Machine Learning Method to Establish the Connection between Age Related Macular Degeneration and Some Genetic Variations

Antonieta Martínez-Velasco[1], Juan Carlos Zenteno[2], Lourdes Martínez-Villaseñor[1], Luis Miralles-Pechúan[1], Andric Pérez-Ortiz[1], Francisco Javier Estrada-Mena[1].

[1] Universidad Panamericana Campus México
Augusto Rodin 498, Col. Insurgentes-Mixcoac, Ciudad de México, México
{amartinezv,lmartine,lmiralles,festrada}@up.edu.mx,
andricc@me.com
[2] Instituto de Oftalmología Conde de Valenciana
Chimalpopoca 14, Col.Obrera, Ciudad de México, México
jczenteno@institutodeoftalmologia.org

Abstract. Medicine research based in machine learning methods allows the improvement of diagnosis in complex diseases. Age related Macular Degeneration (AMD) is one of them. AMD is the leading cause of blindness in the world. It causes the 8.7% of blind people. A set of case and controls study could be developed by machine-learning methods to find the relation between Single Nucleotide Polymorphisms (SNPs) SNP_A, SNP_B, SNP_C and AMD. In this paper we present a machine-learning based analysis to determine the relation of three single nucleotide SNPs and the AMD disease. The SNPs SNP_B, SNP_C remained in the top four relevant features with ophthalmologic surgeries and bilateral cataract. We aim also to determine the best set of features for the classification process.

Keywords: Single Nucleotide Polymorphisms, Macular Degeneration, machine learning, polymorphism relation

1 Introduction

Age related macular degeneration (AMD) is the leading cause of visual dysfunction and blindness in developed countries, and a rising cause in underdeveloped countries. In the United States, its prevalence in the population over age 65 years is 9% and increases to 28% in those over 75 years[1][2]. AMD is characterized by progressive degeneration of the macula, causing central field vision loss. A characteristic feature of AMD is the formation of deposits in the macula, called drusen, which may progress to either geographic atrophy (dry form) or subretinal neovascularization (wet form), manifestations of late AMD. Several genetic and environmental risk factors influence disease susceptibility. AMD is a multifactorial disease, typically caused by many genetic

variants, each with modest effect on the risk and also influenced by non-genetic/environmental factors, such as diet and smoking[3].

Multiple studies have assessed the role of genetic variants on AMD risk and progression. Some of them are consistently associated with the disease in Caucasians[4], [5], and in ethnic groups with a complex admixture of ancestral populations such as Mexican mestizos [6]. Sivakumaran et al. [7] have identified a set of SNPs for relation with AMD. This is evidence that the aforementioned genes may confer increased risk for AMD.

The prevalence of AMD varies widely across different ethnic groups in the world. However, it has been observed that in the United States is estimated there are more than 7 million individuals with early changes in the retina that places them at high risk to develop the disease. Thirty per cent of them could develop macular degeneration consistent with the early form of AMD [3].

In other populations such as Oriental, it has also become a health problem due to demographic changes, senescence and lifestyle [8]–[10]. It is estimated that by 2020, at least 80 million people globally, may be affected with AMD [11].

This substantial increase in public health burden could lead to a collapse in health systems worldwide. Not only envisages the poor response of these systems to care for these patients, and the AMD has a tremendous impact not only on physical health, also in mental health and in the economy of the geriatric population and their families[12], [13].

One of the factors that make this disease a problem of major health is that its incidence continues to rise due to the increasing number of elderly people in the world. AMD is also the leading cause of lower disability, affecting progressively central vision.

There are few studies on Mexican population. It is important to study the relation between the genetic variation in populations not traditionally studied to find the relation between ethnic differences in each genetic variants to modify disease risk[6],[14].

Usually the relation is determined by statistics techniques. Obtaining the odds ratio and p value which indicates the risk or protection against the disease[3].

Machine learning methods have been used to study the relation between SNPs and complex diseases such as cancer[15], schizophrenia[16], even AMD [17] some SNPs for Caucasian population mainly.

Currently medical research is made to relate mutations in some particular genes associated with this disease. In order to establish the relationship between risk factors and AMD, statistical information management is usually used. These studies asses the presence of Single Nucleotide Polymorphisms (SNPs) to establish the relation between them and the disease, but the results of this analysis only give the possibilities of having the disease. An improved model can be generated using machine learning techniques. This will allow more reliable and accurate predictions for the risk of developing the disease.

We constructed a dataset with 119 patients and 137 healthily subjects. Criteria for patient inclusion were as follows: (1) age 60 years or older, (2) diagnosis by a retina specialist of AMD grades 4 or 5 in both eyes or AMD grades 4 or 5 in one eye and any type of drusen in the fellow eye, (3) no relation with other retinal disease, and (4) a negative history of vitreous-retinal surgery. Subjects were classified according to Clinical Age Related Maculopathy Staging System (CARMS) where grade 4 corresponds

to geographic atrophy, while grade 5 corresponds to choroidal neovascularization [18],[3]. The AMD stage assigned was based on the most affected eye at the time of recruitment. Control subjects were enrolled from the outpatient department throughout routine ophthalmic examination. They were aged 60 years or older, had no drusen or retinal pigment epithelium (RPE) changes under dilated fundus examination, and informed a negative family history of AMD. Informed consent was signed by all subjects before they participated in the study

In this paper we present a machine-learning based analysis to determine the relation of three single nucleotide SNPs and the AMD disease. One hundred and nineteen patients and 137 controls — persons without visual injury— were recruited following a standard ophthalmologic examination protocol. This investigation was a hospital-based, case-control relation study done in a Mexican population. We included 30 clinical features for each one of the cases and controls.

We found an accuracy of 85.5% in the supervised analysis for the three Single Nucleotide Polymorphism (SNPs) to predict the disease. It is a promising result which motivates us to continue studying and testing with larger databases to improve the predictability of the disease.

The rest of the paper is as follows. A brief state of the art of the analysis of associations in complex diseases is presented in section 2. In section 3, we propose our machine-based analysis to determine the relation of three single nucleotide SNPs and the AMD disease. In section 4, experiments and results are shown and discussed. Section 5 concludes the paper and highlights future work in this context.

2. Determining Relation between SNPs and Complex Diseases

A complex disease is the medical condition that arises from an intricate interaction of inherited nature and environmental factors. Examples of them are cancer, AMD, bipolar disorder, obesity, and schizophrenia [19].

Traditionally statistical analysis is applied to determining the association between SNPs in genes and complex diseases. Various features should be considered to analyze a complex disease for sick and healthy subjects. In the main, continuous variables are compared with the Student t test, and corrected chi- square statistics are applied for categorical variables. Univariate and multivariate non conditional logistic regressions are developed to determine risk magnitude comparing each allele and genotype with the main effect employed as the binary variable[20].

Machine learning is also used to detect patterns and relations between SNPs and AMD[21]. Machine learning models generate inference rules that serve to generate new knowledge from initial data collections. The accuracy of the model will be proportional to the amount of data. Those models will predict with high accuracy the cases where the disease is present in a patient[22].

Machine learning models consider relationships among input data that cannot always be recognized by conventional analyses. In the future, complex medical diagnostic and treatment decisions will be increasingly based in machine learning models[23].

The growing interest in individualized medicine reinforces the role of predictive models, particularly of diagnostic processes. These are intended to establish a relationship between a set of variables and dependence of them have AMD, to make reliable and accurate predictions of the risks of developing the disease.

It is important to determine the risk of having the disease as soon as possible given that, as it appears in elderly subjects, it leaves us a long time to prevent the disease.

Until today machine learning methods have been used to study the relation between SNPs and complex diseases such as cancer[15], schizophrenia[16] and AMD [21].

3. Machine Learning Based Method for Relation of SNPs and AMD

In order to find the relation of single nucleotide polymorphisms SNP_A, SNP_B, and SNP_C we followed the procedure shown in figure 1.

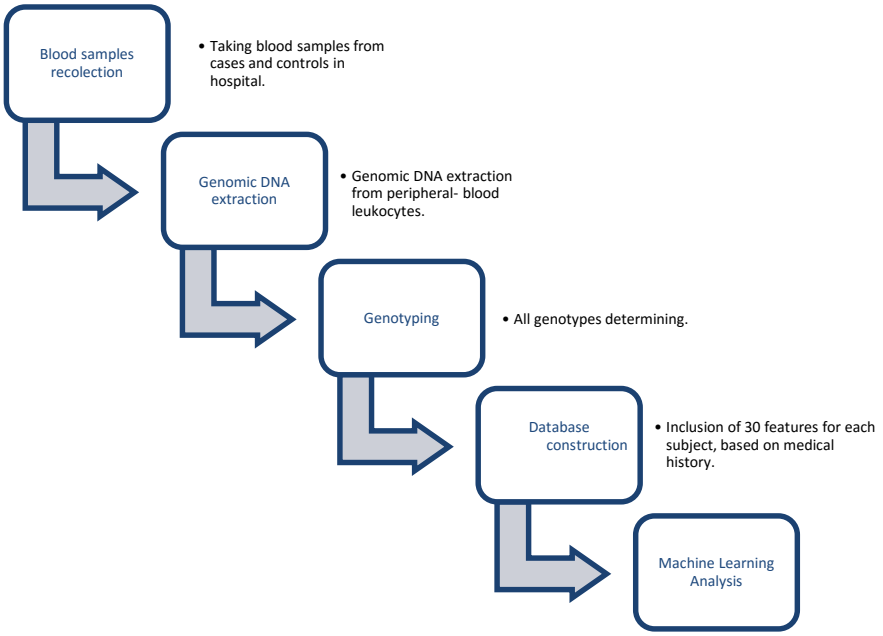

Fig. 1. Procedure to determine the relation

3.1 Data Collection Procedure

The first step was to obtain the blood samples from patients and healthy subjects in the ophthalmologic hospital.

Then genomic DNA was extracted from peripheral- blood leukocytes using the QI-Aamp DNA Blood Maxi Kit (Qiagen, Hilden, Germany).

The next step was to genotyping DNA samples with Taqman probes (Applied Biosystems). All genotypes were determined in a Real Time PCR instrument PikoReal (Thermo Scientific). The probes were acquired directly from tests on demand service Applied Biosystems as quality control verification process for genotyping. The assay was done for all samples twice.

The last step was to build the data base with the data obtained from the clinical history of each one of subjects. The data base includes 30 features as ophthalmic surgery, cataract both eyes, alcoholism, visual acuity left eye, visual acuity right eye, glaucoma, cataract left eye, pterygium, diabetic retinopathy, altered glucose, diabetes, vitreous hemorrhage, age, obesity, sex, hypercholesterolemia, xerosis both eyes, cataract right eye, presbyopia, astigmatism, smoking, hole in macula, blepharitis, posterior vitreous detachment, choroidal fracture, dyslipidemia, ectropion. The last feature is group which values "1" for cases and "0" for controls.

3.2 Analysis and Information Processing

We proposed to improve the analysis with machine learning modeling. The main reason to use machine learning techniques for our investigation is that they can detect patterns and relations than a human, or traditional statistic cannot. In addition, ML methods are able to "learn" automatically without explicit programming, effectively, and with less cost and effort[15]. These models are evaluated by metrics widely used by the scientific community such as accuracy, the root mean square error (RMSE), sensitivity or specificity [24]. These metrics allow not only to effectively evaluating model precision, but also comparing the results with the research community.

As we must predict whether the patient will suffer the disease or not determining the probability with great reliability, we used supervised classification machine learning methods.

The objective of classification methods is to learn relations between entries X and outputs Y, where $Y \in \{1, \ldots, C\}$ and C represents the number of classes.

In our experiment, models predict two outputs. Models predict the class group which has two values. The class has the value "1" for patients and "0" for healthy subjects. It is called binary classification or dual-case [25].

Among the most famous supervised classification models there are Support Vector Machine, Bayesian networks, Naive Bayes, Neural networks, Random Forest, C5.0, C4.5-like Trees, Multivariate Adaptive Regression Spline, Logistic Model Trees, and Boosted Logistic Regression.

Finally, machine learning models are typically created from a set of samples known as training set and evaluated with a different set of samples called test set. The data sets are composed of the training set that usually has the major part of the data set samples and the test set that usually has a small part of the data set samples, depending on the experimentation settings.

4 Experiments and Results

In this work, we designed two sets of experiments. The goal of the first set was to find

the importance of each feature in order to determine the relevance for the classification process of SNP_A, SNP_B, and SNP_C. We aim also to determine the best set of features for the classification process. The second set of experiments compared the performance of ten well-known supervised machine learning techniques in the classification process. The goal of these experiments is to prove the predictive power of all combinations of features, preferably testing SNP's sets of features.

4.1 Data Set

The data set used in the experiment is composed by 256 samples. Each one has 30 features, where "Group" variable represents the output of the model and the remaining 29 represent the model entries. The "Group" column has the value "1" for patients or "0" for healthy subjects. Therefore, we have a binary classification problem. Data were standardized as nominal variables, except age that remained in numeric format.

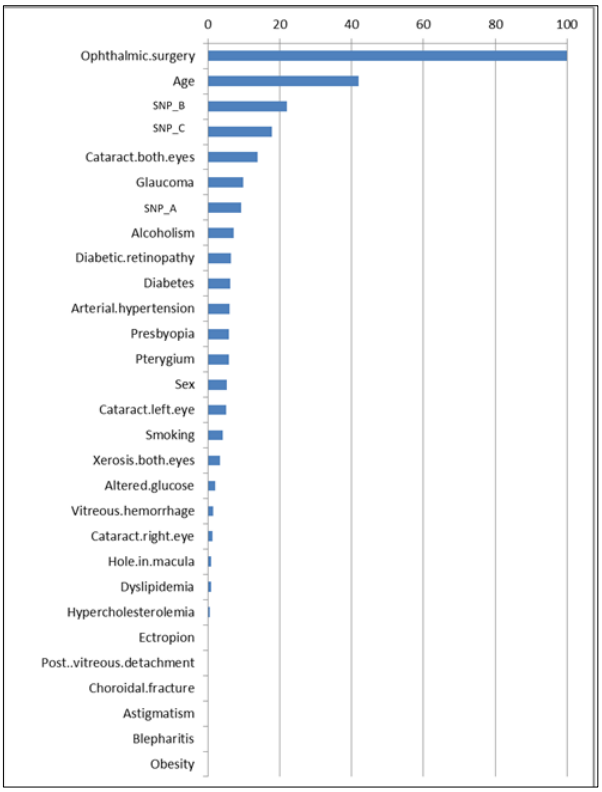

Fig. 2. Ranking of 30 features in relation with AMD.

4.2 Feature Ranking and best feature set selection

Recursive Feature Elimination (RFE) was used in order to perform feature ranking. RFE method provides high quality results. This method is a wrapper type. This technique uses Random Forest to measure the quality of each combination of features. Accuracy of generated models is the evaluation measure to determine the predictive power of the feature set in the classification process. RFE provides high levels of accuracy and optimal amount of time.

Table 1. Feature importance and combination set accuracy

N°	Feature	Accuracy	Kappa	AccuracySD	KappaSD
1	Ophthalmic surgery	0.7969	0.6024	0.05913	0.1130
2	SNP_B	0.8242	0.6537	0.05837	0.1129
3	Cataract both eyes	0.8332	0.6708	0.06078	0.1183
4	SNP_C	0.8245	0.653	0.06297	0.1233
5	Alcoholism	0.8258	0.6549	0.06180	0.1211
6	Visual acuity left eye	0.8184	0.6397	0.06295	0.1236
7	Visual acuity right eye	0.8188	0.6398	0.06689	0.1318
8	Glaucoma	0.8368	0.6755	0.07229	0.1432
9	Cataract left eye	0.8501	0.7012	0.06397	0.1271
10	Pterygium	0.861	0.7233	0.06167	0.1218
11	Diabetic retinopathy	0.8695	0.7405	0.05935	0.1171
12	Altered glucose	0.8683	0.7383	0.05875	0.1158
13	Diabetes	0.8699	0.7411	0.05917	0.1172
14	Vitreous hemorrhage	0.8722	0.746	0.05805	0.1146
15	Age	0.8691	0.7397	0.05900	0.1163
16	Obesity	0.8656	0.7324	0.05705	0.1127
17	Sex	0.8683	0.7377	0.05258	0.1044
18	Hypercholesterolemia	0.8707	0.7426	0.05774	0.1144
19	Xerosis both eyes	0.8715	0.744	0.05671	0.1125
20	Cataract right eye	0.8696	0.7401	0.05762	0.1144
21	Presbyopia	0.8704	0.7418	0.05670	0.1127
22	SNP_A	0.8715	0.7442	0.05557	0.1101
23	ARTERIAL HYPERTENSION	0.8735	0.7479	0.05511	0.1092
24	Astigmatism	0.868	0.7374	0.05862	0.1161
25	Smoking	0.8708	0.7425	0.05397	0.1072
26	Hole in macula	0.8708	0.7424	0.05478	0.1088
27	Blepharitis	0.8715	0.744	0.05447	0.1082
28	Posterior vitreous detachment	0.8747	0.7503	0.05358	0.1063
29	Choroidal fracture	0.8738	0.7486	0.05617	0.1114
30	Dyslipidemia	0.8743	0.7493	0.05606	0.1113
31	Ectropion	0.8758	0.7526	0.05625	0.1115

As is shown in Figure 2, the ranking of the variables was found by RFE method. We can observe that the SNP_B and SNP_C remained in the top four relevant features with ophthalmologic surgeries and bilateral cataract.

In the first place, RFE algorithm fits the model to all variables. Each variable is classified according to the importance to the model as it can be seen in Figure 3. Then, the RFE algorithm begins to create models using the S_i variables from i=1...S. RFE algorithm tries all possible combinations and keeps in a list the variables combination and its performance.

For each iteration, all variables are again classified. At the end of the algorithm execution, a ranking list is done using the results of all iterations as it is shown in Table 1. That explains why the variable order in Figure 2 is different from the variable order in Table 1. Finally, it is selected the combination with the highest accuracy.

As we can see in Figure 3, from the combination of ten features, models do not improve significantly with the inclusion of more features. It could be useful to include only the set of the first ten features. Because the small number of samples we used all features in the classification process in section 4.3.

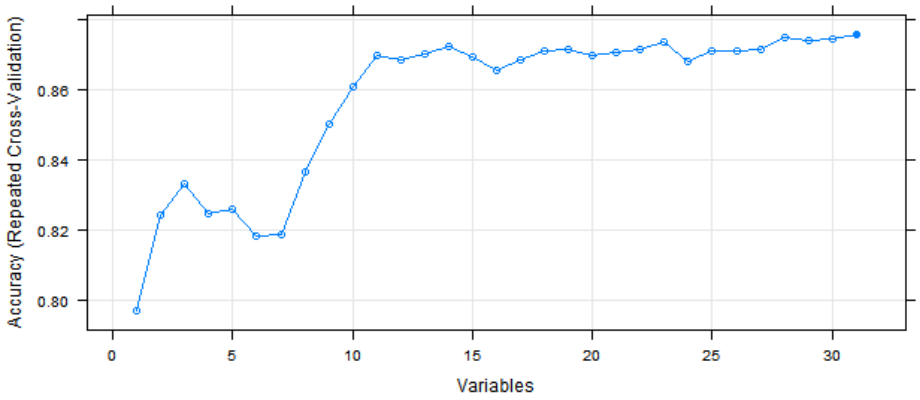

Fig. 3. RFE feature sets accuracy analysis using CV

4.3 Classification Process

The second set of experiments compared the performance of ten well-known supervised machine learning techniques in the classification process: Random Forest, Neural Networks, Naïve Bayes, Multivariate Adaptive Regression Splines, C3.4-like Trees, Stochastic Gradient Boosting and C5.0. The complete analysis were done using R software [27]. We performed the analysis with case and controls samples.

We applied the cross-validation (CV) technique given the small size of our dataset, to obtain more reliable results. CV makes different partitions and combines these to generate many models. The global accuracy is calculated as the average accuracy of all models. Models' performance was analyzed with the means of accuracy, sensitivity and specificity.

To perform the experiments, we used the Caret package of R Studio. We use as measure the accuracy, because it is measuring is the most widely used classification models[28]. Our problem is the classification as we must predict whether a patient has the disease according to the features extracted. We decided to apply CV technique with 10 partitions and 10 repetitions. The algorithm generated 100 models and the model accuracy was calculated as the average of the 100 created models. We used the same methodology and configuration for the generation of models with all selected techniques. We have 4 methods, Random Forest, C5.0, Single C5.0 Ruleset, Single C5.0 Tree, with a number of hits higher than 88%. The results of the rest of the compared

techniques were not very good.

Balance accuracy, is very useful when the samples are not balanced, i.e., when we have a higher number of samples of one class to another. This happens in our case, as we have 119 and 137 by what find us interesting show this value. Giving the same weight to each class. On the other hand, in general gives equal importance to each element. Have also shown the number of configurations that have been tested by the time expressed in thousandths of a second to build each model and each model.

The following comparison experiments were designed to determine the relevance of the three SNPs in the classification process:

1) ALL. – Experiment includes all features.
2) A. - Experiment includes only SNP_A
3) B. - Experiment includes only SNP_B
4) C. - Experiment includes only SNP_C
5) AC. - Experiment includes SNP_A and SNP_C
6) AB. – Experiment includes SNP_A and SNP_B
7) AC. – Experiment includes SNP_A and SNP_C
8) BC. – Experiment includes SNP_B and SNP_C
9) ABC. - Experiment includes SNP_A, SNP_B, and SNP_C

Table 2 shows only the best model resulting of each comparative experiment. Random Forest method is better because it has the highest accuracy, specificity and sensitivity. A good binary classification test always results with high values for all the three factors whereas a poor binary classification test results with low values for all.

"If Sensitivity is high and Specificity is low then, there is no need to worry about the excellent candidates but the poor candidates must be reexamined to eliminate false positives. If Sensitivity is low and Specificity is high, there is no need to bother about the poor candidates but the excellent candidates must be reexamined to eliminate false negatives. An average binary classification test always results with average values which are almost similar for all the three factors."[26]

Table 2. Comparative experiment results (A=SNP_A, B= SNP_B, C= SNP_C)

	Method	Acc	Sens	Spec	Prec	F1	B.Acc	Conf	Time(ms)
ALL	Random Forest	0.8585	0.9045	0.8188	0.8528	0.8779	0.8616	3	231.0
A	Neural Network	0.5666	0.4397	0.6769	0.6103	0.5112	0.5583	9	15.3
B	Naive Bayes	0.5548	0.9579	0.2046	0.5809	0.7232	0.5812	2	20.3
C	Multivariate Adaptive Regression Splines	0.6761	0.5549	0.7814	0.7450	0.6361	0.6682	1	21.2
AC	C4.5-like Trees	0.6757	0.5543	0.7812	0.7446	0.6355	0.6678	1	100.4
AB	Stochastic Gradient Boosting	0.5912	0.4438	0.7192	0.6453	0.5259	0.5815	9	8.4
AC	C4.5-like Trees	0.6757	0.5543	0.7812	0.7446	0.6355	0.6678	1	100.4
BC	Random Forest	0.7317	0.5488	0.8907	0.8525	0.6677	0.7197	1	80.1
ABC	C5.0	0.7305	0.5465	0.8904	0.8516	0.6658	0.7185	12	7.8

It is remarkable the BC experiment (which includes only SNP_B and SNP_C) because it has high predictive power with 73.17% Accuracy, 54.88% Sensitivity, and 89.07% Specificity.

The combination of the SNP_A and SNP_B shows the same behavior of them separately. Which means that they are correlated and show the same information: 59.12% accuracy. It was found that the combination of SNP_B and SNP_C results 67.57% accuracy, which means that 67.57% of the tests in the diagnosis hit AMD. As it was expected.

5 Conclusions

In this work, we used a machine learning method to develop a classification model to determine if an individual is likely to suffer AMD disease. We also proved the relevance of SNP_A, SNP_B, and SNP_C to the classification process and hence to predict if an individual will have the disease or not. For this purpose, we designed two sets of experiments. The first experiment we performed a feature ranking with RFE. The SNPs SNP_B, and SNP_C remained in the top four relevant features with ophthalmologic surgeries and bilateral cataract. The best combination is the one that includes all the variables, based on measurements of accuracy, sensitivity and specificity. It was obtained by the Random Forest technique using all features. The second set of experiments compared the performance of ten well-known supervised machine learning techniques in the classification process. Random Forest method is best because it has the highest accuracy, specificity and sensitivity.

Based on our results, we conclude that models generated with machine learning techniques support the diagnosis of AMD disease. This shows a promising scenario for the handling of large amounts of data and inference of more precise results.

For future work, we will intend to improve our model including a larger data set. It may help to refine the model and allow to improve the prediction disease. Future studies should evaluate the possibility of extending this model to other diseases.

References

1. Hageman, G. S., Gehrs, K., Johnson, L. V. & Anderson, D. Age-Related Macular Degeneration (AMD). (2008).
2. Congdon, N. *et al.* Causes and prevalence of visual impairment among adults in the United States. *Arch. Ophthalmol.* **122,** (2004).
3. Friedman, D. S. *et al.* Prevalence of age-related macular degeneration in the United States. *Arch. Ophthalmol. (Chicago, Ill. 1960)* **122,** 564–72 (2004).
4. Jager, R. D., Mieler, W. F. & Miller, J. W. Age-related macular degeneration. *N Engl J Med* **358,** (2008).
5. Patel, N., Adewoyin, T. & Chong, N. V. Age-related macular degeneration: a perspective on genetic studies. *Eye (Lond).* **22,** 768–76 (2008).
6. Buentello-Volante, B. *et al.* Susceptibility to advanced age-related macular degeneration and alleles of complement factor H, complement factor B, complement component 2, complement component 3, and age-related maculopathy susceptibility 2 genes in a Mexican population. *Mol. Vis.* **18,** 2518–25 (2012).
7. Sivakumaran, T. a. *et al.* A 32 kb critical region excluding Y402H in CFH

mediates risk for age-related macular degeneration. *PLoS One* **6,** (2011).
8. Gupta, S. K. *et al.* Prevalence of Early and Late Age-Related Macular Degeneration in a Rural Population in Northern India: The INDEYE Feasibility Study. *Invest. Ophthalmol. Vis. Sci.* **48,** 1007–1011 (2007).
9. Nirmalan, P. K. *et al.* Prevalence of Vitreoretinal Disorders in a Rural Population of SouthernIndia. *Arch. Ophthalmol.* **122,** 581 (2004).
10. Krishnaiah, S. *et al.* Risk Factors for Age-Related Macular Degeneration: Findings from the Andhra Pradesh Eye Disease Study in South India. *Investig. Opthalmology Vis. Sci.* **46,** 4442 (2005).
11. Clemons, T. E. *et al.* Risk factors for the incidence of Advanced Age-Related Macular Degeneration in the Age-Related Eye Disease Study (AREDS) AREDS report no. 19. *Ophthalmology* **112,** 533–9 (2005).
12. Berman, K. & Brodaty, H. Psychosocial effects of age-related macular degeneration. *Int. Psychogeriatr.* **18,** 415–28 (2006).
13. Rovner, B. W. *et al.* Effect of Depression on Vision Function in Age-Related Macular Degeneration. *Arch. Ophthalmol.* **120,** 1041 (2002).
14. Klein, R. *et al.* Inflammation, complement factor h, and age-related macular degeneration: the Multi-ethnic Study of Atherosclerosis. *Ophthalmology* **115,** (2008).
15. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13,** 8–17 (2015).
16. Cardoso, L. *et al.* Abstract computation in schizophrenia detection through artificial neural network based systems. *Sci. World J.* **2015,** (2015).
17. Fraccaro, P. *et al.* Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: a machine learning approach. *BMC Ophthalmol.* **15,** 10 (2015).
18. Seddon, J. M., Sharma, S. & Adelman, R. A. Evaluation of the clinical age-related maculopathy staging system. *Ophthalmology* **113,** 260–6 (2006).
19. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37,** 413–7 (2005).
20. Hadley, D. *et al.* Analysis of six genetic risk factors highly associated with AMD in the region surrounding ARMS2 and HTRA1 on chromosome 10, region q26. *Investig. Ophthalmol. Vis. Sci.* **51,** 2191–2196 (2010).
21. Simonett, J. M. *et al.* A Validated Phenotyping Algorithm for Genetic Association Studies in Age-related Macular Degeneration. *Sci. Rep.* **5,** 12875 (2015).
22. Dasgupta, A. & Sun, Y. Brief review of regressionâ€• based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genet. Epidemiol.* **35,** 1–13 (2011).
23. Hu, X. *et al.* Artificial neural networks and prostate cancer--tools for diagnosis and management. *Nat. Rev. Urol.* **10,** 174–82 (2013).
24. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45,** 427–437 (2009).
25. Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **31,** 249–268 (2007).

26. The R Foundation. The R Project for Statistical Computing. (2016). Available at: www.r-project.org. (Accessed: 13th June 2016)
27. Kuhn, M. *et al*. caret: classification and regression training. R package version 6.0-24. (2014).
28. Aswathi, B. L. Sensitivity, Specificity, Accuracy and the relationship between them. *Lifescience* (2009). Available at: http://www.lifenscience.com/bioinformatics/sensitivity-specificity-accuracy-and. (Accessed: 13th June 2016)