Automated clinical computational biology: an interpretable machine learning framework to predict disease severity and stratify patients from clinical data

Soumya Banerjee[1,2]
[1]University of Oxford, Oxford, United Kingdom
[2]Ronin Institute, Montclair, USA
E-mail: soumya.banerjee@maths.ox.ac.uk

Abstract

We outline an automated computational and machine learning framework that predicts disease severity and stratifies patients. We apply our framework to available clinical data. Our algorithm automatically generates insights and predicts disease severity with minimal operator intervention. The computational framework presented here can be used to stratify patients, predict disease severity and propose novel biomarkers for disease. Insights from machine learning algorithms coupled with clinical data may help guide therapy, personalize treatment and help clinicians understand the change in disease over time. Computational techniques like these can be used in translational medicine in close collaboration with clinicians and healthcare providers. Our models are also interpretable, allowing clinicians with minimal machine learning experience to engage in model building. This work is a step towards automated machine learning in the clinic.

1 Introduction

The advent of big data and clinical records databases opens possibilities for clinical data science. Machine learning techniques coupled with clinical data is thought to be critical in delivering the next generation of healthcare (Clifton et al., 2012).

Here we present an automated computational framework to derive insights from clinical data. The computational framework presented here can be used to stratify patients, predict disease severity and propose novel biomarkers for disease.

Our approach automatically performs model inference, cross-validation, model selection and generates insights with minimal operator intervention. Our models are also interpretable, allowing domain experts like clinicians (with minimal machine learning experience) to engage in model building.

Insights from machine learning algorithms coupled with clinical data may help guide therapy, personalize treatment and help clinicians understand the change in disease over time. Our approach is a step towards automated machine learning and computational biology in the clinic.

2 Methods

We have developed an automated machine learning framework that performs predictions with minimal operator intervention. First, we perform feature scaling to ensure that all input features are on the same scale. We then look at a suite of different machine learning techniques like neural networks, random forests, regularized generalized linear model (logistic regression) with LASSO (least absolute shrinkage and selection operator), support vector machines, linear regression and principal components analysis. Crucially, we perform inference, cross-validation, model selection and insight generation with minimal operator intervention.

3 Data

We used data from the UCI machine learning repository (Wisconsin breast cancer dataset) (available for download from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29) (Wolberg and Mangasarian, 1990; Zhang, 1992). The dataset consists of 699 patients divided into healthy and patients with breast cancer. The disease status is reported as benign or malignant. The different attributes measured were clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses.

All predictors are numeric (there are no categorical predictors) and were scaled to be within a range of 0 to 10. We replaced missing values with a 0. Future work will look at schemes to impute these values. Finally, we split the data into training, cross-validation and test sets.

4 Results

4.1 Stratifying patients

We used principal component analysis (PCA) to gain insights into the clinical data. The PCA analysis suggests that there are a few clusters that the data can be separated into (Figure 1 and 2). Single epithelial cell size and uniformity of cell shape seem to separate the data into distinct clusters (Figure 2). The attribute mitoses seem to account for many outliers (Figure 2).

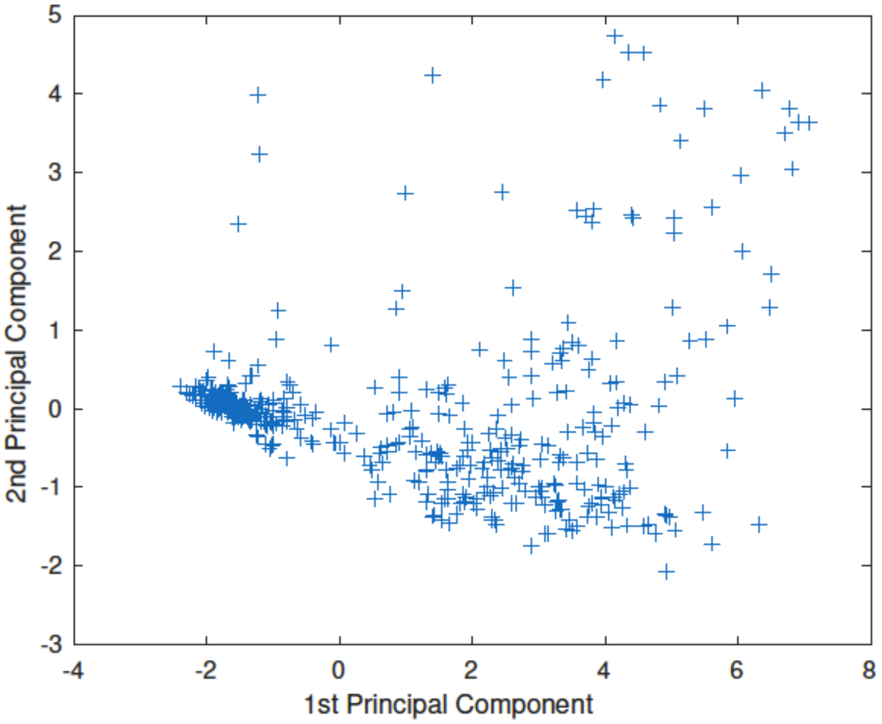

Figure 1. Principal components analysis of data. Analysis shows a few clusters for the first two principal components.

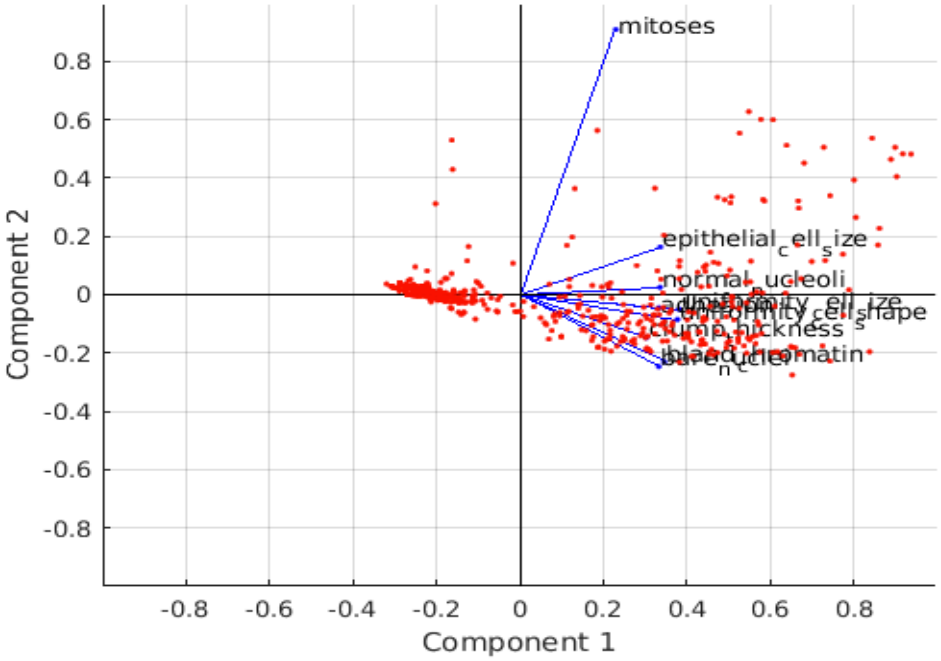

Figure 2. Principal components analysis of the data showing clusters for the first two principal components.

We note that the first principal component explains about 65% variance in the data (Figure 3).

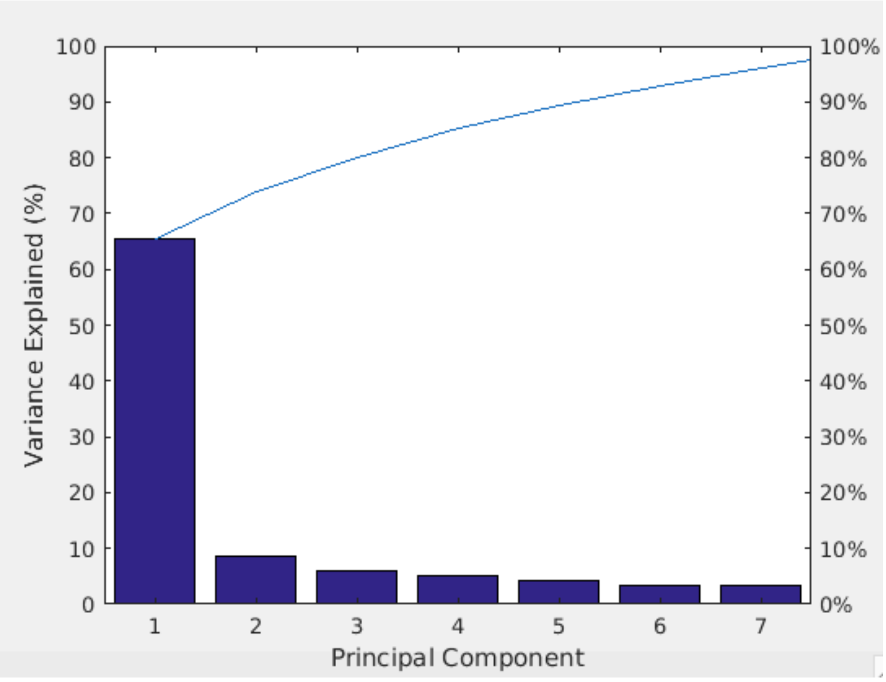

Figure 3. Percentage of variation explained by each principal component in a PCA.

Finally, the PCA analysis suggests the most extreme points in the data (outliers). Five patients with codes 1123061, 1198128, 1147748, 1165926 and 760001 are predicted to be outliers. For example, patients 1123061 and 76001 have a very low value (< 3) for the uniformity of cell shape. Patent 760001 has a very low value of mitoses (value of 1 on a scale of 1 to 10). All patients predicted to be outliers also have low values of the attribute bare nuclei. This kind of analysis can be used to stratify patients.

4.2 Predicting disease severity

We predict disease severity or probability of getting the disease using a suite of different machine learning algorithms.

We looked at artificial neural networks (Figure 4) which are composed of an input layer of features, hidden layers and an output layer that predicts disease severity (on a scale of 0 to 1). We varied the number of hidden layers from 1 to 100. A neural network with 10 hidden layers was found to give the best performance (mean squared error = 0.01) (Figure 5 and 6).

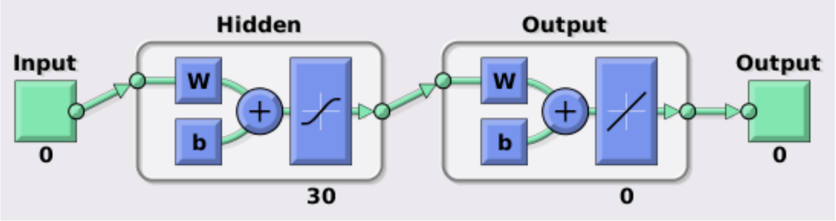

Figure 4. Architecture of neural network used to predict disease severity. The network shown has an input layer, 30 hidden layers and an output layer.

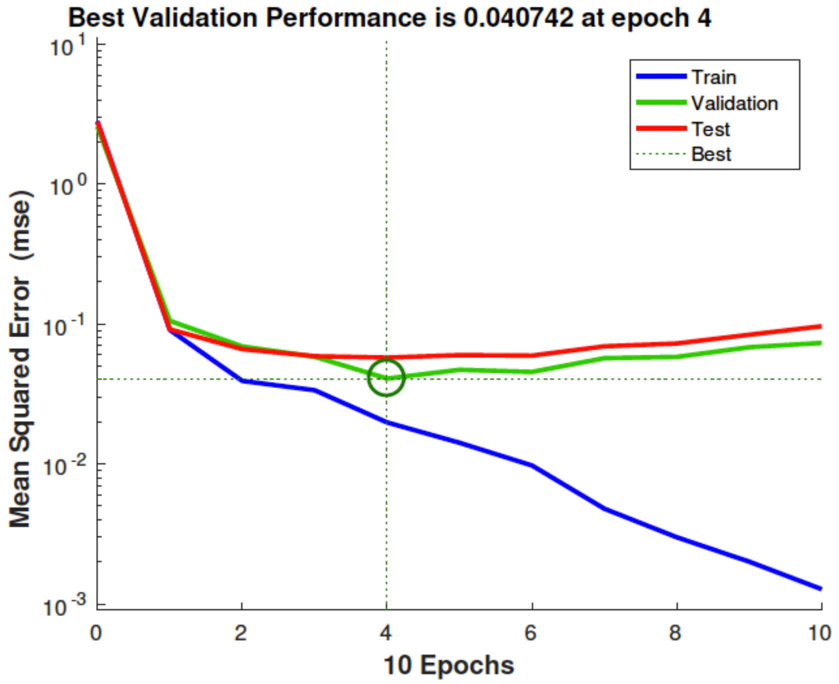

Figure 5. Neural network performance on training, validation and test dataset with 30 hidden layers.

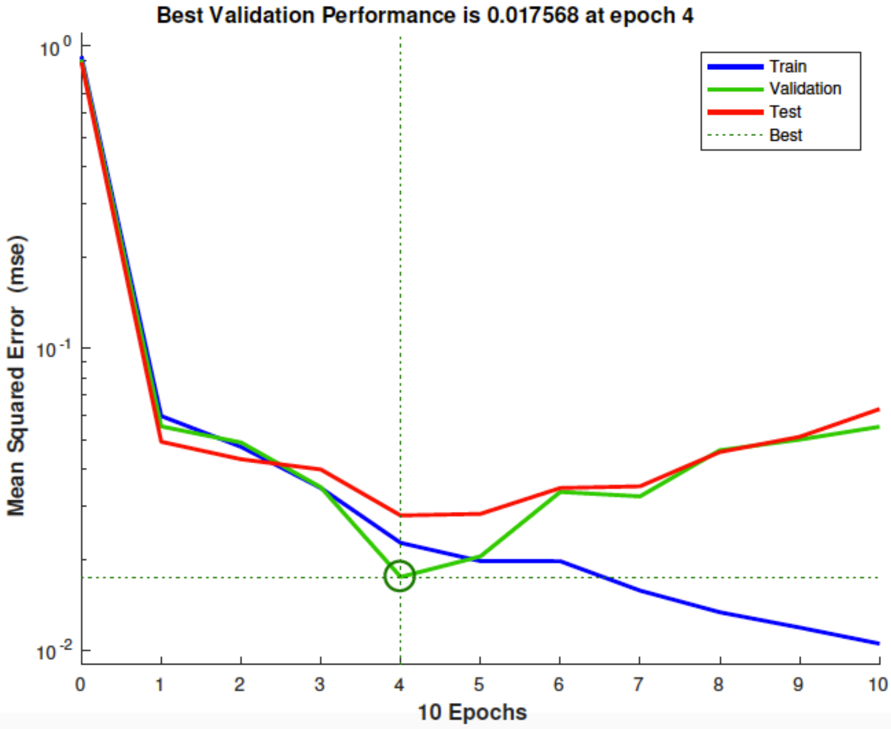

Figure 6. Neural network performance on training, validation and test dataset with 10 hidden layers.

We also used random forests which are collections of trees. Each tree can be interpreted as a set of rules that suggest how to combine the attributes to predict a disease severity. A forest is a collection or ensemble of such trees. We varied the leaf size from 5 to 100 and the number of trees that are grown from 1 to 50 (Figure 7). The best random forest model achieved a cross-validation mean squared error of 0.04.

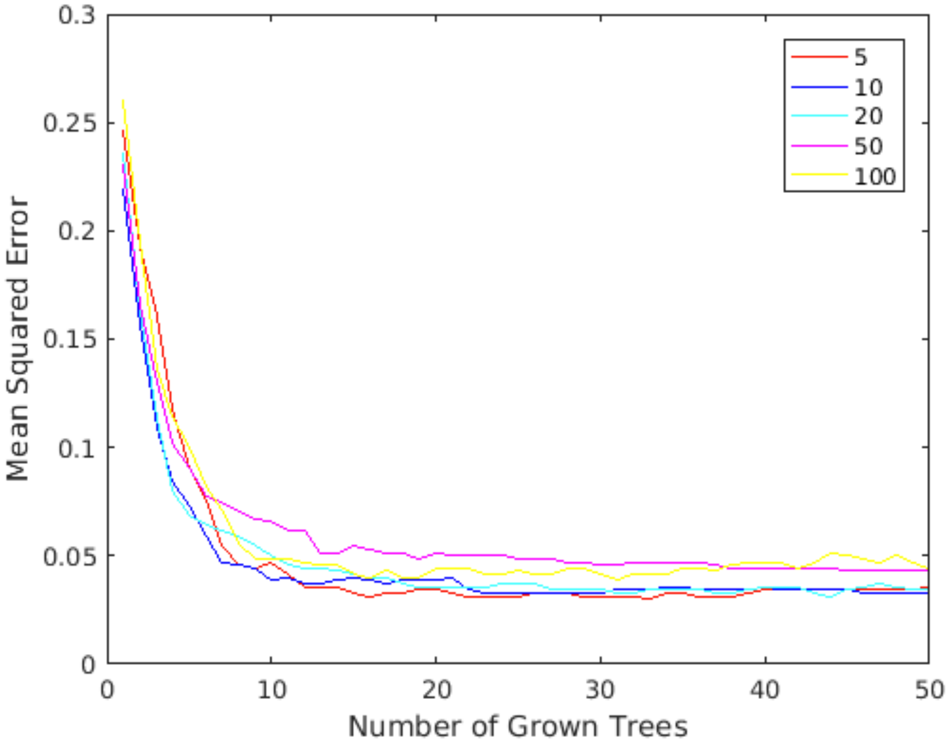

Figure 7. Performance of a random forest algorithm (out of bag prediction error). The leaf sizes are varied from 5 to 100 and up to 50 trees are grown.

Representative trees used for predicting disease severity are shown in Figure 8 (regression) and Figure 9 (classification). Random forests are very interpretable. For example, the tree shown in Figure 9 is very interpretable since it represents a rule of the form:

if single epithelial cell size >= 2.5 and uniformity of cell shape < 1.5
then healthy

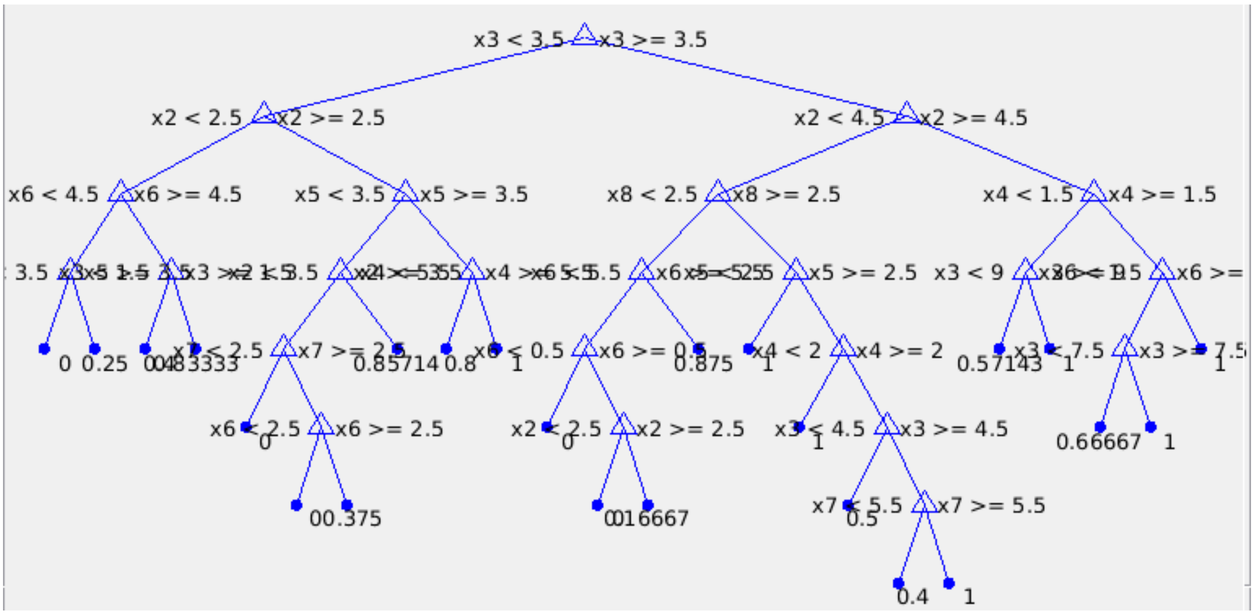

Figure 8. A representative tree from the random forest used in predicting disease severity (regression).

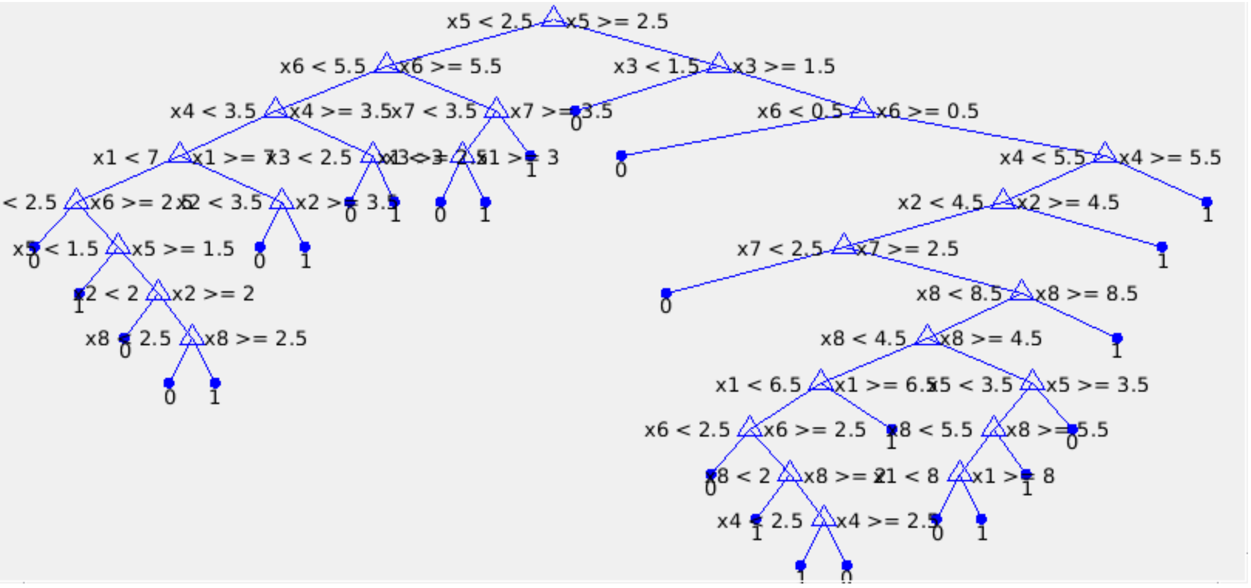

Figure 9. A representative tree from the random forest used in predicting disease severity (classification).

Insights from interpretable machine learning algorithms like random forests can inform decisions in the clinic. The top predictors in random forests are shown in Figure 10. Uniformity of cell size (2nd feature) and bare nuclei (6th feature) are important predictors. Mitoses (9th feature) is the least important predictor. We note however that mitoses separates two different clusters in the PCA plot (Figure 2) and may be useful as a biomarker.

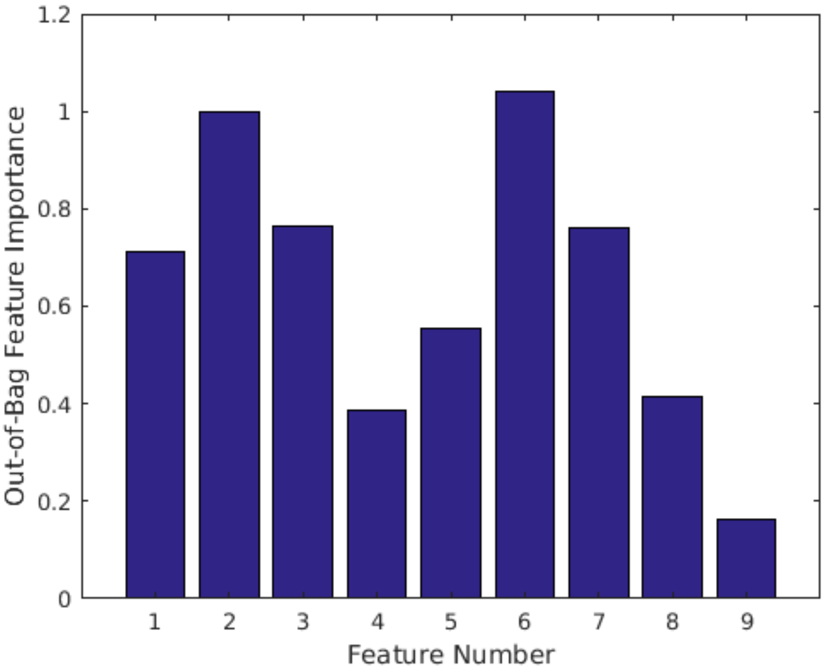

Figure 10. Top predictors in a random forest algorithm.

We note that even though artificial neural networks have the best performance (cross-validation mean squared error = 0.01 for neural networks; cross-validation mean squared error = 0.04 for random forests), the most interpretable models are random forests.

We also used a logistic regression model with LASSO (L1 regularization). We performed 10-fold cross validation to determine the regularization parameter (Figure 11). We found that all predictors are non-zero after cross-validation. Hence the logistic regression model suggests that all the predictors are important.

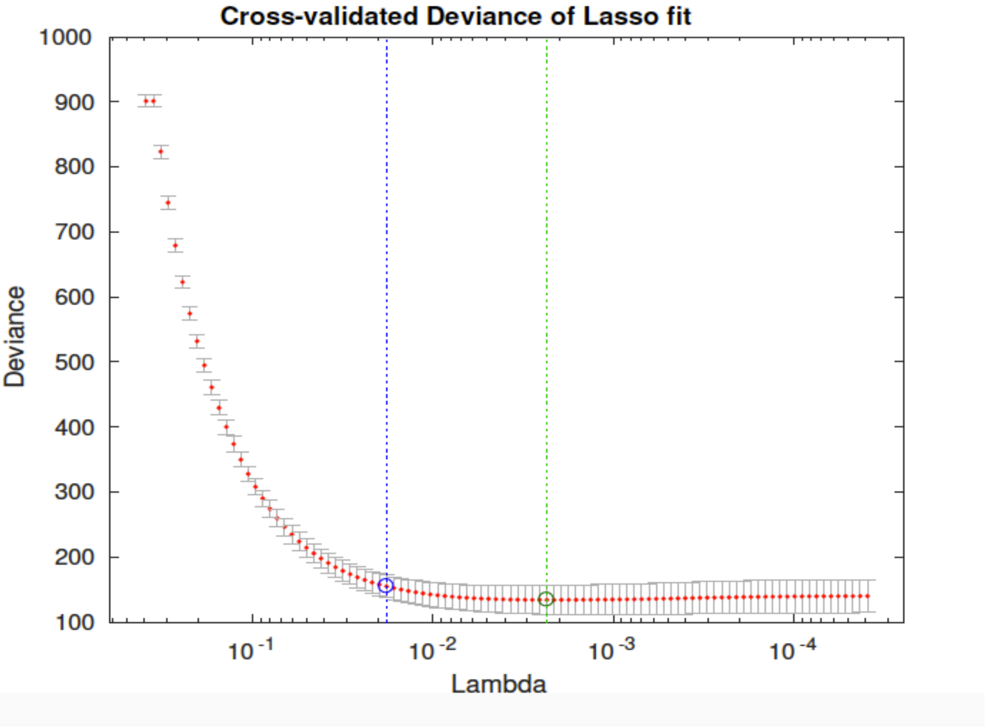

Figure 11. A plot of the effect of changing the regularization parameter (lambda) in a logistic regression model with LASSO (L1 regularization). The cross-validation error is used to find the optimal value of lambda.

Finally, we also looked at linear regression models for correlations of attributes with each other (within patients). We did not observe any meaningful relationships.

4.3 Biomarkers

The predictors uniformity of cell shape and single epithelial cell size separate the data into a few different clusters in the PCA plot (Figure 2). Mitoses separates the data into a third cluster in the PCA plot (Figure 2). Bare nuclei is an attribute that accounts for some outliers in the PCA analysis (see Section 4.1 Stratifying Patients).

Our random forest algorithm suggests that the top predictors are uniformity of cell size and bare nuclei (Figure 10). Taken together, we suggest that uniformity of cell size and bare nuclei maybe important biomarkers for disease.

4 Discussion

Big data technologies coupled with massive clinical records databases opens possibilities for data science in the clinic. Machine learning techniques coupled with clinical big data are thought to be critical in delivering the next generation of healthcare (Clifton et al., 2012).

Here we present an automated machine learning framework that generates insights from clinical data with minimal operator intervention. The computational framework presented here can be used to stratify patients, predict disease severity and propose novel biomarkers for disease. This can be used to guide therapy and intervention in the clinic.

We use a suite of machine learning algorithms to predict disease severity and stratify patients. We found that a PCA analysis combined with random forests can suggest biomarkers and ways to stratify patients. Our analysis suggests that uniformity of cell size and bare nuclei maybe important biomarkers for disease.

Even though artificial neural networks have better performance predicting disease severity than random forests, the most interpretable models are random forests. This is critical in communicating these insights to clinicians and healthcare professionals who may not be machine learning experts. We show a representative rule from a tree in a random forest (Figure 9) which takes the following form:

```
        if single epithelial cell size >= 2.5 and uniformity of cell shape < 1.5
            then healthy
```

Insights from interpretable machine learning algorithms like random forests can be very informative to clinicians. Our framework automatically performs model inference, cross-validation, model selection and generates insights into data with minimal operator intervention. Our models are also interpretable, allowing domain experts like clinicians (with minimal machine learning experience) to engage in model building. Coupling automated and interpretable machine learning techniques with clinical data may help guide therapy, personalize treatment and help clinicians understand the change in disease over time.

Our approach can be combined with multi-scale models (Banerjee and Moses, 2009; Banerjee, 2013; Banerjee, 2015; Banerjee and Moses, 2010a; Banerjee et al., 2016). Hybrid modelling approaches can be combined with machine learning techniques presented in the current work to gain mechanistic insights into disease, as has done previously for infectious diseases (Banerjee et al, 2015; Banerjee, 2015b, 2015c).

In summary, we present an automated and interpretable machine learning framework for generating insights. We demonstrate how this computational framework can be applied to clinical data. Computational techniques like these can be used in translational medicine in close collaboration with clinicians and healthcare providers. Our approach is a step towards automated machine learning and computational biology in the clinic.

Acknowledgment

The author wishes to thank Joyeeta Ghose, Dr. Beverly Kloeppel and Dr. Stephen Haben for fruitful discussions.

References

Soumya Banerjee and Melanie Moses. 2009. A Hybrid Agent Based and Differential Equation Model of Body Size Effects on Pathogen Replication and Immune System Response. In Timmis, J. (ed.) The 8th International Conference on Artificial Immune Systems (ICARIS), 14–18 (Springer, Lecture Notes in Computer Science, 2009). URL http://www.springerlink.com/content/b786g874642q2j37/

Clifton, D. A., Gibbons, J., Davies, J., & Tarassenko, L. (2012). Machine learning and software engineering in health informatics. In 2012 First International Workshop on Realizing AI Synergies in Software Engineering (RAISE) (pp. 37–41). IEEE Press. http://doi.org/10.1109/RAISE.2012.6227968

Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193--9196.

Zhang, J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470--479). Aberdeen, Scotland: Morgan Kaufmann.

Soumya Banerjee. 2013. Scaling in the immune system, PhD Thesis, University of New Mexico (2013)

Soumya Banerjee and Melanie Moses. 2010a. Scale Invariance of Immune System Response Rates and Times: Perspectives on Immune System Architecture and Implications for Artificial Immune Systems. Swarm Intelligence 4, 301–318 (2010). URL http://www.springerlink.com/content/w67714j724486331/

Soumya Banerjee, Drew Levin, Melanie Moses, Fred Koster, & Stephanie Forrest. 2011. The Value of Inflammatory Signals in Adaptive Immune Responses. In Artificial Immune Systems, 1–14 (Springer, 2011). URL http://www.springerlink.com/index/U634HJ83W62W5383.pdf

Soumya Banerjee and Melanie Moses. 2010b. Modular RADAR: An Immune System Inspired Search and Response Strategy for Distributed Systems. In E. Hart (ed.) E. Hart et al. (Eds.) Artificial Immune Systems, 9th International Conference, ICARIS 2010, Lecture Notes in Computer Science, 116–129 (Springer Verlag, Berlin, 2010). URL http://www.springerlink.com/content/9l062344680u6w76/

Melanie Moses & Soumya Banerjee. 2011. Biologically inspired design principles for scalable, robust, adaptive, decentralized search and automated response (RADAR). In Artificial Life (ALIFE), 2011 IEEE Symposium on, 30–37 (2011)

Soumya Banerjee. 2009. An Immune System Inspired Approach to Automated Program Verification, arXiv preprint arXiv:0905.2649, 2009

Soumya Banerjee and Joshua Hecker. 2015. A Multi-Agent System Approach to Load-Balancing and Resource Allocation for Distributed Computing, arXiv preprint arXiv:1509.06420, 2015

Soumya Banerjee and Melanie Moses. 2010c. Immune System Inspired Strategies for Distributed Systems. arXiv preprint arXiv:1008.2799, 2010

Soumya Banerjee, Pascal van Hentenryck and Manuel Cebrian. 2015. Competitive dynamics between criminals and law enforcement explains the super-linear scaling of crime in cities. Palgrave Communications, doi:10.1057/palcomms.2015.22, 2015

Soumya Banerjee. 2015b. Analysis of a Planetary Scale Scientific Collaboration Dataset Reveals Novel Patterns. arXiv preprint arXiv:1509.07313, 2015

Soumya Banerjee. 2015c. Optimal strategies for virus propagation. arXiv preprint arXiv: 1512.00844, 2015

Banerjee, S., Guedj, J., Ribeiro, R. M., Moses, M., & Perelson, A. S. 2016. Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. Journal of the Royal Society Interface, 13(117), 20160130-.http://doi.org/10.1098/rsif.2016.0130

Soumya Banerjee, A Roadmap for a Computational Theory of the Value of Information in Origin of Life Questions, Interdisciplinary Description of Complex Systems, 2016

Soumya Banerjee, A Biologically Inspired Model of Distributed Online Communication Supporting Efficient Search and Diffusion of Innovation, Interdisciplinary Description of Complex Systems, 2016

Shoenfeld, Y. (2004). The idiotypic network in autoimmunity: Antibodies that bind antibodies that bind antibodies. *Nature Medicine*, *10*(1), 17–18. http://doi.org/10.1038/nm0104-17

Soumya Banerjee. Optimal strategies for virus propagation. arXiv preprint arXiv:1512.00844, 2015

Soumya Banerjee, Jeremie Guedj, Ruy Ribeiro, Melanie Moses, Alan Perelson (2016). Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. Journal of the Royal Society Interface, 13(117), 20160130-. http://doi.org/10.1098/rsif.2016.0130