New Trends and Issues Proceedings on Humanities and Social Sciences

ISSN 2421-8030
Issue 3 (2017) 01-10
www.prosoc.eu

Selected paper of 5th Cyprus International Conference On Educational Research (Cyicer-2016) 31 March-02 April 2016, University Of Kyrenia, Kyrenia North Cyprus

Development of New Hybrid Admission Decision Prediction Models Using Support Vector Machines Combined with Feature Selection

Gozde Ozsert Yigit[a] *, Cukurova University, Computer Engineering Department, Adana, Turkey
Mehmet Fatih Akay[b], Cukurova University, Computer Engineering Department, Adana, Turkey
Hacer Alak[c], Cukurova University, Computer Engineering Department, Adana, Turkey

Suggested Citation:
Ozsert-Yigit, G., Akay, M.F. & Alak, H. (2017). Development of New Hybrid Admission Decision Prediction Models Using Support Vector Machines Combined with Feature Selection. *New Trends and Issues Proceedings on Humanities and Social Sciences.* [Online]. 03, pp 01-10. Available from: www.prosoc.eu

Selection and peer review under responsibility of Assist. Prof. Dr. Cigdem Hursen, Near East University

Abstract

The purpose of this paper is to develop new hybrid admission decision prediction models by using Support Vector Machines (SVM) combined with a feature selection algorithm to investigate the effect of the predictor variables on the admission decision of a candidate to the School of of Physical Education and Sports at Cukurova University. Experiments have been conducted on the dataset, which contains data of participants who applied to the School in 2006. The dataset has been randomly split into training and test sets using 10-fold cross validation as well as different percentage ratios. The performance of the prediction models for the datasets has been assessed using classification accuracy, specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV). The results show that a decrease in the number of predictor variables in the prediction models usually leads to a parallel decrease in classification accuracy.

Keywords: machine learning; prediction; physical ability test; feature selection

* ADDRESS FOR CORRESPONDENCE: **Gozde Ozsert Yigit,** Cukurova University, Computer Engineering Department, Adana, Turkey
E-mail address: ozsertg@gmail.com / Tel.: 03223387101

1. Introduction

In order to admit a candidate to the School of Physical Education and Sports at Cukurova University ("http://besyo.cu.edu.tr", 2006), the candidate has to be successful in the physical test applied at the School. There are two parts in the physical ability test. Each of these parts contains two tests. In the first part of the test, the vertical jump test as well as the coordination and skill test are applied. In the vertical jump test, the participant waits for resetting of timing mat with weight equally balanced on both feet. After the mat is arranged, the participant jumps vertically to reach the highest point he can and then the participant steps back on the mat. The score of the vertical jump test is calculated considering the time spent in the air. Figure 1 shows a typical example of the vertical jump test.

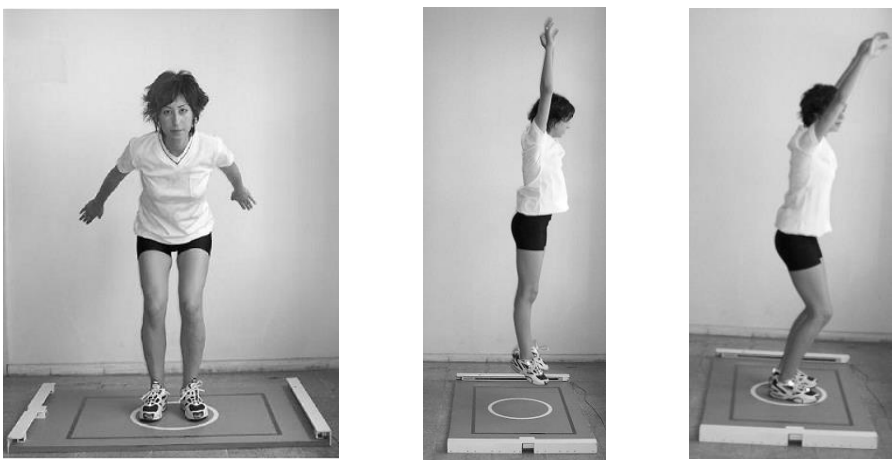

Figure 1. Vertical jump test

The coordination and skill test has several steps. This test starts with a front somersault. Then, the participant gathers a ball, throws the ball up in the air and gets hold of the ball after passing across the horizontal barrier. After this step, the candidate jumps over the balance tool and tries to keep stability while moving. Back somersault follows these steps. The last part of the coordination and skill test consists of sliding down and jumping over the obstacles. The setup of the coordination and skill test is shown in Figure 2.

In the second section, the participant undertakes the 30-meter dash and the 20-meter shuttle run tests. 30-meter dash test evaluates the participant's ability to quickly gain speed on 30 meters. This test contains running one maximum sprint over 30 meters. For the first step, one foot should be one step ahead. The front foot must be on the starting line. This position should be held for two seconds and movements are not permitted. Secondly, the shuttle run test is performed between two end points that are 20 meters apart. In the beginning, the speed of the participant is quite slow. As the test progresses, the speed of the participant is increased with every beep sound and the frequency of beeps is gradually increased. The test is completed when the participant gets two failures in consecutive.

A participant's admission decision is related with the participant's total scores from the physical ability test (PATS) together with his/her scores from National Student Selection Exam (NSSE) and National Student Placement Exam (NSPE), Grade Point Average (GPA) and specialization area at high

school. The overall score of a participant who graduated from a sports branch at high school is computed by Eq. (1)

$$SCORE = (PATS) + (0.52 \times GPA) + (0.36 \times NSPE)\,. \quad (1)$$

Similarly, the overall score of a participant who graduated from a non-sports branch at high school is computed by Eq. (2)

$$SCORE = (PATS) + (0.16 \times GPA) + (0.47 \times NSPE)\,. \quad (2)$$

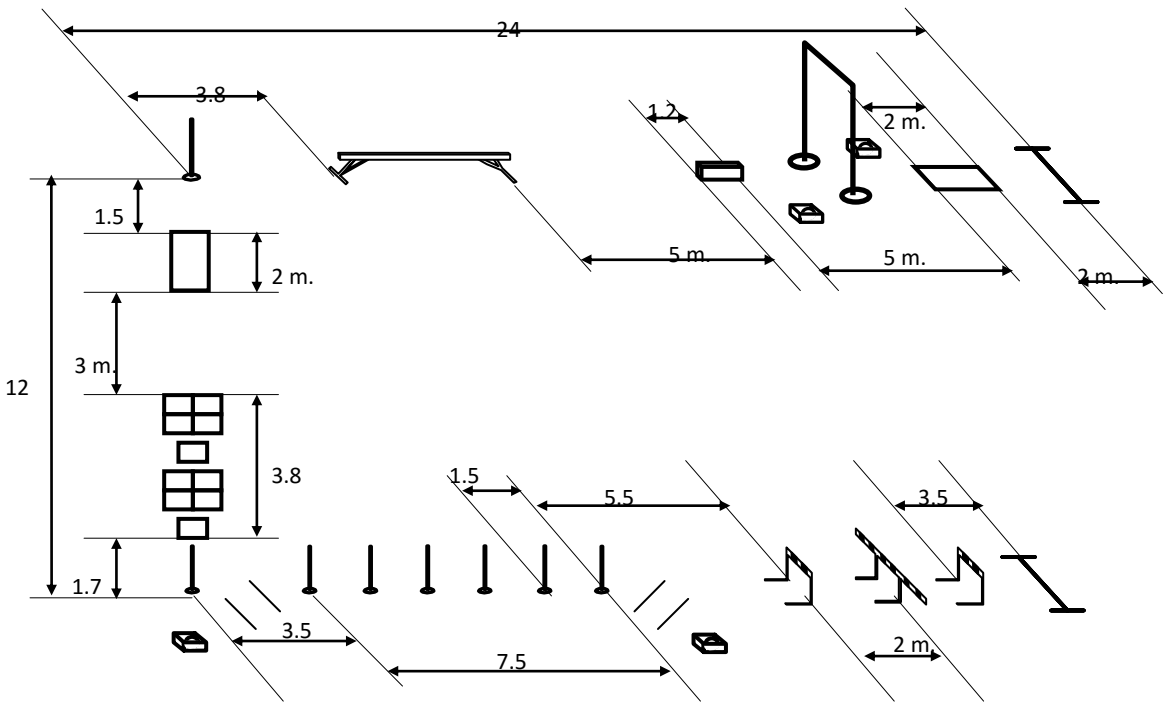

Figure 2. Coordination and skill test

After the overall scores are calculated, the scores are arranged in descending order and a pre-specified number of participants is admitted to the School ("http://besyo.cu.edu.tr/besyo/TR/OzelYetenekSinavi.aspx", 2006).

Developing admission decision prediction models has been an active reserach area for several years. İn this regard, there exist a few studies in literature (Abut, Akay, Turhan & Ozsert, 2015; Acikkar & Akay, 2008; Acikkar & Akay, 2009; Akay, Guler & Acikar, 2014) which have attempted to predict the admission decision of a candidate to the School of Physical Education and Sports of Cukurova University by using different machine learning methods. These methods included SVM, Multilayer Perceptron (MLP), Single Decision Tree, K-Means Clustering, Radial Basis Function Network and Logistic Regression. It has been shown that SVM models usually performed better than other prediction models.

Feature selection is helpful in locating the discriminative features that are the most appropriate to predict the class. Feature selection is used in data mining and statistics. The basic approach of feature selection is to choose a subset of input variables by removing non-relevant features. Therefore, in the preprocessing step, it may be advantageous to pick the relevant and necessary features (Sun, Lou & Bao, 2011).

There is only one study (Acikkar, Akay, Abut & Isoglu, 2014) in literature that uses machine learning methods combined with a feature selection algorithm to develop hybrid admission decision prediction models for the School of Physical Education and Sports at Cukurova University. In this study, MLP combined with a feature selection algorithm has been used to develop prediction models to predict the admission decision. As a feature selection algorithm, Relief-F has been selected. The results have shown that the model including all the predictor variables yielded the best classification accuracies, independent of which activation function has been used at the output layer.

Apparently, more research is required with the help of different machine learning methods combined with a feature selection algorithm. The purpose of this study is to investigate the effect of predictor variables on the admission decision by using SVM combined with a feature selection algorithm. More specifically, using the Relief-F feature selection algorithms, ranking of the predictor variables in the dataset has been calculated. Then, based on the ranking scores, several models have been developed by removing the predictor variable with the lowest score at a time. The performance of the prediction models has been assessed utilizing classification accuracy, specificity, sensitivity, PPV and NPV. The results show that a decrease in the number of predictor variables in the prediction models usually leads to a parallel decrease in classification accuracy.

The rest of the paper is structured as follows: Section 2 summarizes the dataset that is used in the current study. Section 3 gives details of the admission decision prediction models. Section 4 presents results and discussion. The last section, Section 5, concludes the paper.

2. Overview of the dataset

The dataset used in this study is provided from the School of Physical Education and Sports of Cukurova University. The dataset includes data of participants who applied to the school in 2006. The dataset contains nine predictor variables including gender, the scores from the NSSE and NSPE, GPA and the specialization area at high school, the scores from the coordination and skill test, vertical jump test, 30-meter dash test and 20-meter shuttle run test and two classes, which have the values 0 and 1, where 0 means “reject” and 1 is means “admit”.

There are 143 participants (87 males and 56 females) in the dataset. Table 1 shows descriptive statistics of the dataset

Table 1. Descriptive statistics of the dataset

Predictor Variable	Minimum	Maximum	Mean	Standard Deviation
Gender	0	1	0.60	0.49
NSSE	175.65	219.42	205.88	19.48
NSPE	197.81	236.55	226.20	14.90
GPA	79.30	76.13	75.98	10.89
Specialization area	0	76.13	0.14	0.34
Vertical jump test score	32.00	0	39.87	8.07
Coordination and skill test score	33.42	51.00	30.66	3.35
30-meter dash test score	4.83	27.04	4.03	0.38
20-meter shuttle run test score	54.00	3.81	97.44	23.52

3. Prediction models

By using the Relief-F feature selection algorithm (Witten, Frank, & Hall, 2011), ranking of the predictor variables has been calculated. Then, based on these ranking scores, several models have been developed by removing the predictor variable with the lowest score at a time. Table 2 shows the ranking of predictor variables based on their Relief-F scores and Table 3 shows the predictor variables in each admission decision prediction model.

The SVM is a state-of-the-art classifier which is widely utilized in many application areas due to its high accuracy (Jaganathan, Rajkumar, & Kuppuchamy, 2012; Xu, Lemischka & Ma'ayan, 2010; Young, Ridgway, Leung, Barnes & Ourselin, 2011). The type of kernel function, kernel function parameters and the value of cost (C) are the major components that affect the quality and performance of a SVM model. İn this study, grid search technique has been used to find the optimal SVM parameter values and radial basis function (RBF) has been selected as the kernel. Grid search technique looks for values of every parameter using geometric steps across the search range. The values of C and gamma (γ) determine the performance of SVM based models. γ is an important parameter for the RBF kernel. To guarantee that the prediction models developed by using SVM is valid and can be generalized for making new predictions regarding new data, the dataset is partitioned into training and independent test sets via a 10-fold cross validation. For each value of (C, γ), 5-fold cross validation is conducted on the new training subset and root mean squared errors of each prediction are calculated. The (C, γ) pair that yields the lowest overall root mean squared error is chosen for training the train subset. The prediction model is developed after training subset is trained with the optimized values of C and γ. Lastly, the prediction model is used to estimate admission decision value in the test subset.

The values of the SVM parameters that have been used to develop the prediction models are given in Table 4.

Table 2. Relief-F scores of the predictor variables

Variables	Score
30-meter dash test score	0.01755
NSPE	0.01212
Vertical jump test score	0.00686
NSSE	0.00523
Coordination and skill test score	0.00325
Gender	0.0
20-meter shuttle run test score	-0.00282
Specialization area	-0.01099
GPA	-0.01509

Table 3. **Overview of prediction of admission decision models along with the predictor variables**

Models	Prediction Variables
Model 1	Gender, NSSE, NSPE, GPA, SA, VJTS,CSTS, DTS, SRTS
Model 2	Gender, NSSE, NSPE, SA, VJTS, CSTS, DTS, SRTS
Model 3	Gender, NSSE, NSPE, VJTS, CSTS, DTS, SRTS
Model 4	Gender, NSSE, NSPE, VJTS, CSTS, DTS
Model 5	NSSE, NSPE, VJTS, CSTS, DTS
Model 6	NSSE, NSPE, VJTS, DTS
Model 7	NSSE, NSPE, DTS
Model 8	NSPE, DTS
Model 9	DTS

SA: Specialization area, **VJTS**: Vertical jump test score, C**STS**: Coordination and skill test score
DTS: 30-m dash test score, **SRTS**: 20-meter Shuttle run test score

Table 4. Values of the utilized parameters for SVM

Parameter	Value
Cost (*C*)	[1, 100]
Gamma (γ)	[0.001, 50]
Kernel Function	RBF

4. Results and Discussion

The performance of the prediction models has been assessed using classification accuracy, specificity, sensitivity, PPV and NPV, the formulas of which are given in Eq. (3) through Eq. (7)

$$Accuracy\ (\%) = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \tag{3}$$

$$Specificity\,(\%) = \frac{TN}{FP + TN} \times 100 \tag{4}$$

$$Sensitivity\,(\%) = \frac{TP}{TP + FN} \times 100 \tag{5}$$

$$PPV\,(\%) = \frac{TP}{TP + FP} \times 100 \tag{6}$$

$$NPV\,(\%) = \frac{TN}{FN + TN} \times 100 \tag{7}$$

TP is true positive (i.e. number of applicants admitted and correctly classified as "admitted" by the classifier), TN is true negative (i.e. number of applicants rejected and correctly classified as "rejected" by the classifier), FP is false positive (i.e. number of applicants rejected but wrongly classified as "admitted" by the classifier) and FN is false negative (i.e. number of applicants admitted but wrongly classified as "rejected" by the classifier).

Table 5 through Table 9 show classification accuracies, specificities, sensitivities, PPV's and NPV's of the prediction models after applying various validation techniques. Figure 3 shows the average classification accuracies of prediction models.

Regarding the results obtained, the following discussions can be made:

- The model including all predictor variables has the highest classification accuracy regardless of which validation technique has been utilized. In other words, the classification accuracy diminishes as the number of predictor variables decreases.

- The classification accuracies obtained by Model 1, Model 2 and Model 3 are close to each other suggesting that the predictor variables GPA and specialization area have a negligible effect on the prediction of admission decision.

- Observation of the classification accuracies of Model 4 and Model 5 suggests that gender plays a significant role for prediction of admission decision. Similarly, the observed decrease in classification accuracy from Model 8 to Model 9 reveals that NSPE markedly increases the classification accuracy.

- The classification accuracies obtained by Model 6, Model 7 and Model 8 are approximately the same. This, in turn, indicates that the predictor variables NSSE and vertical jump test score have negligible effect on the prediction of admission decision

- A decrease in the number of predictor variables in the prediction models usually leads to a parallel decrease in specificity, sensitivity, PPV and NPV.

Table 5. Classification accuracies of prediction models

Models	Data Split				
	CV	80-20%	70-30%	60-40%	50-50%
Model 1	97.20	97.67	97.67	94.74	97.22
Model 2	95.80	93.02	93.02	91.23	94.44
Model 3	96.50	95.35	95.35	92.98	94.44
Model 4	92.31	93.02	93.02	87.72	87.50
Model 5	84.62	83.72	83.72	80.70	77.78
Model 6	82.52	83.72	83.72	80.70	77.78
Model 7	80.42	79.07	79.07	73.68	77.78
Model 8	80.42	79.07	79.07	73.68	75.00
Model 9	72.03	72.41	74.42	71.93	72.22

Table 6. Specificities of prediction models

Models	Data Split				
	CV	80-20%	70-30%	60-40%	50-50%
Model 1	98.06	100.00	100.00	97.56	98.08
Model 2	97.09	100.00	100.00	92.68	96.15
Model 3	97.09	95.24	100.00	95.12	94.23
Model 4	95.15	85.95	100.00	92.50	90.38
Model 5	97.09	85.71	100.00	81.25	88.46
Model 6	94.17	80.95	90.32	84.09	94.23
Model 7	97.09	80.95	93.55	79.55	94.23
Model 8	96.12	85.71	96.77	76.47	92.31
Model 9	95.15	90.48	100.00	71.93	100.00

Table 7. Sensitivities of prediction models

Models	Data Split				
	CV	80-20%	70-30%	60-40%	50-50%
Model 1	95.00	87.50	91.67	81.25	80.00
Model 2	92.50	87.50	75.00	87.50	90.00
Model 3	95.00	87.50	83.33	87.50	90.00
Model 4	85.00	62.50	75.00	76.47	65.00
Model 5	47.50	62.50	41.67	77.78	45.00
Model 6	52.50	50.00	66.67	69.23	35.00
Model 7	37.50	25.00	41.67	53.85	35.00
Model 8	35.00	37.50	33.33	66.67	30.00
Model 9	7.50	0.00	0.00	0.00	0.00

Table 8. PPV's of prediction models

Models	Data Split				
	CV	80-20%	70-30%	60-40%	50-50%
Model 1	95.00	100.00	100.00	94.74	94.12
Model 2	92.50	100.00	100.00	82.35	90.00
Model 3	92.68	97.50	100.00	87.50	85.71
Model 4	87.18	55.56	100.00	81.25	72.22
Model 5	86.36	62.50	100.00	43.75	60.00
Model 6	77.78	50.00	72.73	56.25	70.00
Model 7	83.33	33.33	71.43	43.75	70.00
Model 8	77.78	50.00	80.00	25.00	60.00
Model 9	37.50	0.00	0.00	0.00	0.00

Table 9. NPV's of prediction models

Models	Data Split				
	CV	80-20%	70-30%	60-40%	50-50%
Model 1	98.06	95.45	96.88	93.02	92.73
Model 2	97.09	95.45	91.18	95.00	96.15
Model 3	98.04	85.24	93.94	95.12	96.08
Model 4	94.23	85.00	91.18	90.24	87.04
Model 5	82.64	85.71	81.58	95.12	80.70
Model 6	83.62	80.95	87.5	90.24	79.03
Model 7	80.00	73.91	80.56	85.37	79.03
Model 8	79.20	78.26	78.95	95.12	77.42
Model 9	72.54	70.37	72.09	100.00	72.22

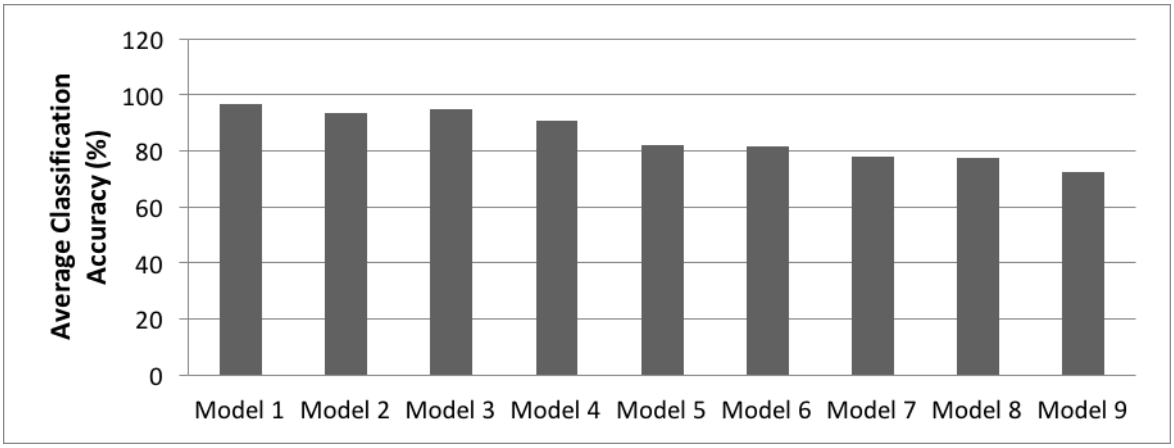

Figure 3. Average classification accuracies of prediction models

5. Conclusion

In this study, SVM combined with the Relief-F feature selection algorithm has been used to analyze the effects of the predictor variables on admission decision of a candidate to the School of Physical Education and Sports at Cukurova University. 10-fold cross validation and random percent splits of training/testing data have been used to ensure that the presented results are valid and can be used in making predictions within acceptable limits of accuracy regarding new data. The results show that gender and NSPE have remarkable effect on prediction of admission decision. On the other hand, the least effective predictor variables are GPA, specialization area, NSSE and vertical jump test score.

Acknowledgment

The authors would like to thank Cukurova University Scientific Research Projects Center for supporting this work under grant no. FYL-2015-3845.

References

Abut, F., Akay, M. F., Turhan, I. & Ozsert, G. (2015). Performance evaluation of different classifiers for predicting the admission decision of a candidate to the school of physical education and sports at Cukurova University. In *Proceedings of the 1st International Symposium on Sport Science, Engineering and Technology* (*ISSSET2015), Istanbul, 10-13 May 2015* , 178-184.

Acikkar, M. & Akay, M. F. (2009). Support vector machines for predicting the admission decision of a candidate to the School of Physical Education and Sports at Cukurova University. Expert Systems with Applications, *36* (3), 7228-7233.

Acikkar, M., Akay, M. F. Abut, F. & Isoglu, O. (2014). Predicting the admission decision of a candidate to the school of physical education and sports at Cukurova University by using Multilayer Perceptron combined with feature selection. In *Proceedings of the Second International Symposium on Engineering, Artificial Intelligence & Applications (ISEAIA2014), North Cyprus, 5-7 Nov 2014,* 15-16.

Akay, M. F., Guler, M. & Acikkar, M. (2014). Multilayer perceptron models for predicting the admission decision of a candidate to the school of physical education and sports at Cukurova University. In *Proceedings of the Second International Symposium on Engineering, Artificial Intelligence & Applications (ISEAIA2014), North Cyprus, 5-7 Nov 2014,* 10.

Jaganathan, P., Rajkumar, N. & Kuppuchamy, R. (2012). A comparative study of improved F-Score with Support Vector Machine and RBF network for breast cancer classification. *International Journal of Machine Learning and Computing*, *2* (6), 741-745.

Retrieved from http://besyo.cu.edu.tr

Retrieved from http://besyo.cu.edu.tr/besyo/TR/OzelYetenekSinavi.aspx

Sun, Y., Lou, X. & Bao, B. (2011). A novel Relief feature selection algorithm based on Mean-Variance model. *Journal of İnformation & Computational Science*, *8* (16), 3921-3929.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xu, H., Lemischka, I. R. & Ma'ayan, A. (2010). SVM classifier to predict genes important for self-renewal and pluripotency of mouse embryonic stem cells. BMC systems biology, *4* (1), 1.

Young, J., Ridgway, G., Leung, K. K., Barnes, J. & Ourselin, S. (2011). Prediction of MCI to Alzheimer's conversion with hippocampal shape features and support vector machine. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *7* (4), 41.