Performance Analysis of a New Computer Aided Detection System for Identifying Lung Nodules on Chest Radiographs

Russell C. Hardie*,†, Steven K. Rogers**,†, Terry Wilson**,† and Adam Rogers†

* (Corresponding author) Department of Electrical and Computer Engineering, University of Dayton, 300 College Park, Dayton OH 45469-0232
Phone: (937) 229-3178, Fax: (937) 229-4529
† Quale, LLC, 4017 Emerald Ct. S, Beavercreek OH 45430
** Air Force Research Laboratory, AFRL/ SN Wright Patterson AFB, Ohio 45433

Abstract. A new computer aided detection (CAD) system is presented for the detection of pulmonary nodules on chest radiographs. Here we present the details of the proposed algorithm and provide a performance analysis using a publicly available database to serve as a benchmark for future research efforts. All aspects of algorithm training were done using an independent data set containing 167 chest radiographs with a total of 181 lung nodules. The publicly available test set was created by the Standard Digital Image Database Project Team of the Scientific Committee of the Japanese Society of Radiological Technology (JRST). The JRST data set used here is comprised of 154 chest radiographs containing one radiologist confirmed nodule each (100 malignant cases, 54 benign cases). The CAD system uses an active shape model for anatomical segmentation. This is followed by a new weighted-multiscale convergence index nodule candidate detector. A novel candidate segmentation algorithm is proposed that uses an adaptive distance-based threshold. A set of 114 features is computed for each candidate. A Fisher linear discriminant (FLD) classifier is used on a subset of 46 features to produce the final detections. Our results indicate that the system is able to detect 78.1% of the nodules in the JRST test set with an average of 4.0 false positives per image (excluding 14 cases containing lung nodules in retrocardiac and subdiaphragmatic regions of the lung).

Keywords: Computer aided detection; chest radiographs; lung nodules; computer aided diagnosis

1. INTRODUCTION

Lung cancer accounts for the most cancer related deaths in the United States. According to the National Cancer Institute, 213,380 new cases are expected in 2007 [1]. Early detection of potentially cancerous pulmonary nodules may be a way to improve a patient’s chances for survival [2]. Computed tomography (CT) is a very sensitive imaging modality for detecting small pulmonary nodules [2]. However, chest radiographs are far more common. They are relatively simple, low cost, and provide only a fraction of x-ray dose of CT. Chest radiographs are also useful for a variety of purposes other than lung cancer detection. Because chest radiographs are so widely prescribed, improvements in the detection of lung nodules in chest radiographs could have a significant impact on early lung cancer detection.

In this paper, we present a new computer aided detection (CAD) algorithm for the detection of lung nodules in chest radiographs. Here we provide the details of the proposed algorithm and present a performance analysis using a publicly available database to serve as a benchmark for future research efforts. While several CAD systems have been presented in the literature including [3–13], our approach has several novel aspects. We employ a nodule candidate detector that represents a novel extension of the previously proposed convergence-index detector [14–16]. A novel nodule segmentation algorithm is then used to segment nodule candidates. For each nodule candidate detected and segmented, 114 features are computed. Several of the features used appear to be unique to our system for this application. We believe these innovative system components are one of the major contributions of this paper. Finally, our recommended system uses a Fisher linear discriminant (FLD) classifier [17–21] with a subset of 46 selected features to produce the final detections. While the FLD classifier is widely used, it does not appear to be a common choice for this application. Notwithstanding this, we have found that it provides surprisingly good performance when combined with the other modules in the proposed CAD system and is attractive because of its computational simplicity. In Section 3.4, we demonstrate the efficacy of the FLD by comparing it to two other classifiers.

In the spirit of [3], the performance of the proposed CAD algorithm is evaluated here using the publicly available database created by the Standard Digital Image Database Project Team of the Scientific Committee of the Japanese
TABLE 1. Distribution of nodules in the JRST database stratified by subtlety, pathology, and size.

<table>
<thead>
<tr>
<th>Size (mm)</th>
<th>Extremely subtle</th>
<th>Very subtle</th>
<th>Subtle</th>
<th>Relatively obvious</th>
<th>Obvious</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (&lt; 10 mm)</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>25 (16.2%)</td>
</tr>
<tr>
<td>Medium (≥ 10 mm and &lt; 20 mm)</td>
<td>18</td>
<td>16</td>
<td>29</td>
<td>20</td>
<td>5</td>
<td>50 (32.5%)</td>
</tr>
<tr>
<td>Large (≥ 20 mm)</td>
<td>5</td>
<td>10</td>
<td>17</td>
<td>17</td>
<td>7</td>
<td>38 (24.7%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathology</th>
<th>Benign</th>
<th>Malignant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>54 (35.1%)</td>
<td>100 (64.9%)</td>
</tr>
</tbody>
</table>

Society of Radiological Technology (JRST) [22] from April 1995 to March 1997. The full JRST data set is comprised of 154 chest radiographs containing one nodule each (100 malignant, 54 benign), and 93 chest radiographs containing no nodules. The JRST images are each 2048 × 2048 pixels in size, with a pixel spacing of 0.175 mm and 4096 gray levels (12 bits). According to the JRST, all the cases in the database have been examined using CT images to provide truth information. Nodules in the JRST database are classified as malignant or benign and graded with the consensus of three chest radiologists as to their visual subtlety. This subtlety rating includes five levels: extremely subtle, very subtle, subtle, relatively obvious, and obvious. The distribution of nodules in the JRST dataset according to subtlety, pathology, and size (effective diameter) is summarized in Table 1. We use these ratings to stratify the CAD performance analysis.

Since generalizability of CAD systems is essential to their utility, we want to perform the most rigorous validation possible of our system with regards to generalizability. To do so, training and tuning of all modules in our CAD system is done here using a separate and independent data set provided courtesy of Riverain Medical. The full training set is comprised of 167 chest radiographs with a total of 181 lung nodules. Using the same size categories shown in Table 1, the training set is comprised of 5.5% (10/181) small, 58.6% (106/181) medium, and 35.9% (65/181) large nodules. This training set represents a sampling from Riverain’s database and consists of 30 computed radiography images, 7 digital radiography images and the remainder are digitized film images. We also present 10-fold cross validation results using only the JRST data. These results are included for reference and to serve as a reproducible benchmark requiring only the JRST data. We believe that by testing on a publicly available database, but training on a diverse and independent dataset, we are able to provide a realistic performance benchmark for future research efforts. We believe this is another distinctive contribution of this paper.

In a fashion similar to that in [4], we have excluded cases in this study containing lung nodules in the “opaque” portions of the chest radiograph that correspond to the retrocardiac and subdiaphragmatic regions of the lung. For training, we use the 160 of the 167 Riverain cases that contain no lung nodules in opaque portions of the radiographs. These 160 cases contain a total of 173 nodules. For testing data, we utilize the 140 JRST cases containing nodules in the non-opaque regions of the radiographs. Detection of nodules in the opaque regions on the chest radiographs is certainly an important issue that needs to be addressed by the research community. Note that 7.6% (76/1000) of the cases in [4] contain nodules in opaque portions of the lung. Opaque cases represent 9.1% (14/154) of the JRST dataset and 4.2% (7/167) of the Riverain training data.

The remainder of this paper is organized as follows. The overall CAD system and its modules are described in Section 2. This includes the preprocessor, nodule candidate detection, nodule candidate segmentation, features and classifier. Experimental results are presented in Section 3. These results include performance comparisons with five other previously published methods. Conclusion are provided in Section 4.

---

1 Riverain Medical, 3020 South Tech Boulevard Miamisburg, Ohio 45342-4860. Maker of the FDA approved RapidScreen® lung CAD system.
2. CAD SYSTEM

In this section, the proposed CAD system is described. A top level block diagram of the CAD system is shown in Fig. 1. Chest radiographs go through a preprocessing step to resize, perform local contrast enhancement, and perform anatomical segmentation. Next, nodule candidates are detected and then segmented. Features for all segmented nodule candidates are computed next. Finally, a classifier is applied to produce final detections. This top level system is fairly conventional. However, we believe that the novelty of our approach lies in the specific choice and design of its modules.

2.1. Preprocessor

The first step in the preprocessing stage is to resample the input images to a common pixel spacing. We have selected a pixel spacing of 0.7 mm, which yields images of size 512 × 512 for the JRST data. This resolution provides a significant reduction in processing time and allows for some noise smoothing [4]. The resampling is accomplished using an 11 × 11 finite-impulse-response low-pass anti-aliasing filter followed by bilinear interpolation. After the resampling, the preprocessor generates three images to be used in subsequent modules: a local contrast enhanced image, the lung segmentation masks, and a normalized image. Computation of the three output images is described below.

2.1.1. Local Contrast Enhancement

To normalize the contrast across different images and within each image, we perform local contrast enhancement (LCE) identical to that in [3, 6]. This operation is given by

\[ y(m, n) = \frac{x(m, n) - \mu(m, n)}{\sigma(m, n)}, \]  

where \( x(m, n) \) is the input image, \( y(m, n) \) is the LCE image, \( \mu(m, n) \) is a local mean estimate, and \( \sigma(m, n) \) is a local standard deviation estimate. The local mean is computed by

\[ \mu(m, n) = x(m, n) * h(m, n), \]  

where \( h(m, n) \) is a filter kernel.
where \( h(m,n) \) is a Gaussian low-pass filter with an impulse response function having a standard deviation of 16 (corresponding to a one standard deviation diameter span of 22.4 mm). The local standard deviation is computed by

\[
\sigma(m,n) = \sqrt{\mu^2(m,n) - h(m,n) * \mu^2(m,n)}.
\]

The local contrast images are used for candidate detection and segmentation. Features are also computed from this LCE image.

### 2.1.2. Lung Segmentation

Another, very crucial, preprocessing step is to segment the lungs from the full chest radiograph. This is done to limit the detection of lung nodules to the segmented lung fields and to prevent false positive (FP) detections outside the lung field. For automated lung segmentations, we employ an ASM [23–26] and closely follow the method described in [27]. This ASM method is the same as that used for lung field segmentation in [3, 6]. A typical lung mask obtained using the ASM is shown in Fig. 2. To train the ASM, manual segmentations of the 93 JRST cases that contain no nodules are used. These independent cases were used exclusively for ASM training. The manual perimeters are all normalized by height, width, and position using landmarks obtained using a rule-based method similar to that in [28]. In particular, the top and bottom of the lung field are identified as shown by the horizontal lines in Fig. 2. These are used to normalize the height of each lung mask. The approximate horizontal middle of each lung is also found and these are shown by the vertical lines in Fig. 2. Note that we do not seek to find the outer extreme of the ribcage on the left and right. We have observed that some chest radiographs used in normal clinical settings, such as those in the Riverain dataset, may not include the entire lung field (something not seen in the well conditioned JRST dataset). Because of this, we have found that the low intensity column that we use to define the “middle” of the lung tends to serve as a more robust landmark to define lung width. These mid-lines are also straightforward and computationally simple to locate. The reference point for all the perimeter coordinates is the intersection of the vertical centerline and the bottom line.

Since the ASM used here follows the method described in [27], we refer the reader to that source for further details. Our tuning parameters vary slightly from those reported in [27]. The ASM used here is applied first on coarse 2.8 mm pixel spacing images and then refined on 1.4 mm pixel spacing images. The perimeter appearance model includes 6 pixels on either side of the lung perimeter at each landmark. The search process spans 15 pixels either side of the current perimeter point. We use 15 principal component dimensions (containing 95% of the variance) and we truncate these components such that they lie no more than 2.5 standard deviations from the mean in each of the 15 dimensions. The ASM lung segmentation approach is quite effective and computationally efficient, as reported in [27]. We have observed that the most common ASM lung segmentation errors occur in the left lung near the diaphragm, heart, or aorta. When these mask errors occur, they tend to include extra area into the lung mask. While this might lead to extra false positive (FP) detections, it generally does not exclude true positive (TP) detections.

### 2.1.3. Normalized Image

The normalized image is computed by using the sample estimates of the mean and standard deviation of the pixel values that lie inside the segmented lung field in the 0.7 mm resampled image. We then subtract this mean from each raw pixel value and then divide each by the standard deviation. The normalized image is used in the computation of a number of nodule features. The normalization step provides consistency across the input images for improved generalization of the features. By limiting the estimate of the statistical parameters for normalization to the lung field, we eliminate many potential outliers due to the edge of the image, various markings, and the variable amount of tissue area outside the lung on the chest radiograph.

### 2.2. Nodule Candidate Detection

After preprocessing, we employ a nodule candidate detection algorithm to locate potential lung nodules. Some examples of nodule candidate detectors presented in the literature include Laplacian of Gaussian (LoG) blob detectors...
FIGURE 2. Typical ASM lung segmentation result (JPCLN003) shown for the 1.4 mm pixel spacing image. The horizontal lines are used to normalize the height and vertical position of the lung mask. The vertical lines normalize the width and horizontal position of the lung mask.

[3, 6, 29–31], the average radial gradient (ARG) detector [4], convergence index (CI) filter based methods [14–16], and filtered image differencing [12, 13]. In our experiments on the Riverain training data, we have observed the best results using CI-based detectors.

The CI filter proposed in [14] is given by

\[ c(m, n) = \frac{1}{|W|} \sum_{(k,l) \in W} \cos(\theta_{m,n}(k,l)), \] (4)

where \( \theta_{m,n}(k,l) \) is the angle between the radial vector pointing from pixel \((m + k, n + l)\) to \((m, n)\) and the intensity gradient vector at pixel \((m + k, n + l)\). Here the gradient is estimated discretely using finite differences. The CI value at pixel \((m, n)\) is the average of \(\cos(\theta_{m,n}(k,l))\) for \((k,l) \in W\). This is computed by taking the inner product of the unit radial vector and unit gradient vector. The set \(W\) typically defines a circular neighborhood as is depicted in Fig. 3. Variations on this CI filter have been proposed in [14–16]. The basic idea of all these CI filters is that structures in the image with circular shape that are brighter than their surroundings will tend to have gradients that point towards the center of the object. This would yield a high CI value. Note that the magnitude of the gradients associated with a structure do not impact the CI value (gradient direction only). This can be helpful in detecting very dim and subtle nodules (which are quite common).

Here we define a generalized CI filter allowing arbitrary weights for each gradient angle term. This weighted convergence index (WCI) filter is given by

\[ \tilde{c}(m, n) = \sum_{(k,l) \in W} w(k,l) \cos(\theta_{m,n}(k,l)), \] (5)

where \(w(k,l)\) is the weight for the gradient angle for position \((m + k, n + l)\) in the window about \((m, n)\). Since it may not be possible to identify one set of weights that can successfully detect the full range of nodule sizes and/or shapes, we propose computing multiple WCI images and selecting the maximum of these at each pixel. Thus, the weights can be optimized for specific applications and could be designed to target specific shapes, orientations, and textures. This gives rise to what we term the weighted multiscale convergence-index (WMCI) filter. The output of the WMCI filter is given by

\[ \tilde{c}(m, n) = \arg \max_j \left\{ \sum_{(k,l) \in W} w_j(k,l) \cos(\theta_{m,n}(k,l)) \right\}. \] (6)
where \( w_j(k, l) \) are the weights for the \( j \)'th scale WCI filter.

The WMCI image is converted into preliminary candidate detections by selecting spatial local-maxima with a WMCI value above a specified threshold. Detection thinning is performed to remove very close adjacent detections. If any preliminary detections are within a specified distance of one another, only the preliminary detection with the larger WMCI value is retained. For the result presented here, we use a distance threshold of 5 mm. As described in Section 2.5, we perform another round of adjacent candidate rejection after the classifier posterior probabilities are computed for each candidate. Thus, it is not critical that complete detection thinning take place at the candidate detection stage. In fact, the extra detections increase the likelihood that a nodule will have a detection near the center. These centralized detections tend to serve as better seed points for nodule segmentation. The threshold, in conjunction with the choice of weights, control the operating point for the candidate detector.

For the results reported in this paper, we employ the weighting functions shown in Fig. 4. There are three sets of weights, generating WCI images at three scales. The outer diameters of the three weight sets are 14 pixels (9.8 mm), 20 pixels (14 mm), and 26 pixels (18.2 mm). The first set of weights corresponds to a standard CI filter with a circular window function. The other two functions have multiple tiers with decreasing weight moving away from the center. The sum of the first set of weights is 1, while the sum of the second and third are 1.47 and 1.88, respectively. We have observed that, much like with scale-space representations, WCI values for larger windows tend to be smaller than their counter parts with smaller windows. Without boosting these larger scale WCI image values, the smaller scales would dominate since we perform a max operation across scales in (6). Note that the weights shown in Fig. 4 have been selected based on empirical results obtained with the Riverain training data exclusively. In fact, these parameters were determined prior to us obtaining the JRST dataset.

A typical candidate detection output using the WMCI detector is shown in Fig. 5 (JPCLN004) superimposed on the corresponding LCE image. A threshold of 0.5 is used for all WMCI detections. Each detection is shown with a circle corresponding to the window size of the WCI scale that gave rise to that detection. The truth cue is also shown, near the middle of the right lung. A closer look at the detection near the truth cue is provided in Fig. 6. Here the contours of the WMCI image are shown along with the WMCI detections and the truth cue and circle. Note that the WMCI detections occur at local maximum of the WMCI image as can be seen from the contours. In this case, the detection is quite close to the actual truth cue. A performance analysis of the WMCI candidate detector is provided in Section 3.2.

### 2.3. Nodule Candidate Segmentation

After candidate detection, we apply an adaptive distance-based threshold (ADT) algorithm to segment each nodule candidate. These segmentations are used as a key part of the computation of most of the candidate features. A block diagram of the ADT algorithm is shown in Fig. 7. The ADT operates on the LCE image, \( y(m, n) \) as defined in (1), and
begins with the following thresholding operation

\[
s(m,n,T_0) = \begin{cases} 
1 & y(m,n) > T(m,n,T_0) \\
0 & \text{otherwise}
\end{cases}.
\]

(7)

Here \(T(m,n,T_0)\) is the adaptive threshold function given by

\[
T(m,n,T_0) = \begin{cases} 
T_0 + T \left( 1 - e^{-d(m,n)/r_{\text{max}}} \right) / (1 - e^{-1}) & d(m,n) < r_{\text{max}}^2 \\
\infty & \text{otherwise}
\end{cases}.
\]

(8)

This adaptive threshold is a function of the distance of a given pixel to the detection cue

\[
d(m,n) = (m - m_0)^2 + (n - n_0)^2,
\]

(9)

where \((m_0,n_0)\) is the detection cue point. The maximum nodule radius for segmentation purposes is specified by \(r_{\text{max}}\) in pixels. The range of thresholds from the cue point to those at \(r_{\text{max}}\) is specified by \(T_\Delta\). The adaptive threshold parameter is the offset \(T_0\). For the CAD results presented here, we use \(r_{\text{max}} = 25\) (17.5 mm) and \(T_\Delta = 1.7\). The distance-based threshold is shown in Fig. 8 for these parameters, where \(T_0 = 0\). The idea behind the distance-based threshold is to exploit a priori knowledge that nodules tend to be approximately circular lesions. Increasing the threshold with distance tends to prevent the segmentations from erroneously growing far from the cue point along a bright structure such as a rib.

Performance Analysis of a New CAD for Lung Nodules July 9, 2007
FIGURE 5. WMCI detector output for JPCLN004. The size of each detection circle indicates the WCI scale that gave rise to that detection (i.e., diameters of 9.8 mm, 14 mm, or 18.2 mm).

FIGURE 6. WMCI output for JPCLN004 near truth cue. Contours of WMCI image are shown along with the truth cue and truth circle.
Next, the binary image $s(m,n,T_0)$ is logically ANDed with the lung mask, to prevent the segmentation from growing outside the defined lung boundary. We then apply a hole filling morphological operation followed by a morphological opening with a small circular structuring element with a radius of 1 pixel. Finally, only ON pixels connected to the cue point, using the 4-connected neighbor criterion, are retained. This yields the mask denoted $\tilde{s}(m,n,T_0)$. The adaptive threshold, $T_0$, is selected as the one that maximizes the average radial gradient for the LCE image averaged over the mask $\tilde{s}(m,n,T_0)$. This average radial gradient objective function is given by

$$ARG(T_0) = \frac{1}{|\tilde{S}(T_0)|} \sum_{(k,l) \in \tilde{S}(T_0)} |g(m_0 + k, n_0 + l)| \cos \left( \theta_{m_0,n_0}(k,l) \right),$$  

(10)

where $\tilde{S}(T_0) = \{(k,l) : \tilde{s}(m_0 + k, n_0 + l, T_0) = 1 \}$ and $g(m_0 + k, n_0 + l)$ is the intensity gradient magnitude of LCE.
FIGURE 9. Average radial gradient as a function of threshold $T_0$ for JPCLN003. The threshold that maximizes this curve is selected to produce the final segmentation mask.

image at pixel $(m_0 + k, n_0 + l)$. We find the $T_0$, in the range $-2 \leq T_0 \leq 0$, that maximizes ARG($T_0$), and use that to generate the final nodule segmentation mask. A typical plot of $T_0$ versus ARG($T_0$) is shown in Fig. 9 (JPCLN003). The ARG value is often at a maximum when the segmentation follows the contours of the lesion and has proven to be a relatively simple and effective objective function with which to identify a good $T_0$. It is interesting to note that we have found that by using the radial gradient averaged over the full mask, compared with averaging only over the perimeter, the algorithm tends to produce more robust segmentations. These segmentations are less likely to erroneously follow along a rib or lung boundary or be too small.

Note that the threshold parameters for the ADT segmentation have been selected based on empirical studies exclusively using the Riverain training data. As with the nodule candidate detector, these parameters were selected prior to working with the JRST data. Examples of segmented nodules using the current ADT system are shown in Fig. 10 for JPCLN003 (obvious), JPCLN015 (relatively obvious), JPCLN087 (subtle) and JPCLN121 (very subtle). Also shown on these plots are the WMCI cue points, the truth cue points, and the corresponding isocontours of the distance-based threshold function as applied to the image data in the ADT algorithm. Note that accurate segmentations are an important part of our CAD system since many of the nodule candidate features, described in Section 2.4, rely on these segmentations. In particular, the ADT segmentations are used for computing features from the normalized image, LCE image, and WMCI image. We believe that one of the main advantages of the ADT segmentation lies in how its limits segmentations from extending too far from the cue point, without abruptly truncating the segmentation. It also conforms better to irregularly shaped nodules and is less sensitive to having non-centralized cue points than other methods we tested. For these reasons, we believe that ADT approach is a good choice here and may also be of value in other segmentation applications.

2.4. Nodule Candidate Features

For each nodule candidate we compute a total of 114 features listed in Table 2. These includes 9 geometric features, 18 intensity features and 17 gradient features. The geometric features are computed using only shape and position information provided by the ADT segmentation. The intensity and gradient features are computed for the normalized, LCE, and WMCI images, using the ADT segmentation mask to define the candidate boundary. Many of these features
FIGURE 10. Nodule segmentation examples for (a) JPCLN003 (obvious), (b) JPCLN015 (relatively obvious), (c) JPCLN087 (subtle) and (d) JPCLN121 (very subtle). Also shown on these plots are the WMCI cue points, the truth cue points, and the corresponding isocontours of the distance-based threshold function for the ADT algorithm.

Let us begin with the geometric features. The geometric Size feature is a unidimensional measurement representing the maximum diameter across the candidate segmentation. The X-Fraction and Y-Fraction features refer to the normalized horizontal and vertical positions of the candidates within the ASM segmented lung containing the candidate. In particular, the horizontal and vertical distances from a reference point at the top center of the lung field is computed. These are divided by the width and height of the lung mask, respectively, to produce X-Fraction and Y-Fraction. Circularity is computed by the segmented area divided by \( \pi r^2 \), where \( r \) is the maximum bounding box radius for Circularity 1 and is the unidimensional radius for Circularity 2. The Distance to Lung Perimeter features is computed using the distance transform. It is normalized by dividing each candidate distance by the maximum distance for the given ASM.
the nodule segmentation mask. Thus, to get the most out of these features, we believe an accurate nodule segmentation in the horizontal and vertical directions within the segmented region. Note that most of these features rely heavily on gradient magnitude separation inside and outside standard deviations. The skew and kurtosis features are computed with the sample skewness and sample kurtosis estimates, respectively. The moment features are the seven invariant moments for 2D functions presented in [32]. These moments are invariant to translation, rotation, and scale change [32].

Three types of gradient features are computed: gradient magnitude, radial deviation, and radial gradient. Radial deviation corresponds to the angle between gradient vectors and corresponding radial vectors. These are the same angles depicted in Fig. 3, and are denoted \( \theta_{m_c,n_c}(k,l) \), where \( (m_c,n_c) \) is the centroid of the relevant candidate. Radial gradient refers to gradient strength along radial vectors and is denoted \( |g(m_c + k, n_c + l)| \cos(\theta_{m,n}(k,l)) \). The XY gradient magnitude separation inside feature is the mean separation between the magnitude of the gradients oriented in the horizontal and vertical directions within the segmented region. Note that most of these features rely heavily on the nodule segmentation mask. Thus, to get the most out of these features, we believe an accurate nodule segmentation is the key.

A subset of the 114 features listed in Table 2 is used for the classifier in our CAD system. The subset of features is chosen using a sequential forward selection (SFS) process [33]. The objective function for the SFS is based on the area under the free receiver operating characteristic (FROC) curve for the chosen classifier. Here, the FROC curve is defined as the fraction of TPs passed versus the average number of FPs per image. The FROC curve is generated exclusively with Riverain training data using 10-fold cross validation. For the 10-fold cross validation, classifier posterior probabilities are determined for candidates in 10% of the training cases, based on a classifier trained on the remaining 90% of cases in the training set. The particular 10% is cycled through in this fashion until posterior probabilities are computed for all candidates. The FROC curve is then generated by thresholding the posterior probabilities and scoring the results. The SFS objective function is the area under this FROC curve from 2 to 4 FPs per image. This range has been selected because we believe it may include suitable operating points for clinical applications. The number of features is chosen so as to maximize the objective function.

We consider three distinct classifiers here and perform a separate SFS for each. The three classifiers are the FLD classifier [17–21], the Gaussian Bayes Linear (GBL) classifier [33], and the quadratic classifier [33]. More will be said about these classifiers in Section 2.5. The FLD classifier uses 46 features and these are marked in Table 2 with an (F). The GBL classifier uses the 42 features marked with a (G), and the quadratic classifier uses the 15 features marked with a (Q). Note that the number of features used by these classifiers lies in the range reported in the literature for other CAD systems. For example, the CAD in [4] uses 71 features, [9] uses 202, and [3] uses a two-stage classifier with 10 and 20 features, respectively.

Some of the features used here can also be found in the system in [4]. These features include several of the geometric and intensity features. However, we believe many of the features listed in Table 2 are unique to our system for this application. Note that many of the features are in groups of three: inside, outside, and a separation feature for a given statistic. We have not seen this done as extensively elsewhere, particularly for the gradient features. We think this approach provides helpful local context for the nodule candidates. For example, if a candidate has a strong radial gradient, but is in a background that also has a strong radial gradient, this would suggest that the candidate is less likely to be a TP. Based on the feature selection results summarized in Table 2, it does appear that these separation features are frequently selected, demonstrating their saliency. Furthermore, we are not aware of any other system employing the invariant moment features for this application. Based on the feature selection results, these features also appear to be very salient. It is interesting to note that the features used in [3] are quite different from those used here and in [4]. The system in [3] uses features primarily derived from a multi-scale Gaussian filterbank. A summary of CAD systems and the key features used by those systems can be found in [10].

### 2.5. Classifier

After the features are computed we form the final detections in the CAD system using a classifier. Our recommended system uses an FLD classifier [17–21]. We have selected this linear classifier here because of its ability to generalize well, even with a relatively small number of TP training samples. It is also very computationally simple for both
TABLE 2. List of features computed for the CAD system. The features are derived from the binary segmentation image (ADT), the normalized image (Norm), the LCE image, and the WMCI image. Of the full set of features, the 46 used by the FLD classifier are marked with an (F). The 42 features used by the GBL classifier are marked with a (G), and the 15 features used by the quadratic classifier are marked with a (Q).

<table>
<thead>
<tr>
<th>Geometric Features</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>ADT</td>
</tr>
<tr>
<td>Area</td>
<td></td>
</tr>
<tr>
<td>X-Fraction</td>
<td></td>
</tr>
<tr>
<td>Y-Fraction</td>
<td></td>
</tr>
<tr>
<td>Eccentricity</td>
<td></td>
</tr>
<tr>
<td>Orientation</td>
<td></td>
</tr>
<tr>
<td>Circularity 1</td>
<td></td>
</tr>
<tr>
<td>Circularity 2</td>
<td></td>
</tr>
<tr>
<td>Distance to Lung Perimeter</td>
<td></td>
</tr>
</tbody>
</table>

**Intensity Features**

| Maximum Value Inside                | FG |
| Minimum Value Inside                | FG |
| Mean Inside                         | G  |
| Mean Separation                     | FG |
| Contrast 1                          | G  |
| Contrast 2                          | G  |
| Standard Deviation Inside           | G  |
| Standard Deviation Outside          | G  |
| Standard Deviation Separation       | G  |
| Skew Inside                         | FG |
| Kurtosis Inside                     | FG |
| Moment 1 Inside                     | FG |
| Moment 2 Inside                     | Q  |
| Moment 3 Inside                     | Q  |
| Moment 4 Inside                     | F  |
| Moment 5 Inside                     | F  |
| Moment 6 Inside                     | G  |
| Moment 7 Inside                     | F  |

**Gradient Features**

| Gradient Magnitude Mean Inside      | FG |
| Gradient Magnitude Standard Deviation Inside | FG |
| XY Gradient Magnitude Separation Inside | FG |
| Radial-Divergence Mean Inside       | FG |
| Radial-Divergence Mean Outside      | FG |
| Radial-Divergence Mean Separation   | FG |
| Radial-Gradient Mean Inside         | FG |
| Radial-Gradient Mean Outside        | FG |
| Radial-Gradient Mean Separation     | FG |
| Radial-Gradient Standard Deviation Inside | FG |
| Radial-Gradient Standard Deviation Outside | FG |
| Radial-Gradient Standard Deviation Separation | FG |
| Radial-Gradient Mean Perimeter      | FG |
| Radial-Gradient Standard Deviation Perimeter | FG |
training and testing and yields surprisingly good performance on the cross-validation experiments we have conducted. In particular, we show in Section 3.3 that the FLD classifier generally outperforms the GBL classifier [33] and the quadratic classifier [33] on these data using comparable feature selection procedures.

The FLD detection statistic is given by

$$d_i = w^T x_i,$$

where

$$w = (C_{TP} + C_{FP})^{-1} (\mu_{TP} - \mu_{FP}).$$

(11)

Here $x_i$ is an $M \times 1$ feature vector for candidate $i$ and $w$ is an $M \times 1$ vector of weights. The variables $C_{TP}$ and $C_{FP}$ represent the class conditional covariance matrices, and $\mu_{TP}$ and $\mu_{FP}$ are the class conditional means. These are obtained using sample estimates with TP and FP feature vector examples, respectively. The FLD weights in (12) maximize the following quantity

$$J(w) = \frac{(w^T (\mu_{TP} - \mu_{FP}))}{w^T (C_{TP} + C_{FP}) w}.$$  
(13)

The objective function in (13) represents the detection statistic’s between-class separation divided by the within-class variation. The FLD classifier is one of the most computationally simple methods, requiring only one floating point multiply and add per features per candidate. This is far less than that required by the classifiers used in most other previously published lung nodule CAD systems. This may be of benefit in fielding a CAD system in a clinical setting.

Note that since the class conditional covariance estimates are added together in (12), the large number of FP examples help to effectively augment the limited number of TP examples in computing the relevant statistics. In contrast, a quadratic classifier, for example, relies on both TP and FP class conditional covariance estimates independently [33]. The blending of the class conditional covariances in (12) allows the FLD to operate robustly with the relatively few TP training examples in our application. Note also that the FLD in (12) is closely related to the GBL classifier. For the GBL classifier, it is assumed that the features for both classes follow a Gaussian distribution with different means, but identical covariance matrices. This gives rise to the following weights

$$\hat{w} = C^{-1} (\mu_{TP} - \mu_{FP}),$$

(14)

where $C$ is the common covariance matrix for the two classes. This common covariance matrix can be estimated by pooling feature vectors after the subtraction of the class conditional means. In our application the TP and FP covariances do appear to be distinct and we have a significant imbalance in training data (approximately 100 times as many FPs as TPs). Thus, feature pooling used by the GBL classifier leads to a covariance dominated by the FPs. As the results in Section 3.3 suggest, this appears to hurt performance in comparison to the FLD.

For the experimental results presented here, the statistics for all the three classifiers are estimated from the Riverain training set and the detection statistics are computed using the JRST testing set. Candidates with detection statistics above a specified threshold are declared to be tentative final detections. Since we have allowed initial candidate detections to be as close as 5 mm apart using the WMCI detector, we are likely to have multiple tentative detections on the same anatomical structure. Thus, some method for eliminating redundant tentative detections is needed. One option is to more aggressively thin the initial WMCI detections prior to the classifier stage. However, this would tend to deprive the segmentation algorithm, and ultimately the classifier, from seeing the best possible candidate for a given anatomical structure. We believe it is better to do the more aggressive detection thinning after the classifier detection statistic is computed in order to take advantage of the full suite of candidate features. Therefore, we seek to identify potentially redundant tentative detections and keep only the one with the largest detection statistic. To identify potentially redundant tentative detections, we have investigated various measures of candidate segmentation overlap as well as using the distance between candidate cue points. Our empirical studies using the training data suggest that a simple 22 mm candidate cue adjacency rule provides the most improvement in operating points near 2-4 FPs per image. Thus, for any candidate detections with cue points within 22 mm of one another, only the candidate with the largest detection statistic is presented as a final system detection.

### 3. EXPERIMENTAL RESULTS

In this section, we present a number of experimental results to demonstrate the efficacy of the proposed CAD system. We begin by describing the labeling and scoring system we employ here. Next we present results for the nodule candidate detector. We then present the overall system performance results. Finally, we compare the performance of the proposed CAD system using the three different classifiers to that reported for several other systems.
3.1. Labeling and Scoring

The truth information provided with our datasets includes cue points near the center of each nodule and size information. However, it is not entirely straightforward how nodule candidates should be labeled in chest radiograph CAD systems, and a variety of schemes can be found in the literature. For example, [4] declares a candidate a TP if its centroid is within 22 mm of the center of the lung nodule (presumably the truth cue point) for candidates in the apical and peripheral regions. They use a different 24 mm criterion for candidates in the hilum region. In [3], a candidate is labeled a TP if there is overlap between the truth circle and the initial candidate detection blob circle (obtained using a LoG multiscale candidate detector). Thus, small nodules require the candidate detection to be closer than for larger nodules. While this method is quite reasonable, it depends on the specific candidate detector used in that system. This makes a perfect benchmark comparison impossible for CAD systems using other candidate detectors. What we believe is of clinical importance is to direct the attention of the radiologist to a place on the radiograph sufficiently close to the actual lesion to prevent them from overlooking it. While no scheme is perfect, we believe that a fixed distance metric is consistent with this clinical goal. Also, since a fixed distance labeling scheme is entirely independent of the CAD system, this scheme facilitates benchmark comparisons between CAD systems.

For the purposes of clinical scoring with our system, we have adopted a 25 mm fixed distance labeling rule. This basic distance criterion has been used by Deus Technologies (now Riverain Medical) to obtain Food and Drug Administration approval of their RapidScreen® lung CAD [34]. In our labeling scheme, candidates with a cue point within 25 mm of a truth cue are considered to be potential TPs. Any candidate with a cue point beyond 25 mm from a truth cue is labeled an FP for our clinical scoring. Note that multiple candidates may lie within 25 mm of a truth cue, and we must not count multiple TPs for the same nodule. Thus, if our CAD system outputs any candidate that lies within 25 mm of a truth, we credit the CAD with detecting that particular nodule. Extra detections within 25 mm of this same nodule are ignored (they are not counted as TPs or FPs for clinical scoring purposes). Since we believe this scoring method is clinically appropriate, we have optimized aspects of our system accordingly, such as the adjacent candidate rejection procedure and WMCI candidate detector.

We have observed that not all candidates labeled as TP, using the clinical method above, constitute ideal examples for training the classifier. While these labels may be clinically appropriate for CAD scoring, we employ a second labeling scheme exclusively for the purposes of classifier training and feature selection. We refer to this secondary labeling scheme as engineering labeling. For engineering labeling, a candidate is considered a TP if the ADT candidate segmentation includes the truth cue. If multiple candidate segmentations include the same truth cue, we select only the one with the detection cue point closest to the truth cue as the only TP for that nodule. The engineering labeling for FPs is the same as that for clinical labeling. Any candidate that does not meet the engineering FP or TP labeling criterion is ignored for classifier training. The idea here is to use only the “best” TP candidates for training, and 161 of the 173 Riverain training nodules meet the engineering TP criterion. We have observed performance gains in clinical scoring when using engineering labeling for training, compared with using clinical labeling for training.

3.2. Nodule Candidate Detection

In this subsection, we examine the performance of the WMCI candidate detector. Figure 11 shows the fraction of the JRST nodules with a WMCI detection within the distance specified on the horizontal axis. These results are obtained with the 14 opaque JRST cases excluded. Notice that all of the nodules have a WMCI detection within 20 mm, and greater than 90% of the nodules have a detection within 7.5 mm. These results are obtained with a WMCI threshold of 0.5 and using the weights shown in Fig. 4. While this distance criterion is certainly relevant, it is perhaps more illuminating to know that 95.0% of the nodules have a WMCI detection within their truth radius. Also, the average closest detection distance and standard deviation are 3.70 mm and 2.95 mm, respectively. This WMCI sensitivity comes with an average of 97.2 detections per image. Since the threshold and other parameters were determined using the Riverain training data, we are pleased to see the WMCI generalize well to the JRST data set.

To put the performance of the nodule candidate detector into context, let us consider the reported operating point for candidate detectors in two previously published systems. Based on the reported figures, the LoG candidate detector in [3] has a sensitivity of 96.4% with 134 detections per image (excluding opaque cases). This sensitivity is based on the labeling method in [3] which looks for any overlap between the LoG detection circle and truth circle. Thus, with comparable sensitivity (based on WMCI detections inside the truth circle), the WMCI detector has 27.5% greater specificity (i.e., 97.2 detections per image versus 134). The ARG detector in [4] has a reported sensitivity of 92.5%
FIGURE 11. Fraction of the JRST nodules with a WMCI detection within the distance specified on the horizontal axis. These results are obtained with the 14 opaque JRST cases excluded. At this sensitivity, the WMCI detector produces an average of 97.2 detections per image.

(Using the 22 mm - 24 mm labeling in [4]) with an average of 60.2 detections per image (according to our calculations based on the reported 59.3 FPs per image). Running the WMCI at this sensitivity and using 22 mm labeling, we obtained 39.1 detections per image. This corresponds to an increase in specificity of 35% for this sensitivity operating point. However, for the overall CAD system results presented here, we use the more sensitive operating point with 97.2 detections per image in order to keep a greater number of TP candidates for the classifier stage.

3.3. Overall System Performance

The overall system performance is quantified using FROC curves. The main result with the FLD classifier is shown in Fig. 12. Here a FROC curve, showing sensitivity versus the average number of FPs per image, is included for the classifier trained exclusively on the Riverain data (including feature selection) and tested on the JRST data. The system is using the 46 features with the label (F) in Table 2 and employs adjacent candidate rejection. Scoring is done using the 25 mm clinical scoring method. This curve includes an operating point of 78.1% sensitivity with and average of 4.0 FPs per image. Figure 12 also shows a 10-fold cross-validation FROC for the system using the JRST data only. For the 10-fold cross-validation results, the JRST data are divided equally into 10 groups by subtlety rating. For this result, the FLD classifier is using 58 selected features using SFS like that described in Section 2.4, but here with JRST data. Note the improved results using the JRST 10-fold cross validation in the 2-4 FPs per image range targeted. This may suggest that the diverse Riverain training set is not fully representative of the JRST data. The independent training set, however, provides a more rigorous validation procedure and may equip the CAD system for more broad application. The remainder of the results in this section are for the CAD system trained on the independent Riverain data set and tested on the JRST data. Except where stated otherwise, these also employ the adjacent candidate rejection procedure.

A comparison of the three classifiers is provided in Fig. 13. This includes a FROC curve for the FLD, GBL, and quadratic classifiers using the features listed in Table 2. It is interesting to note that the simple FLD classifier generally yields the best results in the 2-4 FPs per image range. Note that the quadratic classifier is employing only 15 features, while the FLD uses 46. Thus, the FLD hyperplane decision boundary in 46 dimensional space seems to generalize better than the quadratic boundary in 15 dimensional space (using the features provided here). It is also interesting to
note that the FLD generalizes fairly well even using all of the features (i.e., no SFS). At 2 FPs per image the sensitivity of the system using the FLD classifier with all features is 51.2% (compared with 62.9% using 46 features). At 4 FPs per image, this sensitivity is 66% (compared with 78.1% using 46 features). In contrast, when all features are used with the quadratic classifier, sensitivity drops to 3.6% and 4.8% at 2 and 4 FPs per image, respectively. Clearly such a classifier easily becomes over-tuned to the training set with so many features. We attribute the reduced performance of the GBL classifier, compared with the FLD, to the imbalance in the available training data for the two classes. As mentioned in Section 2.5, the GBL classifier uses a pooled feature covariance matrix and it will be dominated by the FPs. In contrast, the FLD defined in (12) gives equal weight to the class conditional covariance matrices for the TPs and FPs, despite the data imbalance.

To illustrate the benefit of the post-classifier adjacent candidate rejection, FROC curves using the FLD classifier are shown in Fig. 14 with and without this procedure described in Section 2.5. Note the improvement in performance when the adjacent candidate rejection is applied. The CAD performance stratified according to nodule subtlety, pathology, and size is quantified with FROC curves in Figs. 15 - 17. It appears that size is one of the most important factors in predicting CAD sensitivity. It is interesting to note the improved results for malignant nodules versus benign nodules shown in Fig. 16. This may be partially attributable to the fact that the malignant nodules tend to be larger than the benign ones, as shown in Table 1.

A typical CAD output image is shown in Fig. 18 (JPCLN026). This was obtained for the CAD system trained exclusively on Riverain data and operating at a sensitivity of 71.4% and specificity of 3.1 FPs per image on average. Here the ASM lung mask is shown along with the CAD detection points and corresponding segmentations. This result shows one TP and three FPs. The FPs here appear to include anterior/posterior rib intersections. After inspecting all of the CAD results, it is clear that the ribs and anterior/posterior rib intersections are the main source of FPs. Similar FPs are also found along the clavicle. In some cases, the lung segmentation erroneously includes some of the diaphragm or other structures, leaving a nodule sized object that may get detected by the CAD system. Missed nodules tend to be small and have poorly defined boundaries (i.e., small intensity gradient along the perimeter).

We also conducted an experiment to investigate the value of the ADT segmentation algorithm in terms of overall system performance. In this experiment, we replaced the ADT segmentation step with one that produces a simple circle segmentation based on the WMCI detection scale (and then limit to the segmented lung field). Thus, the segmentations would essentially be the circles shown in Fig. 5. Features using these new segmentations have been recomputed and
FIGURE 13. FROC curves showing CAD performance on the JRST data for three classifiers. The CAD system has been trained exclusively on Riverain data with a separate feature selection procedure for each of the classifiers.

FIGURE 14. FROC curves showing CAD performance using the FLD classifier operating on the JRST data with and without post-classifier adjacent candidate rejection.
FIGURE 15. FROC curves showing CAD performance using the FLD classifier operating on JRST data stratified by subtlety rating.

FIGURE 16. FROC curves showing CAD performance using the FLD classifier operating on JRST data stratified by pathology (malignant or benign).
an SFS feature selection process has been implemented for the FLD classifier. This variation of the CAD system has a significantly reduced computational complexity, but we found that performance does suffer. At a specificity of 2 FPs per image, the sensitivity of the system drops from 62.9% to 55.5%. A drop in sensitivity from 78.1% to 72.3% is seen at 4 FPs per image. Thus, it does seem clear that the ADT segmentation is adding value to the system. This result differs, somewhat, from a similar experiment reported for the system in [3]. There, comparable results have been obtained with simple circular segmentations and segmentations obtained with a radial-gradient based method. We suspect that emphasis on inside-outside-separation feature groups makes our system more sensitive to the defined nodule candidate boundary.

3.4. Performance Comparison

It is difficult to make definitive comparisons with previously published CAD systems because of variability in the data sets, validation procedures, labeling and scoring methods, and different optimized and reported operating points. However, we have attempted to identify published results that use the JRST data or a database of comparable difficulty (which helps to mitigate one of variability factors). Unfortunately, labeling and scoring methods still vary in published results, as do the optimized operating points and validation procedures. These factors must be taken into consideration when interpreting the relative performance of the systems. Notwithstanding this, we believe it is important to attempt to put the performance of the proposed CAD system into the context of prior state-of-the-art systems.

Our comparison results are summarized in Table 3. Here, each row represents a published benchmark method and lists the reported operating point (specificity and corresponding sensitivity). For each row we report the sensitivity for our CAD system with the three different classifiers operating at the same specificity as the benchmark system. The results in Table 3 are for our system trained exclusively on the Riverain data and using adjacent candidate rejection and 25 mm scoring. Since we have excluded cases with nodules in the opaque regions of the lung, we must factor that in when comparing against systems that do not similarly exclude such cases. To do so, we adjust our CAD sensitivity to account for the 14 nodules in the JRST data set located in opaque regions of the radiographs. Treating these 14 nodules as false negatives (misses), our calculated sensitivity must be scaled by a factor of $\frac{140}{154}$ = 0.9091. This
FIGURE 18. Typical CAD output (JPCLN026) showing the detection cue points and corresponding ADT segmentations.

TABLE 3. CAD system performance comparison. The right most columns show the sensitivity of the proposed CAD system with 25 mm scoring operating with the reported average number of FPs per image for the benchmark system listed on the left. Except where indicated, the proposed CAD sensitivities are computed here with the 14 opaque JRST cases treated as false negatives (misses).

<table>
<thead>
<tr>
<th>CAD System</th>
<th>Average FPs Per Image</th>
<th>Reported Sensitivity</th>
<th>FLD Classifier</th>
<th>GBL Classifier</th>
<th>Quadratic Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schilham et al [3] 2006</td>
<td>2</td>
<td>51%</td>
<td>57.1%</td>
<td>50.7%</td>
<td>57.8%</td>
</tr>
<tr>
<td>Shiraishi et al [4] 2006×</td>
<td>4</td>
<td>67%</td>
<td>71.0%</td>
<td>68.1%</td>
<td>67.2%</td>
</tr>
<tr>
<td>Coppini et al [7] 2003</td>
<td>5</td>
<td>70.1%</td>
<td>80.1%</td>
<td>79.3%</td>
<td>77.1%</td>
</tr>
<tr>
<td>Freedman et al [8] 2002†</td>
<td>4.3</td>
<td>60%</td>
<td>71.4%</td>
<td>69.9%</td>
<td>67.5%</td>
</tr>
<tr>
<td>Wei et al [9] 2002</td>
<td>5.4</td>
<td>66%</td>
<td>72.8%</td>
<td>72.1%</td>
<td>70.1%</td>
</tr>
</tbody>
</table>

* All results reported in this row are excluding opaque cases. Shiraishi et al results are based on 924 cases that include the JRST cases.
† Freedman et al results are not for JRST data.

The results reported in [3, 7, 9] all use the JRST data. The data set used by Shiraishi et al [4] is the largest among of those listed in Table 3, employing 1000 cases in all (924 with the exclusion of opaque cases). These data include the JRST data, but many other cases as well. We believe that the size and subtlety distribution reported for the nodules in that extended database are sufficiently similar to those of the JRST subset to warrant a CAD performance comparison. The JRST data are not used in [8]. However, the area under the ROC curve for observers of the data set in [8] is 0.835, compared with 0.833 for the JRST data.

On whole, the CAD system proposed here appears to perform well in comparison to the benchmark methods in Table 3. A higher sensitivity is reported for the system in [9]. However, the classifier there employed 202 features and its ability to generalize across independent data sets is not evaluated. Also, note that our system is tuned for an
operating point in the range of 2-4 average FPs per image and the sensitivity reported in [9] is for 5.4 FPs per image.

4. CONCLUSIONS

Here we have presented a new lung nodule CAD system for chest radiographs. We present a detailed performance analysis of the system using the publicly available JRST database that can serve as a benchmark for future research efforts. We test the generalizability of the system by employing a diverse and independent training data set for all aspects of system tuning and training. We also present results for 10-fold cross-validation using only the JRST data. Based on the performance comparison in Section 3.4, the proposed CAD system appears to be promising. We acknowledge that differences in labeling and scoring, as well as other experimental factors must be taken into account when interpreting such comparisons. However, we believe that the current system is certainly competitive and offers some useful innovations including the WMCI detector, the ADT segmentation algorithm, and the selected feature set. In particular, the sensitivity and specificity of the WMCI appears to be quite good compared to some previously published methods, as shown in Section 3.2. As described in Section 3.3, the performance of the system using ADT segmentations is notably improved compared with using simple circular segmentations. The ADT segmentation works together with many of the inside-outside-separation feature groups in our system. Based on the feature selection process, it appears that the such inside-outside-separation features are highly salient.

Several other factors also seem to aid performance within our system. One is allowing a relatively high number of initial candidate detections and then employing a post-classifier adjacent candidate rejection procedure. In Section 3.3, we have demonstrated the improvement in performance obtained using the adjacent candidate rejection procedure. This provides an effective means to deal with the inevitable redundant detections that come with allowing a relatively large number of candidate detections. Another interesting observation relates to the choice of classifier. We have compared three classifiers and found that the FLD classifier generally yields the best results for our data. The FLD classifier generalizes well with a relatively small number of TPs and high number of features. The quadratic classifier feature selection led to the use of only 15 features, whereas the FLD uses 46. So for our system, using more features with the linear classifier provided performance gains over using fewer features with the more complex quadratic classifier. The FLD not only performs well but is very computationally simple. This is significant since processing speed is an important consideration when fielding a CAD system in a clinical setting. We have also observed improved classifier performance by training on a subset of the candidates obtained using our engineering labeling scheme as described in Section 3.1.

A valuable area of future work would be in developing a detection scheme for treating nodules in the opaque portions of the lung. As described in Section 1, these nodules appear to make up approximately 8-9% of the population in large data sets [4]. Detecting even a fraction of these efficiently, could further boost CAD performance on the JRST data set and in the field. While we are generally pleased with the performance of the proposed CAD system, valuable future work might include further optimization of system parameters such as the image resolution, operating point for the candidate detector, weights for the WMCI detector, and parameters for the ADT segmentation algorithm. More anatomical context feature may also help to improve system performance by targeting certain classes of observed FPs. This may include targeting FP resulting from anterior/posterior rib intersections, as mentioned in Section 3.3.

ACKNOWLEDGMENTS

The authors would like to thank Riverain Medical for providing the chest radiographs and corresponding truth to serve here as and independent training data set. In particular, we wish to thank Richard Burns at Riverain for preparing the data and providing advice and clinical context. We also appreciate the constructive critique of this paper provided by Steve Worrell at Riverain. We would like to thank Dr. Bram van Ginneken with Image Sciences Institute, The Netherlands, for kindly making available the manual segmentations of the JRST lung fields from his work in [27]. Some of these manual lung masks were used here for ASM training. We wish to express our appreciation to the JRST and other institutions who contributed to the creation of the JRST chest radiograph database. Finally, we wish to thank the anonymous reviewers for helping to strengthen this paper.
Performance Analysis of a New CAD for Lung Nodules

July 9, 2007

23

REFERENCES


