ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Machine Learning in cancer research: implications for personalised medicine

Alfredo Vellido[1], Elia Biganzoli[2], and Paulo J.G. Lisboa[3] *

1- Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3, Campus Nord. Barcelona, 08034, Spain.

2- Istituto di Statistica Medica e Biometria, Campus Cascina Rosa
Istituto Nazionale Tumori - Università degli Studi di Milano
Via G. Venezian 1, 20133 Milan, Italy.

3- School of Computing and Mathematical Sciences
Liverpool John Moores University
Byrom St., L3 3AF, Liverpool, United Kingdom

Abstract. Driven by the growing demand of personalization of medical procedures, data-based, computer-aided cancer research in human patients is advancing at an accelerating pace, providing a broadening landscape of opportunity for Machine Learning methods. This landscape can be observed from the wide-reaching view of population studies down to the genotype detail. In this brief paper, we provide a sweeping glimpse, by no means exhaustive, of the state-of-the-art in this field at the different scales of data measurement and analysis.

1 Introduction

The growing demands from ever better informed patients, with their increasingly sophisticated expectations from doctors and from health systems as a whole, conform one of the forces driving the current trend towards personalized medicine. The true personalization of medical procedures is a difficult task, and balancing the improvement of healthcare delivery it entails and the corresponding escalating costs in all its phases -prevention, diagnosis, prognosis, and therapy- maybe beyond reach for most health systems.

Cancer research in human patients is, in any case, advancing at an accelerating pace driven by new data modalities, a socioeconomic need for cost containment through prevention and early diagnosis, and the increasing patient demands for the personalisation of therapy. These developments accord with the 4P prospective medicine agenda (predictive, preventive, personalised and participatory) [1], which recognises the pivotal role that experimental and observational data have in systems biology. Whereas this agenda is often closely coupled with the huge potential of pharmacogenomics, which has information

*Alfredo Vellido is a researcher within the Ramón y Cajal program of the Spanish MEC and acknowledges funding from the MEC I+D project TIN2006-08114. Paulo J.G. Lisboa and Elia Biganzoli acknowledge funding from the Biopattern Network of Excellence FP6/2002/IST/1; N. IST-2002-508803, as well as numerous discussions and the active participation of the Biopattern consortium members.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

processing at is very core, 4P is in reality much broader than this, covering the complete range of observation scales from genotype through to population cohorts, as shown in Table 1.

This tutorial is a guided tour of this landscape of clinical need and enabling data-based science, much of which relies on, and requires further advances in, Machine Learning (ML) methods. In order for research to retain clinical relevance, it is necessary to distinguish clearly between methodological studies, which form the basis of this brief review, and the development of medical DSS, for which formal frameworks and guidance are provided elsewhere [2].

This paper is structured top-down following the hierarchy of clinical measurement scales, starting with some of the unfortunately rare examples of diagnostic support algorithms and systems that have been taken as far as prospective evaluation. This is followed by an overview of flexible models for the analysis of survival data, which is a mainstream task to study prognostic outcome over time, a critical factor in determining choice of therapy. Then, we provide a glimpse of the rapidly growing field of multi-modal imaging, where anatomical signals are fused with metabolic data often at different spatial resolution and with high dimensionality. Drilling down to the next level of study, cancer histology is moving towards characterization of disease biology, starting with exploratory clustering and visualization studies looking for tumour subtypes. Finally, there is much interest in mapping out networks of protein activation, possibly the route from genotype to phenotype, or disease expression.

In all these research enterprises, ML experts have to fight the ground of acceptance by medical institutions and practitioners. Therefore, the role of ML methods in human cancer research should be carefully defined and justified. One way to ensure this is by consolidating their development through international large-scale research projects. Some examples of these in the area of cancer research in Europe are described in section 3.

2 Application of machine learning methods to human cancer research

This section illustrates the application of ML methods to the different study types listed in Table 1, emphasizing key factors that influence the clinical relevance of the work by reference to particular case studies. Comprehensive reviews of applications can be found elsewhere [3, 4]. In the following, informative case studies are drawn from the literature with particular reference to current applications developed with funding from the Biopattern European Network of Excellence (www.biopattern.org).

2.1 Predictive diagnostics and rule generation for translational research

One of the first diagnostic projects to thoroughly benchmark linear statistics with neural networks and kernel methods, was the IOTA project (International

Ovarian Tumour Analysis). This project has advanced as far as prospective evaluation of the accuracy of diagnostic algorithms, that is to say where models are kept fixed and applied predictively by clinicians during consultation. The second phase of this project has records from over 2,000 patients.

External validation, i.e. on IOTA data from clinical centres different from those from where the training data were collected, showed similar performances between linear and non-linear logistic regression models, but suggested that kernel models such as the Suport Vector Machine (SVM) and the Relevance Vector Machine (RVM) may be more accurate [5]. It is interesting that this paper reported better prediction for models using measurements that did not require higher ultrasound skills, suggesting the reliability of the measurements may have a significant impact on the results. This points towards the importance of suitable data standards and clinical protocols, a matter sometimes given scant attention by the ML community before embarking on empirical modelling - yet a critical factor the value of the study [2].

The selection of logistic regression for the clinical interface in the former study should not surprise us. It is the interpretability of this model that makes it attractive to clinicians. One of the potential drawbacks affecting the application of ML methods in general is indeed the usually limited interpretability of the results they yield. This is again an extremely sensitive issue in a critical context such as clinical oncology and, therefore, it is of interest to translate the analytical models into the domain language of clinical experts. One way to do this is by explaining the operation of ML models using rule extraction methods. Several authors have, in recent years, resorted to rule extraction from ML models in cancer research. Many of these involve the analysis of breast cancer data [6, 7, 8], although rule extraction from the classification of other cancer pathologies such as ovarian tumours [9] has also been implemented.

2.2 Survival data analysis

Once a cancer is diagnosed, attention turns towards prognosis and the choice of therapy. Prognostic modelling requires a smooth fit of outcome data collected over varying periods of time, ranging from a few days to several decades. This is known as time-event-modelling, where the event of interest may be remission due to the therapy, or recurrence, or mortality, either specifically due to the disease or from any cause. Once again, the definition of the patient cohort to investigate is critical to the clinical value of the study, and preferences vary between clinical domains and even across geographical areas, even within Europe. Nevertheless, from the statistical viewpoint, this task is characterised by the occurrence of censorship, or loss to follow-up.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Scale	Clinical drivers	Study types	Relevance to ML methods
Population	Integrated support for patients.	Epidemiology	Formal acquisition and validation platforms: data standards and clinical protocols for prospective and retrospective studies.
Individuals (system level)	Disease Biology and Computer-Aided Decision Support.	Personalised Decision Support with Interfaces for Clinicians and Patients; Personalised Monitoring; and Point of Care Diagnostics.	Predictive diagnostics and rule generation for translational research
Clinical signs (organ level)	Disease Biology, Computer-Aided Decision Support, Response to Therapy, and Tumour Delineation.	Personalised Decision Support with Interfaces for Clinicians and Patients; Personalised Monitoring; and Point of Care Diagnostics	Longitudinal data analysis for prognostic modelling
Physiological measurement (tissue level, $10^{-3} - 10^{-2}$m)	Disease Biology, Computer-Aided Decision Support, Response to Therapy, and Tumour Delineation.	Multimodal Data Fusion; Virtual Physiological Human.	Fusion of anatomical and functional data including different measurement modalities
Immunohistochemistry (cell level, $10^{-5} - 10^{-4}$m)	Disease Biology, Computer-Aided Decision Support, Response to Therapy, and Tumour Delineation.	Integration with higher level data linked to clinical expertise about the pathology, and Pharmacogenomics.	High-dimensional data modelling for diagnostic and laboratory imaging
Phenotype (pathways, $10^{-7} - 10^{-6}$m)	Disease Biology, Computer-Aided Decision Support, Response to Therapy, and Tumour Delineation.	Integration with higher level data linked to clinical expertise about the pathology, and Pharmacogenomics.	Knowledge discovery from data, clustering and visualization
Genotype (genes, $10^{-9} - 10^{-8}$m)	Disease Biology, Computer-Aided Decision Support, and Response to Therapy.	Exploratory data analysis for hypothesis generation.	Modelling protein networks, with fewer observations than potential predictors

Table 1: Overview of the role of data analysis in human cancer research.

Prognostic modelling for single and multiple competing risks was originally proposed with the Partial Logistic Artificial Neural Network (PLANN) in [10]. This model can be regarded as an extension of the Multi-Layer Perceptron to model censored data within the theoretical framework of Generalised Linear Models. It has since been extended in [11]. Clearly the potential for overfitting is ever present in non-linear modelling, so the well-known Bayesian regularisation framework, with a multivariate normal approximation of the evidence term, was applied in [12]. Alternative Bayesian frameworks have also been proposed for survival modelling, notably the Conditional Hazard Estimation Neural Network (CHENN) model [13] and, more recently, an SVM implementation featured in this special session [14].

In addition to modelling the conditional event probability over time, it is essential to have rigorous measures of the ability of the model to discriminate between patients at different levels of risk. This requires a new index that is closely related to the AUROC but applies over time and in the presence of censorship, known as the time-dependent c-index (C_{TD}) [15]. A double-blind evaluation of the performance accuracy for a range of linear, spline and neural network survival models was carried out for a large data set of uveal melanoma (cancer of the eyeball) in [16].

The rule extraction framework can be applied to severity of risk indices derived for longitudinal data, showing that it is possible to use the neural network as the generating function to derive transparent Boolean risk-allocation rules that stratify patients by risk practically as accurately as the original network model [17].

2.3 Multi-modal data fusion

Over the last two decades, computing-based advances in biomedical engineering have led to ever more complex measurement, especially moving from anatomical imaging, which is restricted to identifying shapes, to functional imaging. Of particular importance for the future is the fusion of anatomical and physiological modalities. For instance, Magnetic Resonance Imaging (MRI) measures only the density of free water in tissue, while MR Spectral Imaging (MRSI) is specific to particular metabolites derived from a multiple localised MR Spectra. Once again, kernel methods have been successful in carrying out this data fusion [18]. However, there have been approaches to decision support based on visualisation, rather than discrimination. This is the so-called nosological imaging, where patient-specific topographic visualisation combining MRI/MRS is created with low-dimensional visualisation mappings of MRS data, using GTM or SOM, and then projected back into the spatial MRI space [19]. These methodological developments are showing potential for clinical decision support for tumour delineation, both pre-operatively and for guided surgery. Finally, imaging of laboratory cytology is understudied by the ML community, given the clinical importance and apparent subjectivity of some of the key tests required to assess cancer progression. For a review on this subject, see [20].

2.4 Knowledge discovery from data, clustering and visualization

Historically, the progression of cancer has been measured clinically from the size ot he tumour (T-stage), involvement of the lymphatic system (N-stage) and Metastatic spread (M-stage), which together form the TNM staging recommended by the World Health Organisation. The next step from this is to evaluate the differentiation between the tumour cells from the original normal cells from which they grew. This requires both identification of relevant markers (e.g. [21]) and the development of robust clustering and visualisation algorithms, preferably with the capacity to produce low-dimensional linear projections with minimal mixing of pre-defined cluster labels.

An orthogonal methodological approach is to define latent spaces on which to characterise high-dimensional data, in particular for identification of outliers. This can benefit from the use of heavy tailed basis functions within accepted Bayesian visualisation models [22]. The aspect of visualisation is also developed further in this special session [23].

New signal modalities tend to be high dimensional and are very difficult to interpret in isolation, thus relying on multivariate analysis. One aspect of this still current research is whether to use multivariate selection or resort to projective dimensional reduction methods, such as Principal Components Analysis (PCA) and Independent Components Analysis (ICA). An early study comparing the two approaches showed that ICA identifies in brain tumour spectroscopy two main signal types, closely resembling necrotic tissue and infiltrating tissue. These two degrees of freedom are useful in the differential diagnosis of high-grade astrocytic tumours [24]. More recent work with kernel methods has explored the concept of bagging to achieve robustness in discrimination [25]. The issue of feature selection is revisited in two papers from this special session [26, 27].

2.5 Modelling protein networks with fewer observations than potential predictors

A related approach to clustering, but set typically within the framework of Bayesian Graphical Models, is model-based clustering. One such approach is to layer the graph to form a latent model that is applicable even to very high dimensional data, such as Tissue Micro Arrays (TMA) which may comprise several thousand measurements yet total a sample size of only a few hundred patients. The approach of Latent Dirichlet Allocation has been applied to a set of breast cancer markers, resulting in an initial differentiation of sub-types which will no doubt be followed-up by studies from other clinical centres [28].

Clearly, probe selection is also a necessary part of the exploratory analysis of data for bioinformatics, and this is the subject of a paper in this special session [29]. Interestingly, this paper is concerned with adjuvant therapy that is applied prior to surgical intervention. The need to early adjuvant therapy is also motivated by the "dormancy" hypothesis, initially substantiated by careful analysis of event-rate curves in longitudinal data modelling for patients with early breast cancer. It was found that a proportion of these succumbed to

distant metastases within a relatively short period of having surgery. The current hypothesis is that specific sub-types will spread to other organs early but emit systemic signals to keep the metastases dormant.

A further and final consideration under this section is the need to careful regularisation of any statistical models with more predictors than observations. In particular, the independence relationships between covariates which form the basis for graphical models, can be difficult to establish robustely since they may not be observed under particular conditional constraints. This is the issue of "faithfulness" for which the Partial Correlation algorithm provides a computationally efficient solution [30]. The impact of approximate deterministic relationships on faithfulness is followed-up in this special session [31].

3 Machine Learning for cancer research in current European research projects

The European Commission Information and Communication Technologies (ICT) for Health Unit of the Information Society and Media Directorate General, manages a series of international research projects in the medical ambit, funded under the Sixth Framework program (FP6). Within the program's 4^{th} call, "Integrated biomedical information for better health", several projects concern cancer research more or less directly. All these projects involve, in one way or another, data analysis, and some of them realize it through data mining or computational intelligence methods, often related to ML. In Table 2, we can find a summary of the general goals of some of these projects, including more specific data analysis goals and the use in them of ML and related methods.

4 Conclusion

The latest data [32] of the EUROCARE research collaboration indicate that the gap between European countries in cancer survival is narrowing, suggesting substantial improvement in cancer care in countries with poor survival.

One of the keys to the improvement of cancer survival figures is discovery-driven translational research and, as part of it, the development of efficient medical DSS for the support of medical diagnosis and prognosis of pathologies. ML can provide, coalescing with more traditional statistical approaches, robust methods to be deployed at the core of these DSS. In such an integrated framework a key role is played by a substantial evaluation component, which, according to the principles of Evidence-Based Medicine should be covered by Biostatistics, to be merged with the ML approach.

In order to pass from adolescence to fruitful maturity, ML studies of medical data must broaden their focus beyond technical detail to pay greater attention to medical requirements. Over the last decade, clinicians and medical researchers have become more aware of what these methods are and what can they achieve, so that closer collaboration with them should help the data analyst to drive data-based studies according to key clinical questions, therefore building into study

design relevance. A further challenge that ML practitioners will have to face in coming years is that of data management, as multi-centre and international databases become more of a standard, while the available biomedical signal for analysis becomes increasingly multivariate, multiscale and multimodal [33].

References

[1] Institute for Systems Medicine (ISM). 4P Medicine Agenda. URL: www.ismedicine.org/Section_P4/P4.aspx

[2] P.J.G. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, *Neural Networks*, 15:11-39, 2002.

[3] P.J.G. Lisboa and A.F.G. Taktak, The use of Artificial Neural Networks in decision support in cancer: a systematic review, *Neural Networks*, 19:408-415, 2006.

[4] A. Vellido and P.J.G. Lisboa, Neural networks and other machine learning methods in cancer research. In J. Cabestany *et al.*, editors, *proceedings of the* 8^{th} *International Work-Conference on Artificial Neural Networks* (IWANN 2007), Springer, pages 964-971, San Sebastian, Spain, LNCS Vol.4507, 2007.

[5] C. Van Holsbeke, B. Van Calster, L. Valentin, *et al.*, External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the International Ovarian Tumor Analysis Group, *Clinical Cancer Research*, 13:4440-4447, 2007.

[6] K.C.Tan, Q.Yu, C.M. Heng, T.H. Lee, Evolutionary computing for knowledge discovery in medical diagnosis, *Artificial Intelligence in Medicine*, 27:129-154, 2003.

[7] A.E. Hassanien, Fuzzy rough sets hybrid scheme for breast cancer detection, *Image and Vision Computing*, 25:172-183, 2007.

[8] P.J.G. Lisboa, T.A. Etchells, I.H. Jarman, *et al.*, Development of a rule based prognostic tool for HER 2 positive breast cancer patients. In *proceedings of the* 29^{th} *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE-EMBC), IEEE Press, pages 5416-5419. Lyon, France, 23-26 Aug 2007.

[9] M.S.H. Aung, P.J.G. Lisboa, T.A. Etchells, *et al.*, Comparing analytical decision support models through Boolean rule extraction: a case study of ovarian tumour malignancy. In *proceedings of the* 4^{th} *International Symposium on Neural Networks, part II* (ISNN 2007), Springer, pages 1177-1186, Nanjing, China, June 3-7, LNCS Vol.4492, 2007.

[10] E. Biganzoli, P. Boracchi and E. Marubini, Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach, *Statistics in Medicine* 17:1169-1186, 1998.

[11] E. Biganzoli, P. Boracchi, F. Ambrogi and E. Marubini, Artificial neural network for the joint modelling of discrete cause-specific hazards, *Artificial Intelligence in Medicine*, 37: 119-130, 2006.

[12] P.J.G. Lisboa, H. Wong, P. Harris and R. Swindell, A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer, *Artificial Intelligence in Medicine*, 28:1-25, 2003.

[13] A. Eleuteri, M.S.H. Aung, A.F.G. Taktak, B. Damato and P.J.G. Lisboa, Continuous and discrete time survival analysis: neural network approaches. In *proceedings of the* 29^{th} *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC 2007), IEEE Press, pages 5420-5423, Lyon, France, 23-26 Aug, 2007.

[14] V. Van Belle, K. Pelckmans, J. Suykens and S. Van Huffel, Survival SVM: a practical scalable algorithm. In M. Verleysen, editor, *proceedings of the* 16^{th} *ESANN* , d-side pub., current volume, April 23-25, Bruges (Belgium), 2008.

[15] L. Antolini, P. Boracchi and E. Biganzoli, A time-dependent discrimination index for survival data, *Statistics in Medicine*, 24:3927-3944, 2005.

[16] A. Taktak, L. Antolini, M.S.H. Aung, *et al.*, Double-blind evaluation and benchmarking of survival models in a multi-centre study, *Computers in Biology and Medicine*, 37:1108-1120, 2007.

[17] P.J.G. Lisboa, E. Biganzoli, A.F. Taktak, *et al.*, Assessing flexible models and rule extraction from censored survival data. In *proceedings of the* 20^{th} *International Joint Conference on Neural Networks* (IJCNN 2007), IEEE Press, pp. 1663-1668, Florida, 2007.

[18] J. Luts , A. Heerschap , J.A.K. Suykens and S. Van Huffel, A combined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection, *Artificial Intelligence in Medicine*, 40:87-102, 2007.

[19] M. De Vos, T. Laudadio, A.W. Simonetti, A. Heerschap, S. Van Huffel, Fast nosologic imaging of the brain, *Journal of Magnetic Resonance*, 184:292-301, 2007.

[20] Z. Theodosiou, I. Kasampalidis, G. Livanos, M. Zervakis, I. Pitas and K. Lyroudia, Automated analysis of FISH and immunohistochemistry images: a review, *Cytometry Part A* 71A:439-450, 2007.

[21] M.E. Blazadonakis and M. Zervakis, Polynomial and RBF kernels as marker selection tools - a breast cancer case study, In *Proceedings of the* 6^{th} *International Conference on Machine Learning and Applications*, Cincinnati, Ohio 13-15 Dec. 2007.

[22] A. Vellido and P.J.G. Lisboa, Handling outliers in brain tumour MRS data analysis through robust topographic mapping, *Computers in Biology and Medicine*, 36:1049-1063, 2006.

[23] E. Romero, M. Julià-Sapé and A. Vellido, DSS-oriented exploration of a multi-centre magnetic resonance spectroscopy brain tumour dataset through visualization. In M. Verleysen, editor, *proceedings of the* 16^{th} *ESANN*, d-side pub., current volume, April 23-25, Bruges (Belgium), 2008.

[24] Y. Huang, P.J.G. Lisboa and W. El-Deredy, Tumour grading from Magnetic Resonance Spectroscopy: a comparison of feature extraction with variable selection, *Statistics in Medicine*, 22:147-164, 2003.

[25] C. Lu, A. Devos, J.A.K. Suykens, C. Arús, S. Van Huffel, Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis, *IEEE Transactions on Information Technology in Biomedicine*, 11:338-347, 2007.

[26] F.F. González, L.A. Belanche, Feature selection in proton magnetic resonance spectroscopy for brain tumor classification. In M. Verleysen, editor, *proceedings of the* 16^{th} *ESANN*, d-side pub., current volume, April 23-25, Bruges (Belgium), 2008.

[27] A. Barla A, S. Mosci, L. Rosasco, A. Verri, A method for robust variable selection with significance assessment. In M. Verleysen, editor, *proceedings of the* 16^{th} *ESANN* 2008, d-side pub., current volume, April 23-25, Bruges (Belgium), 2008.

[28] L. Carrivick, S. Rogers, J. Clark, C. Campbell, M. Girolami and C. Cooper, Identification of prognostic signtures in breast cancer microarray data using Bayesian techniques, *Journal of the Royal Society: Interface*, 3:367-381, 2006.

[29] R. Natowicz, A. Braga, R. Incitti, E. Horta, R. Rouzier, T. Rodrigues and M. Costa, A new method of DNA probes selection and its use with multi-objective neural network for predicting the outcome of breast cancer preoperative chemotherapy. In M. Verleysen, editor, *proceedings of the* 16^{th} *ESANN*, d-side pub., current volume, April 23-25, Bruges (Belgium), 2008.

[30] M. Kalisch and P. Bühlmann, Estimating high-dimensional directed acyclic graphs with the PC-Algorithm, *The Journal of Machine Learning Research*, 8: 613-636, 2007.

[31] S. Rodrigues de Morais, A. Aussem and M. Corbex, Handling almost-deterministic relationships in constraint-based Bayesian network discovery: Application to cancer risk factor identification. In M. Verleysen, editor, *proceedings of the* 16^{th} *ESANN*, d-side pub., current volume, April 23-25, Bruges (Belgium), 2008.

[32] F. Berrino, R. De Angelis, M. Sant, *et al.*, Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995-99: results of the EUROCARE-4 study, *The Lancet Oncology*, 8:773-783, 2007.

[33] S. Cerutti, Multivariate, multiorgan and multiscale integration of information in biomedical signal processing. Keynote lecture. In *proceedings of the* 1^{st} *International Joint Conference on Biomedical Engineering Systems and Technologies*, (BIOSTEC 2008), INSTICC Press, pages IS5-IS8, January 28-31, Funchal (Portugal), 2008.

Project name	General goals	Use of ML and related methods
ACGT	It aims to fill-in the technological gaps of clinical trials for breast cancer and paediatric nephroblastoma. The project will develop a Biomedical GRID infrastructure for sharing clinical and genomic expertise, helping to identify what determines which form of treatment suits which patient.	Data mining tools, using R language in a grid environment, including SOM, k-Means and Sammon's mapping in combination with statistical techniques. These are combined with tools for interactive visualization. It includes a data mining methodological framework.
ASSIST	The project aims to provide medical researchers of cervical cancer with an environment that will unify multiple patient record repositories. They will be able to combine phenotypic and genotypic data and perform association studies on larger multicenter sets of patient records.	It claims that, given a hypothesis that needs validation, the system must be able to process relevant records and mine the collected data. Mixed text and data mining are integrated in a workflow process using weighted fuzzy methods, neural networks, and support vector machines.
Biopattern	This project's goal is to develop a pan-European, intelligent analysis of a citizen's bioprofile; to make it remotely accessible to patients and clinicians; and to exploit it to combat ovarian, breast and brain cancers, leukaemia and melanoma.	It proposes to provide online novel computational intelligent techniques for the analysis of bioprofiles, including ANNs, evolutionary algorithms, SVMs, Bayesian methods, Adaptive Resonance Theory, Tree models, Fuzzy techniques, etc.
Health Agents	This project plans to create a multi-agent distributed DSS for the early diagnosis and prognosis of brain tumours. A distributed Data Warehouse with a network of interconnected databases of clinical, histological, and molecular phenotype data of brain tumour patients will be created.	It aims to develop new pattern recognition methods for a distributed classification and analysis of HR MAS and microarray data. They resort to feature selection, clustering and classification techniques, including support vector machines, mixture models and Bayesian methods.

Table 2: Several current European research projects on cancer and their use of ML techniques.