

Analysing PET scans data for predicting response to chemotherapy in breast cancer patients

Elias Gyftodimos, Laura Moss and Derek Sleeman
Department of Computing Science, University of Aberdeen
Aberdeen

Andrew Welch
School of Medical Sciences, University of Aberdeen
Aberdeen

Abstract

We discuss the use of machine learning algorithms to predict which breast cancer patients are likely to respond to (neoadjuvant) chemotherapy. A group of 96 patients from the Aberdeen Royal Infirmary had the size of their tumours assessed by Positron Emission Tomography at various stages of their chemotherapy treatment. The aim is to predict at an early stage which patients have low response to the therapy, for which alternative treatment plans should be followed. A variety of machine learning algorithms were used with this data set. Results indicate that machine learning methods outperform previous statistical approaches on the same data set.

1 Introduction

Each year, more than 44,000 people are newly diagnosed with breast cancer in the UK [5]. Up to 25% of these patients have large ($>3\text{cm}$) tumours [21]. For these patients, neoadjuvant chemotherapy is sometimes offered in an attempt to reduce the size of the tumour before surgery is carried out to remove the tumour [4, 8]. It is estimated that up to 25% of these patients do not respond to this chemotherapy [21]. Therefore, for this group of patients, it is a waste of time and resources for neoadjuvant chemotherapy to be administered, and these patients should have surgery at an earlier stage. It would be beneficial in a clinical setting to predict which patients with breast cancer will not respond positively to this chemotherapy. At the same time it would be important to ensure that patients who would be positive responders receive the treatment. This prediction of treatment outcome would preferably happen before treatment commences or at least early in the scheduled treatment, thereby avoiding toxic and expensive chemotherapy doses.

Methods to detect the response of a breast cancer tumour to neoadjuvant chemotherapy include the use of Positron Emission Tomography (PET) [11, 13]. PET scans can be used to visualise the concentration of a given trace compound such as $[^{18}\text{F}]$ -fluorodeoxy-D-glucose (or FDG for short) in body tissue. As

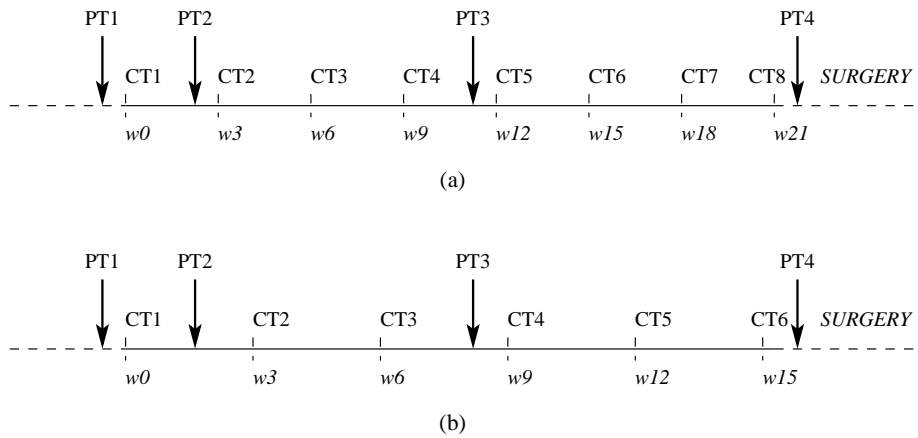


Figure 1: Timeline (in weeks) for a patient receiving (a) 8 doses or (b) 6 doses of chemotherapy treatment, illustrating the timepoints of chemotherapy (CT1, CT2,...) and PET scans (PT1 to PT4).

cancer cells tend to grow more rapidly than other tissue they tend to use more glucose, and hence as a result more of the tracer element is found in such tissue. So this technique allows one to “see” the site and shape of the tumours. In the case of this study the tumour itself is being subjected to an “attack” by the chemotherapy between successive PET scans, and hence the imaging analysis allows the clinician to “see” the site and shape of the remaining tumour. The data that is produced from scans taken before and during a neoadjuvant chemotherapy treatment can be analysed to predict which of the breast cancer patients will be unresponsive to chemotherapy [20, 21].

This paper describes the application of various machine learning algorithms to data acquired and analysed in a previous study [17, 21]. The patients underwent 6 or 8 cycles of chemotherapy treatment before having the tumour removed surgically. The interval between chemotherapy cycles was three weeks. Medical imaging data on the tumour region were gathered for each patient using PET at four time points: before the start of the chemotherapy, after the first chemotherapy cycle, at the midpoint, and at the endpoint of chemotherapy treatment (prior to surgery). Figure 1 (upper) illustrates a timeline of 21 weeks for a patient receiving 8 doses of chemotherapy; and the lower figure illustrates a timeline of 15 weeks for a patient receiving 6 doses of chemotherapy. The diagram shows how chemotherapy sessions are organised, when the 4 PET scans occur, as well as the surgery to remove the tumour. Biopsies after surgery revealed that only about 1 out of every 3 patients (33%) significantly responded to the chemotherapy treatment, while in about 1 out of every 4 (25%) cases there was no measurable response and so there was partial response in about 40% of the population. These biopsies are used to indicate the effectiveness of the chemotherapy treatment through this study. As the results of such pathological analyses are highly reliable these are effectively treated as 100% accurate

and so are treated as a *gold standard*.

The medical imaging data was manually processed by a domain expert who identified the tumour and background regions in the image. Subsequently, for each scan a vector of numerical features was constructed; each such feature is a measure of the activity in the tumour region of interest. The relative change in the value of each feature was derived for each patient between the pre-therapy and each of the three other scans. The initial study concluded that the relative change of the values of each feature between the pre-therapy and the second or third scan correlate well with the tumour response to chemotherapy. In [17] it is shown that these relative changes of values can predict response at the midpoint of therapy (after the third scan is performed). The methodology was to examine each attribute independently, and set a single threshold as a decision criterion which discriminates between high and low responses. The evaluation of each feature in prediction was performed in two ways. Firstly, by setting the threshold value to the point where 100% of the high responders are classified correctly, and measuring the percentage of correctly predicted low responders. This measure is often referred to as “specificity at 100% sensitivity” (referred to as SPS subsequently). Secondly, by shifting the threshold value between the extremes where all cases are classified as high and low responders, and plotting the ratio of correctly classified high responders against incorrectly classified low responders, a 2-dimensional curve is constructed. This is called an ROC curve.¹ The performance of a feature in prediction is quantified by the area under this curve (ranging from 0 to 1). This is often called the “area under the ROC curve” measure (we use the abbreviation AUC for this measure).

This previous work suggests that using data from PET scans during chemotherapy can effectively predict the tumour response at the midpoint of chemotherapy, and on this particular data set their method achieved a success rate of 77% measured as SPS and 93% measured as AUC. However there are certain issues on which this analysis can be improved.

1. Only the change of a single attribute, derived from the medical image, was used for prediction in the initial study [17]. One should investigate the benefits of combining several features (e.g. by specifying conjunctions or joint probabilities), and therefore potentially increasing predictive performance.
2. The use of a single attribute from the pre-therapy and midpoint PET scans excludes from the analysis a significant amount of patients who missed the midpoint scan. So using techniques which allow missing data to be estimated should strengthen the study.
3. The evaluation of prediction is performed on the same data that was used to build the predictive model (i.e. choose a threshold value). This

¹The term stands for Receiver Operating Characteristic and originates from the field of signal processing. ROC analysis is widely used in medical data analysis and is gaining popularity within the machine learning community for evaluating 2-class predictive modelling.

introduces the danger of overfitting the training data. Although the constructed models are very simple, the reported predictive performance is optimistic compared to what would be achieved on unseen data. It is considered methodologically better practice to combine model learning with cross-validation.

4. The data was analysed purely from a predictive perspective, i.e. there is no descriptive modelling which might give the domain experts an understanding of the mechanisms underlying the tumour response process. Various machine learning techniques which provide conceptual models should also be included in the analysis.

The aim of the present study was to address these issues using a variety of widely used machine learning algorithms. Such algorithms are able to combine several features straightforwardly. Most of them can either handle missing values, or can be combined with pre-processing methods that fill in these values. Cross-validation is typically applied when evaluating the predictive performance of classification algorithms. This eliminates the danger of overfitting; additionally, it gives a more realistic estimate of future performance of the method. Finally, certain algorithms, apart from defining a decision boundary in the domain space, have a structure that can be intuitively related to the problem domain, revealing interesting patterns in the data. Part of our analysis consisted of finding and discussing such patterns with the domain expert.

The structure of the paper is as follows: in the next section we discuss previous related research. In section 3 we give a brief overview of relevant machine learning algorithms. In section 4 we discuss the details of the data set and present our experimental approach. Section 5 summarises the results achieved by the best-performing classification algorithms we have tested. We conclude in section 6 by summarising the main contributions of this study.

2 Related research

Previous research has been carried out into the ability of PET scans to predict the pathological response of breast cancer tumours to neoadjuvant chemotherapy. In Smith et al. [21] patients were given 8 doses of chemotherapy and PET imaging using [18F]-FDG took place 3 times throughout treatment and once immediately before surgery. The extracted tumour was analysed for pathological response. Results were that after the first PET scan they were able to predict pathological response with sensitivity of 90% and specificity of 74%, and an area under the ROC curve of 86%.

McDermott et al. [17] investigated optimum times for imaging when using PET to predict response to neoadjuvant chemotherapy. They found that by measuring the mean standard uptake value (SUV) at the midpoint of neoadjuvant chemotherapy, they identified 77% of the low responding patients whilst identifying 100% of high responding patients, achieving an area under the ROC curve of 0.93. However, to achieve this the data was filtered to

include only patients who had an initial (pre-therapy) tumour to background ratio of greater than 5.0 in the first PET scan.²

Schelling et al. [20] evaluated the use of PET for prediction with similar breast cancer patients. They took PET images at the start of treatment and after one and two cycles of chemotherapy. Histopathologic response was classified as gross residual disease (GRD) or minimal residual disease (MRD), where GRD were non-responding tumours and MRD were responding tumours to the neoadjuvantive chemotherapy. They identified responders using the SUV of the baseline scan as a threshold. This achieved a sensitivity of 100% and specificity of 85%, and accuracies of 88% after the first scan and 91% after the second scan.

3 Machine learning

Machine learning [18] is a subfield of artificial intelligence involving the automatic construction of models of a domain from observations. Classification algorithms, a subset of machine learning algorithms which are of interest for the purposes of the present analysis, accept as input a finite set of observations (training examples) each of which is associated with a label (also called a *class*) that takes values from a finite domain.³ A training example typically has the form $\langle a_1, \dots, a_n, c \rangle$, where each $a_i \in A_i$ is the value of the i -th *attribute* of the example and $c \in C$ is the value of the *class attribute* for that example.

The output of a classification algorithm is a model (e.g. a set of rules) that accepts a previously unseen observation (test example) $\langle a'_1, \dots, a'_n \rangle$ and predicts a value c' for its class. Normally in a controlled experiment the entire dataset is partitioned into a training and a test set; the former is used by the classification learning algorithm to create a set of predictive rules, and the latter (withholding the class values) is used to generate predictions. These predictions are then compared against the actual values to measure the performance of the classification algorithm. In the present study, an observation is a vector containing some demographic and some PET related data for a patient, while the respective classes are “high response” or “low response”.

Some classification algorithms produce models that are “black boxes” (for example neural networks or regression models), in the sense that their internal structure and/or parameter values are hard to relate to the problem domain. Other methods such as decision rules, decision trees or Bayesian networks yield models which, as well as providing predictions for unseen observations, are readily interpretable by domain experts.

²See also the related sub-section ‘Contrast selection’ in section 4.2

³This is a simple setting using what is called *propositional* data representations. In *relational* or *first-order* representations attributes may take more complex values such as sets or lists of arbitrary length. In the remainder of this paper we will be dealing strictly with propositional representations and methods.

4 Experimental methodology

This section summarises the main stages of our analysis. We applied the WEKA data mining software tool [25] to our data set. Specifically, we used it to select appropriate descriptors; various classification algorithms were then applied to the remaining data set; finally the outcomes of the classification algorithms (in the form of ROC curves) were analysed.

4.1 Data features

The following data features were initially available for each patient: age, pre-therapy body surface area, survival at five years from diagnosis (binary), pathological response of tumour at the end of the treatment (i.e. tumour shrinkage due to chemotherapy) on a scale 1 to 5 (where 1 indicates no response and 5 indicates complete response); additionally, for each of the four PET scans, the injected activity (i.e. the amount of radioactive tracer administered), image contrast, three different measurements of the observed activity (image intensity) in the tumour region, and the derived metabolic volume of the tumour.

For the present analysis a few additional attributes were derived from the initial data. The percentage change between each pair of successive readings of the intensity attributes has been calculated and added to the dataset; the same procedure was also applied to the contrast, injected activity and volume measurements. The class attribute, i.e. the outcome, was derived from pathological response. In accordance with the earlier study, patients with pathological response values 1, 2, and 3 are assigned to the low response class and those with values 4 and 5 to the high response class. Prior to applying the learning algorithm, the response attribute as well as the survival attribute were removed from the dataset.

4.2 Pre-processing

Several versions of the data were produced to enable learning with different combinations of settings to be performed. The following pre-processing steps were implemented:

4.2.1 Discretisation

We studied the effect of applying discretisation to the data prior to classification. On the one hand, this allowed a larger number of algorithms to be used, as some machine learning algorithms can only handle discrete data. As the original dataset contained non-integer data it was necessary to perform some kind of discretisation prior to using the particular algorithms. On the other hand, when using algorithms that can handle both real-valued and discrete data, we wanted to investigate how pre-discretisation would affect the performance of these algorithms.

The first decision we had to make was to determine the cardinality of the discretised domains. Given the number of available instances, we chose to discretise continuous attributes to three-valued scalar domains; this ensured that a sufficient number of instances would be allocated to each of the three values. The discretisation process works as follows: For each of the attributes to be discretised, the data is split into three ordered bins, each containing the same number of instances, such that all values for the particular attribute in one bin are smaller than the values in the next bin. Then, the range of values in every bin is mapped to the index of that bin, yielding a discrete attribute with three possible values. For example, assume that we have nine instances with values 1, 2.5, 3, 5, 5, 11, 13, 14.5 and 15. The bins are $\langle 1, 2.5, 3 \rangle$, $\langle 5, 5, 11 \rangle$, $\langle 13, 14.5, 15 \rangle$. We have the following mappings of ranges to the set $\{1, 2, 3\}$ of discrete values: $(-\infty, 4) \mapsto 1$, $[4, 12) \mapsto 2$, $[12, +\infty) \mapsto 3$. Note that the discretisation process is independent of the class attribute, and therefore introduces no bias in classification.

We noticed that for the algorithms that can handle discrete as well as real values, discretization prior to classification in fact improves classification performance.

4.2.2 Contrast selection

Previous clinical studies indicate that tumour intensity data correlates better with response for scans with an image contrast value which is greater than 5.0 [3, 7, 17]. In the original study against which we are comparing our results, only patients with contrast values > 5.0 in the pre-therapy scan were included in the analysis; the remaining patient records were ignored. Initially we examined the data in two versions, one containing the entire patient population (96 patients) and one containing only records with contrast values > 5.0 in the first PET scan. This sub-group of the total population consisted of 63 patients. Preliminary results suggested in accordance with previous studies that strong correlation with chemotherapy response could not be achieved within the population of patients with low contrast values. Therefore, consistently with [17] we removed from the data set those patients with contrast values < 5.0 in the pre-therapy scan.

4.2.3 Missing values

The previous analysis [17] ignored records with missing values, i.e. patients who have missed one or more scans. On the other hand, most of the machine learning algorithms we have used, are able to handle missing values. We have created two versions of the data, which were analysed independently. In the first version (we call this version **NoMV**), records with missing values were removed completely from the data set. This approach is consistent to the previous study [17] and is used for direct comparison of performance. In the second version (we call this version **MV**), all records are retained and missing values are handled accordingly by the algorithms.

4.2.4 Prediction at different points of treatment

Two versions of the data set were created with respect to the timings of the scans. One version (**A**) contained all the information from the pre-therapy scan, the first scan after the start of chemotherapy, and the changes between the two scans. The second version (**B**) contained information from the pre-therapy, the first scan and the mid-point scan plus the changes between the successive scans. In an actual clinical setting, using different versions would correspond to making predictions at different time-points of the chemotherapy treatment.

We also examined a further third version (**C**) which contained all the scan data, including the final scan after the end of the chemotherapy. Preliminary experiments revealed that using this data gives no improvement to the predictive performance of the learning algorithms. This was an initially surprising observation; although prediction at the end of the treatment would not be useful from a clinical viewpoint, we would expect that having the entire data set would provide the gold standard for data analysis. However this was considered a reasonable result by the domain expert; as the uptake of FDG by the tumour reduces as the chemotherapy treatment progresses. There is minimal change shown in the uptake of FDG between the midpoint and endpoint of chemotherapy. This leads to a decrease in the quality of the image data which shows only the FDG uptake, not necessarily the status of the actual tumour. This is a consistent finding as similar studies have shown the same effect [2, 24], however the reason for the fall off of FDG is not yet properly understood. It has been suggested that chemotherapy induces vascular damage, resulting in a drop in blood supply to the tumour and therefore a drop in FDG reaching the tumour [15, 23]. This implies that FDG-PET becomes less sensitive to metabolic changes as chemotherapy progresses. Another process that could effect the quality of imaging in the late stages of treatment is the presence of immune or stromal cells removing chemo-sensitive cells. These agents result in an increased activity observed through the PET scan in the tumour region of interest, which can be misleading [22].

In short, since the use of the data from the final PET scan could neither improve predictive performance nor give any clinical benefit, in our study we analysed thoroughly only versions **A** and **B**.

4.2.5 Attribute selection

Independently from classification, attribute selection was carried out to select the most informative attributes from the data set. Attribute selection was based on an attribute evaluation method which calculates the chi-squared statistic with respect to the class. The attributes were ranked and the highest ranking attributes were chosen until the chi-squared statistic indicated that including additional attributes was having little effect.

Attribute selection was combined with cross-validation to ensure no overfitting was introduced in the process, and that we do not bias the attribute

selection in favour of the classifier algorithms. The data was split into the same 10 folds which were later used in classification. For each of the folds, only the training partition of the data was used to derive the selected attributes.

In the case of the data sets with pre-therapy and first therapy scans, we selected 6 out of 19 attributes which scored the highest on the chi-squared evaluation. For all folds, these attributes were:

- The three attributes corresponding to percentage changes in image intensity between scans
- The percentage change of the derived tumour volume
- the absolute tumour volume at the first therapy scan
- the percentage change in image contrast

In the case of the data sets containing the additional mid point scan information 13 out of 30 attributes were selected and were used in 7 out of the 10 folds. In these 7 folds although the same attributes were used, the features often ranked differently in the various folds. In the remaining 3 folds either one or two attributes were different (from the above 13 attributes). The following 9 attributes were common to all ten folds:

- Two of the measurements of image intensity at the mid point scan
- Percentage changes in two of the measurements of image intensity between the first and the second scans and the second and the third scans
- Percentage change in tumour volume between first and second scans
- Tumour volume of the third scan
- Image contrast of the third scan

4.3 Classification

A variety of classification algorithms were trained and evaluated using the WEKA software, including decision trees, Bayesian methods, and instance-based learning. Each classifier was applied to each of the variations of the data produced in pre-processing. These were derived from combining the options of first versus second point in the treatment, and including versus excluding records with missing values; therefore there were 4 versions of the data. We refer to experimental settings where records with missing values were retained as **MV** and to settings where they were discarded as **NoMV**. Index **A** denotes the use of data from pre-therapy and start of therapy scans, and index **B** the use of pre-therapy, start, and mid-point data. Table 1 shows the total number of instances and the distribution between the high and low response classes in the various settings. Note that the patients in the group **NoMV_B** are a

Table 1: Low response, high response, and total number of instances in different settings.

	NoMV _A	NoMV _B	MV _A , MV _B
Low response	34	25	45
High response	13	10	18
Total	47	35	63

sub-group of those in NoMV_A, which in turn are a sub-group of the ones in the MV groups.

Initial experiments suggested that contrast selection, and attribute selection, invariably improved the performance of classification algorithms. So these were applied in all the different settings. The same was observed with discretisation, where it was applicable.

4.4 Evaluation

In order to test how well a particular model (output of a classifier learning algorithm) would perform on unseen data, stratified 10-fold cross validation was performed. This type of validation randomly divides the data set into 10 folds (preserving the same ratio of class values in all folds). The classifier trains on nine-tenths of the data and tests the classifier on the remaining data. This is then repeated 10 times, each time testing on a different fold of data, and the performances over the different folds are averaged to yield the overall performance on the entire data. The performance of each classifier was evaluated as both specificity at 100% sensitivity (SPS) and area under the ROC curve (AUC), as explained previously.

5 Results

In this section we summarise the most interesting results obtained by the classification algorithms from various experimental settings. We report both the area under the ROC curve measure (AUC) and the specificity rate at 100% sensitivity (SPS).

Using the data from the pre-therapy and start of therapy scans, the performance was in general significantly worse than when using additionally the mid-point of therapy data. In addition we observed that numerical models (such as linear regression models, neural networks and support vector machines) did not match the best performance obtained by other algorithms. Table 2 summarises the performance of the best-performing algorithms we have tested. NB is the Naive Bayes classifier [14]. Bayesian network classifiers [12] were tested using three different methods for structure learning: Cooper and Herskovits' K2 algorithm [6], referred to as BN/K2; tabu search referred to as BN/Tabu; and the tree-augmented naive Bayes algorithm [10], referred to as BN/TAN.

Table 2: Classification performances (percentage) of various algorithms and results of previous research [17]. Note that previous results are optimistic as no cross-validation was performed. The abbreviations for the algorithms are given in the text.

	NoMV_A		NoMV_B		MV_A		MV_B	
	AUC	SPS	AUC	SPS	AUC	SPS	AUC	SPS
NB	88.0	70.6	94.4	84.0	78.6	60.0	92.2	68.9
BN/K2	88.5	70.6	94.4	84.0	79.0	62.2	93.0	82.2
BN/Tabu	85.6	67.6	85.6	72.0	72.3	42.2	91.9	80.0
BN/TAN	90.1	70.6	96.0	88.0	74.1	48.9	87.8	77.8
C4.5	79.9	52.3	63.8	0.0	47.0	17.8	65.6	15.6
ADTree	86.0	47.1	78.6	4.0	80.9	40.0	74.4	42.2
NBTree	89.8	73.5	93.2	76.0	78.3	40.0	90.6	66.7
5-NN	86.7	67.6	92.0	80.0	71.5	51.1	88.1	62.2
<i>previous</i>	88	66	93	77				

C4.5 is the well-known decision tree algorithm [19]; alternating decision trees (ADTree) [9] and naive Bayes trees (NBTree) [16] were used as well. 5-NN is the k -nearest neighbour algorithm [1] with $k = 5$ and no distance weighting. Note that the setting **NoMV_B** is the easiest of all experimental settings, since it includes more information per patient (data from three PET scans), and records with missing values (which are harder to handle), are discarded. In contrast, the setting **MV_A** is the hardest of the experimental settings: it contains measurements of only the first two PET scans, and also includes patient records with missing values (essentially patients who missed the second scan, i.e. for whom only the pre-therapy scan was available).

We observe that Bayesian classifiers perform consistently well. Among these, BN/TAN performed better than the rest in the absence of missing values while BN/K2 outperformed the others when missing values were present. Decision trees give interesting results, but have worse performance measured as SPS; this is due to the fact that they are less flexible than probabilistic classifiers in handling varying mis-classification costs. Reasonably high performance was also obtained by naive Bayes trees and by the k -nearest neighbour algorithm in all experimental settings. (Further, the run-time for most algorithms did not exceed a few seconds; however, naive Bayes trees took up to a minute).

Combining the figures from Tables 1 and 2 we observe that, in the setting **NoMV_A**, BN/TAN classifies correctly 25 out of 34 low responders at the 100% sensitivity point. In **NoMV_B**, the ratio is 22/25 for BN/TAN. Note also that the ability to handle missing values gives a strong advantage to machine learning methods: When records with missing values are included in the analysis, the BN/K2 algorithm classifies correctly 28 out of 45 low responders at 100% sensitivity in the setting **MV_A**, and 37 out of 45 low responders in the setting **MV_B**. Therefore, it is important to note that while the SPS rates are lower

than the cases where missing values are discarded for the same time points, in fact a greater number of low response patients can be correctly identified, without mis-classifying any of the high response patients.

6 Conclusions

In this paper we have discussed the use of machine learning algorithms for analysing PET imaging to predict response to chemotherapy in breast cancer patients. We have evaluated several algorithms using real-world clinical data. Our methodology has shown clear advantages compared to previous approaches. Firstly, some machine learning algorithms outperform previous methods applied to the same data. Secondly, an important advantage of the machine learning algorithms is that they may be applied to clinical cases where missing values occur. These factors suggest that machine learning algorithms are highly suitable for constructing predictive models in this domain.

Additionally, some of the models produced by probabilistic and symbolic machine learning algorithms have been interpreted by the domain expert, who found them to be of clinical interest. Due to the small size of the available data set, no statistically significant conclusion could be obtained. We believe that using data from larger-scale clinical studies in this domain can lead to the development of clinically insightful models.

7 Acknowledgements

This work is supported by the EPSRC sponsored Advanced Knowledge Technologies project, GR/NI5764, which is an Interdisciplinary Research Collaboration involving the University of Aberdeen, the University of Edinburgh, the Open University, the University of Sheffield and the University of Southampton. Additionally we would like to acknowledge helpful feedback from a reviewer.

References

- [1] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.
- [2] Avril N., Sassen S., Schmalfeldt B. et al. Prediction of response to neoadjuvant chemotherapy by sequential F-18-fluorodeoxyglucose positron emission tomography in patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology*, 23:7445–7453, 2005.
- [3] Black Q.C., Grills I.S., Kestin L.L., et al. Defining a radiotherapy target with positron emission tomography. *Int J Radiation Oncology Biol Phys*, 60:1272–82, 2004.
- [4] Bonadonna G., Valagussa P., Zucali R., et al. Primary chemotherapy in surgically resectable breast cancer. *CA Cancer J Clin*, 45:227–243, 1995.
- [5] Cancer Research UK. <http://www.cancerresearchuk.org/>.

- [6] Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
- [7] Erdi Y.E., Mawlawi O., Larson S.M., et al. Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding. *Cancer*, 80:2505–9, 1997.
- [8] Fisher B., Brown A., Mamounas E., et al. Effect of preoperative chemotherapy on loco-regional disease in women with operable breast cancer: Findings from National Surgical Adjuvant Breast and Bowel Project B-18. *J Clin Oncol*, 15:2483–2493, 1997.
- [9] Y. Freund and L. Mason. The alternating decision tree learning algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, pages 124–133, Bled, Slovenia, 1999.
- [10] Nir Friedman, Dan Geiger, and Moisés Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.
- [11] Gennari A., Donati S., Salvadori B., Giorgetti A., Salvadori P. A., Sorace O., Puccini G., Pisani P., Poli M., Dani D., Landucci E., Mariani G., Conte P. F. Role of 2-[18F]-fluorodeoxyglucose(FDG) positron emission tomography(PET) in the early assessment of response to chemotherapy in metastatic breast cancer patients. *Clin Breast Cancer*, pages 156–61, 2000.
- [12] Heckerman D. Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery*, 1:79–119, 2004.
- [13] Jansson T., Westlin J.E., Ahlstrom H., Lilja A., Langstrom B. and Bergh J. Positron emission tomography studies in patients with locally advanced and/or metastatic breast cancer: a method for early therapy evaluation? *Journal of Clinical Oncology*, 13:1470–1477, 1995.
- [14] George H. John and Pat Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, 1995.
- [15] Kaushal V., Kaushal G.P., Mehta P. Differential toxicity of anthracyclines on cultured endothelial cells. *Endothelium*, 11:253–258, 2004.
- [16] Ron Kohavi. Scaling up the accuracy of naive Bayes classifiers: a decision tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.
- [17] McDermott G.M., Welch A., Staff R.T., Gilbert F.J., Schweiger L., Semple S.I.K., Smith T.A.D., Hutcheon A.W., Miller I.D., Smith I.C., Heys S.D. Monitoring primary breast cancer throughout chemotherapy using FDG-PET. *Breast Cancer Research and Treatment*, 2006.
- [18] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
- [19] Quinlan J.R. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [20] Schelling M., Avril N., Nährig J., Kuhn W., Römer W., Sattler D., Werner M., Dose J., Jänicke F., Graeff H. . Positron Emission Tomography Using [18F]-Fluorodeoxyglucose for Monitoring Primary Chemotherapy in Breast Cancer. *Journal of Clinical Oncology*, 18:1689–1695, 2000.
- [21] Smith I.C., Welch A.E., Hutcheon A.W., Miller I.D., Payne S., Chilcott F., Waikar S., Whitaker T., Ah-See A.K., Eremin O., Heys S.D., Gilbert F.J., Sharp

- P.F. Positron Emission Tomography Using [18F]-Fluorodeoxy-D-Glucose to Predict the Pathologic Response of Breast Cancer to Primary Chemotherapy. *Journal of Clinical Oncology*, 18:1676–1688, 2000.
- [22] Spaepen K., Stroobants S., Dupont P., et al. [18F]FDG PET monitoring of tumour response to chemotherapy: does [18F]FDG uptake correlate with the viable tumour cell fraction? *Eur J Nucl Med Mol Imaging*, 30:682–688, 2003.
- [23] Wakabayashi I., Groschner K. Vascular actions of anthracycline antibiotics. *Curr Med Chem*, 10:427–436, 2003.
- [24] Wieder H.A., Brucher B.L.D.M., Zimmermann F. et al. Time course of tumour metabolic activity during chemoradiotherapy of esophageal squamous cell carcinoma and response to treatment. *Journal of Clinical Oncology*, 22:900–908, 2004.
- [25] Witten I.H., Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.