Electronic Supplementary Material (ESI) for Analyst.

Infrared Micro-spectroscopy Coupled with Multivariate and Machine Learning Techniques for Cancer Classification in Tissue: A Comparison of Classification Method, Performance, and Pre-processing Technique

Dougal Ferguson,*[ab] Alex Henderson[ab], Elizabeth F. McInnes [c], Rob Lind [c], Jan Wildenhain [c], and Peter Gardner [ab]

Supplementary material: Calculating and comparing F_1-Scores from a simple two-class classification problem

For the benefit of the reader, two fictitious studies working on an identical classification problem will be presented that allow for direct comparisons as to the calculation and interpretation of the F_1-Score. This serves to highlight for those unfamiliar with the F_1-Score what gives rise to a good score and how they can be interpreted.

The study will take the form of a two-class classification problem for the preliminary diagnosis of cancer at the patient level through the interrogation of tissue samples using FT-IR spectroscopy. In this case the classification results are simply the level of correct prediction/classification of cancer or non-cancer status and can be represented by the following confusion matrix:

		Predicted Class	
		Cancer	Non-Cancer
Actual Class	Cancer	True Positive (**TP**)	False Negative (**FN**)
	Non-Cancer	False Positive (**FP**)	True Negative (**TN**)

For the example, the initial comparison for both fictitious studies will have the same number of patient predictions (100) with an equal number of patients in each class (50 cancerous and 50 non-cancerous respectively) and the only difference between the two is the Machine Learning (ML) method employed for predictions. In a clinical context it can be assumed that the ideal model will maximise the number of true positive (correct classification of cancer) and minimise false negatives (false classification of non-cancer). For equal and balanced classes the results of the studies are as follows:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad , \quad F_1 \epsilon [0, 1]$$

Method 1

		Predicted Class: Cancer	Predicted Class: Non-Cancer
Actual Class	Cancer	38	12
	Non-Cancer	24	26
F1_score		**0.68**	

Method 2

		Predicted Class: Cancer	Predicted Class: Non-Cancer
Actual Class	Cancer	38	12
	Non-Cancer	12	38
F1_score		**0.76**	

Although of equal predictive power regarding the correct classification of cancerous patients, the second method obtains a larger F_1-Score due to its lower mis-classification rate of non-cancer.

To highlight the main criticism of this score, a second comparison can be conducted but with one example having unbalanced classes. This time the studies will be shown to have equal F_1-Scores even when one method is a stronger classifier. One method will maintain the balanced class as previously mentioned, whereas the other method will have triple the amount of non-cancerous samples (150). The results of the studies are as follows:

Method 1

		Predicted Class: Cancer	Predicted Class: Non-Cancer
Actual Class	Cancer	40	10
	Non-Cancer	24	26
F1_score		**0.70**	

Method 2

		Predicted Class: Cancer	Predicted Class: Non-Cancer
Actual Class	Cancer	40	10
	Non-Cancer	24	136
F1_score		**0.70**	

Although the second method is far superior regarding its correct classification of non-cancerous samples, the returned F_1-Scores are identical. To combat this, you can simply calculate two F_1-Score by simply calculating a second F_1-Score but reversing the calculation to incorporate the false negatives and then taking the mean:

$$F_{1-alt} = \frac{TN}{TN + \frac{1}{2}(FP + FN)} \quad , \quad F_1 \epsilon [0, 1]$$

This provides a new score that reflects the better classification power of the model, and is used in the parent study (referred to as the micro and macro F_1-Scores:

Method 1		**Method 2**	
F1_score	0.70	F1_score	0.70

F1-alt_score	0.60	F1-alt_score	0.89
Mean	**0.65**	Mean	**0.80**

Supplementary material: Calculating F_1-Score from Sensitivity and Specificity metrics

In the paper published by Berisha et al. [1] results are presented as a collection of sensitivity and specificity metrics over standard and high-definition datasets, across two machine learning classifiers. The calculation of the F_1-Score for the CNN(HD) method is provided as an illustration for the reader. To calculate an F_1-Score, a confusion matrix must be generated that results in the same sensitivity and specificity metrics as the distribution of predictions is not known. Firstly, an empty confusion matrix is generated, with the diagonals populated with the sensitivity values. As the metrics are reported as percentages, these are converted to a proportion (however this does not change any results in practice). Since the calculation of sensitivity is the division of the true positive values over the sum of all class predictions for that class, as expressed in Table 2 of the study, an initial guess of other prediction counts is needed. This is calculated as:

$$Initial\ guess = \frac{(1 - TP)}{n - 1}$$

Where the value 1 reflects the upper limit of the sensitivity metric (if using percentage values this would remain 100), and n is the number of classes. This initial confusion matrix is tabulated below.

Class	Sensitivity	Specificity
Adipocytes	91.86	99.58
Blood	87.50	99.98
Collagen	98.22	98.54
Epithelium	90.90	96.41
Myofibroblasts	93.39	97.29
Necrosis	91.89	99.97

		Predicted Class					
		Adipocytes	Blood	Collagen	Epithelium	Myofibroblasts	Necrosis
True Class	Adipocytes	**91.86**	1.63	1.63	1.63	1.63	1.63
	Blood	2.50	**87.50**	2.50	2.50	2.50	2.50
	Collagen	0.36	0.36	**98.22**	0.36	0.36	0.36
	Epithelium	1.82	1.82	1.82	**90.90**	1.82	1.82
	Myofibroblasts	1.32	1.32	1.32	1.32	**93.39**	1.32
	Necrosis	1.62	1.62	1.62	1.62	1.62	**91.89**

From this confusion matrix containing the initial guesses, the estimated sensitivity and specificity metrics obtained by this matrix can then be calculated. From these estimations, the sum of squared error is then calculated, with the average of the pair being the objective of Microsoft Excel's solver.

	Sensitivity	
	Real	Est
Adipocytes	91.86	91.86
Blood	87.50	87.50
Collagen	98.22	98.22
Epithelium	90.90	90.90
Myofibroblasts	93.39	93.39
Necrosis	91.89	91.89

SSE_{Sens}	**0**

	Specificity	
	Real	Est
Adipocytes	99.58	98.48
Blood	99.98	98.65
Collagen	98.54	98.22
Epithelium	96.41	98.51
Myofibroblasts	97.29	98.41
Necrosis	99.97	98.47

SSE_{Spec}	**0.11**

Mean SSE	**0.06**

The solver is then run to find the correct confusion matrix makeup that would provide the same reported sensitivity and specificity values as reported in the study, from which the F_1-Score can be calculated.

		Predicted Class					
		Adipocytes	Blood	Collagen	Epithelium	Myofibroblasts	Necrosis
True Class	Adipocytes	**83.30**	0.00	1.34	3.62	2.41	0.00
	Blood	0.00	**67.07**	0.39	5.80	3.39	0.00
	Collagen	0.60	0.09	**125.03**	0.74	0.69	0.13
	Epithelium	0.20	0.00	1.21	**42.03**	2.80	0.00
	Myofibroblasts	0.69	0.00	1.35	2.84	**68.93**	0.00
	Necrosis	0.26	0.00	1.26	3.54	2.45	**85.16**

	Sensitivity	
	Real	Est
Adipocytes	91.86	91.86
Blood	87.50	87.50
Collagen	98.22	98.22
Epithelium	90.90	90.90
Myofibroblasts	93.39	93.39
Necrosis	91.89	91.89

SSE_{Sens}	**0.00**

	Specificity	
	Real	Est
Adipocytes	99.58	99.58
Blood	99.98	99.98
Collagen	98.54	98.54
Epithelium	96.41	96.41
Myofibroblasts	97.29	97.29
Necrosis	99.97	99.97

SSE_{Spec}	**0.00**

Mean SSE	**0.00**

Class	F1-Score
Adipocytes	0.95
Blood	0.93
Collagen	0.97
Epithelium	0.80
Myofibroblasts	0.89
Necrosis	0.96

Micro-F1	**0.93**
Macro-F1	**0.92**
Median-F1	**0.94**

References

[1] S. Berisha, M. Lotfollahi, J. Jahanipour, I. Gurcan, M. Walsh, R. Bhargava, H. Van Nguyen and D. Mayerich, "Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks," *Analyst,* vol. 144, no. 5, pp. 1642-1653, 2019.