Final Project IEOR 265

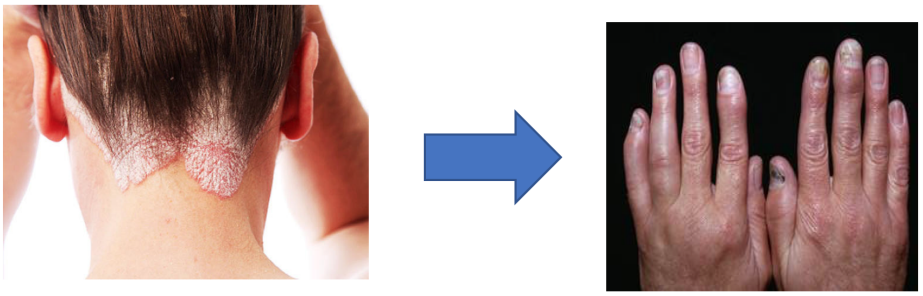
Pierre-Habté Nouvellon, instructor: Pr. Aswani

April 24, 2018

1 Introduction

1.1 Description of the problem

Psoriasis is an immune-mediated inflammatory skin disorder that affects 2-3% of the population. Up to 30% of patients with psoriasis also develop psoriatic arthritis, a debilitating condition affecting the joints that significantly lowers patients' quality of life. As psoriatic arthritis is characterized by a diverse set of clinical features, referral to a rheumatologist, diagnosis, and treatment is often delayed. For example, irreparable joint damage may result in just six month. Thus, it is crucial to predict which psoriasis patients will develop psoriatic arthritis before onset of symptoms.



Psoriasis Psoriatic arthritis

1.2 Goal of the project:

The goal of this project is to predict Psoriatic arthritis (PsA)based on genomic information and historical medical records of about 80,000 patients. For that we will implement some supervised models on the cleaned data set. There are several problems we need to consider beforehand. First is to perform data exploratory analysis and feature engineering. We need to take into account of the effects of different feature combination techniques and also consider the approach we could take according to different tests performed. For example, some tests involved UV treatment for patients suffering from severe skin disorder, who will have a higher probability to develop Psoriatic Arthritis. Whereas other tests results such as the glucose test are applicable to most patients. Another task is to select useful features among the existing ones that will lead to psoriatic arthritis. This will require thorough consideration on all the variables we currently have and to take into consideration different ways we could combine all the datasets. Once the data is ready, we will have to train different supervised model such as logistic regression, SVM and neural network. The goal would be to have the least number of false positive, with less than 20% of false positive. A good prediction model would be a model with an AUC bigger than 0.75.

In the first part of the report, we will discuss the steps of the preprocessing of the data without going to much into the technical details. In the second part we will focus on the use of deep learning. We will study the different methods and tunning used to come up with the best PsA and PsO (psoriasis) prediction model.

2 Preprocessing the data

2.1 Understanding the phenotype data

2.1.1 Overview of the phenotypic data

The phenotype data consists of the observable characteristics of the patients (age, BMI, sex, race, whether the patient smokes, etc...). We received 73 phenotypic features from UCSF for each of the 86,000 subjects.

However, we couldn't use them directly for our Machine Learning models.

2.1.2 Treating the missing values

One of the first issues we faced while exploring the data is the missing values. When we looked at our data, we noticed that the amount of missing values depends on the features: some of the features have 97% of missing values, and others have no missing values at all. We had to choose a threshold: we decided to keep only features where there are less than 20% of missing values. All the other features were deleted. This first choice made us go from 73 initial features to only 56 features. Obviously this shrinking of features' number partially solves the problem the overfitting of the models since they are less likely to fit the noise of the training set. However, among the remaining 56 features, we still had columns with missing values. We had to replace those cells. But how? A first idea would be replacing the missing values simply by the mean value of each column. But this was not a good idea; in fact, it assumes that the data has a symmetric distribution. To check this assumption, we looked at the distribution of each column, and we noticed that the majority of the data is clearly not symmetric and the skewness (it is the asymmetry in a statical distribution, in which the curve appears distorted or skewed either to the left or to the right) was very high. As a consequence, the idea of the average was not suitable to apply in our project. In this situation, according to Lee, H.,Rancourt, E., & Särndal, C. E. (2002), one solution would be to randomly sample from the existing values and replace the missing values. This method would make sure to keep the initial data distribution after replacing the missing values.

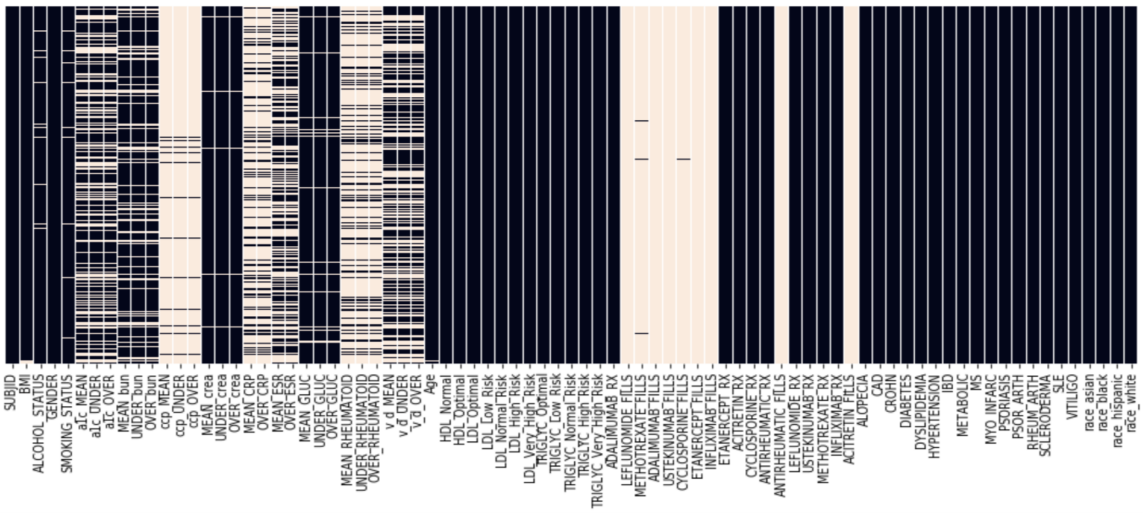


Figure 1: Table of missing values for 73 phenotype features(white = missing, black = not missing)

2.1.3 Treating the outliers

The second part of the data exploration was to handle the outliers (observations that fall out of the normal range of total data points). As for the definition of outliers, we used the Goldilocks rule of interquartile. This rule says that any data points that are smaller than or larger than 1.5*the interquartile range are outliers. The interquartile is the difference between first and third quartile. Another issue that we faced is that not all the outliers are replaceable: let's assume that 60 is an outliers in the column "Age". How would we replace 60? There is no way to that since 60 is the age of the patient, and we need to keep it as it is. As a consequence, we handled at the features case by case to see each time if it makes sense to replace the outliers.

2.1.4 Features selection

As we started with 73 features and kept only 56 after the deletion of certain columns while treating the missing values, we continue in this part decreasing the number of features by selecting the most important in predicting the output. To do so, we combined the results of two methods. The first method is univariate: for each feature, it looks at how it's dependent from the predicted variable. We computed the Pearson coefficient (a measure of the relationship between variables and ranges from -1 to 1, which means from negatively correlated to no correlation, and to positively correlated)for each feature and ranked them in a descending order. We chose a threshold of 0.1 and we kept only the features that have a Pearson coefficient higher than this value. The choice of this coefficient can justified by the fact it detects the linear relationship between vectors. The second method is multivariate: it is called regularization. This method penalizes some coefficients of the regression model to make it less likely to fit the noise, and as a consequence, decreases the variance and make the model more generalizable. The first regularizer we

used is L1 (LASSO). This regularizer add a penalization term to the cost function to be minimized and plays the role of feature selection and deleting others since, it set some coefficient to zero. One advantage of this method is model sparsity since some features completely disappear which make the model simpler and easier to interpret. We combined the results from those two methods and selected 15 features that we used in our final predicting models.

BMI	ALCOHOL_STATUS	GENDER	SMOKING_STATUS	MEAN_crea	UNDER_crea	OVER_crea	MEAN_GLUC	...	RHEUM_ARTH	SCLERODERMA
29.0	1.0	1	2.0	3.529375	0.0	0.79375	91.428571	...	0	0
27.0	1.0	1	1.0	1.092308	0.0	0.00000	92.692308	...	0	0
31.0	1.0	2	2.0	0.942857	0.0	0.00000	103.333333	...	0	0

Figure 2: Example of phenotype features after preprocessing

2.2 Understanding the Genetic data

2.2.1 Overview of the genetic data

UCSF gave us access to the genetic data of the 86,000 subjects. For every person, the genetic data is made of the values of roughly 1.5 million Single Nuclear Polymorphisms (SNPs, pronounced "snips"). A SNP is a single nucleotide that can change from human to human. In effect, we have about a tenth of the variations in each subject's genome. The goal of this section is to explain how we selected the SNPs that are most correlated to the occurrence of PsO and PsA. USCF also provided the position of each SNP in the genome (particularly in which chromosome they are located) as well as the possible variances that can be found in the population. We were also aware of the ancestry of each subject (european, african, latino or asian). This is extremely important since some SNPs only appear in a precise population (for example, many SNPs only appear in the african population).

2.2.2 Testing SNP correlation to PsO with Genome Wide Association Study (GWAS)

We can't keep all 1.5 million SNPs in our machine learning model. Therefore, choosing the most important features is crucial to predicting PSO. To do this, we used an algorithm called Genome wide Association (GWAS) to test how much a particular SNP is linked to the occurence of PsO. For every SNP, a logistic regression is trained with the value of the SNP (A, C, T, G or missing) as input and whether each subject has PsO (case) or not (control). From these logistic regression models, we obtain a p-value for each SNP (the null hypothesis being that the particular SNP isn't linked to PsO, the smaller the p-value, the more the SNP is correlated to PsO). Logistic regression is a simple model that computes the probability that the test subject is case (meaning that he has PsO) given the known data (the value of the SNP). A threshold is selected to decide whether to predict if the tested subject is case or control. Its simplicity allows it to be run on every SNP in a short amount of time (about 4 hours).
In order to make a decision on which SNPs to keep for the machine learning models, we generate what we call a Manhattan plot. It represents each SNP's p-value in a simple understandable way (see figure below). The x-axis represents the chromosome in which each SNP is located. The y-axis is -log(p-value), which means, the smaller the p-value the higher a SNP will score. This way we can determine which zones of the genome is linked to PSO. To select the useful SNPs, a threshold value is chosen (we keep SNPs which p-value is lower than the threshold). Usually, the threshold is established at 0.05/k where k is the number of SNPs (Bush & Moore, 2012). However, to be sure we don't lose any SNP that could be correlated to PSO, we set this threshold at 5x10-6 (the red line in the figure).

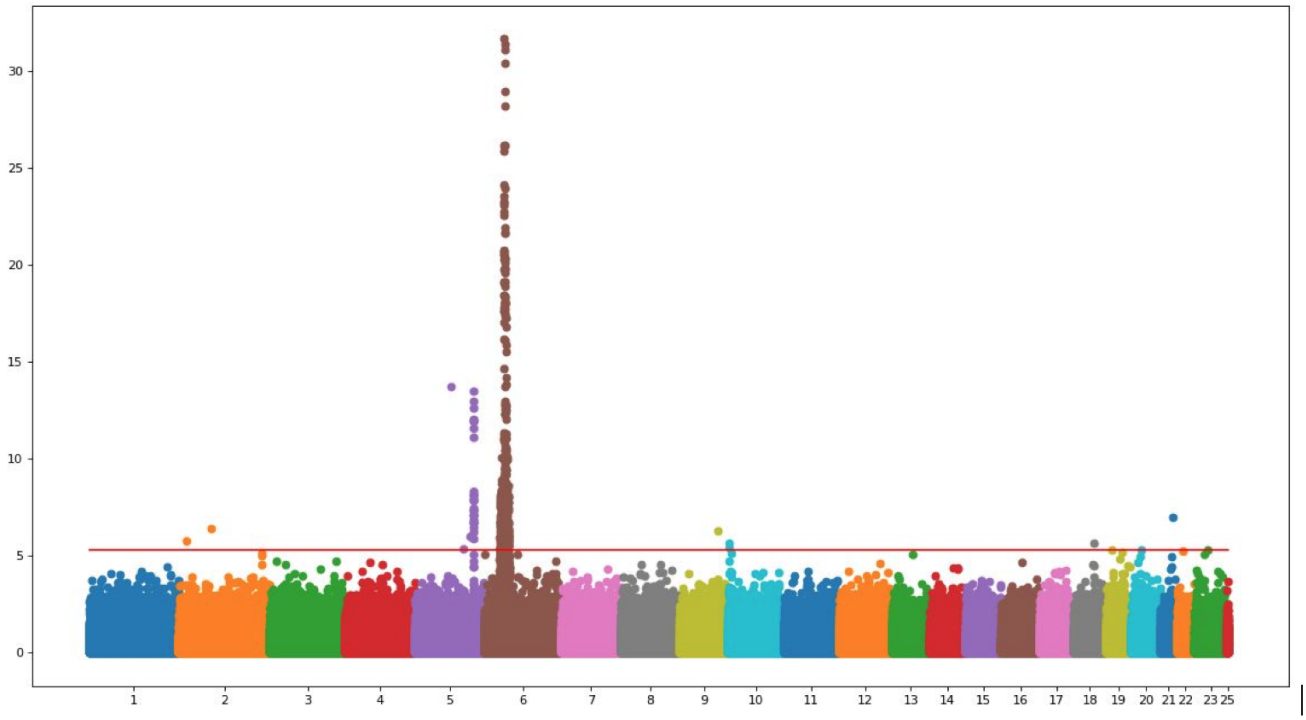


Figure 3: -log of the p values of SNP's on each chromosome. The red line is the threshold that we used to keep SNPs. We can see that most SNP related to PSO and PSA are on the chromosome 6. That is normal because most of human immune system is encoded in this chromosome, and PSO/PSA are auto-immune diseases.

2.2.3 Accounting for shared ancestry through Principal Component Analysis

There can be differences in the frequencies of different alleles (an allele is a possible value for a SNP, for example, if a SNP can be A or C, A and C are two alleles of the SNP) between different sub-population groups. This is because people of a subpopulation share a common ancestry which is different from the rest of the subjects. This must be taken into account when running GWAS, otherwise the results will not be representative of the population studied. To take sub-populations into account, we run a PCA (Principal Component Analysis) on the subject group. PCA is a process that aims to compute a reduced number of features from that starting input. Here, we want to infer three features from the original 1.5 million SNPs. This new features will represent whether a subject belongs to a particular sub-population group. By adding these new features as well as the ancestry of each subject to the GWAS, we may obtain results representative of the studied population (Price et Al., 2006).

2.2.4 Results of studying the genetic data

From running the PCA and the GWAS and then comparing the results between ancestry groups (european, african, latino, asian), we were able to select about a two thousand SNPs most correlated to PSO in order to use them in our subsequent machine learning models.

2.3 Overview of the final data ready to be used for training

BMI	ALCOHOL_STATUS	GENDER	SMOKING_STATUS	MEAN_crea	UNDER_crea	OVER_crea	MEAN_GLUC	UNDER_GLUC	OVER_GLUC
42	1	2	2	0.900000	0	0	102.250000	0	0.625000
24	1	1	1	0.975000	0	0	88.666667	0	0.000000
22	1	1	1	0.877778	0	0	98.454545	0	0.454545
32	1	2	1	0.848485	0	0	101.833333	0	0.500000
19	1	2	1	0.763636	0	0	101.923077	0	0.692308

Age	HDL_Normal	HDL_Optimal	LDL_Optimal	LDL_Low_Risk	LDL_Normal_Risk	LDL_High_Risk	LDL_Very_High_Risk	TRIGLYC_Optimal
70	0	1	1	0	0	0	0	0.000000
39	0	1	1	0	0	0	0	0.000000
86	0	1	1	0	0	0	0	1.000000
66	0	1	1	0	0	0	0	0.000000
71	0	1	1	0	0	0	0	0.333333

rs182279833_1	rs6432390_1	rs6432390_2	rs2082412_1	rs2082412_2	rs28409559_1	rs28409559_2	rs1549922_1	rs1549922_2	rs
0	1	0	0	0	0	0	0	1	
0	1	0	0	0	0	0	0	1	
0	1	0	0	0	0	0	1	0	
0	0	0	1	0	1	0	1	0	
0	0	0	0	0	0	0	0	1	

Figure 4: Example of phenotype and genotype data after preprocessing

At the end of the preprocessing, we have a dataset of 60192 raws, corresponding to the patients, and 1075 columns, which corresponds to the phenotype and genotype features selected. We also have their corresponding outputs: 1 if the patient has PsA, 0 otherwise (a vector of size (60192, 1)). Thanks to this preprocessed dataset, we can build a neural network to predict PsA. We only have 302 cases of PsA among those 60192 patients, that is only 0.5% of all patients.

3 Use of Neural Network for prediction:

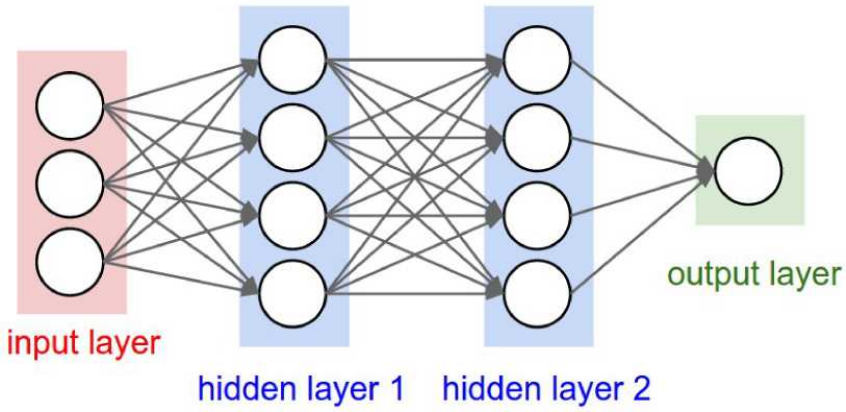


Figure 5:

3.1 Prediction of PsO:

In this part we are going to use Neural networks to predict the occurrence of psoriasis. The neural network is very sensitive to the hyperparameters that we use in our model. Since the number of parameters is big, it is not possible to train the models in a simple CPU. We had access to UCSF GPUs for the computations. The hyperparameters that we are tunning are:

- The number of features $n_{features}$ used: In order to keep the best features we use a univariate method called chi-squared method. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. For each of the 1075 features we are going to compute the value of χ^2. We then keep the $n_{features}$ features with the highest χ^2 value.
- the number of epochs n_{epochs}
- the number of nodes n_{nodes} per hidden layer
- the number if hidden layer n_{hidden}
- the weight C that we give for a case in the loss-function (see code)

.

The last hyperparameter C is crucial. In fact, the dataset is very umbalanced. There are only 6% of cases in the dataset (which correspond to the ratio of people with PsO in the population). If we keep the regular cross-entropy loss function, our model will tend to predict non-PsO for all patients. Thus, we use a different loss function:

$$Loss(y^{true}, y^{predict}) = -\sum_i C * y_i^{true} log(y_i^{predict}) + (1 - y_i^{true}) log(1 - y_i^{predict})$$

where $C > 0$

We use the area under the receiver operating characteristic curve (ROC), also called AUC, as a performance metric for our Neural network. In statistics, a receiver operating characteristic curve, i.e. ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. A perfect model has an AUC of 1, and a model that predicts always the same class and AUC of 0.5. The best result was yield for:

- $n_{features} = 30$

  The phenotype features that were kept are: BMI, Mean glucose, age,antirheumatic, alopecia, cad, crohn, diabetes, dyslipidemia, hypertension, ibd, Rheumatic arthritis, scleroderma ,sle, vitiligo, B57, Cw6

  The SNP's selected are: rs28894977, rs2873208, rs28732092, rs17196961, rs28732101, rs13203895, rs12189871, rs9468933, rs4406273, rs10484554, rs28894993, rs12212594, rs28732138.
- the number of epochs $n_{epochs} = 5$
- the number of nodes $n_{nodes} = 50$ per hidden layer
- the number if hidden layer $n_{hidden} = 2$
- the weight $C = 30$

.

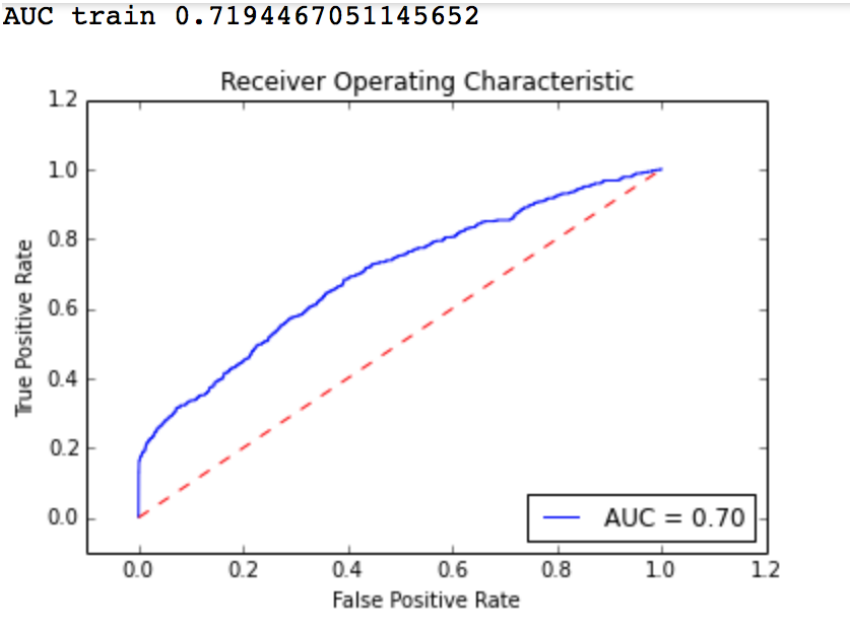


Figure 6: The AUC is a common metric for model performance.

The best neural network model yields and AUC of 0.7

3.2 Prediction of PsA among PSO:

Now we want to predict patients that will develop PsA among the people that already have PsO. Neural networks are not well suited for this problem because we have too few data. In fact, the training set has only 207 cases of PsA for a total number of patients of 3128.

The best result was yield for:

- $n_{features} = 10$

  The phenotype features that were kept are: BMI, Age, rheumatic arthritis. The genotype features kept are: rs7741091, rs17200386, rs9348876, rs9378164, rs9267677, rs13199524, rs6916062

- the number of epochs $n_{epochs} = 5$
- the number of nodes $n_{nodes} = 20$ per hidden layer
- the number if hidden layer $n_{hidden} = 2$
- the weight $C = 3$

.

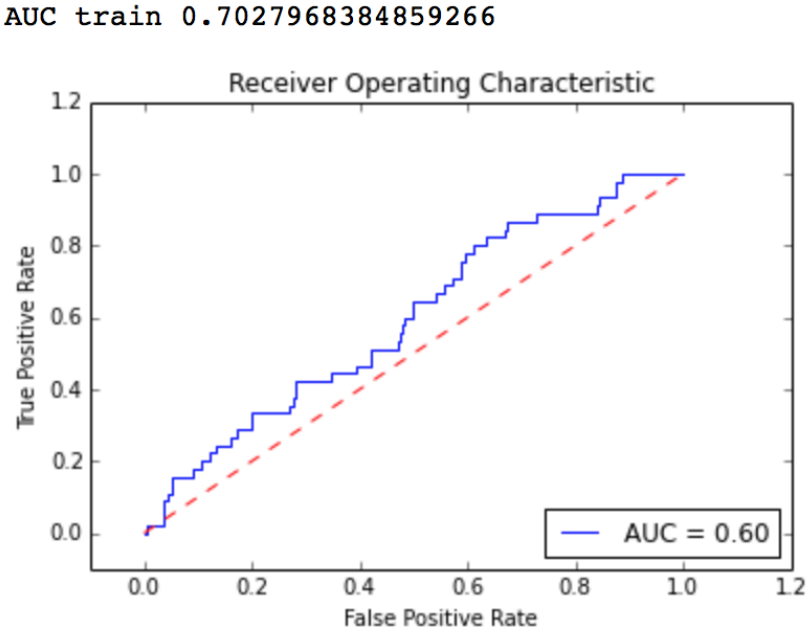


Figure 7: Best AUC for PsA among PsO prediction model

This result is very bad. This is not a surprise if we consider the number of data points used to train our model !

3.3 Prediction of PsA among all patients:

This time our training dataset has size 48153 and contains 252 cases of PsA. Our best result is given for:

- $n_{features} = 32$
- the number of epochs $n_{epochs} = 15$
- the number of nodes $n_{nodes} = 20$ per hidden layer
- the number if hidden layer $n_{hidden} = 2$
- the weight $C = 3$

.

The good result yield in this part can be explained by the high number of data that we have. However, even if with the AUC metric the performance of the neural network looks quite promising, when we look at the confusion matrices for different thresholds the result is not that good. In fact, the ratio of false negative is higher than 50%. That means that we are not detecting enough PsA among patients.

```
AUC train 0.774525492889932
```

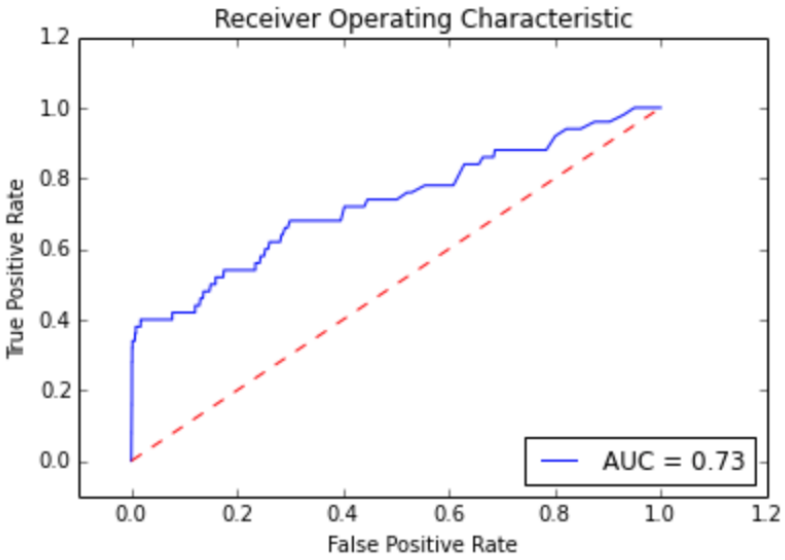


Figure 8: The best AUC for the PsA prediction model

```
========== CONFUSION MATRIX =========
threshold= 0.005
[[11347   642]
 [   30    20]]
threshold= 0.01
[[11803   186]
 [   31    19]]
threshold= 0.02
[[11842   147]
 [   31    19]]
threshold= 0.05
[[11926    63]
 [   33    17]]
threshold= 0.1
[[11964    25]
 [   33    17]]
threshold= 0.15
```

Figure 9: Confusion matrices for different thresholds

4 Conclusion

In conclusion, during this project, we applied a field that we are comfortable with, machine learning, to a field that we were completely strangers to: genomics. By using the genetic and phenotype data of 86000 patients, we were able to yield prediction results on PSO and PSA. This implied to preprocess and understand the raw data, build machine learning algorithms, train them and fine-tune them. One of the limitation of this study was the fact that the raw dataset contains uncertainties with respect to who had Psoriasis and who didn't. In fact, some people that were diagnosed with PsO might actually not have it, and vice versa.

The merge between machine learning and genomics is quickly expanding. We expect this kind of study using machine learning and biology to have promising results in the coming years with the prediction and prevention of diseases.