

CLRC—TR—08—01
MRC UKOPS: CLRC data analysis report*

D. Devetyarov, B. Burford, Z. Luo, I. Nouretdinov,
V. Vovk, A. Chervonenkis, A. Gammerman
Royal Holloway, University of London
and
S. Camuzeaux, R. Hallett, J. Ford,
A. Gentry, J. Timms, U. Menon, I. Jacobs
University College London
and
R. Cramer, A. Tiss, C. Smith
University of Reading

October 16, 2008 †

Abstract

This report presents data analysis on the serum mass spectrum data collected in the UKOPS trial. The main goal of this analysis is to discover new potential bio-markers that can be used either alone or in combination with CA125 to produce models with greater predictive power. We conclude that despite trying a large number of classification models only small improvements can be made to the the current ability of CA125. Several peaks when used in combination with CA125 produce modest but reproducible results when validated on a blind test set.

1 Data

Serum samples were collected from patients in the UKOPS trial. These patients are from 3 main groups: Healthy, Benign and Malignant. In addition, there were borderline (BL) samples: 1 fallopian tube BL and 10 ovarian BL, which were treated in a different way in different experiments. The serum samples were prepared and analysed independently using two MALDI-TOF mass spectrometers at UCL and at the University of Reading. Therefore, we have two sets of mass spectrum and will refer to these two sets simply as UCL data and Reading data, respectively.

The samples were analysed in 2 batches. These batches were used to construct a training set and a test set for classification in the following way:

*See also Supplements 1 and 2 to CLRC—TR—08—01

†The date of version 1 - June 18, 2008

Class	Reading Data	UCL Data
Training set, before elimination of samples without CA-125		
Healthy	106	100
Benign	63 + 1 fallopian tube BL	62 + 1 fallopian tube BL
Malignant	41	42
Training set, after elimination of samples without CA-125		
Healthy	104	98
Benign	62 + 1 fallopian tube BL	62 + 1 fallopian tube BL
Malignant	38	42
Test set		
Healthy	66	66
Benign	22	22
Malignant	29 + 10 ovarian BLs	29 + 10 ovarian BLs

Table 1: The breakdown of training and test tests into Healthy, Benign and Malignant classes.

- The training set contains all of batch 1 and some randomly selected samples from batch 2.
- The test set contains the remaining samples from batch 2.

Later in this report we will consider the stability of control (healthy) samples. This stability will be tested using the training set only; we will compare: the control samples in the training set that came from batch 1 with the control samples in the training set that came from batch 2.

Table 1 represents the breakdown of training and test sets of both Reading and UCL data into Healthy, Benign and Malignant. Borderline samples were originally labelled as Benign or Malignant. For this reason, borderline samples are shown within groups of samples (Benign or Malignant) they were labelled as.

Some serum samples were deemed to be of low quality and were removed during MS experiments. As a result, the number of serum samples analysed successfully in UCL data and in Reading data differs slightly. Furthermore, each serum sample analysed by UCL had a maximum of 18 replicates; each sample analysed by the University of Reading has a maximum of 3 replicates. In order to produce one peak intensity per sample, we took the following action:

- We generated peak groups based on all samples (see Section 2 for details).
- In order to deal with replicates we take the following action: for every peak in each sample average the intensity of the peak over all the replicates.

Please note that we did not simply take the average of the processed spectra before peak identification, as this would result in the intensities being affected by shifts in the m/z -axis.

2 Pre-processing

This section briefly describes the pre-processing applied to the UCL data and Reading data. The data is provided in a two-column format: the first column is a list of m/z -values, the second column is a list of corresponding intensities. Calibration had been performed prior to the data being distributed; therefore, all our further pre-processing steps were applied only to the intensities, m/z -values remained unchanged. The pre-processing steps which were applied to both data sets are described below:

1. **Down-sampling** was performed in order to decrease the number of m/z -ratios for computational optimization purposes.
2. For noise elimination, we performed **smoothing** by averaging the intensities within a moving window.
3. **Baseline subtraction** ensured that the spectra sit on the intensity = 0 axis. The algorithm applied is based on finding the lowest points between some dominant local maxima (we refer to these lowest points as *troughs* later on). We define a dominant local maximum as the point with the highest intensity within some range, the width of this range is a parameter of the algorithm. We then apply Piecewise Cubic Hermite Interpolating Polynomial to all the points marked as troughs in order to construct a baseline. Further steps are applied to correct the baseline for any points where the baseline is above the spectra.
4. We performed **normalisation** to make sure that the total amount of ions across different samples were the same. The algorithm involved dividing the intensity of each point in a spectra by the sum of all intensities.
5. The goal of the **peak identification** step was to generate a list of peaks for each sample. This was achieved by identifying all local maxima in the mass spectra above an intensity threshold and above a certain signal-to-noise ratio threshold. As a result of this algorithm we had a table for each sample with the following columns:
 - (a) Unique sample ID — for any single sample this column has the same number for each entry, the reason for this column will become apparent in the next step.
 - (b) Number of peak in the initial array of m/z -values.
 - (c) M/z -value of a local maximum.
 - (d) Signal-to-noise ratio within a window of the certain width.
 - (e) Corresponding intensity.
6. Peaks obtained in the previous step do not strictly coincide because of the noise and possible presence of different isotopes. Therefore, the next step (called **peak alignment**) is to find common peaks (that is, peaks with m/z -values close to each other) among the samples. We combined all peak lists constructed for individual samples into one list and sorted in descending order relative to column 5 — peak intensity. We then worked down the list taking one of the following actions for each peak:

- (a) If the peak is within some predefined range of an existing peak group and there is no other peak from the same sample in that group, then the current peak can be added to the group; else the peak is ignored.
- (b) If there are no groups close to the current peak, then a new peak group is created containing the single peak.

See Appendix A for the lists of peaks for Reading and UCL data.

7. The result of the previous step is a set of peak groups, each of which can potentially contain between 1 peak and N peaks, where N is the number of samples in the data set. Now we have to compute peak intensities for each sample for each peak group. For those peak groups that have less than N peaks, we need to estimate the intensities for the remaining samples. For this purpose, we set a mass separation parameter which we use to define a range in the m/z-vector given a single m/z-value — the maximum m/z-ratio of all the peaks in the group. Then the intensity of each missing sample is defined as the maximum intensity within this range in the spectrum of the sample.

3 Classification: Description

The output from the pre-processing described above was a table containing a list of peak intensities for each sample. Each peak intensity could be used as a feature in a classification problem. However, we selected the subset of features and eliminated several samples as follows:

1. The only peak groups considered were those that appeared in greater than 10% of the total number of samples. This reduced the number of peaks from 431 to 108 in the Reading data and 665 to 132 in the UCL data.
2. We calculated Mann-Whitney U-test p-values checking the hypothesis that Healthy samples from the training set draw from batch 1 and Healthy samples from the training set drawn from batch 2 come from the same distribution. Peaks with Bonferroni¹ adjusted p-values lower than the threshold = 0.01 (we refer to them as "unstable peaks") were eliminated. The motivation behind this was to produce a list of peaks that were stable with respect to the controls (healthy samples).
As a result, we identified 20 unstable peaks for Reading data and 36 unstable peaks for UCL data and reduced our peak sets to 88 and 96 peaks for Reading and UCL data, respectively. The eliminated unstable peaks are represented in Tables 15 and 16 of Appendix B.
3. Since we used values of the biomarker CA-125 for classification, 6 samples from Reading data and 2 samples from UCL data without CA-125 values were eliminated (Table 1).

¹Bonferroni adjustment involves multiplication of each p-value by the total number of peaks in the comparison, see for example H, Abdi (2007). "Bonferroni and Sidak corrections for multiple comparisons." In N. J. Salkind (ed.): *Encyclopedia of Measurement and Statistics*.

The aim of our analysis was to identify classification rules that when applied to small number of peaks can provide high-quality discrimination between the Benign and Malignant classes and between the Healthy and Malignant classes.

At first, a fallopian tube borderline sample in a training set was treated as Benign and 10 ovarian borderline samples were treated as Malignant. For this reason, we tested two types of discrimination:

- Healthy vs (Malignant + ovarian BLs)
- Benign + fallopian tube BL vs Malignant + ovarian BLs

After excluding borderline samples, we compared the following classes:

- Healthy vs Malignant
- Benign vs Malignant

And finally, the following type of discrimination was tested:

- (Healthy + Benign + fallopian tube and ovarian BLs) vs Malignant

We constructed models for different types of discrimination independently, and the prediction model comprised a classification method and a subset of peaks.

The general outline of identifying the best prediction models on the training set was as follows. Given a certain classification method, we went over all subsets of peaks of certain cardinality (usually a small number) and applied the classification method only to the subset of peaks. In the following, we measured the quality using leave-one-out cross-validation. For each sample we trained our algorithm on all other samples and generated a prediction for the *left out* sample by using it as a test set of size 1. This resulted in a prediction for each sample based on a rule generated by all other samples in the training set. Therefore, we could construct a confusion matrix and compute such values as specificity and sensitivity. As we put sensitivity first in this study, the quality measure of prediction was calculated as a linear combination of sensitivity and specificity with weights in favour of sensitivity:

$$\text{Quality} = \frac{2}{3} \times \text{Sensitivity} + \frac{1}{3} \times \text{Specificity}.$$

To avoid confusion, we will distinguish the measure called Quality introduced above from the usual meaning of the term “quality” through capitalisation.

Out of all the subsets of peaks considered, we selected the one that provided the best value of Quality. After that, we selected the classification rules with the best Quality on certain combinations of peaks with the biomarker CA-125 and applied the identified prediction models to the test set, training the classifier on the whole training set.

As the peak intensities were incommensurable, before applying classification rules we took a logarithm of all values and then normalized them across the peaks by subtracting means and dividing by standard deviations, where the means and standard deviations were determined from the whole training set. When considering a sample from the test set in validation, the mean and standard deviation were determined from the union of the training set and the sample.

CA-125 is a well established biomarker for ovarian cancer and provides good discrimination in Healthy vs Malignant case just using a cut-off model. In this report we will compare our results to a ‘CA125 30 cut-off rule’, this rule is defined as:

$$\text{prediction} = \begin{cases} \text{Malignant} & \text{if the level of CA125 is } > 30 \\ \text{Healthy / Benign} & \text{otherwise.} \end{cases}$$

For this reason, we intended to find prediction models that outperform this benchmark model.

At first we tried to construct models on the basis of the mass spectra information only. But the models obtained did not appear to be stable: when parameters of classification methods were slightly changed, selected peaks differed significantly. Also when we attempted to validate models on the test set it was clear the the results were not reproducible. Therefore, we considered CA-125 as one of the attributes (together with peak intensities) and applied classification methods to combinations of CA-125 with several peaks.

In our analysis we considered such classification methods as weighted k-nearest neighbours algorithm (kNN), logical combinations of cut-off rules, cut-off rules for linear combinations, SVM with various kernels. In summary, all the methods produced worse results than weighted kNN. For this reason, we focused on application of the weighted k-nearest neighbours algorithm, but we will also demonstrate some SVM results.

The weighted kNN is based on observing the closest samples to the current sample when represented by the peak intensities in a vector space. The parameter k is the number of closest samples that will be observed. The additional parameter j is the necessary and sufficient number of Malignant nearest neighbours out of k nearest neighbours. Thus, if at least j out of k nearest neighbours are Malignant, the sample is predicted to be Malignant, otherwise it is predicted to be Benign/Healthy. The following classification rule is referred to as “ j out of k weighted NN rule” in our result tables.

In order to show the significance of any classification results, we calculated p-values. Our null hypothesis is that we can obtain the Quality as small or better than the Quality of classification when randomly permuting the labels. Intuitively, this tests whether the results could be obtained by chance. We use the Monte-Carlo method for generating p-values. The method is based on an iterative process: we randomise the labels at each iteration in order to test if the Quality obtained with labels randomly permuted is equal or better than the Quality with the correct labels. In our case, given certain values of parameters j and k for the weighted nearest neighbours classification method and a certain permutation of labels, we looked for the combination of peaks that provided the best Quality on one of the two kNN classification methods (one of them is weighted in favour of Malignant class, another - in favour of Benign/Healthy class):

1. If at least j out of k nearest neighbours are Malignant, the sample is predicted to be Malignant, otherwise it is predicted to be Benign/Healthy.
2. If at least j out of k nearest neighbours are Benign/Healthy, the sample is predicted to be Benign/Healthy, otherwise it is predicted to be Malignant.

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	97.4%	99.0%	98.6%	97.9%	89.7%	98.5%	95.2%	92.6%
CA-125	94.7%	91.4%	92.3%	93.6%				
Threshold model (CA-125 < 30)	87.8%	96.2%	93.9%	90.6%	87.2%	98.5%	94.3%	91.0%

Table 2: Reading data, Healthy vs Malignant (including BLs), Model 1 (1 out of 2-NN, peaks 2660.9 Da, 2770.1 Da, CA-125).

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	94.7%	100.0%	98.6%	96.5%	84.6%	100.0%	94.3%	89.7%
CA-125	92.1%	96.2%	95.1%	93.5%				
Threshold model (CA-125 < 30)	87.8%	96.2%	93.9%	90.6%	87.2%	98.5%	94.3%	91.0%

Table 3: Reading data, Healthy vs Malignant (including BLs), Model 2 (2 out of 5-NN, peaks 972.07 Da, 2366.8 Da, CA-125).

4 Classification: Results

In all types of discrimination we only present results for pairs of peaks as increasing the number of peaks (≥ 3) did not provide stable or reproducible results.

4.1 Healthy vs Malignant (Including Borderline Samples)

In case of Healthy vs Malignant discrimination including borderline samples, we managed to find models outperforming the CA-125 cut-off model on both training and test sets.

Tables 2–6 show the best models in terms of Quality for this discrimination for Reading data and UCL data. The tables represent Sensitivity, Specificity, Accuracy and Quality on both training and test sets. The second row shows results of applying the kNN method to CA-125 as a single attribute without peak intensities, and the third row—for the CA-125 cut-off method. The classification method and peaks involved are shown in the captions of the tables.

All the p-values were equal to 0.001 for 1000 iterations, that is, no random permutation of labels resulted in a quality equal to or better than the quality of real labels.

As you can see from the tables, three models were identified for Reading data. They surpass the cut-off model on the training set significantly and,

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	100.0%	92.3%	94.4%	97.4%	89.7%	87.9%	88.6%	89.1%
CA-125	97.4%	87.5%	90.1%	94.1%				
Threshold model (CA-125 < 30)	87.8%	96.2%	93.9%	90.6%	87.2%	98.5%	94.3%	91.0%

Table 4: Reading data, Healthy vs Malignant (including BLs), Model 3 (1 out of 4-NN, peaks 3952 Da, 1114.4 Da, CA-125).

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	95.2%	98.0%	97.1%	96.1%	87.2%	95.5%	92.4%	90.0%
CA-125	95.2%	90.8%	92.1%	93.8%				
Threshold model (CA-125 < 30)	90.5%	96.0%	94.4%	92.3%	87.2%	98.5%	94.3%	91.0%

Table 5: UCL data, Healthy vs Malignant (including BLs), Model 4 (1 out of 2-NN, peaks 1420.3 Da, 7779.3 Da, CA-125).

although their performance deteriorates on the test set, they produce better or approximately the same Quality results than the benchmark method on the test set:

- Model 1 (Table 2) outperforms the cut-off model (with respect to sensitivity, specificity and accuracy) on the test set.
- Model 2 (Table 3) has the same accuracy as the cut-off model and 100% specificity on the test set.
- Model 3 (Table 4) has better sensitivity but is inferior to the cut-off model on the test set in terms of the other criteria.

For UCL data, two models were identified (Tables 5 and 6). Both models outperform the cut-off model on the training set but are inferior to it on the test set. Nevertheless, Model 1 has very close values of the criteria to those of the benchmark model.

For all models selected we can notice that performance of the kNN method on the single attribute CA-125 is worse than the performance of the cut-off model. However, addition of a pair of peak intensities improves the performance and allows achieving the same or better prediction quality.

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	100.0%	83.7%	88.6%	94.6%	97.4%	60.6%	74.3%	85.1%
CA-125	97.6%	82.7%	87.1%	92.6%				
Threshold model (CA-125 < 30)	90.5%	96.0%	94.4%	92.3%	87.2%	98.5%	94.3%	91.0%

Table 6: UCL data, Healthy vs Malignant (including BLs), Model 5 (1 out of 5-NN, peaks peak 709.98 Da, 2771.4 Da, CA-125).

4.2 Benign vs Malignant (Including Borderline Samples)

In case of Benign vs Malignant (including BLs) discrimination, models that provide stable predictions of high quality were not found. On both Reading and UCL data, we could identify the models based on various prediction rules that significantly outperform the cut-off model on the training set. But the performance of all the models dramatically deteriorated on test sets and happened to be inferior to the cut-off model.

Tables 7 and 8 represent several prediction rules (one based on kNN classifier, one based on SVM classifier) that demonstrated the best results on both training and test sets. The first row shows the performance of the cut-off model for comparison purposes.

In summary, for discrimination between Benign and Malignant classes, the improvement detected on the training set analysis did not retain on the test set.

Model	Peaks	Training set				Test set			
		Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
Threshold model (CA-125 < 30)	-	87.8%	65.6%	74.3%	80.4%	87.2%	59.1%	77.1%	77.8%
1 out of 3 weighted kNN	2366.8, 4787.4	97.4%	77.8%	85.2%	90.9%	82.1%	59.1%	73.8%	74.4%
SVM, linear kernel	1021.3, 3507.3	84.2%	93.7%	90.1%	87.4%	71.8%	68.2%	70.5%	70.6%

Table 7: Reading data, Benign vs Malignant (including BLs), the best models.

4.3 Healthy vs Malignant (Excluding Borderline Samples)

After excluding borderline samples, we did no change the training set, but eliminated 10 ovarian BLs from Malignant class in the test set. Hence, the

Model	Peaks	Training set				Test set			
		Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
Threshold model (CA-125 < 30)	-	90.5%	65.1%	75.2%	82.0%	87.2%	59.1%	77.1%	77.8%
1 out of 2 weighted kNN	3954.7, 4969.9	95.2%	88.9%	91.4%	93.1%	79.5%	36.4%	63.9%	65.1%
SVM, rbf kernel, $\sigma = 2$	2771.4, 4093.4	85.7%	93.7%	90.5%	88.4%	56.4%	81.8%	65.6%	64.9%

Table 8: UCL data, Benign vs Malignant (including BLs), the best models.

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	97.4%	99.0%	98.6%	97.9%	100.0%	98.5%	99.0%	99.5%
CA-125	94.7%	91.4%	92.3%	93.6%				
Threshold model (CA-125 < 30)	87.8%	96.2%	93.9%	90.6%	100.0%	98.5%	99.0%	99.5%

Table 9: Reading data, Healthy vs Malignant (excluding BLs), Model 1 (1 out of 2-NN, peaks 2660.9 Da, 2770.1 Da, CA-125).

same models were selected as the best ones on the training set. They just needed to be applied to the new test set.

Tables 9–11 represent results for Models 1–3 for Reading data. Results for UCL data can be found in slides ‘UKOPS October 16th 2008’.

For Reading data the cut-off model provides high accuracy on the test set: 100% Sensitivity and 98.5% Specificity. For this reason, only small improvement can be made in comparison with the cut-off model:

- Model 1 matches the cut-off model.
- Model 2 has 100% Accuracy and outperforms the cut-off model

For UCL set, both Models 1 and 2 are inferior to the cut-off model on the test set.

4.4 Benign vs Malignant (Excluding Borderline Samples)

We excluded a fallopian tube BL from the training set and 10 ovarian BLs in the test set and repeated the analysis. All the models selected for both Reading and UCL datasets are inferior to the cut-off model on the test set. The description of models and the characteristics of their performance on the training and test sets can be found in slides ‘UKOPS October 16th 2008’.

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	94.7%	100.0%	98.6%	96.5%	100.0%	100.0%	100.0%	100.0%
CA-125	92.1%	96.2%	95.1%	93.5%				
Threshold model (CA-125 < 30)	87.8%	96.2%	93.9%	90.6%	100.0%	98.5%	99.0%	99.5%

Table 10: Reading data, Healthy vs Malignant (excluding BLs), Model 2 (2 out of 5-NN, peaks 972.07 Da, 2366.8 Da, CA-125).

Model	Training set				Test set			
	Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
2 peaks and CA-125	100.0%	92.3%	94.4%	97.4%	100.0%	86.4%	90.5%	95.5%
CA-125	97.4%	87.5%	90.1%	94.1%				
Threshold model (CA-125 < 30)	87.8%	96.2%	93.9%	90.6%	100.0%	98.5%	99.0%	99.5%

Table 11: Reading data, Healthy vs Malignant (excluding BLs), Model 3 (1 out of 4-NN, peaks 3952 Da, 1114.4 Da, CA-125).

4.5 Healthy + Benign + Fallopian Tube and Ovarian Borderlines vs Malignant

When trying to discriminate between Malignant samples and all the others (that is, Healthy, Benign and two types of borderlines samples), we obtained the following results.

For Reading data, two models were identified as ones performing better than the cut-off model. Their performance on both training and test sets are given in Table 12. As the table shows, only Model 2 (2 of 7-nearest neighbours with peaks 3769 Da and 1114.4 Da) outperforms the cut-off model on the test set in terms of accuracy, and it is still beaten by the benchmark model in terms of sensitivity.

Model	Peaks	Training set				Test set			
		Sensitivity	Specificity	Accuracy	Quality	Sensitivity	Specificity	Accuracy	Quality
Threshold model (CA-125 < 30)	-	94.7%	84.4%	86.3%	91.3%	100.0%	84.7%	88.2%	94.9%
2 out of 9 weighted kNN	1546.8, 4466.1	97.4%	90.4%	91.7%	95.1%	93.1%	82.7%	85.0%	89.6%
2 out of 7 weighted kNN	3769, 1114.4	94.7%	92.8%	93.2%	94.1%	96.6%	86.7%	89.0%	93.3%

Table 12: Reading data, Healthy + Benign + Borderlines vs Malignant, the best models.

For UCL set, all models that outperformed the cut-off model on the training set were beaten by the cut-off model on the test set. The description of the models and the characteristics of their performance on the training and test sets can be found in slides ‘UKOPS October 16th 2008’.

5 Conclusion

Healthy vs Malignant

It is clear that no single peak or combination of peaks could be found to provide a clear improvement over CA125; and indeed the application of several models resulted in a new peak combination each time a new model was tried. This instability in the peaks is not reassuring however there are instances where very modest improvements over a CA125 cut-off method can be made. There are two examples where some improvement exists. However, the case for further investigation is weak, because a CA125 cut-off method provides 99% on the test set when borderlines are excluded. The results included below were achieved when using the Reading data:

- Model 2 (2 of 5-nearest neighbours with features: CA125, peak 972.07Da and peak 2366.8Da) shows some stability with respect to reproducibility from the training set onto the test set. Specifically, we achieve 100%

specificity on the training set and reproduce this on the test set regardless of including or excluding borderline samples. The overall accuracy of this model on the test set is identical to that of CA125 using a 30 cut-off when we include borderlines. However, when excluding borderlines, we can achieve 100% accuracy, which still can't be considered as a significant improvement in comparison with 99% accuracy for the CA125 cut-off method.

- Model 1 (1 of 2-nearest neighbours with features: CA125, peak 2660.9Da and peak 2770.1Da) produced 98.6% overall accuracy on the training set outperforming the CA125 30 cut-off, which achieved an overall accuracy of 93.9%. This improvement is also present, to a lesser degree, in the test set when the borderlines are excluded: the model achieves an accuracy of 95.2% compared to 94.3% for the CA125 cut-off—very modest, however this stands as one of the main results in this report. When the borderlines are included, Model 1 matches CA125 but-off model in both sensitivity and specificity.

Models based on the UCL data have also been analysed, however, not as much improvement over the CA125 models were made. Tables 5 and 6 show the results of this analysis with borderlines included.

Benign vs Malignant

Unfortunately the results for the Benign vs Malignant classification problem did not show a good discrimination (between the two classes) across both UCL and Reading data regardless of including or excluding borderline samples. Once again some tables demonstrating the lack of reproducibility of the results are shown in this report. In general models appear to work well on the training set but fail to perform on the test set, this is a clear case of overfitting caused by over tuning of parameters in a situation where no real discrimination between the two classes can be made.

Further Analysis

Results of further analysis of UKOPS data are represented in Supplements 1 and 2. In general, the results support the conclusions made in this report.

Supplement 1 covers the analysis of UKOPS data with Quality weighted towards specificity rather than sensitivity as it is done this report. In addition, the analysis of Benign vs Malignant discrimination for samples with CA125 levels greater than 30 was carried out. Supplement 2 provides the results of the analysis of UKOPS Reading data set randomly divided into new training and test sets, since the division of UKOPS data into training and test sets analysed in this report was biased.

Appendix A: Tables of peaks

Table 13: List of peaks in Reading data.

Peak number	M/z-value
1	2755.2
2	2660.9
3	7764.9
4	9288.8
5	5903.8
6	8141.5
7	3263.3
8	1618
9	1466.9
10	1207.6
11	4209.9
12	6630
13	1351.5
14	905.96
15	2554.8
16	4963.4
17	5336.3
18	4643.8
19	1546.8
20	1420
21	4053.8
22	4091.1
23	1741.5
24	1021.3
25	3192.3
26	6431
27	740.77
28	7923.2
29	1897.7
30	2933.1
31	1262
32	1078.2
33	2379.9
34	3448.5
35	7564.4
36	5064.3
37	2863.1
38	3086.5
39	3883
40	3952
41	2023.1
42	8602.5
43	4787.4
44	869.93

Peak number	M/z-value
45	1866.6
46	8933.1
47	852.93
48	2106.1
49	8766.4
50	1130.9
51	6089.3
52	7469.3
53	5752.9
54	9132.2
55	2093.5
56	1520.4
57	3634.9
58	2295
59	5633.9
60	3769
61	9714.1
62	3364.7
63	2210.8
64	6303.2
65	9422.9
66	8915.3
67	2468.4
68	3965.5
69	3507.3
70	4396.3
71	972.07
72	5487
73	2429.8
74	7191.5
75	2770.1
76	6803
77	4466.1
78	2569
79	2164.5
80	3241.2
81	6909.6
82	3475.6
83	1779.4
84	7219.2
85	6206.5
86	2495.5
87	5595.4
88	808.26
89	9060.6
90	6734.7
91	4527
92	3681.5
93	6225.2

Peak number	M/z-value
94	3349.4
95	2272.9
96	3710
97	7004.8
98	2366.8
99	4362.5
100	8565.3
101	5521.9
102	1114.4
103	5160.2
104	3607
105	4438.5
106	3970.5
107	2191.2
108	2230.4

Table 14: List of peaks in UCL data.

Peak number	M/z-value
1	1467.4
2	2023.5
3	1352.2
4	2771.4
5	2556
6	1208
7	2934.6
8	906.33
9	3265.2
10	2662.2
11	3194.1
12	1867.2
13	2381
14	1021.7
15	709.98
16	4212.8
17	4056.8
18	1520.6
19	2107.1
20	1265
21	1742.1
22	1078.6
23	3954.7
24	1618.7
25	4093.4
26	4647
27	2864.4

Peak number	M/z-value
28	3450.3
29	3368.3
30	1780.1
31	1692.8
32	5070.8
33	7779.3
34	6641.3
35	4969.9
36	1563.7
37	6441.9
38	5913.2
39	4793.8
40	9308.4
41	8779.4
42	5344.4
43	1898.5
44	8931.1
45	9148.3
46	5336.1
47	2127.7
48	2425.1
49	1628.7
50	3637.1
51	2274.7
52	3509.6
53	9442.2
54	943.63
55	5902.1
56	1061.8
57	6428.1
58	3772
59	4317.2
60	1125.8
61	6625.7
62	5762.9
63	7756.9
64	855.28
65	6185
66	9282.7
67	8616.5
68	1420.3
69	8752
70	2211.6
71	2192.3
72	8154
73	5062.1
74	5162.3
75	7308.4
76	8272

Peak number	M/z-value
77	5753.3
78	4961.8
79	9831.2
80	9118.2
81	3038.5
82	3967.1
83	2496.9
84	8903.4
85	2954.9
86	9410.1
87	6194.1
88	5494
89	2468.9
90	8025.9
91	8590.4
92	1134.1
93	853.79
94	5446.8
95	9796.9
96	5637.2
97	3971.2
98	5603.2
99	4652.1
100	6016.2
101	3351.1
102	5549
103	7289
104	3713.1
105	7540.6
106	3434.9
107	4398.8
108	2278.4
109	8455.5
110	4785.6
111	5248.6
112	4365.9
113	8127.2
114	7552.6
115	7576.9
116	6744.8
117	3438.2
118	811.18
119	7013.4
120	4529.6
121	6919.9
122	7143.9
123	9160.7
124	7927.6
125	9269

Peak number	M/z-value
126	773.86
127	8248.8
128	9455.4
129	8002.2
130	782.21
131	3243.4
132	2231.5

Appendix B: Tables of unstable peaks

Peak number in Table 13	M/z-value	p-value
5	5903.8	0.0004
7	3263.3	0.0001
14	905.96	1.1e-10
25	3192.3	0.0001
28	7923.2	0.0084
31	1262	0.0011
36	5064.3	0.0003
45	1866.6	1.5e-7
46	8933.1	3.3e-5
49	8766.4	2.1e-7
53	5752.9	1.1e-7
54	9132.2	1.3e-11
55	2093.5	0.0002
59	5633.9	1.6e-7
61	9714.1	4.1e-7
65	9422.9	6.7e-10
66	8915.3	1.7e-5
89	9060.6	2.0e-7
99	4362.5	0.0001
101	5521.9	0.0001

Table 15: Unstable peaks in Reading data.

Peak number in Table 14	M/z-value	p-value
1	1467.4	0.0018
2	2023.5	4.1e-13
9	3265.2	0.0001
11	3194.1	0.0001
12	1867.2	2.8e-8
14	1021.7	0.0009
29	3368.3	0.0012
47	2127.7	5.1e-14
53	9442.2	0.0023
54	943.63	2.1e-6
59	4317.2	2.7e-15
60	1125.8	0.0003
65	6185	1.8e-10
74	5162.3	1.0e-5
76	8272	4.3e-9
79	9831.2	2.0e-18
81	3038.5	2.7e-8
85	2954.9	0.0003
87	6194.1	1.6e-10
90	8025.9	1.9e-6
94	5446.8	1.4e-16
95	9796.9	3.9e-17
100	6016.2	1.9e-17
101	3351.1	1.8e-5
102	5549	6.1e-6
104	3713.1	2.6e-6
111	5248.6	3.3e-6
112	4365.9	9.6e-10
113	8127.2	0.0041
115	7576.9	0.0009
116	6744.8	4.8e-13
118	811.18	0.0044
124	7927.6	4.2e-14
127	8248.8	1.6e-5
129	8002.2	0.0061
131	3243.4	0.0001

Table 16: Unstable peaks in UCL data.