HANDWRITTEN ARABIC CHARACTER RECOGNITION USING MULTIPLE CLASSIFIERS BASED ON LETTER FORM

Gheith A. Abandah, Khaled S. Younis
Computer Engineering Department, University of Jordan, Amman 11942, Jordan
abandah@ju.edu.jo, younis@ju.edu.jo
Mohammed Z. Khedher
Electrical Engineering Department, University of Jordan, Amman 11942, Jordan
khedher@ju.edu.jo

ABSTRACT
Users are still waiting for accurate optical character recognition solutions for Arabic handwritten scripts. This research explores best sets of feature extraction techniques and studies the accuracy of well-known classifiers for Arabic letters. Depending on their position in the word, Arabic letters are drawn in four forms: Isolated, Initial, Medial, and Final. The principal component analysis technique is used to select best subset of features out of a large number of extracted features. We used parametric and non-parametric classifiers and found out that a subset of 25 features is needed to get 84% recognition accuracy using a linear discriminant classifier, and using more features does not substantially improve this accuracy. However, for features fewer than 25 features, a quadratic discriminant classifier is more accurate than the linear classifier. Classifiers that are parameterized for the individual four forms score better accuracy than classifiers that do not make use of this input information.

KEY WORDS
Optical character recognition, principal component analysis, classification techniques, Arabic script

1. Introduction

Optical character recognition (OCR) is computer software designed to translate images of typewritten or handwritten text into machine-editable text encoded in a standard encoding scheme (e.g. ASCII or Unicode) [1]. In the 1950s, the early solutions for recognizing typewritten English text started to appear. And now, there are many solutions for accurately recognizing typewritten Latin text. Moreover, modern operating systems include built-in support for OCR.

Although the progress in developing OCR solutions for the Arabic language is slower than the progress in developing solutions for Latin and Asian languages, there has been some success in developing solutions for recognizing typewritten Arabic text [2]. There are even some commercial products for this application that offer more than 99% accuracy in some cases.

Character recognition of text images is usually called offline OCR to distinguish it from online character recognition. In online character recognition, characters are recognized on the fly as they are entered [3], [4]. Online character recognition is widely used now in mobile devices such as tablet PCs, smart phones, and personal digital assistants. The user of these devices draws either cursive text or isolated characters using special pen on a sensitive screen. Online character recognition solutions utilize order, speed, and direction of individual pen strokes to achieve good accuracy in recognizing handwritten text.

Offline recognition of handwritten cursive text is more difficult than online recognition because the former must deal with two-dimensional images of the text after it has already been written [5]. Offline recognition of unconstrained handwritten cursive text must overcome many difficulties such as unlimited variation in human handwriting, similarities of distinct character shapes, character overlaps, and interconnections of neighboring characters. Although offline systems are less accurate than online systems, they are now good enough for specialized systems such as interpreting handwritten postal addresses on envelopes and reading currency amounts on bank checks.

Users are still waiting for reliable and accurate solutions for recognizing handwritten cursive text such as Arabic text [6]. Some researchers are experimenting with many approaches for this problem. However, only modest accuracies have been achieved in recognizing Arabic handwritten text samples collected using special forms (e.g. IFN/ENIT database [7]). In ICDAR 2005 competition, accuracies not better than 76% have been achieved on the IFN/ENIT database [8].

The research described in this paper is a contribution toward reliable OCR systems for Arabic handwritten text. The main objectives of this research are:

1. Study the performance of well-known classification techniques on a database of Arabic handwritten samples using increasing numbers of best features.

2. Evaluate the effect of using the letter form information as a basis for classification or as additional feature in classification.

This paper is organized in five sections. Section 2 is an introduction on the Arabic writing system. Section 3 describes the experimental setup used in this research; it includes a description of the used database of Arabic letter samples and the feature extraction tools. Section 4 analyzes the character recognition accuracy as function of the feature subset size using multiple classifiers. Finally, Section 5 summarizes our findings.

2. Arabic Writing System

Arabic is written from right to left and is always cursive [9]. It has 29 basic letters and eight diacritics [9], [10]. Table 1 shows the 29 letters and their various forms. Each letter has multiple forms depending on its position in the word. Each letter is drawn in an isolated form when it is written alone, and is drawn in up to three other forms when it is written connected to other letters in the word. For example, the letter Ain has four forms: *Isolated* form (ع) and *Initial*, *Medial*, and *Final* forms (ععع), respectively from right to left. Moreover, letters Hamza, Teh, and Alef have other forms, as shown in Table 1.

Within a word, every letter can connect from the right with the previous letter. However, there are six letters that do not connect from the left with the next letter (see Table 1). These letters have only the Isolated and Final forms. When one of these six letters is present in a word, the word is broken into *sub-words*. For example, the word Arabic (عربية) has two sub-words: the first sub-word consists of Initial Ain and Final, left-disconnecting, Reh; and the second sub-word consists of Initial Beh, Medial Yeh, and Final Teh.

Some letter sequences have special composite ligatures when they come in one word. For example, Lam followed by Alef is usually drawn (لا) not (ل ا), and Meem followed by Hah is often drawn (محمد) rather than (محمد).

3. Experimental Setup

Our experimental setup comprises a database of handwritten Arabic samples and feature extraction tools [11]. These elements are described in the following subsections.

Table 1: Arabic Letters and Their Four Forms

No	Letter Name [a]	Isolated Form	Initial Form	Medial Form	Final Form
1	Hamza [b]	ء	ئ	ـئـ	أ
2	Beh	ب	بـ	ـبـ	ب
3	Teh [c]	ت ة	تـ	ـتـ	ة ت
4	Theh	ث	ثـ	ـثـ	ث
5	Jeem	ج	جـ	ـجـ	ج
6	Hah	ح	حـ	ـحـ	ح
7	Khah	خ	خـ	ـخـ	خ
8	Dal [d]	د	-	-	د
9	Thal [d]	ذ	-	-	ذ
10	Reh [d]	ر	-	-	ر
11	Zain [d]	ز	-	-	ز
12	Seen	س	سـ	ـسـ	س
13	Sheen	ش	شـ	ـشـ	ش
14	Sad	ص	صـ	ـصـ	ص
15	Dad	ض	ضـ	ـضـ	ض
16	Tah	ط	طـ	ـطـ	ط
17	Zah	ظ	ظـ	ـظـ	ظ
18	Ain	ع	عـ	ـعـ	ع
19	Ghain	غ	غـ	ـغـ	غ
20	Feh	ف	فـ	ـفـ	ف
21	Qaf	ق	قـ	ـقـ	ق
22	Kaf	ك	كـ	ـكـ	ك
23	Lam	ل	لـ	ـلـ	ل
24	Meem	م	مـ	ـمـ	م
25	Noon	ن	نـ	ـنـ	ن
26	Heh	ه	هـ	ـهـ	ه
27	Waw [d]	و	-	-	و
28	Alef [d, e]	ى ا	-	-	ى ا
29	Yeh	ي	يـ	ـيـ	ي

[a] Letter names are as in the Unicode Standard.

[b] In addition to these forms, the Hamza has the Isolated forms (أ إ ؤ) and the Final forms (ـئ ـؤ).

[c] Teh has open forms (ت) and closed forms (ة).

[d] Letters that do not connect from the left.

[e] Alef has straight forms (ا) and curly forms (ى).

3.1 Database

Our database of handwritten Arabic samples was collected from 48 persons. These persons were selected to represent various age, gender, and educational background groups. The samples were collected by asking the participants to write on a blank paper one page of Arabic text. This text was carefully selected so that it contains all Arabic letter forms.

We have extracted from these page samples about 440 collections of individual words, sub-words, and letter forms. Each collection comprises 48 samples from 48 different persons. Fig. 1 shows the collection of 48 samples of the Isolated Ain form.

Fig. 1: A Set of 48 Samples of the Isolated Ain Form

We used in this research 104 collections of letter forms: 30 isolated forms, 22 initial forms, 22 medial forms, and 30 final forms. These collections contain all forms shown in Table 1 devoid of Hamza forms.

3.2 Feature Extraction Tools

To allow easy experimentation of OCR algorithms on this database of handwritten Arabic samples, we developed a desktop application using Microsoft Visual Studio C++. This application is an expandable tool that allows developers to easily add various preprocessing, feature extraction, and recognition OCR routines. It enables the user to select the order of the OCR routines to be applied on the sample collections. This application allows the user to visualize the results of preprocessing routines and obtain the results of the feature extraction and recognition routines. Fig. 2 shows this application with its dialog box for selecting what routines to apply on the collection of the Isolated Beh samples.

This application also features batch processing where the selected routines can be applied on multiple sample collections. The results of feature extraction routines can be exported from this application into an Excel spreadsheet.

We have implemented in this application feature extraction routines for extracting 95 features [12], [13], [14]. The details of these features are in [15]. We start by detecting the secondary parts of the Arabic letters and extracting features from these parts. Then we remove the secondary parts and extract additional features from the raw main body, the main body's skeleton, and main body's boundary. These routines where applied to the 104 collections of letter forms and the 95 feature vectors where used to find the recognition accuracy as described in the next section.

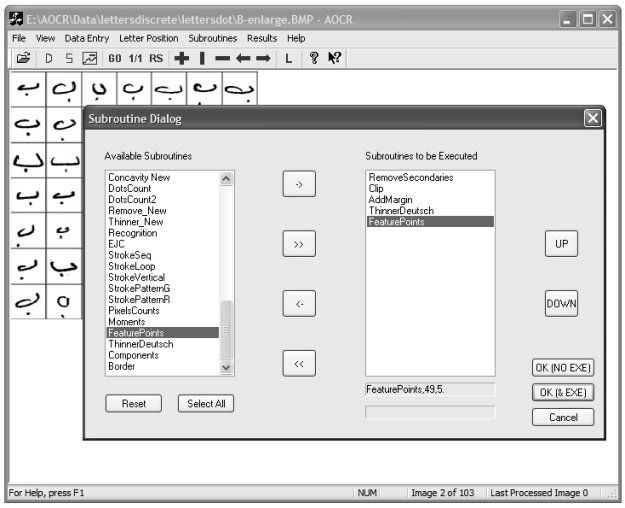

Fig. 2: Arabic OCR Application Tool and its Routine Selection Dialog Box

The used 95 features are summarized as follows:

Secondary components features: More than half the Arabic letters are composed of main body and secondary components. The secondary components are letter components that are disconnected from the main body. We find the type and position of the secondary components with respect to the main body.

Main body features: include the size features area, width, and height; width to height ratio; eight pixel distribution ratios; normalized center of mass; seven normalized central moments; letter orientation; roundness ratio; and number of main body loops.

Skeleton features: are features found from the thinned image of the main body and include the numbers of vertical and horizontal crossings and the numbers of end points, branch points, and cross points.

Boundary features: are features found from the outer contour of the main body and include number of boundary pixels, perimeter length, perimeter to diagonal ratio, compactness ratio, bending energy, and sixty elliptic Fourier descriptors.

4. Recognition Accuracy

To find how many features are needed to achieve good character recognition accuracy, we used the five classifiers described below. We found the classification accuracy as a function of the number of features used in the classifier.

We applied the Principal Component Analysis (PCA) [16] as a preprocessing step to transform the data to a new space where the features are uncorrelated. In this space, the features are ordered in decreasing variance order such

that the first transformed feature accounts for the most variability in the data. Hence, PCA overcomes the problems of high-dimensionality and co-linearity. Then we choose the first q transformed features to be used in classification. These are the best q transformed features.

For multivariate normal class densities, Bayes rule generally becomes a quadratic rule; hence the name Quadratic Discriminant Analysis (QDA) classifier. However, when the class densities are assumed to have the same covariance matrix, the discriminant classifier is linear; therefore it is called Linear Discriminant Analysis (LDA) classifier [17].

If no correlation between features is assumed, i.e., the covariance matrix is diagonal, and we get Naive Bayes rule. We can estimate a diagonal sample covariance matrix for each class, and this yields the Diagonal Quadratic Discriminant Analysis (DQDA) classifier. Alternatively, we estimate a diagonal common covariance matrix that yields a Diagonal Linear Discriminant Analysis (DLDA) classifier [17].

The k-nearest neighbor (kNN) classifier is one of the simplest and most attractive nonparametric classifiers. However, it faces serious challenges when patterns of different classes overlap in some regions in the feature space, especially if we use cross-validation [18].

Each classifier was tested using the 10-fold cross-validation to find the classification accuracy [19], [20]. Fig. 3 shows the classification accuracy of these five classifiers. The QDA classifier has the best classification accuracy when using up to 20 features. This is expected because the nonlinear decision surface obtained with the QDA classifier separates classes better than the linear surface. The QDA classifier's accuracy falls with features more than 20 because the number of free parameters to be estimated approaches the number of available training data points. For $q > 40$, it is not possible to estimate a sample covariance matrix because the number of training samples is not sufficient.

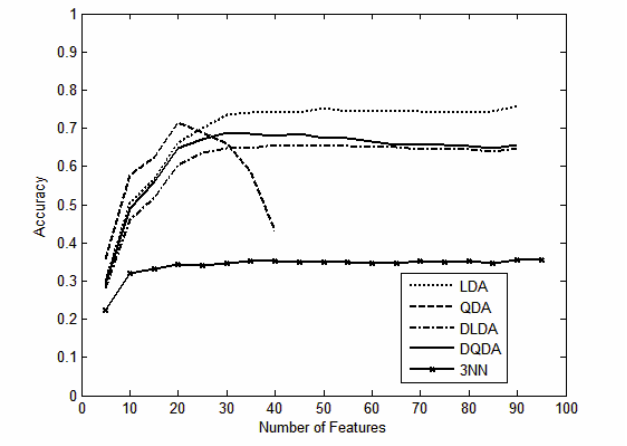

Fig. 3: Classification Accuracy using Five Classifiers

The accuracy of the DQDA classifier is lower than the QDA for small number of features because the implied shape of the diagonal covariance matrix of orthogonal features does not describe the data accurately. However, with fewer parameters to estimate; q instead of q^2, the DQDA is able to tie the QDA at q=28 and continues to give good results afterwards. However, the DQDA's accuracy slightly falls with more features due to the increased number of parameters to estimate with limited number of training samples.

The linear discriminant surface offered by the LDA classifier initially gives bad results compared to the nonlinear discriminant surface offered by the QDA classifier. However, as q increases, the number of free parameters to estimate q^2, from the whole dataset is easily computed and the LDA's accuracy improves with the addition of more features.

Similar to LDA, the DLDA classifier estimates the global covariance matrix using the whole dataset, so there is no accuracy degradation as q increases. However, because the correlation among features is not taken into account, the direction of linear decision surface does not separate the classes as good as in LDA.

The 3NN classifier gives 90% accuracy for any q using resubstitution. However, its accuracy was much worse when we used cross validation. We tried k=1 and k=3 and found that k=3 gives better results. However, its performance is much worse than the parametric classifiers due to the existence of large number of classes and class overlap.

Fig. 4 shows the QDA classification accuracy for the four letter forms taken separately. The four curves climb fast when the number of features is increased, then the curves fall down. The Final form has the highest accuracy of 86% and the Isolated form has the lowest accuracy of 76%. The curves of the Final and Initial forms rise faster than the curves of Medial and Isolated forms.

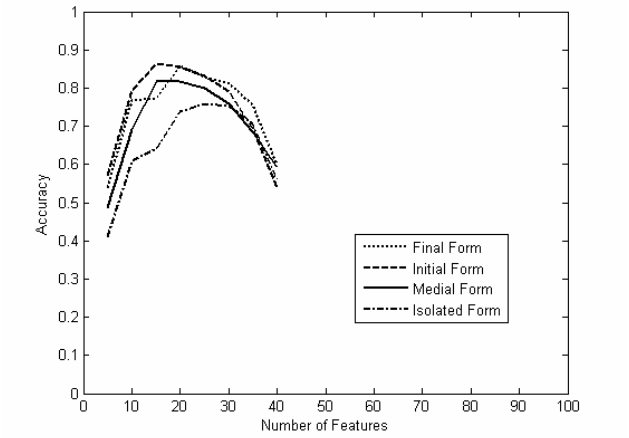

Fig. 4: QDA Classification Accuracy for the Four Forms Taken Separately

Fig. 5 shows the LDA classification accuracy for the four letter forms taken separately. The four curves also climb fast when the number of features is increased from 5 to 20, then the curves almost cease to rise after using 30 features. The Final form has the highest accuracy of 91% and the Isolated form has the lowest accuracy of 84%. The curves of the Final and Initial forms rise faster than the curves of Medial and Isolated forms, and the former two curves reach higher accuracies.

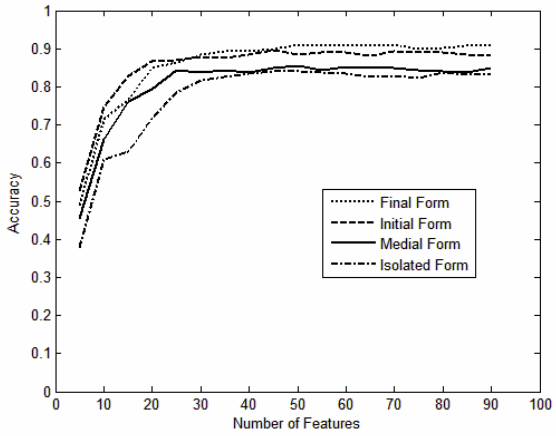

Fig. 5: LDA Classification Accuracy for the Four Forms Taken Separately

From Figures 4 and 5, we conclude that Final and Initial forms are easier to recognize than Medial and Isolated forms.

Fig. 6 shows the curve of the average of these four curves which has a maximum of 87%. The lower curve in this figure shows the LDA classification accuracy when the classifier is used to recognize all four forms. With a maximum of 76%, the lower curve indicates that we get lower accuracy when the classifier is required to recognize all forms.

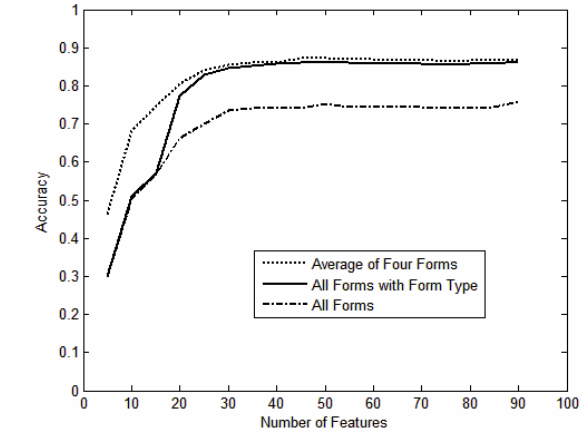

Fig. 6: LDA Classification Accuracy for the Average of the Four Forms Taken Separately, All Forms with the Form Type Feature, and All Forms without the Form Type.

The third curve shows the LDA classification accuracy for all forms with one additional feature. This feature is a categorical feature that specifies the letter form. This information is readily available from the segmentation process that comes before feature extraction and letter classification. Although this curve rises almost as high as the average curve (maximum of 86%), it rises slower than the average curve.
This implies that, with fewer features, four classifiers each tuned for one form are more accurate than one classifier with context (letter form) information.

Using best set of 30 features, we examined the classification accuracy of every letter. We used the data of one of the four experiments that produced the solid curve in Fig. 6. The classification accuracy was 100% for easy to recognize letters such Alef, Lam, and Reh and the classification accuracy was as low as 56% for the worst letter. Table 2 shows the worst 10 letters recognized and their corresponding classification accuracy.

Table 2: Worst Ten letters Recognized

No	Character	Accuracy	Often Mistaken For
1	Isolated Qaf (ق)	56%	ت ف
2	Isolated Teh (ت)	58%	ف ث
3	Isolated Theh (ث)	63%	ش ت
4	Isolated Feh (ف)	63%	ت ن ق
5	Medial Hah (ـحـ)	65%	ـعـ
6	Medial Feh (ـفـ)	67%	ـقـ ـغـ
7	Medial Ain (ـعـ)	69%	ـمـ ـصـ ـحـ
8	Medial Ghain (ـغـ)	69%	ـقـ ـخـ ـفـ
9	Medial Heh (ـهـ)	69%	ـكـ ـصـ
10	Initial Theh (ـثـ)	71%	ـقـ ـفـ ـطـ ـتـ

Nine of the worst 10 letters are of the Isolated or Medial forms, consistent with the results shown in Fig. 5. Letters with dots above the main body tend to have low classification accuracy because the variations in drawing the dots give inaccuracies in extracting the important secondary type feature. Note that some Medial letters that have loops such as Ghain (ـغـ) and Feh (ـفـ) have subtle difference in the way the loop is drawn and, consequently, they have low classification accuracy.

5. Conclusion

Among the 5 studied classifiers, the QDA classifier gives best accuracy with 20 features or less. The LDA classifier gives best accuracy with more features. The DQDA and DLDA classifiers' accuracies are about 10% lower than the LDA classifier's accuracy. The 3NN classifier has very low accuracy.

Using four classifiers each tuned for one letter form gives about 11% better accuracy than using one classifier for all forms. And, with small number of features, the four classifiers are also more accurate than one classifier with the additional letter form information.

We have noticed that final and initial forms are easier to recognize than medial and isolated forms. The highest recognition accuracy we have achieved is 87% using LDA classifier. This accuracy was limited by low recognition of some medial and isolated forms. Better feature extraction techniques are needed for letters with dots above the main body because of the variations in drawing these dots. Also better classification techniques are needed for these forms.

Acknowledgements

This work was supported in part by the Deanship of Academic Research, The University of Jordan.

References

[1] S. Mori, H. Nishida, & H. Yamada, *Optical character recognition* (New York, NY: John Wiley & Sons Inc., 1999).

[2] M. Khorsheed, Off-line Arabic character recognition - a review, *Pattern Analysis & Applications, 5*(1), 2002, 31-45.

[3] R. Plamondon & S. Srihari, On-line and off-line handwriting recognition: a comprehensive survey, *IEEE Trans. on Pattern Analysis and Machine Intelligence, 22*(1), 2000, 63-84.

[4] S. Al-Emami & M. Usher, On-line recognition of handwritten Arabic characters, *IEEE Trans. Pattern Analysis and Machine Intelligence, 12*(7), 1990, 704-710.

[5] N. Arica & F. Yarman-Vural, Optical character recognition for cursive handwriting, *IEEE Trans. Pattern Analysis and Machine Intelligence, 24*(6), 2002, 801-813.

[6] L. Lorigo & V. Govindaraju, Offline Arabic handwriting recognition: a survey, *IEEE Trans. on Pattern Analysis and Machine Intelligence, 28*(5), 2006, 712-724.

[7] M. Pechwitz, S. Snoussi Maddouri, V. Märgner, N. Ellouze, & H. Amiri, IFN/ENIT-database of handwritten Arabic words, *Proc. 7th Colloque Int'l Francophone sur l'Ecrit et le Document, CIFED 2002*, Hammamet, Tunis, 2002, 129-136.

[8] V. Märgner, M. Pechwitz, & H. ElAbed, ICDAR 2005 Arabic handwriting recognition competition, *Proc. Int'l Conf. Document Analysis and Recognition*, 2005, 70-74.

[9] A. Amin, Arabic character recognition, *in Handbook of character recognition and document image analysis, H. Bunke and P. Wang, Eds.* (World Scientific, 1997), 397-420.

[10] G. Abandah & F. Khundakjie, Issues concerning code system for Arabic letters, *Dirasat Engineering Sciences Journal, 31*(1), 2004, 165-177.

[11] G. Abandah & M. Khedher, Printed and handwritten Arabic optical character recognition - initial study, *A report on research supported by The Higher Council of Science and Technology*, Jordan, Aug 2004.

[12] R. Jain, R. Kasturi, & B. Schunck, *Machine vision* (New York, NY: MacGraw-Hill Inc., 1995).

[13] O. Trier, A. Jain, & T. Taxt, Feature extraction methods for character recognition - a survey, *Pattern Recognition, 29*(4), 1996, 641-662.

[14] F. Kuhl & C. Giardina, Elliptic Fourier features of a closed contour, *Computer Graphics and Image Processing, 18*(3), 1982, 236-258.

[15] G. Abandah, M. Khedher, & K. Younis, Evaluating and selecting features for recognizing handwritten Arabic characters, *Technical Report, Computer Eng. Dept.*, The University of Jordan, http://www.abandah.com/gheith, Sep 2007.

[16] K. Fukunaga, *Introduction to statistical pattern recognition* (San Diego, CA: Academic Press Professional Inc., 1990).

[17] R. Duda, P. Hart, & D. Stork, *Pattern classification, 2nd ed.* (Wiley Interscience, 2000).

[18] B. Dasarathy, *Nearest neighbor (NN) norms: NN pattern classification techniques* (IEEE Computer Society Press, 1990).

[19] M. Stone, Cross-validatory choice and assessment of statistical predictions, *J. Royal Statist. Soc., 36*(2), 1974, 111-147.

[20] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proc. 14th Int'l Joint Conf. on Artificial Intelligence*, 1995, 1137-1143.