CONFIDENTIAL - 25

TITLE: AUTOMATED CHEST X-RAY ANALYSIS

AUTHOR(S): E. L. Hall, B. K. Rouge, USC
R. P. Kruger, LASL

SUBMITTED TO: Proceedings SPIE Conference, August 1976

MAST

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Energy Research and Development Administration, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

By acceptance of this article for publication, the publisher recognizes the Government's (license) rights in any copyright and the Government and its authorized representatives have unrestricted right to reproduce in whole or in part said article under any copyright secured by the publisher.

The Los Alamos Scientific Laboratory requests that the publisher identify this article as work performed under the auspices of the USERDA.

los alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87545

An Affirmative Action/Equal Opportunity Employer

MASTER

ef

AUTOMATED CHEST X-RAY ANALYSIS[†]

E. L. Hall
B. K. Rouge
R. P. Kruger*

Departments of Electrical Engineering & Radiology
University of Southern California
Los Angeles, California

*Los Alamos Scientific Laboratory
Los Alamos, New Mexico

Abstract

A computer hardware-software system has been developed for extracting optical and digital measurements from chest x-rays and using these measurements to classify the severity of lung disease. The system consists of a digital image scanner and an optical system consisting of an x-y controller for film transportation and an RSI Fraunhofer diffraction pattern sampling unit. Initially the classification has been of opacity textural patterns associated with coal workers pneumoconiosis (CWP). However, the system appears to be expandable to all diseases described in the international standard which include lung disease, heart disease, and cancer with only software modification. Recently extensions to large opacity detection, measures of lung vascularity and lung and heart volume have been made. Results indicate classification accuracy rates for pneumoconioses comparable to visual readings of the films by expert radiologists, and significantly better than non-radiologist physicians who often read these films.

Introduction

The successful use of computers in many diverse applications depends largely upon the development of a successful software system. The purpose of this paper is to describe recent developments in a software system called DR. ADAM (Diagnostic Radiology Aid to Diagnosis by Automatic Measurements) which was developed to provide measurements and decisions from chest x-ray films.

Previous studies on the Dr. Adam system have been reported in detail^(1, 2, 3). This paper will report recent results and updates while explaining the system in general. References will be made to the in-depth discussions of algorithmic justification. Studies are currently in progress for the detection of coal workers pneumoconiosis (CWP), pulmonary edema, and abnormal heart and lung volume. This paper will discuss the results of the CWP studies. Other capabilities will be reported at seminar 9 of session 4 of the SPIE symposium by Dr. A. Franklin Turner.

The extent of disease of CWP is judged on 12 minor and 4 major categories which are based on the profusion of small regular and irregular opacities per unit area. A fifth major profusion category consists of still larger, regular and irregular shaped lesions associated with progressive massive fibrosis (PMF). This fifth category often occurs when smaller opacities coalesce.

To determine the overall film profusion category, a decision as to the extent of the opacities in the lung field must first be made by the physician. Each lung field is vertically divided into three zones. Overall film profusion category is established by visually averaging the profusion in the affected zones. The result is a visually derived indication of zonal extent, opacity type, and profusion category for each film.

Each film is read by at least two government certified physicians according to the International Labor Organization (ILO) 1971 Classification of Radiographs of the Pneumoconiosis⁽⁴⁾. Depending on the agreement of these initial readers, a film decision could require a panel of readers. It is anticipated that the Dr. Adam system could assist in handling the mounting quantities of CWP x-rays being taken. Specifically, this automated system could screen out numerous normal films and indicate the abnormal ones which require a physicians interpretation. As mentioned previously, the measurements could reflect abnormalities of many types other than pneumoconiosis. For example, lung vascularity measurements can be made. Abnormal vascularity conditions are indicative of both pulmonary edema and cancer.

Prototype System

Using experience gained from a continually growing computer system, the final configuration is being trained as the Dr. Adam prototype system. Originally, data acquisition was implemented on a Hewlett Packard 2100 which controls the Dicomed image digitizer, the Recognition Systems optical spectrum analyzer, the x-y film transport and storage peripherals. The data obtained was analysed by using programs on an IBM 370 and a PDP 10. The changeover to a stand alone Dr. Adam system began with the acquisition of a PDP11 process controller. Additional hardware and software was incorporated and modified to provide greater speed and data handling capabilities. After training is completed, the system will be independent with data acquisition, storage and display.

While the initial prototype system consists of physically distinct but electronically integrated subsystems, this discussion will present a feasible scenario for their physical integration. A block diagram is shown in Figure 1.

[†] This work performed under NIOSH Contract HSM 99-73-92

Figure 1. Block Diagram of Hybrid System

A chest radiograph is inserted into the computer-controlled film transport. Behind this transport is placed a diffuse illumination source. The entire 35 by 35 centimeter film is scanned by an image disector camera at low resolution by the digital subsystem. The right and left lung regions are then detected and centroids of the lung boundary within each of six zones are computed. The film can then be transported under computer control into the path of the optical subsystem where spectral measurements are obtained about each of these six zonal centroids. A digital scan at high resolution can provide a data base for texture measurements within the lung region or other analysis.

A total of 191 digital spectral measurements have been gathered. They are both raw and derived measurements as shown in Table 1.

A diagnostic classification based on these measurements is then rendered. A graphical representation of the detected lung boundaries and centroid locations as well as film quality and diagnostic decisions are available in hard copy form on a terminal with hard copy output.

Table 1. Digital and Optical Measurements Summary

	1 - 5	Means (5)
	6 - 10	Ranges (5)
	11 - 15	Variances (5)
	16 - 20	Maximums (5)
	21 - 25	Minimums (5)
	26 - 45	Raw Features (20)
A. Conditional Probability Measurements (1-45)	46 - 53	Histogram Moments (8)
B. Moment Measurements (46-111)	54 - 61	Edge Histogram Moments (8)
C. Fourier Transform Geometric Samples (112-191)	62 - 71	Gray Level Spatial Moments (10)
	72 - 81	Edge Spatial Moments (10)
	82 - 89	Central Gray Level Moments (8)
	90 - 97	Central Edge Moments (8)
	98 - 104	Derived Gray Moment Measurements (7)
	105 - 111	Derived Edge Measurements (7)
	112 - 143	Annular Ring Measurements (32)
	144 - 175	Annular Wedge Measurements (32)
	176 - 191	Derived Fourier Measurements (16)

Film Quality

The first logical step in the reading of a chest x-ray is determining film quality. During the low resolution digital scan, a global histogram, which is a compilation of the number of image picture elements residing in each of 64 gray levels, can be computed. When compared with known histograms of underexposed, overexposed, fogged and normally exposed films, the exposure quality of the film to be processed can be determined. A detailed explanation of this determination process has been presented in detail⁽²⁾. The probability of each class of exposure is displayed on the terminal along with a decision of the exposure quality. Using this information, the operator has the option to continue with the diagnostic process or to abort the film at this point. Poor patient positioning can be determined by comparing the vertical signature to the vertical signature of a properly positioned chest x-ray, Figure 2a. The white backbone peak should be centered in the frame, and the peaks showing the chest cavity should be the appropriate distance from the center and edge of the film. Signature measurements are described further in the discussion of large

opacity detection. Improper film screen contact, another consideration of film quality, results in an unsharp film.

Large Opacity Detection (Signatures)

Another global feature obtained from the low resolution scan is the x profile or signature. This curve is the sum of all pixels in the y direction. The function is instrumental in determining the presence of progressive massive fibrosis (PMF), also referred to as a large opacity¹⁰. Large opacities are obvious to the eye (radiologist) and one would expect the same to be true for the computer^{10, 11}.

The rationale in looking for features is that the presence of large opacities should lead to a blurred image where the 'white' (i.e., the high-level densities) is predominant and where the amount of edge is reduced. An edge is the border between two adjacent surfaces within a small array of image points, if such a border exists at all.

The distribution of the 'white' within the lung for a chest radiograph can be obtained by taking the projection of the x-ray image with respect to a horizontal axis below the thoracic cavity. The projection of a digital image (signature) is defined as the vertical summation of the gray-level densities for each point on the horizontal axis.

In the case of the normally diagnosed x-ray image there might be clearly distinguishable for each one of the lungs, a left and right peak, separated by a well-defined valley, whereas for an abnormally diagnosed x-ray image, the valleys which characterize the signature flatten up. The difference between a positive and a negative signature can be seen in Figures 2a and 2b.

Figure 2a. Negative Chest Radiograph and Signature

Figure 2b. Positive Chest Radiograph and Signature

Also shown are the references for deriving the first two discriminant features. The points shown are as follows: R1 - outer peak of right lung, R2 - valley in right lung, R3 - inner peak of right lung, L1 - inner peak of left lung, L2 - valley in left lung, and L3 - outer peak in left lung.

The first two features defined as

$$S1 = \frac{y_{R1} - y_{R2}}{x_{R2} - x_{R1}} + \frac{y_{R3} - y_{R2}}{x_{R3} - x_{R2}} \quad (1) \quad S2 = \frac{y_{L1} - y_{L2}}{x_{L2} - x_{L1}} + \frac{y_{L3} - y_{L2}}{x_{L3} - x_{L2}} \quad (2)$$

are an approximation for the first derivative of the signature. S1 and S2 are similar features and they are a measure of the flatness of the right and left lung's signature, respectively. If the first derivative is high the film would be considered a negative one with respect to large opacities, while if the approximation for the first derivative is low, i.e., the signature is almost flat, the film would be considered positive with respect to large opacities.

These features were computed for 36 films which were read by three radiologists and a majority rule used to determine the classification. To select the best features from the normalized data a stepwise discriminant analysis was performed. Using the two best features resulted in a true positive and a true negative rate of 100% and 75% respectively.

The presence of large B and C size opacities can distort film features so badly that lung boundaries cannot be detected. For this reason, the large B&C opacities are indicated to the operator at this point. The large A size opacities may be detected by the optical subsystem which will be described in the Optical Measurement Section.

Lung Boundary Detection and Measurement

Once the digital film scan has been completed, a quality assessment made from its histogram, and a large capacity check has been made from the signature, it is necessary to detect and characterize the right and left lung boundaries. This is necessary to position the optical subsystem at zone centroids and to reduce the size of the digital data blocks. It was decided to approach lung boundary characterization with a combination of global and local criteria. The global criteria consisted of a threshold computed from the histogram previously used in film quality assessment. This thresholding operation created a binary mask which served as a first estimate of the lung region. This estimate was refined with local binary operators to complete the lung detection process. Details of this process can be found elsewhere.⁽⁶⁾

FIGURE 3. ZONAL REGIONS
RIGHT LUNG

The output of this lung detection process, shown in Figure 3 consists of three anatomical constraint points and three least squares polynomials per lung field. The constraint points (11, J1), (12, J2), and (13, J3) represent the lung apex, cardiac-diaphragm intercept, and costophrenic angle vertex respectively. Curves 1, 2, and 3 represent polynomial representations of the right side lateral chest, mediastinal-heart and diaphragm boundaries respectively. Note that mutual endpoint constraints of the polynomials insure a closed area. The left lung is characterized in an equivalent manner. Thus, each lung is characterized by 14 polynomial coefficients and three constraint points.

Using this information it is possible to calculate the centroid of each of the six lung zones and transfer these to the computer-controlled x-y transport of the optical subsystem. Higher resolution digital analysis of just the lung zonal regions is also possible because the interpolated polynomial values constitute lateral turn on and off indicators as the film is raster-scanned.

Image Textural Pattern Discrimination

After lung areas are determined, the high resolution scan is made. The digital image within the lung boundary is then stored. Also, the optical measurements may be taken in each of the six zones of the lungs. Thus, there are two approaches in the investigation of the texture patterns.

The problem of manually detecting and grading simple pneumoconioses from radiographs appears to be largely one of discrimination between normal pulmonary vascularity (lung marking) patterns and partial or complete obliteration of this normal tree-like structure by opacities of various sizes and profusions, which themselves exhibit a more or less textural nature.

Image texture refers to the visual sensation that one receives about the structure arrangement of an image region. The textural properties of a scene are often qualitatively described as coarse, grainy, striated, or rough. Computer simulations

have been done in which overlapping particles have been placed at random locations. When only a few particles are present, the visual impression is of discrete countable objects. However, when the number of particles is increased, the visual impression is of texture rather than countable objects.

Digital Measurements of Texture Patterns

The a priori specification of a set of measurements upon which to base a texture recognition decision is a difficult problem. No theoretical solution to this measurement discovery problem has been reported. However, one may objectively evaluate the effectiveness of a specified set using experimental classification accuracy. This suggests the conservative approach used in this study of computing a large number of measurements and evaluating their effectiveness using supervised classification techniques. Both first and second order probability distributions as well as spatial moments of the picture function and derived picture functions corresponding to edges had been used for texture measurements. A set of 111 such measurements are computed from the high resolution digital image. They are shown as the first 111 measurements in Table 1.

Spatial moments may be used for texture classification^(8,9) and are commonly used to describe a mass distribution. The total mass, centroid, and radius of gyration are often used in mechanics to describe the properties of a single object in a collection of objects. The generality of the method of moments

and the use of moment invariants in pattern recognition was described by Deutsch⁽¹⁰⁾. He proposed a theory that the process of human perception measures moment-type information in an image. Furthermore, he proposed that the information contained in the relation of a point to the weights of its neighbors can be hypothesized to demonstrate the invariances to size, orientation, mirror image, and symmetry which have been observed in human perception experiments.

The mathematical generality of the set of moment invariants has been pointed out by Hu⁽¹¹⁾ and Duda and Hart⁽¹²⁾. The set of infinite moments of any bounded probability density function exactly, completely, and uniquely describes the density function. All possible measurable features of an image are represented in the set of moments. Various normalization procedures may be performed on the computed moments to obtain matrices which are invariant to certain transformations such as translation, rotation, size change, or symmetry.

The moment computation may also be extended to reflect edge structure of texture patterns. The moments at any point (x, y) may be approximated by a linear combination of the moments of the mass function $f(x, y)$ and the moments of the partial derivative images. Thus moment computations for texture measurements may also be related to the intuitive and experimental concepts of edge structure and edge per unit area type measurements. The computational advantage of this procedure is that only low-order moments of the original and derivative pictures may be required for discrimination.

First order measurements such as the histogram, raw moments, central moments, variance, skewness coefficient, and the coefficient of excess are computed. First-order measures similar to these were initially used by Sprawls⁽¹³⁾ to detect abnormal thyroid conditions. However, first-order statistics are often insufficient to discriminate textural properties.

Some promising second-order statistical measures of texture have been developed for automated terrain classification by Haralick⁽¹⁴⁾. Ausherman's definitive work⁽¹⁵⁾ applied these same features in both the radiographic and terrestrial classification with considerable success. These features are all based on the computation of a measurement of relative frequencies $p(i, j)$ termed spatial gray-level dependence matrices. More specifically, it is assumed that texture-context information is adequately specified by the symmetric matrix of relative frequencies with which two neighboring picture elements are separated by a distance (d) and an angle (a) for each (i, j) gray level in a field. Thus p is a function of a , d , i , and expressed by $p(i, j, a, d)$ where

$$i = 0, \dots, n-1; j = 1, 111, n-1; a = 0^\circ, 45^\circ, 90^\circ, 135^\circ; d = 1, 2, \dots$$

and its size is defined by the number, n , of gray-level quantization values.

If appropriate a priori normalization is applied to the image, these matrices are approximately joint probability matrices. The number of operations necessary to produce this matrix for a given angle and distance is proportional to the number of picture elements in the field of view. Measures formed from these matrices which can be simply implemented on a minicomputer are the second angular moment, the moment of inertia about the diagonal of the p matrix, the entropy, a correlation measurement, a local homogeneity measurement, the mean, the range and the variance. The first and second order measures are described in detail in an earlier work⁽¹⁶⁾.

Optical Measurements of Texture Patterns

Within the optical approach to textural tone discrimination, various pattern recognition algorithms have been utilized to achieve classification on a digital computer.

The most salient of the optical systems was developed by Lendaris and Stanley for terrain classification⁽¹⁷⁾. The sensor or feature extractor measured aspects of the optical Fourier transform. This initial feasibility system utilized coherent optical techniques to extract measurements (features) of the Fraunhofer diffraction pattern which were classified via an interactive nonstatistical classification method. In the present application, the image $f(x, y)$ is transformed optically into its Fourier space $F(v_x, v_y)$ via the well-known property that coherent light passed through an image $f(x, y)$ and then through a thin convex lens produces the diffraction pattern in the back focal plane. The coordinates (u, v) of this transform plane are directly related to spatial frequencies and of the input image. The sensor is composed of 32 annular rings and 32 angular wedges which sample the squared modulus of the diffraction pattern (power spectrum) and produces appropriate signatures. The annular ring signature provides a rotationally insensitive means of detecting one or two dimensional spatial periodicity within apertured areas. In contrast, the wedge signatures are insensitive to periodic structure but sensitive to directions. Both frequency signatures subsequently have yielded valuable features for pneumoconiosis detection.

Each of six lung zones are sequentially illuminated by a 7.62 cm. diameter aperture, and at each zonal centroid, the 64 initial spectral measurements are collected as features for diagnostic classification. Each 64-member feature set of the 130 training samples was normalized to unit spectral energy, and the features were then normalized to zero mean with a standard deviation of one without reference to class. The first of these operations was performed to remove average gray level and linear contrast differences between films. The second and third operations were performed to increase subsequent machine classification accuracy. The averaged spectral signatures for each zone of the normal training set were graphed over those of the abnormal training set showing a distinct crossover frequency at ring 9 corresponding to objects with approximately 1.5 mm diameters. A very similar crossover frequency was also noted in the ring signatures of each of the other five lung zones. Upon inquiry with several B and C level readers, it was learned that the average small pneumoconiosis opacity was estimated to be between 1 and 2 mm in diameter. Thus, what is seen in these ring spectral signatures is most probably a higher average energy component in abnormal zones due to multiple diffractions of the opacities. Because the opacities vary in size and shape, their individual diffraction patterns when superimposed would lead to the observed broad

band phenomenon. For this reason it was decided to create several derived features in addition to the 64 features available from the diffraction pattern sampling unit. The mean (μ_1) and standard deviation (σ_1) of the first 9 rings were computed. In addition, the mean (μ_2) and standard deviation (σ_2) of the normalized log energies of the 24 rings above ring 9 were computed. The mean to mean (μ_1/μ_2), mean to standard deviation (μ_1/σ_1 , μ_2/σ_2), standard deviation to standard deviation (σ_1/σ_2) ratios were computed. The averaged energy signatures of the 32 angular wedge measurements for the training set yielded similar difference in their signature patterns. However, since these measurements are regularly dependent, there was greater inter-zone variability. This probably reflects the different angulation and quantity of the vascular tree as viewed in the individual lung zones. In any case, it is the partial or complete obliteration of this normal vascular structure by pneumoconiosis opacities that is being measured. A total of 8 derived mean and standard deviation measurements were computed from the 32 wedge signature values in a manner similar to that of the rings. Thus, the final feature set consisted of 80 features including 64 original and 16 zone specific derived features for each lung zone. They are measurements 112-191 in Table 1.

Large Opacity Detection

The Fourier transform measurements have been studied with respect to large opacity detection. As mentioned before, the presence of large opacities should lead to a decrease in the amount of edge in the image. The amount of the edge which is present in a given image is directly related to the distribution of the energy in the frequency domain between the lower and the higher frequencies.

A blurred image, where the amount of edge is small, should be characterized by an energy distribution (power spectra) which is shifted towards the lower frequencies more than it would be in the case of a normal (negative diagnosed) film. If this is the case, the power spectra should reveal a difference between a negative (normal) and positive (abnormal) blurred image. This can be seen visually in photographs taken of the power spectra from a negative and positive region.

Classification was performed on training data containing large opacity and normal lung zonal regions. The resulting scatter diagram showed wide separation. Using the best feature resulted in a true positive and true negative rate of 100% and 80% respectively.

Device Training

Before a device can successfully diagnose, it must be properly trained to recognize a specific disease pattern. In this instance it must be trained to recognize the presence or absence of a multitude of small opacities from several profusion categories in one or more of the six lung zones specified in the grading procedure. There are three significant tasks in device training. First, the feature space for each zone must be significantly reduced to prevent device overtraining. This is referred to as the feature selection process. Secondly, these reduced feature sets must be used to classify their respective zones, normal or abnormal. Thirdly, these zonal decisions must be integrated into an overall film decision.

Since the diagnosis is rendered from the film by observational nonquantitative means, there exists an inter-reader inconsistency. This inconsistency is characterized both by differences in the number and location of diseased and normal lung zones in a film and the manner in which these zonal decisions are integrated into the overall film classification by the individual physician. Because of this inconsistency, it was decided to initially train the device to render only a normal or abnormal decision on the 465 film-set from training data gathered from a subset of 130 films. Each of the 130 films had been independently read by the 6 physicians and judged by each of them either completely normal or abnormal in all 6 lung zones. There were 57 normal and 73 abnormal films in this training set. Thus, individual zonal reading patterns and integration of these zonal decisions into an overall film decision were removed as inter-reader variables during device training. With a training scheme outlined, the feature selection process is the next step. Although several feature selection criteria were considered, one of the most effective used was the Hotelling trace criteria.⁽¹⁷⁾ This reduced the dimensionality of each zonal feature set to the 10 features for which zonal classification accuracy was maximized.

Classification (Using Digital and Optical Measurements)

The classifier was chosen as a Bayesian decision process with multi-variate Gaussian class assumption.⁽¹⁾ Note that this assumption has meant that one only needs to compute a sample mean vector and covariance matrix from training data for each of m classes. The use of the resulting a posteriori probabilities allows not only the determination of class membership, but also the probabilistic certainty of that decision. Their use in the diagnostic classification will now be discussed.

Each zone was trained with a zone specific 10-dimensional feature vector x_i and the a posteriori probabilities were computed. For the initial prototype system, correct zonal training classification averaged 80 percent correct over the six zones with a low of 77 percent for the left middle zone and a high of 81.5 percent for the left upper zone. It was decided to produce an overall film decision by averaging the a posteriori probabilities of normal for the six zones and thresholding the result. If the overall probability (P) is greater than the threshold (T) the film is assigned to the normal class.

Initial Results

Device results may be presented in the form of confusion matrices of the form shown in Table 2. True negative percentage refers to correct normal reading while false positives refer to normals falsely judged abnormal. In a similar manner, a true positive represents abnormalities judged correctly while false negatives are abnormalities falsely judged normal.

Using P and threshold decision logic that $\bar{P} > T$ implies that the film is normal, experimental trials yielded maximum training classification accuracy at $T = .483$. This yielded the confusion matrix, shown in Table 3, for the 130 film device training set. Thus, 119 out of 130 or 91.5 percent of the film were classified correctly. This is comparable to the unanimous vote of the 6 manual readers.

As has been discussed earlier, reader inconsistency presupposes an accuracy level above which a trainable classifier which relies on reader information cannot be expected to exceed. The 130 film training set constituted the least inconsistent diagnostic information available. As a less biased test of the system against unknowns, it was decided to force a decision on all 465 films through the use of a consensus of 3 out of 5 manual readers. Neither the least consistent manual reader nor the prototype device were used to formulate the consensus. The a posteriori probability was computed for all 465 films using mean vectors and covariance matrices computed during device training. The threshold T was experimentally set to $T = .56$ in order to maximize device testing classification accuracy relative to the forementioned consensus. There were 102 consensus normals with 45 of these not completely free of disease. There were 363 consensus abnormal more than half of which indicated category one or first stage disease. The film library therefore represented a relative continuum between normal (category 0) and abnormal (category > 0).

The most inconsistent visual reader (reader 1) yielded the confusion matrix shown in Table 4. Reader 6 had the highest consistency, as shown in Table 5. If one averages the results of the 6 visual readers, the confusion matrix of Table 6 results.

The digital system yielded the results shown in Table 7 when compared to the consensus. The hybrid system results are shown in Table 8.

Table 2. General Form of Two Class Confusion Matrices

		Individual Reader	
		N	A
True Class	N	True Negative	False Positive
	A	False Negative	True Positive

Table 3. Training Results at $T = .483$

		Computer Class	
		N	A
True Class	N	86%	14%
	A	4%	96%

Percent Correct 92

Table 4. Most Inconsistent Reader
Reader 1

		N	A
Consensus Defined Truth	N	100% (102)	0% (0)
	A	33.9% (123)	66.1% (240)

Percent Correct 73.5

Table 5. Most Consistent Reader
Reader 6

		N	A
Consensus Defined Truth	N	65.1% (97)	4.9% (5)
	A	1.0% (2)	99% (361)

Percent Correct 96.5

Table 6. Averaged Reader Performance
Averaged Reader Results

		N	A
Consensus Defined Truth	N	96.3% (591)	3.7% (21)
	A	10.8% (212)	89.2% (1966)

Average Percent Consensus Correct Defined Truth 91.6

Table 7. Digital System Performance with
 $T = .56$
Digital System

		N	A
	N	48%	52%
	A	16%	84%

Percent Correct 74

Table 8. Hybrid System Performance
with $T = .56$

Hybrid System		Percent Correct 78.5
Consensus Defined Truth		
N	N	60.8 (62)
	A	39.2 (40)
A	N	16.5 (60)
	A	83.5 (303)

Table 9. Hypothetical B-Reader Performance
with Fixed True Positive Percentage
Estimated B-Reader Performance

Estimated B-Reader Performance		Percent Correct 82.5
Truth		
N	N	81.5
	A	18.5
A	N	16.5
	A	83.5

RECEIVER OPERATING CHARACTERISTIC

LEGEND

- ▲ Digital Training Results
- △ B Readers
- ▼ Independent Digital Test Results
- ▽ A Readers

Figure 4.

These initial results are significantly accurate to encourage refinements in the existing program design.

A recent paper by Goodenough⁽¹⁷⁾ compared performance of the general population of A and B level readers on approximately 14 thousand films for which a C level interpretation was necessary. These films could be considered difficult to diagnose. He used receiver operation curves (ROC) commonly applied in radar detection problems. The resulting curves labeled A and B in Figure 4 refer to the relative performance of A and B level readers when compared to truth as established by at least one C level reader. (Truth represents 100 percent true positive and 0 percent false positive.) A third curve refers to the performance of the prototype system. This curve was traced by varying T between 0 and 1 for the 465 film test set used in the study. A fourth curve represents the reading level subsequently established by the National Institute for Occupational safety and Health administration for certification of all future B level readers. This later curve was established by allowing 20 high proficiency B readers to specify the indicated operating curve relative to this truth. It should be pointed out that the curves shown were often derived from different film sets. However, it is hoped that they provide some approximate guidelines for this system's intermediate performance level.

Stand Alone Dr. Adam Prototype

Using the set of 130 films for both training and testing, the following results are on an overall film basis with no zonal classification being considered.

	Percent Correct
Digital training	93%
Digital testing	84%
Optical training	92%
Optical testing	80%

Conclusions

The results just presented yield evidence that 5 out of 6 of our readers diagnose substantially better than the general population of B readers on the 465 films used in this study. Even the least consistent of our readers diagnoses at a level above a B level curve. The computer system presently reads at a proficiency between that of an A and B reader but favoring the B curves.

In preliminary medical screening programs the cost associated with false negatives are much greater than that associated with a false positive. An abnormal judged normal will not be treated. However, a normal judged abnormal will at least benefit from further tests. In the present situation the device is designed to serve as a preliminary screening tool at a single central film collection and remailing facility. Thus, a higher false positive rate is much more acceptable than a high false negative rate since the former will always undergo at least one more reading but the latter may not.

Table 6 represents the average result of our readers. The percent correct of 91.6 represents a reasonable upper bound of inter-reader consistency for the films used in this study, and thus a trainable classifier could not be expected to exceed this bound. This conjecture is further strengthened by the device training results. It should also be noted that the five readers used to establish the consensus could all agree on only 81.5 percent of the 465 films. If one were to fix the true positive rate at 83.5 percent as in Table 8 and then estimate the false positive rate from the higher performance B reader curve of Figure 4, the following confusion matrix, shown in Table 9, would be formed.

Thus, with an equal true positive rate, the relative performance of the device and the accepted B reader population is put into better perspective. The results presented represent the analysis of the first large scale film reading test of this prototype system. Thus, the performance figures are expected to improve in subsequent tests.

Acknowledgments

The authors wish to acknowledge contract HSM-99-73-92 and its monitors Dr. Alan Purdy, Mr. Earle Shoub, and Dr. Robert Reger of the National Institute for Occupational Safety and Health for their continuing support. In addition to these we would acknowledge Dr. William B. Thompson, formerly a staff member at USC, and Dr. Alexander Sawchuk of USC for their invaluable aid. Finally Steve Dashiell, Jim Willis, Alan Bormuth, Rich Bagdazian, Bob Erbe, Eileen Jurak, Dave Peck, Harvey Modlin, and Rick Shiffman, have provided such key support to this project that it is difficult to imagine its completion without them.

References

1. Ernest L. Hall, Richard P. Kruger and A. Franklin Turner, "An Optical - Digital System for Automatic processing of Chest X-rays," Optical Engineering, Vol. 13 No. 3, pp. 250-257, May/June 1974.
2. Richard P. Kruger, Ernest L. Hall, and A. Franklin Turner, "Automated Measurements from Chest X-rays for Lung Disease Classification," American College of Radiology Conference, Las Vegas, Nevada, March 1975.
3. E. L. Hall, R. P. Kruger, A. F. Turner, "Automated Measurements from Chest X-rays for Lung Disease Classification," Proceedings Computer Graphics, Pattern Recognition and Data Structure Conference, Beverly Hills, California, May 1975.
4. G. Jacobson and W. S. Lainhart (eds.), "ILO U/C 1971 International Classification of Radiographs of the Pneumoconioses," Medical Radiology and Photography, Vol. 48, pp. 66-113 (1972).
5. C. E. Rossiter, D. Rivers, I. Bergman, et al, "Dust Content, Radiology and Pathology in Simple Pneumoconiosis of Coal Workers," Inhaled Particles and Vapors, II, Pergamon Press, New York, 1967, pp. 419-437.
6. D. H. Ballard and J. Sklansky, "Tumor Detection in Radiographs," Computers and Biomedical Research, Vol. 6, pp. 229-321, August 1973.
7. H. Wechsler, E. L. Hall, and A. F. Turner, "Automatic Detection of Large Opacities in Chest Radiographs," USC Technical Report, December, 1975.
8. R. N. Sutton and E. L. Hall, "Texture Measures for Automated Classification of Pulmonary Disease," IEEE Trans. Computers, Vol. C-21, pp. 667-676 (1972).
9. E. L. Hall, W. O. Crawford, K. Preston and F. E. Roberts, "Computer Classification of Coal Work-Pneumoconiosis," Proceedings of the First International Conference on Pattern Recognition, Washington, D. C., pp. 77-87, October 1975.