JAFROC Analysis Software Version 2.3a

Overview

JAFROC software is intended to analyze free-response or "FROC" data (1-5). It should run on any WINDOWS computer. The software does not have an install or setup file as it runs under a separate program called "IDL", which you do not need to purchase. See "running the software" section below for details on how to run JAFROC.

FROC data consists of mark-rating pairs, where a mark is an image location that was considered suspicious (e.g., for cancer) and the rating is the associated confidence level or degree of suspicion. By adopting a proximity criterion, each mark is classified by the investigator as a true positive (TP) - if it is close to a real lesion - or a false positive (FP) otherwise. The choice of proximity criterion is up to the investigator.

JAFROC compares the performances of one or more readers interpreting the same set of images or cases in two or more modalities. Readers could be radiologists, mammographers, or computer algorithms designed to find lesions, like CAD (computer aided detection). It is intended to answer questions like whether or not CAD improves diagnostic performance, or whether digital mammography is better than conventional mammography, etc.

JAFROC is specifically designed for studies that generate localization information. If localization is not a factor, you should use ROC methods. Background information on FROC (6-11) including some applications (12-15) of JAFROC can be found in the cited references in this document.

JAFROC calculates a figure-of-merit defined as the probability that a lesion is rated higher than any mark on a normal image. It is estimated using the non-parametric Mann-Whitney-Wilcoxon U-statistic applied to the lesion ratings and the highest rating on normal images, where unmarked lesions and unmarked normal images are assigned the negative infinity rating (operationally this is -2000.0 in the program, so your ratings have to be bigger than this value).

JAFROC calculates 95% confidence intervals for the inter-modality figure-of-merit differences. If a confidence interval includes zero the modalities are not different at the 5% level and otherwise they are significantly different.

JAFROC requires approximately equal numbers of normal and abnormal cases, typically 50 or more of each. It is best applied to human observer FROC studies. Future enhancements will include the ability to analyze data sets with abnormal images only and allow a bootstrapping option for significance testing (currently the jackknife is used). For CAD data the IDCA (16) or non-parametric methods (17, 18) should be used, software for which is under development by Dr. Chakraborty.

A note on the figure or merit: users of JAFROC are sometimes surprised that a modality (mod-A) on which the observer marks *some* of the lesions without marking any normal image does not yield a perfect figure of merit ($A_{1J} = 1$) and conversely a modality (mod-B) on which the observer marks *some* of the normal images and does not mark any of the lesions does not yield a zero figure of merit ($A_{1J} = 0$). In fact it is observed that $0 < (A_{1J})_{\text{mod-B}} < (A_{1J})_{\text{mod-A}} < 1$. The observer who marks only lesions is obviously better than the observer who marks only normal images. The operative word is "some". If the observer marked *every* lesion and did not mark any normal image, the figure of merit would be unity and if the observer marked *every* normal image and did not mark any lesion, the figure of merit would be zero. When neither of these conditions is true, each unmarked lesion and unmarked normal image comparison yields a tie, i.e., contributes 0.5 to the Mann-Whitney-Wilcoxon

U-statistic kernel function, and for mod-A each marked lesion vs. unmarked normal image comparison contributes unity and for mod-B each marked normal image vs. unmarked lesion comparison contributes zero.

The reported values will be slightly different from 2.3 because Hillis-Berbaum pseudo-value normalization is now implemented (2/15/2008). Also the degrees of freedom for the F-statistic are reported, as required by some journals.

Acknowledgements

Dr. Chakraborty is grateful to Hong-Jun Yoon, MS, for implementation of the JAFROC software and to many JAFROC users, especially Mr. Glen Maitz, for bringing issues to our attention. This work is currently supported by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB006388. Future enhancements that I am working on, some quite sophisticated compared to JAFROC, will require continued funding. If you find this software useful I would appreciate a letter of support, so that I may be able to justly claim that my work is making a difference.

Data Format

Text file format: historically this came first; it is a difficult data entry method and is retained only for compatibility with previous software versions. See file **TextFileFormat.doc** for description of this data format.

The Excel format: A sample Excel format data set is provided in **DataFile.xls** to be found in the **USER_IO_FILES** directory. It consists of three worksheets embedded within a single workbook. The worksheets ***must*** be named **Truth, TP** and **FP**. The first row of each worksheet is reserved for data labels. **Truth** denotes ground truth information for each image. **TP** = the ratings "true positives", i.e., lesions that are correctly localized (termed lesion localizations in the recent papers); **FP** = ratings for "false positive", i.e., ratings of marked normal regions (termed non-lesion localizations in the recent papers).

Truth worksheet

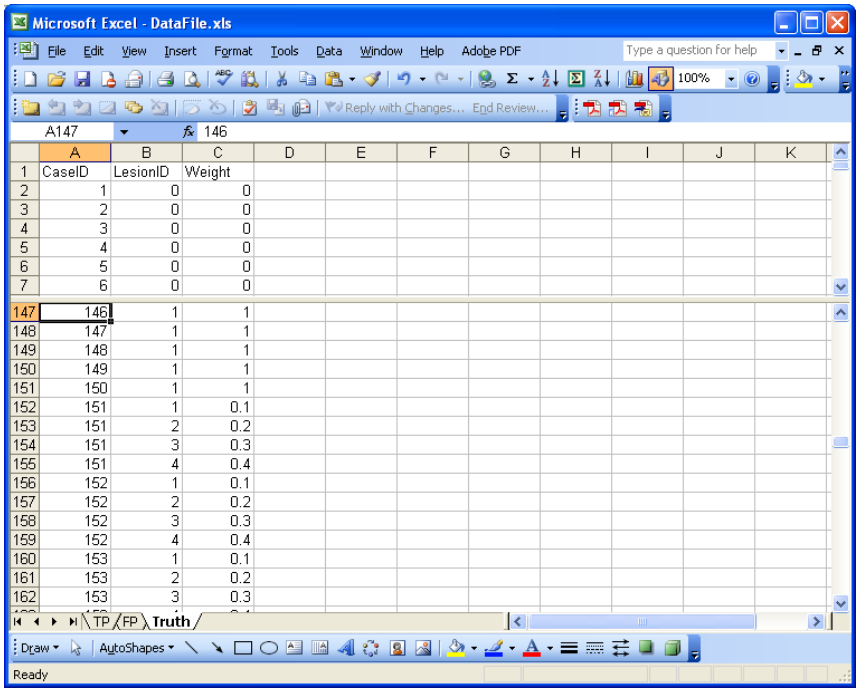


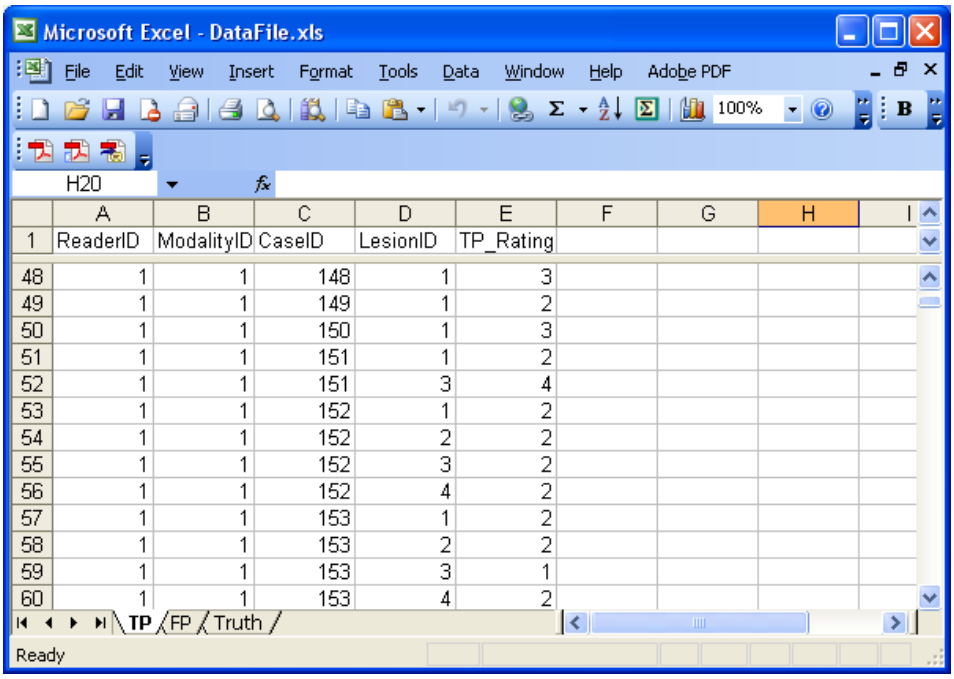
Figure 1: Truth worksheet example in **DataFile.xls**

Table 1: This table describes the contents of the truth worksheet shown in Fig. 1.

Label	**Meaning**
CaseID	A **string** (**first 8 characters must be unique,** maximum 256 characters, spaces are OK) identifying the case, or patient or image; Examples: 1, 2, 3, …, 60 case 1, case2, …, case60 0026_LMLO, 0073_LCC, ...
LesionID*	This **integer** field tells the program how many lesions are in a particular image and uniquely identifies them. If this number is zero the case is considered to be normal. If a case has multiple lesions the numbers must be unique integers but they don't have to be sequential. Examples: 1, 3, 2 and 9 for an abnormal case with 4-lesions 1 and 2 for an abnormal case with 2-lesions 0 for a normal case
Weight*	Dummy column: this field must be filled in with **zeroes**.

* These are not used per-se but must be included for future enhancements to work. The original figure-of-merit definition (6) included a weighting factor for each lesion which allowed for the possibility that detection of lesions in an image might have varying clinical significances. It was found subsequently (19) that this definition failed the NH test when jackknifing was used for significance testing. An alternate definition was proposed which involved performing a weighted average of the ratings for each abnormal image and comparing it to the highest rated noise on normal images (19). This had correct NH behavior but since the number of comparisons is smaller the statistical power is lower. The version of JAFROC (2.0) implementing lesion weighting has been withdrawn. A version employing the original definition but using bootstrapping for significance testing is in preparation. Bootstrapping should correct the NH behavior without sacrificing power.

TP Worksheet



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	TP_Rating				
48	1	1	148	1	3				
49	1	1	149	1	2				
50	1	1	150	1	3				
51	1	1	151	1	2				
52	1	1	151	3	4				
53	1	1	152	1	2				
54	1	1	152	2	2				
55	1	1	152	3	2				
56	1	1	152	4	2				
57	1	1	153	1	2				
58	1	1	153	2	2				
59	1	1	153	3	1				
60	1	1	153	4	2				

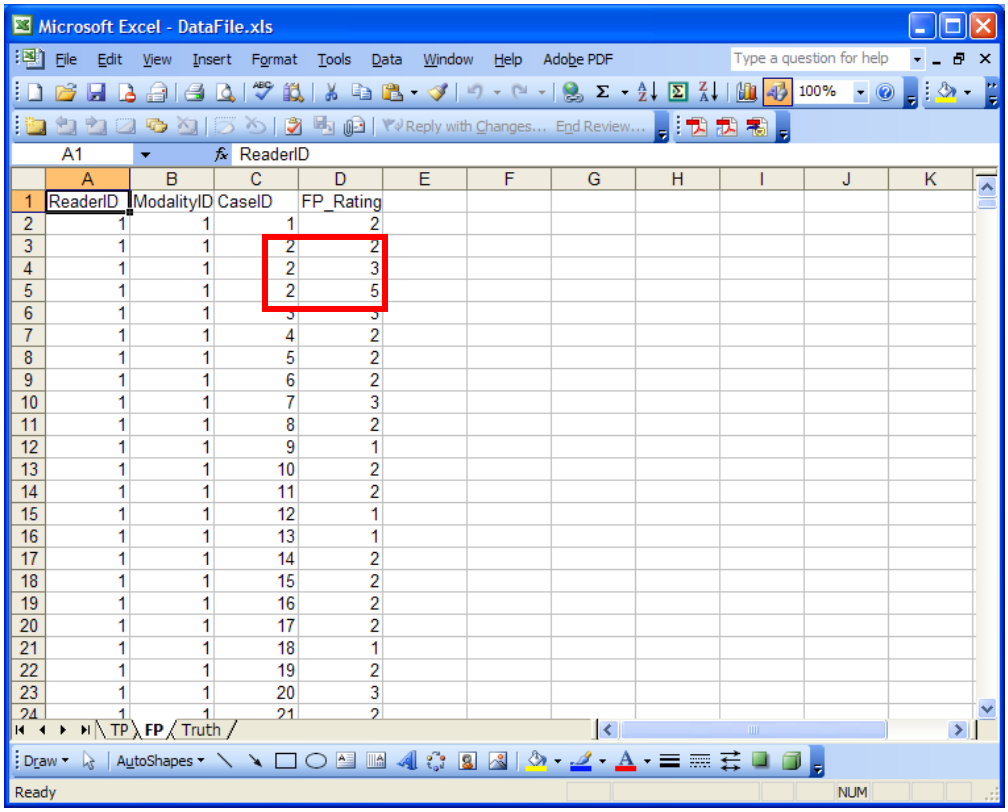
Figure 2: TP worksheet example in **DataFile.xls**

Table 2: This table describes the contents of the TP worksheet shown in Figs. 2.

Entry	**Meaning**
ReaderID	This **string** (**first 8 characters must be unique**, spaces OK, maximum length 256 characters) identifies the reader. Examples: 1, 2, 3, 4, ... reader 1, reader 2, reader 3, ... Dr. Tom, Dr. Dick, Dr. Harry, Resident 1,...
ModalityID	This **string** (**first 12 characters must be unique**, spaces OK, maximum length 256 characters) field identifies the modalities. Examples: 1, 2, 3, ... mod1, mod2, mod3, ... digital mammo, conventional mammo, breast tomosynthesis, ...
CaseID	As in Table 1
LesionID	As in Table 1
TP_Rating*	This **integer or floating-point** value (> -2000.0) is the rating assigned to a particular TP mark. **Higher numbers represent greater confidence** that the location is a lesion. For images with multiple lesions the ratings must correspond to the LesionIDs in the truth table. Examples: 1,4,3,2,... -3.1, 1, 2.5, 0,...

* At one time I disallowed zero ratings; to include output of CAD algorithms I now allow any reasonable value (greater than -2000, since the latter is used internally by the program).

FP Worksheet



	A	B	C	D	E	F	G	H	I	J	K
1	ReaderID	ModalityID	CaseID	FP_Rating							
2	1	1	1	2							
3	1	1	2	2							
4	1	1	2	3							
5	1	1	2	5							
6	1	1	3	3							
7	1	1	4	2							
8	1	1	5	2							
9	1	1	6	2							
10	1	1	7	3							
11	1	1	8	2							
12	1	1	9	1							
13	1	1	10	2							
14	1	1	11	2							
15	1	1	12	1							
16	1	1	13	1							
17	1	1	14	2							
18	1	1	15	2							
19	1	1	16	2							
20	1	1	17	2							
21	1	1	18	1							
22	1	1	19	2							
23	1	1	20	3							
24	1	1	21	2							

Figure 3: FP worksheet example in **DataFile.xls**

Table 3: This table describes the contents of the FP worksheet in Fig. 3.

Entry	**Meaning**
ReaderID	As in Table 2.
ModalityID	As in Table 2.
CaseID	As in Table 1.
FP_Rating	The rating assigned to a particular FP mark. Otherwise as in Table 2. Multiple false positives on an image are indicated on multiple lines. See for example rows 3,4,5 in Fig. 3, which says reader 1 produced 3 false positives in modality 1 for case 2.

Running the software

1. The free IDL Virtual Machine (IDL VM) is available at http://www.ittvis.com/idlvm . Download and install the latest version of IDL VM.
2. Fig. 4 shows the (typical) distribution files.

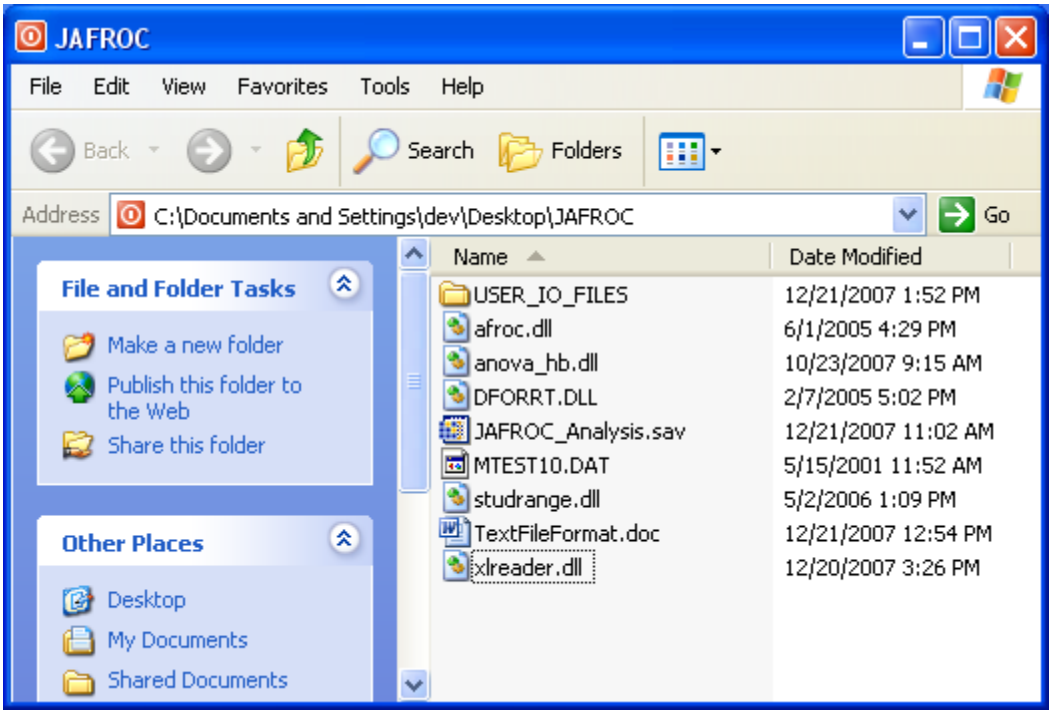


Figure 4: JAFROC distribution files

3. The **JAFROC_Analysis.sav** file and the **DLL** files must be in the same directory. It is convenient to have all input and output files in a subdirectory (called USER_IO_FILES in the distribution), see Fig. 4. This can be achieved by opening the zip file and extracting all files to the desktop, or any desired location.

4. Open **JAFROC_Analysis.sav**. The following screen may appear (Fig. 5). If it does click the "**click to continue**" button.

Figure 5: IDL Virtual Machine splash window

5. Select the **USER_IO_FILES** folder, Fig. 6 (a) and click **Open**.

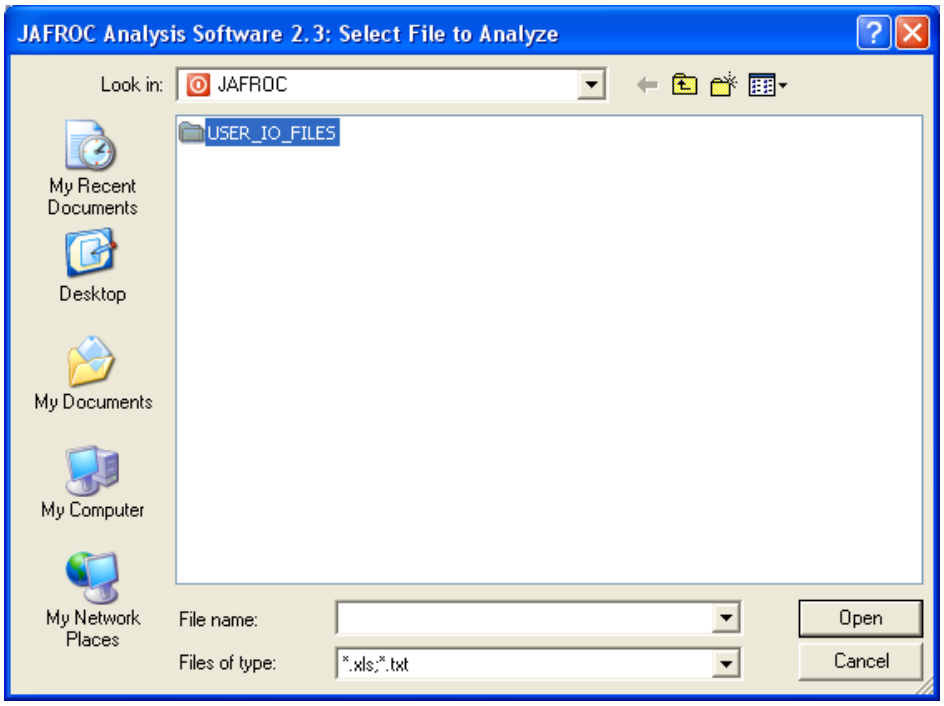


Figure 6 (a)

6. Select the **Excel** or **Text** file containing the data to be analyzed, Fig. 6 (b) and click **Open**.

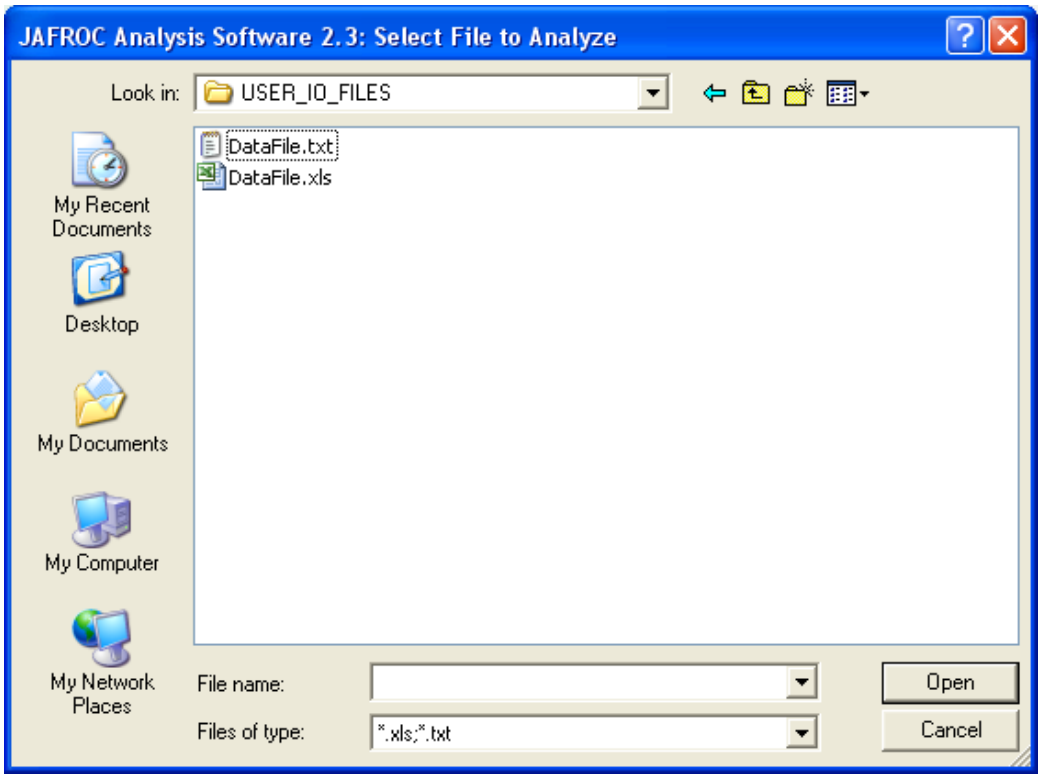


Figure 6 (b)

7. **Do not select an output file as the input file. If you do, an error screen, see Fig. 6(c), will result.**

Figure 6(c): Incorrect input file was selected

8. The results of the analysis are saved in the file **DataFile_JAFROC.txt** in folder **USER_IO_FILES and** displayed on the screen (see Fig. 7). The output file name is constructed from the input file name by appending the string "_JAFROC".

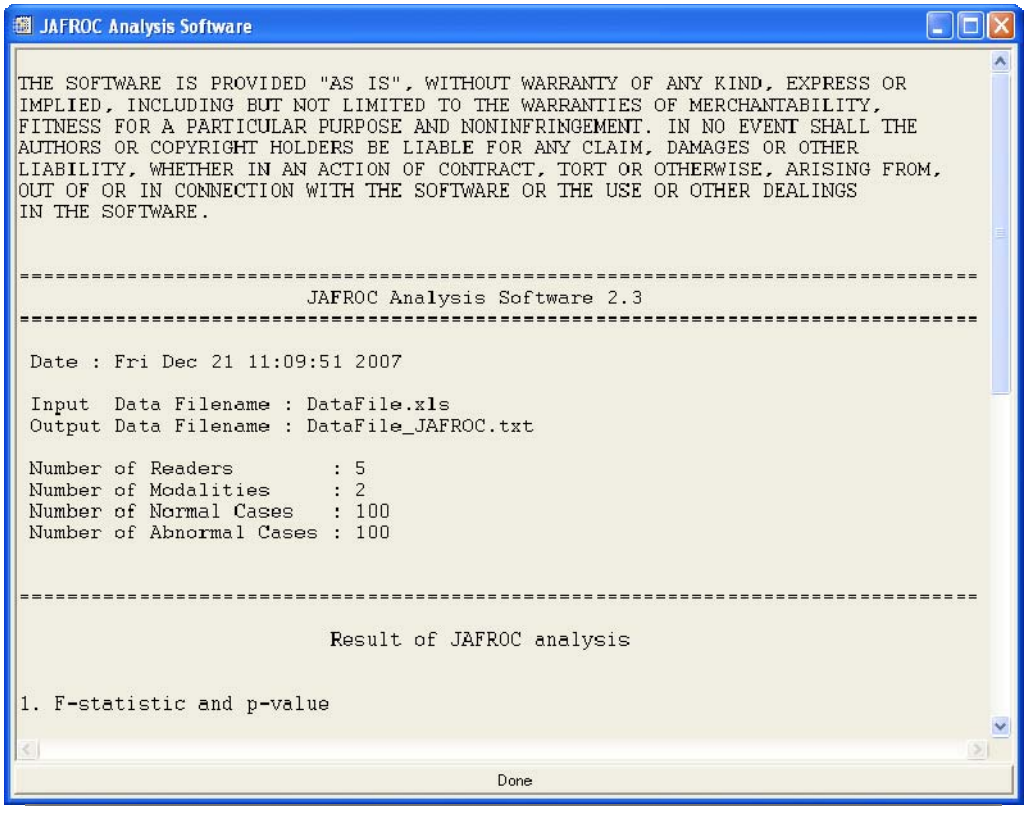


Figure 7: Report screen of JAFROC Analysis Software

9. Click **Done** (see bottom of frame in Fig. 7) to exit the software.

10. The next page shows the output using the supplied **DataFile.xls** data file:

==========

JAFROC Analysis Software 2.3a

==========

Date : Fri Feb 15 13:05:49 2008

Input Data Filename : ExcelDataFile.xls
Output Data Filename : ExcelDataFile_JAFROC.txt

Number of Readers : 5
Number of Modalities : 2
Number of Normal Cases : 100
Number of Abnormal Cases : 100
Significance level of test (alpha) = 0.05

==========

Result of JAFROC analysis

1. F-statistic, p-value and degrees of freedom (numerator, denominator)

F-statistic : 0.381621
p-value : 0.546871
DF (num,den) : (1 , 13.696902)

N.B. If the p-value is less than 0.05 there is a significant difference
between at least one pair of modalities

2. FOM = JAFROC Figure of Merit
(a) Reader Averaged FOMs and confidence intervals
Modality 1 : 0.690064, and 95% CI= [0.646018 , 0.731538]
Modality 2 : 0.710672, and 95% CI= [0.672858 , 0.746296]

(b) FOMs for modalities (columns) and readers (rows)

Modality :		1	2
Reader ID	1 :	0.665440	0.648600
Reader ID	2 :	0.686720	0.696340
Reader ID	3 :	0.744720	0.706680
Reader ID	4 :	0.711820	0.760160
Reader ID	5 :	0.641620	0.741580

3. Inter-modality differences and 95% confidence intervals (CI)
If CI DOES NOT include zero, the differences is significant
at alpha = 0.05

FOM(Modality 1) - FOM(Modality 2) = -0.0206080
and 95% CI= [-0.0926807 , 0.0514637]

* The CI may be "mildly" inconsistent with the p-value
* This is due to sampling effects and is quite normal.

Explanation of the output

There are three parts to the result of JAFROC analysis.

1) The F-stat, degrees of freedom and p-value, and whether or not the test revealed one or more significant differences between the modalities.
2) FOM = JAFROC Figure-of-merit
   a) Reader Averaged FOMs and confidence intervals.
   b) FOMs for modalities (columns) and readers (rows).
3) Inter-modality differences and 95% confidence intervals: For all pairings of modalities, the differences in reader-averaged FOMs and the corresponding 95% confidence intervals. If any one of the CI's does not include zero, then the corresponding difference is significant. The CI may be "mildly" inconsistent with the p-value (e..g., CI barely includes zero, but p-value is less than 0.05). This is due to sampling effects and is quite normal.

Important notes

We have tried to catch as many data entry errors as we can. But there are two errors that we cannot catch using the Excel format. If there is a misspelling of a reader name, then this reader will appear as an unexpected additional reader in the output in section 2 (b). So please check output to make sure you don't have any unexpected readers. **The reader names must be exactly correct.** The same applies to a misspelled modality name. **The modality names must be exactly correct.**

History of JAFROC versions (most recent on top)

2.2→2.3

1. HB pseudo-value normalization implemented (2/15/2008).
2. F-statistic and degrees of freedom are provided.
3. An error affecting the section 2(a) of the output was fixed (the reader-averaged modality FOMs and the corresponding 95% confidence intervals). We thank the user who brought this to our attention.
4. The figure-of-merit reverts to that in the 2004 Med Phys paper. Lesion weighting removed.
5. The ratings can be arbitrary numbers (but > -2000).
6. More conditions resulting in 'crashes' identified and fixed (e.g., error internal to ANOVA routine).

2.1→2.2

7. **Better error checking:** Because of the variable number of data-units per case, the JAFROC data format is necessarily more complex than that for an ROC study, and is more prone to user input errors. As noted by some users in some cases a user-input error can cause version 2.1 to "crash". This update addresses some of these issues and the program provides more helpful feedback which should allow the user to correct the input errors. Please bring other "crash" issues to my attention – send me the data file.
8. **The zero-rating is not allowed in the JAFROC data format**: if the observer marks a location it is because there was something suspicious about it, and the confidence level no matter how small cannot be zero. This error was not being detected in some cases. The error checking in the new code will catch this condition. The sample Excel sheet has been replaced - the old Excel data sheet, as some of you have noticed, had a few zero-ratings.

2.0→2.1

1. An improved method of performing significance testing has been implemented. See Hillis SL, Berbaum KS: "Power Estimation for the Dorfman-Berbaum-Metz Method". Acad. Radiol. 2004; 11:1260-1273.
2. For datasets with <u>equal</u> lesions-weights (see definition below) the <u>program reverts to 1.05</u>, excepting for the above change. Most users use equal lesion-weights. The program automatically detects if the dataset has equal or unequal weights. For <u>unequal</u> weights it reverts to 2.0, excepting for the change in #1 above.
3. The single reader analysis, which should be of interest to CAD developers, is also applicable to multiple modalities (e.g., different CAD algorithms applied to a common image set).
4. The Excel data format, while not necessary for equal weights, may be easier for some users. Both original text format (see ReadMe-1.05.pdf) and the Excel format are supported.

Earlier versions→2.0

1. **It is not necessary to have an IDL license in order to run the JAFROC program.** The program should run on any Windows PC.
2. **Lesion "weights": most users can ignore this feature, and it is wise to present the equally-weighted analysis as the primary result, since the weighted analysis is not as widely accepted (it can be presented as a secondary result)**
    a. The program implements a procedure (see Reference 4) for handling multiple lesions per image, and where the lesions on any particular image can have variable "weights". This capability is probably of interest to a very limited number of users. For data in that format all lesions are assumed to be equally weighted. Users can also use the Excel format described in this document and specify equal weights.
    b. The "weight" of a lesion can be thought of as the clinical significance of detection of that lesion. For example, an image with two lesions could have two assigned weights; e.g., 0.9 and 0.1 if detection of the first lesion is nine times as important as detection of the second lesion. <u>It is up to</u>

the user (a clinician should be consulted) to assign the weights, keeping in mind that the weights must add up to unity (to within 1%). For example an image with three equally weighted lesions would have weights 0.33, 0.33, and 0.33.

c. The data entry for the weighted analysis is necessarily more complex. To ease the burden, the program accepts this in the form of three Excel worksheets embedded within a single Excel workbook. The Excel data format is described below.

3. The output is sent to a text file – this is necessitated by the capability described in #1 above. The output is "prettier", and 95% confidence intervals are provided for the reader-averaged figure-of-merit value for each modality (this quantity was not provided in earlier versions).

Troubleshooting

The program incorporates various error-checking routines that check the DLLs and the consistency of the input Excel data file or the input text file (i.e., it checks for several possible user entry errors). Error messages are posted to the output file (most are applicable to text input file only).

Table 4: Error messages

Message	Meanings
DLL Error	DLL version mismatch. Report this error to devchakraborty@devchakraborty.com. Include your data file.
Error opening excel file: [filename]	Worksheets were not properly named. Must have the following worksheets: **Truth, TP** and **FP**. Invalid data format in the Excel worksheets (user entry errors). Microsoft Excel ODBC driver not found (this should be rare).
Number of lesions for Case # are inconsistent for Reader # Modality #	Error of the number of lesions for the specified Reader, Modality and Case combination (applicable to text input file only).
There are two entries for Reader #, Modality # and Case # at Line #.	The specified Case ID for the Reader and Modality combination appears twice in the dataset (applicable to text input file only).
There are two TP entries for the same lesion ID. Check Row # in TP worksheet.	The specified Reader, Modality, Case and Lesion ID combination appears twice in the TP worksheet (applicable to Excel input file only).
Number of normal(abnormal) image for Reader # ad Modality # are inconsistent with rest of the data.	self explanatory (applicable to text input file only).
No normal(abnormal) cases in the dataset.	The dataset must contain normal and abnormal cases.
Data entry error at Line #.	General data entry error, mostly due to the mismatch of the number of marks (applicable to text input file only); (ex: "1 1 1 0 0 3 1 2": three FP marks expected, two actual FP marks listed).
Inconsistent data for Reader #. Does this reader exist?	self explanatory; typographical error of Reader ID;e.g., reader ID 6 occurs someplace in a 5 reader study with reader IDs 1 – 5 (applicable to text input file only).
Does Modality # exist? Check Modality entry for Reader #.	self explanatory ; typographical error of Modality ID (applicable to text input file only).
Does Case # exist? Check Case ID for Reader # and Modality #.	self explanatory ; typographical error of Case ID (applicable to text input file only).
Blank cell detected at row # column # in ## worksheet.	Data entry error in Excel worksheet.
Any other error	Report this error to devchakraborty@devchakraborty.com. Include your data file.

Revision Notes

Friday, April 16, 2004
In this version, the first revision since SPIE-04, the supporting code is in a binary file; also this version is distributed through the web at http://jafroc.radiology.pitt.edu.

Friday, April 30, 2004
A serious error in the analysis code was corrected. Cross checking of analysis code with simulator generated data was implemented. Updated this file.

Tuesday, October 05, 2004
Implemented (1) floating point ratings, (2) one reader case – for CAD optimization, (3) DLL checking, (4) updated this file.

Monday, October 11, 2004
Implemented (1) arbitrary numbering of modalities, cases and readers – to be consistent with this file; (2) no need to specify number of images, readers, modalities – gets all this from data file, (3) web-site download tracking, (4) updated this file.

Thursday, June 02, 2005
In this new version (1.04) the main changes are (1) data entry checking on the input file, (2) the 95% confidence intervals (CI) are output showing the CI for all modality pairings and (3) the results of conventional ROC analysis (with trapezoidal area) of the highest rated mark are also output.

Thursday, September 29, 2005
In this new version (1.05) a listing error was fixed (individual reader values were not being listed for the JAFROC case). Also a new section "Notes on Output (dpc: 9/27/05)" on the interpretation of the listing was added. The website address has changed to http://www.devchakraborty.com. My e-mail address is: devchakraborty@devchakraborty.com.

Tuesday, March 21, 2006
In this new version (1.06) a display error was fixed (Display mean FOM for single-reader).

Saturday, July 01, 2006
JAFROC Analysis Software version 2.0 released.

Wednesday, November 08, 2006
JAFROC Analysis Software version 2.1 released.

Thursday, June 21, 2007
JAFROC Analysis Software version 2.2 released.

12/28/2007
JAFROC Analysis Software version 2.3 released.

2/15/2008
JAFROC Analysis Software version 2.3a released.

Last revision date: 2/15/2008

FROC REFERENCES

1. Egan JP, Greenburg GZ, Schulman AI. Operating characteristics, signal detectability and the method of free response. J Acoust Soc. Am. 1961; 33:993-1007.
2. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. J of Appl Photogr. Eng. 1978; 4:166-171.
3. Chakraborty DP, Breatnach ES, Yester MV, Soto B, Barnes GT, Fraser RG. Digital and Conventional Chest Imaging: A Modified ROC Study of Observer Performance Using Simulated Nodules. Radiology 1986; 158:35-39.
4. Chakraborty DP. Maximum Likelihood analysis of free-response receiver operating characteristic (FROC) data. Med. Phys. 1989; 16:561-568.
5. Chakraborty DP, Winter LHL. Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment. Radiology 1990; 174:873-881.
6. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis and validation. Medical Physics 2004; 31:2313-2330.
7. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. Phys. Med. Biol. 2006; 51:3449–3462.
8. Chakraborty DP. ROC Curves predicted by a model of visual search. Phys. Med. Biol. 2006; 51:3463–3482.
9. Chakraborty DP. FROC curves using a model of visual search. In: Sahiner YJaB, ed. Proc. SPIE Medical Imaging. San Diego: SPIE, 2007.
10. Yoon HJ, Zheng B, Sahiner B, Chakraborty DP. Evaluating computer-aided detection algorithms. Medical Physics 2007; 34:2024-2038.
11. Chakraborty DP, Yoon HJ. Operating characteristics predicted by models for diagnostic tasks involving lesion localization. Med. Phys. 2008; 35:435-445.
12. Penedo M, Souto M, Tahoces PG, et al. Free-Response Receiver Operating Characteristic Evaluation of Lossy JPEG2000 and Object-based Set Partitioning in Hierarchical Trees Compression of Digitized Mammograms. Radiology 2005; 237:450-457.
13. Zheng B, Chakraborty DP, Rockette HE, Maitz GS, Gur D. A comparison of two data analyses from two observer performance studies using Jackknife ROC and JAFROC. Medical Physics 2005; 32:1031-1034.
14. Timberg P, Ruschin M, Båth M, et al. Potential for lower absorbed dose in digital mammography: a JAFROC experiment using clinical hybrid images with simulated dose reduction. In:Proc. SPIE Medical Imaging. San Diego, CA, 2006.
15. Ruschin M, Timberg P, Bath M, et al. Dose dependence of mass and microcalcification detection in digital mammography: free response human observer studies. Med. Phys. 2007; 34:400 - 407.
16. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys 2002; 29:2861-2870.
17. Samuelson FW, Petrick N, Paquerault S. Advantages and examples of resampling for CAD evaluation. In:2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2007; 492-495.
18. Samuelson FW, Petrick N. Comparing image detection algorithms using resampling. 2006 IEEE International Symposium on Biomedical Imaging: From Nano to Micro 2006:1312-1315.
19. Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. Acad Radiol. 2006; 13:1187-1193.