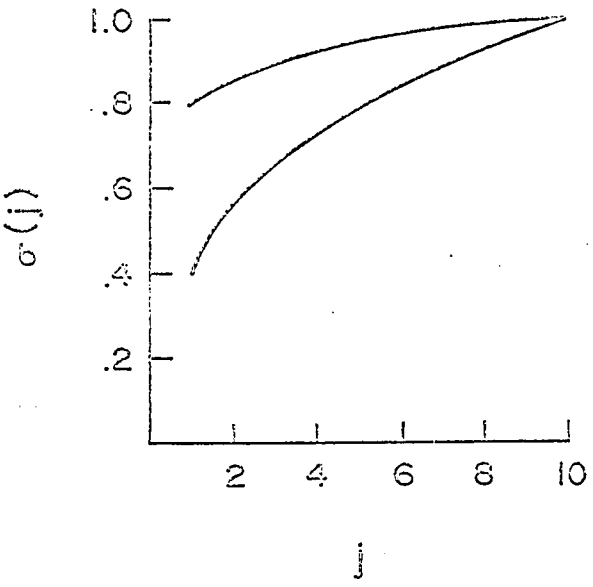

Fig. 1. Fidelity threshold σ vs. subspace dimension.

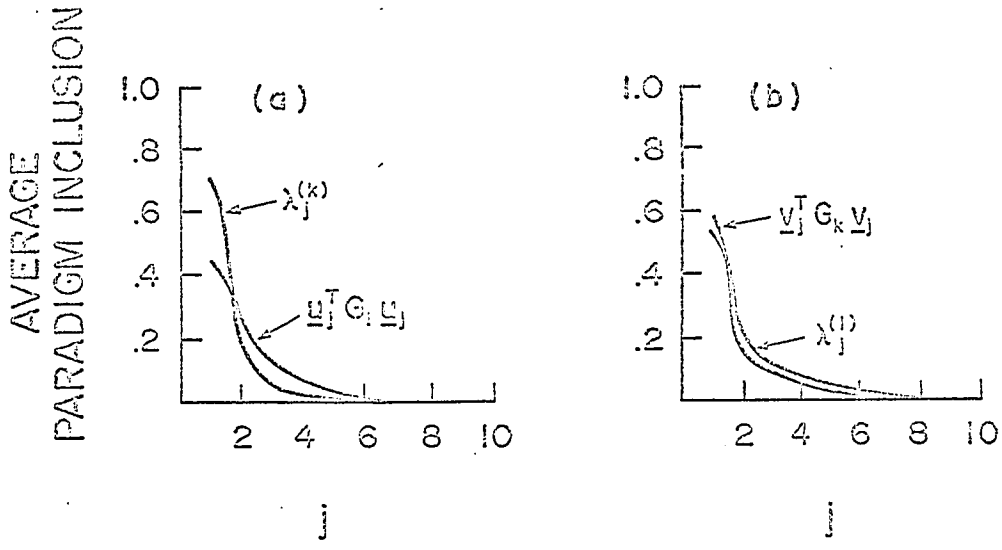

Fig. 2. Curves of average paradigm inclusion in the subspaces of the j-th eigenvector. (a) subspace of class k; (b) subspace of class ℓ.

This can be seen as a widening of the type of discriminatory features allowed in a purely linear model. If, on the other hand, a fixed dimensionality for all class subspaces is used as a criterion instead of a fixed σ, the boundaries between classes are hyperplanes.

It is important to gauge which σ will produce best discrimination. Unfortunately, this quantity is not related in a simple manner to discriminatory ability, because the CLAFIC method results in features which are not necessarily the best discriminators.

What follows refers first to two-class discrimination. In general, it can be said that the larger the σ and the larger the subspace dimension, the larger will be the overlap between the subspaces of two classes, until, for $\sigma = 1$, the whole space is shared by both classes and no discrimination is achieved by the subspace method. If only a few of the eigenvectors are used to define a subspace (σ is small) recognition will be good, provided that these subspaces are at very large angular separation and with little overlap. A good measure of the degree of overlap between two subspaces S_ℓ and S_k is given by the two quantities representing the average inclusion of the paradigms of one in the subspace of the other, $\mathrm{tr}(G_\ell P_k)$ and $\mathrm{tr}(G_k P_\ell)$. For good discrimination it might be expected that $\mathrm{tr}(G_k P_k) > \mathrm{tr}(G_\ell P_k)$ and $\mathrm{tr}(G_\ell P_\ell) > \mathrm{tr}(G_k P_\ell)$. This can be restated as $\mathrm{tr}(G_k P_k) = r_{k\ell}\,\mathrm{tr}(G_\ell P_k)$ for all ℓ, where $r_{k\ell}$ is a constant that depends on k and ℓ and the amount of discrimination between the classes. Hence,

$$r_{k\ell} = \mathrm{tr}(G_k P_k) \;/\; \mathrm{tr}(G_\ell P_k) \tag{3.19}$$

ought to be a good measure of the ability of the subspace of class k to discriminate vectors of its own class from those of another class ℓ.

It is noted from (3.10) that $tr(G_k P_\ell)$ can be decomposed into its components along the eigenvectors that span S_ℓ :

$$tr(G_k P_\ell) = \sum_i^m \sum_j^m G_{kij} \sum_{s=1}^{m_\ell^*} v_{si} v_{sj} = \sum_{s=1}^{m_\ell^*} \underline{v}_s^T G_k \underline{v}_s$$

This quantity can be considered as a function of the number of dimensions j of the subspace of class ℓ:

$$\mu_{\ell/k}(j) = \sum_{s=1}^{j} \underline{v}_s^T G_k \underline{v}_s \tag{3.20}$$

Similarly,

$$\mu_{k/\ell}(j) = \sum_{s=1}^{j} \underline{u}_s^T G_\ell \underline{u}_s$$

and

$$\mu_{k/k}(j) = \sum_{s=1}^{j} \lambda_s^{(k)} = \sigma_k(j)$$

$$\mu_{\ell/\ell}(j) = \sum_{s=1}^{j} \lambda_s^{(\ell)} = \sigma_\ell(j)$$

In Fig. 2a hypothetical curves for $\lambda_s^{(k)}$ and $\underline{u}_s^T G_\ell \underline{u}_s$ are drawn, while Fig. 2b illustrates those for the subspace of class ℓ: $\lambda_s^{(\ell)}$ and $\underline{v}_s^T G_k \underline{v}_s$. The sums of these quantities as functions of j are shown in Fig. 3. The optimal coordinate expansion used for each class guarantees that $\mu_{k/k}(j) \geq \mu_{\ell/k}(j)$ for every value of j. There is no need, however, for $\mu_{\ell/\ell}(j) \geq \mu_{\ell/k}(j)$, even though in practice this is usually the case. The ratios $r_{k\ell}$ and $r_{\ell k}$ can, therefore, be less than unity. This is shown in Fig. 4b for $r_{\ell k}$. In this case the subspace of class ℓ fits the paradigm vectors of class k better than those of

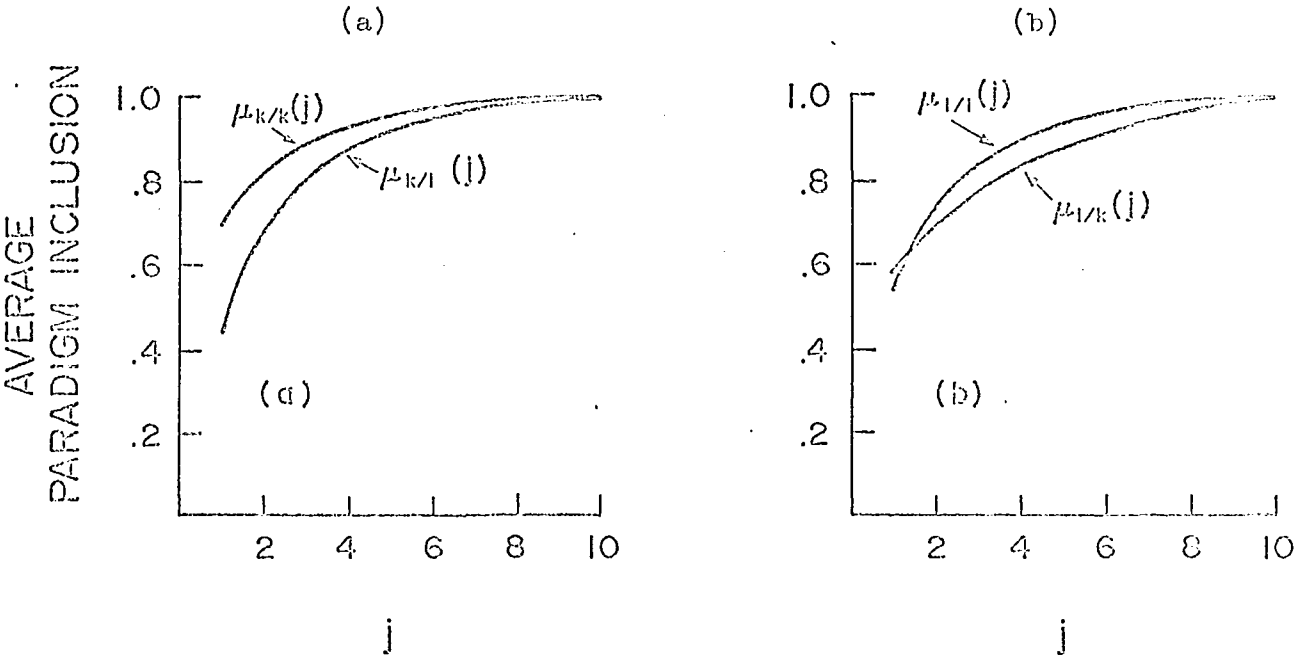

Fig. 3. Curves of average paradigm inclusion in the subspaces of the first j eigenvectors; (a) Subspace of Class k; (b) Subspace of Class ℓ.

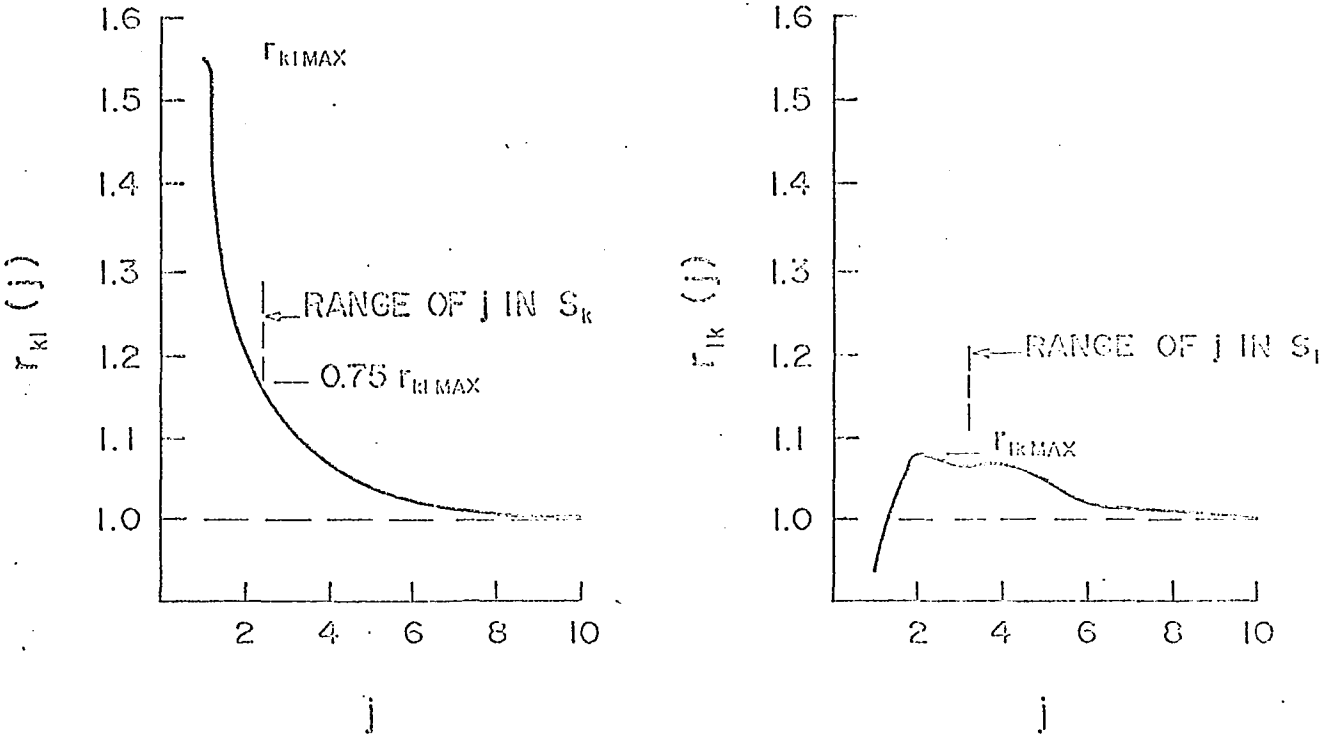

Fig. 4. Curves of $r_{k\ell}(j)$ and $r_{\ell k}(j)$ calculated from the curves of Fig. 3. A threshold of $0.75r_{k\ell MAX}$ results in $m_\ell^* = 2$, $\sigma_\ell(m_\ell^*) = 0.75$, $\sigma_k(m_k^*) = 0.70$ and $m_k^* = 1$.

its own class. The range of j for which the ratio is less than unity cannot be expected to define discriminating subspaces.

A method of selecting σ to yield discriminating subspaces is as follows:

1. Calculate all the $\mu_{k/\ell}(j)$ for all k and ℓ.
2. Calculate the ratios $r_{k\ell}(j) = \mu_{k/k}(j)/\mu_{k/\ell}(j)$.
3. Choose one of the ratios, say $r_{k\ell}(j)$, and require that it be greater than a certain percentage of its maximum. It must also be greater than unity. There will be a certain range of j for which this holds.
4. Determine the range of $\mu_{k/k}(j) = \sigma_k(j)$ corresponding to the range of j chosen in 3. The value of $\sigma_\ell(j)$ will be required to fall within this same range so that both classes are represented with equal fidelity.
5. Find the range of j corresponding to the range of $\sigma_\ell(j)$.
6. Find the j corresponding to the maximum of the ratio $r_{\ell k}(j)$ within the range determined by 5. Set the dimension m_ℓ^* of the subspace S_ℓ equal to this value of j.
7. Calculate the $\sigma_\ell(m_\ell^*)$ corresponding to m_ℓ^*, and determine the m_k^* such that $\sigma_\ell(m_\ell^* - 1) < \sigma_k(m_k^*) \leq \sigma_\ell(m_\ell^*)$.

This method defines the subspaces so as to guarantee a minimum required ratio of correct estimation of the vectors of class k to those of class ℓ in the subspace of the former. Under this constraint the ratio of correct estimation of the vectors of class ℓ to those of class k in the subspace S_ℓ is maximized. It can be expected that this method will result in less subspace overlap than would an arbitrary

choice of σ. To predict the effect on discriminatory performance a relationship must be established with the criterion of classification.

A vector of class k will be correctly classified if $d_k(\underline{x}) - d_\ell(\underline{x}) > 0$. The correct classification of paradigms of class k will be made more unequivocal by the selection of a fidelity σ that maximizes this quantity on the average:

$$D_{k\ell} = E_k\{d_k(\underline{x}) - d_\ell(\underline{x})\} = \mu_{k/k}(m_k^*) - \mu_{\ell/k}(m_\ell^*) \tag{3.21}$$

This quantity is also required to be greater than some positive constant:

$$D_{k\ell} > K_k \tag{3.22}$$

Similarly, it would be desirable to maximize the unequivocalness with which elements of class ℓ are correctly classified. This would require maximizing

$$D_{\ell k} = E_\ell\{d_\ell(\underline{x}) - d_k(\underline{x})\} = \mu_{\ell/\ell}(m_\ell^*) - \mu_{k/\ell}(m_k^*) \tag{3.23}$$

subject to the constraint that

$$D_{\ell k} > K_\ell \tag{3.24}$$

Unfortunately, (3.21) and (3.23) cannot be maximized simultaneously. It is possible, however, to maximize one of them subject to the constraint that the other be greater than a certain minimum difference. The differences $D_{k\ell}$ and $D_{\ell k}$ are functions of m_k^* and m_ℓ^*. These, in turn, are determined by σ. Because of the discreteness of the eigenvalues, $\sigma_k(m_k^*)$ and $\sigma_\ell(m_\ell^*)$ will, in general, be different. A convention must be established to relate them consistently. A convention guaranteeing a fixed fidelity for class k as good as or better than that of class ℓ will be

$$\sigma_\ell(m_\ell^* + 1) > \sigma_k(m_k^*) \geq \sigma_\ell(m_\ell^*) \tag{3.25}$$

The levels of σ will be taken as

$$\sigma = \sigma_k(m_k^*) \tag{3.26}$$

Thus, σ is an upper bound for the true values of fidelity with which the classes are represented. For every value of σ the closest smaller or equal values of $\sigma_k(i)$ and $\sigma_\ell(j)$ will determine m_k^* and m_ℓ^*. The differences $D_{k\ell}$ and $D_{\ell k}$ can then be expressed as functions of σ. These are illustrated in Fig. 5.

A procedure can now be stated for selecting σ:

1. Choose one of the differences, say $D_{k\ell}$, and find the range of σ for which $D_{k\ell}$ is positive and greater than some specified percentage of its maximum.
2. Choose the value of σ that maximizes the other difference $D_{\ell k}$ within the range found in 1.

Though the optimality of the Karhunen-Loève expansion guarantees that $D_{\ell k} > 0$ and $D_{k\ell} > 0$ for $m_k^* = m_\ell^*$ there is no need for this to be true when $m_k^* \neq m_\ell^*$. Thus, it is necessary to place the restrictions that $D_{\ell k} > 0$ and $D_{k\ell} > 0$ on the range of σ within which the maximum of $D_{\ell k}$ is sought.

This procedure maximizes the average margin by which members of class ℓ are correctly classified subject to a minimum required average margin for correct classification of members of class k. The σ resulting from this procedure will give good discriminatory subspaces if the separation between $D_{\ell k}$ and $D_{k\ell}$ is large (reflected by a high value of $D_{\ell k} + D_{k\ell}$) and if the variances of $d_k(\underline{x})-d_\ell(\underline{x})$ and

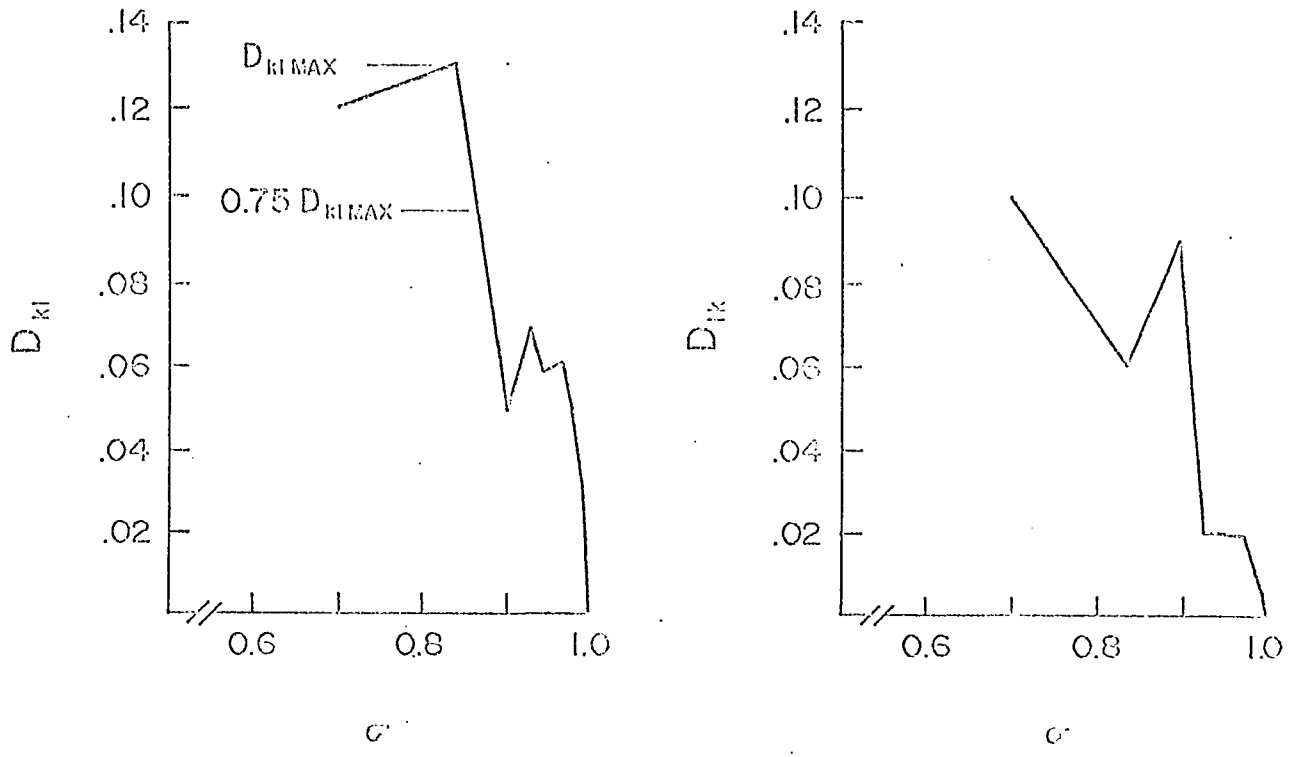

Fig. 5. Curves for $D_{k\ell}$ and $D_{\ell k}$ calculated from the curves of Fig. 3. A threshold of 0.75 $D_{k\ell MAX}$ results in $\sigma = 0.70$ or $\sigma = 0.83$. To maximize $D_{\ell k}$, $\sigma = 0.70$ is chosen.

$d_\ell(\underline{x}) - d_k(\underline{x})$ are small for the paradigms of classes k and ℓ, respectively. The value of σ derived by this procedure will be, in general, different from that obtained by the ratio-maximization procedure. In Chapter V it will be shown how both procedures were tested to compare their performances empirically. Both procedures are well-suited to the problem of medical diagnosis where the categories of ill and healthy patients must be discriminated. It is natural to want to specify a required level for the margin by which ill patients are correctly classified, and, under this condition, to maximize the margin for true negatives.

In multiclass recognition (n classes) the above procedures have a direct extension. Constraints are required on (n - 1) pairwise ratios or differences between classes and the remaining ratio or difference is maximized.

CHAPTER IV

THE COMPUTER-AIDED DIAGNOSIS OF THYROID DYSFUNCTION

Background of the Problem

The goal of the research being reported was to develop one or more computer programs that could be used to aid a clinician in the diagnosis of thyroid dysfunction. Specifically, the usefulness of several pattern recognition techniques was investigated. Their performance was compared to that of a simple Bayes' method and to that of a linear discriminant function.

The computer programs are based on data obtained in a specialty diagnosis situation as described in Chapter I. The patients considered in this study had been referred to Dr. R. A. Nordyke of the Department of Nuclear Medicine at the Straub Clinic in Honolulu. The patients were referred because of the possibility of their suffering from some thyroid disorder, the referrals being obtained either from other departments of the same clinic or from private physicians. Thus, the population on which this study is based was preselected by the fact that all persons were under suspicion of having thyroid disease. The complete diagnosis for each patient included both functional and pathological studies (for carcinoma, nodules, etc.) of the thyroid gland. This dissertation is concerned only with the functional diagnosis. The diagnosis of pathological conditions was not included because of the lack of uniform data. The interpretation of scans and X-ray films is a separate and more complicated visual pattern recognition problem. Therefore, the present study was limited to an analysis of the

patient's functional diagnosis, based on his first visit to the specialist. At the time of the first visit the doctor recorded the patient's answers to a standard series of questions related specifically to signs and symptoms indicative of thyroid dysfunction. The patient next received a standardized physical examination. When laboratory tests were required their results were also recorded in a systematic manner. Thus, complete and uniform data was obtained for almost all patients up to the stage of laboratory testing at their first visit. The problem of follow-up diagnosis has not been considered since uniform data was lacking. The final diagnoses used in this study for classifying patients were confirmed by laboratory tests or response to treatment. Those cases for which no certain diagnosis could be reached are grouped separately and are not included in the model.

The sequence of this study falls into three stages: 1) analysis and preselection of data; 2) development of pattern recognition and other programs to test performance on large numbers of diagnosed patients; 3) development of a sequential program for clinical use.

A brief summary of the mechanism of thyroid function and the symptoms of dysfunction will introduce the data analysis.

Thyroid Function

Hyperthyroidism and hypothyroidism are diseases caused by the overactivity and underactivity, respectively, of the thyroid gland. Though initially not dangerous, they can cause great discomfort to the patient, and, if untreated, can lead to serious illness and even death.

The thyroid gland is located in the neck area over the larynx. It is about 25 grams in weight and butterfly-shaped, consisting of two lobes connected by an isthmus. The thyroid gland processes iodine taken into the body in food, releasing several hormones into the blood. Chief of these hormones is thyroxine, followed by triiodothyronine. The hormones affect cellular metabolic rates and, consequently, energy production. The production and release of thyroid hormones is controlled by the pituitary gland through a feedback mechanism. Thyroid stimulating hormone (TSH) is released by the pituitary when the level of circulating thyroid hormones falls below a certain level. TSH causes the thyroid to step up its synthesis and release of thyroid hormones. When these again reach a normal level the pituitary is no longer stimulated to release TSH.

In thyroid dysfunction the normal feedback mechanism is disrupted. Hyperthyroidism is marked by the excessive secretion and release of thyroid hormones. This results in an above-normal metabolic rate and increased energy release. The symptoms thus produced are nervousness, excessive perspiration, irritability, weight loss, increased appetite, and others. Hypothyroidism is marked by the below-normal secretion and release of thyroid hormones. Since cellular metabolism is thereby decreased, symptoms include a decreased appetite, weight gain, and lethargy. The number of apparent symptoms and their severity depend on how long the disease has been present.

A simple physical examination will strengthen or lessen the suspicion of thyroid dysfunction. The Achilles heel reflex time will be faster in a hyperthyroid patient, slower in a hypothyroid one.

Pulse rates tend to be higher in hyperthyroid persons. The skin of a hyperthyroid person tends to be warm and moist, while that of a hypothyroid patient is cool and dry. A hyperthyroid person might have a fine tremor, best seen in the fingers. The size and consistency of the thyroid gland must also be considered.

However, symptoms and a simple physical examination alone are not conclusive enough to diagnose thyroid dysfunction, unless the disease has progressed to a very serious and obvious stage. Patients who exhibit a sufficient number of suspicious symptoms and findings are sent to undergo laboratory tests that will indicate more accurately the state of activity of the thyroid gland. Most of these tests measure the amount of thyroid hormones circulating in the blood--T_3 red cell uptake or resin test, for example, or protein-bound iodine test. Others measure the rate at which the thyroid picks up radioactively marked iodine--6 and 24 hour I^{131} uptakes. Still another test measures the metabolism of the patient at rest, the metabolic rate.

A test of the correctness of the final diagnosis is treatment. If the patient responds to medication his disorder was probably correctly diagnosed. In those cases where even laboratory tests are inconclusive suppression tests can be conducted.

Description of the Data

The data initially available for this study consisted of the records of 3832 patients seen by Dr. R. A. Nordyke between 1963 and 1968. For classification purposes the data was divided according to the final diagnosis. The diagnostic categories that were of interest

for this study are listed below:

1. Hyperthyroid: a) definite with Graves' disease (exophthalmic goiter).
   b) probable with various pathological conditions.
2. Hypothyroid: definite, subcategorized as mild, moderate, or severe, with varied pathology.
3. Euthyroid I: patients with normal thyroid function and pathology.
4. Euthyroid II: patients with normal function but with pathological abnormalities of the thyroid.

In Table 2 are shown the distributions of patients in each of the four categories. The breakdown by years is given in Appendix I, including 1969 data which was used as an additional independent test sample.

Not all patients had complete records. To gather a consistent set of statistics those records with many missing items were excluded from the study. The results of this first selection of data are shown in Table 2. A second selection was made in which only those records having at least three laboratory tests (T_3RCU, 6 and 24 hour I^{131} uptakes) were kept. The patient distribution resulting from this is shown in Table 2.

The original data on the thyroid questionnaire (see Appendix II) included many qualitative questions with multiple answers. For the pattern recognition and discriminant analysis programs quantitative data were required. This meant that the questions had to be divided

TABLE 2

DISTRIBUTION OF PATIENTS BY DISEASE CATEGORIES

Category	All Patients	Patients with Complete Records to Physical Exam. Inclusive	Patients with Complete Records to Laboratory Tests Inclusive*
1. Hyperthyroid	327	263 : 196 confirmed cases 67 probable cases	149 : 126 confirmed cases 23 probable cases
2. Hypothyroid	640	515 : 208 confirmed cases 307 probable cases	273 : 117 confirmed cases 156 probable cases
3. Euthyroid I	1616	1284	800
4. Euthyroid II	1249	999	424
Total in All Categories	3832	3061	1646

*T_3RCU and 6 and 24 hour I^{131} uptakes.

and reformulated to make the answers binary (yes-no) variables. A subset of these was selected by the physician as good indicators of the diseases and as easiest to record by paramedical personnel. This subset is listed in Table 3. The Bayes' probabilistic model could use the qualitative questions directly, resulting in some altered questions as listed in Table 4. In these tables the data are divided into the major sections of History, Physical Examination, and Laboratory Tests.

Statistics for the variables listed in Tables 3 and 4 are given in the form of a matrix of frequencies in Tables 5 and 6 for the three categories used as paradigms. Histograms of the continuous variables for each of the diagnostic categories are shown in Figs. 6-12. Table 7 shows the means, standard deviations, and maximum and minimum values of these variables.

Pattern Recognition Computer Programs for Diagnosis

Several computer programs were developed to test the effectiveness of the pattern recognition subspace methods described in Chapters II and III. These were designed as batch process programs to efficiently test the methods on the large amount of available data. On the basis of the results from these programs a sequential program was designed for clinical use. This will be described in the next section. While details of the programs are relegated to Appendix III some of their principal characteristics will be described here.

The specification of parameters is the first step of the CLAFIC-type programs. These parameters control the operation of the program in the following areas:

TABLE 3

QUESTIONS ON THYROID HISTORY AND PHYSICAL EXAMINATION SELECTED AND BINARIZED FOR ANALYSIS

	History
1.	Sex: male or female?[a]
2.	Has there been previous thyroid surgery?[b]
3.	Has there been a recent increase in weight over 5%?
4.	Has there been a recent decrease in weight over 5%?
5.	Has there been a recent appetite increase?
6.	Has there been a recent appetite decrease?
7.	Have rapid heart beatings or palpitations been noticed recently?
8.	Has nervousness increased recently?
9.	Has irritability increased recently?
10.	Has dry skin been noticed recently (within the 6 months prior to examination)?
11.	Has there been recent intolerance to heat?
12.	Has there been recent intolerance to cold?
13.	Has there been recent diarrhea?
14.	Has there been recent weakness or fatigue?
15.	Is there a family history of hyperthyroidism?
16.	Is there a family history of other thyroid disease?
17.	Has there been a recent increase in perspiration?
	Physical Examination
18.	Can a fine tremor be detected in the fingers?
19.	Is the skin warm and moist?
20.	Is the skin cool and dry?
21.	Is the thyroid gland enlarged?

a male = 0; female = 1.

b all the remaining questions are answered by no = 0, yes = 1.

TABLE 4

MULTIPLE ANSWER QUESTIONS USED IN THE SIMPLE BAYES' MODEL
IN PLACE OF THE CORRESPONDING ITEMS IN TABLE 3

History
1. How has weight changed recently? a) loss over 5% b) gain over 5% c) unchanged or other (3)*
2. How has appetite changed recently? a) decreased b) increased c) unchanged (4)
3. Is there any indication of thyroid disease in the family? a) hyperthyroidism b) other thyroid disease c) none (15 & 16)
Physical Examination
4. Skin observations: a) warm and moist b) cool and dry c) other (19 & 20)

*numbers in parentheses indicate the corresponding questions in Table 3.

TABLE 5

DATA MATRIX FOR THE PRINCIPAL DIAGNOSTIC CATEGORIES (BINARY VARIABLES)

		Hyperthyroid with Graves' Number of Cases (%)	Hypothyroid Number of Cases (%)	Euthyroid I Number of Cases (%)
	Total Cases	196 (100.0)	208 (100.0)	1284 (100.0)
1.	Sex: male	50 (25.5)	12 (5.6)	410 (31.9)
	female*	146 (74.5)	196 (94.2)	874 (68.1)
2.	Previous Thy. Surg.	18 (9.2)	34 (16.3)	4 (0.3)
3.	Weight Increase	21 (10.7)	48 (23.1)	276 (21.5)
4.	Weight Decrease	106 (54.1)	19 (9.1)	201 (15.7)
5.	Appetite Increase	75 (38.3)	14 (6.7)	98 (7.6)
6.	Appetite Decrease	28 (14.3)	20 (9.6)	148 (11.5)
7.	Rapid Heart/Palpit.	129 (66.9)	28 (13.4)	220 (17.1)
8.	Nervousness	121 (61.7)	40 (19.2)	277 (21.6)
9.	Irritability	85 (43.5)	50 (24.0)	309 (24.1)
10.	Dry Skin	14 (7.2)	82 (39.4)	160 (12.4)
11.	Heat Intolerance	87 (44.4)	13 (6.3)	102 (7.9)
12.	Cold Intolerance	2 (1.0)	16 (7.7)	45 (3.5)
13.	Diarrhea	37 (18.9)	11 (5.3)	69 (5.4)
14.	Weakness/Fatigue	130 (66.3)	121 (58.2)	532 (41.4)
15.	Family Hist. Hyper.	37 (18.9)	13 (6.3)	45 (3.5)
16.	Family Hist. Other	26 (13.2)	54 (25.9)	187 (14.6)
17.	Perspiration Incr.	93 (47.5)	20 (9.6)	175 (13.6)
18.	Fine Tremor	137 (69.9)	0 (0.0)	141 (11.0)
19.	Warm, Moist Skin	60 (30.6)	4 (1.9)	70 (5.5)
20.	Cool, Dry Skin	1 (0.5)	33 (15.9)	49 (3.8)
21.	Enlarged Thyroid	181 (92.3)	96 (46.2)	62 (4.9)

*for all but question 1 the counts for "yes" answers only are recorded.

TABLE 6

DATA MATRIX FOR THE PRINCIPAL DIAGNOSTIC CATEGORIES (QUALITATIVE VARIABLES)

	Hyperthyroid with Graves' Number of Cases (%)	Hypothyroid Number of Cases (%)	Euthyroid I Number of Cases (%)
Total Cases	196 (100.0)	208 (100.0)	1284 (100.0)
1. Weight Change			
a) loss over 5%	106 (54.1)	19 (9.1)	201 (15.7)
b) gain over 5%	21 (10.7)	48 (23.1)	276 (21.5)
c) unchanged/other	69 (35.2)	141 (67.8)	807 (62.8)
2. Appetite Change			
a) decreased	28 (14.3)	20 (9.6)	148 (11.5)
b) increased	75 (38.3)	14 (6.7)	98 (7.6)
c) unchanged	93 (47.5)	174 (83.7)	1037 (80.8)
3. Family Hist. Thyroid			
a) hyperthyroidism	37 (18.9)	13 (6.3)	45 (3.5)
b) other thy. disease	26 (13.2)	54 (25.9)	187 (14.6)
c) none	133 (67.9)	141 (67.8)	1052 (81.9)
4. Skin Observations			
a) warm/moist	60 (30.6)	4 (1.9)	70 (5.5)
b) dry/cool	1 (0.5)	33 (15.9)	49 (3.8)
c) other	135 (68.9)	161 (82.2)	1165 (90.7)

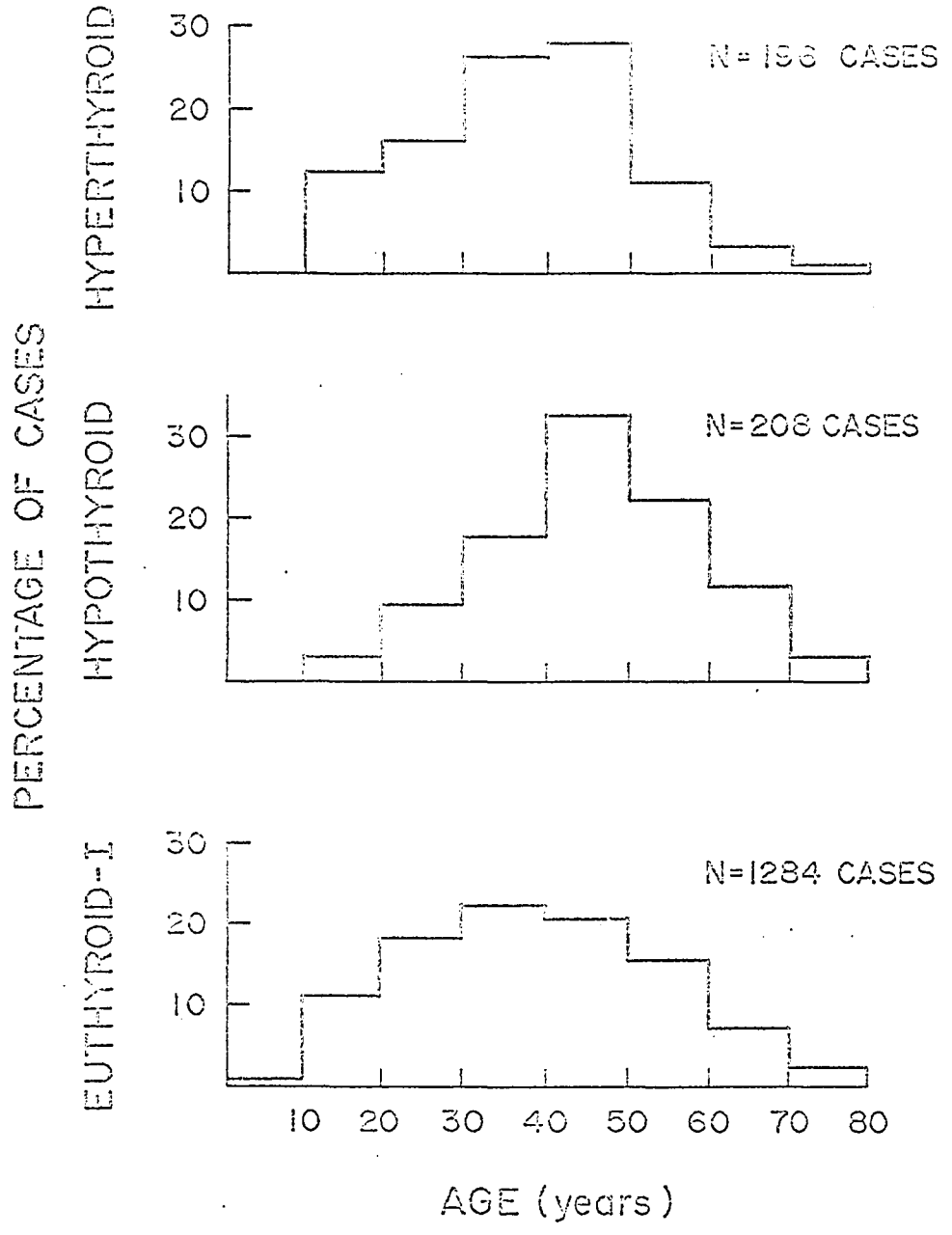

Fig. 6. Age distributions.

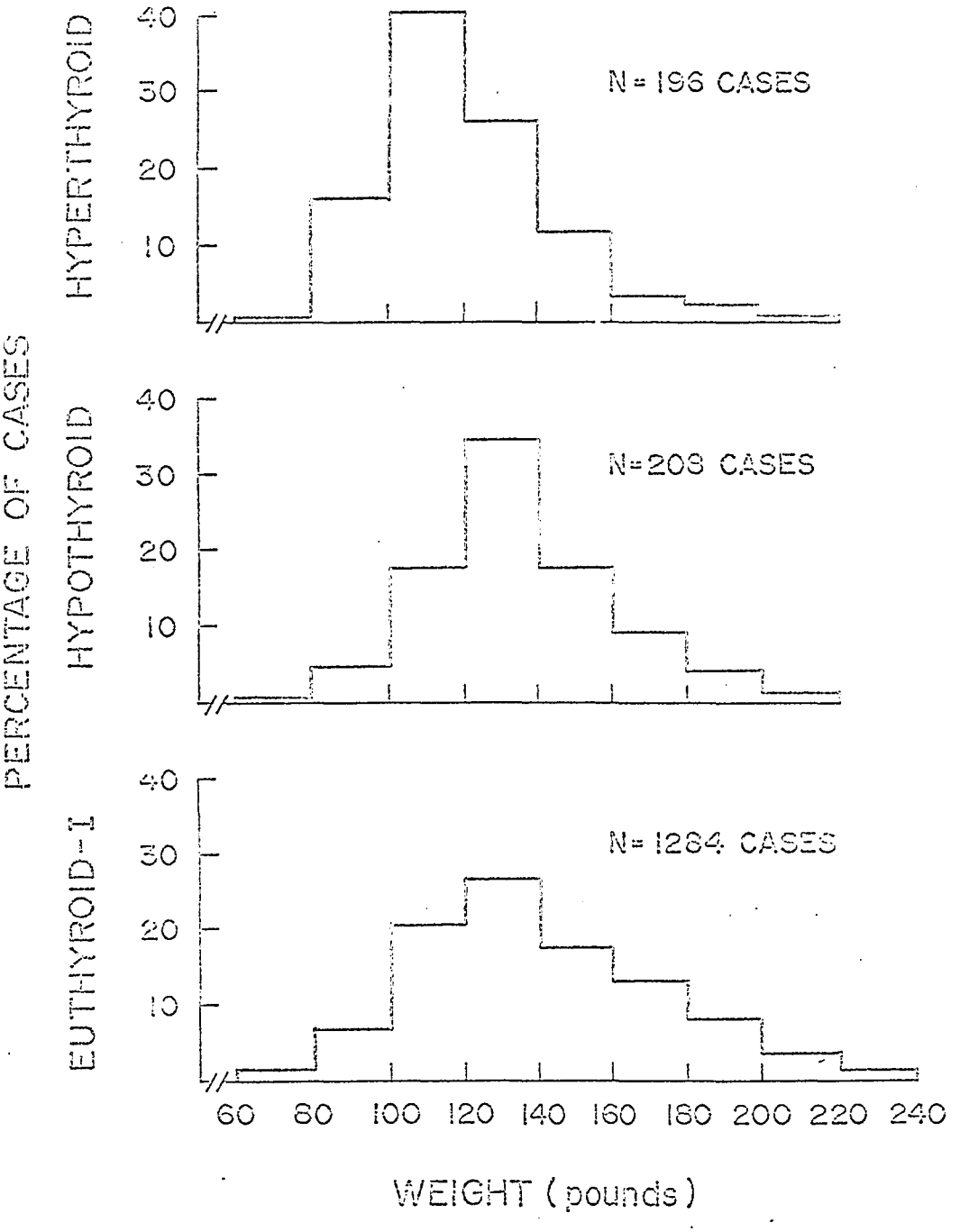

Fig. 7. Weight distributions.

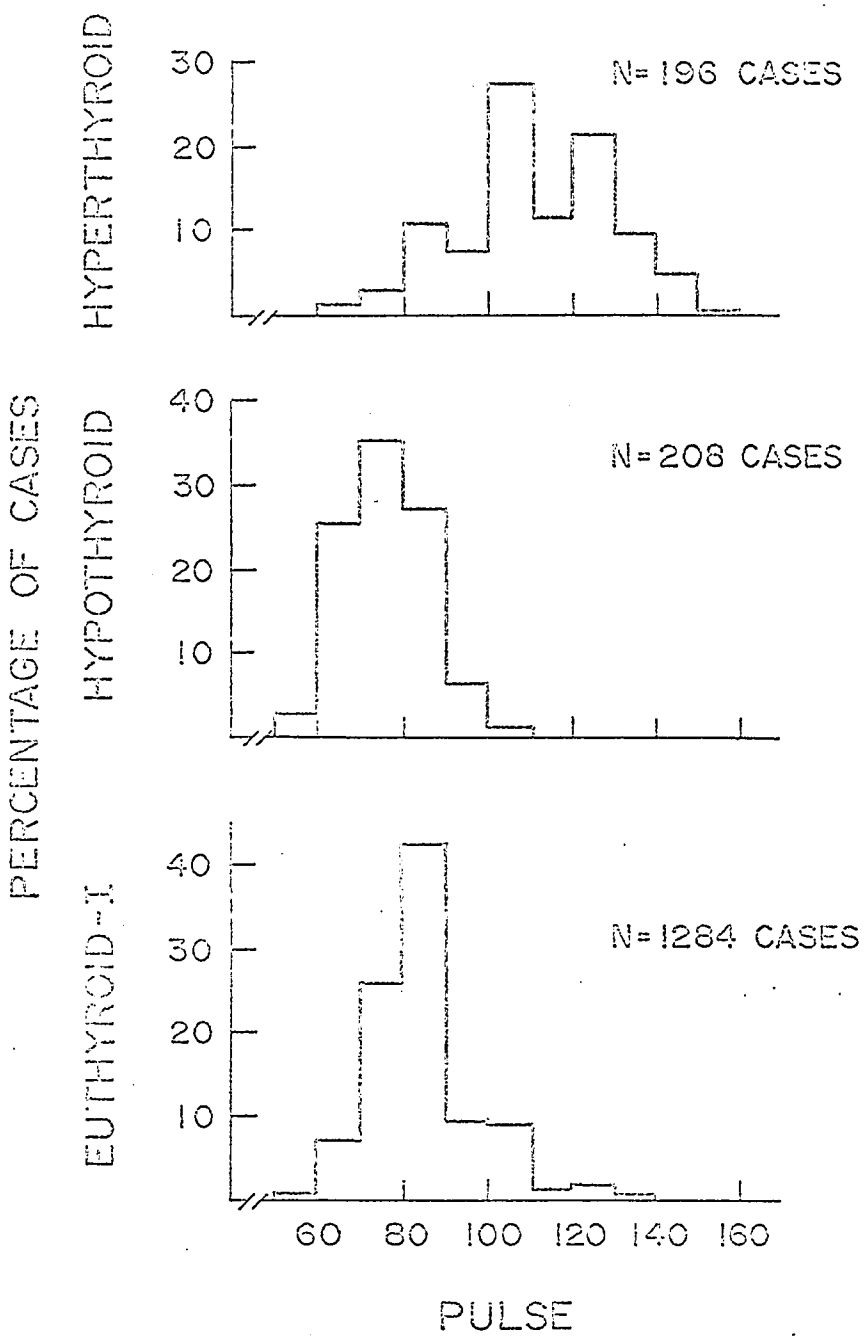

Fig. 8. Pulse distributions.

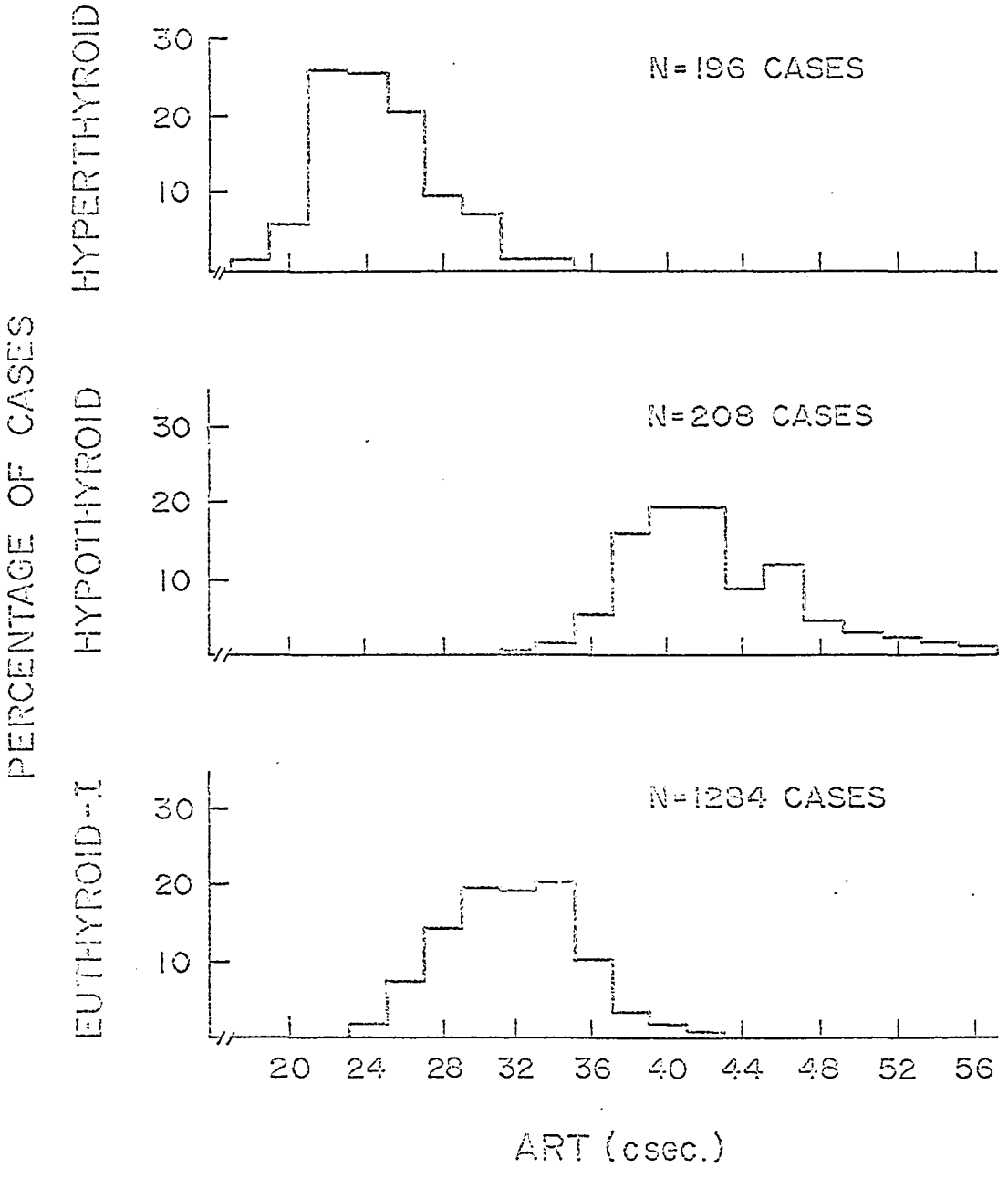

Fig. 9. Achilles heel reflex time distributions.

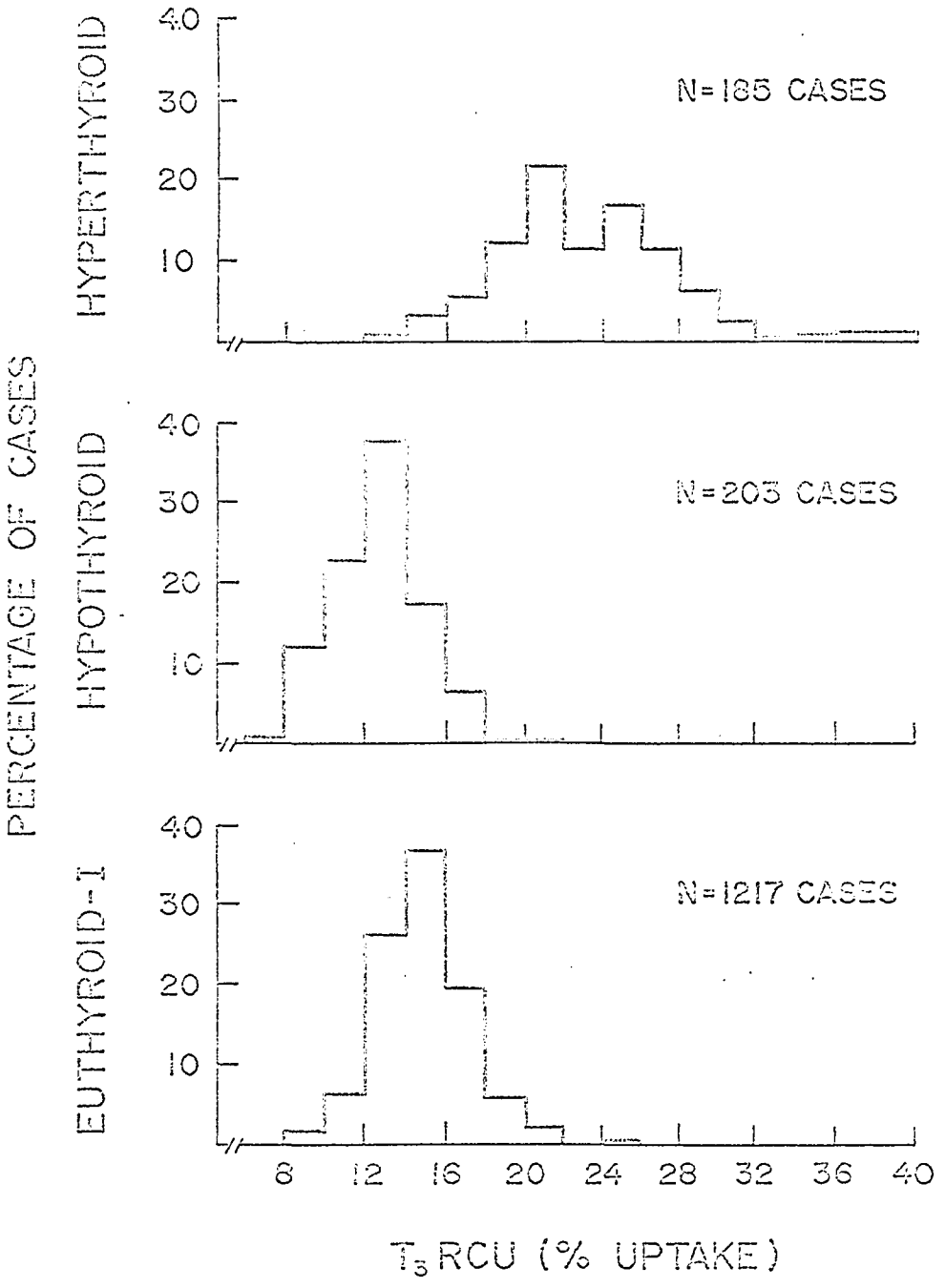

Fig. 10. T_3 red cell uptake distributions.

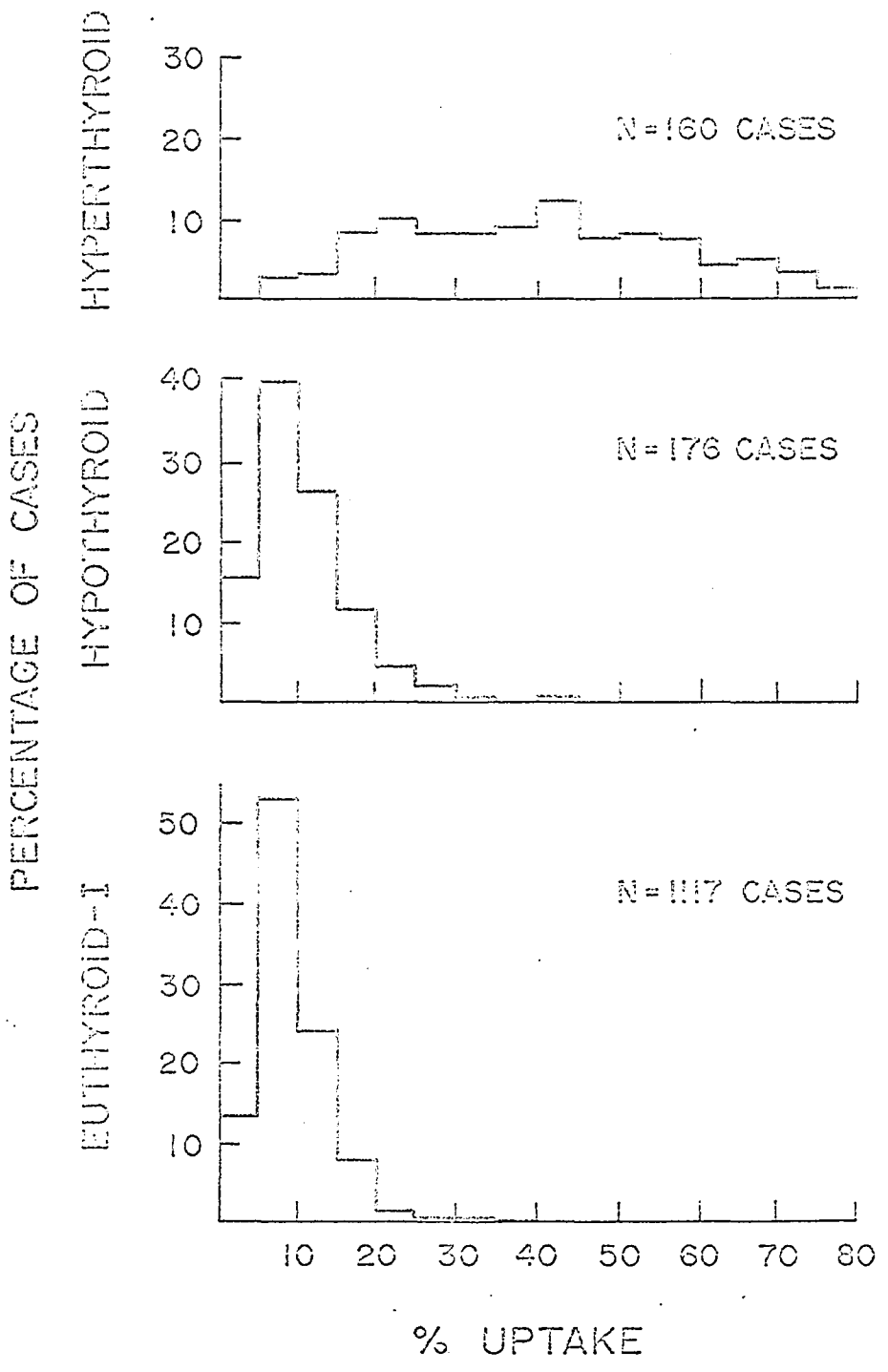

Fig. 11. 6 hour I^{131} uptake distributions.

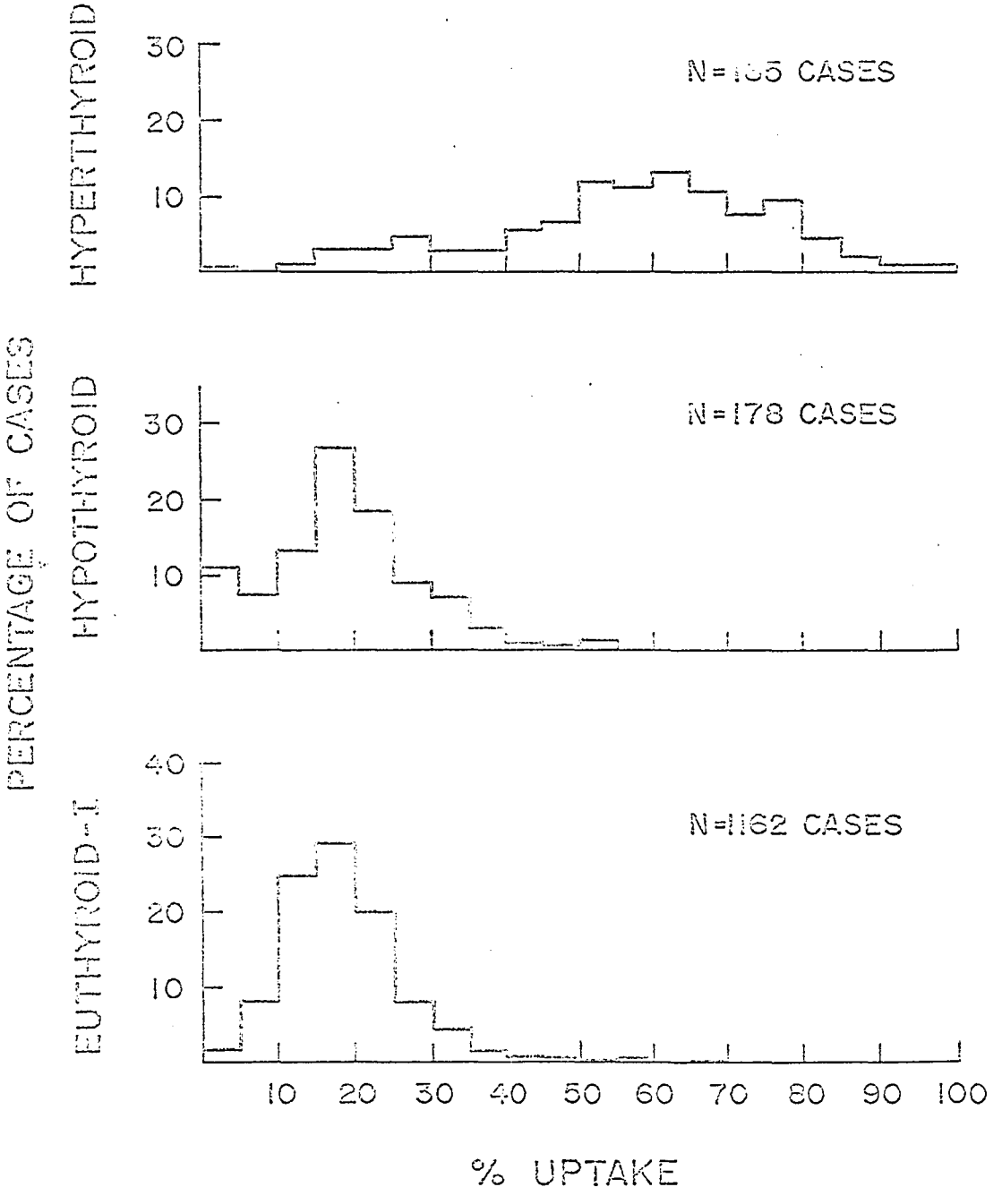

Fig. 12. 24 hour I^{131} uptake distributions.

TABLE 7

STATISTICS FOR CONTINUOUS VARIABLES

Variable		Mean	Standard Deviation	Minimum Value	Maximum Value
Age	Hyper.	37.5	13.07	12.0	74.0
	Hypo.	45.2	13.33	10.0	78.0
	Euth. I	38.4	15.43	5.0	82.0
Weight	Hyper.	118.9	22.94	72.0	199.0
	Hypo.	134.7	26.26	75.0	234.0
	Euth. I	139.5	33.11	41.0	245.0
Pulse	Hyper.	109.0	17.45	60.0	150.0
	Hypo.	75.3	10.04	48.0	108.0
	Euth. I	84.2	12.89	48.0	140.0
ART	Hyper.	24.7	3.10	18.0	34.0
	Hypo.	42.6	4.94	32.0	60.0
	Euth. I	31.7	3.55	24.0	44.0
T_3RCU	Hyper.	23.3	4.88	13.4	39.3
	Hypo.	12.6	2.24	7.6	20.7
	Euth. I	14.8	2.29	8.4	24.8
6 hour I^{131}	Hyper.	40.3	17.66	7.0	78.0
	Hypo.	10.9	6.11	1.0	42.0
	Euth. I	9.8	4.78	1.0	65.0
24 hour I^{131}	Hyper.	56.4	18.79	2.0	98.0
	Hypo.	19.4	10.24	1.0	55.0
	Euth. I	18.7	7.54	1.0	66.0

1. the manner in which the data is to be read in.
2. the specification of model statistics--if already available they can be read in directly; if not, they are calculated.
3. the choice of variables and their characterization as binary or continuous.
4. the selection of artificial features (combinations of variables).
5. the definition of subspace representation of classes by the choice of a "fidelity criterion" for average representation of paradigms.
6. the printing and other output options.

The second step of the programs is to read in the data and perform a normalization. The data are read in successively by classes (diagnostic categories) by subroutine INPUT. There is a provision for a "class" of unknown classification (hereafter referred to as unknowns) which allows the testing of the program on samples independent of the samples of paradigms upon which the class models are built. Immediately after allowing this the data can be augmented (if this is indicated by one of the control parameters) to include quadratic terms, thus extending the class of models generated by this program. In order to define the class subspaces it is necessary to set the norm of every data vector equal to unity. In addition, prescaling of the binary and continuous variables separately is carried out to give each type a unity norm. This is accomplished for all data vectors by subroutine NORM.

The optimal coordinate system calculations for each class are

carried out, when required, by subroutine KLEXP. First, the auto-correlation matrix of a class is calculated. The eigenvalues and eigenvectors of this matrix are calculated by subroutines EIGEN and ARRAY which are standard IBM Scientific Package Subroutines. The former uses the Jacobi diagonalization method for eigenvalue calculation. The eigenvectors are stored on tape and punched on cards for future use. They are the principal components of each class in terms of which the subspace representation is desired. After the eigenvectors of each class have been calculated the main program calls on subroutine TRAN to calculate the number of eigenvectors that will be used for each class in order that the vectors of that class be represented on the average with a specified fidelity (THR).

The last part of the program is comprised of the classification routines for the paradigms (subroutine COMP) and unknowns (subroutine RECOG). Both subroutines call on subroutine SELCT to compute the cosine squared of the angle between a data vector and its projection on the subspace of a class. This measure of subspace inclusion is used by the calling subroutines which choose the maximum inclusion and print out the resulting classification. In COMP a confusion matrix is printed illustrating the classification of the paradigms and serving as a check of the fit between the data and the model.

Sequential Diagnosis (On-Line Program)

The programs described in the preceding section were used in the research stage. For clinical application an on-line program was needed to be interactive with the doctor and aid in his diagnosis. It required a structure different from that of the batch programs

described before. As a preliminary step a sequential program to operate in the batch mode was designed. This contained all the characteristics desired in the eventual on-line program and was, in effect, a modelling of such a program. The principal difference from the programs described in the preceding section is that all statistics and class representation models are resident on disk or supplied at the start of the program. Several representations in terms of different subsets of variables must be available to cover the possibility of incomplete information for a given patient. The classification of a patient is required at the end of each major stage of data collection: history and physical examination, and after every laboratory test. This sequential nature of the program meant that thresholds based on tolerable levels of misclassification had to be set at every stage. Such thresholds were chosen empirically based on the performance of the program on sets of paradigms and unknown samples. In addition, diagnostic rules were incorporated into the program so that, depending on the results of the patient's examination at a given stage, different tests or treatment would be prescribed.

Comparison With Other Diagnostic Programs

Two other diagnostic programs were compared to those based on the pattern recognition models. They were the simple Bayes' model and a stepwise discriminant program. The latter is the standard program available in the Biomedical Computing Library (BMD-Vol. II-M7) and its results will be described in the next chapter. The Bayesian model was implemented by a program called BAYES (see Appendix IV for details) and calculated the posterior probabilities of the various diseases

for a given symptom set. This program had the convenience of being able to use qualitative variables directly; because of the zero-order model of the joint distributions it could also tolerate incomplete data with the greatest ease of all the models.

CHAPTER V

RESULTS AND DISCUSSION

Subspace Models for the Diagnosis of Thyroid Dysfunction

The CLAFIC method was first used in diagnosis to discriminate hyperthyroid from non-hyperthyroid patients. Data from 1963 to 1967 were used as paradigms to determine the statistics of the diagnostic categories and 1968 data were used as independent test samples. The optimal coordinate systems were generated for each of the main stages of the diagnostic process. The subspaces were defined by a fidelity criterion that was chosen by trial and error to give good discrimination. Recognition of both paradigms and test samples was carried out first in a single stage classification program, and then in a sequential program that allowed for a deferred-judgment category. With only variables from the medical questionnaire and physical examination included the programs satisfactorily screened out about 90% of non-hyperthyroid patients while correctly diagnosing over 95% of hyperthyroid patients. When laboratory test results were included over 98% of hyperthyroid and 94% of other patients were correctly recognized. The detailed results of this work have been reported elsewhere.[4]

Some of the variables used in the above study were very dependent on the experience and methods of the doctor making the observations. It was felt that results of wider applicability would be obtained by including only those variables which are easily reproducible by other physicians and paramedical personnel. Such a set of variables was

described in Chapter IV and has been chosen for the work reported in this dissertation. Here the data from 1963 to 1968 have been used as paradigms and the 1969 data as independent test samples. The computer programs for diagnosis have been extended to include the three-way recognition of hyperthyroid/hypothyroid/euthyroid cases.

The main result of this work in pattern recognition has been the development of methods for choosing discriminatory subspace representations of classes. The most practical procedures were ones for the adjustment of the fidelity threshold σ as described in Chapter III. The thyroid data were used to test the efficacy of these procedures. Results with different values of σ for two-way classifications into hyperthyroid/euthyroid and hypothyroid/euthyroid and for the three-way recognition of all categories were obtained at the stages of medical history, physical examination, and three completed laboratory tests. The approach taken to test the σ adjustment procedures will be described in detail for the stage that includes medical history and physical examination items. Classification results at this stage are probably most significant in clinical practice since they include the basic information on which healthy (normal thyroid function) patients can be screened without the need of expensive and time-consuming laboratory tests. Twenty-five variables are used at this stage: the 21 binary variables of Table 5 and the first four continuous variables of Table 7.

Before any classifications were attempted a program was run to calculate the average inclusions of the paradigms of each class in the subspaces of every other class. These inclusions are shown in Fig. 13.

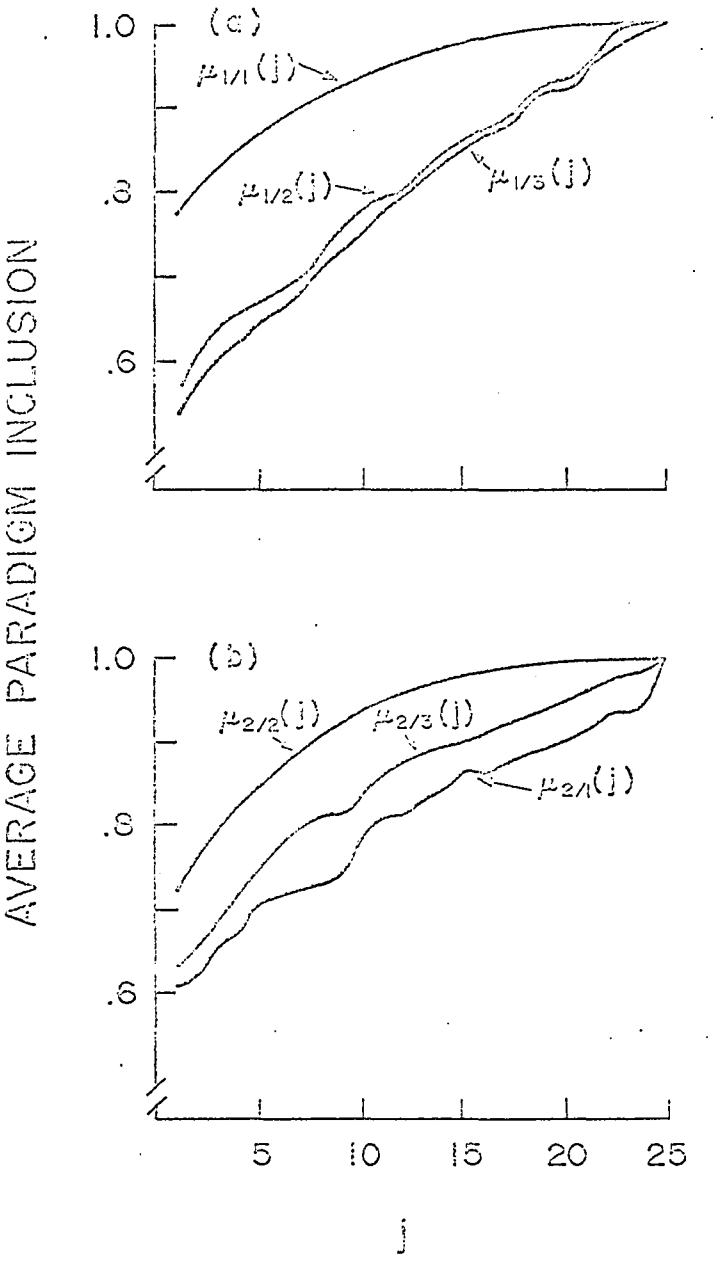

Fig. 13. Average inclusion of paradigms in the class subspaces versus subspace dimension j. (a) Class 1: hyperthyroid; (b) Class 2: hypothyroid.

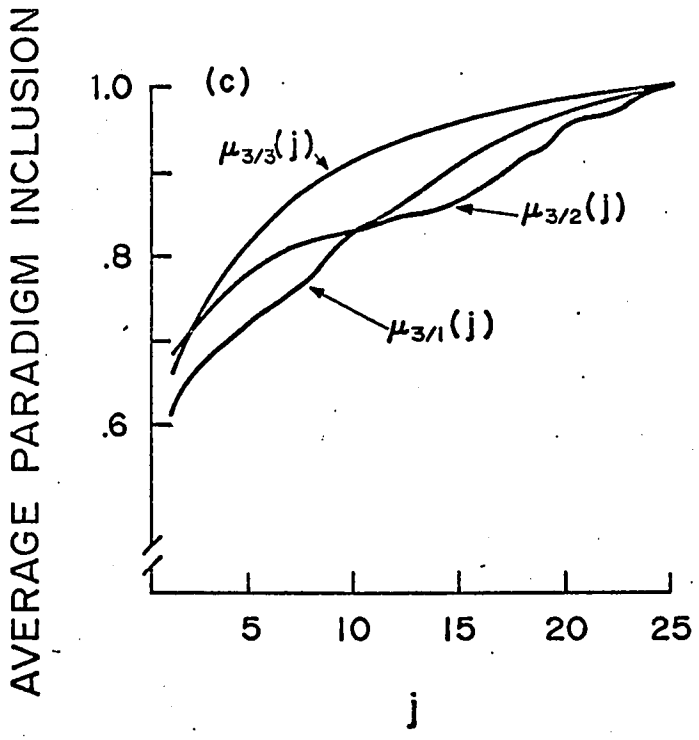

Fig. 13 cont. (c) Class 3: euthyroid I. m = 25.

In addition, the ratios $r_{k\ell}$ of the paradigm inclusions and the average margins of correct classification $D_{k\ell}$ were calculated for all classes and subspaces. The ratios are shown in Fig. 14 as functions of subspace dimensionality. The average margins $D_{k\ell}$ are illustrated in Figs. 15 and 16 for hyperthyroid/euthyroid and hypothyroid/euthyroid discrimination only, since these are most interesting in diagnosis. The overlap of hypothyroid and hyperthyroid categories is no problem in practice.

In Chapter III a rule was given to choose the dimensionality m_k^* of the subspaces in such a way that the paradigms of each are represented with comparable fidelity: for every value of m_k^* the dimensionality of the other subspaces S_ℓ is chosen such that m_ℓ^* includes as many eigenvectors as necessary to make its fidelity $\sigma_\ell(m_\ell^*)$ just smaller or equal to $\sigma_k(m_k^*)$. This rule was derived by experience in classification. It gave consistently better recognition for class C_k than the only other rule that yields comparable values of σ for all classes (for this rule choose m_ℓ^* such that $\sigma_\ell(m_\ell^*)$ is just greater or equal to $\sigma_k(m_k^*)$). Thus, if the best performance is required for a given class, as is the case in diagnosing ill patients where a high level of correct diagnoses is necessary, the rule given in Chapter III results in higher average margins of correct classification and in lower error rates.

The maximization of the ratios $r_{k\ell}$ was expected to reflect less subspace overlap while large margins $D_{k\ell}$ were expected to correlate with low error rates Pe_k for the paradigms of class k. To test this hypothesis a number of classifications were performed for different

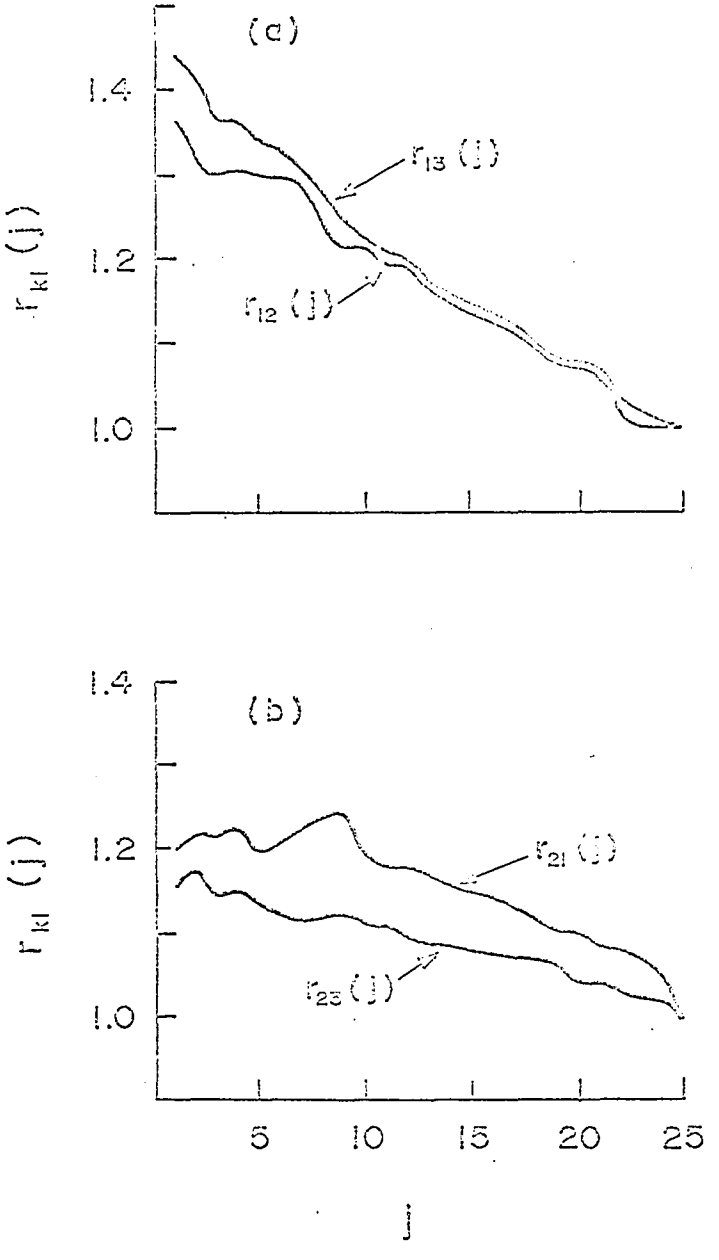

Fig. 14. Ratios of average paradigm inclusions in the class subspaces vs. subspace dimension j. (a) Class 1: hyperthyroid; (b) Class 2: hypothyroid.

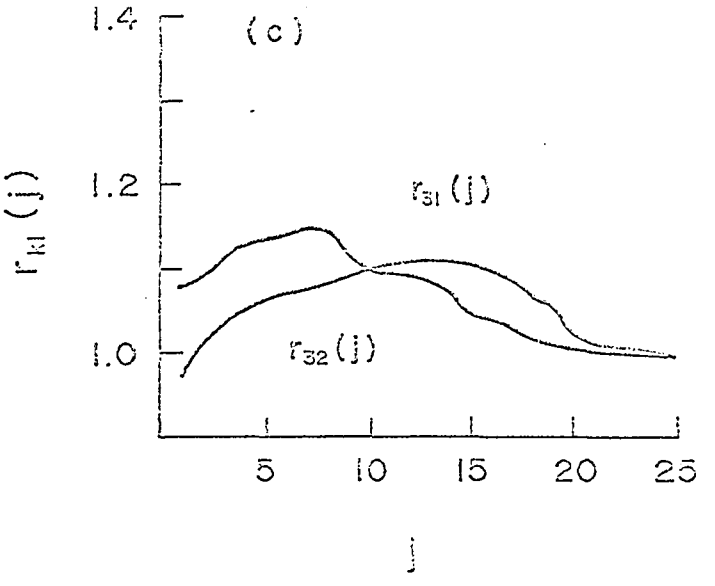

Fig. 14 cont. (c) Class 3: euthyroid I. m = 25.

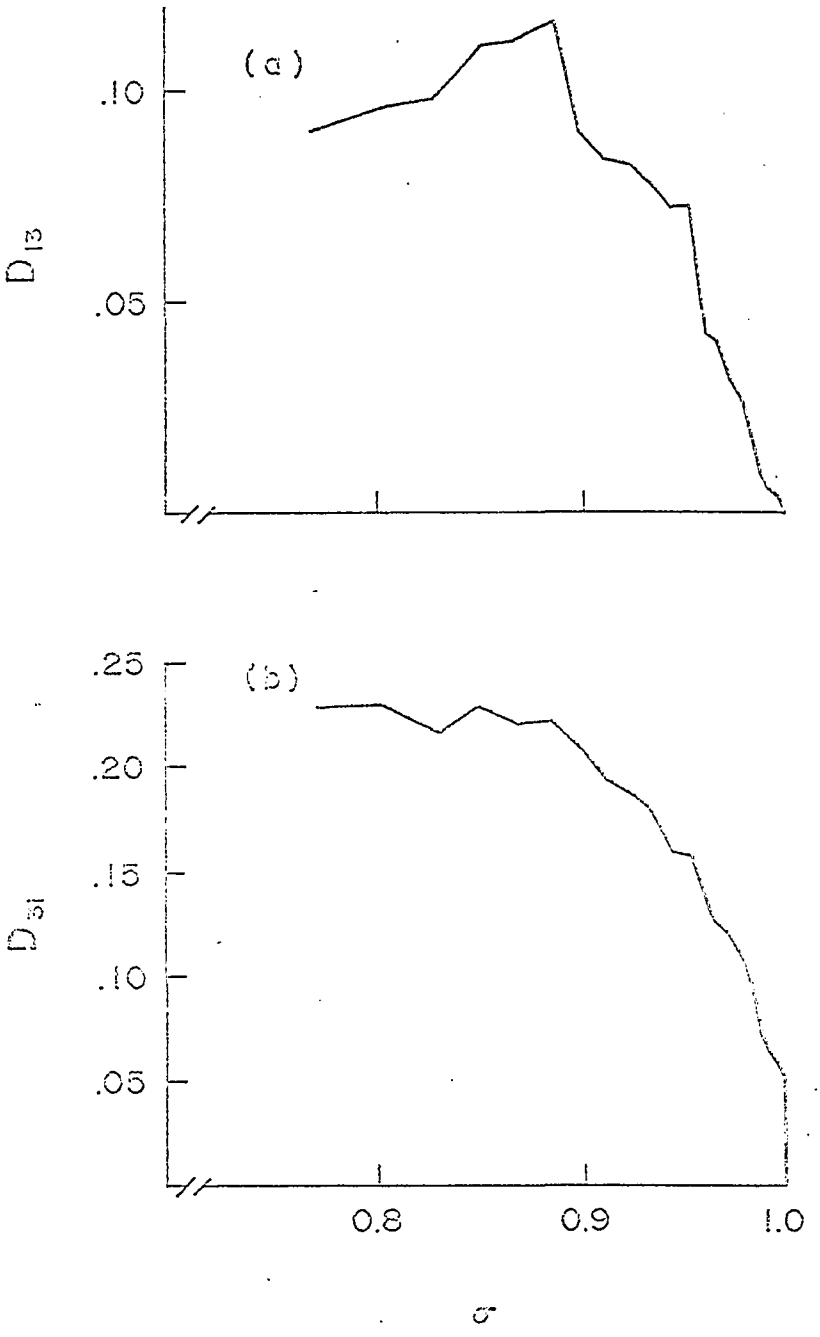

Fig. 15. Average margins of correct classification vs. fidelity σ for discrimination between classes 1 and 3. (a) for paradigms of Class 1; (b) for paradigms of Class 3.

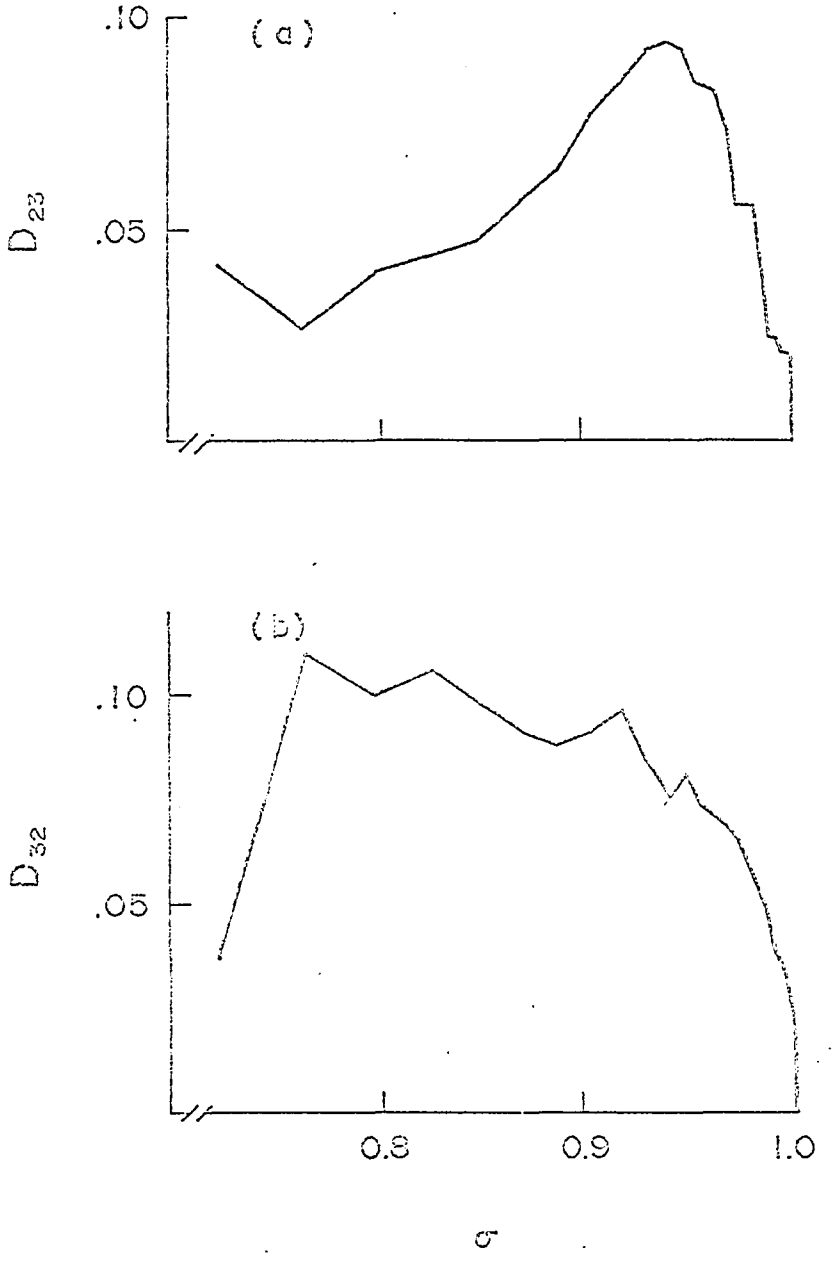

Fig. 16. Average margins of correct classification vs. fidelity σ for discrimination between Classes 2 and 3. (a) for paradigms of Class 2; (b) for paradigms of Class 3.

values of σ chosen to coincide with maxima, minima, and other points of interest on the curves of $D_{k\ell}$ and $r_{k\ell}$ versus σ. The error rates for the paradigms in hyperthyroid (Class 1)/euthyroid (Class 3) discrimination are plotted as functions of σ in Fig. 17. The detailed results of classification are listed in Table 8. Those for hypothyroid (Class 2)/euthyroid discrimination are listed in Table 9 and illustrated in Fig. 18 as functions of σ. Table 10 shows the results for three-way discrimination. It was found that the error rate of paradigms of class C_k was correlated with the ratio $r_{\ell k}$ (see Tables 8, 9, 11, and 12). The maximum of $r_{\ell k}$ coincides or lies close to the minimum of Pe_k. The average error rate for a group of classes correlates with the ratio sum $R_{k\ell} = r_{k\ell} + r_{\ell k}$. This is shown in Fig. 19. Particularly striking is the graph of $R_{123} = \sum_{k=1}^{3} \sum_{j=1}^{3} r_{ij}$ versus $Pe_{123} = \sum_{k=1}^{3} P(k)Pe_k$. The margins $D_{k\ell}$ correlate well with the error rate for hyperthyroid and hypothyroid classes and less well with the euthyroid class, as is shown in Figs. 20 and 21. The characteristics illustrated in Figs. 13 to 21 will now be considered more thoroughly so that the effectiveness of the rules for σ adjustment can be evaluated.

The average inclusions of the paradigms in all subspaces are the basic data used to predict performance. They define the structure of each class and its relationship to the other classes. Much can be learned about the diagnostic categories by analyzing the curves of Fig. 13. It can be seen that $\mu_{1/2}(j)$ and $\mu_{1/3}(j)$ are consistently smaller than $\mu_{2/1}(j)$ and $\mu_{2/3}(j)$ and smaller than $\mu_{3/1}(j)$ and $\mu_{3/2}(j)$, indicating that the paradigms of other classes are much less liable to be confused within the Class 1 subspace than within the Class 2 and

TABLE 8

ERROR RATES FOR PARADIGMS IN
HYPERTHYROID/EUTHYROID DISCRIMINATION
(STAGE 2)

σ_1	σ_3	Pe_1	Pe_3
0.771	0.762	0.097	0.034
0.827	0.822	0.056	0.061
0.849	0.846	0.036	0.065
0.867	0.866	0.031	0.073
0.885	0.883	0.031	0.075
0.912	0.910	0.082	0.065
0.935	0.933	0.061	0.070
0.953	0.949	0.061	0.060
0.968	0.963	0.122	0.070
0.980	0.975	0.133	0.074
0.989	0.985	0.332	0.043

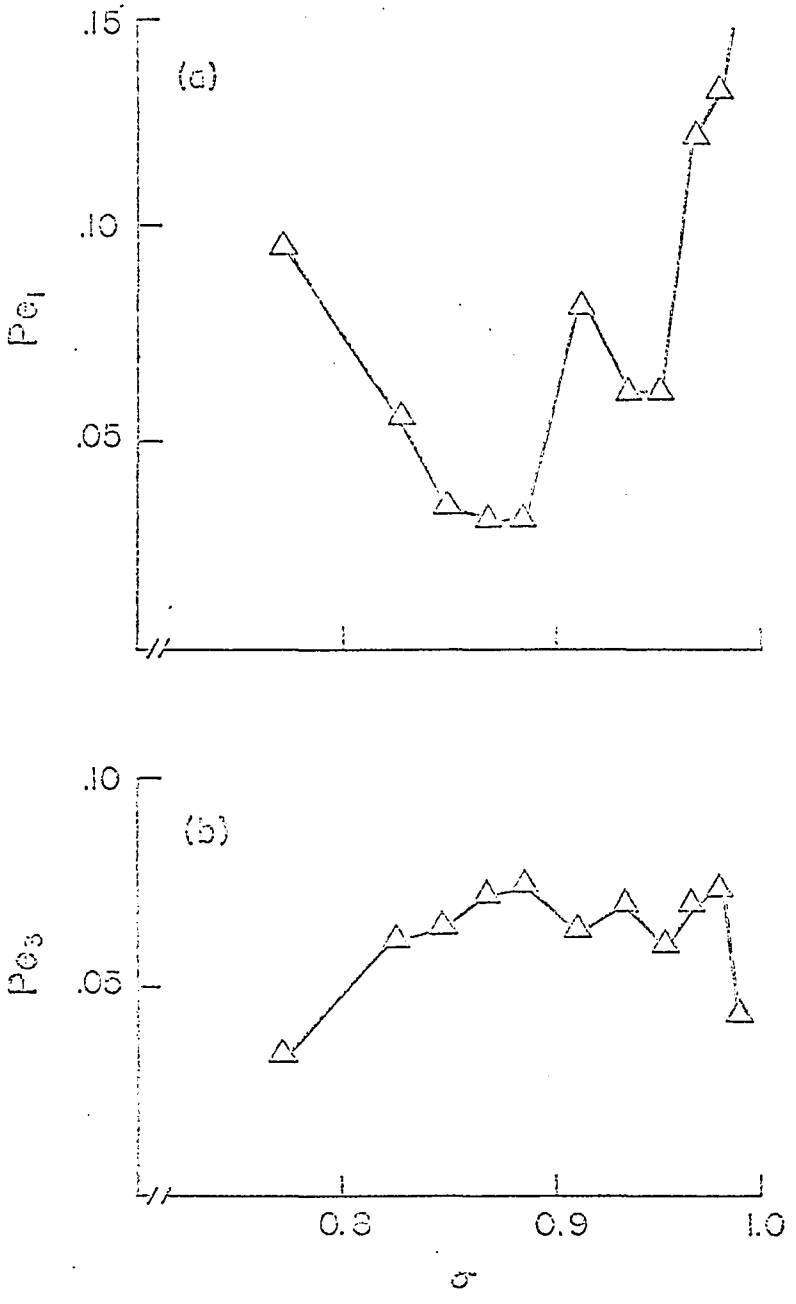

Fig. 17. Error rates for paradigms vs. fidelity σ.
(a) Class 1 paradigms; (b) Class 3 paradigms.

TABLE 9

ERROR RATES FOR PARADIGMS IN
HYPOTHYROID/EUTHYROID DISCRIMINATION
(STAGE 2)

σ_2	σ_3	Pe_2	Pe_3
0.764	0.762	0.414	0.053
0.824	0.822	0.394	0.077
0.849	0.846	0.288	0.084
0.870	0.866	0.274	0.093
0.888	0.883	0.254	0.092
0.904	0.910	0.230	0.098
0.941	0.933	0.149	0.153
0.950	0.949	0.168	0.155
0.958	0.963	0.187	0.173
0.976	0.975	0.226	0.160
0.985	0.985	0.279	0.142

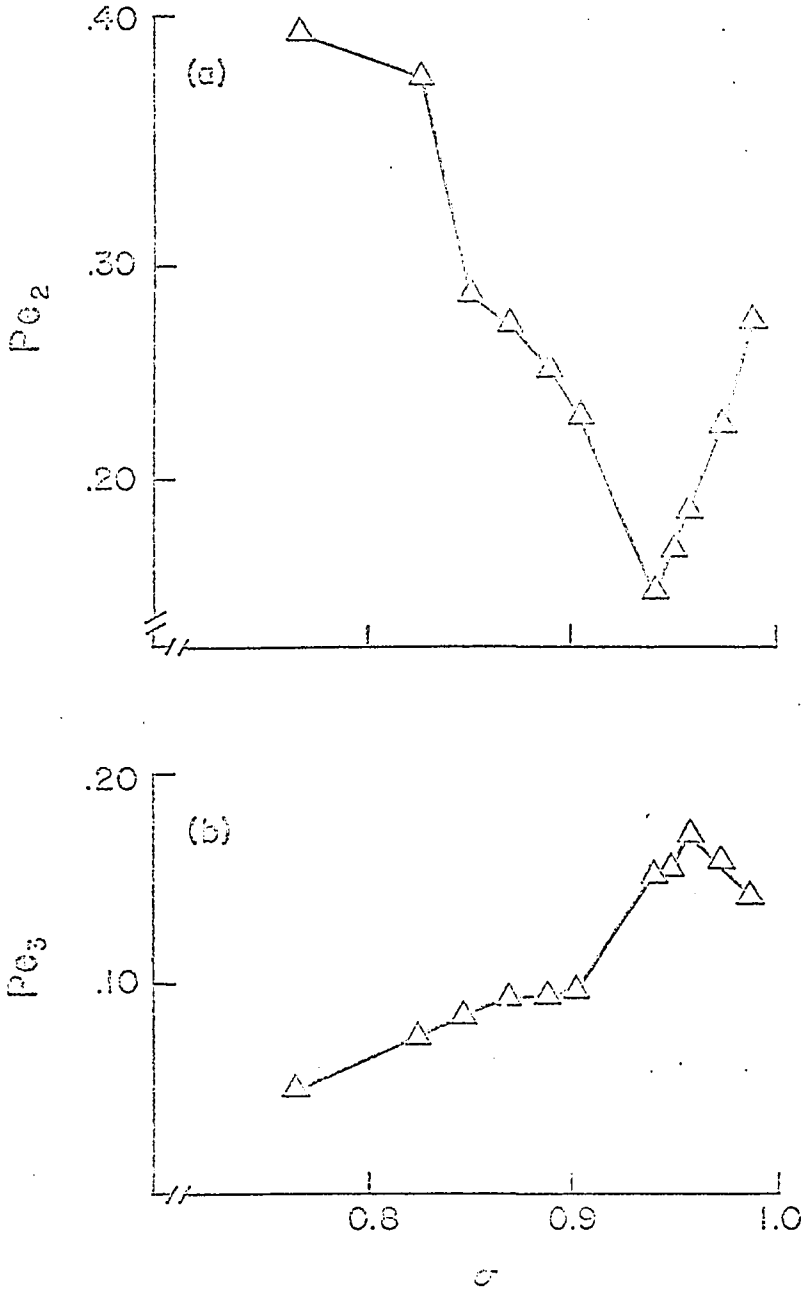

Fig. 18. Error rate for paradigms vs. fidelity σ.
(a) Class 2 paradigms; (b) Class 3 paradigms.

TABLE 10

RECOGNITION OF THYROID DYSFUNCTION AS A FUNCTION OF FIDELITY* (STAGE 2)

σ_3	0.762			0.822			0.846		
Classification	1	2	3	1	2	3	1	2	3
True Category									
1	0.852	0.097	0.051	0.918	0.046	0.036	0.929	0.056	0.015
2	0.029	0.572	0.399	0.034	0.591	0.375	0.029	0.688	0.284
3	0.026	0.050	0.924	0.044	0.072	0.883	0.046	0.078	0.876

*Only the value of σ_3 is given here. The corresponding values of σ_1 and σ_2 can be obtained from Tables 8 and 9.

TABLE 10. (Continued) RECOGNITION OF THYROID DYSFUNCTION AS A FUNCTION OF FIDELITY (STAGE 2)

σ_3	0.866			0.883			0.910		
Classification	1	2	3	1	2	3	1	2	3
True Category									
1	0.944	0.041	0.015	0.949	0.036	0.015	0.893	0.031	0.077
2	0.053	0.692	0.255	0.053	0.712	0.236	0.053	0.740	0.207
3	0.060	0.084	0.856	0.061	0.084	0.855	0.048	0.091	0.861

TABLE 10. (Continued) RECOGNITION OF THYROID DYSFUNCTION AS A FUNCTION OF FIDELITY (STAGE 2)

σ_3	0.933			0.949			0.963		
Classification	1	2	3	1	2	3	1	2	3
True Category									
1	0.903	0.036	0.061	0.918	0.020	0.061	0.867	0.020	0.112
2	0.038	0.817	0.144	0.063	0.779	0.159	0.048	0.774	0.178
3	0.048	0.143	0.808	0.036	0.146	0.819	0.045	0.157	0.798

TABLE 10. (Continued) RECOGNITION OF THYROID DYSFUNCTION AS A FUNCTION OF FIDELITY (STAGE 2)

σ_3	0.975			0.985		
Classification	1	2	3	1	2	3
True Category						
1	0.847	0.026	0.128	0.653	0.020	0.327
2	0.067	0.736	0.197	0.048	0.688	0.264
3	0.051	0.145	0.804	0.030	0.129	0.841

TABLE 11

AVERAGE ERROR RATES* VS. $r_{k\ell}$, $R_{k\ell}$, and $D_{k\ell}$ IN HYPERTHYROID/EUTHYROID DISCRIMINATION (STAGE 2)

σ_1	D_{13}	D_{31}	r_{13}	r_{31}	R_{13}	Pe_{13}
0.771	0.090	0.227	1.441	1.119	2.560	0.043
0.827	0.098	0.218	1.370	1.129	2.499	0.061
0.849	0.110	0.227	1.373	1.144	2.517	0.062
0.867	0.111	0.220	1.344	1.146	2.490	0.068
0.885	0.117	0.221	1.337	1.150	2.487	0.069
0.912	0.083	0.195	1.277	1.098	2.375	0.068
0.935	0.079	0.180	1.241	1.090	2.331	0.069
0.953	0.074	0.157	1.203	1.068	2.271	0.067
0.968	0.040	0.125	1.155	1.038	2.193	0.077
0.980	0.027	0.109	1.132	1.023	2.155	0.082
0.989	0.010	0.075	1.087	1.007	2.094	0.074

*Individual error rates Pe_1 and Pe_3 can be found in Table 8 for the corresponding values of σ_1.

TABLE 12

AVERAGE ERROR RATES* VS. $r_{k\ell}$, $R_{k\ell}$, and $D_{k\ell}$ IN HYPOTHYROID/EUTHYROID DISCRIMINATION (STAGE 2)

σ_2	D_{23}	D_{32}	r_{23}	r_{32}	R_{23}	Pe_{23}
0.764	0.027	0.110	1.172	1.034	2.206	0.103
0.824	0.044	0.107	1.151	1.056	2.207	0.121
0.849	0.048	0.098	1.134	1.056	2.190	0.112
0.870	0.058	0.091	1.123	1.066	2.189	0.119
0.888	0.066	0.088	1.116	1.075	2.191	0.114
0.904	0.061	0.105	1.123	1.093	2.216	0.117
0.941	0.094	0.077	1.099	1.101	2.200	0.153
0.950	0.093	0.082	1.095	1.108	2.203	0.157
0.958	0.078	0.082	1.088	1.095	2.183	0.175
0.976	0.057	0.067	1.074	1.063	2.137	0.170
0.985	0.026	0.058	1.061	1.027	2.088	0.162

*Individual error rates Pe_2 and Pe_3 can be found in Table 9 for the corresponding values of σ_2.

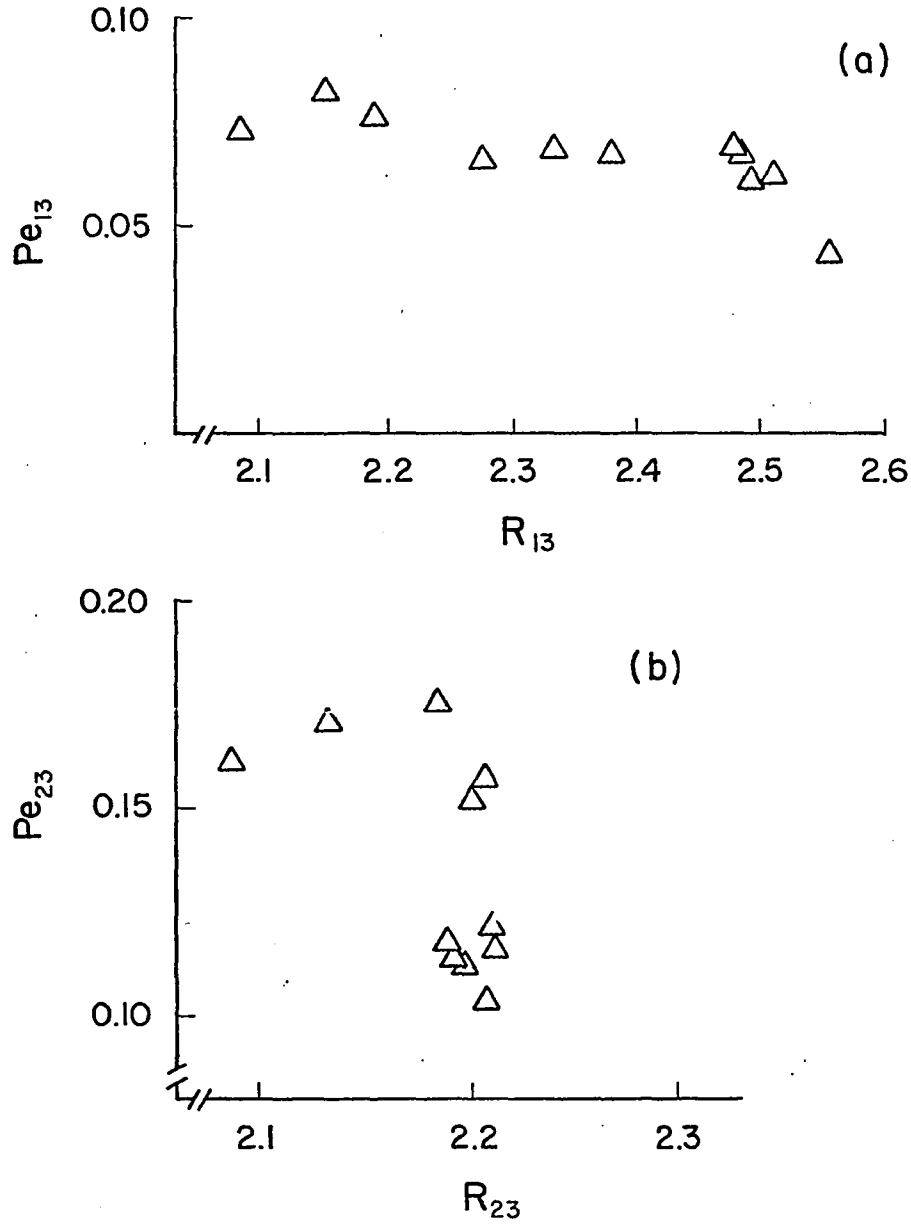

Fig. 19. Average error rate for paradigms vs. the ratio sum $R_{k\ell}$. (a) discrimination of Classes 1 and 3. (b) discrimination of Classes 2 and 3.

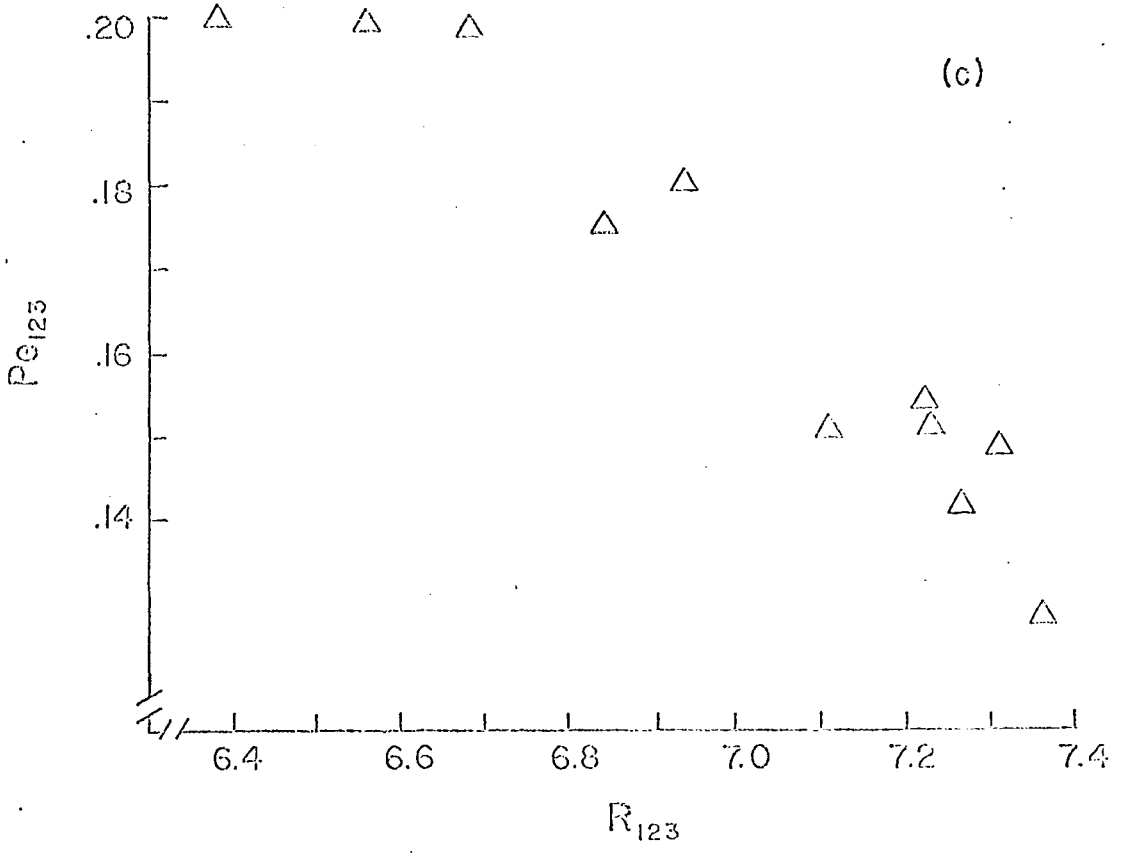

Fig. 19 cont. (c) discrimination of Classes 1, 2, and 3.

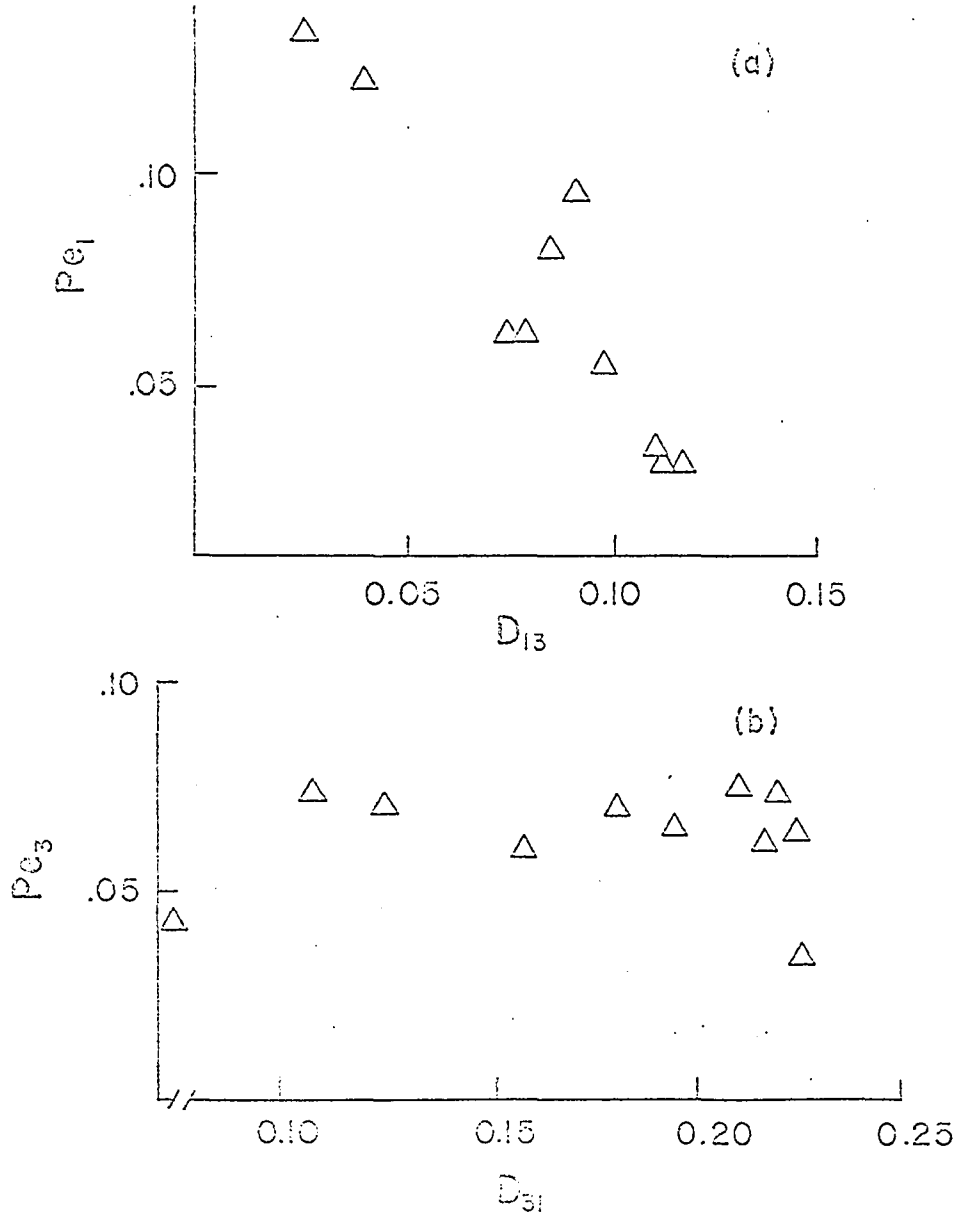

Fig. 20. Error rate for paradigms vs. the average margin of correct classification. (a) Class 1 paradigms; (b) Class 3 paradigms.

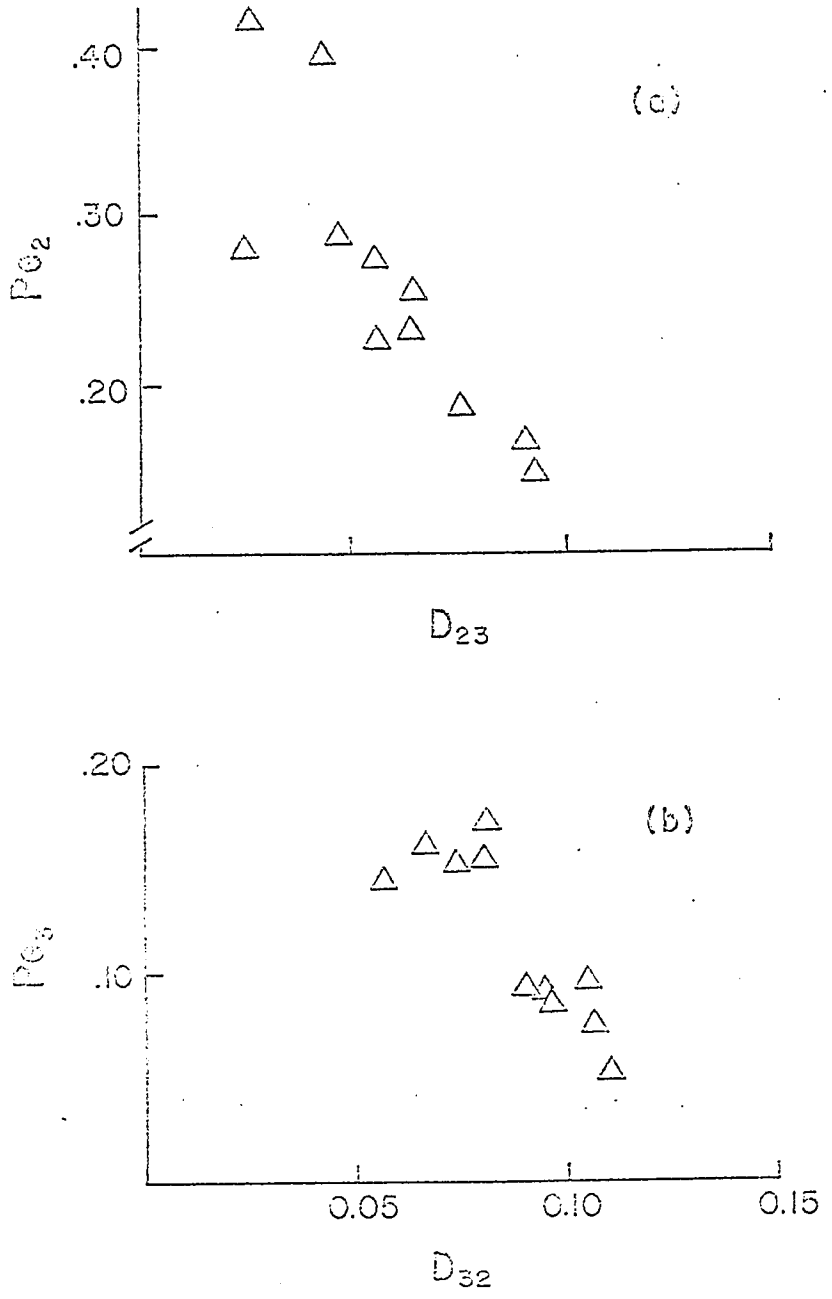

Fig. 21. Error rate for paradigms vs. the average margin of correct classification. (a) Class 2 paradigms; (b) Class 3 paradigms.

Class 3 subspaces. The paradigms of Class 2 are included in the subspace of Class 3 to the extent that for j = 1 (the first component subspace) their average inclusion in the subspace of Class 3 is greater than the inclusion of the Class 3 paradigms themselves. This could be expected in the sense that for all history and physical examination variables the hypothyroid and euthyroid distributions overlap more than those of any other pair of classes. In addition, the euthyroid class is very numerous and nonhomogeneous, including any person without functional thyroid disease, so that a smaller average fidelity of representation could be expected for its paradigms when compared to the paradigms of another, more homogeneous class if the subspaces of both classes are of the same dimensionality. This would occur if the mean square variation of a class were widely spread over the components of the optimal coordinate system, as can be expected for nonhomogeneous data. Such is the case for the Class 3 paradigms whose first eigenvalue is the smallest of the three classes: $\lambda_1^{(3)}$ (=0.665) < $\lambda_1^{(2)}$ (=.723) < $\lambda_1^{(1)}$ (=0.771).

The error-minimizing property of the optimal expansion guarantees that the mean square variation of the paradigms of a class will be greatest in the direction of the principal eigenvector. However, another category may be even better represented by the same principal eigenvector. This occurs in the euthyroid category subspace where it can be observed that $\mu_{3/2}(1)$ (=0.682) > $\mu_{3/3}(1)$ (=0.665). Thus, because they are more homogeneous and less widely spread, the data of Class 2 fit better into the subspace of Class 3 than do the data of Class 3 itself. How this affects discrimination will be observed later.

The ratios $r_{k\ell}$ measure the contrast between the optimal estimation of the paradigms of class C_k in their own subspace and the suboptimal estimation of the paradigms of another class C_ℓ in the subspace of C_k. If this suboptimal estimation is too good it can be expected that the paradigms of C_ℓ will be more confused on the average with those of C_k so that the error rate Pe_ℓ for class C_ℓ ought to be large. This would occur for values of $r_{k\ell}$ close to and less than unity. High values of $r_{k\ell}$, on the contrary, indicate a poor estimation of the paradigms of C_ℓ in S_k and, therefore, the probability of confusing them with the paradigms of C_k is much smaller. A smaller error rate Pe_ℓ can then be expected. The curves for $r_{k\ell}$ differ widely for the various classes. In the subspace of Class 1 the ratios r_{12} and r_{13} are almost monotonically decreasing functions of subspace dimensionality. In the subspace of Class 2 the ratio r_{23} shows a similar trend, but r_{21} presents several peaks with a maximum at $j = 9$. The curves of r_{31} and r_{32} in the subspace of C_3 are relatively smooth and have unique maxima; r_{32} shows that a ratio can assume a value smaller than unity. On the basis of these curves it is expected that in hyperthyroid/euthyroid discrimination performance for the former will be best at the maximum of r_{31} ($m_3^* = 8$) and for the latter at the maximum of r_{13} ($m_1^* = 1$). This is exactly what happened when the paradigms were classified with fidelities corresponding to these values of m_1^* and m_3^* as can be seen in Table 8 for $\sigma_1 = 0.771$ and $\sigma_1 = 0.885$. In this case, as in general, the best performance for both classes is not achieved simultaneously for one value of σ. The procedure for constrained ratio maximization proposed in Chapter III attempts to give a reasonable compromise.

In trying to discriminate between ill and healthy patients it is reasonable to try to guarantee a high percentage of correct classification for the former while minimizing the error rate for the latter. This hypothesis testing approach can be taken when the curves of $r_{k\ell}$ are used as predictors of performance. To obtain a low error rate for hyperthyroid patients a restriction must be placed on the allowable range for r_{31}. This range should be comprised of values close to the maximum of r_{31} which is expected to coincide with the minimum of Pe_1. If 95% of the maximum value of r_{31} is to be the smallest r_{31} tolerated, the corresponding range of m_3^* is $2 \leq m_3^* \leq 11$. This gives a range of $0.725 \leq \sigma \leq 0.922$ for σ, limiting the dimensionality of the subspace S_1 of Class 1 to the range $1 \leq m_1^* \leq 8$. Within this range the maximum of r_{13} corresponds to $m_1^* = 1$, which is also the absolute maximum of r_{13}. Since the classification results are relatively poor for hyperthyroid recognition ($Pe_1 = 0.097$), it might be suspected that the tolerable range of r_{31} was too large. Restricting the minimum allowable value of r_{31} to above 98% of its maximum gives $5 \leq m_3^* \leq 8$, $0.823 \leq \sigma \leq 0.883$, and $3 \leq m_1^* \leq 5$. The maximum of r_{13} in this range yields $m_1^* = 4$ or $\sigma_1 = 0.849$. The corresponding dimension of S_3 is $m_3^* = 6$. From Table 8 it can be seen that this coincides with low error rates for both classes ($Pe_1 = 0.036$ and $Pe_3 = 0.065$), though not with the absolute minimum error for either. The average error rate Pe_{13} for Class 1/3 discrimination is minimized when $\sigma_1 = 0.771$ and $\sigma_3 = 0.762$. This corresponds to the maximum of the sum $R_{13} = r_{13} + r_{31}$ as seen in Fig. 19a. Such a choice of subspaces was also obtained by the maximization of r_{13} with r_{31} constrained to above 95% of its

maximum value. With a 98% level the second best average error was obtained, but the error rate for ill patients was reduced substantially to within limits that are acceptable in clinical practice.

Table 11 lists the average error rate Pe_{13}; the ratios r_{13} and r_{31}; the ratio sum R_{13}; and the margins D_{13} and D_{31} for the tested values of σ_1. The corresponding values of σ_3, Pe_1, and Pe_3 are in Table 8. In Table 12 a similar listing is given for hypothyroid/euthyroid discrimination. It can be seen that the maximum of r_{23} corresponds to the minimum of Pe_3 (see Table 9), while the maximum of r_{32} almost coincides with the minimum of Pe_2. The slight discrepancy is to be expected since the $r_{k\ell}$ are ratios of average quantities which need not be very sensitive to small fluctuations of performance. It can be observed, however, that Pe_2 has a negative correlation with r_{32} as does Pe_3 with r_{23}. The minimum average error rate Pe_{23} also nearly coincides with the maximum of the sum R_{23}. It can be noted that in the recognition of the paradigms of Classes 1 and 2 the negative correlation of the error rates with the ratios becomes less marked as σ increases, and as σ approaches unity it can even reverse. This is particularly true for Class 3 recognition where performance improves for very high values of σ. It is in this range that the fidelity σ_3 becomes comparable to that of the other subspaces for the same dimensionality. The paradigms of C_3 are better included in their own subspace, giving a lower error rate Pe_3. For the range of σ close to unity the subspaces of all classes are of very high dimension, overlapping greatly. The inclusion of a vector in any subspace is very large, and all the average inclusions of the paradigms approach

unity. Hence, for this range of σ the ratios $r_{k\ell}$, which also tend to unity, are not good indicators of performance.

For three-way classification the average error rates calculated from the results of Table 10 are listed in Table 13 together with the sum of ratios R_{123}. It can be seen that the maximum of R_{123} coincides with the minimum average error rate Pe_{123} and vice versa. The best Pe_{123} is not, however, a good indicator of desired performance, since it mainly reflects a good screening rate for healthy patients, while the error rates for ill patients are unacceptably high. An extension of the constrained ratio-maximization procedure to this multiclass situation can be expected to give more reasonable results. The procedure has already been given for discrimination of hyperthyroid/euthyroid categories; it is now necessary to follow the same procedure for hypothyroid recognition so that a compatible set of subspaces with comparable values of σ will result. The ratio r_{32} must be constrained to above a certain level of its maximum to assure good hypothyroid recognition. Choosing a 95% level results in $5 \leq m_3^* \leq 19$ and $0.823 \leq \sigma \leq 0.981$. The allowable range of dimensionality for the subspace S_3 compatible with both hyperthyroid and hypothyroid constraints is $5 \leq m_3^* \leq 9$. This gives $0.822 \leq \sigma \leq 0.898$, $3 \leq m_1^* \leq 6$, and $4 \leq m_2^* \leq 7$. The maxima for r_{13} and r_{23} within these ranges correspond to $m_1^* = 4$ and $m_2^* = 4$. These dimensions do not yield comparable values of fidelity for both Classes 1 and 2. A decision preference must be made to maximize one of the ratios subject to a constraint on the other. Since $r_{13}(3)$ differs from $r_{13}(4)$ by less than 1% the choice of $m_1^* = 3$, $m_2^* = 4$ will give the maximum of r_{23} subject to the

TABLE 13

AVERAGE ERROR RATE Pe_{123} VS. R_{123} IN THREE-WAY DISCRIMINATION (STAGE 2)

σ_3	R_{123}	Pe_{123}
0.762	7.362	0.128
0.822	7.307	0.148
0.846	7.274	0.141
0.866	7.232	0.154
0.883	7.235	0.151
0.910	7.101	0.150
0.933	6.940	0.180
0.949	6.846	0.175
0.963	6.689	0.198
0.975	6.560	0.199
0.985	6.389	0.200

constraint that r_{13} be above 99% of its maximum. The subspace S_3 with $m_3^* = 5$ corresponds to the above choices of S_1 and S_2. This set of subspaces gives satisfactory results for hyperthyroid recognition with only slight degradation of the screening rate. Performance in hypothyroid recognition, however, remains poor. This reflects the very large overlap between the hypothyroid and euthyroid categories and means that the 95% level of the maximum of r_{32} is too low a tolerance for this ratio. If a 98% level is used it results in a range of $9 \leq m_3^* \leq 16$ which overlaps the hyperthyroid constraint range only for $m_3^* = 9$. Although classification results were not obtained for this choice of subspace, results for $m_3^* = 10$ and $m_1^* = m_2^* = 8$ show a distinct improvement for Pe_2 while still keeping the screening rate to a reasonable 0.861. The best improvement in hypothyroid recognition corresponds to $\sigma_2 = 0.941$ and is obtained at the expense of the screening rate which drops to 0.808. The sharp tradeoff effect between hypothyroid and euthyroid recognition can be predicted from the $r_{k\ell}$ curves where the maximum of r_{32} occurs for a high value of $\sigma = 0.949$, whereas the maximum for r_{23} occurs at the much lower value of $\sigma = 0.764$. Furthermore, the minimum of r_{32} almost coincides with the maximum of r_{23}. The best choice of S_3 for hypothyroid recognition also differs considerably from the best choice for hyperthyrcid recognition. This is illustrated by the curves of r_{31} and r_{32} with maxima at $m_3^* = 8$ and $m_3^* = 14$, respectively (Fig. 14). The preceding results can be summarized as follows: 1) the ratios $r_{k\ell}$ are negatively correlated with the error rates Pe_ℓ; 2) the maximum of $r_{k\ell}$ usually coincides with the minimum of Pe_ℓ; 3) the maximum of the ratio sum $R_{k\ell}$

usually coincides with the minimum average error rate $Pe_{k\ell}$; 4) discrepancies from the above results can be expected for values of σ close to unity; 5) the procedure of constrained ratio-maximization gave satisfactory results in both two- and three-way classifications; 6) maximization of the sum of ratios $R_{k\ell m}$ is less effective in three-class recognition than the constrained ratio-maximization technique.

An alternative method of adjusting the fidelity σ is to maximize the average margin of correct classification $D_{k\ell}$ for the paradigms of C_k subject to constraints on the margins for the other classes. From the classification results listed in Tables 11 and 12 the error rates Pe_k can be plotted against the margins $D_{k\ell}$. Figures 20 and 21 show that in all instances the maximum margin $D_{k\ell}$ defines a subspace for which the minimum error rate for the paradigm of C_k is obtained. In the case of hyperthyroid and hypothyroid categories the maximum error rate coincides either exactly or almost with the minimum $D_{k\ell}$. The negative correlation between the trends of D_{13} and Pe_1 as functions of σ is evident in Figs. 15a and 17a. The error rate for hyperthyroids (Pe_1) has a large range (0.031 to 0.332) and is well-predicted by the values of D_{13}. The error rate of euthyroids, on the other hand, has a small range (0.034 to 0.074) and is consistently small. The error rate Pe_3 is not correlated with D_{31} although the minimum of the former does coincide with the maximum of the latter. Disagreement in the trend is particularly noticeable for the high value of $\sigma = 0.989$, where Pe_3 decreases considerably while D_{31} decreases also. As noted in the analysis of ratio trends this range of σ is not adequate to insure a tolerable level of recognition for the ill categories so that

the improvement of Pe_3 is not of practical significance. In hypothyroid/euthyroid discrimination the error rate Pe_3 also shows this trend for high values of σ(Fig. 18b). In this case, however, the range of Pe_3 is much larger (0.053 to 0.173) and the low values of Pe_3 correlate well with higher ones for D_{32}. The margin D_{23} for hypothyroid paradigms is negatively correlated with Pe_2. The range of Pe_2 is the largest for all error rates and has the most pronounced minimum at $\sigma_2 = 0.941$, coinciding with the maximum of D_{23}. Furthermore, the curves of Fig. 18 demonstrate very clearly why it is impossible to obtain best (or even good) results for hypothyroid and euthyroid recognition concurrently. The maximum of Pe_2 coincides with the minimum of Pe_3 and vice versa. The first fact can be predicted from the curves of D_{23} and D_{32} versus σ (Fig. 15), but the effect of decreasing D_{32} at high values of σ prevents prediction of the coincidence of the maximum of Pe_3 with the minimum of Pe_2.

From the curves of Figs. 16 and 17 the procedure of constrained margin-maximization can be seen to yield the choice of subspaces that follows. In hyperthyroid/euthyroid discrimination a constraint level to above 90% of the maximum value of D_{13} leads to $0.849 \leq \sigma \leq 0.885$, within which the maximum D_{31} gives $\sigma_1 = 0.849$. This is the same result that was obtained by the ratio-maximization procedure with a 98% tolerance level. In hypothyroid/euthyroid discrimination a constraint that D_{23} be above 90% of its maximum yields $0.920 \leq \sigma_2 \leq 0.958$. The maximum of D_{32} in this range is at $\sigma_3 = 0.910$ and $\sigma_2 = 0.920$. Although classification results for this threshold were not obtained the results for the next highest D_{32} (at $\sigma_2 = 0.941$) were the

best possible for hypothyroid recognition while reducing Pe_3 from its worst level. This result is also obtained if D_{23} is constrained to above 99% of its maximum. With a greater range for D_{23}, restricted to only above 50% of its peak value, the range for σ_2 becomes $0.849 \leq \sigma_2 \leq 0.985$. The maximum of D_{32} is then at $\sigma_2 = 0.849$, giving a more acceptable screening rate of 0.916 for healthy patients, but increasing the error rate for hypothyroid cases to an unacceptable 0.288. This is also a reflection of the large overlap between the hypothyroid and euthyroid classes. Within the limitations of the data the choice of subspaces given by the procedure is good. In three-way recognition the constraints on the ranges of D_{13} and D_{23} must be relaxed to give a compatible set of subspaces. If D_{13} and D_{23} are restricted to above 80% and 50% of their maxima, respectively, the fidelities fall within the ranges $0.802 \leq \sigma_1 \leq 0.885$ and $0.849 \leq \sigma_2 \leq 0.985$. The permissible range for σ_3 is then $0.846 \leq \sigma_3 \leq 0.883$. The maxima of both D_{31} and D_{32} within this range occur for $\sigma_3 = 0.846$. The corresponding classification results reflect the low tolerance placed on D_{23}, since the error rate for the hypothyroid category is unacceptably high ($Pe_2 = 0.312$). The screening rate, however, remains high, and the error rate for hyperthyroid cases is low ($Pe_1 = 0.071$). If the tolerance on D_{13} is lowered to 70% and if the tolerance on D_{23} is also taken at 70% the allowable range for σ_3 is $0.883 \leq \sigma_3 \leq 0.922$. The maximum of D_{32} within this range corresponds to $\sigma_3 = 0.910$, but that of D_{31} is at $\sigma_3 = 0.883$. To minimize the hypothyroid/euthyroid overlap the former is chosen. The classification results show that a compromise has been reached between the recognition of ill patients and the

screening of healthy ones. Lower error rates Pe_1 and Pe_2 for the ill patients can be obtained at the cost of a lower screening rate for healthy ones by choosing the thresholds $\sigma_1 = 0.935$, $\sigma_2 = 0.941$, and $\sigma_3 = 0.933$. It can be concluded that the constrained margin-maximization procedure chooses subspaces that are discriminating representations of the classes within the limitations of the model.

From the preceding results it can be observed that the levels of constraint for the average margins of correct classification differ from those that give comparable results by the method of constrained ratio-maximization. Thus, a 99% level for r_{31} gives the same choice of subspaces as does the 90% level for D_{13}; the 97% level for r_{32} yields the same results as a 90% level for D_{23}. From these and other similar results it can be seen that the level of constraint on the ratios has to be higher than that on the margins to obtain the same subspaces. This reflects the differences in magnitude and range between the ratios and the margins. A high level of constraint on the margins, however, gives best classification results for hyperthyroid and hypothyroid discrimination: a 99% level of D_{13} and D_{23} gives $\sigma_1 = 0.885$, $\sigma_3 = 0.883$ for Class 1/3 discrimination and $\sigma_2 = 0.941$, $\sigma_3 = 0.933$ for Class 2/3 discrimination. These result in the minimum error rates $Pe_1 = 0.031$ and $Pe_2 = 0.149$. Screening rates, on the other hand, are poor for this choice of subspaces. The lower tolerances on the margins give a better compromise between screening rate and recognition of ill patients. If very low tolerances (30% to 60%) are taken, both methods of constrained maximization yield subspaces that give best screening rates but at the expense of very high error

in classifying ill patients. For this data the most acceptable results in two-way discrimination were obtained with a 99% tolerance level for the ratios $r_{k\ell}$ and a 90% tolerance level for the margins $D_{k\ell}$. In three-way discrimination tolerance levels of about 70% give a compatible set of subspaces with a good compromise between the recognition rates of all categories.

The conclusions reached from the above detailed analyses for a single set of variables have been supported by recognition results at the two other stages of diagnosis. These were the medical history stage (with a total of 19 variables, the first 17 of Table 5 and the first two of Table 7) and the stage of completed laboratory tests (the T_3RCU, 6 and 24 hour I^{131} uptakes, and the variables of the second stage). The results of classifications at different levels of σ are compared with the average margins $D_{k\ell}$ and the ratios $r_{k\ell}$ in Tables 14 and 15 for Class 1/3 and Class 2/3 discrimination, respectively. It can be seen that the minima for the error rates Pe_1 and Pe_2 coincide with the maxima of D_{13}(and r_{31}) and D_{23} (and r_{32}), respectively, for the laboratory test stage. At the first stage (medical history) the best hyperthyroid recognition can be predicted by the maximum of D_{13}, though not by r_{31}. Hypothyroid recognition at this stage agrees only poorly with both D_{23} and r_{32}. While the range of Pe_2 is large the ranges of D_{23} and r_{32} are very small and the actual values of these variables are close to zero and unity. This means that the margins of correct classification are consistently very small and that they cannot be expected to reflect performance with much sensitivity. Euthyroid recognition is also poorly predicted by both D_{31} and r_{13}, as at the

TABLE 14

ERROR RATES VS. $r_{k\ell}$, $R_{k\ell}$ and $D_{k\ell}$ IN HYPERTHYROID/EUTHYROID DISCRIMINATION

σ_1	D_{13}	D_{31}	r_{13}	r_{31}	R_{13}	Pe_1	Pe_3	Pe_{13}
Stage 1								
0.746	0.079	0.072	1.220	1.025	2.245	0.168	0.265	0.252
0.785	0.063	0.093	1.200	1.035	2.235	0.214	0.236	0.233
0.818	0.077	0.064	1.132	1.060	2.192	0.117	0.269	0.249
0.846	0.080	0.068	1.127	1.070	2.197	0.112	0.282	0.260
0.892	0.079	0.079	1.121	1.076	2.197	0.174	0.187	0.185
0.909	0.074	0.071	1.101	1.074	2.175	0.199	0.191	0.192
0.973	0.021	0.043	1.050	1.019	2.069	0.326	0.169	0.190
Stage 3								
0.790	0.158	0.229	1.479	1.208	2.687	0.032	0.056	0.053
0.896	0.170	0.236	1.377	1.220	2.597	0.016	0.079	0.071
0.922	0.126	0.217	1.326	1.146	2.472	0.032	0.053	0.050
0.952	0.098	0.195	1.259	1.114	2.373	0.048	0.046	0.046
0.984	0.049	0.097	1.110	1.050	2.160	0.079	0.051	0.055

TABLE 15

ERROR RATES VS. $r_{k\ell}$, $R_{k\ell}$ and $D_{k\ell}$ IN HYPOTHYROID/EUTHYROID DISCRIMINATION

σ_2	D_{23}	D_{32}	r_{23}	r_{32}	R_{23}	Pe_2	Pe_3	Pe_{23}
Stage 1								
0.744	0.019	0.017	1.116	0.943	2.059	0.284	0.354	0.344
0.788	-0.011	0.065	1.092	0.984	2.076	0.549	0.146	0.203
0.848	0.025	0.031	1.076	0.995	2.071	0.336	0.375	0.370
0.895	0.021	0.041	1.073	1.000	2.073	0.385	0.208	0.233
0.912	0.025	0.056	1.084	1.010	2.094	0.466	0.147	0.192
0.977	0.041	0.021	1.029	1.036	2.065	0.432	0.226	0.255
Stage 3								
0.782	0.034	0.115	1.208	1.019	2.227	0.369	0.073	0.115
0.902	0.071	0.111	1.164	1.066	2.230	0.274	0.091	0.117
0.916	0.076	0.124	1.163	1.086	2.249	0.265	0.096	0.120
0.957	0.095	0.092	1.114	1.103	2.217	0.128	0.160	0.155
0.985	0.041	0.063	1.071	1.041	2.112	0.154	0.131	0.134

second stage, and the same reasons of consistently small error rates (for the third stage), and interference of results for large σ (for the first stage) can be given as explanations. In hypothyroid/euthyroid discrimination, however, the minimum error rate was correctly predicted by the maximum of D_{32} (for stage 1) and by the maximum of r_{23} (for stage 3).

The correlation coefficients between the error rates and the quantities $D_{k\ell}$, $r_{k\ell}$, and $R_{k\ell}$ were calculated to obtain a numerical measure of the observed relationships. These are listed for the three main stages of diagnosis in Table 16 for hyperthyroid/euthyroid discrimination and in Table 17 for hypothyroid/euthyroid discrimination. The most reliable results correspond to the second stage where most classifications were carried out. The best correlations of error rates with their predictors are also obtained at the second stage. With the exception of the correlation of Pe_3 with D_{31} and r_{13} all correlations are statistically significant at the 1% level. The euthyroid error rates are not negatively correlated with either D_{31} or r_{13}, in agreement with the detailed results for the second stage. In hypothyroid discrimination there is a definite negative correlation between Pe_3, D_{32}, and r_{23}. Though not statistically significant at stages 1 and 3 it is quite substantial. The same is true for the correlation of the average error rate Pe_{23} with the sum of the ratios R_{23}. For hyperthyroid discrimination the average error Pe_{13} is significantly correlated with the ratio sum R_{13} at the second stage only. At the other stages the preponderant effect of the euthyroid error (which is not correlated with the ratios) explains the lack

TABLE 16

THE CORRELATION OF ERROR RATES WITH $D_{k\ell}$, $r_{k\ell}$, and $R_{k\ell}$ FOR HYPERTHYROID/EUTHYROID DISCRIMINATION

	Correlation of Pe_1 with		Correlation of Pe_3 with		Correlation of Pe_{13}
	D_{13}	r_{31}	D_{31}	r_{13}	with R_{13}
Stage 1 (5 d.f.)	-0.917**	-0.616	0.262	0.627	0.576
Stage 2 (9 d.f.)	-0.853**	-0.790**	0.081	-0.137	-0.770**
Stage 3 (3 d.f.)	-0.966**	-0.941*	0.469	0.425	0.314

** $P < .01$.
* $P < .05$.

TABLE 17

THE CORRELATION OF ERROR RATES WITH $D_{k\ell}$, $r_{k\ell}$, and $R_{k\ell}$ FOR HYPOTHYROID/EUTHYROID DISCRIMINATION

	Correlation of Pe_2 with		Correlation of Pe_3 with		Correlation of Pe_{23}
	D_{23}	r_{32}	D_{32}	r_{23}	with R_{23}
Stage 1 (4 d.f.)	-0.475	0.454	-0.774	0.212	-0.677
Stage 2 (9 d.f.)	-0.843**	-0.793**	-0.809**	-0.893**	-0.549
Stage 3 (3 d.f.)	-0.486	-0.574	-0.674	-0.823	-0.316

** $P < .01$.

of negative correlation. The consistently best correlation is that between Pe_1 and D_{13}, significant at the 1% level for all stages. Pe_1 is also well-correlated with r_{31}, being significant at the 5% and 1% levels for the third and second stages, respectively. The correlation for the first stage, while not significant, is still quite large. The hypothyroid error rate is also highly correlated with D_{23} and r_{32}, but less significantly than is the hyperthyroid rate with its predictors. In three-class discrimination the correlation between Pe_{123} and R_{123} was calculated for the second stage resulting in a correlation of - 0.956, significant at the 1% level. These correlation results agree well with the previously observed relationships between the error rates and their predictors.

From the preceding results and discussion it is possible to summarize the main advantages and limitations of the fidelity adjustment procedures. These rules are good predictors of performance 1) when the ranges of $r_{\ell k}$ and $D_{k\ell}$ are large with well-pronounced maxima and minima in the range σ away from unity; 2) when the $r_{\ell k}$ and $D_{k\ell}$ are not close to unity or zero, respectively.

There are situations where the rules are poor predictors of performance: 1) when $D_{k\ell}$ and $r_{\ell k}$ are consistently large, outside the range of σ close to unity, which has been shown to coincide with consistently low error rates for Pe_k; 2) when $D_{k\ell}$ and $r_{\ell k}$ are consistently small, corresponding to large error rates for Pe_k, such as occur in the case of hypothyroid recognition.

The performance of the pattern recognition method has been tested thus far by classification results of the paradigms. While the

relationships between the paradigm error rates and the ratios $r_{\ell k}$ and margins $D_{k\ell}$ have been established, it is necessary to show that comparable performance can be obtained on an independent test sample. Data from 1969 were used for this purpose and the results of classification are given in Tables 18 and 19 for five different values of σ. It can be seen that they agree well with the performance of the paradigm cases. In Class 1/3 discrimination the trends for Pe_1 and Pe_3 agree with those for the paradigms. In hypothyroid recognition the trend for screening rate is in agreement with that of the paradigms but the error rate Pe_2 is not. However, because of the very small sample of hypothyroid cases available in 1969 these results are not likely to reflect performance in a larger independent sample. It can be concluded that for all the significant data available, error rates in the independent sample agree with those in the paradigm sample. Hence, the procedures of constrained ratio-maximization and constrained margin-maximization will lead to the choice of good discriminating subspaces for this sample also.

A Comparison of Methods

Results at the three main stages of diagnosis were obtained with the Bayes' conditional probability method and the linear discriminant method. In Tables 20, 21, and 22 these results are listed together with the results of the subspace method for the specified value of fidelity. At the first stage the CLAFIC subspace method gives the best recognition results for ill patients. It does so, however, at the expense of good screening. The worst results for hyperthyroid and hypothyroid recognition are given by the Bayes' method which uses

TABLE 18

ERROR RATES FOR AN INDEPENDENT TEST SAMPLE AS A FUNCTION OF σ FOR HYPERTHYROID/ EUTHYROID DISCRIMINATION

σ_1	σ_3	Pe_1	Pe_3
0.771	0.762	0.135	0.033
0.849	0.846	0.027	0.061
0.885	0.883	0.000	0.094
0.935	0.933	0.108	0.028
0.980	0.975	0.108	0.072

TABLE 19

ERROR RATES FOR AN INDEPENDENT TEST SAMPLE AS A FUNCTION OF σ FOR HYPOTHYROID/ EUTHYROID DISCRIMINATION

σ_2	σ_3	Pe_2	Pe_3
0.764	0.762	0.385	0.028
0.849	0.846	0.231	0.056
0.888	0.883	0.231	0.067
0.941	0.933	0.308	0.144
0.976	0.975	0.231	0.206

TABLE 20

COMPARATIVE DIAGNOSTIC RESULTS (STAGE 1)

Method	Bayes'			Linear Discr.			CLAFIC*		
Classification	1	2	3	1	2	3	1	2	3
True Class									
1	132	5	59	148	16	32	160	8	28
2	11	47	150	11	116	81	24	120	64
3	106	24	1154	146	203	935	231	243	810

* Fidelity $\sigma_3 = 0.875$.

TABLE 21

COMPARATIVE DIAGNOSTIC RESULTS (STAGE 2)

Method	Bayes'			Linear Discr.			CLAFIC*		
Classification	1	2	3	1	2	3	1	2	3
True Class									
1	181	1	14	186	1	9	186	7	3
2	3	188	17	0	186	22	11	148	49
3	35	32	1217	25	39	1220	78	108	1098

* Fidelity $\sigma_3 = 0.883$.

TABLE 22

COMPARATIVE DIAGNOSTIC RESULTS (STAGE 3)

Method	Bayes'			Linear Discr.			CLAFIC*		
Classification	1	2	3	1	2	3	1	2	3
True Class									
1	121	0	5	118	0	8	118	2	6
2	0	101	16	0	98	19	5	98	14
3	6	14	780	9	25	766	22	124	654

* Fidelity $\sigma_3 = 0.950$.

TABLE 23

COMPARATIVE HYPERTHYROID/EUTHYROID DISCRIMINATION

Method	Bayes'		Linear Discr.		CLAFIC***	
Correct recognition for cases of Class:	1	3	1	3	1	3
Stage 1*	135	1178	156	1135	173	938
Stage 2*	182	1249	186	1250	190	1190
Stage 3**	121	794	118	790	122	758

* Total number of cases is 196 for Class 1, 1284 for Class 3.
** Total number of cases is 126 for Class 1, 800 for Class 3.
*** Fidelity σ_1 at Stage 1 is 0.818, at Stage 2 is 0.867, at Stage 3 is 0.922.

a low order approximation of the disease likelihoods. Linear discriminants give intermediate results. At the second stage, which includes the results of the physical examination, the CLAFIC and linear discriminant methods give the best hyperthyroid recognition, but the Bayes' program greatly improves in performance with almost comparable recognition for hyperthyroids and even better recognition for hypothyroids. The best screening rates are again obtained by the discriminant and Bayes' methods, while the subspace method gives too high a rate of false positives. At the third diagnostic stage the Bayes' program gives the best results in all categories of recognition. The CLAFIC and linear discriminant methods give the same recognition results for ill persons, though the former diagnoses more euthyroid cases incorrectly.

Some of the preceding results could be expected because of the nature of the data and the fit of each of the models to it. At the first stage 17 out of 19 variables were binary and the subspace method proved superior in recognition accuracy for ill cases. The poor performance of the Bayes' method could be expected since it does not use the correlations between variables, and these contain much discriminating information at this stage. The linear discriminant and CLAFIC methods, which are based on the second order properties of the data, perform considerably better. The former, optimal for normal data, gives results somewhat worse for hyperthyroid and hypothyroid recognition than does the subspace method which takes advantage of the class subspaces defined by the binary variables. At the second stage of diagnosis four continuous, unimodal variables are included

and the advantage of the discriminant method, designed to handle such data, becomes apparent. The CLAFIC method still performs creditably in hyperthyroid recognition but a high hypothyroid error rate reflects the large overlap of euthyroid and hyperthyroid subspaces as described in the last section. The consistently poorer screening rates obtained by the subspace method are also due largely to the great overlap of the euthyroid subspace with the other subspaces. At stage 3 the Bayes' method performs best. It appears to be least affected by the overlap between the hypothyroid and euthyroid classes, particularly noticeable in the distributions of the I^{131} uptakes (Figs. 11 and 12).

The results of this comparison indicate that the subspace method is a good discriminator for predominantly binary data and can be useful at the early stages of diagnosis. It performs best in the diagnosis of hyperthyroidism as can be seen from the classifications listed in Table 23. For the diagnosis of hypothyroidism the simple Bayes' method performs best at all but the first stage, and is to be preferred over the other, more complicated methods for the final stage of recognition.

Sequential Recognition Results

The subspace methods described in the first section of this chapter were implemented sequentially to simulate the clinical situation in which a final diagnosis is not made until a reliable amount of data for a patient has been accumulated. Subspace fidelities were chosen at each stage of recognition by the method of constrained margin-maximization after the predicted performance had been verified by the results of single-stage classifications. The goal of the sequential programs was to obtain good screening and diagnosis of

healthy patients subject to constraints on tolerable error in the classification of ill patients.

Three stages of recognition were used, coinciding with the principal stages of diagnostic data acquisition: 1) medical history; 2) physical examination and 3) laboratory tests. At the stage of medical history a 60% tolerance level on the values of D_{13} and D_{23} led to the ranges $0.819 \leq \sigma_3 \leq 0.850$ and $0.897 \leq \sigma_3 \leq 0.928$. The maxima of D_{23} and D_{31} within these ranges are at $\sigma_3 = 0.897$ and $\sigma_3 = 0.850$, respectively. The maximum of D_{32} is chosen to minimize hypothyroid/euthyroid overlap, subject to a 90% level of constraint on D_{31}. This results in subspaces defined by $\sigma_1 = 0.909$, $\sigma_2 = 0.912$, and $\sigma_3 = 0.897$. The margin-maximization procedure carried out at the second stage has been described in the first section of this chapter: a 70% constraint level on D_{13} and D_{23} led to $\sigma_1 = 0.912$, $\sigma_2 = 0.904$, and $\sigma_3 = 0.910$. At the third stage, which includes laboratory tests, the same procedure with the same level of constraint on D_{13} and D_{23} gives a range of $0.866 \leq \sigma_3 \leq 0.924$ with the maximum of D_{32} at $\sigma_3 = 0.912$. From the classification results (Tables 14 and 15) it can be seen that the corresponding choice of subspaces ($\sigma_1 = 0.922$ and $\sigma_2 = 0.916$) leads to good hyperthyroid/euthyroid recognition rates, but a too-large error rate. It was found that $\sigma_1 = 0.984$, $\sigma_2 = 0.983$, and $\sigma_3 = 0.981$ led to the best compromise of recognition rates for all three categories in three-way discrimination. Since this is in the range of σ close to unity where margins are small such a result could not have been predicted from the margins.

Two kinds of sequential recognition were carried out. The first

was fully sequential: only those cases falling in a reject region of deferred judgment at every stage were sent on for further recognition. Such a procedure is of interest to see the improvement in classification as stages are added on. It is not acceptable, however, for clinical practice because a diagnosis would rarely, if ever, be made on the basis of medical history findings alone. Thus, the first stage is mainly of experimental interest for the evaluation of the medical history variables. Furthermore, after screening for normal patients is performed at the second stage, all patients not screened out as euthyroid are sent on for laboratory tests, rather than just those who are of undecided diagnosis. This must be done to obtain confirmation of the tentative diagnosis. After laboratory test results are considered a final classification must allow for a no-decision category to include those cases which must be referred to the doctor for treatment/diagnosis. These may be cases which are difficult to recognize by the subspace model or those which can only be detected by suppression or other test results.

The results of classification for the first stage of diagnosis are listed in Table 24. The reject region, denoted as Class 4 in this and all the following tables, is unacceptably large for all categories. Lowering the thresholds that determine the reject region increases error rates excessively. The thresholds shown were chosen empirically to reflect the margin by which classifications into a given class can be considered correct. For example, if the maximum inclusion of a data vector $\underline{x}$ is within subspace S_k it will be classified into this class only if the differences $d_k(\underline{x}) - d_\ell(\underline{x})$ between the inclusion of S_k and

TABLE 24

SEQUENTIAL RECOGNITION OF THYROID DYSFUNCTION (STAGE 1)

Classification	1	2	3	4
True Class				
1	85	2	7	32
2	7	26	7	77
3	89	31	235	455

TABLE 25

PARTLY SEQUENTIAL RECOGNITION OF THYROID DYSFUNCTION

Stage	2				3			
Classification	1	2	3	4	1	2	3	4
True Class								
1	103	2	7	14	107	0	4	8
2	6	71	11	29	4	85	8	9
3	30	49	597	124	11	64	97	31

the inclusions in all other subspaces S_ℓ are greater than a certain positive threshold δ_k. The hyperthyroid class, which fits its subspace well and has little overlap with other subspaces, can be defined with more precision by smaller values of δ_k than can a nonhomogeneous and poorly-fitting class, such as that of the euthyroid cases. The thresholds for stage 1 were $\delta_1 = 0.02$, $\delta_2 = 0.03$, and $\delta_3 = 0.05$. The resulting percentage of euthyroid cases misclassified as hyperthyroid was large even though the rejection rate was very high. These results confirm the wisdom of the usual clinical procedure in which the doctor proceeds to the physical examination before making even tentative diagnostic conclusions.

In Table 25 are listed the results of the partly sequential program in which all patients are sent on to stage 3 except those screened as euthyroid at stage 2. In Table 26 the overall results of this procedure are compared with a single-stage final classification at stage 3. The savings in processing that are obtained by the sequential program can be weighed against the expected increase in error rates. The thresholds at stage 2 were $\delta_1 = \delta_2 = \delta_3 = 0.02$ and at stage 3 were $\delta_1 = \delta_2 = \delta_3 = 0.005$. Only 203 out of a total of 800 euthyroid cases are sent beyond the physical examination stage for laboratory testing (Table 25). This includes those cases falling in the reject region as well as false positives, giving a 0.746 screening rate for true euthyroids at this stage. The true rate of false negatives at stage 2 is 0.055 for Class 1 and 0.094 for Class 2, an average of 0.071. The 0.055 rate for hyperthyroid cases is within the accepted range of error for this study. Though 11 out of 117

TABLE 26

COMPARISON OF THE PARTLY SEQUENTIAL AND SINGLE-STAGE RECOGNITIONS

Method	Partly Sequential				Single Stage (at Stage 3)		
Classification	1	2	3	4	1	2	3
True Class							
1	107	0	11	8	116	1	9
2	4	85	19	9	5	96	16
3	11	64	694	31	30	96	674

TABLE 27

THE PARTLY SEQUENTIAL RECOGNITION OF AN INDEPENDENT TEST SAMPLE

Stage	2				3			
Classification	1	2	3	4	1	2	3	4
True Class								
1	34	0	0	3	34	0	1	2
2	0	8	2	3	1	5	4	1
3	5	9	126	40	10	11	22	11

hypothyroid patients were misclassified these included 8 mild cases for which detection would be difficult under any circumstances.

At the third stage recognition for hyperthyroid cases remained good, giving a final rate of 0.850, with 0.063 left in doubt, and 0.087 false negatives. Recognition of hypothyroids was worse, reflecting the great overlap between this class and the euthyroid category even when laboratory tests are included. The final results gave a 0.762 correct recognition rate for hypothyroid cases, 0.197 for false negatives, and 0.077 for the deferred-judgment category (Table 26). The overall euthyroid recognition rate was 0.867, even better than that obtained by a single-stage classification at the third stage (0.843). The overlap of the euthyroid with the hyperthyroid class is small (0.014). While that of the euthyroid with the hypothyroid class is larger (0.080), it remains smaller than the corresponding overlap in the single-stage program (0.120). A comparison of the two classifications (Table 26) indicates that only the hypothyroid class has a slightly worse error rate in the sequential program than in the single-stage one (up to 0.197 from 0.181). Both hyperthyroid error rates remain about equal and euthyroid recognition is improved by the sequential procedure. This procedure has the added advantage of requiring 615 fewer laboratory tests to be run out of a total of 1043 possible ones--a savings factor of about 59%. Of these 615, 594 correspond to actual savings for truly euthyroid patients and only 19 are false negatives. These results for paradigm recognition are closely paralleled by the results for hyperthyroid and euthyroid classes in the independent test sample. There were not enough new

hypothyroid cases to observe comparable classification trends. The classification of the 1969 data for stage 2 and 3 is given in Table 27.

On the basis of the described results it can be concluded that the sequential diagnostic program is a valuable aid in screening euthyroid cases from a population suspected of thyroid dysfunction. The program is especially good in recognizing hyperthyroid cases without the need of laboratory tests.

Summary

This dissertation dealt with the problem of diagnosing thyroid dysfunction by a pattern recognition method. It was found that the method of class featuring information compression can serve as a useful model of the diagnostic process. In comparison with two other methods it performed well in the recognition of ill patients but gave a lower screening rate for healthy ones. A Bayes' rule method that uses a simple approximation of the joint distribution of symptom patterns in terms of marginal distributions gave the best results at the last stage of diagnosis where laboratory tests were included. At the first stage CLAFIC performed best in the classification of hyperthyroid and hypothyroid cases. A linear discriminant method performed well at the stage of the physical examination, but otherwise gave intermediate results.

The principal result of this work within the pattern recognition field was the development of methods for choosing discriminatory class subspaces. In the original CLAFIC method the choice of a fidelity criterion for defining the class subspaces was arbitrary. No relationship had been established linking the subspace fidelities to

discriminatory performance. In Chapter III several methods were proposed to select subspaces that included discriminatory class features. The most practical results were some procedures developed for the selection of the subspace fidelities. It was found that recognition performance could be related to the average inclusions of the paradigms in the class subspaces. These are easily calculated in terms of the autocorrelation and projection matrices of the classes. The average inclusion of paradigms in their own subspace is the fidelity of the class. Two functions of these quantities were found to be good predictors of discrimination between a pair of classes. One was the ratio of the fidelity of a class to the average inclusion of the vectors of a second class within the subspace of the first. This ratio was found to be significantly correlated with the error rate of the second class. The other predictor was the average margin of correct classification for the paradigms of a class. This was defined as the average of the difference between the inclusion of the paradigms in their own subspace and their inclusion in the subspace of another class. Under certain restrictions it was found that the average margin of correct classification was highly correlated with the error rate of the paradigms. The minimum error rate also coincided with the maximum of the margin. Taking advantage of these relations, two methods were specified by which the fidelities of the class subspaces could be determined to give best paradigm recognition without the need of actually performing any classifications. A constraint was placed on the average margin for one class that it be greater than a certain level. Under this condition the fidelities were chosen so that

the margin for the other class was maximized. If the margins were negatively correlated with the error rates this resulted in subspaces for which the error rate of the second class was minimized, subject to a constraint on the error rate of the first class. This was verified in practice with the thyroid data. Conditions were also found under which the rule did not yield the expected results: whenever the margins were consistently small for all values of fidelity and whenever the range of fidelity was close to unity. When the range of the average margin was large, with distinct maxima and minima, the method of constrained margin-maximization was used to select discriminating subspaces on the basis of the inclusion ratios. The average margins, however, were found to be better predictors of performance in more cases than the ratios. Comparable results were obtained by both methods, although different levels of constraint were used. This reflects the high positive correlation between the inclusion ratios and the margins of classification. The correlation could be indicative of an underlying relationship between the average level of correct estimation of the paradigms and their discriminatory ability which deserves further study.

The margin-maximization procedure led to the choice of discriminatory subspaces at each stage in the diagnosis of thyroid dysfunction. On this basis a sequential recognition program was designed which simulated the clinical application of the pattern recognition method. Thresholds were placed on the margins of classification so that only those cases with the most unequivocal maximum inclusion within a subspace would be definitely recognized as belonging to the corresponding

class. In this way a category of deferred-judgment or rejection was created so that cases falling within it could proceed to further testing. Suspected cases of thyroid dysfunction were also sent on for confirmation. Adjustment of the margin thresholds allowed control of the tradeoff between error and reject rates. The margins were chosen in such a way that the lowest reject rates were obtained subject to minimum tolerable error rates. In practice it was found that substantial savings in the processing of healthy patients could be achieved at the expense of very few false negatives. These results based on paradigm classification were confirmed by testing an independent sample of data.

In conclusion, it seems that the improved method of class featuring information compression has a definite place in clinical practice. Final confirmation of its efficacy must await the on-line implementation of the program.

APPENDIX I

ANNUAL DISTRIBUTION OF THE THYROID DATA

TABLE 28

ANNUAL DISTRIBUTION OF THE THYROID DATA

Year	1963	1964	1965	1966	1967	1968	1969	Total
Category								
Hyperthyroid*	68	70	68	28	48	45	59	386
Hypothyroid*	76	130	140	87	133	74	69	709
Euthyroid I	302	234	284	225	310	261	213	1829
Euthyroid II	238	182	213	186	230	200	224	1473
Total	684	616	705	526	721	580	565	4397

* Includes suspected and confirmed cases.

APPENDIX II

QUESTIONS INCLUDED ON THE ORIGINAL QUESTIONNAIRE
FOR THYROID DYSFUNCTION

Name_____ Project Number_____

Account of_____ Card Number_____

Date_____ Clinic Number_____

Referring Doctor
00) Outside_____
__) Straub Number

History

Age

Sex 0) Male 1) Female

Race
00) Caucasian
01) Japanese
02) Chinese
03) Hawaiian
04) Filipino
05) Negro
06) Japanese-Cauc.
07) Japanese-Haw.
08) Japanese-Other
09) Hawaiian-Cauc.
10) Hawaiian-Chinese
11) Hawaiian-Other
12) Korean
13) Other

Chief Complaint
00) No complaints
01) Fatigue
02) Weight loss
03) Obesity
04) Weight gain
05) Nervousness
06) Heat intolerance
07) Cold intolerance
08) Palpitations
09) Perspiration
10) Dry skin
11) Thyroid lump
12) Thyroid enlargement
13) Thyroid tenderness
14) Exophthalmos
15) Other

Previous Thyroid Surgery
00) None
01) Not known
02) Graves'
03) Physiologic enlargement
04) Enlargement, diffuse, unknown type
05) Enlargement, irregular, unknown type
06) Multinodular, unknown type
07) Multinodular, colloid
08) Single nodule, unknown type
09) Cyst
10) Adenoma
11) Nodule, toxic
12) Hashimoto's, probable
13) Hashimoto's
14) Reidel's
15) Subacute thyroiditis
16) Carcinoma
17) Other

Previous I-131	0) None 1) Hyperthyroidism 2) Toxic nodule 3) Cancer	4) Cardiac failure 5) Angina 6) Other
Antithyroid Drugs	0) None 1) 1 Rx, Hyperthyroidism 2) Pre-cardiac trial	3) With I-131 4) Other
Medications	0) None 1) Thyroid, 1 grain or less 2) Thyroid, 2-3 grains 3) Thyroid, over 3 grains	4) Iodine, in food 5) Iodine, in medication 6) GB or IVP in 6 months 7) Other

Present Weight

Weight Change	0) None 1) Decrease less than 5% 2) Decrease over 5%	3) Increase less than 5% 4) Increase over 5%
Appetite	0) No change 1) Decrease	2) Increase
Rapid Heart	0) None 1) Yes, recent	2) Yes, years
Palpitations	0) None 1) Yes, recent	2) Yes, years
Nervousness	0) None 1) Yes, recent	2) Yes, years
Irritability	0) None 1) Yes, recent	2) Yes, years
Temperature Intolerance	0) None 1) Heat, recent 2) Heat, years	3) Cold, recent 4) Cold, years
Dry Skin	0) None 1) Yes, recent	2) Yes, years
Weakness, Fatigue	0) None 1) Weak, recent 2) Weak, years	3) Fatigue, recent 4) Fatigue, years
Bowel Changes	0) None 1) Diarrhea	2) Constipation

Family History of Thyroid	0) None 1) Hyperthyroidism 2) Hypothyroidism	3) Goiter 4) Other
Family History of Diabetes	0) None	1) Present
Menses	0) No change 1) No menstrual periods 2) Increased flow 3) Decreased flow	4) Other 5) Pregnant 6) Not asked 8) Male
Perspiration (excessive)	0) Normal 1) Yes, Recent	2) Yes, years

Physical Examination

Blood Pressure

Pulse

Tremor	0) None 1) Fine	2) Gross 3) Fine and gross
Skin	0) Normal 1) Dry, normal temperature 2) Dry, cool 3) Dry, warm 4) Moist, normal temperature	5) Moist, cool 6) Moist, warm 7) Other
Lid Retraction	0) None	1) Present
Lid Lag	0) None	1) Present
EOM Weakness	0) None	1) Present
Chemosis Conjunctivitis	0) None	1) Present

Exophthalmic Reading Setting_____ Left Eye_____ Right Eye_____

Thyroid Palpation

Size	0) Normal 1) Not felt 2) Enlarged, slightly	3) Enlarged, moderately 4) Enlarged, greatly
Consistency	0) Normal 1) Soft 2) Firm	3) Hard 4) Cystic 5) Other

Form	0) Normal	3) Single nodule
	1) Smooth	4) Multiple nodules
	2) Irregular	5) Other
Lymph Nodes Neck	0) Normal	1) Abnormal
Reflexes	sec/100_____	
Clinical Impression (Functional)	0) Normal	5) Hyperthyroid, probable
	1) Hypothyroid, mild	6) Hyperthyroid
	2) Hypothyroid, moderate	7) Hypothyroid, possible
	3) Hypothyroid, severe	8) Hypothyroid, probable
	4) Hyperthyroid, possible	
Clinical Impression (Pathological)	00) Normal	08) Single nodule, type unknown
	01) Not known	09) Cyst
	02) Graves'	10) Adenoma
	03) Physiologic enlargement	11) Toxic nodule
	04) Diffuse enlargement, type unknown	12) Hashimoto's, probable
	05) Irregular enlargement, type unknown	13) Hashimoto's
		14) Reidel's struma
	06) Multinodular, type unknown	15) Subacute thyroiditis
		16) Carcinoma
	07) Multinodular, colloid goiter	17) Other

Laboratory

6 hour I^{131} Uptake	Cholesterol
24 hour I^{131} Uptake	PBI
T3RCU	Thyroid Hemagglutination
Adam's Corp.	Thyroid Agglutination
T3 Resin	BMR
Suppression Result (24 hour)	Scan
Stimulation TSH	

Summary & Treatment

Final Diagnoses: Pathological and Functional. See above for code.

Treatment	0) None	3) Surgery
	1) Thyroid or equivalent	4) Antithyroids
	2) Radioiodine	5) Other

APPENDIX III

CLAFIC COMPUTER PROGRAM

Listing of Variables

This is a listing supplementary to the description of Chapter IV. The variables used in this program are listed within each subroutine in the order in which they are required for execution. A variable is relisted only if its meaning changes. The printed copy of a comprehensive version of the CLAFIC program includes some variables which were used in other versions. To avoid confusion these redundant variables are identified by an asterisk and grouped together at the end of each subroutine listing. The comment statements in the program listing are superseded by the descriptions given below.

Main Program

NC Number of classes (diagnostic categories).

M Number of variables.

NGP Number of subsets of variables for normalization.

NPT Total number of paradigms (cases used in the construction of the model).

NU Total number of cases of unknown classification.

SW Option switch to generate artificial variables. If SW = 1 all quadratic combinations of variables are generated. If SW = 0 only original variables are used.

SWTP Option to obtain punched card output of eigenvalues and eigenvectors of all classes, if SWTP = 1. Otherwise, SWTP = 0.

SWKL Option for card input of eigenvalues and eigenvectors if SWKL = 1. In this case subroutine KLEXP which generates them is bypassed. If it is desired to generate the eigenvalues and

eigenvectors SWKL = 0.

SWP*

NGO(I) Number of variables in I-th subset for normalization.

NSS1 If equal to 1, print all data in subroutine INPUT. Otherwise NSS2 = 0.

NSS2 If equal to 1, print all normalized data in subroutine NORM. Otherwise NSS2 = 0.

NSS3 If equal to 1 print all eigenvalues and eigenvectors in subroutine KLEXP. Otherwise NSS3 = 0.

THR Fidelity threshold.

PI(I)*

NPL01 If equal to 1, plot the distributions of all subspace inclusions. Otherwise NPL01 = 0.

NPL02 If equal to 1, plot the distributions of all differences between subspace inclusions. Otherwise NPL02 = 0.

NPL03 If equal to 1, plot the distributions of all ratios of subspace inclusions. Otherwise NPL03 = 0.

SUMNF Index of the first case of the KX-th class.

KX Index for the classes.

NO(KX) = NF:

Number of cases in class KX.

NFU Index of the last case of the KX-th class.

EIG(I) I-th eigenvalue of the KX-th class.

V(J,I) J-th component of the I-th eigenvector of the KX-th class.

VV(I) Single dimension array equivalent to V(J,I).

LM(KX) = LIM:

Dimensionality of the subspace of the KX-th class.

KL*

*Redundant variables: SWP, PI(I), KL.

Subroutine INPUT

FMT Variable for format specification (2 cards maximum allowed).

NO(K) Number of cases in class K.

FV(I) I-th variable of the data vector. I = 37, 38, 39, and 40 are identification codes.

IT Subscript for quadratic variables.

FS(IT) IT-th quadratic variable.

MT Total number of variables including quadratics.

Subroutine NORM

IL(I) Recoding for variables so they can be grouped for normalization.

FX(I) I-th recoded variable of the data vector.

NGL Lower index over which to sum the variables of the data vector for normalization.

NGU Upper index for the same.

IZ(I) Identification variables.

SUM Sum of squares of the variables.

FV(I) I-th normalized variable after NORM.

Subroutine KLEXP

IROW Number of variables (dimension of matrix of eigenvectors).

MM Number of nonredundant terms in the symmetric autocorrelation matrix.

NP, NPM1:

Indeces for eigenvalue and eigenvector printing, specifying

the number of items per row.

A(IJ) Autocorrelation matrix terms.

IJ Index for nonredundant terms of A.

SUM Sum of eigenvalues and eigenvectors to check normalization.

Subroutine COMP

NOXO Total number of pairs of classes.

PW(I,J,K):

Distribution of inclusion of elements of class I in the subspace of class J at inclusion quantization level K (for incorrect classification).

PL(I,J,K):

As above but for correct classifications.

PR(K,KC,I):

Distribution of the KC-th ratio in the subspace of the K-th class at the I-th quantization level.

PF(K,KC,I):

Distribution of the KC-th difference in the K-th subspace at the I-th quantization level.

CM(I,J) Percentage of cases of class I classified as class J.

CN(I,J) Number of cases of class I classified as class J.

KX Index for the classes.

JX Index for the cases.

FZ(I) Identification variables.

ID Index of the class into which a case is classified.

DM Maximum inclusion of a data vector into the class subspaces.

D(K) Inclusion of a data vector into the subspace of class K.

DR(K) K-th ratio of subspace inclusions.

DF(K) K-th difference of subspace inclusions.

KO Index for the distribution of ratios.

I = DS/10. = 1000 * D(K)/10.:

I-th quantization of the subspace inclusion.

J = DRS/10. = 1000 * DR(K)/10.:

J-th quantization for the KO-th ratio.

J = DFS/10. = 1000 * DF(K)/10.:

J-th quantization for the KO-th difference.

LX Index for the class into which a data vector is classified.

ITX Index for the quantization level.

Z The true quantization level for output printing.

IT The total number of correct classifications of cases of class KX distributed at the ITX level of quantization in the subspace of class K.

ITW The total number of incorrect classifications of cases of class KX distributed at the ITX level of quantization in the subspace of class K, plus one case.

The other distribution plots use sets of variables similar to the last four listed above.

Subroutine PROJ

P(K,I,J):

An element in the I-th row, J-th column of the projection matrix of the K-th class.

Subroutine SUBSE

PG(J) Product of data vector FV with the projection matrix P.

NX Index which is equal to 1 if the vector FV is of ambiguous classification (its inclusion in two subspaces is equal).

Subroutine RECOG

FU(I) I-th variable of the data vector FU of unknown classification.

Subroutine EIGEN

R Matrix of eigenvectors; equivalent to matrices VV and V in other subroutines.

For other variables and details of this subroutine see the Scientific Subroutine Package, IBM Publication H20-0205-2, p. 49.

Subroutine ARRAY

This subroutine converts array data from single to double dimension and vice versa. See page 62 of the above referenced IBM manual for details.

Subroutine MLEW

This subroutine calls on EIGEN and ARRAY to calculate the eigenvalues and eigenvectors. Since the eigenvalues are stored in the matrix A, which is destroyed by EIGEN, MLEW assigns them a separate location EIG for transmission to KLEXP. All variables are described in KLEXP or MAIN.

Subroutine TRAN

SUM Sum of eigenvalues.

```
C
C
C REVISED VERSION OF CLAFGEN. SUBSPACE METHOD OF CLAFIC.JUNE 12,1970
C      PROGRAM FOR CALCULATING THE OPTIMAL COORDINATE SYSTEMS OF NC CLASSES
C      FROM A SET OF PARADIGMS ,CHECKING THEIR CLASSIFICATION AND THAT OF
C     NU OTHER UNKNOWN VECTORS.  THIS IS REPEATED FOR NR DATA SETS.
C
      DOUBLE PRECISION FV
      DIMENSION FV(40),FX(40),EIG(40),V(40,40)
      COMMON/S1/THR,LM(7),D(7),DR(7),DF(7)
      COMMON/S2/A(820),VV(1600)
      COMMON/S3/NSS1,NSS2,NSS3
      COMMON/S4/NC,M,NU,NGP,NPT,NC(7),NGO(10)
      COMMON/S6/PI(7),PO(7,7)
      COMMON/S7/SWP
       EQUIVALENCE(V(1),VV(1))
      INTEGER SW,SUMNF,SWTP,SWKL
C
C      VARIABLE PARAMETERS OF THE PROGRAM
C
C      NC = NUMBER OF CLASSES
C      M = DIMENSION OF INPUT VECTORS
C      NGP = NUMBER OF PREDICATE SUBSETS
C      NPT = TOTAL NO. OF PARADIGMS
C      NU = NUMBER OF UNKNOWN VECTORS TO BE CLASSIFIED
C      NGO(I) = NUMBER OF PREDICATES IN SUBSET I
C      NO(I) = NUMBER OF PARADIGMS USED AS INPUT FOR CLASS I
C      THR(I) = I-TH THRESHOLD FOR SELECTION OF MOST SIGNIFICANT EIGENVECTORS
C      SW  = QUADRATIC VECTOR OPTION SWITCH(IF SW =0 LINEAR,IF SW =1 QUADRATIC)
C      SWTP = SWITCH TO MAKE TAPE OF EIGENVECTORS(IF SWTP=1,DO NOT MAKE IF=0)
C      SWKL = OPTION FOR DIRECT TAPE READING OF EIGENVECTORS(IFSWKL=1,READTAPE;
C      IF SWKL = 0,PROCEED WITH KLEXP SUBROUTINE)
C      IF SWTP = 1 OR SWKL = 1,USE EXTRA CONTROL CARD FOR TAPE
C     NSS1 TO NSS3 ARE PRINT OPTION PARAMETERS
C      NSS3 IN KLEXP=EIGENFUNCTIONS OF ALL EIGENVALUES. ALL LATTER PRINTED
C      NSS1 IN INPUT= ALL INPUT AND UNKNOWN VECTORS IN FLOATING POINT
C      NSS2 IN NORM= NORMALIZED INPUT VECTORS BY SUBGROUP AND ENTIRE
C
      READ(5,1) NC,M,NGP,NPT,NU,SW,SWTP,SWKL,SWP
    1 FORMAT(10I5)
C
C      CONSTRAINTS ON PROGRAM PARAMETERS
C     1)M MAXIMUM IS 40 ,BUT IF SW=1(QUADRATICS INCL.) THEN M MAXIMUM IS 5
C     2)NC MAXIMUM IS 3
C      3)NGP MAXIMUM IS 10
C     4)CONSTRAINT FOR NON-DEGENERATE SOLUTION OF EIGENVALUE PROBLEM IS THAT NO.
C       OF SAMPLES BE GREATER THAN NO. OF PREDICATES FOR ALL CLASSES(NO.GT.M)
C     5)THRESHOLDS MUST BE SPECIFIED IN ASCENDING ORDER(THR(I).LT.THR(I+1),ALL I)
C     7)SWTP & SWKL CANNOT BE 1 SIMULTANEOUSLY
C
      PRINT 2,NC,M,NGP,NPT,NU,SW,SWTP,SWKL,SWP
    2 FORMAT(//20X,'PROGRAM PARAMETERS'//20X,'NUMBER OF CLASSES NC ='I5/
     120X,'NUMBER OF VARIABLES M ='I5/20X,'NUMBER OF GROUPS OF VARIABLES
     2 NGP ='I5/20X,'TOTAL NUMBER OF PARADIGMS NPT ='I5/20X,'NUMBER OF U
     3NKNOWN OBJECTS NU ='I5/20X,'QUADRATIC OPTION SW ='I5/20X,'GENERATE
     4K-L EXPANSION SWTP ='I5/20X,'READ K-L EXPANSION SWKL ='I3/20X,'PRO
```

```
     5BABILITIES CALCULATION OPTION SWP='I3)
      READ(5,1)(NGO(I),I=1,NGP)
      PRINT 3,(NGO(I),I=1,NGP)
    3 FORMAT(//20X,'NUMBER OF PREDICATES IN EACH GROUP ='10I5)
      READ(5,4) NSS1,NSS2,NSS3
    4 FORMAT(3I1)
      PRINT 5,NSS1,NSS2,NSS3
    5 FORMAT(//20X,'PRINT OPTION SWITCHES='3I5)
      READ(5,6) THR
    6 FORMAT(F6.3)
      PRINT 7,THR
    7 FORMAT(//20X,'THRESHOLD ='F10.3)
      READ(5,8) (PI(I),I=1,NC)
    8 FORMAT(5F10.5)
      PRINT 9,(PI(I),I=1,NC)
    9 FORMAT(//20X,'PRIOR PROBABILITIES'/20X,'1:HYPERTHYROID ='F7.3/20X,
     1'2:HYPOTHYROID ='F7.3/20X,'3:EUTHYROID ='F7.3)
       IF((SW.EQ.1).AND.(M.GT.15)) GO TO 149
      CALL INPUT(SW)
       SUMNF=1
       REWIND 2
       DO 50 KX=1,NC
       NF=NO(KX)
       NFU=SUMNF+NF-1
      IF(SWKL) 30,30,20
   20 READ(5,25) (EIG(I),I=1,M)
      DO 22 I=1,M
   22 READ(5,25) (V(J,I),J=1,M)
   25 FORMAT(6E13.6,2X)
      GO TO 33
   30 CALL KLEXP(NF,EIG,SUMNF,NFU,SWTP,KX)
   33 CALL TRAN(EIG,LIM,M,THR)
      PRINT 35,KX,LIM,THR
   35 FORMAT(///20X,'CLASS'I3,3X,'HAS LIM='I3,3X,'COMPONENTS IN ITS SUBS
     1PACE FOR THR ='F5.3)
      LM(KX)=LIM
   36 FORMAT(40E14.6)
      WRITE(4,36) (EIG(I),I=1,M)
       DO 48 I=1,M
   48 WRITE(4,36) (V(J,I),J=1,M)
   50  SUMNF= NFU+1
       REWIND 2
       REWIND 4
       WRITE(6,65) NC
   65  FORMAT(///'  OPTIMAL SYSTEM FOR ALL'I3,'  CLASSES OBTAINED')
   95  FORMAT(///'  PARADIGM CLASSIFICATION')
  100  WRITE(6,95)
  110 CALL COMP
       IF(NU) 120,150,120
  120  WRITE(6,140)
  140  FORMAT(1H1,/' UNKNOWN VECTOR RECOGNITION')
  146  CALL RECOG
       GO TO 150
  149  WRITE(6,1149) KL,M
 1149 FORMAT(/' RUN'I3,' DELETED SINCE M='I3,' GT.15')
  150  CONTINUE
```

```
STOP
END
```

```
      SUBROUTINE INPUT(SW)
C
C       SUBROUTINE INPUT TO READ  PARADIGMS AND UNKNOWN VECTORS
C
      DOUBLE PRECISION FV
      COMMON/S3/NSS1,NSS2,NSS3
      DIMENSION FV(40),FMT(40),FS(40)
      COMMON/S4/NC,M,NU,NGP,NPT,NO(7),NGO(10)
       INTEGER SW
       READ(5,4) FMT
    4 FORMAT(20A4/20A4)
      REWIND 4
       WRITE(6,5)
    5 FORMAT(/1H1,'  INPUT DATA'/)
       NCU=NC
C
C       NCU=NC+1 INCLUDES IN INPUT THE "CLASS" OF UNKNOWN VECTORS (NO(NCU)=NU)
C
       IF(NU.NE.0) NCU=NC+1
    8  DO 21 K=1,NCU
       READ(5,10) NO(K)
   10  FORMAT(I5)
       NF=NO(K)
  101  WRITE(6,901) NF,K
  901  FORMAT(I5,'  ELEMENTS IN CLASS'I3)
  109  DO 20 J=1,NF
       READ(5,FMT) (FV(I),I=37,39),(FV(I),I=1,M),FV(40)
1160    MT=M
       IF(SW-1) 14,12,12
   12  IT=0
       DO 1250 I=1,M
       DO 1250 JK=I,M
       IT=IT+1
 1250  FS(IT)=FV(I)*FV(JK)
       MTL=M+1
       MT=IT+M
       DO 1270 I=MTL,MT
       IS=I-M
 1270  FV(I)=FS(IS)
   14 CONTINUE
      IF(NSS1) 19,19,15
   15  WRITE(6,16) K,J
   16  FORMAT(2I4)
       WRITE(6,17) (FV(I),I=1,M)
   17  FORMAT(6X,20F6.2)
   19 CONTINUE
      WRITE(4,18) (FV(I),I=1,MT),(FV(I),I=37,40)
   18 FORMAT(40D14.6)
   20  CONTINUE
   21 CONTINUE
       IF(SW-1) 25,22,22
   22  M=M*(M+3)/2
       IF(MT.EQ.M) WRITE(6,13) MT
   13  FORMAT(/' TOTAL DIMENSIONS INCLUDING SECOND ORDER TERMS IS M='I4)
   25 REWIND 4
      CALL NORM
```

```
RETURN
END
```

```
      SUBROUTINE NORM
C
C      SUBROUTINE NORM TO NORMALIZE ALL INPUT VECTORS
      COMMON/S3/NSS1,NSS2,NSS3
      COMMON/S4/NC,M,NU,NGP,NPT,NO(7),NGO(10)
      DIMENSION FV(40),IZ(4),FX(40),IL(40)
      DOUBLE PRECISION FV,SUM
       READ(5,1) (IL(I),I=1,M)
1      FORMAT(40I2)
       REWIND 3
      IF(NSS2) 15,15,5
    5  WRITE(6,10)
   10  FORMAT(//' NORMALIZED INPUT DATA')
      PRINT 81,M
   81 FORMAT(/' M='I3)
   15  NCU=NC
       IF(NU.NE.0) NCU=NC+1
       DO 50 K=1,NCU
       NOK=NO(K)
       DO 50 J=1,NOK
   17 READ(4,1750) (FV(I),I=1,M),(FV(I),I=37,40)
 1750 FORMAT(40D14.6)
       DO 98 I=1,M
       IN=IL(I)
98     FX(I)=FV(IN)
       DO 99 I=1,M
  99    FV(I)=FX(I)
   19 NGL = 1
       NGW=NGP
      DO 1680 I = 1,4
 1680 IZ(I) = FV(I+36)
       IF(NGP.GT.1) NGW=NGP+1
       NZU=NGO(NGP)
       DO 40 IK=1,NGW
       NZX=NGO(IK)
27     SUM=0.0
       IF(IK-NGW) 25,20,20
   20  NGL=1
       NGU=M
       DO 22 I=NGL,NGU
22     SUM=SUM+FV(I)**2
       GO TO 32
   25  NGU=NGL+NGO(IK)-1
       DO 30 I=NGL,NGU
   30 SUM = SUM + FV(I)**2
32     SUM=DSQRT(SUM)
       IF(SUM) 33,38,33
   33  DO 35 I=NGL,NGU
      FV(I) = FV(I)/SUM
   35 CONTINUE
       IF(NSS2.EQ.0.OR.J.GT. 5) GO TO 38
   37 PRINT 45,(IZ(I),I=1,4),(FV(I),I=1,M)
   45 FORMAT(/2X,2(I5,1X),I2,2X,I3,6X,8(D11.5,2X)/3(1X,10(D11.5,2X)/))
38     NGL=NGU+1
   40    CONTINUE
   49 WRITE(3,1750) (FV(I),I=1,M),(FV(I),I=37,40)
```

```
50      CONTINUE
        REWIND 3
    REWIND 4
        RETURN
        END
```

```
      SUBROUTINE KLEXP(NF,EIG,SUMNF,NFU,SWTP,KX)
C
C  SUBROUTINE KLEXP TO CALCULATE CORRELATION MATRIX AND EXTRACT EIGENVECTORS
C
      DOUBLE PRECISION FV
      DIMENSION FV(40),FX(40),EIG(40),V(40,40)
      COMMON/S2/A(820),VV(1600)
      COMMON/S3/NSS1,NSS2,NSS3
      COMMON/S4/NC,M,NU,NGP,NPT,NC(7),NGO(10)
         EQUIVALENCE(V(1),VV(1))
       INTEGER SUMNF,SUM1
       WRITE(6,10)
10     FORMAT(//'INITIATE SUBROUTINE KLEXP'//)
       IROW=40
         FNUM = NF
         MM = (M*(M+1))/2
         NP = 6
         NPM1 = NP - 1
       SUM1=SUMNF-1
C        THE CORRELATION MATRIX A(I,J)
   25    DO 30  IJ = 1,MM
   30    A(IJ) = 0.0
         IF (SUMNF-1)  35, 35, 33
  33   DO 34 J=1,SUM1
   34 READ (3,18) (FV(I),I=1,M)
   18 FORMAT(4D14.6)
   35    DO 40  K = SUMNF,NFU
      READ(3,18) (FV(L),L=1,M)
         DO 40  J = 1,M
         DO 40 I = 1,J
         IJ = (J*(J-1))/2 + I
         A(IJ) = A(IJ) + FV(J)*FV(I)
   40    CONTINUE
         DO 42  IJ = 1,MM
   42    A(IJ) = A(IJ)/FNUM
         REWIND 3
C
C        CALCULATE EIGENVALUES AND EIGENVECTORS
C
  160  CALL MLEW(M,EIG,IROW)
  163    WRITE(6,5)
         SUM = 0.0
         DO 165  I = 1,M
  165    SUM = SUM + EIG(I)
         DO 45 I = 1,M,NP
         LU = I + NPM1
   45  WRITE(6,58)I,(EIG(L),L=I,LU)
      WRITE(6,59) SUM
      IF(NSS3) 54,54,46
   46    DO 50 I = 1, M
         SUM = 0.0
         DO 48  L = 1,M
   48    SUM = SUM + V(L,I)**2
         WRITE (6,15) EIG(I),I,SUM
         DO 50 J = 1,M,NP
         LU = J + NPM1
```

```
   49 WRITE(6,58) J,(V(L,I),L=J,LU)
   50    CONTINUE
   54     CONTINUE
         REWIND 3
      IF(SWTP) 60,60,55
   55 PUNCH 19,(EIG(I),I=1,M)
       DO 56 I=1,M
   56 PUNCH 19,(V(J,I),J=1,M)
   19 FORMAT(6E13.6,2X)
   60 CONTINUE
C
C       FORMAT STATEMENT 320 LIMITS M
C
  320 FORMAT(/1H0,I3,3X,10E11.4/(6X,10E11.4))
    5    FORMAT (12H EIGENVALUES)
   58    FORMAT(1H0, I4,6E20.8)
   59    FORMAT(////20H SUM OF EIGENVALUES=E15.8)
   15    FORMAT(////33H EIGENFUNCTION FOR THE EIGENVALUE,E16.8,6H
     1   I5/33H SUM OF SQUARES OF EIGENFUNCTION=E16.8)
         RETURN
         END
```

```
      SUBROUTINE COMP
C     SUBROUTINE COMP TO COMPARE PARADIGM CLASSIFICATION WITH TRUE CLAS
C     BY THE SUBSPACE METHOD
C
      DOUBLE PRECISION FV
      COMMON/S1/THR,LM(7),D(7),DR(7),DF(7)
      COMMON/S3/NSS1,NSS2,NSS3
      COMMON/S4/NC,M,NU,NGP,NPT,NO(7),NGO(10)
      COMMON/S5/P(3,40,40)
      COMMON/S8/NPLO1,NPLO2,NPLO3
      DIMENSION FV(40), CN(7,7),CM(7,7),FZ(4)
      INTEGER SUMNF,FZ,CN
      DIMENSION PL(3,3,100),DT(600),DW(400),PF(3,2,200),PR(3,2,200),PW(3
     1,3,100)
      INTEGER PL,PR,PF,PW
      DATA DT/'|',599*'*'/,DW/'|',399*'$'/
      NOXO=NC*(NC-1)/2
      NCN=NC-1
      FNPT=NPT
      NCP = NC + 1
      NCL=0
      SUMNF=1
      DO 5 I=1,NC
      DO 5 J=1,NC
      DO 5 K=1,100
      PW(I,J,K)=0
    5 PL(I,J,K)=0
      DO 6 K=1,NC
      DO 6 KC=1,NCN
      DO 6 I=1,200
      PR(K,KC,I)=0
    6 PF(K,KC,I)=0
C        INITIALIZE CONFUSION MATRIX
      DO 10 I = 1,NCP
      DO 10 J = 1,NCP
      CM(I,J) =0.0
   10 CN(I,J)=0
      CALL PROJ
       WRITE(6,7) (J,J=1,NC)
    7 FORMAT(//3X,'IDENTIFICATION NOS.',2X,'CODE',2X,'CLASS',4X,'NO.',
     14X,'CLFC',6X,'DM',2X,8(4X,I2,2X)/49X,8(4X,I2,2X))
C
C     BEGIN COMPUTATIONS FOR RECOGNITION.
C
         DO 40 KX = 1,NC
         NF = NO(KX)
         NFU = SUMNF + NF - 1
      DO 35 JX = SUMNF, NFU
      READ(3,18) (FV(I),I=1,M),(FV(I),I=37,40)
   18 FORMAT(40D14.6)
      CALL SUBSE(FV,ID,DM)
       DO 25 I=1,4
   25 FZ(I) = FV(I+36)
      PRINT 27,(FZ(I),I=1,4),KX,JX,ID,DM,(D(K),K=1,NC),(DR(K),K=1,NOXO),
     1(DF(K),K=1,NOXO)
   27 FORMAT(5X,I3,1X,2I5,6X,I3,3X,I2,4X,I4,6X,I2,4X,8F8.4/53X,8F8.4)
```

```
      CN(KX,ID)=CN(KX,ID)+1
      IF(NPLO1.NE.1) GO TO 32
      KO=0
      DO 30 K=1,NC
      DS=1000.*D(K)
      I=DS/10.
      IF(ID.NE.KX) PW(KX,K,I)=PW(KX,K,I)+1
      NC1X=NC+1-KX
      IF(K.EQ.NC1X) GO TO 30
      KO=KO+1
      IF(NPLO2.NE.1) GO TO 29
      DRS=DR(K)*1000.
      J=DRS/10.
      IF(J.GT.200) J=200
      PR(KX,KO,J)=PR(KX,KO,J)+1
   29 IF(NPLO3.NE.1) GO TO 30
      DFS=DF(K)*1000.
      J=DFS/10.+100
      PF(KX,KO,J)=PF(KX,KO,J)+1
   30 PL(KX,K,I)=PL(KX,K,I)+1
   32 IF(ID.EQ.(NC+1))NCL=NC+1
   35 CONTINUE
      FNUM = NF
      DO 38 LX=1,NCP
   38 CM(KX,LX)=100.*CN(KX,LX)/FNUM
   40  SUMNF=NFU+1
      PRINT 42,THR
   42 FORMAT(///20X,'CLASSIFICATION MATRIX, THR='F6.3)
      PRINT 43
   43 FORMAT(//12X,'CLASS'17X,'CLASSIFICATION'//18X,'1:HYPERTHYROID'3X,'
     12:HYPOTHYROID'6X,'3:EUTHYROID')
       IF(NCL.NE.(NC+1)) NCL=NC
      DO 45 I=1,NC
   44 FORMAT(/12X,I3,3X,6(F5.1,'('I4,')'5X))
   45 PRINT 44,I,(CM(I,J),CN(I,J),J=1,NCL)
       REWIND 3
   50  CONTINUE
C
C     DISTRIBUTIONS GENERATION
C
      DO 60 KX=1,NC
      DO 60 K=1,NC
      PRINT 53,KX,K
   53 FORMAT(1H1,/20X,'DISTRIBUTION OF CLASS 'I3,'  IN THE SUBSPACE OF C
     1LASS'I3)
      DO 60 I=1,60
      ITX=100-I
      R=I
      Z=1.0-R/100.0
      IT=PL(KX,K,ITX)
      IT1=IT+1
      ITW=PW(KX,K,ITX)+1
   54 FORMAT(3X,F4.2,2X,I4,1X,110A1/4(15X,90A1/))
      IF(ITW-1) 55,55,56
   55 PRINT 54,Z,IT,(DT(J),J=1,IT1)
      GO TO 60
```

```
56 IF(IT1.EQ.ITW) GO TO 57
   ITZ=ITW+1
   PRINT 54,Z,IT,(DW(J),J=1,ITW),(DT(J),J=ITZ,IT1)
   GO TO 60
57 PRINT 54,Z,IT,(DW(J),J=1,ITW)
60 CONTINUE
   IF(NPLO3.NE.1) GO TO 70
   DO 66 KX=1,NC
   DO 66 K=1,NCN
   PRINT 62,K,KX
62 FORMAT(1H1,/20X,'DISTRIBUTION OF DIFFERENCE 'I3,' FOR CLASS 'I3)
   DO 66 I=1,200
   ITX=201-I
   IF(ITX.LT.160) GO TO 63
   PF(KX,K,160)=PF(KX,K,160)+PF(KX,K,ITX)
   GO TO 66
63 IF(ITX.GT.40) GO TO 64
   PF(KX,K,40)=PF(KX,K,40)+PF(KX,K,ITX)
   IF(ITX.NE.1) GO TO 66
64 R=ITX-100
   Z=R/100.0
   IT=PF(KX,K,ITX)
   IT1=IT+1
65 PRINT 54,Z,IT,(DT(J),J=1,IT1)
66 CONTINUE
70 IF(NPLO2.NE.1) GO TO 80
   DO 76 KX=1,NC
   DO 76 K=1,NCN
   PRINT 72,K,KX
72 FORMAT(1H1,/20X,'DISTRIBUTION OF RATIO 'I3,' FOR CLASS 'I3)
   DO 76 I=1,200
   ITX=201-I
   IF(ITX.LT.160) GO TO 73
   PR(KX,K,160)=PR(KX,K,160)+PR(KX,K,ITX)
   GO TO 76
73 IF(ITX.GT.40) GO TO 74
   PR(KX,K,40)=PR(KX,K,40)+PR(KX,K,ITX)
   IF(ITX.NE.1) GO TO 76
74 R=ITX
   Z=R/100.0
   IT=PR(KX,K,ITX)
   IT1=IT+1
75 PRINT 54,Z,IT,(DT(J),J=1,IT1)
76 CONTINUE
80 CONTINUE
   RETURN
    END
```

```
      SUBROUTINE PROJ
C
C      SUBROUTINE PROJ TO CALCULATE THE PROJECTION MATRICES FOR ALL CLASSES
C
      COMMON/S1/THR,LM(7),C(7),CR(7),DF(7)
      COMMON/S2/A(820),VV(1600)
      COMMON/S4/NC,M,NU,NGP,NPT,NO(7),NGO(10)
      COMMON/S5/P(3,40,40)
      DIMENSION EIG(40),V(40,40)
      EQUIVALENCE(V(1),VV(1))
      DO 10 K=1,NC
      LIM=LM(K)
    1 FORMAT(40E14.6)
      READ(4,1) (EIG(I),I=1,M)
       DO 3 I=1,M
    3  READ(4,1) (V(J,I),J=1,M)
      DO 10 I = 1,M
      DO 10 J = 1,M
      P(K,I,J) = 0.0
      DO 10 L=1,LIM
   10 P(K,I,J) = P(K,I,J) + V(I,L)*V(J,L)
      REWIND 4
      DO 20 K = 1,NC
      WRITE(6,15) K
   15 FORMAT(//' PROJECTION OPERATORS FOR CLASS'I3)
      DO 20 J = 1,M
   20 WRITE(6,25) (P(K,I,J),I=1,M)
   25 FORMAT(10(2X,E11.5))
      RETURN
      END
```

```
      SUBROUTINE SUBSE(FV,ID,DM)
C
C       SUBROUTINE SUBSE TO CLASSIFY VECTORS BY THE SUBSPACE DIFFERENCE M
C
      DOUBLE PRECISION FV
      COMMON/S1/THR,LM(7),D(7),DR(7),DF(7)
      COMMON/S2/A(820),VV(1600)
      COMMON/S4/NC,M,NU,NGP,NPT,NO(7),NGO(10)
      COMMON/S5/P(3,40,40)
      DIMENSION EIG(40),FV(40),FX(40),V(40,40),PG(40)
       EQUIVALENCE (V(1),VV(1))
       DO 20 K=1,NC
      D(K)=0.0
       DO 15 J=1,M
       PG(J)=0.0
       DO 15 L=1,M
   15  PG(J)=PG(J)+P(K,J,L)*FV(L)
       DO 16 L=1,M
   16 D(K)=D(K)+FV(L)*PG(L)
  122  IF(K-1) 13,13,14
   13 DM=D(K)
      ID=1
      GO TO 20
   14 DM=AMAX1(D(K),DM)
      IF(DM.EQ.D(K)) ID=K
   20  CONTINUE
      DR(1)=D(1)/D(2)
      DF(1)=D(1)-D(2)
      IF(NC.EQ.2) GO TO 22
      DR(2)=D(1)/D(3)
      DF(2)=D(1)-D(3)
      DR(3)=D(2)/D(3)
      DF(3)=D(2)-D(3)
   22 NX=0
       DO 25 K=1,NC
      IF(DM.EQ.D(K)) NX=NX+1
      IF(NX.GT.1) ID=NC+1
   25  CONTINUE
       RETURN
      END
```

```
      SUBROUTINE RECOG
C
C     SUBROUTINE RECOG TO CLASSSIFY UNKNOWN VECTORS BY THE SUBSPACE METHOD
C
      DOUBLE PRECISION FU
      COMMON/S1/THR,LM(7),D(7),DR(7),DF(7)
      COMMON/S4/NC,M,NU,NGP,NPT,NO(7),NGO(10)
      COMMON/S5/P(3,40,40)
      DIMENSION FU(40),FZ(4)
      INTEGER FZ
      DO 3 KU=1,NPT
    3 READ(3,18) (FU(I),I=1,M),(FU(I),I=37,40)
   18 FORMAT(40D14.6)
      WRITE(6,10) (J,J=1,NC)
   10 FORMAT(//3X,'IDENTIFICATION NOS.',2X,'CODE',3X,'NO.',
     12X,'CLFC',5X,'DM',2X,8(4X,I2,2X)/49X,8(4X,I2,2X))
      DO 100 KU=1,NU
      READ(3,18) (FU(I),I=1,M),(FU(I),I=37,40)
      CALL SUBSE(FU,ID,DM)
      DO 19 I = 1,4
   19 FZ(I) = FU(I+ 36)
   20 PRINT 25,(FZ(I),I=1,4),KU,ID,DM,(D(I),I=1,NC),(DR(I),I=1,NC),(DF(I
     1),I=1,NC)
   25 FORMAT(5X,2(I5,1X),I2,6X,I3,3X,2(I4,3X),6F8.4/48X,8F8.4)
  100 CONTINUE
      REWIND 3
      RETURN
      END
```

```
      SUBROUTINE EIGEN(N,MV)
      DIMENSION A(820),R(1600)
      COMMON/S2/ A, R
    5 RANGE=1.0E-6
      IF(MV - 1)  10, 25, 10
   10 IQ=-N
      DO 20 J=1,N
      IQ=IQ+N
      DO 20 I=1,N
      IJ=IQ+I
      R(IJ)=0.0
      IF(I-J) 20,15,20
   15 R(IJ)=1.0
   20 CONTINUE
C
C        COMPUTE INITIAL AND FINAL NORMS (ANORM AND ANORMX)
C
   25 ANORM=0.0
      DO 35 I=1,N
      DO 35 J=I,N
      IF(I-J) 30,35,30
   30 IA=I+(J*J-J)/2
      ANORM=ANORM+A(IA)*A(IA)
   35 CONTINUE
      IF(ANORM) 165,165,40
   40 ANORM=1.414*SQRT(ANORM)
      ANRMX=ANORM*RANGE/FLOAT(N)
C
C        INITIALIZE INDICATORS AND COMPUTE THRESHOLD, THR
C
      IND=0
      THR=ANORM
   45 THR=THR/FLOAT(N)
   50 L=1
   55 M=L+1
C
C        COMPUTE SIN AND COS
C
   60 MQ=(M*M-M)/2
      LQ=(L*L-L)/2
      LM=L+MQ
   62 IF( ABS(A(LM))-THR) 130,65,65
   65 IND=1
      LL=L+LQ
      MM=M+MQ
      X=0.5*(A(LL)-A(MM))
   68 Y=-A(LM)/ SQRT(A(LM)*A(LM)+X*X)
      IF(X) 70,75,75
   70 Y=-Y
   75 SINX=Y/ SQRT(2.0*(1.0+( SQRT(1.0-Y*Y))))
      SINX2=SINX*SINX
   78 COSX= SQRT(1.0-SINX2)
      COSX2=COSX*COSX
      SINCS =SINX*COSX
C
C        ROTATE L AND M COLUMNS
```

```
C
      ILQ=N*(L-1)
      IMQ=N*(M-1)
      DO 125 I=1,N
      IQ=(I*I-I)/2
      IF(I-L) 80,115,80
   80 IF(I-M) 85,115,90
   85 IM=I+MQ
      GO TO 95
   90 IM=M+IQ
   95 IF(I-L) 100,105,105
  100 IL=I+LQ
      GO TO 110
  105 IL=L+IQ
  110 X=A(IL)*COSX-A(IM)*SINX
      A(IM)=A(IL)*SINX+A(IM)*COSX
      A(IL)=X
  115 IF(MV-1) 120,125,120
  120 ILR=ILQ+I
      IMR=IMQ+I
      X=R(ILR)*COSX-R(IMR)*SINX
      R(IMR)=R(ILR)*SINX+R(IMR)*COSX
      R(ILR)=X
  125 CONTINUE
      X=2.0*A(LM)*SINCS
      Y=A(LL)*COSX2+A(MM)*SINX2-X
      X=A(LL)*SINX2+A(MM)*COSX2+X
      A(LM)=(A(LL)-A(MM))*SINCS+A(LM)*(COSX2-SINX2)
      A(LL)=Y
      A(MM)=X
C
C        TESTS FOR COMPLETION
C
C        TEST FOR M = LAST COLUMN
C
  130 IF(M-N) 135,140,135
  135 M=M+1
      GO TO 60
C
C        TEST FOR L = SECOND FROM LAST COLUMN
C
  140 IF(L-(N-1)) 145,150,145
  145 L=L+1
      GO TO 55
  150 IF(IND-1) 160,155,160
  155 IND=0
      GO TO 50
C
C        COMPARE THRESHOLD WITH FINAL NORM
C
  160 IF(THR-ANRMX) 165,165,45
C
C        SORT EIGENVALUES AND EIGENVECTORS
C
  165 IQ=-N
      DO 185 I=1,N
```

```
      IQ=IQ+N
      LL=I+(I*I-I)/2
      JQ=N*(I-2)
      DO 185 J=I,N
      JQ=JQ+N
      MM=J+(J*J-J)/2
      IF(A(LL)-A(MM)) 170,185,185
  170 X=A(LL)
      A(LL)=A(MM)
      A(MM)=X
      IF(MV-1) 175,185,175
  175 DO 180 K=1,N
      ILR=IQ+K
      IMR=JQ+K
      X=R(ILR)
      R(ILR)=R(IMR)
  180 R(IMR)=X
  185 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE ARRAY(MODE,I,J,N,M)
C     SUBROUTINE ARRAY
      DIMENSION S(1), D(1)
     COMMON/S2/A(820),S
      EQUIVALENCE (S(1), D(1))
      NI = N - I
     IF(MODE-1) 100,100,120
  100 IJ=I*J+1
      NM=N*J+1
      DO 110 K=1,J
      NM=NM-NI
      DO 110 L=1,I
      IJ=IJ-1
      NM=NM-1
  110 D(NM)=S(IJ)
      GO TO 140
C
C     CONVERT FROM DOUBLE TO SINGLE DIMENSION
C
  120 IJ=0
      NM=0
      DO 130 K=1,J
      DO 125 L=1,I
      IJ=IJ+1
      NM=NM+1
  125 S(IJ)=D(NM)
  130 NM=NM+NI
C
  140  RETURN
       END
```

```
C       SUBROUTINE MLEW
      SUBROUTINE MLEW(M,EIG,IROW)
   20 FORMAT(E10.4)
      DIMENSION A(820),EIG(40),V(40,40),VV(1600)
        COMMON/S2/A,VV
        EQUIVALENCE (V(1),VV(1))
        CALL EIGEN(M,0)
      CALL ARRAY(1,M,M,IROW,IROW)
        DO 10 J = 1,M
        JJ = (J*(J+ 1)) /2
        A(J) = A(JJ)
        EIG(J) = A(JJ)
   10 CONTINUE
        RETURN
        END
```

```
C         SUBROUTINE TRAN
        SUBROUTINE TRAN(EIG,K,M,THR)
       DIMENSION EIG(40)
          SUM = 0.0
          K = 0
          DO 10 I = 1,M
          K = K + 1
          SUM = SUM + EIG(I)
          IF(SUM-THR) 10,30,30
   10     CONTINUE
   30     RETURN
          END
```

APPENDIX IV

BAYES' CONDITIONAL PROBABILITY COMPUTER PROGRAM

Description

This program calculates the posterior probabilities of diseases on the basis of the symptom conditional probabilities and prior probabilities. The printed version of this program, called POCOD, accepts punched card input. The first part of the program consists of the input of parameters and conditional probability distributions. The variables may be binary, qualitative, or continuous. In the last two cases they are quantized and the range and interval of quantization must be specified. There is a provision for reordering the data if only subsets of variables are needed in a given classification. There is also a provision for combining two binary variables which are related. Thus, two questions, such as "Is the skin warm and moist?" and "Is the skin cool and dry?", can be combined into a single question with three possible answers.

The data on patients is read in according to the specifications on the variable format card(s). The identification number, classification, posterior probabilities, and coded variables for each case are printed out and the results of classification are summarized in a table at the end of the program.

Listing of Variables

NC The number of classes in the model.

M The number of binary variables.

MN The number of qualitative and continuous variables to be quantized.

NU Number of classes to be classified in addition to NC.

MRM The total number of variables after recodings and combinations.

NCP The total number of classes to be classified.

A(I) The prior probability of the I-th class.

PR(K,J) The conditional probability of the J-th binary variable for the K-th class.

ND(J) The number of quantization intervals for the J-th qualitative or continuous variable.

BAS(J) The initial value of the range of quantization.

EX(J) The interval of quantization.

CNAME(J,I):

The name of the J-th variable (I = 1 allows four characters; I = 2 extends this to eight).

PN(K,J,I):

The conditional probability of the J-th quantizable variable, at the I-th level, of the K-th class.

FMAX The upper bound of the J-th interval of quantization, except for the last interval.

P(K,J,I):

The conditional probability of the J-th binary variable, at the I-th level, of the K-th class.

CN(K,J) The number of cases of class K classified into class J.

MMN The total number of variables.

IL(I) The code for ordering the I-th variable as read off the input data cards.

FMT The variable format.

MRC The number of binary variables to be combined.

IX(I) The code of the first variable to be combined into the I-th new variable.

IY(I) The code of the second variable to be combined into the I-th new variable.

MMP The index of the first new variable.

KL The index of a class.

KK The case number for the KL-th class.

FV(I) The I-th variable of the KK-th case of the KL-th class. I = 37 to 40 are identification variables.

FX(I) The I-th recoded variable.

IFV(I) The I-th binary variable for I = 1 to M; the I-th quantized variable for I = M + 1 to MRM; the identification variable for I = 37 to 40.

IM The index of the class into which the case is classified.

SUM The sum of the PC(KJ).

PC(KJ) The product of the prior probability with the conditional probabilities of the variables for class KJ. At statement 60 it becomes the posterior probability of class KJ.

DM The maximum posterior probability.

NE The total number of cases in class KL.

```
C       MAIN PROGRAM OF POCOD
C       CONDITIONAL PROBABILITY PROGRAM FOR DIAGNOSIS OF THYROID DYSFUNCTION
C
        DIMENSION PR(5,40),A(5),PC(5),PV(40,2),P(5,40,2),PN(5,15,26),
       1FV(40),FX(40),FMT(40),IL(40),IFV(40),CN(10,10),ND(15),BAS(15),EX(1
       25)
        DIMENSION CX(3),IX(6),IY(6),CNAME(15,2)
        DOUBLE PRECISION PR,PC,PV,A,PN,DM,SUM,P
        INTEGER CN
        DATA CX/'HYPE','HYPO','NORM'/
        READ(5,1) NC,M,MN,NU,MRM
      1 FORMAT(10I5)
C       NC= NUMBER OF CLASSES
C       M =NUMBER OF TRANSFORMED (CODED0-1) VARIABLES
C       MN= NUMBER OF NUMERICAL VARIABLES
C       NV(I)=NUMBER OF VALUES OF I-TH NUMERICAL VARIABLE
C       PR(I,J)=PROBABILITY OF A "1" FOR J-TH VARIBALE OF I-TH DISEASE
C       P(K,I,J)=PROBABILITY OF J-TH VALUE OF I-TH VARIABLE OF K-TH CLASS
C       PN(K,I,J)=PROBABILITY OF J-TH VALUE OF I-TH NUMERICAL VARIABLE OF K-TH CL
C       A(I)= APRIORI PROBABILITY OF DISEASE I
C       PV(J,I)= UNCONDITIONAL PROBABILITY OF J-TH VARIABLE,I-TH VALUE
C       NE=NUMBER OF CASES IN A CLASS
C       FV(I)= PATIENT SYMPTOM VECTOR
C       CN(I,J)= CONFUSION MATRIX ELEMENT
C
C                          MAXIMUM PARAMETER VALUES
C       M=40
C       NV(I)=15
C       NC=5
C       MN=10
C       MF=80
C
        NCP=NC+NU
        READ (5,2) (A(I),I=1,NC)
2       FORMAT(5F10.5)
        DO 4  J = 1,M
      3 FORMAT(F10.5,10X,F10.5,10X,F10.5)
      4  READ (5,3) (PR(K,J),K=1,NC)
         PRINT 5, NC, M, MN
      5  FORMAT(//20X,'PROGRAM PARAMETERS'/20X,'NC='I3,'  M='I3,'  MN='I3)
         PRINT 1050, (I,A(I),I=1,NC)
 1050  FORMAT(//20X,'A PRIORI PROBABILITIES OF DISEASES' //3(/20X,'CLASS
       1 NO.'I3,2X,' PROB='F10.5))
         DO 9 J=1,MN
        READ(5,6) ND(J),BAS(J),EX(J),(CNAME(J,I),I=1,2)
        PRINT 505,J,(CNAME(J,I),I=1,2),ND(J)
505     FORMAT(//20X,'CONDITIONAL PROBABILITIES OF CONTINUOUS VARIABLE NO.
       1:'I2,2A4,5X,'NO. OF VALUES ND='I5)
      6 FORMAT(I5,2F5.0,2A4)
        FMAX=BAS(J)-EX(J)
        NDV=ND(J)
        DO 9 I=1,NDV
        FMAX=FMAX+EX(J)
        READ(5,3) (PN(K,J,I),K=1,NC)
      9 PRINT 705,I,FMAX,(PN(K,J,I),K=1,NC)
    705 FORMAT(20X,I2,3X,F5.1,5X,5F10.5)
```

```
      DO 10 K=1,NC
      DO 10 J=1,M
      P(K,J,1)=1.0-PR(K,J)
   10 P(K,J,2)=PR(K,J)
      PRINT 11
   11 FORMAT (//20X,'SYMPTOM CONDITIONAL PROBABILITIES')
      DO 12 J=1,M
      DO 12 I=1,2
   12 PRINT 14,J,(P(K,J,I),K=1,NC)
   14 FORMAT(30X,I5,5F10.5)
      DO 15 K=1,NCP
      DO 15 J=1,NC
   15 CN(K,J)=0
      MM = M+1
      MMN = MN + M
      READ(5,25) (IL(I),I=1,MMN)
25    FORMAT(40I2)
      PRINT 26,(IL(I),I=1,MMN)
   26 FORMAT(5X,40I2)
      READ(5,30) FMT
30    FORMAT(20A4/20A4)
      PRINT 30, FMT
      READ(5,25) MRC
      READ(5,25) (IX(I),IY(I),I=1,MRC)
      PRINT 25,MRC,(IX(I),IY(I),I=1,MRC)
      MMP=MMN-MRC+1
      WRITE (6,33)  (I,I = 1,NC)
   33 FORMAT (//20X, ' CLASS', 3X,'NO.',3X,'CLAF.',3X,5('PR',I3,11X))
      DO 130 KL=1,NCP
      KK=0
   34 READ(5,FMT)(FV(I),I=37,39),(FV(I),I=1,MRM),FV(40)
      IF(FV(37).GT.900) GO TO 100
      KK=KK+1
      DO 35 I=1,MMN
      IN=IL(I)
35    FX(I)=FV(IN)
      DO 3520 I=MMP,MMN
      IQ=IX(I-MMP+1)
      IJ=IY(I-MMP+1)
      FX(I)=FV(IJ)+1
 3520 IF(FV(IQ).EQ.1) FX(I)=3
      DO 36 I=1,M
36    IFV(I)=FX(I)
      IF(MN) 38,38,3650
 3650 DO 37 J=MM,MMN
      JW=J-M
      FMAX=BAS(JW)-EX(JW)
      NDV=ND(JW)-1
      DO 3680 I=1,NDV
      FMAX=FMAX+EX(JW)
      IF(FX(J)-FMAX) 3660,3680,3680
3660  IFV(J)=I
      GO TO 37
3680  CONTINUE
      IFV(J)=NDV+1
37    CONTINUE
```

```
38       CONTINUE
         DO 3910 IK=37,40
3910     IFV(IK)=FV(IK)
C
C
C        RECOGNITION ROUTINE
C
         IM=1
        SUM=0.0
         DO 50 KJ=1,NC
         PC(KJ)=A(KJ)
          DO 40 J = 1,M
          IP =IFV(J)+1
40       PC(KJ)=PC(KJ)*P(KJ,J,IP)
         IF(MN) 50,50,42
42       DO 45 J=MM,MMN
         IP=IFV(J)
        JW=J-M
45      PC(KJ)=PC(KJ)*PN(KJ,JW,IP)
 50      SUM=SUM+PC(KJ)
         DO 70 KJ=1,NC
         IF(KJ.EQ.2) DM=PC(1)
 60      PC(KJ)=PC(KJ)/SUM
          IF(KJ.EQ.1) GO TO 70
         DM=DMAX1(PC(KJ),DM)
         IF(PC(KJ).EQ.DM) IM=KJ
 70       CONTINUE
C
C     END ROUTINE
C
    80 CN(KL,IM)=CN(KL,IM)+1
       PRINT 95,(IFV(I),I=37,40),KL,KK,CX(IM),(PC(K),K=1,NC),(IFV(I),I=1,
      1MMN)
    95 FORMAT(1X,I3,2I5,2X,I3,2X,I2,3X,I4,2X,A4,1X,3(2X,D13.5),24I2)
       GO TO 34
100     CONTINUE
        NE=KK
130     CONTINUE
        PRINT 135
  135 FORMAT(////////10X,'CLASSIFICATION RESULTS')
  140 FORMAT(/' CLASS',6X,'CLASSIFICATION'//5X,5(7X,I1,3X))
        PRINT 140, (K,K=1,NC)
        DO 150 I=1,NCP
150     PRINT 160,I,(CN(I,J),J=1,NC)
  160 FORMAT(/2X,I3,3(4X,I6))
        STOP
        END
```

REFERENCES

1. G. A. Gorry, "A system for computer-aided diagnosis," thesis presented to Sloan School of Management, M.I.T., Project MAC Rept. MAC-TR-44, 1967.

2. S. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxton, and R. Walker, "Evaluation and selection of variables in pattern recognition," in Computer and Information Sciences, vol. 2. New York: Academic Press, 1967, pp. 91-122.

3. S. Watanabe, "Karhunen-Loeve expansion and factor analysis," Trans. of the Fourth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes, pp. 635-660.

4. C. A. Kulikowski, "Pattern recognition approach to medical diagnosis," Proc. of the 1969 IEEE-SSCG Conf., Philadelphia. Accepted for publication in IEEE Transactions of SSCG, July 1970.

5. M. Lipkin and J. D. Hardy, "Mechanical correlation of data in differential diagnosis of hematological disease," J. Am. Med. Assoc., vol. 166, no. 2, pp. 113-125, 1958.

6. R. S. Ledley, "Logical aid to systematic medical diagnosis (and operational simulation in medicine)," J. Operations Res. Soc. of Am., vol. 4, no. 3, p. 392 (F7), 1956.

7. F. A. Nash, "Differential diagnosis, an apparatus to assist the logical faculties," Lancet, vol. 266, no. 6817, pp. 874-875, 1954.

8. R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis," Science, vol. 130, no. 3366, pp. 9-21, 1959.

9. R. S. Ledley and L. B. Lusted, "The use of electronic computers in medical data processing," IRE Trans. on Med. Electronics, vol. ME 7, no. 1, pp. 31-47, 1960.

10. H. R. Warner, A. F. Toronto, L. G. Veasy, and R. Stephenson, "A mathematical approach to medical diagnosis. Application to congenital heart disease," J. Am. Med. Assoc., vol. 177, no. 3, pp. 177-183, 1961.

11. H. R. Warner, A. F. Toronto, and L. G. Veasy, "Experience with Bayes' theorem for computer diagnosis of congenital heart disease," Ann. N. Y. Acad. Sci., vol. 115, pp. 558-567, 1964.

12. C. R. Bishop and H. R. Warner, "A mathematical approach to medical diagnosis: application to polycythemic states utilizing clinical findings with values continuously distributed," Computers and Biomedical Res., vol. 2, pp. 486-493, 1969.

13. R. A. Bruce and S. R. Yarnall, "Computer-aided diagnosis of cardiovascular disorders," J. Chronic Dis., vol. 19, no. 4, pp. 473-484, 1966.

14. A. W. Templeton, J. L. Lehr, and C. Simmons, "The computer evaluation and diagnosis of congenital heart disease using roentgenographic findings," Radiology, vol. 87, no. 4, pp. 658-670, 1966.

15. J. E. Overall and C. M. Williams, "Conditional probability program for diagnosis of thyroid function," J. Am. Med. Assoc., vol.183, no. 5, pp. 307-313, 1963.

16. L. T. Fitzgerald, J. E. Overall, and C. M. Williams, "A computer program for diagnosis of thyroid disease," Am. J. Roentgen., vol. 97, 1966.

17. J. E. Overall and C. M. Williams, "Models for medical diagnosis," Behavioral Sci., vol. 6, no. 2,pp. 134-141, 1961.

18. J. E. Overall and C. M. Williams, "Models for medical diagnosis: factor analysis. I. Theoretical," Med. Documentation, vol. 5, no. 2, pp. 51-56, 1961.

19. J. E. Overall and C. M. Williams, "Models for medical diagnosis: factor analysis. II. Experimental," Med. Documentation, vol. 5, no. 3, pp. 78-80, 1961.

20. J. Crooks, I. P. C. Murray, and E. J. Wayne, "Statistical methods applied to the clinical diagnosis of thyrotoxicosis," Quart. J. Med., vol. 28, New Series, no. 110, pp. 211-234, 1969.

21. M. F. Collen, L. Rubin, J. Neyman, G. B. Dantzig, R. M. Baer, and A. B. Siegelaub, "Automated multiphasic screening and diagnosis," Am. J. Pub. Health, vol. 54, no. 5, pp. 741-750, 1964.

22. L. Rubin, M. F. Collen, and G. E. Goldman, "Frequency decision theoretical approach to automated medical diagnosis," Proc. of 5th Berkeley Symp. on Math. Stat. and Probability, Berkeley, vol. IV, pp. 867-886, 1967.

23. G. A. Gorry and G. O. Barnett, "Experience with a model of sequential diagnosis," Computers and Biomedical Res., vol. 1, no. 5, pp. 490-507, 1968.

24. G. A. Gorry and G. O. Barnett, "Sequential diagnosis by computer," J. Am. Med. Assoc., vol. 205, no. 24, pp. 849-854, 1968.

25. A. S. Ginsburg and F. L. Offensend, "An application of decision theory to a medical diagnosis-treatment problem," IEEE Trans. on Systems Sci. and Cybernetics, vol. SSC-4, no. 3, pp. 355-362, 1968.

26. R. S. Ledley, "High speed automatic analysis of biomedical pictures," Science, vol. 146, no. 3641, pp. 216-223, 1964.

27. S. Abraham and C. A. Caceres, "Statistical computer methods for

diagnosis," Data Acquisition and Processing in Bio. and Med., vol. 3, New York: Pergammon Press, 1963, pp. 277-288.

28. P. L. Reichertz, "Comparison of the results of a diagnostic thyroid program in two different populations," Radiology, vol. 91, no. 1, pp. 32-36, 1968.

29. J. E. Overall and C. M. Williams, "Comparison of alternative computer models for thyroid diagnosis," San Diego Symp. for Biomedical Eng., vol. 3, p. 141, 1963.

30. K. Brodman and A. J. van Woerkom, "Computer-aided diagnostic screening for 100 common diseases," J. Am. Med. Assoc., vol. 197, no. 11, pp. 179-183, 1966.

31. S. Watanabe, Knowing and Guessing. New York: Wiley, 1969.

32. G. S. Sebestyen, Decision-Making Processes in Pattern Recognition. New York: Macmillan, 1962.

33. K. Fukunaga and W. L. G. Koontz, "Application of the Karhunen-Loeve expansion to feature extraction and classification," IEEE Trans. on Computers, vol. C-19, no. 4, pp. 311-318, 1970.

34. C. A. Kulikowski and S. Watanabe, "Multiclass subspace methods in pattern recognition," to be presented at the 1970 National Electronics Conference.