Logistic Regression on the BIOSCAD data set

Introduction

Overview:
The aim of the project was to create a model that, based on biomarkers measured on the skin, would diagnose atopic dermatitis (AD) in a patient. We used the data from the AMC BIOSCAD study.

Method

Data set:
The AMC BIOSCAD study data is first pre-processed to make it suitable for training – see pre-processing report (../../Preprocessing/preprocessing-report.pdf) for details. One additional step was performed to the pre-processed data – the SCORAD values were replaced with a Boolean to indicate if the patient suffered from AD. After pre-processing, there were 100 observations (53 AD and 47 control) available for training, 35 features and one dependent variable.

Model building:
Building each model consisted of a training, validation, and testing phase. To train the model, 60% of the data was used. Using the validation data (20%), the best prediction threshold was found. This is the value above which a patient is predicted to have AD, and below which not. Once the best prediction threshold has been found, the data is tested using the remaining 20% to allow the performance of the model to be evaluated. This 3 step process is repeated 100 times using different data points for the training, testing, and validation data.

Performance evaluation:
Several performance metrics were calculated to test the models. Since the two classes are balanced in the dataset, our main metric is the accuracy – the percentage of predictions that were correct. We also calculated the sensitivity (true positive rate, TPR), and the specificity (true negative rate, TNR). We also computed the F1-score, the Matthews correlation coefficient and the diagnostic odds ratio.

Input variable selection:
The first model built used all 35 attributes as inputs. To see the relative importance of these attributes, we also built models using each attribute individually. Using this information, we combined the top *n* performing variables (based on accuracy) together to create input data for new models.

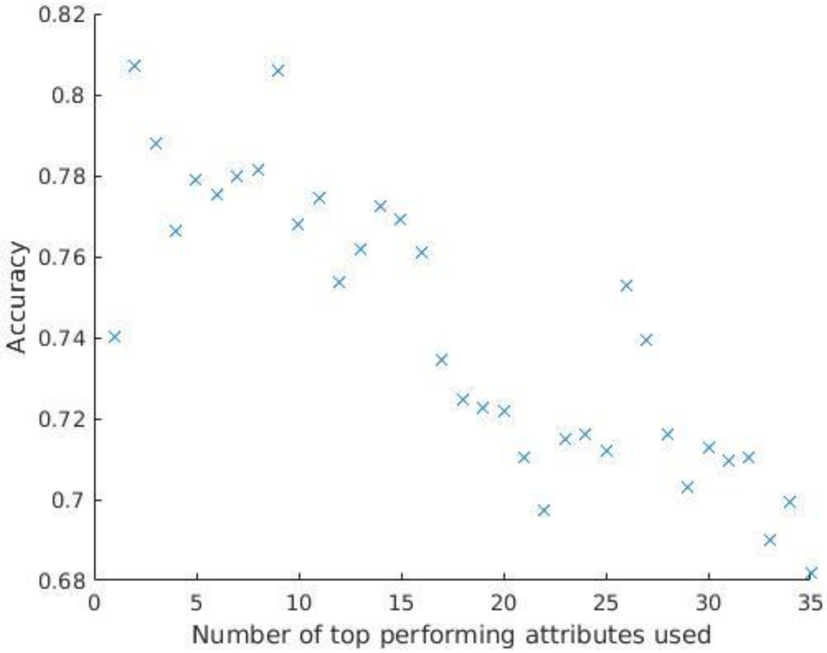

Figure 1

As increasing the number of top performing attributes used decreases the performance of the model (figure 1), we chose to limit future models to 4 input attributes. We then built every possible model using up to 4 attributes (approximately 60,000 models).

We also performed logistic regression using only IL-1a and IL-1β as these were the only two attributes in which most of data was above the detection limit.

Results

Top performing models:

Attribute 1	Attribute 2	Attribute 3	Attribute 4	Accuracy	Threshold value
FLGCarrier	IL-7	IL-2	-	0.8365	0.4109
FLGCarrier	IP-10	IL-7	IL-2	0.8350	0.4314
FLGCarrier	MCP-1	IL-7	IL-2	0.8345	0.4051
FLGCarrier	IL-7	IL-2	IL-13	0.8340	0.4340
FLGCarrier	MCP-4	IL-7	IL-2	0.8335	0.3768
FLGCarrier	IL-7	IL-16	IL-2	0.8330	0.4583
FLGCarrier	IL-7	IL-2	IL-17A	0.8315	0.4332
FLGCarrier	MIP-1A	IL-7	IL-2	0.8310	0.3664
AgeAtVisit	FLGCarrier	IL-7	IL-2	0.8305	0.4072
FLGCarrier	IL-7	IL-2	IL-6	0.8305	0.4172
FLGCarrier	IL-7	IL-15	IL-2	0.8270	0.3980
FLGCarrier	IL-16	IL-2	IL-10	0.8270	0.3811
skinType6	IL-5	IL-16	IL-2	0.8270	0.4102
IL-5	IL-7	IL-16	IL-2	0.8255	0.4214
FLGCarrier	IL-7	IL-18	IL-2	0.8250	0.4014

Table 1

Most accurate model:

Model trained using FLGCarrier, IL-7, and IL-2.

Performance metric	Value
Accuracy	0.837
Sensitivity	0.891
Specificity	0.779
Precision	0.817
F-Score	0.852
Mathews Correlation Coefficient	0.677
Diagnostic odds ratio	28.940

Table 2

Attribute	Odd ratio	P-value
IL-2	0.287	0.001
IL-7	0.328	0.008
FLGCarrier	39.425	0.056

Table 3

The coefficients of the most accurate model show that FLG-Carrier had the most important contribution (table 3) for the prediction. This is because in the original data set, there are no patients with FLG mutations in the control group. As such, the model indicates that any patient with an FLG mutation must have AD. However, we know from other analysis of other data sets that this is not the case. As a result, this variable is overfitting the input data and will not generalise well to external data.

To avoid the problem of overfitting, we looked at models that did not use FLGCarrier. This initially led us to a model using skin type 6, IL-2, IL-5, and IL-16 however the use of skin type 6, for which there are only a few values in the data set, means that there is presumed overfitting to match those data points. As such, we must also remove this attribute.

IL-2, IL-5, IL-7, and IL-16 model:
The best model to not use FLGCarrier or skin type 6 used IL-2, IL-5, IL-7, and IL-16 as input attributes.

Performance metric	Value
Accuracy	0.826
Sensitivity	0.898
Specificity	0.733
Precision	0.785
F-Score	0.838
Mathews Correlation Coefficient	0.643
Diagnostic odds ratio	24.222

Table 4

Attribute	Odd ratio	P-value
IL-2	0.128	0.000
IL-5	2.478	0.005
IL-7	0.372	0.004
IL-16	2.825	0.008

Table 5

The use of IL-2 and IL-7 in this model is unexpected as previous analysis (cytokine analysis, currently unpublished) has shown that these attributes have little impact on SCORAD. When we look at the original data set, we find that the majority of data points for these attributes are marked as below the detection range (as they are for most other attributes too) and are thus unreliable. By the garbage in garbage out principle, we must conclude that the model will not generalise well to data outside of the data set (overfitting).

To overcome this issue, we used only attributes that had most of the values above the detection limit. These were, IL-1a and IL-1β.

IL-1a and IL-1β model evaluation:

Model trained using IL-1a and IL-1β.

Performance metric	Value
Accuracy	0.541
Sensitivity	0.575
Specificity	0.505
Precision	0.561
F-Score	0.568
Mathews Correlation Coefficient	0.080
Diagnostic odds ratio	1.378

Table 6

Discussion

Final model selection and performance:

After elimination of all models that used FLGCarrier or skin type 6 due to overfitting, and elimination of all models trained on large amounts of data marked as below the detection limit for similar reasons, we were forced to consider models trained only on combinations of IL-1a and IL-1β. Of the 3 possible options, the best performer was the one that used both as input attributes.

This model showed accuracy levels of 54% (table 6). As this is only 1% higher than the prevalence, the model is no better than a random predictor.

Conclusion:

Despite the high accuracies of the top performing models, they cannot be considered good predictors as they all exhibit signs of overfitting and would not generalise well. The models which do not show signs of overfitting (those trained on data above the detection limit) have poor performance. As such, there are no models, trained with this dataset, which can be used to reliably diagnose a patient in the real world.