



Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών  
και Μηχανικών Υπολογιστών

ΔΠΜΣ ΕΔΕΜΜ

**Πρόβλεψη φαινοτύπου από δεδομένα γονότυπου με χρήση  
Polygenic risk scores και Μηχανικής Μάθησης**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΟΥΖΟΥΝΙΔΗΣ ΓΕΩΡΓΙΟΣ

Επιβλέπων : Καράντζαλος Κωνσταντίνος

Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2022



Εθνικό Μετσόβιο Πολυτεχνείο  
Σχολή Ηλεκτρολόγων Μηχανικών  
και Μηχανικών Υπολογιστών  
ΔΠΜΣ ΕΔΕΜΜ

## Πρόβλεψη φαινοτύπου από δεδομένα γονότυπου με χρήση Polygenic risk scores και Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΟΥΖΟΥΝΙΔΗΣ ΓΕΩΡΓΙΟΣ

Επιβλέπων : Καράντζαλος Κωνσταντίνος

Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1/2/2023.

.....  
Aris Anagnostopoulos  
Professor University of Rome

.....  
Κωνσταντίνος Καράντζαλος  
Αναπλ.Καθηγητής Ε.Μ.Π.

.....  
Maria Vakalopoulou  
Assistant Professor CentraleSupélec

Αθήνα, Δεκέμβριος 2022



## Περιεχόμενα

ΕΙΣΑΓΩΓΗ .....	6
ΥΠΟΒΑΘΡΟ.....	12
GWAS .....	12
Polygenic Risk Scores .....	14
Μηχανική Μάθηση και πρόβλεψη φαινοτύπου .....	20
Τεχνητό Νευρωνικό Δίκτυο.....	21
Μονοδιάστατο Συνελικτικό Δίκτυο .....	22
Αναδραστικό δίκτυο .....	23
Μακροπρόθεσμη βραχυπρόθεσμη μνήμη .....	24
Το Τυχαίο Δάσος .....	25
Μηχανή διανυσματικής υποστήριξης .....	26
Η κατάρα της διαστατικότητας.....	28
ΜΕΘΟΔΟΛΟΓΙΑ .....	30
OpenSNP .....	30
Συλλογή και επεξεργασία δεδομένων.....	32
Σετ δεδομένων .....	34
Σετ συνθετικών δεδομένων .....	35
Κατασκευή συνθετικών δεδομένων .....	36
Κωδικοποίηση δεδομένων .....	37
Επιλογή των SNPs και σημαντικά SNPs .....	39
Πειράματα .....	43
PRS και Μηχανική Μάθηση .....	44
ΑΠΟΤΕΛΕΣΜΑΤΑ .....	47
Κωδικοποίηση.....	47
Επιλογή των SNPs .....	49
Σημαντικά SNPs.....	55
PRS και Μηχανική Μάθηση .....	57
ΣΥΜΠΕΡΑΣΜΑΤΑ.....	61
Γενετικά δεδομένα και προβλήματα.....	61
Συλλογή γενετικών δεδομένων .....	63

Κωδικοποίηση των δεδομένων .....	65
Επιλογή των SNPs .....	68
Επιλογή Χαρακτηριστικών SNPs .....	70
PRS και Μηχανική Μάθηση .....	74
Επεξηγησιμότητα των μεθόδων .....	79
ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ.....	80
ΑΝΑΦΟΡΕΣ.....	82

## Εικόνες

Εικόνα 1 : Επικρατές και υπολειπόμενο αλληλόμορφο .....	7
Εικόνα 2 : Πολυγονιδιακό χαρακτηριστικό.....	8
Εικόνα 3: Πολυγονιδιακά σκορ ρίσκου (PRS) .....	9
Εικόνα 4: Πρόληψη με την χρήση PRS.....	11
Εικόνα 5: Πρόληψη με την χρήση PRS.....	11
Εικόνα 6 : GWAS .....	13
Εικόνα 7: Υπολογισμός PRS .....	18
Εικόνα 8 : Τεχνητό Νευρωνικό δίκτυο .....	22
Εικόνα 9 : Μονοδιάστατο ΣΥνελικτικό Δίκτυο .....	23
Εικόνα 10 : Αναδραστικό δίκτυο.....	24
Εικόνα 11 : Μακροπρόθεσμη βραχυπρόθεσμη μνήμη .....	25
Εικόνα 12: Τυχαίο Δάσος .....	26
Εικόνα 13: Μηχανή διανυσματικής υποστήριξης .....	27
Εικόνα 14: Η κατάρα της διαστατικότητας.....	29
Εικόνα 15: OpenSNP .....	31
Εικόνα 16: Βήματα συλλογής δεδομένων .....	33
Εικόνα 17: Σετ δεδομένων .....	35
Εικόνα 18: Διαδικασία κατασκευής συνθετικών δεδομένων .....	37
Εικόνα 19: Κωδικοποίηση δεδομένων .....	38
Εικόνα 20: Επιλογή SNPs – GWAS .....	40
Εικόνα 21: Relief .....	41
Εικόνα 22 GMM .....	42
Εικόνα 23 GMM 2 .....	42
Εικόνα 24 Πίνακας συχνοτήτων.....	43
Εικόνα 25: Δίκτυα Μηχανικής-Βαθιάς Μάθησης .....	45
Εικόνα 26: Διαδικασία πρόβλεψης με PRS.....	46
Εικόνα 27: Αποτελέσματα κωδικοποίησης .....	48
Εικόνα 28: Αποτελέσματα επιλογής SNP,σετ 1 .....	50

Εικόνα 29: Αποτελέσματα επιλογής SNP,σετ 2 .....	51
Εικόνα 30: Αποτελέσματα επιλογής SNP,σετ 3 .....	52
Εικόνα 31: Αποτελέσματα επιλογής SNP,σετ 4 .....	53
Εικόνα 32: Αποτελέσματα επιλογής SNP,σετ 5 .....	54
Εικόνα 33: Σημαντικά SNPs ανά αριθμό χαρακτηριστικών .....	56
Εικόνα 34: Μηχανική Μάθηση και PRS μέθοδοι.....	57
Εικόνα 35: Διάγραμμα loss Dense μοντέλου.....	58
Εικόνα 36: Διάγραμμα accuracy Dense μοντέλου .....	58
Εικόνα 37: Κατανομή προβλέψεων .....	59
Εικόνα 38: Κατανομή προβλέψεων με διαφορετικά κατώφλια .....	60
Εικόνα 39: Αποτελέσματα περιαιμάτων κωδικοποίησης.....	67
Εικόνα 40: Σύγκριση σχημάτων κωδικοποίησης .....	67
Εικόνα 41: Συγκριτική ακρίβεια ανά μέθοδο .....	69
Εικόνα 42: Μέγιστη ακρίβεια ανά μέθοδο.....	70
Εικόνα 43: Σημαντικά SNPs ανά μέθοδο, συγκριτικά με την βιβλιογραφία .....	72
Εικόνα 44: Σύγκριση ακρίβειας στις προβλέψεις, GMM vs Relief .....	73
Εικόνα 45:Σύγκριση ακρίβειας στις προβλέψεις, RF vs SVM .....	73
Εικόνα 46: Κατανομή των προβλέψεων .....	75
Εικόνα 47: Διαφορές ML-PRS .....	78
Εικόνα 48: Μοντελοποίηση interactions.....	78

## ΠΕΡΙΛΗΨΗ

Η τεράστια τεχνολογική εξέλιξη γύρω από τον τομέα της Μηχανικής Μάθησης απαιτεί την άμεση ενσωμάτωση της σε όλους τους πιθανούς τομείς. Σχετικά με τα γενετικά δεδομένα, μέχρι και σήμερα η Μηχανική Μάθηση δεν έχει καταφέρει να εισχωρήσει και να κερδίσει την εμπιστοσύνη της ιατρικής κοινότητας. Σε αντίθεση, η μέθοδος PRS είναι η ενδεδειγμένη σχετικά με τις προβλέψεις φαινοτυπικών χαρακτηριστικών από γενετικά δεδομένα.

Στην συγκεκριμένη εργασία αναζητήθηκαν γενετικά δεδομένα μέσα από την διαδικτυακή πύλη OpenSNP. Έπειτα από διάφορα στάδια συλλογής και προεπεξεργασίας των δεδομένων, κατασκευάστηκαν σετ δεδομένων που συνδέονται με το χρώμα των ματιών και τον διαβήτη τύπου Β. Τα σετ αυτά κατασκευάστηκαν με τέτοιο τρόπο ώστε να εξυπηρετήσουν το πρόβλημα της δυαδικής ταξινόμησης σε κατηγορίες φαινοτύπων (πχ γαλάζια/καστανά μάτια)

Πάνω σε αυτά τα σετ δεδομένων έγιναν διάφορα πειράματα. Πρώτος στόχος των πειραμάτων αυτών ήταν να προσδιοριστεί ποια είναι η καλύτερη μορφή κωδικοποίησης των δεδομένων προκειμένου εφαρμοστούν ταξινομητές Μηχανικής Μάθησης. Στη συνέχεια, εξετάστηκαν διάφορες μέθοδοι μείωσης των διαστάσεων των χαρακτηριστικών, πάνω στην ίδια λογική με την επιλογή των σημαντικών SNPs στις μεθόδους PRS. Ακόμα, έγινε σύγκριση των καθοριστικών SNPs που εντόπισαν οι αλγόριθμοι Μηχανικής Μάθησης για το χρώμα των ματιών σε σχέση με αυτά που αναφέρει η βιβλιογραφία.

Τέλος, δημιουργήθηκε ένα συνθετικό σετ δεδομένων προκειμένου να συγκριθούν οι μέθοδοι PRS με δίκτυα Μηχανικής Μάθησης. Πάνω στην σύγκριση αυτή, τα δίκτυα της Μηχανικής Μάθησης αποδείχθηκαν πολύ αποτελεσματικά και ξεπέρασαν σε ακρίβεια τα αντίστοιχα μοντέλα PRS. Προέκυψε πως η προβλεπτική ικανότητα των μοντέλων Μηχανικής Μάθησης είναι αρκετά μεγαλύτερη, αρκεί να υπάρχει μεγάλη διαθεσιμότητα δεδομένων.





## ABBREVIATION

The huge technological development around the field of Machine Learning requires its immediate integration into all possible areas. Regarding genetic data, to this day Machine Learning has not been able to penetrate and gain the trust of the medical community. In contrast, the PRS method is appropriate for predicting phenotypic traits from genetic data.

In this paper, genetic data were searched through the OpenSNP portal. After various stages of data collection and preprocessing, data sets linked to eye color and type B diabetes were constructed.

Various experiments were carried out on these data sets. The first objective of these experiments was to determine what is the best form of data coding in order to apply Machine Learning classifiers. Subsequently, various methods of reducing the dimensions of the characteristics were examined, based on the same logic as the selection of important SNPs in the PRS methods. In addition, the decisive SNPs identified by machine learning algorithms for eye color were compared to those reported in the literature.

Finally, a synthetic data set was created in order to compare PRS methods with Machine Learning networks. Based on this comparison, the Machine Learning networks proved to be very effective and surpassed in accuracy the corresponding PRS models. It emerged that the predictive capacity of Machine Learning models is much higher, as long as there is high data availability.



## ΕΙΣΑΓΩΓΗ

Όλοι οι άνθρωποι διαφέρουμε μεταξύ μας. Συνεπώς διαφέρουν και τα φυσικά χαρακτηριστικά μας. Οι λόγοι για τους οποίους διαφέρουμε εμείς και τα χαρακτηριστικά μας, ως ανθρώπινο είδος, θα μπορούσαν να κατηγοριοποιηθούν σε 2 βασικούς. Στους βιολογικούς και στους περιβαλλοντικούς.

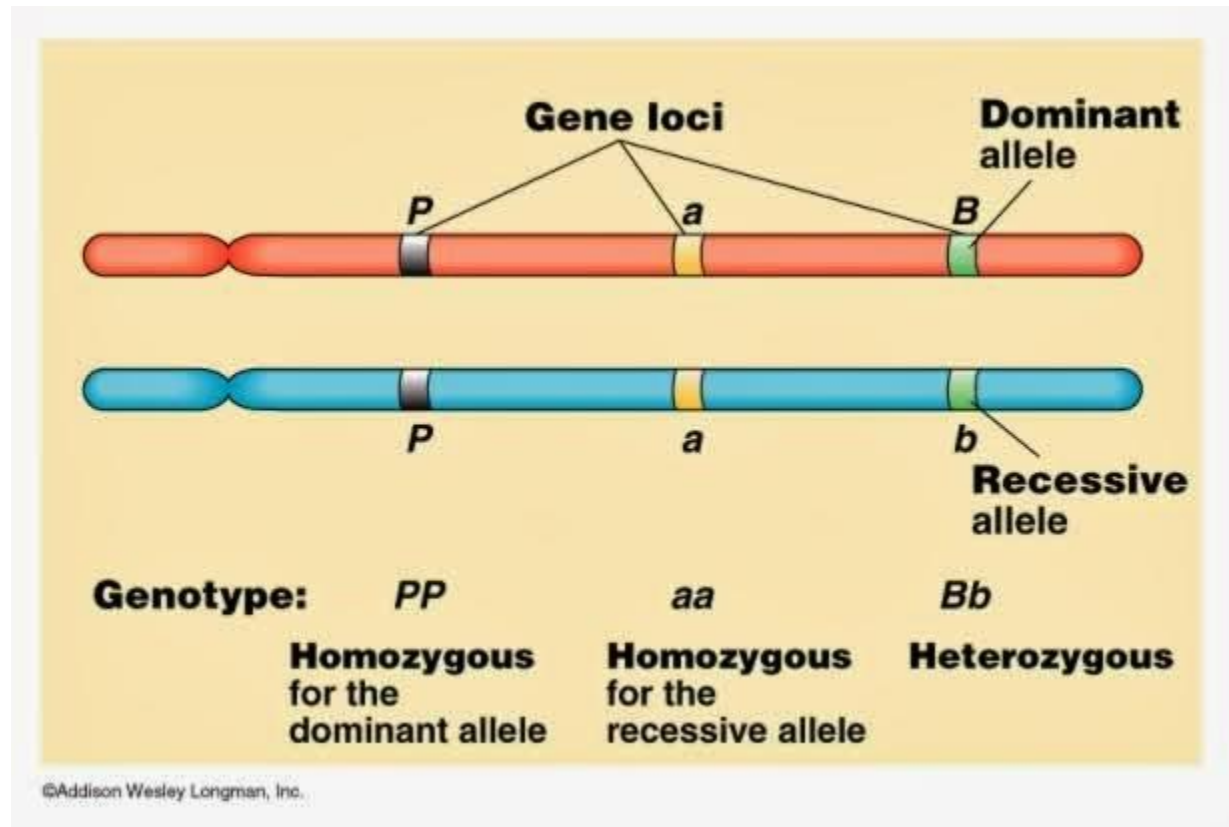
Εξετάζοντας του περιβαλλοντικούς λόγους, υπάρχουν περίπου 53 διαφορετικοί πληθυσμοί σε όλο τον κόσμο. Μπορεί κανείς να παρατηρήσει πως οι εξωτερικές διαφορές μεταξύ των ανθρώπων είναι λιγότερες ανάμεσα στους ανθρώπους του ίδιου πληθυσμού και περισσότερες ανάμεσα στους πληθυσμούς ή φυλές ( πχ Αφρικανοί και Ευρωπαίοι). Η εξελικτική θεωρία λέει πως αν δυο άνθρωποι έχουν μεγαλώσει σε διαφορετικά περιβάλλοντα, τότε υπάρχει μεγάλη πιθανότητα η συμπεριφορά τους, τα χαρακτηριστικά τους και η εξέλιξή τους να είναι διαφορετική. Έτσι λοιπόν, γίνεται κατανοητός ο ρόλος που παίζει το περιβάλλον στην εξέλιξη του είδους και στις διαφορές που προκύπτουν στους οργανισμούς που ζουν εντός αυτού.

Οι βιολογικοί λόγοι για τους οποίους διαφέρουμε μεταξύ μας είναι και αυτοί καθοριστικοί. Η γενετική δομή του ανθρώπου εκφράζεται με τις αλλαγές στις συχνότητες αλληλόμορφων ή γενικότερα του γενοτύπου μέσα στο χρόνο. Αυτό εκφράζεται μέσω της μετάλλαξης, της μετανάστευσης, της φυσικής επιλογής, της γενετικής απόκλισης ή και με το μη τυχαίο ζευγάρωμα. Συνεπώς, οι αλλαγές του DNA δημιουργούν νέα αλληλόμορφα και αποτελούν την πηγή της γενετικής ποικιλότητας.

Στην παρούσα εργασία θα εστιάσουμε στους βιολογικούς λόγους και στις γενετικές διαφορές που παρουσιάζονται στο ανθρώπινο είδος. Το ανθρώπινο DNA αποτελείται από 3 δισεκατομμύρια βάσεις, και περισσότερες από το 99,9% από αυτές τις βάσεις είναι οι ίδιες σε όλους τους ανθρώπους. Η μεταξύ αυτών διαφορά εντοπίζεται μόλις στο 0,1%.

Το αλληλόμορφο είναι μία από τις 2 ή περισσότερες μορφές ενός γονιδίου. Ο άνθρωπος αποτελεί ένα διπλοειδή οργανισμό, δηλαδή διαθέτει 2 αλληλόμορφα που βρίσκονται στην ίδια θέση των ομόλογων χρωμοσωμάτων. Το γονότυπο αποτελείται από το ζευγάρι των αλληλόμορφων, ενώ η έκφραση τους συνιστά τον φαινότυπο. Από τα 2 αλληλόμορφα, συνήθως το ένα επικρατεί έναντι του άλλου και καθορίζει το φαινότυπο. Συνεπώς, υπάρχουν το επικρατές και το υπολειπόμενο αλληλόμορφο. Ενώ υπάρχουν διαφορετικά μεγέθη αλληλόμορφων, το μικρότερο δυνατό μέγεθος που μπορεί να έχει ένα αλληλόμορφο ονομάζεται Πολυμορφισμός μονού νουκλεοτιδίου (SNP). Αυτό θα

είναι το μέγεθος που θα χρησιμοποιήσουμε στην παρούσα διπλωματική εργασία, ώστε να προσδιορίσουμε το φαινότυπο του κάθε ανθρώπου μέσα από τη γενετική του πληροφορία.



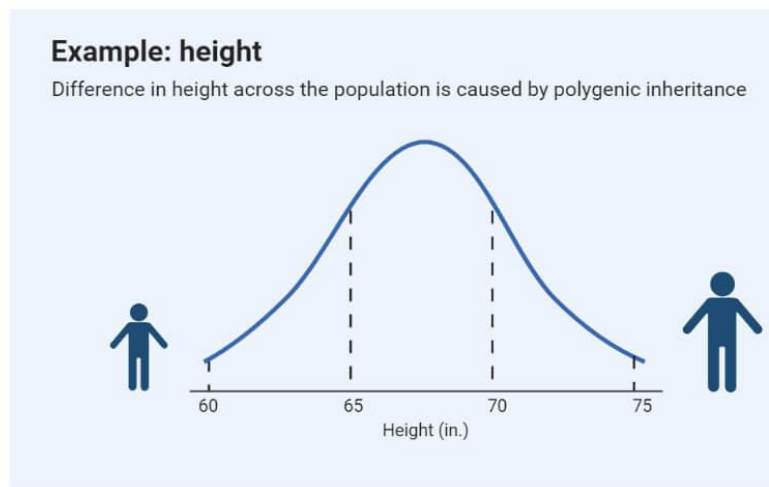
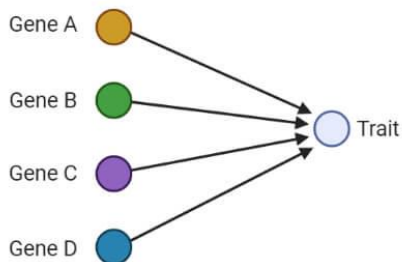
Εικόνα 1 : Επικρατές και υπολειπόμενο αλληλόμορφο

Μέσα από μελέτες συσχέτισης σε όλο το γονιδίωμα μπορούμε να καταλάβουμε τη σχέση μεταξύ των γενοτύπων και των φαινοτύπων. Έτσι η θετική συσχέτιση μεταξύ ενός πολυμορφισμό μονού νουκλεοτιδίου και ενός φαινοτύπου μπορεί να μας βοηθήσει στην ταξινόμηση αυτή. Συγκεκριμένα, οι μελέτες συσχέτισης σε όλο το γονιδίωμα (GWAS) περιλαμβάνουν ταχεία σάρωση γενετικών δεικτών σε ολόκληρο το γονιδίωμα, πάνω σε έναν πληθυσμό για την εύρεση γενετικών παραλλαγών που σχετίζονται με μια συγκεκριμένη ασθένεια. Τα αποτελέσματα του GWAS είναι χρήσιμα για τους ερευνητές να αναπτύξουν καλύτερες στρατηγικές για την ανίχνευση, τη θεραπεία και την πρόληψη της νόσου.

Τα περισσότερα φαινότυπα ή ασθένειες στο ανθρώπινο είδος είναι πολυγονιδιακά. Αυτό σημαίνει ότι δεν επηρεάζονται μόνο από ένα στοιχείο του γονιδιώματος, αλλά συνδυαστικά από περισσότερα. Τα GWAS ήταν ιδιαίτερα χρήσιμα στην αποκρυπτογράφηση της γενετικής αρχιτεκτονικής των διαταραχών, που προκαλούνται από μεταλλάξεις σε συγκεκριμένο μέρος του γονιδιώματος. Ωστόσο, οι πολύπλοκες ασθένειες δεν καθορίζονται από έναν παράγοντα στο γονιδίωμα, αλλά επηρεάζονται από την αλληλεπίδραση πολλαπλών παραγόντων. Στην περίπτωση αυτή, κάθε παράγοντας θα πρέπει να έχει μόνο μια μικρή ατομική συμβολή, και η επιρροή του ασκείται μέσω της αλληλεπιδράσεως με τους άλλους παράγοντες. Επιπλέον, οι πολύπλοκες ασθένειες τείνουν να συνεπάγονται μεγαλύτερες δυσκολίες στον ορισμό του φαινοτύπου, έτσι ώστε να απαιτείται να μετρηθούν πολλαπλά και διάφορα χαρακτηριστικά-παράγοντες προκειμένου να εντοπιστεί μια διαταραχή.

## POLYGENIC INHERITANCE

### Multiple genes control a single trait














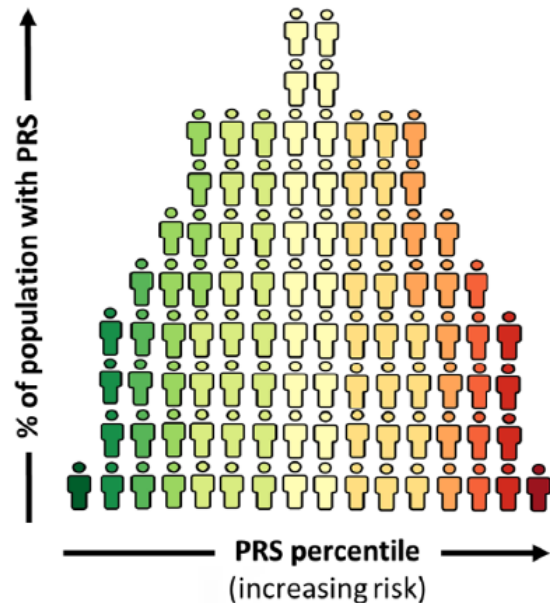
Εικόνα 2 : Πολυγονιδιακό χαρακτηριστικό

Η συνήθης διαδικασία για τη μελέτη της γενετικής συσχέτισης σχετίζεται με την επεξήγηση των γραμμικών συσχετίσεων και αλληλεπιδράσεων μεταξύ γενετικών παραλλαγών και φυσιολογικών χαρακτηριστικών. Τα μοντέλα γραμμικής παλινδρόμησης χρησιμοποιούνται συνήθως για την κατανόηση της σχέσης μεταξύ των γενετικών παραγόντων και μιας κλινικής εξόδου (ασθένεια ή σχετικό χαρακτηριστικό). Έτσι, η συμπεριφορά αυτών των τεχνικών θα εξαρτηθεί σε μεγάλο βαθμό από το βαθμό

της γραμμικότητας ή μη γραμμικότητας στη χαρτογράφηση μεταξύ τόπων και χαρακτηριστικών. Μέχρι και σήμερα, οι τεχνικές αυτές δεν έχουν καταφέρει να αποκρυπτογραφήσουν τις πολύπλοκες συσχετίσεις μεταξύ των γονιδιακών παραγόντων ώστε να υπάρχουν σταθερές και αποτελεσματικές προβλέψεις. Ίσως οι ήδη υπάρχουσες τεχνικές να μην μπορούν να επιτελέσουν τον σκοπό αυτό. Ως εκ τούτου, η διερεύνηση των μη γραμμικών συσχετίσεων μεταξύ γενετικών ουσιών και κλινικών χαρακτηριστικών μπορεί να παρέχει χρήσιμες πληροφορίες για την κατανόηση της γενετικής δομής σύνθετων ασθενειών.

Τα πολυγονιδιακά σκορ ρίσκου (PRS) και η μηχανική-βαθιά μάθηση είναι οι δυο βασικοί ανταγωνιστές προσδιορισμού του φαινοτύπου ή της ασθένειας μέσα από το γονιδίωμα. Τα PRS είναι η παραδοσιακή στατιστική μέθοδος, όπου το ρίσκο προσδιορίζεται με βάση την άθροιση όλων των γενετικών δεικτών οι οποίοι έχουν συσχέτιση με το εξεταζόμενο φαινότυπο ή ασθένεια. Χρησιμοποιεί τις μελέτες συσχέτισης στο γονιδίωμα ώστε να προσδιορίσει το πόσο επηρεάζει ως προς το τελικό χαρακτηριστικό ο κάθε γενετικός δείκτης.

	PRS percentile	Risk of disease vs. reference group
	0-1	Lowest ↓
	1-5	
	5-10	
	10-20	
	20-40	
	<b>40-60 (reference)</b>	1
	60-80	↑ Highest
	80-90	
	90-95	
	95-99	
	99-100	



Source: RGA

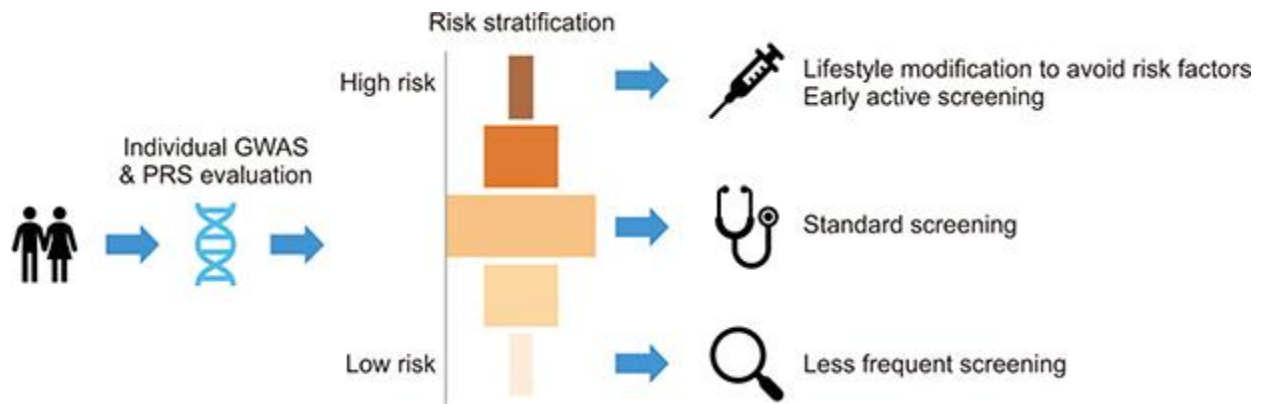
Εικόνα 3: Πολυγονιδιακά σκορ ρίσκου (PRS)

Η μηχανική μάθηση και η βαθιά μάθηση χρησιμοποιούν μοντέλα τα οποία κάνουν προβλέψεις διερευνώντας σε βάθος όλες τις γραμμικές και μη γραμμικές συσχετίσεις μεταξύ του γονιδιώματος και του φαινοτύπου. Τα μοντέλα αυτά έχουν αυξημένες δυνατότητες αφού μπορούν να εξάγουν πολύ αποτελεσματικά τα πρότυπα που παρατηρούνται, από δεδομένα μεγάλης διάστασης και πολυπλοκότητας όπως το ανθρώπινο γενότυπο.

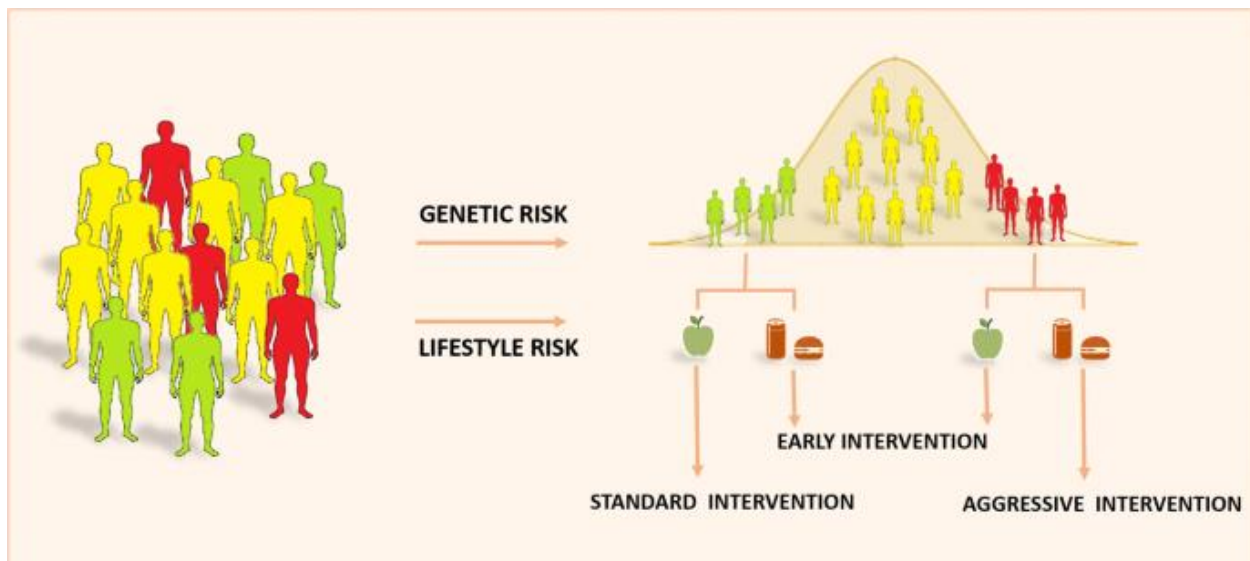
Ενώ η τεράστια τεχνολογική εξέλιξη που σχετίζεται με τη μηχανική και βαθιά μάθηση απαιτεί την ολοένα και εντονότερη εφαρμογή της σε κάθε επιστημονικό πεδίο, οι συγκεκριμένες μέθοδοι στον κλάδο της υγείας δεν προσδίδουν την απαραίτητη εμπιστοσύνη ώστε να τις υιοθετήσει η ιατρική κοινότητα. Οι βασικοί λόγοι είναι οι εξής: Αρχικά έστω ένα αξιόπιστο και αποτελεσματικό μοντέλο μηχανικής μάθησης, το οποίο γενικεύει τα αποτελέσματα των προβλέψεων του με ικανοποιητική ακρίβεια. Αυτό θα απαιτεί ένα τεράστιο μέγεθος δεδομένων προκειμένου να εκπαιδευτεί. Από τη στιγμή που τα γενετικά δεδομένα αποτελούν ευαίσθητα προσωπικά δεδομένα, δεν είναι εύκολο ούτε να συλλεχθούν σε μια μεγάλη βάση αλλά ούτε να τα προμηθευτεί κανείς. Επιπρόσθετα, είναι αδύνατον για κάποιον που παίρνει ιατρικές αποφάσεις να στηριχθεί στην απόφαση ενός μοντέλου μηχανικής μάθησης χωρίς να έχει επαρκείς εξηγήσεις και γενικότερη κατανόηση για την απόφαση αυτή. Αυτά τα μοντέλα βασίζονται σε υπερπαραμέτρους και αρχιτεκτονικές, οι οποίες επιλέγονται με διαφορά συγκριτικά τεστ και τελικό γνώμονα την ακρίβεια. Πολλές φορές όμως το ότι ένα μοντέλο αποδίδει με ακρίβεια δεν σημαίνει απαραίτητα ότι η απόφαση που πήρε είναι αντικειμενική και ορθή. Σε πολύ σημαντικές ιατρικές αποφάσεις που μπορεί να κοστίσουν και ζωές, πάντα θα προτιμάτε για απόφαση ενός εξειδικευμένου ατόμου από την απόφαση ενός μοντέλου και την ελλιπή επεξηγησιμότητα της.

Εν αντιθέσει με τα μοντέλα μηχανικής μάθησης, τα PRS είναι πιο απλά και κατανοητά. Μπορεί να μην περιέχουν τις δυνατότητες κατανόησης όλων των πολύπλοκων και μη γραμμικών συσχετίσεων μεταξύ του γονιδιώματος, αλλά η μέθοδος αυτή εξηγείται πολύ πιο εύκολα όντας πιο απλή. Βασίζεται στις μελέτες συσχέτισης στο γονιδίωμα, που γίνονται ανάμεσα σε ομογενείς πληθυσμούς που μοιράζονται το προς μελέτη χαρακτηριστικό.





Εικόνα 4: Πρόληψη με την χρήση PRS



Εικόνα 5: Πρόληψη με την χρήση PRS

## ΥΠΟΒΑΘΡΟ

### GWAS

Κατά τη διάρκεια των τελευταίων δύο δεκαετιών, υπήρξε ένα αυξανόμενο ενδιαφέρον για τη διερεύνηση στις επιδράσεις των γενετικών παραγόντων στην ποικιλότητα των ανθρώπινων χαρακτηριστικών. Τα τεχνικά και αναλυτικά εργαλεία που απαιτούνται για τη διεξαγωγή γενετικών μελετών έχουν γίνει ολοένα και πιο προσβάσιμα. Αυτή η αυξημένη προσβασιμότητα οδηγεί σε μεγάλες υποσχέσεις, καθώς ερευνητές εκτός του τομέα στις γενετικής μπορεί να φέρουν νέα εμπειρογνωμοσύνη στον τομέα (π.χ. πιο βαθιά γνώση στις νοσολογίας των ψυχιατρικών χαρακτηριστικών). Ωστόσο, η διεξαγωγή μελετών γενετικής συσχέτισης με σωστό τρόπο απαιτεί ειδικές γνώσεις γενετικής, στατιστικής και (βιο)πληροφορικής.

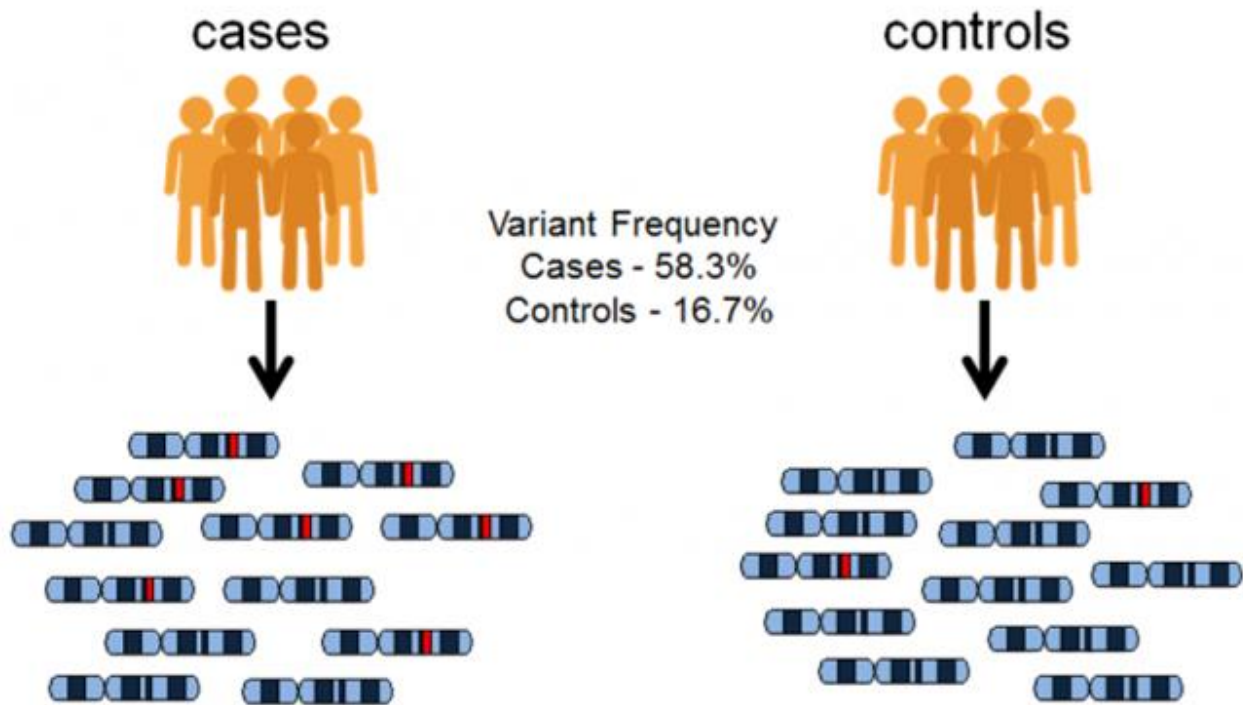
Ο στόχος των μελετών συσχέτισης σε επίπεδο γονιδιώματος (GWAS) είναι να εντοπιστούν μονονουκλεοτιδικοί πολυμορφισμοί (SNPs) οι οποία παίζουν έναν καθοριστικό ρόλο για την εξέλιξη των τιμών των φαινοτύπων ή χαρακτηριστικών στις εξέταση. Ο προσδιορισμός των SNPs που σχετίζονται με χαρακτηριστικά μπορεί στη συνέχεια να αποκαλύψει νέες γνώσεις σχετικά με στις βιολογικούς μηχανισμούς που διέπουν την προέλευση αυτών των χαρακτηριστικών.

Οι τεχνολογικές εξελίξεις επιτρέπουν τη διερεύνηση μεγάλου αριθμού SNPs που κατανέμονται σε όλο το γονιδίωμα. Μέχρι σήμερα, τα GWAS έχουν επιτύχει να αποκαλύψουν SNPs που συμβάλλουν στον κίνδυνο ψυχιατρικών χαρακτηριστικών, συμπεριλαμβανομένης στις σχιζοφρένειας, των διαταραχών του φάσματος του αυτισμού, στις διαταραχές ελλειμματικής προσοχής και υπερκινητικότητας, στις μείζονος καταθλιπτικής διαταραχής και στις διπολικής διαταραχής. Η συνολική εικόνα αυτών των αποτελεσμάτων υποδηλώνει ότι τα ψυχιατρικά χαρακτηριστικά επηρεάζονται από τα SNPs, που έχουν το καθένα μικρά ατομικά μεγέθη επίδρασης.

Τα GWAS βασίζονται σε μεγάλο βαθμό στη βαθιά γνώση της γενετικής αρχιτεκτονικής του ανθρώπινου γονιδιώματος, η οποία παρασχέθηκε από δύο σημαντικές ερευνητικές πρωτοβουλίες, δηλαδή το Διεθνές Πρόγραμμα HarMap και το πρόγραμμα 1000 Genomes. Το Διεθνές Έργο HarMap περιέγραψε τα πρότυπα των κοινών SNPs εντός στις αλληλουχίας του ανθρώπινου DNA, ενώ το πρόγραμμα 1000 Genomes (1KG) παρείχε έναν χάρτη τόσο κοινών όσο και σπάνιων SNPs. Επειδή τα αποτελέσματα του GWAS



έδειξαν ότι τα μεγέθη επίδρασης των μεμονωμένων SNPs είναι μικρά, οι ερευνητές ανέπτυξαν ενδιαφέρον για μεθόδους που συγκεντρώνουν την επίδραση των SNPs.



Εικόνα 6 : GWAS

## Polygenic Risk Scores

Στη συνέχεια θα συγκεκριμένα αναφορά στην ανάλυση στις βαθμολογίας πολυγονιδιακού κινδύνου (PRS), καθώς ότι αυτή είναι η πιο σχετική μέθοδος, καθώς είναι αρκετά εύκολη στη διεξαγωγή, ενώ μπορεί να εφαρμοστεί σε δείγματα-στόχους με σχετικά μέτρια μεγέθη δειγμάτων χωρίς την απαίτηση μεγάλου όγκου δεδομένων. Η PRS μέθοδος συνδυάζει τα μεγέθη επίδρασης πολλαπλών SNP σε μία συγκεντρωτική βαθμολογία που μπορεί να χρησιμοποιηθεί για την πρόβλεψη του κινδύνου ασθένειας.

Η PRS είναι μια βαθμολογία σε ατομικό επίπεδο που υπολογίζεται με βάση τον αριθμό των παραλλαγών κινδύνου που φέρει ένα άτομο, σταθμισμένες με τα μεγέθη επίδρασης SNP που προκύπτουν από ένα ανεξάρτητο μεγάλης κλίμακας GWAS. Ως εκ τούτου, η βαθμολογία αποτελεί ένδειξη του συνολικού γενετικού κινδύνου στις συγκεκριμένου ατόμου για ένα συγκεκριμένο χαρακτηριστικό, το οποίο μπορεί να χρησιμοποιηθεί για κλινική πρόβλεψη ή προσυμπτωματικό έλεγχο. Ωστόσο, η διακριτική του ακρίβεια δεν είναι επαρκής για κλινικές εφαρμογές.

Η PRS έχει συμβάλει στη γνώση μας για τη γενετική αρχιτεκτονική των διάφορων χαρακτηριστικών των ασθενειών με την ικανότητά στις να προβλέπει την κατάσταση στις νόσου. Έχει χρησιμοποιηθεί περαιτέρω για να διερευνηθεί εάν τα μεγέθη γενετικού αποτελέσματος που λαμβάνονται από ένα GWAS συγκεκριμένου φαινοτύπου ενδιαφέροντος μπορούν να χρησιμοποιηθούν για την πρόβλεψη του κινδύνου στις άλλου φαινοτύπου.

Βασικό βήμα που θα πρέπει να αποτελεί μέρος οποιουδήποτε GWAS είναι η χρήση του κατάλληλου ποιοτικού ελέγχου. Χωρίς εκτεταμένο ποιοτικό έλεγχο, το GWAS δεν θα παράγει αξιόπιστα αποτελέσματα. Σφάλματα στα δεδομένα μπορεί να προκύψουν για πολλούς λόγους, λόγω κακής ποιότητας δειγμάτων DNA, κακού υβριδισμού DNA στη συστοιχία, ανεπαρκούς απόδοσης ανιχνευτών γονότυπου και ανάμειξης δειγμάτων ή μόλυνσης. Τα επτά συνήθεις βήματα ποιοτικού ελέγχου αποτελούνται από φιλτράρισμα των SNP και των ατόμων με βάση τα ακόλουθα:

- I. Ατομική έλλειψη πληρότητας ή έλλειψη σε επίπεδο SNP
- II. ασυνέπειες στο εκχωρημένο και γενετικό φύλο των ατόμων (διαφορά φύλου)
- III. μικρή συχνότητα του υπολειπούμενου αλληλόμορφου (MAF)
- IV. αποκλίσεις από την Hardy-Weinberg Equilibrium (HWE)
- V. ποσοστό ετεροζυγωτικότητας
- VI. συγγένεια
- VII. φυλετικές ακραίες τιμές (διαφορά πληθυσμών).

- Πολυμορφισμός στις νουκλεοτιδίου (SNP): Αυτή είναι μια παραλλαγή σε ένα μόνο νουκλεοτίδιο (δηλαδή, A, C, G ή T) που εμφανίζεται σε μια συγκεκριμένη θέση στο γονιδίωμα. Ένα SNP υπάρχει συνήθως ως δύο διαφορετικές μορφές (π.χ. A εναντίον T). Αυτές οι διαφορετικές μορφές ονομάζονται αλληλόμορφα. Ένα SNP με δύο αλληλόμορφα έχει τρεις διαφορετικούς γονότυπους (π.χ. AA, AT και TT).
- Ατομική έλλειψη πληρότητας: Στις είναι ο αριθμός των SNP που λείπουν για ένα συγκεκριμένο άτομο.
- Έλλειψη επιπέδου SNP: Στις είναι ο αριθμός των ατόμων στο δείγμα για τα οποία λείπουν πληροφορίες σχετικά με ένα συγκεκριμένο SNP. Στο βήμα αυτό, εξαιρούνται τα αρχικά τα SNPs εκείνα που απουσιάζουν σε ένα μεγάλο ποσοστό από τον στις μελέτη πληθυσμό. Ακόμα εξαιρούνται τα άτομα εκείνα του πληθυσμού που παρουσιάζουν υψηλά ποσοστά έλλειψης SNP από αυτά που στις αντιστοιχούν.
- Διαφορά φύλου: Αυτή είναι η διαφορά μεταξύ του εκχωρημένου φύλου και του φύλου που καθορίζεται με βάση τον γονότυπο. Μια ασυμφωνία πιθανότατα υποδεικνύει ανάμειξη δειγμάτων στο εργαστήριο. Αυτή η δοκιμή μπορεί να διεξαχθεί μόνο όταν έχουν αξιολογηθεί τα SNPs στα φυλετικά χρωμοσώματα (X και Y). Στο βήμα αυτό αφαιρούνται τα δείγματα εκείνα που έχουν αναντίστοιχο εκχωρημένο και καθορισμένο φύλο.

- Μικρή συχνότητα αλληλομόρφων (MAF): Αυτή είναι η συχνότητα του λιγότερο συχνά εμφανιζόμενου αλληλόμορφου σε μια συγκεκριμένη θέση. Οι περισσότερες μελέτες δεν είναι ικανές να ανιχνεύσουν συσχετίσεις με SNP με χαμηλό MAF και ως εκ τούτου αυτά τα SNPs πρέπει να αφαιρεθούν. Στο βήμα αυτό αφαιρούνται τα δείγματα εκείνα που έχουν MAF χαμηλότερο από ένα συγκεκριμένο όριο.
- Ο νόμος Hardy-Weinberg ισορροπίας (HWE): Ο νόμος στις αφορά τη σχέση μεταξύ των συχνοτήτων αλληλόμορφου και γονότυπου. Υποθέτει έναν απεριόριστα μεγάλο πληθυσμό, χωρίς επιλογή, μετάλλαξη ή μετανάστευση. Ο νόμος ορίζει ότι ο γονότυπος και οι συχνότητες των αλληλομόρφων είναι σταθερές για γενιές. Η παραβίαση του νόμου HWE δείχνει ότι οι συχνότητες γονότυπου διαφέρουν σημαντικά από στις προσδοκίες (π.χ. εάν η συχνότητα του αλληλόμορφου A = 0,20 και η συχνότητα του αλληλόμορφου T = 0,80, η αναμενόμενη συχνότητα του γονότυπου AT είναι  $2 * 0,2 * 0,8 = 0,32$ ) και η παρατηρούμενη συχνότητα δεν πρέπει να διαφέρει σημαντικά. Στο GWAS, θεωρείται γενικά ότι οι αποκλίσεις από το HWE είναι αποτέλεσμα σφαλμάτων συλλογής του γονότυπου. Τα κατώτατα όρια HWE σε περιπτώσεις είναι συχνά λιγότερο αυστηρά από εκείνα των ελέγχων, καθώς η παραβίαση του νόμου HWE σε περιπτώσεις μπορεί να είναι ενδεικτική πραγματικής γενετικής συσχέτισης με τον κίνδυνο ασθένειας. Στο βήμα αυτό αφαιρούνται τα δείγματα εκείνα που έχουν HWE χαμηλότερο από ένα συγκεκριμένο όριο.
- Ετεροζυγωτικότητα: Αυτή είναι η μεταφορά δύο διαφορετικών αλληλόμορφων στις συγκεκριμένου SNP. Το ποσοστό ετεροζυγωτικότητας στις άτομου είναι το ποσοστό των ετερόζυγων γονότυπων. Τα υψηλά επίπεδα ετεροζυγωτικότητας μέσα σε ένα άτομο μπορεί να αποτελούν ένδειξη χαμηλής ποιότητας δείγματος, ενώ τα χαμηλά επίπεδα ετεροζυγωτικότητας μπορεί να οφείλονται στην ενδογαμία. Στο βήμα αυτό αφαιρούνται τα δείγματα εκείνα που διαθέτουν ακραίες τιμές.

- **Σχετικότητα:** Αυτό δείχνει πόσο έντονα ένα ζευγάρι ατόμων σχετίζεται γενετικά. Ένα συμβατικό GWAS υποθέτει ότι όλα τα θέματα είναι άσχετα (δηλαδή, κανένα ζευγάρι ατόμων δεν είναι πιο στενά συνδεδεμένο από στις συγγενείς δεύτερου βαθμού). Χωρίς κατάλληλη διόρθωση, η συμπερίληψη συγγενών θα μπορούσε να οδηγήσει σε μεροληπτικές εκτιμήσεις των τυπικών σφαλμάτων των μεγεθών των επιπτώσεων SNP. Στο βήμα αυτό αφαιρούνται τα δείγματα εκείνα που παρουσιάζουν έντονη σχετικότητα με βάση στατιστικά τεστ.
- **Διαστρωμάτωση πληθυσμού:** Αυτή είναι η παρουσία πολλαπλών υποπληθυσμών (π.χ. άτομα με διαφορετικό φυλετικό υπόβαθρο) σε μια μελέτη. Επειδή οι συχνότητες των αλληλομόρφων μπορεί να διαφέρουν μεταξύ των υποπληθυσμών, η διαστρωμάτωση του πληθυσμού μπορεί να οδηγήσει σε ψευδώς θετικές συσχετίσεις ή/και να συγκαλύψει στις αληθείς συσχετίσεις. Μια σημαντική πηγή συστηματικής μεροληψίας στο GWAS είναι η διαστρωμάτωση του πληθυσμού. Είναι αποδεδειγμένο πως μπορεί να υπάρχει και ακόμα διαστρωμάτωση του πληθυσμού μεταξύ ενιαίων εθνικών πληθυσμών. Για τον λόγο αυτό η βέλτιστη πρακτική είναι πάντα το σύνολο των δειγμάτων να αναφέρεται στον ίδιο πληθυσμό.

Μετά τον ποιοτικό έλεγχο, τα δεδομένα είναι έτοιμα για στις επακόλουθες δοκιμές συσχέτισης. Ανάλογα με το αναμενόμενο γενετικό μοντέλο του χαρακτηριστικού ή στις ασθένειας που ενδιαφέρει και τη φύση του φαινοτυπικού χαρακτηριστικού που μελετήθηκε, μπορεί να επιλεγεί η κατάλληλη στατιστική εξέταση.

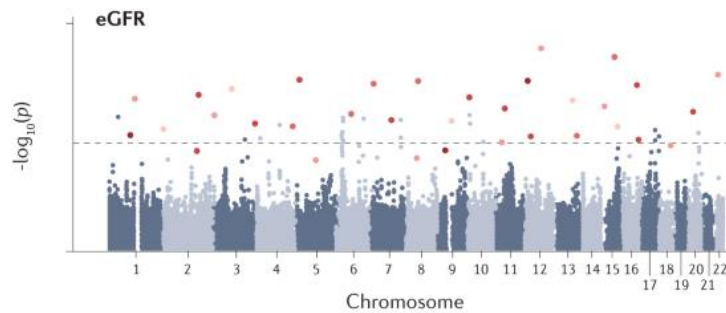
Η ανάλυση συσχέτισης στις SNP ήταν από παλαιότερα η κύρια μέθοδος στο GWAS, αλλά απαιτεί πολύ μεγάλα μεγέθη δειγμάτων για την ανίχνευση λίγων σε αριθμό SNPs για πολλά σύνθετα χαρακτηριστικά. Αντίθετα, η ανάλυση PRS δεν στοχεύει στον εντοπισμό μεμονωμένων SNPs, αλλά αντίθετα συσσωματώνει τον γενετικό κίνδυνο σε όλο το γονιδίωμα σε μια μεμονωμένη πολυγονιδιακή βαθμολογία για ένα χαρακτηριστικό ενδιαφέροντος.

Σε αυτή την προσέγγιση, απαιτείται ένα μεγάλο δείγμα ανακάλυψης για να προσδιοριστεί αξιόπιστα πόσο κάθε SNP αναμένεται να συμβάλει στην πολυγονιδιακή

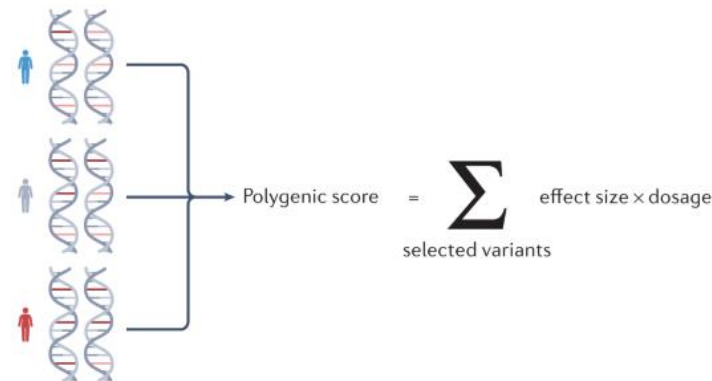
βαθμολογία («βάρη») στις συγκεκριμένους χαρακτηριστικού. Κατά κανόνα στις αναλύσεις αυτές, αρκούν 5000 δείγματα συσχετισμένα με το χαρακτηριστικό μελέτης.

Για τη διεξαγωγή ανάλυσης PRS, λαμβάνονται ειδικά βάρη χαρακτηριστικών από μια GWAS. Στο σύνολο των δειγμάτων που συσχετίζονται με το στις εξέταση χαρακτηριστικό, υπολογίζεται ένα PRS για κάθε άτομο με βάση το σταθμισμένο άθροισμα του αριθμού των αλληλόμορφων κινδύνου που φέρει πολλαπλασιαζόμενο με τα ειδικά βάρη χαρακτηριστικών.

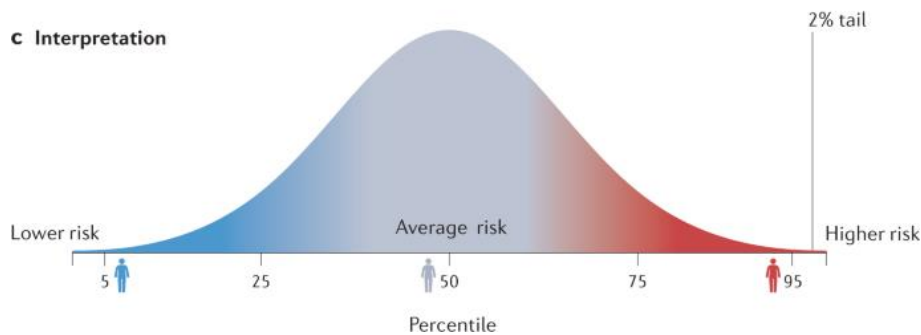
**a Selection of genetic variants from GWAS**



**b Calculation of the polygenic score**



**c Interpretation**



Εικόνα 7: Υπολογισμός PRS

Αν και καταρχήν όλα τα κοινά SNP θα μπορούσαν να χρησιμοποιηθούν σε μια ανάλυση PRS, είναι συνηθισμένο να συσσωρεύονται πρώτα τα αποτελέσματα GWAS πριν υπολογιστούν οι βαθμολογίες κινδύνου. Η συσσώρευση είναι η διαδικασία με την οποία επιλέγονται συγκεκριμένα SNP στο εύρος του γονιδιώματος τα οποία αντιπροσωπεύουν τα γειτονικά στις. Εφαρμόζεται σε συγκεκριμένους τόπους και ομάδες στο γονιδίωμα, οι οποίες καθορίζονται από τα LD blocks.

Ανισορροπία σύνδεσης (LD): Αυτό είναι ένα μέτρο μη- τυχαία συσχέτιση μεταξύ αλληλόμορφων σε διαφορετικούς τόπους στο ίδιο χρωμόσωμα σε έναν δεδομένο πληθυσμό. Τα SNP είναι σε LD όταν η συχνότητα συσχέτισης των αλληλόμορφων στις είναι μεγαλύτερη από το αναμενόμενο σε τυχαία ποικιλία. LD ανησυχίες μοτίβα συσχετισμών μεταξύ SNP.

Τα κατώτατα όρια τιμών p-value χρησιμοποιούνται συνήθως για την κατάργηση των SNP που εμφανίζουν ελάχιστα ή καθόλου στατιστικά στοιχεία για συσχέτιση (π.χ. διατηρούν μόνο SNP με τιμές  $p < 0,5$  ή  $< 0,1$ ). Συνήθως, πραγματοποιούνται πολλαπλές αναλύσεις PRS, με ποικίλα όρια για στις τιμές p. Μόλις υπολογιστεί η PRS για όλα τα άτομα στα δείγματα, οι βαθμολογίες μπορούν να χρησιμοποιηθούν σε μια λογιστική ανάλυση παλινδρόμησης για να προβλέψουν το στις μελέτη χαρακτηριστικό. Η αύξηση στις ακρίβειας πρόβλεψης PRS δείχνει την αύξηση στις ακρίβειας πρόβλεψης που εξηγείται από γενετικούς παράγοντες κινδύνου. Η ακρίβεια πρόβλεψης στις PRS εξαρτάται κυρίως από τη κληρονομικότητα των αναλυθέντων χαρακτηριστικών, τον αριθμό των SNPs και το μέγεθος του δείγματος ανακάλυψης.

## Μηχανική Μάθηση και πρόβλεψη φαινοτύπου

Καθώς ζούμε στην εποχή των μεγάλων δεδομένων, η μετατροπή των μεγάλων δεδομένων σε πολύτιμη γνώση έχει γίνει σημαντική περισσότερο από ποτέ. Σίγουρα, η βιοπληροφορική δεν αποτελεί εξαίρεση σε τέτοιες τάσεις. Διάφορες μορφές βιοϊατρικών δεδομένων, συμπεριλαμβανομένων των δεδομένων του γονιδιώματος, στις εικόνες και του σήματος, έχουν συσσωρευτεί σημαντικά και οι μεγάλες δυνατότητές στις στη βιολογική έρευνα και την έρευνα στον τομέα στις υγειονομικής περίθαλψης έχουν προσελκύσει τα συμφέροντα στις βιομηχανίας καθώς και στις ακαδημαϊκής κοινότητας. Για παράδειγμα, η IBM παρείχε το Watson for Oncology, μια πλατφόρμα που αναλύει στις ιατρικές πληροφορίες των ασθενών και βοηθά στις κλινικούς ιατρούς με θεραπευτικές επιλογές. Επιπλέον, η Google DeepMind, επιτυγχάνοντας μεγάλη επιτυχία με το AlphaGo στο παιχνίδι go, ξεκίνησε πρόσφατα το DeepMind Health για την ανάπτυξη αποτελεσματικών τεχνολογιών υγειονομικής περίθαλψης.

Για την εξαγωγή γνώσης από τεράστια δεδομένα στη βιοπληροφορική, η μηχανική μάθηση ήταν μια από τις πιο ευρέως χρησιμοποιούμενες μεθοδολογίες. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν δεδομένα εκπαίδευσης για να αποκαλύψουν υποκείμενα μοτίβα, να δημιουργήσουν ένα μοντέλο και, στη συνέχεια, να κάνουν προβλέψεις για τα νέα δεδομένα με βάση το μοντέλο. Μερικοί από τις γνωστούς αλγόριθμους – διανυσματική μηχανή υποστήριξης, κρυφό μοντέλο Markov, Bayesian, δίκτυα – έχουν εφαρμοστεί στη μελέτη των γονιοδιωμάτων, των πρωτεϊνών, τη βιολογία συστημάτων και πολλούς στις τομείς. Οι συμβατικοί αλγόριθμοι μηχανικής μάθησης έχουν περιορισμούς στην επεξεργασία στις ακατέργαστης μορφής δεδομένων, οπότε οι ερευνητές κατέβαλαν τεράστια προσπάθεια για τη μετατροπή στις ακατέργαστης μορφής σε κατάλληλα χαρακτηριστικά υψηλού επιπέδου αφαίρεσης με σημαντική τεχνογνωσία στον τομέα.

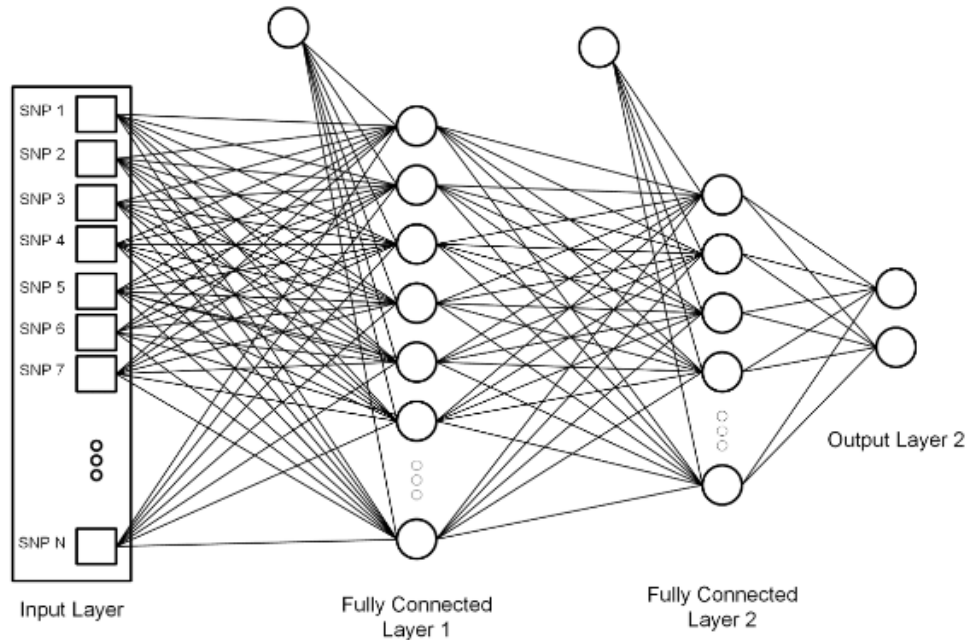
Από την άλλη, η βαθιά μάθηση, στις στις τύπος αλγορίθμου μηχανικής μάθησης, εμφανίστηκε με βάση τα μεγάλα δεδομένα, τη δύναμη του παράλληλου και κατανομημένου υπολογισμού και στις εξελιγμένους αλγόριθμους. Οι αλγόριθμοι βαθιάς μάθησης έχουν ξεπεράσει στις προηγούμενους περιορισμούς και σημειώνουν σημαντικές προόδους σε διάφορους τομείς στις η αναγνώριση εικόνας, η αναγνώριση ομιλίας και η επεξεργασία φυσικής γλώσσας. Σίγουρα, η βιοπληροφορική δεν αποτελεί εξαίρεση στις εφαρμογές βαθιάς μάθησης.



## Τεχνητό Νευρωνικό Δίκτυο

Ένα ΤΝΔ έχει εκατοντάδες ή χιλιάδες ολοκληρωμένους, τεχνητούς νευρώνες που ονομάζονται μονάδες επεξεργασίας. Οι μονάδες εισόδου και εξόδου αποτελούνται από αυτές τις μονάδες επεξεργασίας. Με βάση ένα εσωτερικό σχήμα στάθμισης, οι μονάδες εισόδου λαμβάνουν ποικίλες πηγές και δομές πληροφοριών και το νευρωνικό δίκτυο στοχεύει να μάθει από τις πληροφορίες που παρέχονται για τη δημιουργία μιας αναφοράς εξόδου.

Τα ANNs χρησιμοποιούν συχνά μια σειρά από αρχές μάθησης που ονομάζονται backpropagation, μια συντομογραφία για την προς τα πίσω διάδοση του λάθους, για να βελτιώσουν τις επιδόσεις τους, ακριβώς όπως οι άνθρωποι χρειάζονται οδηγίες και οδηγίες για να καταλήξουν σε ένα συμπέρασμα ή έξοδο. Ένα ANN αρχικά περνάει από την διαδικασία της εκπαίδευσης όπου μαθαίνει να εντοπίζει τις τάσεις στα SNPs. Το δίκτυο αντιπαραβάλλει την πραγματική παραγωγή του με αυτό που προοριζόταν να επιτύχει την επιθυμητή έξοδο κατά τη διάρκεια αυτής της ελεγχόμενης διαδικασίας. Χρησιμοποιώντας backpropagation, η διαφορά μεταξύ όλων των επιδράσεων τροποποιείται. Αυτό υποδηλώνει ότι το δίκτυο λειτουργεί προς τα πίσω, μετακινούμενο από τη μονάδα εξόδου στις μονάδες εισόδου για να αλλάξει το βάρος των αλληλεπιδράσεών του μεταξύ των μονάδων έως ότου δημιουργηθεί το χαμηλότερο δυνατό σφάλμα από την απόκλιση μεταξύ του πραγματικού και του αναμενόμενου αποτελέσματος. Κάθε σύνολο δεδομένων μεταβιβάζεται στο ANN με ετικέτες εξόδου. Το κύριο πλεονέκτημα της χρήσης του ANN είναι η μη γραμμικότητα που παράγεται από τη λειτουργία ενεργοποίησης.

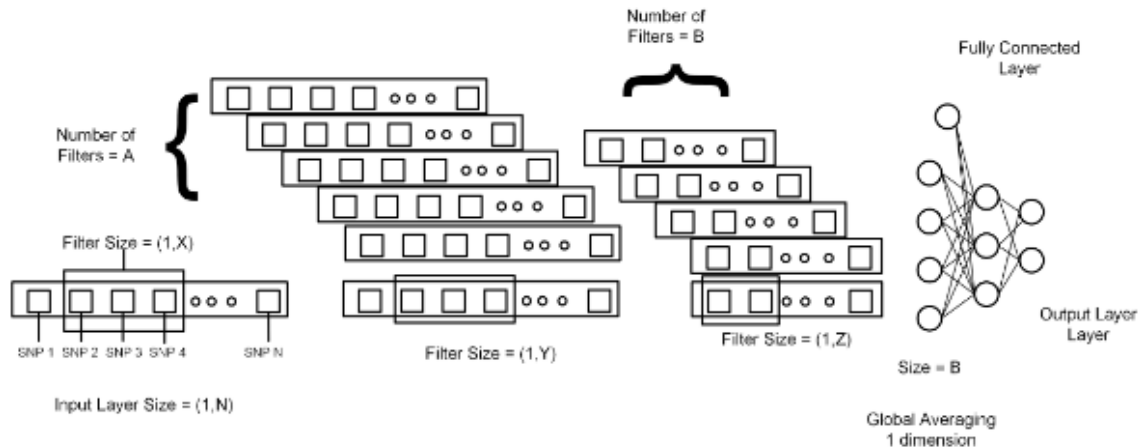


Εικόνα 8 : Τεχνητό Νευρωνικό δίκτυο

### Μονοδιάστατο Συνελικτικό Δίκτυο

Η βαθιά μάθηση είναι μέρος της μηχανικής μάθησης και μπορεί να διαδραματίσει σημαντικό ρόλο σε πραγματικές εφαρμογές, όπως η βιοπληροφορική και η υπολογιστική βιολογία, η τηλεπισκόπηση, η φωτογραμμετρία, η ιατρική και η 3D μοντελοποίηση. Η ψηφιακή ανάλυση σήματος και εικόνας με τη χρήση μεθόδων βαθιάς μάθησης, ιδιαίτερα συνελκτικών νευρωνικών δικτύων, είναι ένα πεδίο που έχει αναπτυχθεί σε μεγάλο βαθμό και συνεχίζει. Για την αναγνώριση εικόνας, έχουν δημιουργηθεί μοντέλα νευρωνικού δικτύου συνέλιξης (CNN). Ο αλγόριθμος δέχεται μια δισδιάστατη είσοδο που αντιπροσωπεύει τα εικονοστοιχεία και τα κανάλια χρώματος μιας εικόνας σε μια διαδικασία που ονομάζεται εκμάθηση χαρακτηριστικών. Είναι δυνατή η επέκταση της ίδιας μεθόδου σε μονοδιάστατες ακολουθίες δεδομένων. Το μοντέλο αντλεί χαρακτηρισμούς από δεδομένα ακολουθίας και χαρτογραφεί τα εσωτερικά χαρακτηριστικά της ακολουθίας. Ένα 1DCNN είναι πολύ επιτυχημένο στην εξαγωγή χαρακτηριστικών από το τμήμα σταθερού μήκους του συνόλου δεδομένων. Τα γενετικά δεδομένα είναι διαδοχικές πληροφορίες, επομένως είναι δυνατή η χρήση του 1DCNN

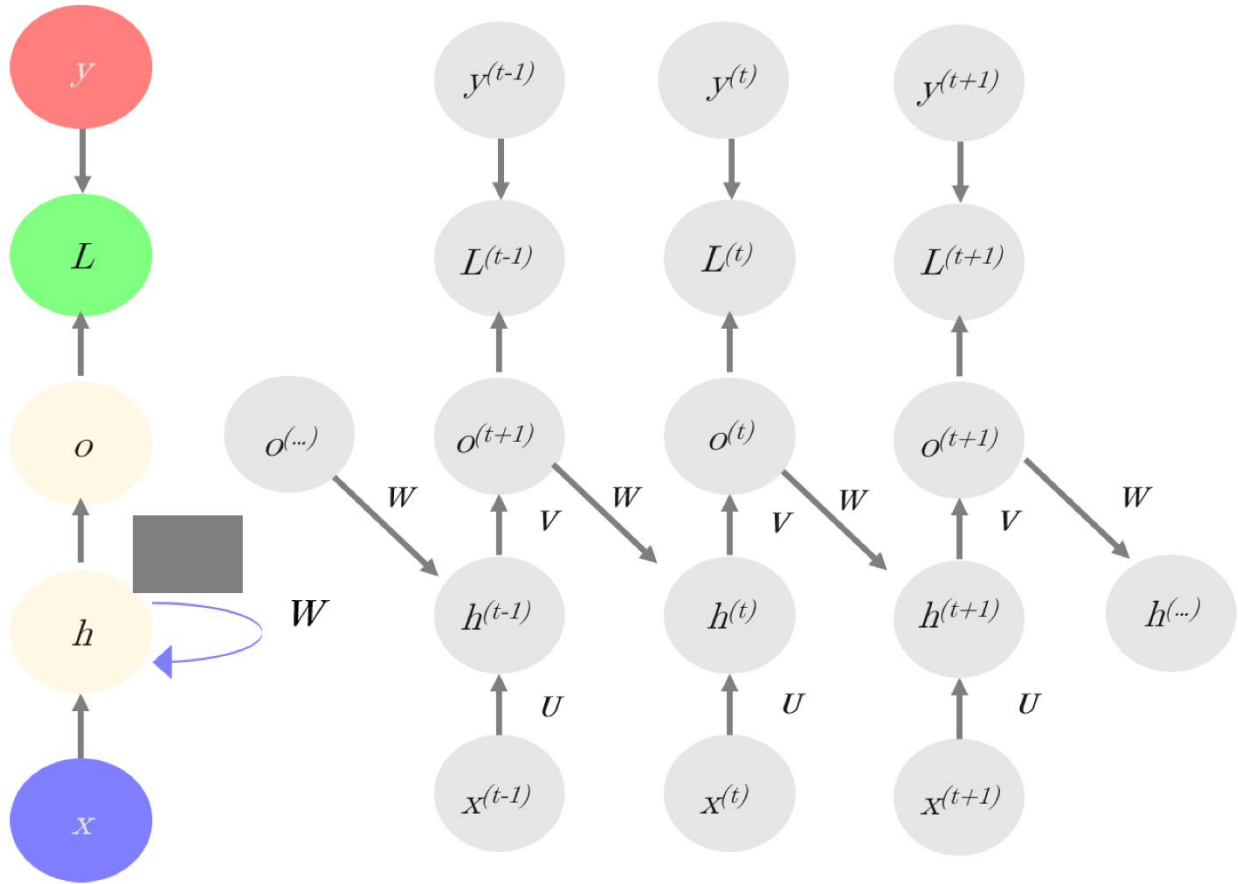
για την πρόβλεψη ενός φαινοτύπου. Αυτό το μοντέλο ενσωματώνει πληροφορίες από πολλά SNP και βασίζεται στο μέγεθος του φίλτρου σε κάθε επίπεδο στον αριθμό των SNP που θα συγχωνευθούν.



Εικόνα 9 : Μονοδιάστατο ΣΥνελκτικό Δίκτυο

## Αναδραστικό δίκτυο

Για διαδοχικά δεδομένα ή δεδομένα χρονοσειρών, μπορεί να χρησιμοποιηθεί το Αναδραστικό νευρωνικό δίκτυο (RNN). Αυτού του τύπου τα δίκτυα χρησιμοποιούνται ευρέως, όπως μετάφραση γλώσσας, επεξεργασία φυσικής γλώσσας (NLP), αναγνώριση φωνής, κλπ. Αναγνωρίζουν τον εαυτό τους από τη "μνήμη" τους επειδή λαμβάνουν δεδομένα από προηγούμενες εισόδους για να επηρεάσουν την τρέχουσα είσοδο και έξοδο. Αν και τα συμβατικά βαθιά νευρωνικά δίκτυα υποθέτουν ότι το ένα το άλλο είναι ανεξάρτητο από εισόδους και εξόδους, η απόδοση των αναδραστικών νευρωνικών δικτύων εξαρτάται από τα προηγούμενα στοιχεία της ακολουθίας.



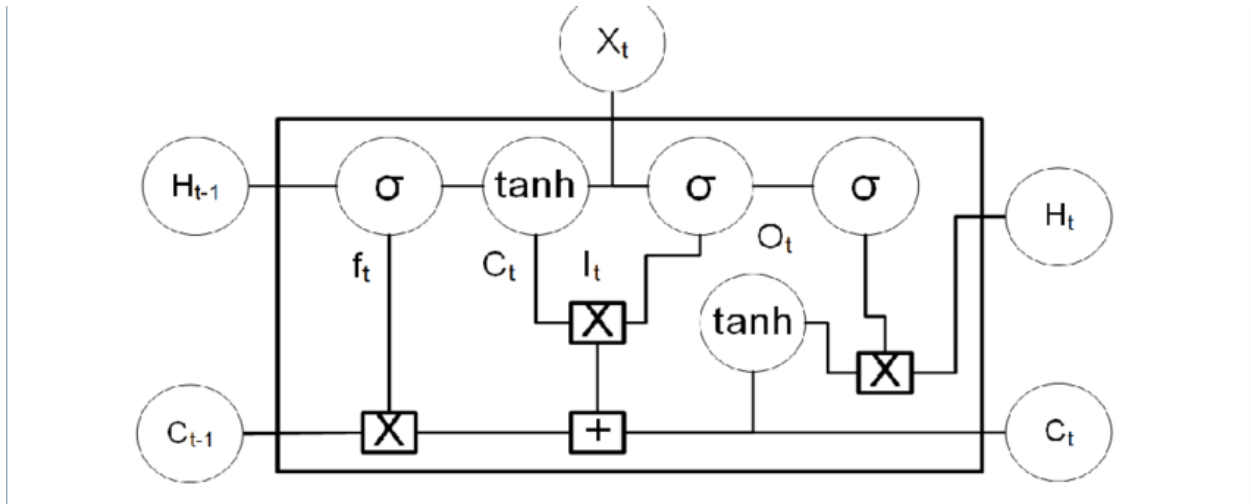
Εικόνα 10 : Αναδραστικό δίκτυο

### Μακροπρόθεσμη βραχυπρόθεσμη μνήμη

Ως αναδραστικό νευρωνικό δίκτυο, μια μακροπρόθεσμη βραχυπρόθεσμη μνήμη (LSTM) έχει παρόμοια ροή ελέγχου. Χειρίζεται πληροφορίες που μεταβιβάζουν δεδομένα καθώς διαδίδονται προς τα εμπρός. Τα γεγονότα είναι τα γεγονότα μέσα στα κύτταρα του LSTM. Τέτοιες λειτουργίες χρησιμοποιούνται για να επιτρέψουν στο LSTM να διατηρήσει ή να ξεχάσει πληροφορίες. Η κατάσταση κυττάρων χρησιμεύει ως η μνήμη ενός δικτύου.

Αρχικά, η κατάσταση των κυττάρων θα διατηρεί σχετικά δεδομένα κατά τη διάρκεια της επεξεργασίας της ακολουθίας. Έτσι, ακόμη και τα δεδομένα από τα προηγούμενα χρονικά βήματα θα επιτρέψουν σε μεταγενέστερα χρονικά βήματα να μειώσουν τον αντίκτυπο της βραχυπρόθεσμης μνήμης. Τα δεδομένα προστίθενται ή αφαιρούνται στην

κατάσταση κελιού μέσω πυλών καθώς η κατάσταση κελιού συνεχίζει το ταξίδι της. Οι πύλες είναι διαφορετικά νευρωνικά δίκτυα που αποφασίζουν ποια γνώση σχετικά με την κυτταρική κατάσταση επιτρέπεται. Οι πύλες θα μάθουν ποια δεδομένα κατά τη διάρκεια της εκπαίδευσης είναι απαραίτητα για να διατηρηθούν ή να ξεχαστούν.

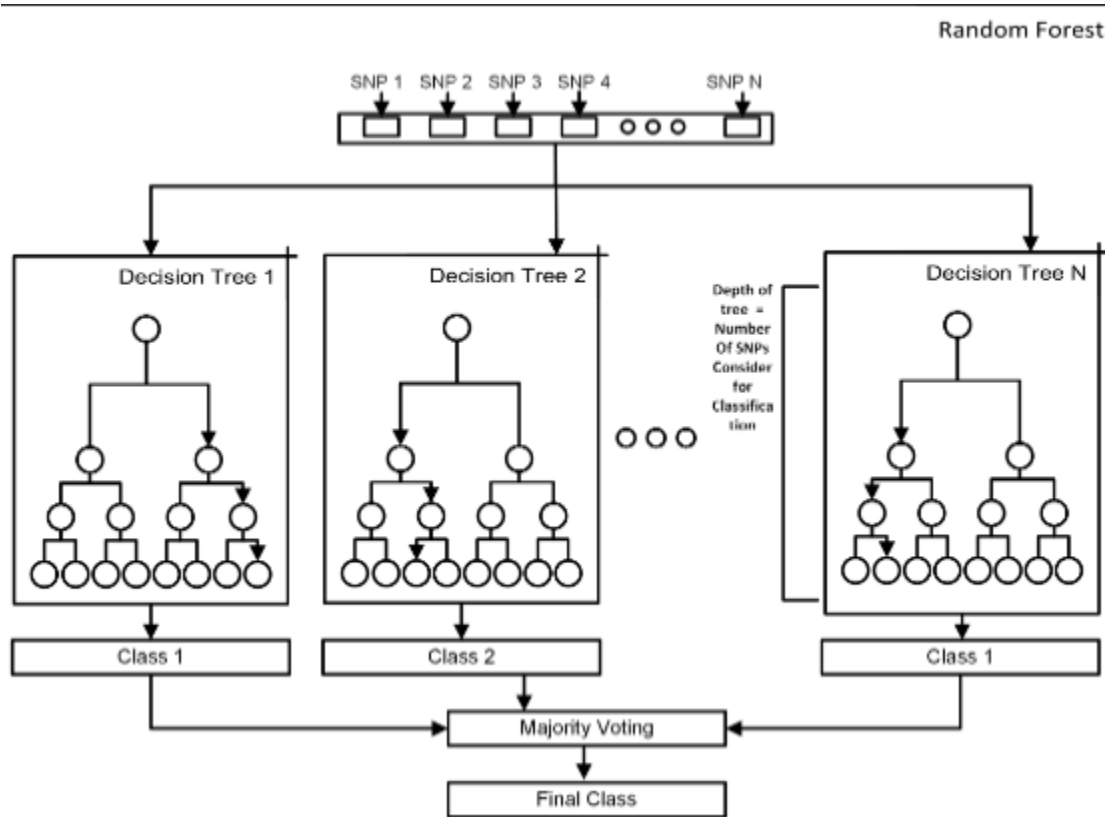


Εικόνα 11 : Μακροπρόθεσμη βραχυπρόθεσμη μνήμη

## Το Τυχαίο Δάσος

Τυχαίο Δάσος είναι ένας συνδυασμός πολλών δέντρων αποφάσεων. Ένα δέντρο αποφάσεων είναι μια τεχνική για τη δημιουργία μοντέλων ταξινόμησης ή παλινδρόμησης. Ονομάζονται δέντρα αποφάσεων αφού πολλά κλαδιά του αν... τότε...» οι διαχωρισμοί αποφάσεων χρησιμοποιούνται για την πρόβλεψη - παρόμοιοι με τους κλάδους ενός δέντρου. Ο πιο συχνός δείκτης για τον προσδιορισμό της καλύτερης διάσπασης είναι η εντροπία Gini και το κέρδος πληροφοριών για εργασίες ταξινόμησης. Το Bagging και Boosting είναι δύο βασικοί τρόποι ενσωμάτωσης των εκροών σε ένα τυχαίο δάσος διαφορετικών δέντρων αποξήρανσης.

Το Random Forest είναι κατάλληλο για δεδομένα γονότυπου. Χρησιμοποιούνται ήδη για προβλέψεις γονότυπου-φαινοτύπου και είναι καλά στο χειρισμό θορυβωδών δεδομένων. Τα SNP που δεν περιέχουν χρήσιμες πληροφορίες απορρίπτονται και η τελική πρόβλεψη βασίζεται μόνο στα χρήσιμα SNP.

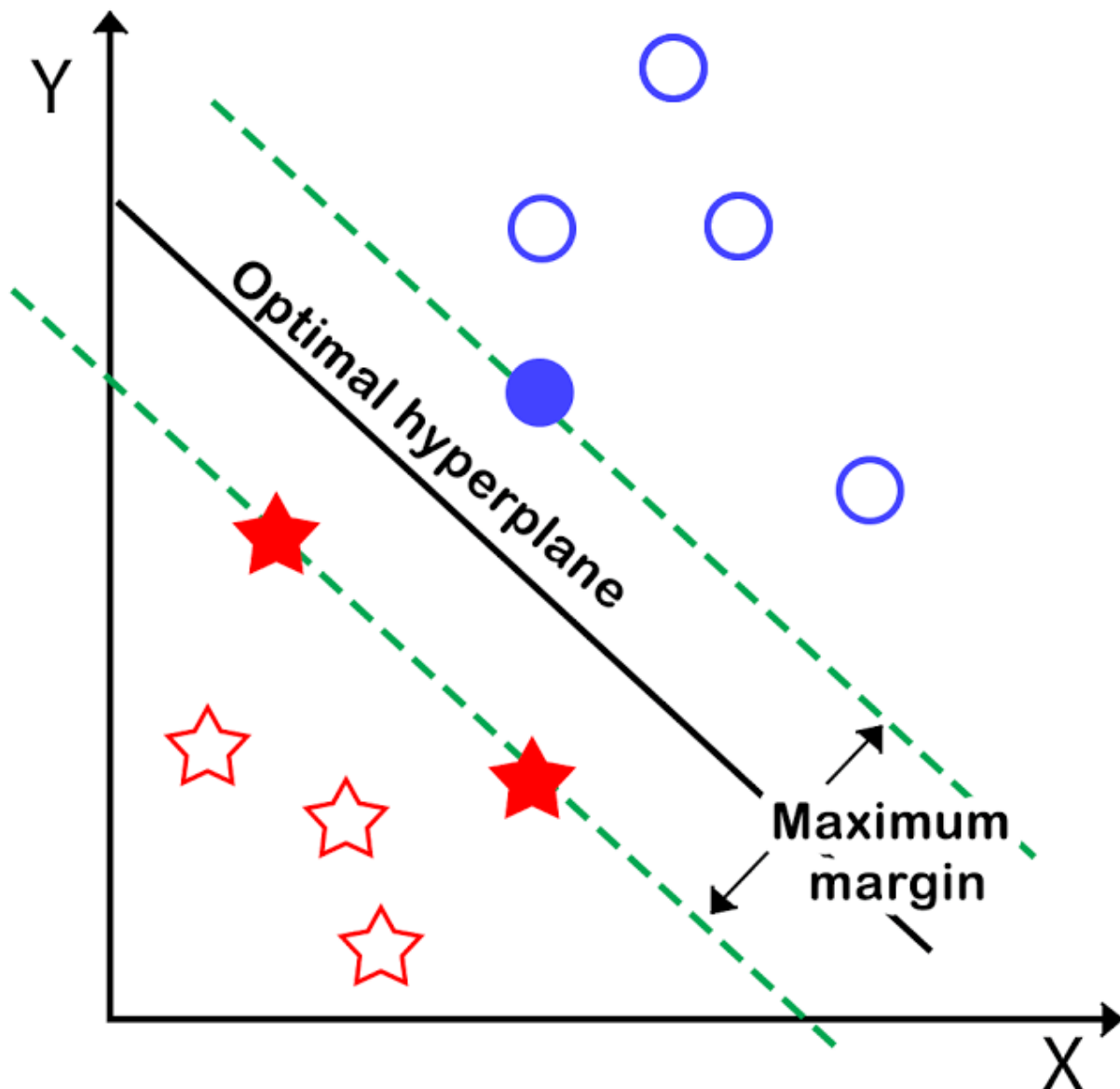


Εικόνα 12: Τυχαίο Δάσος

## Μηχανή διανυσματικής υποστήριξης

Θεωρώντας ένα πρόβλημα δυαδικής ταξινόμησης με θετικά και αρνητικά παραδείγματα, αναζητείται εκείνη η επιφάνεια που καταφέρνει να αποτελέσει το σύνορο εκείνο που διαχωρίζει τα παραδείγματα αυτά. Η μέθοδος SVM επιδιώκει να βρει το σύνορο που απέχει όσο το δυνατόν περισσό-τερο από τα παραδείγματα των κλάσεων που

διαχωρίζει. Το σύνορο αυτό ορίζεται από έναν πεπερασμένο αριθμό παραδειγμάτων του συνόλου εκπαίδευσης που ονομάζονται διανύσματα υποστήριξης (support vectors). Στον χώρο των SNPs, ο SVM επιχειρεί να κατασκευάσει την επιφάνεια εκείνη που θα μπορέσει να διαχωρίσει τα δείγματα σε θετικά και αρνητικά ανάλογα με το χαρακτηριστικό προς εξέταση βασισμένος στα SNPs εκείνα που μπορούν να χαρακτηριστούν ως διανύσματα υποστήριξης.



Εικόνα 13: Μηχανή διανυσματικής υποστήριξης

## Η κατάρα της διαστατικότητας

Η κατάρα της διαστατικότητας περιγράφει την εκρηκτική φύση της αύξησης των διαστάσεων των δεδομένων και την επακόλουθη εκθετική αύξηση των υπολογιστικών προσπαθειών που απαιτούνται για την επεξεργασία ή/και την ανάλυσή τους. Αυτός ο όρος εισήχθη για πρώτη φορά από τον Richard E. Bellman, για να εξηγήσει την αύξηση του όγκου του Ευκλείδειου χώρου που σχετίζεται με την προσθήκη επιπλέον διαστάσεων, στον τομέα του δυναμικού προγραμματισμού. Σήμερα, αυτό το φαινόμενο παρατηρείται σε τομείς όπως η μηχανική μάθηση, η ανάλυση δεδομένων, η εξόρυξη δεδομένων για να αναφέρουμε μερικά. Η αύξηση των διαστάσεων μπορεί θεωρητικά, να προσθέσει περισσότερες πληροφορίες στα δεδομένα βελτιώνοντας έτσι την ποιότητα των δεδομένων, αλλά πρακτικά αυξάνει τον θόρυβο και τον πλεονασμό κατά την ανάλυσή τους.

Στη μηχανική μάθηση, ένα χαρακτηριστικό ενός αντικειμένου μπορεί να είναι ένα χαρακτηριστικό ή ένα χαρακτηριστικό που το ορίζει. Κάθε δυνατότητα αντιπροσωπεύει μια διάσταση και μια ομάδα ιδιοτήτων δημιουργεί ένα σημείο δεδομένων. Αυτό αντιπροσωπεύει ένα διάνυσμα χαρακτηριστικών που ορίζει το σημείο δεδομένων που θα χρησιμοποιηθεί από έναν αλγόριθμο μηχανικής μάθησης. Όταν λέμε αύξηση της διάστασης συνεπάγεται αύξηση του αριθμού των χαρακτηριστικών που χρησιμοποιούνται για την περιγραφή των δεδομένων.

Για παράδειγμα, στον τομέα της έρευνας για τον καρκίνο του μαστού, η ηλικία, ο αριθμός των καρκινικών κόμβων μπορούν να χρησιμοποιηθούν ως χαρακτηριστικά για τον καθορισμό της πρόγνωσης της ασθενούς με καρκίνο του μαστού. Αυτά τα χαρακτηριστικά αποτελούν τις διαστάσεις ενός διανύσματος χαρακτηριστικών. Αλλά άλλοι παράγοντες όπως προηγούμενες χειρουργικές επεμβάσεις, ιστορικό ασθενούς, τύπος όγκου και άλλα τέτοια χαρακτηριστικά βοηθούν έναν γιατρό να καθορίσει καλύτερα την πρόγνωση. Σε αυτήν την περίπτωση προσθέτοντας χαρακτηριστικά, αυξάνουμε θεωρητικά τις διαστάσεις των δεδομένων μας.

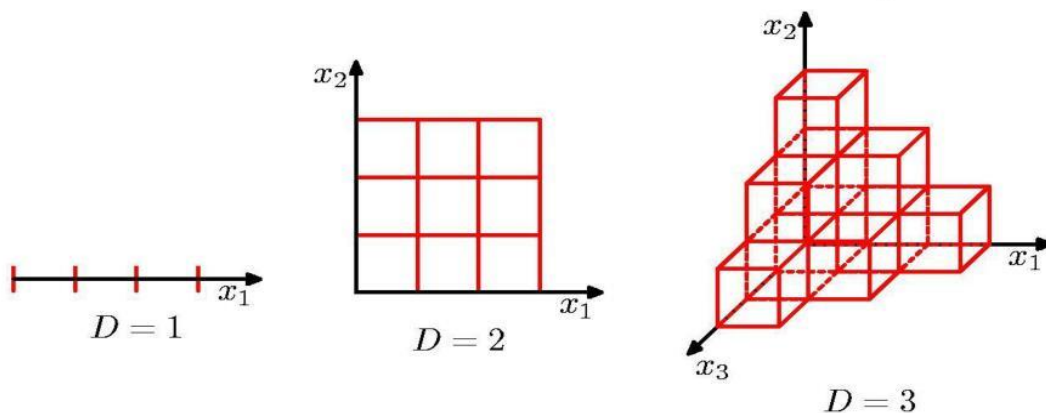
Καθώς η διάσταση αυξάνεται, ο αριθμός των σημείων δεδομένων που απαιτούνται για την καλή απόδοση οποιουδήποτε αλγορίθμου μηχανικής μάθησης αυξάνεται εκθετικά. Ο λόγος είναι ότι, θα χρειαζόμασταν περισσότερο αριθμό σημείων δεδομένων για κάθε



δεδομένο συνδυασμό χαρακτηριστικών, για να είναι έγκυρο οποιοδήποτε μοντέλο μηχανικής μάθησης. Για παράδειγμα, ας υποθέσουμε ότι για να έχει καλή απόδοση ένα μοντέλο, χρειαζόμαστε τουλάχιστον 10 σημεία δεδομένων για κάθε συνδυασμό τιμών δυνατοτήτων. Αν υποθέσουμε ότι έχουμε ένα δυαδικό χαρακτηριστικό, τότε για τις 21 μοναδικές τιμές του (0 και 1) θα χρειαζόμασταν  $2^1 \times 10 = 20$  σημεία δεδομένων. Για 2 δυαδικά χαρακτηριστικά, θα είχαμε μοναδικές τιμές  $2^2$  και θα χρειαζόμασταν  $2^2 \times 10 = 40$  σημεία δεδομένων. Έτσι, για k-αριθμό δυαδικών χαρακτηριστικών θα χρειαζόμασταν  $2^k \times 10$  σημεία δεδομένων.

Ο Hughes (1968) στη μελέτη του κατέληξε στο συμπέρασμα ότι με έναν σταθερό αριθμό δειγμάτων εκπαίδευσης, η προγνωστική δύναμη οποιοδήποτε ταξινομητή αυξάνεται πρώτα καθώς αυξάνεται ο αριθμός των διαστάσεων, αλλά μετά από μια ορισμένη τιμή αριθμού διαστάσεων, η απόδοση επιδεινώνεται. Έτσι, το φαινόμενο της κατάρτας της διάστασης είναι επίσης γνωστό ως φαινόμενο Hughes.

## Curse of Dimensionality



- ▶ No. of cells grow exponentially with D
- ▶ Need exponentially large no. of training data points
- ▶ Not a good approach for more than a few dimensions!

Reference: Christopher M Bishop: Pattern Recognition & Machine Learning, 2006 Springer

Εικόνα 14: Η κατάρτα της διαστατικότητας



## ΜΕΘΟΔΟΛΟΓΙΑ


### OpenSNP

Καθώς το κόστος για γενετικές αναλύσεις συνεχίζει να μειώνεται, οι γενετικές εξετάσεις γίνονται όλο και περισσότερο διαθέσιμες και προσιτές για αυξανόμενο αριθμό ανθρώπων - μια τάση που μπορεί να φανεί στον αυξανόμενο αριθμό πελατών που χρησιμοποιούν υπηρεσίες γενετικών δοκιμών Direct-To-Consumer (DTC) όπως η 23andMe και η AncestryDNA. Η μείωση του κόστους και η αυξημένη διαθεσιμότητα έχουν οδηγήσει στη δημιουργία μίας βάσης γενετικών δεδομένων, όπως το OpenSNP.

Το OpenSNP είναι μοναδικό σε σχέση με διάφορους ανταγωνιστές, διότι προσφέρει ανοιχτή συμμετοχή και ανοικτή πρόσβαση στα δεδομένα: Οι συμμετέχοντες του OpenSNP μπορούν να χρησιμοποιήσουν την πλατφόρμα για να μοιραστούν ανοιχτά τα υπάρχοντα δεδομένα γενετικών δοκιμών DTC, θέτοντας τα δεδομένα τους στον δημόσιο τομέα. Επιπλέον, οι συμμετέχοντες μπορούν να μοιραστούν φαινοτυπικά χαρακτηριστικά, όπως το χρώμα των ματιών, το χρώμα των μαλλιών ή το ύψος. Από την έναρξή της το 2011, πάνω από 5.000 άτομα 11 έχουν χρησιμοποιήσει την πλατφόρμα για να καταστήσουν διαθέσιμα τα γενετικά τους δεδομένα. Διαγωνισμοί ανάλυσης δεδομένων που προέρχονται από το πλήθος έχουν γίνει όλο και πιο δημοφιλείς τα τελευταία χρόνια, επιτρέποντας σε ειδικούς και λάτρεις της επιστήμης δεδομένων να λύσουν συνεργατικά προβλήματα του πραγματικού κόσμου, μέσω διαδικτυακών προκλήσεων.


Αυτή η προσέγγιση επιτρέπει την ευρεία εξερεύνηση του χώρου του μοντέλου σε ένα συγκεκριμένο σύνολο δεδομένων από άτομα με δεξιότητες ανάλυσης δεδομένων που προέρχονται από πολύ διαφορετικά υπόβαθρα. Στο πλαίσιο της γενετικής πρόβλεψης σύνθετων ασθενειών, είναι πρωτοφανής. Ενώ η πιο ευρέως χρησιμοποιούμενη πλατφόρμα, kaggle.com, προσφέρει χρηματικές ανταμοιβές, η crowdai.org είναι πιο ακαδημαϊκή και προσέφερε στον νικητή την ευκαιρία να παρουσιάσει τη δουλειά του σε επιστημονικό συνέδριο.

openSNP News Data Latest Data About  Giorgos Ouzounidis 




**Upload Your Genotyping File**

Upload the genotyping raw-data you got from *23andMe*, *deCODEme* or *FamilyTreeDNA* to the *openSNP* database to share it with other personal-genomics customers and scientists from around the world.




**Enter Your Variations**

Let us, and the other *openSNP* users, know some of your characteristics—like hair or eye color! Or how about some diseases? *Whatever you feel like sharing!*



**Enter a New Phenotype**

Got an idea for a phenotype that has not been asked about yet and that could have genetical roots? *Great, you can add one* to the *openSNP* database.





**All Genotypes** 

Python libraries to parse the provided files:

- [SNPy](#) courtesy of [Sergei Lebedev](#)
- [snps](#) courtesy of [Andrew Riha](#)

[Download all data](#)

Includes all genotyping files, a CSV with all phenotypes of those users, and all picture phenotypes.

User	ID	Created	Type	
 <a href="#">Aylish</a>	9384	09.11.2022 16:54	23andme	<a href="#">Download</a>
 <a href="#">jasmine921</a>	9382	09.11.2022 13:11	23andme	<a href="#">Download</a>
 <a href="#">ashleykrichard</a>	9381	09.11.2022 01:19	ancestry	<a href="#">Download</a>
 <a href="#">marcs78</a>	9380	07.11.2022 23:01	23andme	<a href="#">Download</a>

Εικόνα 15: OpenSNP

Αρχικά, έπειτα από μια σχετική μελέτη των δεδομένων θα επιλεγούν τα πιο δημοφιλή φαινότυπα, ώστε να κατασκευαστούν σεντ δεδομένων που θα περιέχουν όσο το δυνατόν περισσότερα άτομα. Τα περισσότερα άτομα είναι απαραίτητα σε μία τέτοια μελέτη. Το φαινόμενο της κατάρας της διαστατικότητας είναι πολύ εμφανές στην περίπτωση των γενετικών δεδομένων.

Τα γενετικά δεδομένα περιλαμβάνουν μεγάλο πλήθος χαρακτηριστικών. Τα χαρακτηριστικά αυτά, ανάλογα και το σχήμα κωδικοποίησής τους για το οποίο θα αναφερθούμε παρακάτω, μπορεί να είναι και πάνω από 2 εκατομμύρια. Από την άλλη, η φύση των δεδομένων αυτών τα καθιστά ως ευαίσθητα και απρόσιτα για τον καθένα.

Η πρωτοβουλία του OpenSNP είναι αρκετά χρήσιμη, παρόλα αυτά δεν είναι αρκετή ώστε να κατασκευαστεί ένα σετ δεδομένων με τόσα πολλά δείγματα όσα και τα χαρακτηριστικά του ανθρώπινου γονιδιώματος. Στην πραγματικότητα, το να είναι διαθέσιμος ένας τόσο μεγάλος αριθμός δειγμάτων φαντάζει αδύνατο στην σημερινή εποχή.

Η τράπεζα δεδομένων UK Biobank είναι μία περίπτωση με 500000 συνολικά δείγματα, η οποία όμως απαιτεί πληρωμή για την πρόσβαση σε αυτήν. Επίσης, κατά τη διαδικασία κατασκευής ενός σετ δεδομένων συσχετισμένων με ένα φαινότυπο, τα δείγματα που μοιράζονται αυτό τον συγκεκριμένο φαινότυπο θα είναι αρκετά λιγότερα. Χαρακτηριστικό παράδειγμα το μελάνωμα του δέρματος, όπου για 500000 κόσμο που συμμετέχει στο UK Biobank, υπάρχουν 3000 δείγματα ανθρώπων που πάσχουν και μπορούν να χρησιμοποιηθούν ως δεδομένα για την συγκεκριμένη ασθένεια. Έτσι λοιπόν, στόχος είναι να επιλεγθούν τα φαινότυπα εκείνα που θα τα μοιράζονται τα περισσότερα δείγματα από την βάση δεδομένων OpenSNP.

### Συλλογή και επεξεργασία δεδομένων

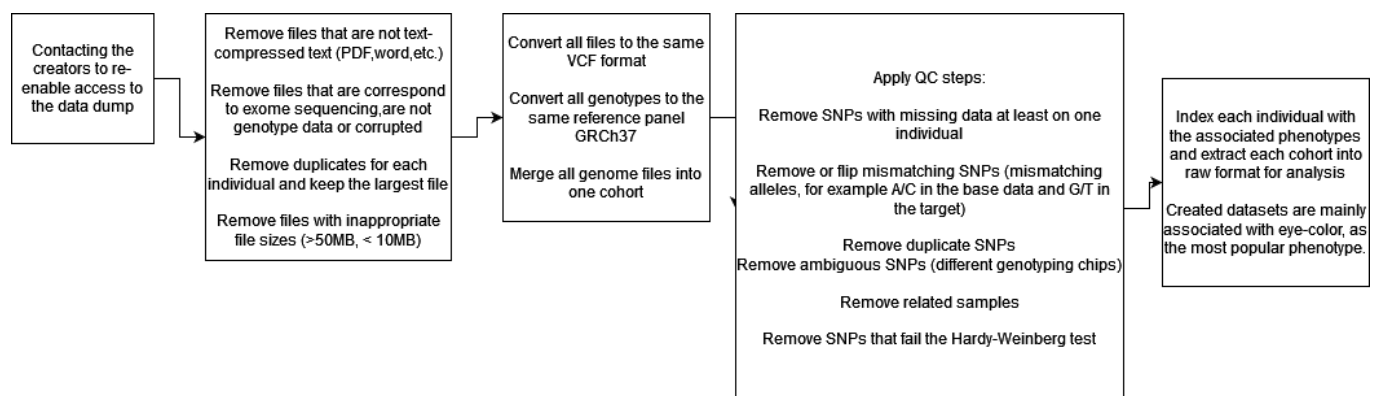
Προκειμένου να συλλεχθούν τα δεδομένα, πραγματοποιήθηκε η μεταφόρτωση του αρχείου που περιλαμβάνει το σύνολο των γενετικών δεδομένων όλων των χρηστών του ιστοτόπου. Οι χρήστες που έχουν επιλέξει να μοιραστούν τη γενετική τους πληροφορία αλλά και τα φαινοτυπικά χαρακτηριστικά τους με την πύλη OpenSNP, είναι 6420. Στη συνέχεια πραγματοποιούνται διάφορα βήματα ελέγχου ποιότητας ώστε να διασφαλιστεί ότι τα δεδομένα θα είναι κατάλληλα προς επεξεργασία από τους διάφορους αλγορίθμους.

Αφού αποσυμπιεστεί το αρχείο, εφαρμόζεται το πρώτο κριτήριο διαλογής που είναι η αφαίρεση των αρχείων των γενοτύπων που το μέγεθος αυτών δεν βρίσκεται ανάμεσα σε ορισμένα τυπικά όρια. Τα αρχεία των γενοτύπων που υπερβαίνουν τα 50mb ή είναι μικρότερα από 10mb έχουν τεράστια πιθανότητα να αποτελούν προβληματικά αρχεία.

Στη συνέχεια, αφαιρούνται όλα εκείνα τα αρχεία που δεν είναι κείμενο και συνεπώς μη κατάλληλης μορφής ή άσχετα για την συγκεκριμένη ανάλυση.

Έχοντας απομακρύνει τα μη κατάλληλα αρχεία, πραγματοποιείται μία ταξινόμηση των αρχείων-χρηστών ώστε να προσδιοριστεί αν κάποιος χρήστης έχει ανεβάσει αρχείο γενοτύπου παραπάνω από μία φορά. Στην περίπτωση αυτή, κρατιέται μόνο το μεγαλύτερο σε μέγεθος αρχείο και όλα τα άλλα αρχεία που αναφέρονται στον ίδιο χρήστη απομακρύνονται. Έτσι αποφεύγονται τα διπλότυπα στην ανάλυση. Επιπρόσθετα, μετατρέπονται όλα τα αρχεία στο κατάλληλο build reference, ώστε το κάθε SNP να αναφέρεται στην ίδια τοποθεσία εντός του γονιδιώματος. Τα γενετικά δεδομένα με συντεταγμένες που βασίζονται στο NCBI36 αναβαθμίζονται ώστε να ταιριάζουν με την αναφορά GRCh37 39 με το liftOver. Το PLINK χρησιμοποιείται για τη μετατροπή των μορφών των αρχείων. Τα VCFtools χρησιμοποιούνται για την ταξινόμηση των παραλλαγών (αλληλόμορφα). Το BCFtools χρησιμοποιείται για την αναφορά των αλληλόμορφων στο σύστημα αναφοράς GRCh37, για την μετονομασία δειγμάτων και την τελική συγχώνευση όλων των ατόμων σε ένα αρχείο.

Τέλος αφαιρούνται από το τελικό αρχείο τα χαρακτηριστικά εκείνα που εμφανίζουν τιμές μεγαλύτερες από το  $1e-10$  στο Hardy-Weinberg equilibrium test. Αφαιρούνται επίσης τα δείγματα εκείνα που προβλέπεται πως διαθέτουν οικογενειακό συσχετισμό μεταξύ τους. Σαν τελευταίο βήμα της προεπεξεργασίας αυτής είναι η αφαίρεση κάθε χαρακτηριστικού εκείνου (SNP) που απουσιάζει πληροφορία σε τουλάχιστον ένα δείγμα από το κάθε σετ δεδομένων που θα κατασκευαστούν. Συνεπώς από τα 6420 άτομα-δείγματα, τελικώς απομένουν τα 4666 από τις διάφορες επεξεργασίες και αφαιρούνται άλλα 3 λόγω οικογενειακής συσχέτισης.



Εικόνα 16: Βήματα συλλογής δεδομένων

Σχετικά με τα χαρακτηριστικά, ο ακριβής αριθμός τους ποικίλει ανάλογα το σετ δεδομένων. Γενικότερα, προσεγγιστικά μπορούμε να πούμε ότι αφαιρέθηκαν περίπου 3591475 χαρακτηριστικά λόγω απουσίας σε μεμονωμένα δείγματα και 18468 λόγω του Hardy-Weinberg τεστ. Οι διάφορες τεχνικές imputation των δεδομένων ώστε να συμπληρωθούν οι τιμές που απουσιάζουν αποτελεί ένα σπουδαίο ξεχωριστό κεφάλαιο στις εφαρμογές αυτές, με αποτέλεσμα να μην ασχοληθούμε με το κομμάτι αυτό αφού τα χαρακτηριστικά ήδη υπερβαίνουν κατά πολύ τον αριθμό των δειγμάτων. Οπότε απομένουν περίπου 39162 χαρακτηριστικά και 4663 άνθρωποι-δείγματα. Τα σετ δεδομένων που κατασκευάστηκαν περιλαμβάνουν κατηγορίες που σχετίζονται με το χρώμα των ματιών και τον διαβήτη τύπου Β.

### Σετ δεδομένων

Τα σετ δεδομένων που θα χρησιμοποιηθούν επιλέχθηκαν με τέτοιο τρόπο έτσι ώστε να υπάρχουν όσο το δυνατόν περισσότερα δείγματα. Η πιο συχνά συμπληρωμένη κατηγορία από τους χρήστες της βάσης OpenSNP ήταν η κατηγορία του χρώματος των ματιών. Το χρώμα των ματιών είναι ένα χαρακτηριστικό το οποίο είναι αρκετά εύκολο να συμπληρωθεί από τους χρήστες και αποτελεί μία σχετικά μη ευαίσθητη πληροφορία, γεγονός που επιτρέπει σε αυτούς να την μοιραστούν εύκολα και χωρίς δισταγμό. Έτσι, προκειμένου να παραχθούν σετ δεδομένων που να περιέχουν αρκετά δείγματα, σχηματίστηκαν σετ δεδομένων με βάση αυτό το χαρακτηριστικό. Τα σετ δεδομένων κατασκευάζονται ως δυαδικά σετ για το πρόβλημα της ταξινόμησης.

Ακόμα, θα χρησιμοποιηθεί και ο διαβήτης τύπου Β ως χαρακτηριστικό για ένα επιπλέον σετ δεδομένων, το οποίο όμως δεν θα χρησιμοποιηθεί σε όλα τα πειράματα καθώς οι ακρίβεια που προκύπτει είναι απόλυτη τις περισσότερες φορές και δεν προσφέρει κάποια χρήσιμη πληροφορία. Αυτό συμβαίνει γιατί τα δείγματα είναι αρκετά λίγα και η ταξινόμηση των δειγμάτων δεν αποτελεί πρόκληση.

Eye colour		Number of samples		Dataset name
Hazel	Brown	128	331	1
Blue	Brown	180	331	2
Hazel	Blue	128	180	3
Brown	Bluegrey	331	97	4
Brown	Bluegreen	331	75	5
Diabetes type		Number of samples		Dataset name
Diabetic	Non diabetic	15	114	6

Εικόνα 17: Σετ δεδομένων

### Σετ συνθετικών δεδομένων

Γνωρίζοντας ότι τα δεδομένα από την πύλη OpenSNP είναι πραγματικά αλλά μικρά σε μέγεθος, προκύπτει η ανάγκη ύπαρξης ενός μεγαλύτερου σετ δεδομένων. Βασικός λόγος γι' αυτή την ανάγκη είναι η αντικειμενική συγκριτική ανάλυση των διάφορων μεθόδων και τεχνικών που θα δοκιμαστούν για την πρόβλεψη του φαινοτύπου μέσα από τα γενετικά δεδομένα. Ο μικρός αριθμός δειγμάτων σε σχέση με τον αριθμό των χαρακτηριστικών μπορεί να δημιουργήσει πρόβλημα στην λειτουργία των μοντέλων. Οπότε, κατασκευάζεται ένα νέο σετ συνθετικών δεδομένων, που προσομοιώνει το γενότυπο και το φαινότυπο 10000 ατόμων και τα SNPs επιλέγονται με βάση στατιστικό τεστ.

Πιο συγκεκριμένα, όπως θα αναφερθεί και στην ενότητα με τα πειράματα, προκειμένου να υπάρξει μία αντικειμενική σύγκριση μεταξύ τεχνικών μηχανικής μάθησης και PRS θα πρέπει τα δείγματα να είναι αρκετά και σίγουρα μεγαλύτερα από τον αριθμό των δειγμάτων που συλλέχθηκαν από την πύλη openSNP.

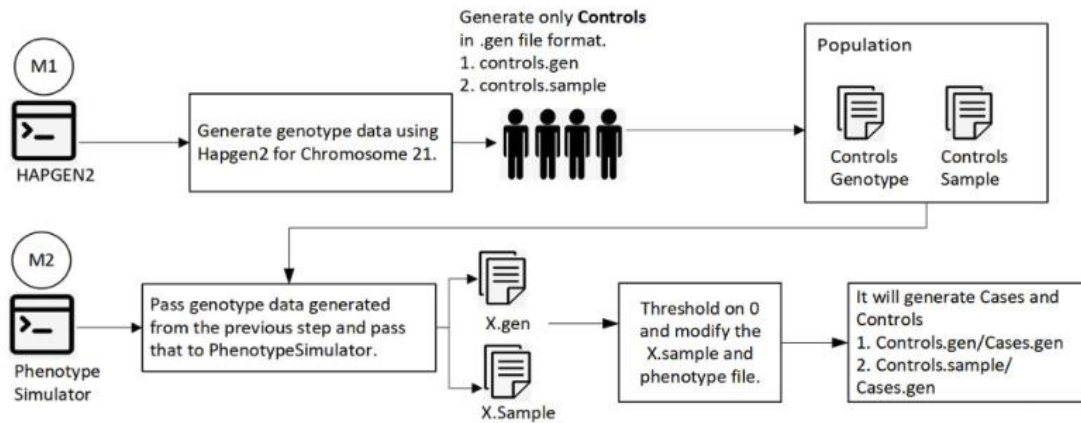
## Κατασκευή συνθετικών δεδομένων

Τα συνθετικά δεδομένα κατασκευάστηκαν με την βοήθεια των PhenotypeSimulator και hargen2. Χρησιμοποιώντας το hargen 2 , κατασκευάστηκαν δεδομένα γονότυπου σχετικά με το χρωμόσωμα 21 για 10000 άτομα. Έπειτα, ο πληθυσμός αυτόν πέρασε στον PhenotypeSimulator για να εξωμοιωθεί το φαινότυπο ( ύπαρξη ή μη ύπαρξη) . Το φαινότυπο αυτό που παράγεται είναι σε συνεχή μορφή, και μετατρέπεται σε ύπαρξη ή μη ύπαρξη ανάλογα με το πρόσημο του.

Το HAPGEN2 είναι μια ενημερωμένη έκδοση του προγράμματος HAPGEN, η οποία προσομοιώνει σύνολα δεδομένων ελέγχου περιπτώσεων( ύπαρξη ή μη ύπαρξη ασθένειας) σε δείκτες SNP. Η νέα έκδοση μπορεί τώρα να προσομοιώσει πολλαπλές ασθένειες SNPs σε ένα μόνο χρωμόσωμα, με την υπόθεση ότι κάθε ασθένεια SNP δρα ανεξάρτητα και τα SNPs βρίσκονται σε ισορροπία Hardy-Weinberg. Η υποκείμενη προσέγγιση προσομοίωσης μπορεί να χειριστεί δείκτες σε ανισορροπία σύνδεσης (LD) και να προσομοιώσει σύνολα δεδομένων σε μεγάλες περιοχές όπως ολόκληρα χρωμοσώματα.

Το PhenotypeSimulator επιτρέπει την προσομοίωση σύνθετων φαινοτύπων υπό διαφορετικά μοντέλα, συμπεριλαμβανομένων των επιδράσεων γενετικών παραλλαγών και των απειροελάχιστων γενετικών επιδράσεων (αντανακλώντας τη δομή του πληθυσμού) καθώς και σχετιζόμενων, μη γενετικών συμμεταβλητών και επιδράσεων θορύβου. Διαφορετικές επιδράσεις μπορούν να συνδυαστούν σε έναν τελικό φαινότυπο, ενώ ελέγχονται για το ποσοστό της διακύμανσης που εξηγείται από κάθε ένα από τα συστατικά. Για κάθε στοιχείο, ο αριθμός των μεταβλητών, η κατανομή τους και ο σχεδιασμός της επίδρασής τους μεταξύ των χαρακτηριστικών μπορούν να προσαρμοστούν.





Εικόνα 18: Διαδικασία κατασκευής συνθετικών δεδομένων

## Κωδικοποίηση δεδομένων

Κωδικοποίηση δεδομένων Τα δεδομένα SNP μπορούν να αναπαρασταθούν ως ονομαστικά χαρακτηριστικά, π.χ. AA, AG ή GG, ή αριθμητικά, π.χ. 0, 1 και 2. Ενώ ορισμένοι αλγόριθμοι ταξινόμησης μπορούν να λειτουργήσουν με κατηγορικά χαρακτηριστικά, όπως το Τυχαίο Δάσος, σχεδόν όλοι μπορούν να λειτουργήσουν σε αριθμητικά χαρακτηριστικά και μερικοί μπορούν να λειτουργήσουν μόνο σε αυτά, όπως η Μηχανή Διανύσματος Υποστήριξης ή το Τεχνητό νευρωνικό δίκτυο. Αυτό καθιστά απαραίτητη την κωδικοποίηση των SNP ως αριθμητικά χαρακτηριστικά. Υπάρχουν διάφοροι τρόποι κωδικοποίησης των SNP και κάθε κωδικοποίηση μπορεί να αντιπροσωπεύει διαφορετικές βιολογικές υποθέσεις. Η κωδικοποίηση μπορεί επίσης να επηρεάσει την ικανότητα των αλγορίθμων μηχανικής μάθησης κατά την εκπαίδευση του μοντέλου.

Στο προσθετικό μοντέλο (additive model), κάθε γονότυπος κωδικοποιείται ως ένα ενιαίο αριθμητικό χαρακτηριστικό που αντικατοπτρίζει τον αριθμό των εναλλακτικών αλληλόμορφων. Τα ομόζυγα επικρατή, τα ετερόζυγα και τα ομόζυγα υπολειπόμενα κωδικοποιούνται ως 0, 1 και 2, αντίστοιχα. Αυτό έχει ως αποτέλεσμα έναν ελάχιστο αριθμό χαρακτηριστικών, διατηρώντας παράλληλα όλες τις πληροφορίες. Από την άλλη, οι μη προσθετικές επιδράσεις όπως η ετεροζυγωτικότητα που οδηγούν σε υψηλότερο κίνδυνο νόσου από την ομοζυγωτικότητα δεν μπορούν να μοντελοποιηθούν.

Στο υπολειπόμενο/κυρίαρχο μοντέλο(rec/dom model), κάθε γονότυπος κωδικοποιείται ως δύο δυαδικά χαρακτηριστικά, ένα για κάθε πιθανό αλληλόμορφο. Ένα χαρακτηριστικό ορίζεται σε 0 εάν το αντίστοιχο αλληλόμορφο δεν υπάρχει και ορίζεται σε 1 εάν υπάρχει τουλάχιστον μία φορά. Αυτό διπλασιάζει τον αριθμό των χαρακτηριστικών, γεγονός που απαιτεί περισσότερη μνήμη και μπορεί να αυξήσει το υπολογιστικό κόστος. Ωστόσο, αυτή η κωδικοποίηση μπορεί να είναι ανώτερη από την πρόσθετη κωδικοποίηση κατά τη μοντελοποίηση των αλληλεπιδράσεων μεταξύ των SNPs.

Ένας άλλος δυαδικός τρόπος κωδικοποίησης των πληροφοριών γονότυπου είναι η δημιουργία τριών χαρακτηριστικών για κάθε SNP, ένα για κάθε γονότυπο, το οποίο έχει επίσης αποδειχθεί χρήσιμο για την ανίχνευση αλληλεπιδράσεων. Σε αυτή την κωδικοποίηση, κάθε χαρακτηριστικό αντιπροσωπεύει εάν η αντίστοιχη μετάλλαξη είναι παρών ή όχι, πράγμα που σημαίνει ότι ακριβώς ένα από τα τρία χαρακτηριστικά είναι 1 και τα άλλα δύο είναι 0. Ενώ αυτό το σχήμα ονομάζεται επίσης κωδικοποίηση one hot(genotypic model), δηλώνουμε αυτό το σχήμα ως γονοτυπικό. Δεδομένου ότι κάθε γονότυπος αντιπροσωπεύεται από ένα ξεχωριστό χαρακτηριστικό, οι ταξινομητές μπορούν να δημιουργήσουν πιο λεπτομερή μοντέλα. Το μειονέκτημα είναι μια ακόμη υψηλότερη κατανάλωση μνήμης λόγω του πλεονασμού αυτής της αναπαράστασης χαρακτηριστικών.

SNP <sub>i</sub>	Add count	Rec		Gen		
		A	B	AA	AB	BB
AA	0	1	0	1	0	0
AB	1	1	1	0	1	0
BB	2	0	1	0	0	1

Εικόνα 19: Κωδικοποίηση δεδομένων

## Επιλογή των SNPs και σημαντικά SNPs

Τα SNPs είναι οι σημαντικότεροι δείκτες που χρησιμοποιούνται για τη γενετική χαρτογράφηση ασθενειών. Αν και τα περισσότερα από αυτά είναι ουδέτερα, πρόσφατες μελέτες έχουν δείξει ότι ορισμένα SNPs είναι λειτουργικά και επηρεάζουν τον φαινότυπο, π.χ. ύψος, χρώμα δέρματος, αντίσταση, λοίμωξη ή αποκρίσεις σε φάρμακα κ.λ.π.

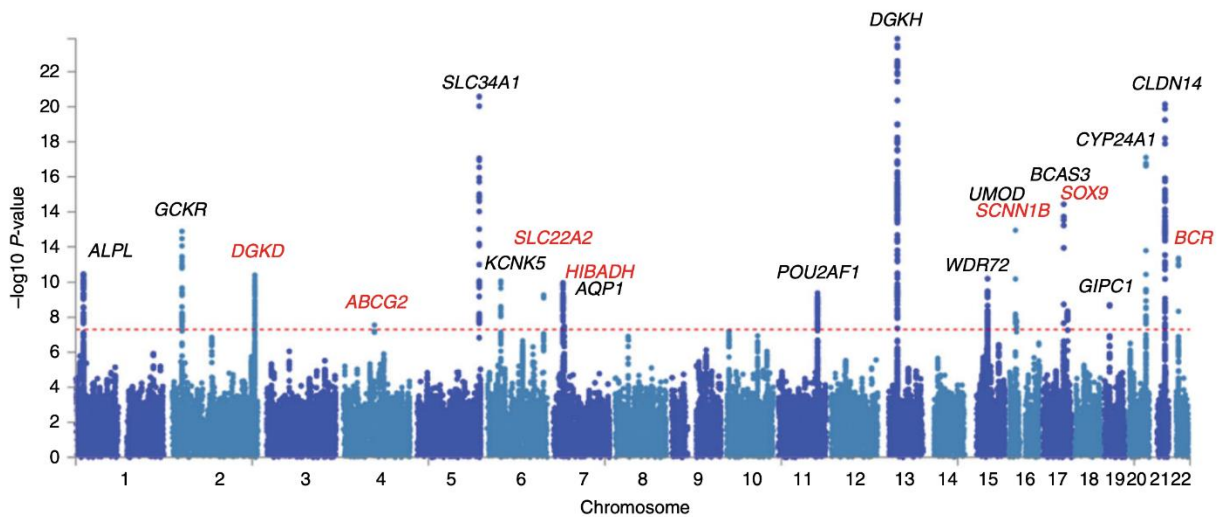
Σε αυτό το πλαίσιο, πολλοί αλγόριθμοι μηχανικής μάθησης έχουν εφαρμοστεί ευρέως για την ταξινόμηση δεδομένων SNP. Ωστόσο, η "κατάρα της διάστατικότητας" είναι η κύρια πρόκληση που συναντάται, στις περισσότερες μελέτες, λόγω του ότι ο αριθμός των δειγμάτων (μερικές εκατοντάδες) είναι σημαντικά μικρότερος από τον αριθμό των SNP (έως και μερικά εκατομμύρια) . Η δημιουργία ενός μοντέλου για την ταξινόμηση των δειγμάτων ως ανήκοντα σε ένα υγιές ή επηρεαζόμενο άτομο είναι ένας από τους κύριους στόχους της ανάλυσης των SNPs . Ωστόσο, ο τεράστιος αριθμός SNPs εμποδίζει την ανάπτυξη ακριβών αλγορίθμων πρόβλεψης.

Η επιλογή ενός υποσυνόλου περιγραφικών και ουσιαστικών SNP είναι ζωτικής σημασίας για τη μείωση της χρονικής πολυπλοκότητας και για την αύξηση της ακρίβειας. Ως αποτέλεσμα, το αρχικό στάδιο της ανάλυσης δεδομένων SNP θα πρέπει να είναι η επιλογή του πιο διακριτικού και ενημερωτικού υποσυνόλου των SNP, προκειμένου να βελτιωθεί η απόδοση του αλγορίθμου ταξινόμησης και να μειωθούν οι χρονικές και υπολογιστικές απαιτήσεις . Για το σκοπό αυτό έχουν χρησιμοποιηθεί πολλές μέθοδοι επιλογής χαρακτηριστικών.

Η επιλογή μιας κατάλληλης μεθόδου επιλογής χαρακτηριστικών είναι και αυτή πολύ βασική για την επιτυχία μοντέλου που βασίζεται στη μηχανική μάθηση αλλά και στις συμβατικές μεθόδους PRS. Η επιλογή χαρακτηριστικών είναι η διαδικασία μείωσης της διάστασης του χώρου χαρακτηριστικών, διατηρώντας παράλληλα μια ακριβή αναπαράσταση των αρχικών δεδομένων. Τα κύρια πλεονεκτήματά της είναι η βελτιωμένη απόδοση ταξινόμησης, οι μειωμένες ταχύτητες μάθησης, η διευκόλυνση της ερμηνείας δεδομένων και η βελτιωμένη ικανότητα γενίκευσης των προβλέψεων.

Επίσης η επιλογή των χαρακτηριστικών πέρα από ένα μέσο μείωσης των διαστάσεων και της υπολογιστικής δύναμης που απαιτείται στην ταξινόμηση, μπορεί να χρησιμοποιηθεί και σαν ένα μέσο ανάλυσης των SNPs σχετικά με την εξεταζόμενη ασθένεια ή

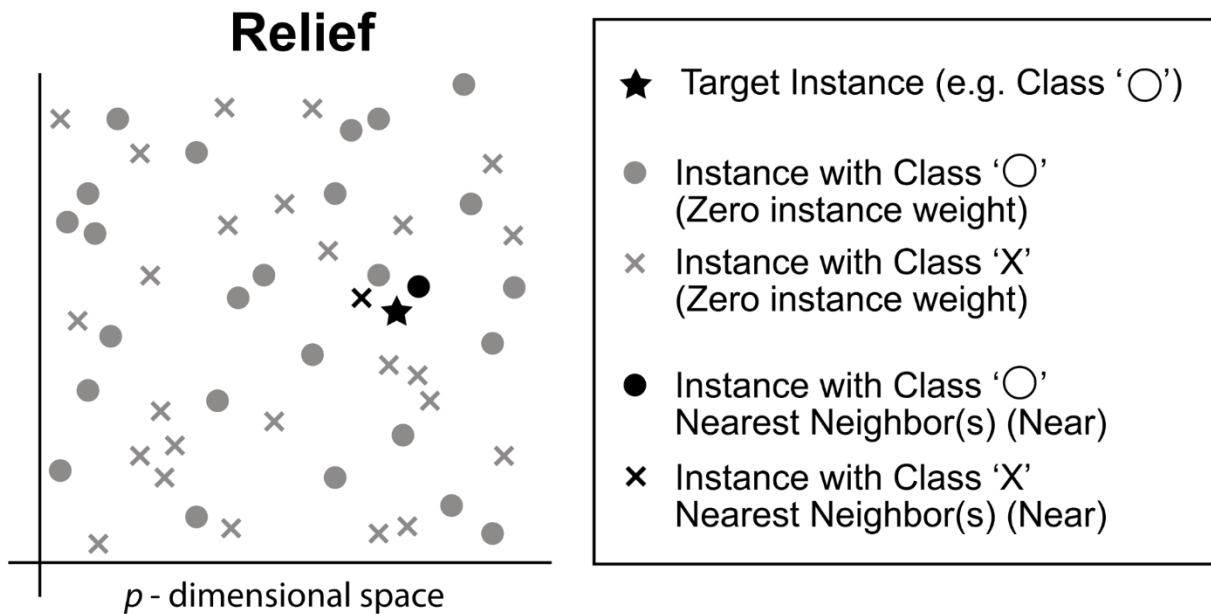
χαρακτηριστικό. Συγκεκριμένα, μπορεί να μελετηθεί το ποια SNPs σχετίζονται με την κάθε ασθένεια και επηρεάζουν σε μεγάλο βαθμό την ταξινόμηση. Έτσι, γνωρίζοντας τους κρίσιμους γενετικούς δείκτες μπορούν να γίνουν πιο άμεσα και εύκολα προβλέψεις σχετικά με χαρακτηριστικά ή πιθανές ασθένειες. Υπάρχουν διαθέσιμα πολλά αποτελέσματα από διάφορες αναλύσεις που αναδεικνύουν τα σχετιζόμενα SNPs με την κάθε ασθένεια.



Εικόνα 20: Επιλογή SNPs – GWAS

Στην συνέχεια των πειραμάτων θα χρησιμοποιηθεί ο αλγόριθμος Relief. Ο Relief είναι ένας αλγόριθμος που αναπτύχθηκε από την Kira και τον Rendell το 1992 και ακολουθεί μια προσέγγιση φίλτρου-μεθόδου για την επιλογή χαρακτηριστικών που είναι ιδιαίτερα ευαίσθητη στις αλληλεπιδράσεις χαρακτηριστικών. Αρχικά σχεδιάστηκε για εφαρμογή σε προβλήματα δυαδικής ταξινόμησης με διακριτά ή αριθμητικά χαρακτηριστικά. Το Relief υπολογίζει μια βαθμολογία χαρακτηριστικών για κάθε χαρακτηριστικό, η οποία μπορεί στη συνέχεια να εφαρμοστεί για την κατάταξη και την επιλογή κορυφαίων χαρακτηριστικών βαθμολόγησης για την επιλογή χαρακτηριστικών. Εναλλακτικά, αυτές οι βαθμολογίες μπορούν να εφαρμοστούν ως συντελεστές στάθμισης χαρακτηριστικών για την καθοδήγηση της μοντελοποίησης. Η βαθμολόγηση χαρακτηριστικών ανακούφισης βασίζεται στον προσδιορισμό των διαφορών τιμής χαρακτηριστικών μεταξύ ζευγών παρουσίας πλησιέστερου γείτονα. Εάν παρατηρηθεί διαφορά τιμής χαρακτηριστικού σε γειτονικό ζεύγος παρουσιών με την ίδια κλάση (hit), η βαθμολογία

της δυνατότητας μειώνεται. Εναλλακτικά, εάν παρατηρηθεί διαφορά τιμής χαρακτηριστικού σε γειτονικό ζεύγος παρουσιών με διαφορετικές τιμές κλάσης ('miss'), η βαθμολογία της δυνατότητας αυξάνεται.



Εικόνα 21: Relief

Έπειτα θα χρησιμοποιηθεί και θα συγκριθεί μία μέθοδος που βασίζεται στο υπολειπόμενο/κυρίαρχο γενετικό μοντέλο, η μέθοδος μεγιστοποίησης γενετικής διαφοράς (GMM). Η μέθοδος αυτή, αναλύοντας τον πληθυσμό, προσπαθεί να διαλέξει εκείνα τα SNPs τα οποία παρέχουν την μεγαλύτερη γενετική πληροφορία που θα μπορούσε να ταξινομήσει τον συγκεκριμένο πληθυσμό με βάση το φαινότυπο του. Πιο συγκεκριμένα, μεταξύ του πληθυσμού που έχει το φαινότυπο και του πληθυσμού που δεν έχει το φαινότυπο θέλουμε να κρατήσουμε τα SNPs που παρουσιάζουν την μεγαλύτερη διαφορά σε μεταλλάξεις. Δηλαδή αν σε ένα συγκεκριμένο SNP οι μεταλλάξεις είναι πολλές στον πληθυσμό που έχει το φαινότυπο ενώ στον συμπληρωματικό πληθυσμό είναι λίγες, σημαίνει ότι αυτό το SNP μπορεί να είναι σημαντικό.

Στους παρακάτω τύπους ως πλήρης μετάλλαξη θεωρείται η ύπαρξη των δυο υπολειπόμενων αλληλόμορφων, ως μερική μετάλλαξη η ύπαρξη ενός υπολειπόμενου

αλληλόμορφου και ως μη μετάλλαξη η ύπαρξη των δυο επικρατών αλληλόμορφων στο SNP. Όλες οι τιμές αναφέρονται σε ποσοστά στο δείγμα μελέτης.

$$\begin{aligned} \mathit{maxFullMutation} &= \mathit{max}(\mathit{phenotype}_{AFM}, \mathit{phenotype}_{BFM}) \\ \mathit{minFullMutation} &= \mathit{min}(\mathit{phenotype}_{AFM}, \mathit{phenotype}_{BFM}) \\ \mathit{maxPartialMutation} &= \mathit{max}(\mathit{phenotype}_{APM}, \mathit{phenotype}_{BPM}) \\ \mathit{minPartialMutation} &= \mathit{min}(\mathit{phenotype}_{APM}, \mathit{phenotype}_{BPM}) \\ \mathit{maxNoMutation} &= \mathit{max}(\mathit{phenotype}_{ANM}, \mathit{phenotype}_{BNM}) \\ \mathit{minNoMutation} &= \mathit{min}(\mathit{phenotype}_{ANM}, \mathit{phenotype}_{BNM}) \end{aligned}$$

Εικόνα 22 GMM

Έπειτα, μέσω μιας μεταβλητής χειρισμού επιλέγεται η ένωση από τα SNPs εκείνα που έχουν την μεγαλύτερη ποσοστιαία διαφορά σε όλες τις περιπτώσεις μεταλλάξεων. Ακολουθούν οι τύποι για την πρώτη κατηγορία μετάλλαξης, οι οποίοι είναι οι ίδιοι και για τις υπόλοιπες.

$$\begin{aligned} \mathit{Threshold} &= \mathit{slope} * \mathit{maxFullMutation} + \mathit{intercept} \\ \mathit{LowerThreshold} &= (1 - \mathit{Threshold}/100) * \mathit{maxFullMutation} \\ \mathit{selectedSNPs} &= \mathit{minFullMutation} \leq \mathit{LowerThreshold} \end{aligned}$$

Εικόνα 23 GMM 2

Το ίδιο αντίστοιχα συμβαίνει με την απουσία μεταλλάξεων και την εν μέρη μετάλλαξη (0 και 1 αντίστοιχα στο προσθετικό μοντέλο). Τελικά επιλέγονται τα SNPs εκείνα που ταυτόχρονα και στις 3 κατηγορίες έχουν την μεγαλύτερη ποικιλία μεταξύ του πληθυσμού. Έτσι μεγιστοποιείται η γενετική διαφορά μεταξύ των δειγμάτων και τα δείγματα ταξινομούνται πιο εύκολα.

Στη συνέχεια θα χρησιμοποιηθεί και η μέθοδος PCA. Η ανάλυση κύριων συνιστωσών (PCA) είναι μια δημοφιλής τεχνική για την ανάλυση μεγάλων συνόλων δεδομένων που περιέχουν μεγάλο αριθμό διαστάσεων/χαρακτηριστικών ανά παρατήρηση, αυξάνοντας την δυνατότητα της ερμηνείας των δεδομένων διατηρώντας παράλληλα τη μέγιστη ποσότητα πληροφοριών. Εφαρμόζεται με γραμμικό μετασχηματισμό των δεδομένων σε ένα νέο σύστημα συντεταγμένων όπου το μεγαλύτερο μέρος της διακύμανσης των δεδομένων μπορεί να εκφραστεί με λιγότερες διαστάσεις από τα αρχικά δεδομένα.

Τέλος, άλλη μία μέθοδος επιλογής χαρακτηριστικών που θα εξετασθεί είναι ο έλεγχος  $\chi^2$ . Το  $\chi^2$ -τεστ του Pearson είναι επίσης γνωστό ως  $\chi^2$ -τεστ για την ανεξαρτησία. Αναπτύχθηκε κατά το έτος 1900. Το  $\chi^2$ -τεστ του Pearson αποτελεί μια στατιστική δοκιμή η οποία εφαρμόζεται σε σύνολα κατηγοριοποιημένων δεδομένων για να αξιολογήθει πόσο πιθανό είναι οποιαδήποτε παρατηρούμενη διαφορά μεταξύ των συνόλων να προέκυψε κατά τύχη. Είναι κατάλληλο για ασύζευκτα δεδομένα από μεγάλα δείγματα. Είναι το πιο ευρέως χρησιμοποιούμενο από όλα τα  $\chi^2$ -τεστ (π.χ., Yates, λόγω πιθανοφάνειας, κλπ.)

Έχοντας δυο κατηγορίες για το χαρακτηριστικό προς εξέταση ( case/control) και τρεις κατηγορίες (0,1,2 – Προσθετικό μοντέλο ) ο πίνακας συχνοτήτων είναι κάπως έτσι :

	<b>AA=0</b>	<b>Aa=1</b>	<b>aa=2</b>	<b>Total</b>
<b>healthy</b>	$N_{11}$	$N_{12}$	$N_{13}$	$R_1$
<b>disease</b>	$N_{21}$	$N_{22}$	$N_{23}$	$R_2$
<b>Total</b>				

Εικόνα 24 Πίνακας συχνοτήτων

## Πειράματα

Για να συγκρίνουμε όλες αυτές τις μεθόδους μεταξύ τους θα χρησιμοποιήσουμε δυο ταξινομητές, τον SVM και τον RF. Θα χωρίσουμε το κάθε σετ δεδομένων σε σετ εκπαίδευσης και επαλήθευσης με αναλογία 70%/30%, καθώς εκπαιδεύοντας τους και εφαρμόζοντας τους στο σετ επαλήθευσης θα συγκρίνουμε τα αποτελέσματα. Προκειμένου να μην επηρεάσει το ποσοστό των δειγμάτων τους ταξινομητές, πραγματοποιείται τυχαία δειγματοληψία με επανάθεση πριν τον διαχωρισμό των σετ δεδομένων ώστε τα δείγματα να είναι αντίστοιχα σε αριθμό και να μην επηρεαστεί η εκπαίδευση .Δεν θα υπάρξει κάποια αναζήτηση η επιλογή των βέλτιστων



υπερπαραμέτρων για τους ταξινομητές αυτούς. Θα ακολουθηθούν οι προεπιλεγμένοι παράμετροι της βιβλιοθήκης Scikit-learn.

Στην περίπτωση των πειραμάτων για την κωδικοποίηση ακολουθείτε η παραπάνω διαδικασία, ενώ στα πειράματα για την επιλογή χαρακτηριστικών η εκάστοτε μέθοδος εκπαιδεύεται/λαμβάνει υπόψιν μόνο το σετ εκπαίδευσης και δεν «βλέπει» καθόλου το σετ επαλήθευσης έως την πρόβλεψη. Το ίδιο συμβαίνει και για τα πειράματα σχετικά με την σημαντικότητα των χαρακτηριστικών.

## PRS και Μηχανική Μάθηση

Για να συγκρίνουμε την μέθοδο PRS με της δυνατότητες των μοντέλων της μηχανικής μάθησης θα πραγματοποιηθεί ένα συγκριτικό πείραμα. Για το πείραμα αυτό, θα χρησιμοποιηθούν τα συνθετικά δεδομένα, έτσι ώστε να μην αποτελέσει η έλλειψη δεδομένων έναν παράγοντα που μπορεί να επηρεάσει τα αποτελέσματα και την σύγκριση των μεθόδων. Τα 10000 άτομα του συνθετικού σετ δεδομένων θα χωριστούν στο 75% και στο 25%, όπου θα χρησιμοποιηθούν για την εκπαίδευση και την επαλήθευση αντίστοιχα.

Στα δεδομένα πραγματοποιούνται της οι τυπικές διαδικασίες προεπεξεργασίας που έχουν αναφερθεί μέχρι στιγμής. Πιο συγκεκριμένα, δεν χρειάστηκε να εφαρμοστούν όλα εκείνα τα βήματα καθώς τα δεδομένα είναι συνθετικά και δεν εμφανίζουν προβλήματα της αναντιστοιχίας στο σύστημα γονιδιακό σύστημα αναφοράς, διπλότυπα SNPs ή αναντιστοιχίες στο χρωμόσωμα XY σε σχέση με το φύλο του δείγματος ( τα δεδομένα παράχθηκαν μόνο για το χρωμόσωμα 21). Επιλέγεται ένα κατώφλι της p-value τιμής,  $5e-40$ , ώστε να επιλεγθούν τα SNPs εκείνα που συνδέονται με το εικονικό της μελέτη χαρακτηριστικό. Τέλος, προκειμένου να μετατραπεί το πρόβλημα σε πρόβλημα δυαδικής ταξινόμησης, γίνεται η θεώρηση πως τα PRS scores ανήκουν σε κάθε αντίστοιχη κατηγορία case/control αν είναι μικρότερα ή μεγαλύτερα του 50%.

Η έξοδος των δικτύων καθώς και η έξοδος από τις μεθόδους PRS θα ανοιχτούν στο διάστημα 0-1 εκφράζοντας την πιθανότητα του δείγματος να συνδέεται με το προς εξέταση χαρακτηριστικό. Στη συνέχεια με κατώφλι το 50% τα δείγματα που έχουν ως



έξοδο αριθμό μικρότερο του 50 θα θεωρούνται μη συσχετισμένα ενώ τα δείγματα με αριθμό μεγαλύτερο συσχετισμένα.

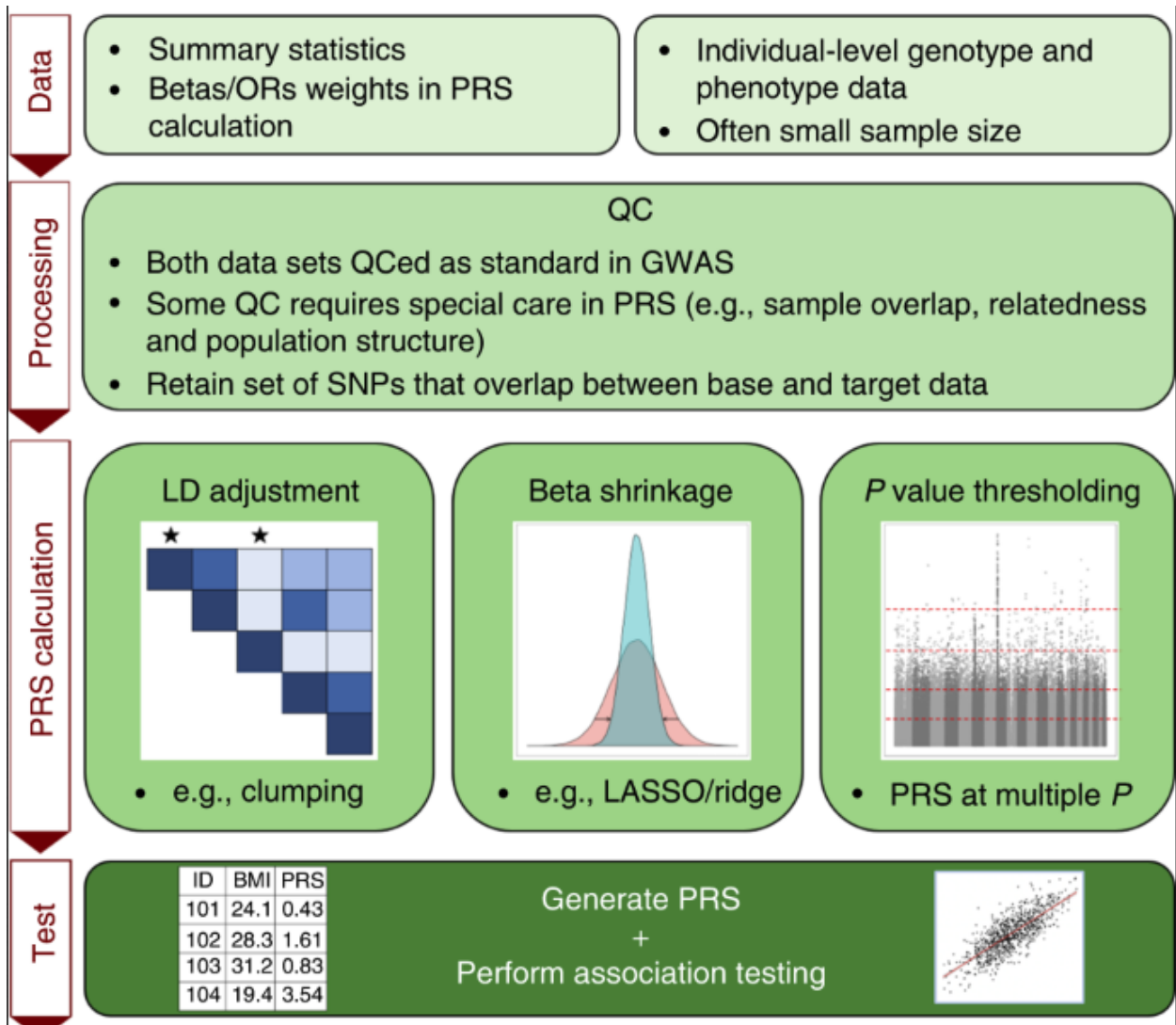
Θα χρησιμοποιηθούν 3 δίκτυα Βαθιάς μάθησης προκειμένου να συναγωνιστούν την μέθοδο PRS. Στην πρώτη περίπτωση υπάρχει το κλασσικό μοντέλο με τα Fully connected στρώματα, εξερευνώντας της της μη γραμμικούς συσχετισμούς των δεδομένων. Στην δεύτερη περίπτωση υπάρχει ένα στρώμα συνέλιξης που προσπαθεί να αντιληφθεί τα πρότυπα στο γενετικό σήμα, ενώ στην Τρίτη περίπτωση υπάρχει ένα αναδραστικό στρώμα που ενδείκνυται για προβλήματα με ακολουθίες δεδομένων.

Dense model		LSTM model		Conv1D model	
Layers	Output Shape	Layers	Output Shape	Layers	Output Shape
Dense	(batch_size, 100)	LSTM	(batch_size, 2319,3)	Conv1D	(batch_size, 2300, 5)
Dropout	(batch_size, 100)	Flatten	(batch_size, 11595)	Flatten	(batch_size, 11500)
Dense	(batch_size, 50)	Dense	(batch_size, 50)	Dense	(batch_size, 50)
Dropout	(batch_size, 50)	Dropout	(batch_size, 50)	Dropout	(batch_size, 50)
Dense	(batch_size, 1)	Dense	(batch_size, 1)	Dense	(batch_size, 1)
Activation	Sigmoid(only on training)	Activation	Sigmoid(only on training)	Activation	Sigmoid(only on training)
Capacity		Capacity		Capacity	
237k		579k		575k	

Εικόνα 25: Δίκτυα Μηχανικής-Βαθιάς Μάθησης

Σχετικά με της μεθόδους PRS, εξεταστούν δυο διαφορετικές προσεγγίσεις, η PRSice και η lassosum. Η πρώτη μέθοδος είναι η κλασσική μέθοδος C+T , δηλαδή clumping και thresholding. Σε αυτή τη μέθοδο αφαιρούνται τα SNPs με βάση το Linkage disequilibrium (LD) ώστε να μην υπερεκτιμηθεί η επιρροή των κοντινών και σχετικών μεταξύ της SNPs. Έπειτα επιλέγονται τα καθοριστικά εκείνα SNPs με βάση το στατιστικό τεστ. Η lassopred προσέγγιση ουσιαστικά διαφέρει στο γεγονός ότι χρησιμοποιώντας την μέθοδο lasso μειώνει το βάρος της επιρροής του κάθε SNP προκειμένου να είναι πιο αντικειμενική η πρόβλεψη.

Η έξοδος των δικτύων καθώς και η έξοδος από τις μεθόδους PRS θα ανοιχτούν στο διάστημα 0-1 εκφράζοντας την πιθανότητα του δείγματος να συνδέεται με το προς εξέταση χαρακτηριστικό. Στη συνέχεια με κατώφλι το 50% τα δείγματα που έχουν ως έξοδο αριθμό μικρότερο του 50 θα θεωρούνται μη συσχετισμένα ενώ τα δείγματα με αριθμό μεγαλύτερο συσχετισμένα.



Εικόνα 26: Διαδικασία πρόβλεψης με PRS

## ΑΠΟΤΕΛΕΣΜΑΤΑ

### Κωδικοποίηση

Προκειμένου να συγκρίνουμε τα μοντέλα κωδικοποίησης των γενετικών δεδομένων, πραγματοποιήθηκαν ορισμένα πειράματα. Τα πειράματα αυτά περιλαμβάνουν προβλέψεις πάνω στα σετ δεδομένων με τους 3 διαφορετικούς τρόπους κωδικοποίησης, το additive model, το recessive/dominant model και το genetic model (one-hot encoding). Για τις προβλέψεις αυτές, θα χρησιμοποιηθούν οι δυο από τους πιο διαδεδομένους αλγορίθμους πρόβλεψης στην Μηχανική Μάθηση, οι SVM και Random Forest. Σκοπός των πειραμάτων αυτών είναι το να εξεταστεί ποια μέθοδος κωδικοποίησης επιφέρει καλύτερα αποτελέσματα στις προβλέψεις. Δεν χρησιμοποιήθηκε κάποιο fine-tuning των υπερπαραμέτρων, με σκοπό την αποφυγή του άσκοπου overfit στα δεδομένα και στο πρόβλημα. Χρησιμοποιήθηκαν οι προεπιλεγμένοι υπερπαραμέτροι της βιβλιοθήκης scikit-learn.

Dataset	N_features	SVM test accuracy	RF test accuracy	SVM training time(s)	RF training time(s)
1	119555	92%	90%	124.74	0.83
	80320	91%	92%	5.11	0.32
	40160	92%	91%	4.74	0.28
2	117286	84%	82%	135.08	0.91
	78766	79%	84%	4.58	0.45
	39383	84%	83%	3.27	0.35
3	108620	77%	73%	44.13	0.54
	73596	76%	76%	1.28	0.29
	36798	77%	74%	0.99	0.25
4	123852	93%	93%	145.03	0.82
	83236	87%	93%	4.74	0.45
	41618	93%	93%	5.36	0.32
5	125685	95%	95%	123.11	7.24
	84490	85%	95%	4.71	0.85
	42245	94%	95%	4.81	0.43
6	144834	100%	100%	16.7	0.33
	99584	100%	100%	1.07	0.25
	49792	100%	100%	0.67	0.2

Εικόνα 27: Αποτελέσματα κωδικοποίησης

Στα αποτελέσματα των προβλέψεων με τον αλγόριθμο πρόβλεψης SVM:

Αρχικά παρατηρείται πως το σετ δεδομένων που σχετίζεται με τον διαβήτη τύπου Β δεν αφήνει περιθώρια σύγκρισης των διαφορετικών τεχνικών κωδικοποίησης, καθώς οι ακρίβειες είναι τέλειες εξίσου και στους τρεις διαφορετικούς τρόπους. Από την άλλη πλευρά εξαιρώντας το συγκεκριμένο σετ δεδομένων μπορεί να διακριθεί πως το recursive/dominant model καταφέρνει σταθερά μειωμένες ακρίβειες σε σχέση με τους άλλους τρόπους κωδικοποίησης. Κατά τα άλλα, το genotypic model έχει παρόμοιες ακρίβειες με το additive model με μοναδική εξαίρεση το τελευταίο σετ δεδομένων όπου επικρατεί με ελάχιστη διαφορά.

Στα αποτελέσματα των προβλέψεων με τον αλγόριθμο πρόβλεψης RF:

Το ίδιο φαινόμενο σε αντιστοιχία με τον SVM παρατηρείται και στον RF, όπου το σετ δεδομένων που σχετίζεται με τον διαβήτη Β έχει ουσιαστικά τέλεια ακρίβεια πρόβλεψης κι δεν ενδείκνυται για σύγκριση. Παρόλα αυτά, σε αντίθεση με τον SVM το recursive/dominant model καταφέρνει να επικρατήσει των υπολοίπων μοντέλων στα τρία από τα πέντε συνολικά σετ που αναφέρονται στο χρώμα των ματιών.

Στα συνολικά αποτελέσματα των προβλέψεων:

Παρατηρώντας συνολικά τα αποτελέσματα προκύπτει πως παρόλο που οι επιδόσεις και η επικρατέστερη μέθοδος κωδικοποίησης διαφέρει ανάμεσα στα μοντέλα πρόβλεψης, η μέγιστη ακρίβεια επιτυγχάνεται σταθερά μέσω του *genotypic model*. Η κυριαρχία του *recursive/dominant model* στον RF επισκιάζεται από την επιτυχία των υπολοίπων μοντέλων στον SVM. Ακόμα το *additive model* ακολουθά σταθερά το *genotypic model* με ελάχιστες διαφορές στην μέγιστη ακρίβεια, μόνο στο τελευταίο σετ δεδομένων. Επιπρόσθετα, δείχνει να επικρατεί πάλι με ελάχιστες διαφορές στις προβλέψεις του RF, έναντι του *genotypic model*.

Τελικά φαίνεται ότι οι συνολικές ακρίβειες μεταξύ μοντέλων, δηλαδή του είναι ελάχιστες. Ωστόσο, οι διαφορές στον αριθμό των χαρακτηριστικών (κίνδυνος *overfit*) και στους χρόνους εκπαίδευσης δεν αντιστοιχούν με τις διαφορές στην ακρίβεια. Η κατανομή των διαφορών αυτών είναι εντελώς διαφορετική και δείχνει ότι το *additive model* καταφέρνει παρόμοια συνολική ακρίβεια με τα άλλα μοντέλα δίχως τα αρνητικά στοιχεία που τα χαρακτηρίζουν.

### Επιλογή των SNPs

Στην συγκεκριμένη κατηγορία πειραμάτων θα εξεταστούν όλες οι μέθοδοι επιλογής/μείωσης των χαρακτηριστικών του γενότυπου. Ο μικρός αριθμός των δειγμάτων σε σχέση με τα χαρακτηριστικά αποτελεί ένα τεράστιο κίνδυνο για παραγωγή μοντέλων πρόβλεψης που κάνουν *overfit*. Ακόμα, όπως και στις μεθόδους PRS, πρέπει να προσδιοριστούν τα SNPs εκείνα που παίζουν κάποιο σημαντικό ρόλο στην κάθε ασθένεια ώστε να διατηρηθούν μόνο οι χρήσιμες εκείνες πληροφορίες που απαιτούνται για την πρόβλεψη.

Οι μέθοδοι PRS χρησιμοποιούν σαν ενδεδειγμένη διαδικασία προεπεξεργασίας διάφορα στατιστικά test προκειμένου να επιλέξουν έναν ικανό αριθμό SNPs και έπειτα να προβλέψουν με βάση το γενότυπο. Αυτή η διαδικασία αποτελεί ένα απαραίτητο βήμα για την αποτελεσματική λειτουργία τους. Για παράδειγμα τα μοντέλα Βαθιάς Μάθησης αποτελούνται από τόσες πολλές παραμέτρους προς εκπαίδευση που είναι αδύνατον να

παραχθούν αξιόπιστα μοντέλα με περίπου 500 δείγματα σε ένα χώρο 1000000 χαρακτηριστικών.

Στα επόμενα πειράματα εφαρμόζεται όλη η τυπική διαδικασία της εκπαίδευσης και επικύρωσης ενός μοντέλου, ώστε να εξετασθεί ποια μέθοδος επιλογής ή μείωσης χαρακτηριστικών θα οδηγήσει σε αποτελεσματικότερα μοντέλα. Και πάλι χρησιμοποιώντας τους SVM και RandomForest εξετάζεται η τελική ακρίβεια συγκριτικά με τις μεθόδους. Οι μέθοδοι αυτοί για λόγους επικύρωσης εφαρμόζονται στο σετ εκπαίδευσης και έπειτα στο σετ επικύρωσης. Έτσι εξετάζεται η αποτελεσματικότητά τους στο να μειωθεί η διάσταση του χώρου χαρακτηριστικών και σε νέα δείγματα που δεν θα έχουν πληροφορία γι' αυτά, δηλαδή θα τους είναι άγνωστα.

Επιλέγεται σε κάθε σετ δεδομένων ένα σετ των αριθμών χαρακτηριστικών, ώστε ή κάθε μέθοδος να συγκρίνεται πάνω στον ίδιο αριθμό χαρακτηριστικών. Η μέθοδος PCA προφανώς δεν μπορεί να υπερβεί έναν αριθμό παραγόμενων συνιστωσών. Επίσης ενδιαφέρον αποτελεί και η συνολική σύγκριση των ακριβειών ανεξάρτητα του αριθμού χαρακτηριστικών αλλά και η σύγκριση μεταξύ των αποτελεσμάτων όλων των χαρακτηριστικών σε σχέση με τα σετ δεδομένων με μειωμένα χαρακτηριστικά. Τέλος, στην περίπτωση που παράγονται νέες συνιστώσες και δεν επιλέγονται χαρακτηριστικά, διατηρείται η χρησιμότητα της εκάστοτε μεθόδου. Ο εκπαιδευμένος αλγόριθμος μπορεί να εξάγει τις απαραίτητες συνιστώσες σε αντίστοιχα νέα δείγματα και να προβεί σε προβλέψεις χωρίς απαραίτητα να γίνει επιλογή χαρακτηριστικών. Το σετ δεδομένων με τον διαβήτη τύπου Β δεν συμμετέχει σε αυτά τα πειράματα.

Method	N_features	SVM train acc.	SVM test acc.	RF train acc.	RF test acc.
Chi squared test	67	98%	84%	100%	90%
Chi squared test	126	99%	83%	100%	93%
Chi squared test	309	100%	86%	100%	92%
PCA	67	94%	82%	100%	92%
PCA	126	94%	82%	100%	87%
PCA	309	100%	72%	100%	77%
Relief	67	96%	81%	100%	90%
Relief	126	97%	83%	100%	91%
Relief	309	100%	88%	100%	90%
GMM	67	94%	76%	99%	82%
GMM	126	97%	80%	100%	85%
GMM	309	99%	87%	100%	88%
No method	40160	100%	92%	100%	91%

Εικόνα 28: Αποτελέσματα επιλογής SNP, σετ 1

Στο πρώτο σετ δεδομένων παρατηρείται ότι τελικώς η όποια προσπάθεια μείωσης των χαρακτηριστικών δεν προσδίδει κάποιο συγκριτικά θετικό αποτέλεσμα σε σχέση με την χρήση όλων των χαρακτηριστικών, σχετικά με τον SVM. Τα αποτελέσματα δεν είναι ικανοποιητικά σε σχέση με τον RF. Στον RF, το στατιστικό τεστ αποδίδει την μέγιστη ακρίβεια όλων των περιπτώσεων γενικότερα με την επιλογή 127 χαρακτηριστικών. Αξίζει να σημειωθεί ότι η μέθοδος PCA επιτυγχάνει αντίστοιχη ακρίβεια με την χρήση όλων των χαρακτηριστικών αλλά και την αμέσως καλύτερη επίδοση με μόλις 6 συνιστώσες. Επιπρόσθετα, στην συνολική ακρίβεια και στις δυο μεθόδους φαίνεται να κυριαρχεί η επιλογή χαρακτηριστικών με το στατιστικό τεστ, και η μέθοδος Relief ακολουθεί με μικρή διαφορά.

Method	N_features	SVM train acc.	SVM test acc.	RF train acc.	RF test acc.
Chi squared test	77	99%	80%	100%	84%
Chi squared test	134	100%	80%	100%	84%
Chi squared test	273	100%	82%	100%	82%
Chi squared test	424	100%	81%	100%	85%
PCA	77	85%	70%	100%	84%
PCA	134	86%	70%	100%	80%
PCA	273	100%	84%	100%	73%
PCA	424	100%	70%	100%	71%
Relief	77	84%	69%	100%	80%
Relief	134	88%	70%	100%	77%
Relief	273	94%	81%	100%	80%
Relief	424	95%	80%	100%	83%
GMM	77	90%	71%	99%	73%
GMM	134	95%	72%	100%	78%
GMM	273	98%	76%	100%	77%
GMM	424	100%	79%	100%	75%
No method	39383	100%	84%	100%	83%

Εικόνα 29: Αποτελέσματα επιλογής SNP, σετ 2

Τις ίδιες περίπου παρατηρήσεις μπορεί να κάνει κανείς και για το επόμενο σετ δεδομένων, όπου στην περίπτωση του SVM και με τον συγκεκριμένο αριθμό χαρακτηριστικών προς εξέταση δεν παρουσιάστηκε μεγαλύτερη ακρίβεια συγκριτικά με την χρήση όλων των χαρακτηριστικών. Στην περίπτωση του RF κυριαρχεί και πάλι η μέθοδος του στατιστικού τεστ όπως και γενικότερα.

Method	N_features	SVM train acc.	SVM test acc.	RF train acc.	RF test acc.
Chi squared test	102	100%	67%	100%	68%
Chi squared test	183	100%	70%	100%	73%
Chi squared test	247	100%	69%	100%	72%
PCA	102	91%	74%	100%	57%
PCA	183	100%	77%	100%	56%
PCA	247	100%	78%	100%	56%
Relief	102	98%	76%	100%	75%
Relief	183	100%	85%	100%	79%
Relief	247	100%	79%	100%	75%
GMM	116	100%	76%	100%	76%
GMM	183	100%	75%	100%	78%
GMM	257	100%	72%	100%	72%
No method	36798	100%	77%	100%	74%

Εικόνα 30: Αποτελέσματα επιλογής SNP, σετ 3

Στο συγκεκριμένο σετ δεδομένων παρατηρείται ότι η μείωση/ επιλογή χαρακτηριστικών είναι ένα βήμα που έδωσε αρκετή παραπάνω ακρίβεια στην πρόβλεψη σε σχέση με την πρόβλεψη στο σύνολο των χαρακτηριστικών. Ο SVM πέτυχε περίπου 10% επιπλέον ακρίβεια στην πρόβλεψη με την επιλογή των 183 SNPs του αλγόριθμου Relief, και 7% περίπου παραπάνω ακρίβεια από το κυρίαρχο στα προηγούμενα πειράματα στατιστικό τεστ. Το ίδιο γεγονός θα προκύψει και στον RF αλγόριθμο, με την μέθοδο Relief και GMM να επικρατούν χωρίς να αγγίζουν όμως τα επίπεδα ακρίβειας της καλύτερης πρόβλεψης του SVM. Αξίζει να σημειωθεί πως η μέθοδος PCA δεν έπιασε υψηλές ακρίβειες στο συγκεκριμένο σετ δεδομένων σε σχέση με τα προηγούμενα.



Method	N_features	SVM train acc.	SVM test acc.	RF train acc.	RF test acc.
Chi squared test	172	100%	91%	100%	92%
Chi squared test	236	100%	89%	100%	93%
Chi squared test	459	100%	91%	100%	93%
PCA	172	97%	87%	100%	89%
PCA	236	100%	93%	100%	89%
PCA	459	100%	79%	100%	87%
Relief	172	89%	74%	100%	92%
Relief	236	93%	79%	100%	93%
Relief	459	98%	84%	100%	93%
GMM	189	98%	85%	100%	91%
GMM	236	99%	86%	100%	92%
GMM	465	99%	89%	100%	92%
No method	41618	100%	93%	100%	93%

Εικόνα 31: Αποτελέσματα επιλογής SNP, σελ 4

Η κατάσταση σε αυτό το σετ δεδομένων είναι η αντίστοιχη με τα πρώτα 2 σετ. Η μεγαλύτερη ακρίβεια έρχεται από το σύνολο των χαρακτηριστικών, όμως στον RandomForest η ίδια ακριβώς ακρίβεια μπορεί να προκύψει από το στατιστικό τεστ και την μέθοδο Relief με τα 459 χαρακτηριστικά.

Method	N_features	SVM train acc.	SVM test acc.	RF train acc.	RF test acc.
Chi squared test	140	97%	90%	98%	95%
Chi squared test	251	98%	93%	98%	95%
Chi squared test	457	98%	93%	98%	95%
PCA	140	96%	86%	98%	95%
PCA	251	98%	95%	98%	95%
PCA	457	98%	94%	98%	95%
Relief	140	95%	84%	98%	95%
Relief	251	97%	89%	98%	95%
Relief	457	97%	92%	98%	95%
GMM	140	96%	87%	98%	92%
GMM	251	97%	89%	98%	94%
GMM	457	97%	94%	98%	96%
No method	42245	98%	94%	98%	95%

Εικόνα 32: Αποτελέσματα επιλογής SNP, σελ 5

Στο τελευταίο σετ δεδομένων παρατηρούμε ότι η μέθοδος pca αλλά και η GMM ξεπερνούν τα αποτελέσματα του συνόλου των χαρακτηριστικών. Μάλιστα, η μέθοδος GMM καταφέρνει την μέγιστη ακρίβεια στον RF με μόλις 457 χαρακτηριστικά SNPs. Στην περίπτωση του SVM η μέθοδος PCA καταφέρνει και αυτή να ξεπεράσει σε ακρίβεια την χρήση όλων των χαρακτηριστικών με 457 συνιστώσες, δίχως όμως να φτάσει την αντίστοιχη της GMM. Γενικότερα παρατηρείται πως ενώ συνολικά στα πειράματα πάνω σε διάφορους αριθμούς χαρακτηριστικών το στατιστικό τεστ εμφανίζει την καλύτερη ακρίβεια, την μέγιστη ακρίβεια καταφέρνουν οι άλλες μέθοδοι.

## Σημαντικά SNPs

Οι διάφορες μελέτες που έχουν γίνει σε GWAS σε διάφορους πληθυσμούς σχετικά με το χρώμα των ματιών, έχουν καταλήξει στο συμπέρασμα ότι κάποια συγκεκριμένα SNPs και συνεπώς γονίδια στα οποία ανήκει το κάθε SNP παίζουν καθοριστικό ρόλο στο τελικό χρώμα των ματιών του κάθε ανθρώπου. Βέβαια εξετάζοντας το κάθε φαινοτυπικό χαρακτηριστικό προκύπτουν και άλλα συμπεράσματα, όπως το ότι κάποια χαρακτηριστικά ή ασθένειες καθορίζονται από πολλά γονίδια μαζί συνθετικά και δεν είναι ένα γονίδιο μόνο ο κυρίαρχος παράγοντας. Ακόμα πρόσφατες μελέτες δείχνουν πως όχι μόνο παίζουν πολλά γονίδια ρόλο αλλά και οι διάφοροι συνδιασμοί γονιδίων μεταξύ τους.

Στα επόμενα πειράματα γίνεται η προσπάθεια να προσδιοριστούν τα γονίδια εκείνα που παίζουν τελικά τον πιο σημαντικό ρόλο στην ταξινόμηση και συνεπώς μπορούν να θεωρηθούν καθοριστικοί παράγοντες για το εκάστοτε χαρακτηριστικό. Θα χρησιμοποιηθούν τα σετ δεδομένων με το χρώμα των ματιών. Αναζητώντας διάφορες μελέτες μπορεί κανείς να παρατηρήσει ότι τα πιο συνηθισμένα γονίδια που αναφέρεται ότι σχετίζονται με το χρώμα των ματιών είναι τα εξής :

'ASIP', 'IRF4', 'SLC24A4', 'SLC24A5', 'SLC24A2', 'TPCN2', 'TYR', 'TYRP1', 'OCA2', 'HERC2' με τα OCA2 και HERC2 ως βασικότερα.

Τα σετ δεδομένων που είναι διαθέσιμα στα πλαίσια της εργασίας αυτής διαθέτουν έναν περιορισμένο αριθμό SNPs, δεδομένου ότι πολλά SNPs έχουν αφαιρεθεί για να υπάρχει πληρότητα τιμών σε όλα τα δείγματα. Από τα παραπάνω χαρακτηριστικά γονίδια τα σετ δεδομένων περιέχουν τα εξής 3 :

'SLC24A4', 'SLC24A2', 'OCA2'

Στη συνέχεια, θα εξεταστούν οι τρόποι επιλογής χαρακτηριστικών SNPs ώστε να συγκριθούν τα αποτελέσματα μεταξύ των στατιστικών τεχνικών και των τεχνικών μηχανικής μάθησης.

Dataset	RF	SVM	Relief	GMM	No of SNPs
1	-	-	-	SLC24A2	n=10
2	-	OCA2	-	-	
3	OCA2	OCA2	OCA2	OCA2,SLC24A2,SLC24A4	
4	-	OCA2	-	SLC24A2,SLC24A4	
5	-	-	-	SLC24A4	
1	SLC24A4	-	-	SLC24A2	n=100
2	-	OCA2	-	-	
3	OCA2	OCA2	OCA2	OCA2,SLC24A2,SLC24A4	
4	-	OCA2	-	OCA2,SLC24A2,SLC24A4	
5	OCA2	OCA2	-	SLC24A2,SLC24A4	
1	SLC24A4		-	SLC24A2,SLC24A4	n=1000
2	OCA2,SLC24A2	OCA2,SLC24A4	-	SLC24A4	
3	OCA2,SLC24A2	OCA2,SLC24A2	OCA2	OCA2,SLC24A2,SLC24A4	
4	OCA2	OCA2,SLC24A2	-	OCA2,SLC24A2,SLC24A4	
5	SLC24A2', 'OCA2'	OCA2,SLC24A4	-	OCA2,SLC24A2,SLC24A4	

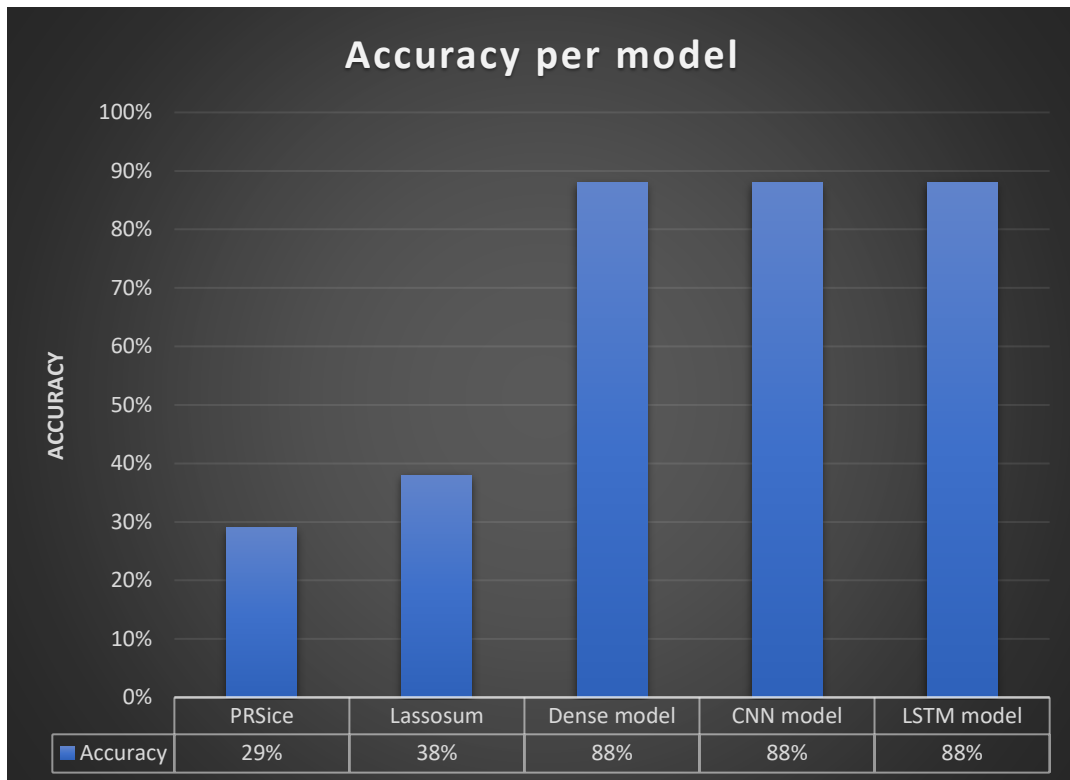
Εικόνα 33: Σημαντικά SNPs ανά αριθμό χαρακτηριστικών

Από τα παραπάνω πειράματα μπορεί κανείς να παρατηρήσει ότι η επιλογή χαρακτηριστικών με βάση τα βάρη του SVM αντιλαμβάνεται από τα 10 πρώτα χαρακτηριστικά την ύπαρξη του γονιδίου OCA2. Από εκεί και πέρα στα 100 χαρακτηριστικά και ο αλγόριθμος RandomForest επιτυγχάνει να αντιληφθεί τη σημαντικότητα του OCA2, όπως και ο SVM στο τελευταίο σετ δεδομένων. Πλέον στα 1000 χαρακτηριστικά και οι 2 αυτοί αλγόριθμοι καταφέρνουν να αντιληφθούν στα διάφορα σετ δεδομένων τα ίδια χαρακτηριστικά με αυτά που αναφέρουν οι μελέτες, χωρίς όμως να υπάρχει σετ δεδομένων τέτοιο που να τα περιλαμβάνει όλα.

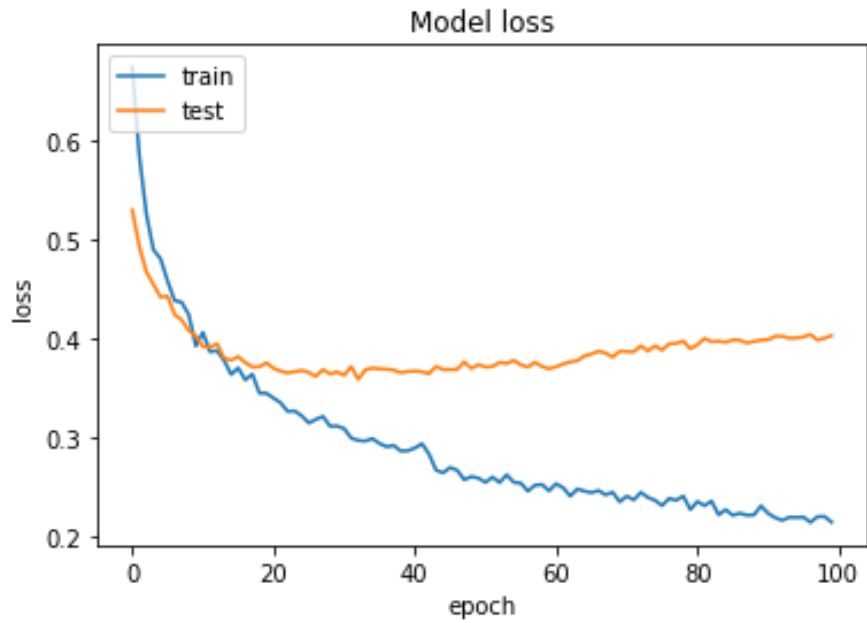
Αξίζει κανείς να παρατηρήσει πως η μέθοδος Relief κατάφερε να επιλέξει το γονίδιο OCA2 μόνο σε ένα σετ δεδομένων από τα 10 χαρακτηριστικά και δεν κατάφερε κάτι πρόσθετο στην αναζήτηση σε άλλο σετ ή άλλου γονιδίου μέχρι και τα 1000 χαρακτηριστικά. Επίσης το πρώτο σετ δεδομένων αποτελεί εξαίρεση και από τον SVM, αφού μόνο ο RandomForest κατάφερε να βρει τα αντίστοιχα γονίδια της βιβλιογραφίας.

## PRS και Μηχανική Μάθηση

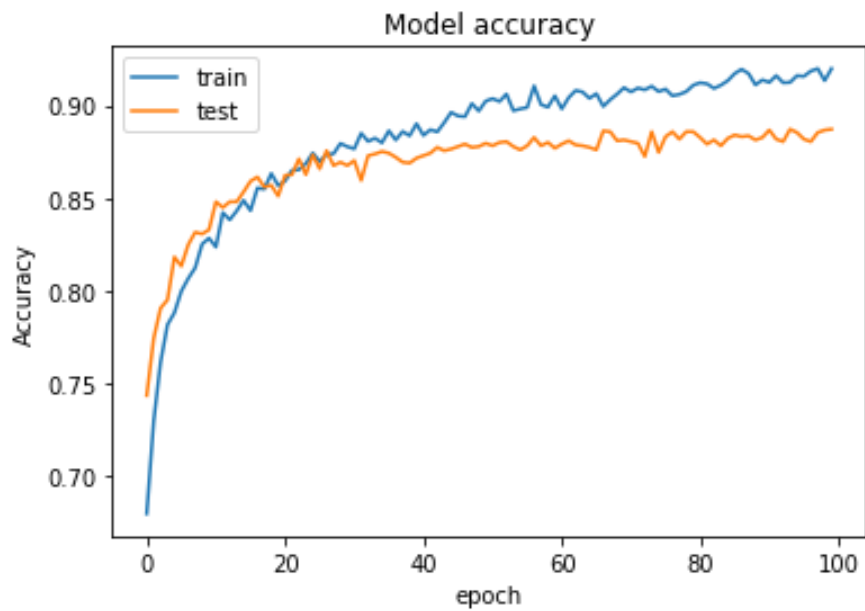
Στη συνέχεια ακολουθούν τα πειράματα που έγιναν ώστε να συγκριθούν οι διάφορες μέθοδοι PRS με ένα δίκτυο βαθιάς μάθησης. Το σετ δεδομένων που θα χρησιμοποιηθεί είναι το συνθετικό σετ δεδομένων που δημιουργήθηκε με 2500 άτομα σαν σετ επαλήθευσης και 7500 σαν σετ εκπαίδευσης. Ο αριθμός των δειγμάτων αυτών είναι ένας σχετικά ικανός αριθμός δειγμάτων ώστε να εκπαιδευτούν τα μοντέλα αυτά. Στη συνέχεια θα γίνει σύγκριση των PRS μεθόδων με τα δίκτυα βαθιάς μάθησης. Υπενθυμίζεται πως το σετ δεδομένων έχει δεχθεί τα ίδια βήματα προεπεξεργασίας ανάμεσα στις PRS μεθόδους και στα δίκτυα βαθιάς μάθησης, όπως αναφέρεται στο κεφάλαιο της μεθοδολογίας. Τα δείγματα έχουν χωριστεί με τέτοιο τρόπο ώστε τα σετ δεδομένων να είναι ισορροπημένα σχετικά με τον πληθυσμό case/control. Παρακάτω φαίνονται τα αποτελέσματα των πειραμάτων καθώς και τα γραφήματα με την πορεία της εκπαίδευσης για το επικρατές μοντέλο.



Εικόνα 34: Μηχανική Μάθηση και PRS μέθοδοι



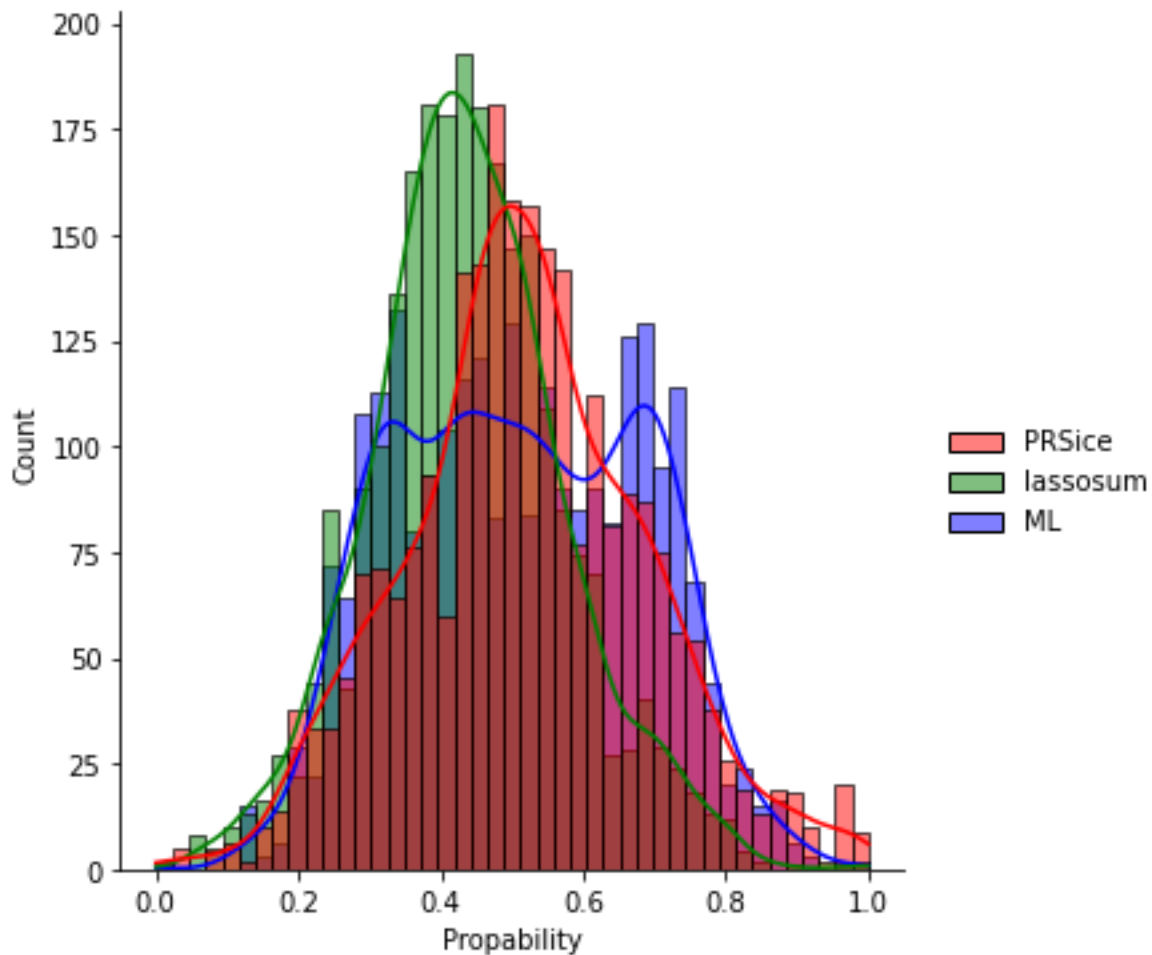
Εικόνα 35: Διάγραμμα loss Dense μοντέλου



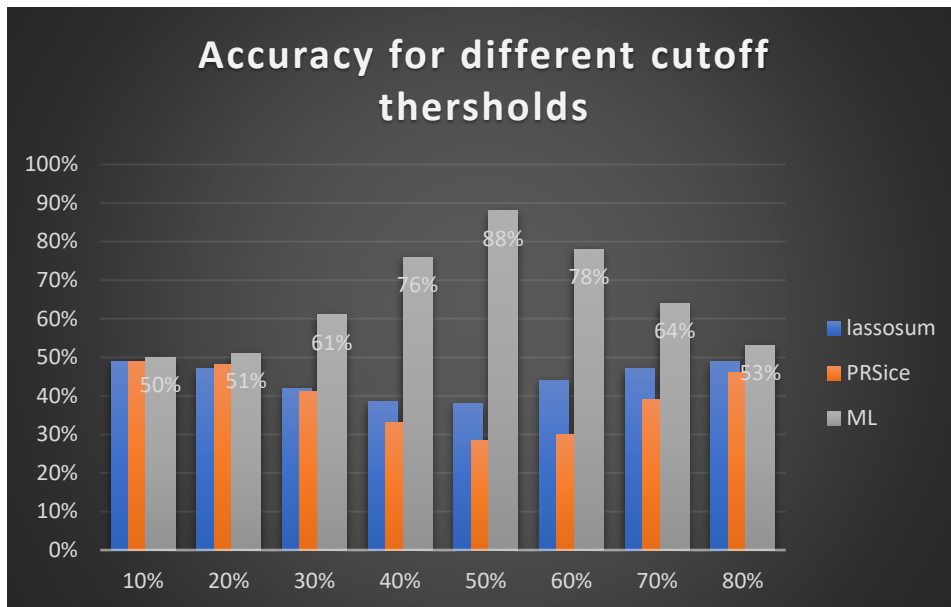
Εικόνα 36: Διάγραμμα accuracy Dense μοντέλου

Τα 3 διαφορετικά δίκτυα δεν παρουσίασαν κάποια σημαντική διαφορά στα αποτελέσματά τους. Η συνέλιξη πάνω στο γενότυπο αλλά και το LSTM στρώμα δεν

έδειξαν να προσδίδουν κάποια ουσιαστική διαφοροποίηση στα αποτελέσματα του μοντέλου. Αντίθετα είχαν περισσότερο υπολογιστικό κόστος, χωρίς να προσφέρουν κάτι πρόσθετο. Η μέθοδος PRS με τις διάφορες παραλλαγές της απέχει αρκετά σε ακρίβεια από τις μεθόδους μηχανικής μάθησης. Τέλος στο παρακάτω γράφημα μπορεί να φανεί η κατανομή των αποτελεσμάτων των διαφορετικών μεθόδων αλλά και η κατανομή της ακρίβειας των μεθόδων PRS αν άλλαζε ο τρόπος με τον οποίο θα χωριζόντουσαν τα δείγματα με βάση τις πιθανότητες σε case/control.



Εικόνα 37: Κατανομή προβλέψεων



Εικόνα 38: Κατανομή προβλέψεων με διαφορετικά κατώφλια



## ΣΥΜΠΕΡΑΣΜΑΤΑ

### Γενετικά δεδομένα και προβλήματα

Η όλη διαδικασία συγγραφής τις παρούσας εργασίας και η ενασχόληση με το συγκεκριμένο αντικείμενο του κλάδου τις υγείας, από την σκοπιά τις εξερεύνησης και εφαρμογής διάφορων σύγχρονων τεχνικών σε αυτό, οδήγησε σε διάφορα συμπεράσματα. Τα ιατρικά δεδομένα και γενικότερα ο κλάδος υγείας είναι ένας αρκετά δύσκολος κλάδος για την εφαρμογή τεχνολογικών καινοτομιών. Η υγεία αποτελεί το βασικότερο κομμάτι τις ανθρώπινης ζωής και απαιτούνται αργά και σταθερά βήματα για την εξέλιξη. Προκειμένου να εφαρμοστεί μία νέα ιδέα θα πρέπει να έχουν εξεταστεί ενδελεχώς εκείνες οι παράμετροι που την επηρεάζουν, διότι το αποτέλεσμα αυτό αφορά άμεσα την ανθρώπινη ζωή και μπορεί να αποβεί καταστροφικό.

Παρατηρείται λοιπόν η τεράστια τεχνολογική εξέλιξη των τελευταίων χρόνων και ιδιαίτερα στον κλάδο τις Τεχνητής Νοημοσύνης-Μηχανικής Μάθησης-Βαθιάς μάθησης σε διάφορους τομείς. Πλέον είναι γεγονός πως υπάρχουν διαθέσιμα στο ευρύ κοινό αυτόνομα αυτοκίνητα που οδηγούν μέχρι τον τελικό προορισμό χωρίς την παραμικρή βοήθεια του επιβάτη-χρήστη, χρησιμοποιώντας την όραση υπολογιστών. Ακόμα υπάρχουν εικονικοί “βοηθοί”, ολοκληρωμένα προγράμματα που μπορούν να καθοδηγήσουν πλήρως έναν πελάτη σχετικά κάποια αγορά του σε ένα εμπορικό μαγαζί, ή και να αναλάβουν την οργάνωση και συντήρηση ολόκληρου νοικοκυριού μέσω φωνητικών εντολών.

Η εμπιστοσύνη των τεχνολογικών εξελίξεων που έχει κατακτηθεί σε διάφορους κλάδους, δεν είναι αντίστοιχη με την εμπιστοσύνη στον ιατρικό κλάδο. Ίσως ο κλάδος αυτός να είναι ο τελευταίος που θα τις υιοθετήσει. Βασικό χαρακτηριστικό που απαιτείται για την εμπιστοσύνη αυτή είναι η επεξηγησιμότητα των προβλέψεων που παράγονται. Σε έναν αυτόματο μεταφραστή, δεν ενδιαφέρει τον χρήστη το πως προέκυψε τελικώς ή πρόβλεψη τις λέξης στην γλώσσα τις επιθυμίας του. Το μοντέλο πρόβλεψης μπορεί να αποτελεί γι’ αυτόν ένα «μαύρο κουτί» και δεν τίθεται κανένα θέμα εμπιστοσύνης. Στην περίπτωση που η πρόβλεψη αυτή αναφέρεται σε κάποια βασική απόφαση που θα

πρέπει να πάρει ένας αρμόδιος γιατρός για την υγεία του ασθενούς, πολύ δύσκολα θα εμπιστευτεί το «μαύρο κουτί» συγκριτικά με τον υπεύθυνο γιατρό. Γι' αυτό μέχρι στιγμής όλες οι εφαρμογές υγείας που σχετίζονται με μοντέλα πρόβλεψής τεχνητής νοημοσύνης λειτουργούν πάντα υποστηρικτικά στις αποφάσεις των ειδικών και όχι καθοριστικά.

Οι μέθοδος PRS είναι μία μέθοδος που παράγει ένα σκορ επί τις εκατό, το οποίο φανερώνει την σύνδεση τις γενετικής πληροφορίας του ατόμου με το εκάστοτε χαρακτηριστικό. Το σκορ αυτό αποτελεί από την φύση του μία μετρική που μπορεί να βοηθήσει υποστηρικτικά κάποια κλινική απόφαση και όχι καθοριστικά. Μία συνηθισμένη εφαρμογή του συγκεκριμένου σκορ που πραγματοποιούν στις ιδιωτικές εταιρίες σε όλο τον πλανήτη είναι να λαμβάνεται η γονιδιακή πληροφορία του ενδιαφερόμενου ατόμου και να κατασκευάζονται διάφορα σκορ που θα αναδεικνύουν τις πιθανές ασθένειες που μπορεί να εμφανίσει το συγκεκριμένο άτομο με στόχο την πρόληψη. Η ερμηνεία του σκορ αυτού είναι ακαθόριστη σε μεγάλο βαθμό, αφού πρακτικά ένα σκορ 40% συγκριτικά με ένα 60% για μία σοβαρή ασθένεια δεν προσδίδει κάποια ουσιαστική και χρήσιμη πληροφορία στον ενδιαφερόμενο.

Βέβαια η «ακαθόριστη» λογική των σκορ αυτών ταιριάζει απόλυτα με την πραγματικότητα και την ικανότητα τις αντικειμενικής πρόβλεψης πάνω στα γενετικά δεδομένα. Αρκεί κανείς να αναλογιστεί τα εξής: Αρχικά, ο ανθρώπινος πληθυσμός αποτελείται από τις διαφορετικές φυλές, με διαφορετικά γονιδιακά χαρακτηριστικά η κάθε μία. Συνεπώς ένα τοίχος που εμποδίζει την αποτελεσματικότητα των προβλέψεων είναι ότι τις γονιδιακές διαφορές οφείλονται στο γεγονός αυτό και δεν αποτελούν πάντα δείκτες για την ύπαρξη κάποιας συσχέτισης με το εξεταζόμενο χαρακτηριστικό.

Επίσης τα γενετικά δεδομένα είναι δυσεύρετα. Αποτελούν ευαίσθητα προσωπικά δεδομένα, είναι αρκετά σπάνιο φαινόμενο να βρίσκονται ελεύθερα και διαθέσιμα σε όλους και ακόμα κι αν συμβεί αυτό δεν εγγυάται κανείς την ποιότητα και την πληρότητα τις. Επιπρόσθετα ο μικρός αριθμός των διαθέσιμων δεδομένων καθιστά ακόμα μεγαλύτερη πρόκληση την διαφορετικότητα των φυλών και το πως επηρεάζει αυτή την κάθε πρόβλεψη.

Ακόμα ένας βασικός παράγοντας που μειώνει την αποτελεσματικότητα των προβλέψεων είναι η φύση του χαρακτηριστικού ή τις ασθένειας τις εξέταση. Αυτό συμβαίνει διότι τα περισσότερα χαρακτηριστικά έχουν τεράστια συσχέτιση με το περιβάλλον, την αλληλεπίδραση του περιβάλλοντος με τον άνθρωπο και την εξελικτική διαδικασία στο ανθρώπινο είδος. Στην εργασία αυτή γίνεται η προσπάθεια να βρεθούν μοτίβα πάνω στη βιολογική σύσταση του ατόμου και δεν λαμβάνονται υπόψιν οι αλληλεπιδράσεις με το

περιβάλλον του. Είναι τις προφανές ότι ανεξάρτητα από την βιολογική «ταυτότητα» του καθενός, η έκθεση σε επικίνδυνη ακτινοβολία μπορεί να επιφέρει καταστάσεις υγείας ανεξάρτητες από την γενετική προδιάθεση.

### Συλλογή γενετικών δεδομένων

Στο σημείο αυτό αξίζει να αναφερθεί το πόσο μεγάλη πρόκληση είναι το να συγκεντρωθεί ένα σετ γενετικών δεδομένων, καθώς και να είναι κατάλληλο για έρευνα. Από την μία πλευρά, το OpenSNP είναι ουσιαστικά η μόνη πύλη που μπορεί κάποιος να προμηθευτεί γενετικά δεδομένα χαρακτηρισμένα με κάποιο φαινότυπο εύκολα. Σίγουρα υπάρχουν διάφοροι τρόποι να αποκτήσει κάποιος πρόσβαση σε παρόμοια δεδομένα, αλλά ο τρόπος τις είναι δύσκολος. Ενδεικτικά η μεγαλύτερη βάση γενετικών δεδομένων αυτή τη στιγμή, η UK Biobank, απαιτεί χρηματικό αντίτιμο για την πρόσβαση στα δεδομένα τις καθώς και πολυήμερη διαδικασία για την αίτηση και τελική αποδοχή τις πρόσβασης. Η dbGaP, η βάση δεδομένων Genotypes and Phenotypes απαιτεί και αυτή αρκετές διαδικασίες και επίσημα αιτήματα από το εκάστοτε εκπαιδευτικό ίδρυμα για την πρόσβαση στα δεδομένα.

Αυτές οι διαδικασίες λειτουργούν ως δικλίδες ασφαλείας. Η μη ορθή χρήση αυτών των δεδομένων μπορεί να έχει τις αρνητικές συνέπειες. Ο γενότυπος του κάθε ατόμου είναι ξεχωριστός, με συνέπεια η αναγνώριση του μέσω αυτού να μην αποτελεί πλέον πρόκληση. Σε συνδυασμό με τις τεχνολογικές εξελίξεις, εγείρονται διάφορα ζητήματα διακρίσεων. Αν υποθετικά ανακαλυφθεί μέσα από την έρευνα ένα SNP που να καθορίζει την αποδοτικότητα του ατόμου σε μια συγκεκριμένη εργασία, τότε είναι πολύ πιθανό να λαμβάνεται υπόψιν και τις ο παράγοντας κατά τη διαδικασία επιλογής υποψηφίων. Το ίδιο γεγονός μπορεί να προκύψει και στην περίπτωση των ασφαλειών υγείας, εξετάζοντας το γενότυπο του κάθε ατόμου για ευαισθησία σε διάφορες ασθένειες. Οι προεκτάσεις που μπορούν να προκύψουν είναι εύκολα αντιληπτές.

Από την άλλη πλευρά, η υποστήριξη τις έρευνας με ανοιχτά και μεγάλα δεδομένα θα μπορούσε να οδηγήσει σε σημαντικές ανακαλύψεις. Ενδεικτικά, τα αποτελέσματα των γενετικών ελέγχων μπορούν να προσφέρουν μια αίσθηση ανακούφισης από την αβεβαιότητα και να βοηθήσουν τους ειδικούς αλλά και τους καθημερινούς ανθρώπους να λάβουν τεκμηριωμένες αποφάσεις σχετικά με τη διαχείριση της υγειονομικής της

περίθαλψης. Για παράδειγμα, ένα αρνητικό αποτέλεσμα μπορεί να εξαλείψει την ανάγκη για περιττούς ελέγχους και εξετάσεις ελέγχου σε ορισμένες περιπτώσεις. Ένα θετικό αποτέλεσμα μπορεί να κατευθύνει ένα άτομο για τις διαθέσιμες επιλογές πρόληψης, παρακολούθησης και θεραπείας. Ορισμένα αποτελέσματα δοκιμών μπορούν τις να βοηθήσουν τις ανθρώπους να λάβουν αποφάσεις σχετικά με την απόκτηση παιδιών. Ο προληπτικός έλεγχος νεογνών μπορεί να εντοπίσει γενετικές διαταραχές νωρίς στη ζωή, ώστε η θεραπεία να μπορεί να ξεκινήσει όσο το δυνατόν νωρίτερα.

Τα αποτελέσματα τις παρούσας εργασίας σίγουρα περιορίζονται από την στιγμή που τα διαθέσιμα δεδομένα αναφέρονται σε δεκάδες ανθρώπινα δείγματα. Ακόμα, ο μεγαλύτερος περιορισμός είναι ότι προκειμένου να κατασκευαστεί ένα σετ δεδομένων έστω με αυτές τις δεκάδες των δειγμάτων, αναγκαστικά εξετάζεται το χαρακτηριστικό του χρώματος των ματιών. Τα υπόλοιπα χαρακτηριστικά δεν συγκεντρώνουν αρκετά δείγματα ώστε να στηθούν κάποια βασικά πειράματα. Το σετ δεδομένων με τους πάσχοντες από διαβήτη τύπου Β, περιέχει τόσα λίγα δείγματα που οι αλγόριθμοι προσαρμόζονται στα λίγα αυτά δεδομένα και η πρόβλεψη δεν έχει πραγματική αξία. Ενώ έχουν αναπτυχθεί τεχνικές για συνθετικά σετ δεδομένων, ο ευαίσθητος ιατρικός τομέας σχετικά με αυτή την οικογένεια πειραμάτων έχει ανάγκη από πραγματικά δεδομένα.

Η εύκολη πρόσβαση σε σετ δεδομένων με περισσότερα δείγματα και σε διάφορα χαρακτηριστικά αλλά και ασθένειες, θα επέτρεπε πολλά χρήσιμα συμπεράσματα. Αρκεί κανείς να παρατηρήσει την τεράστια εξέλιξη που έχει προκύψει στην όραση υπολογιστών και στην αντίληψη τις φωνής και τις γλώσσας από τις υπολογιστές. Πολύ σημαντικός παράγοντας στην εξέλιξη αυτή αποτελεί η ανοικτή διαθεσιμότητα των διάσημων, μεγάλων και οργανωμένων σετ δεδομένων, τις το ImageNet. Είναι προφανές πως η προσοχή της ερευνητικής κοινότητας θα στραφεί προς τα εκεί που υπάρχει αυτή η διαθεσιμότητα. Η εισαγωγή στις εξελιγμένες τεχνικές μηχανικής και βαθιάς μάθησης αλλά και γενικότερη εφαρμογή πραγματοποιείται σε αυτά τα πεδία, λόγω τις διαθεσιμότητας των δεδομένων. Δυστυχώς η διαθεσιμότητα αυτή δεν υπάρχει στα γενετικά δεδομένα και η έρευνα και η εφαρμογή αυτών των τεχνικών αποτελεί ιδιαίτερη πρόκληση και απαιτεί πολύπλευρες γνώσεις από όλες τις επιστήμες που εμπλέκονται.

Στην παρούσα εργασία έχουν γίνει διάφορες συμβάσεις προκειμένου να προκύψουν ορισμένα αντικειμενικά συμπεράσματα. Η μεταφορά του προβλήματος πρόβλεψης του χαρακτηριστικού σε δυαδική ταξινόμηση και η αναγκαιότητα μείωσης των διαστάσεων είναι κάποια βήματα που μπορεί να αποφευχθούν με την ύπαρξη μεγαλύτερου όγκου δεδομένων. Έτσι θα μπορούσαν να χρησιμοποιηθούν πιο «βαριά» και περίπλοκα μοντέλα χωρίς το φαινόμενο του overfitting, συγκριτικά με την μέθοδο PRS που γι' αυτήν αποτελεί αναγκαίο βήμα η επιλογή χαρακτηριστικών.

## Κωδικοποίηση των δεδομένων

Προκειμένου να χρησιμοποιηθούν διάφορες μέθοδοι μηχανικής μάθησης ή και δίκτυα βαθιάς μάθησης, τα δεδομένα πρέπει να μεταφραστούν σε αριθμητικά καθώς αυτή η μορφή απαιτείται σαν εισαγωγή. Στα πειράματα πραγματοποιήθηκε η σύγκριση μεταξύ τριών διαφορετικών τύπων αριθμητικής κωδικοποίησης του κάθε SNP. Μέσα από αυτή τη σύγκριση, επιδιώκεται να απαντηθεί το ερώτημα αν ο τρόπος κωδικοποίησης επηρεάζει την πρόβλεψη του χαρακτηριστικού μέσα από το γενότυπο και τελικώς ποιος είναι ο βέλτιστος τρόπος κωδικοποίησης.

Η *genotypic model* κωδικοποίηση κατάφερε να πετύχει καλύτερες ακρίβειες από αυτές του *additive model*, αλλά με μικρές διαφορές. Υποθετικά, η αναπαράσταση του κάθε SNP από 3 κατηγορίες θα μπορούσε να χτίσει πιο λεπτομερή μοντέλα. Παρόλα αυτά, το γεγονός ότι το προσθετικό μοντέλο αναφέρεται στον αριθμό των υπολειπόμενων στο διπλότυπο δείχνει να ανταποκρίνεται στην πραγματικότητα και στο γενετικό μοντέλο ρίσκου που διέπει το φαινότυπο προς εξέταση. Ο αριθμός των υπολειπόμενων δρα προσθετικά στην επιρροή και συνεπώς τα δυο υπολειπόμενα αποκτούν διπλάσια αξία επιρροής συγκριτικά με την ύπαρξη του ενός. Οπότε, η *genotypic* κωδικοποίηση παρότι πιο λεπτομερής ήταν το σχεδόν το ίδιο αποτελεσματική με το προσθετικό μοντέλο με ουσιαστικές όμως διαφορές σε άλλους τομείς που θα αναφερθούν στη συνέχεια

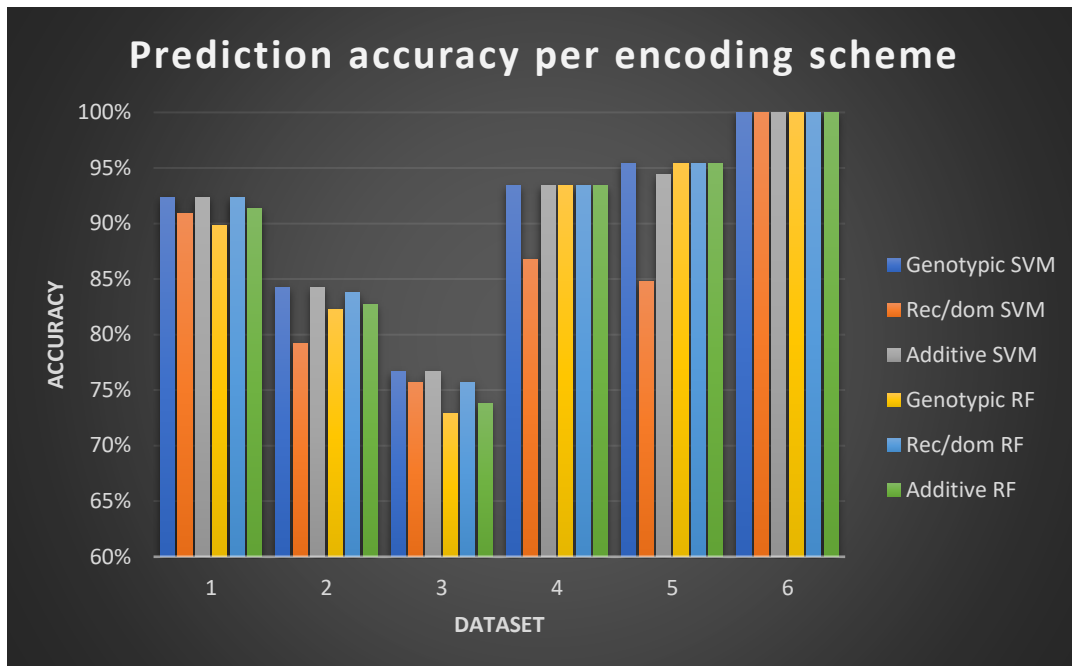
Σχετικά με την *recursive-dominant* κωδικοποίηση, παρατηρήθηκε η κυριαρχία της στα μοντέλα πρόβλεψης εκείνα που σχετίζονται με δέντρα αποφάσεων. Συγκριτικά με τα υπόλοιπα μοντέλα, στον RF ταξινομητή φαίνεται να ανταποκρίνεται εμφανώς καλύτερα η συγκεκριμένη κωδικοποίηση. Σε αντίθεση όμως, στον SVM ταξινομητή τα πράγματα αντιστρέφονται. Σε όλα τα πειράματα η *recursive-dominant* κωδικοποίηση αποτυγχάνει να είναι ανταγωνιστική συγκριτικά με τις άλλες μεθόδους, απέχοντας αρκετή απόσταση στα ποσοστά ακρίβειας. Επίσης, παρότι μπορεί ο καθένας με βάση τα πειράματα να θεωρήσει ότι η *recursive-dominant* κωδικοποίηση ενδείκνυται για ταξινομητές δέντρων αποφάσεων, το *additive model* μέσω του SVM πάντα κατάφερε να πιάσει την μέγιστη ακρίβεια για το κάθε σετ δεδομένων.

Έχοντας εξετάσει τις ιδιαιτερότητες του κάθε μοντέλου κωδικοποίησης, μπορεί να προκύψει το γενικό συμπέρασμα ότι το *genotypic* μοντέλο κωδικοποίησης είναι το πιο

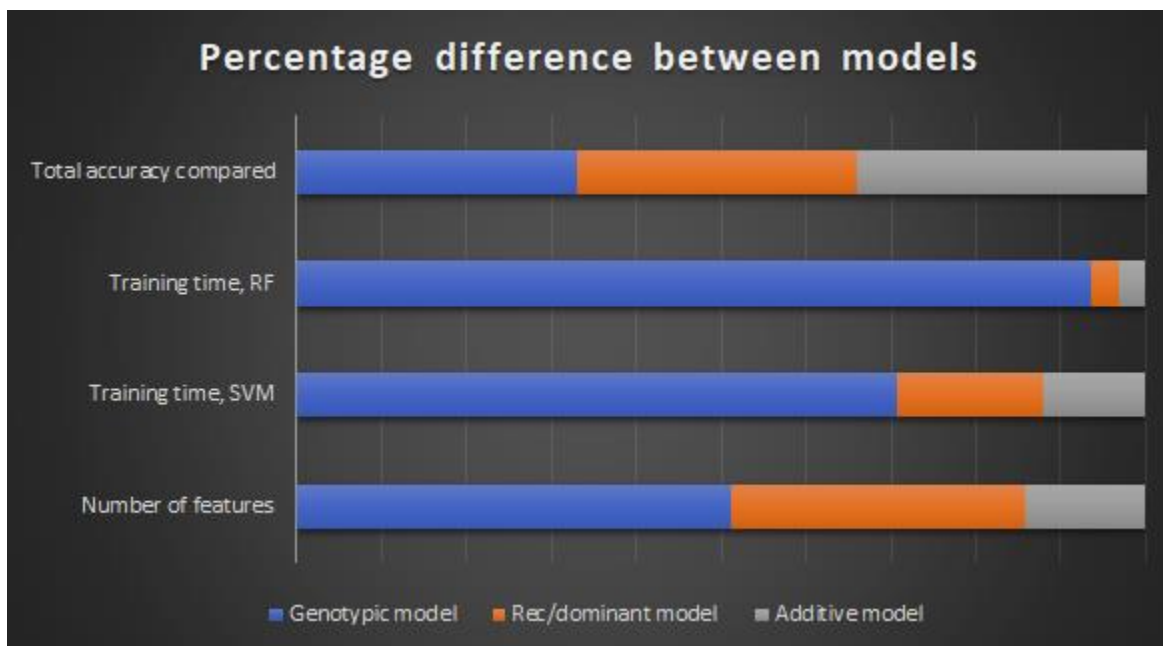
λεπτομερές και δίνει τις καλύτερες ακρίβειες, με πολύ κοντινό ανταγωνιστή το additive model. Αν όμως κανείς αναλογιστεί τους χρόνους εκπαίδευσης των ταξινομητών και γενικότερα το αρκετά μεγαλύτερο υπολογιστικό κόστος που απαιτείται για την επεξεργασία αυτού του μοντέλου σε σχέση με την επιπλέον απόδοση που προσδίδει, τότε το additive model μπορεί ξεκάθαρα να θεωρηθεί το καταλληλότερο μοντέλο. Με ουσιαστικά παρόμοια ακρίβεια και διαφορές της τάξης του ένα τοις εκατό, το additive μοντέλο είναι πολύ πιο ελαφρύ, εύχρηστο και κατάλληλο για την επεξεργασία τέτοιου είδους δεδομένων. Δεδομένα με τεράστιους χώρους χαρακτηριστικών και συνήθως λίγα συγκριτικά δείγματα απαιτούν το additive model για τον ελάχιστο δυνατό χώρο χαρακτηριστικών που δημιουργεί, για την ταχύτητα των υπολογισμών και επιπλέον για την αποφυγή του overfit που οι τεχνικές Μηχανικής Μάθησης είναι επιρρεπείς.

Επιπρόσθετα, με βάση τα πειράματα μπορούν να γίνουν δυο παρατηρήσεις. Η πρώτη παρατήρηση είναι πως το σετ δεδομένου που σχετίζεται με τον διαβήτη τύπου Β εμφανίζει απόλυτες ακρίβειες και δεν είναι χρήσιμο για την σύγκριση των μεθόδων. Ακόμα, σε επίπεδο ταξινομητών φαίνεται πως ο SVM μπορεί να ανταποκριθεί καλύτερα στο πρόβλημα της εύρεσης του φαινοτύπου σε σχέση με τον RF , που παρουσιάζει μία αστάθεια στην ακρίβεια των προβλέψεων.

Εδώ αξίζει να αναφερθεί πως το γεγονός ότι στα πειράματα που πραγματοποιήθηκαν επικράτησε το προσθετικό μοντέλο δεν συνεπάγεται ότι θα επικρατεί πάντα και σε όλες τις περιπτώσεις. Προφανώς τα αποτελέσματα αυτά είναι ανάλογα του αλγορίθμου πρόβλεψης, των δεδομένων( διαθέσιμα SNPs), του χαρακτηριστικού προς εξέταση αλλά και τον ορισμό του προβλήματος (στην συγκεκριμένη περίπτωση δυαδική ταξινόμηση). Ενδέχεται αν εξεταζόταν ένα άλλο χαρακτηριστικό, η επιρροή του κάθε SNP να ανταποκρινόταν σε ένα διαφορετικό μοντέλο με διαφορετικά πειραματικά αποτελέσματα. Γενικότερα, το γεγονός ότι το Additive model φαίνεται να είναι το επικρατέστερο και καταλληλότερο μοντέλο, δεν σημαίνει απαραίτητα ότι και το γενετικό μοντέλο ρίσκου που διέπει τα δεδομένα είναι προσθετικό. Μπορεί να γίνει η παραδοχή πως το Additive μοντέλο είναι το καταλληλότερο για την χρήση τεχνικών Μηχανικής Μάθησης.



Εικόνα 39: Αποτελέσματα περιεργμάτων κωδικοποίησης



Εικόνα 40: Σύγκριση σχημάτων κωδικοποίησης



## Επιλογή των SNPs

Η μέθοδος PRS είναι μία μέθοδος που απαιτεί την επιλογή των στατιστικά σημαντικών SNPs προκειμένου να λειτουργήσει αποτελεσματικά. Οπότε, χρησιμοποιώντας κάποιο στατιστικό τεστ επιλέγεται ένας αριθμός SNP με βάση ένα κατώφλι της  $p$ -value τιμής. Έτσι κατασκευάζονται διάφορα σετ δεδομένων με τα SNPs εκείνα που επιλέγονται ανάλογα την  $p$ -value που χρησιμοποιείται σαν κατώφλι. Στην συνέχεια συγκρίνονται τα αποτελέσματα της μεθόδου PRS σε όλα τα σετ δεδομένων προκειμένου να επιλεχθεί το μοντέλο που περιγράφει καλύτερα το χαρακτηριστικό προς εξέταση και συνεπώς να επικρατήσει το αντίστοιχο κατώφλι εκείνο με την μεγαλύτερη αποτελεσματικότητα. Έχοντας ξεχωρίσει το κατάλληλο κατώφλι και τα σημαντικά SNPs για το χαρακτηριστικό προς εξέταση, μπορεί η πρόβλεψη να γίνει και σε ένα άγνωστο δείγμα με τις ίδιες παραμέτρους (βάρος των SNPs, κατώφλι  $p$ -value).

Στην προσέγγιση με Μηχανική Μάθηση, έγινε η εξής προσπάθεια: Από την μία πλευρά να προσδιοριστεί το αν μειώνοντας τα χαρακτηριστικά των δειγμάτων μπορούν να προκύψουν καλύτερα αποτελέσματα στην ακρίβεια της ταξινόμησης, και από την άλλη αν άλλες τεχνικές μείωσης των διαστάσεων – μείωσης των χαρακτηριστικών μπορούν να βελτιώσουν ακόμα παραπάνω τα αποτελέσματα σε σχέση με τα στατιστικά τεστ. Πιθανώς αυτές οι τεχνικές της επιλογής χαρακτηριστικών να μπορούν να αντικαταστήσουν την επιλογή SNPs με το στατιστικό τεστ στις PRS μεθόδους, αλλά και να διερευνηθεί η αποτελεσματικότητά τους συγκριτικά με το στατιστικό τεστ στις τεχνικές Μηχανικής Μάθησης.

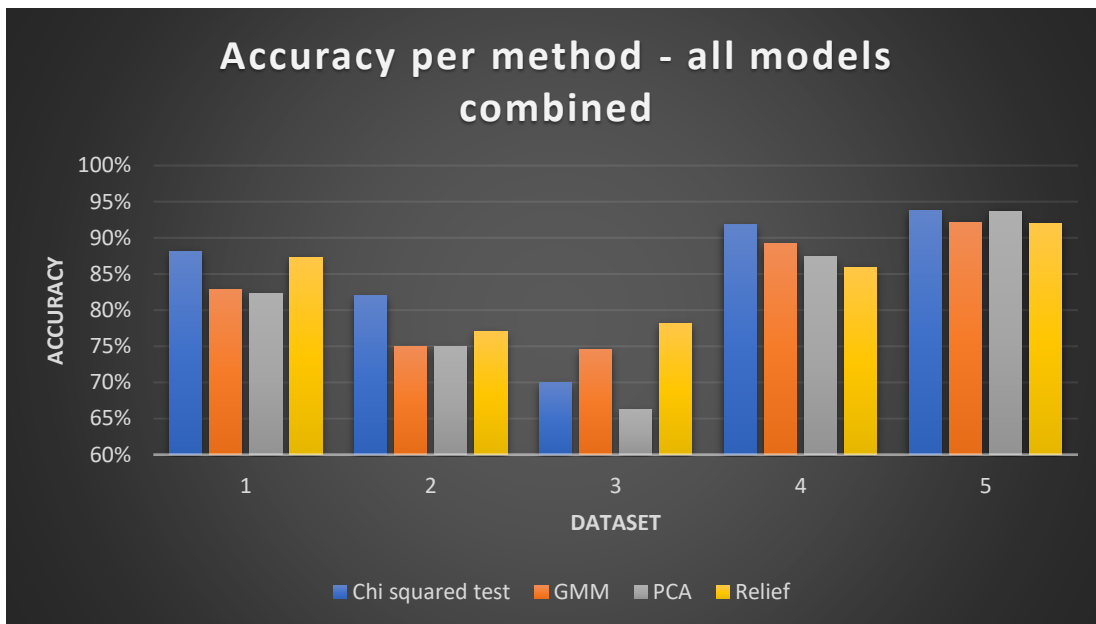
Αρχικά, μπορεί να θεωρηθεί πως το στατιστικό τεστ  $\chi^2$  είναι μία ασφαλής μέθοδος επιλογής SNPs με βάση τα πειράματα. Το γεγονός ότι χρησιμοποιείται στις PRS μεθόδους δεν σημαίνει ότι δεν θα έχει αποτελεσματικότητα και στις μεθόδους Μηχανικής Μάθησης. Σε μία συνολική εικόνα επικρατεί έναντι των υπολοίπων μεθόδων οδηγώντας στις υψηλότερες συγκεντρωτικές ακρίβειες.

Στο τρίτο σετ δεδομένων, στις συγκριτικές αλλά και στις μέγιστες ακρίβειες φαίνεται πως οι πιο περίπλοκες μέθοδοι ανταποκρίνονται αρκετά καλύτερα σε σχέση με το στατιστικό τεστ. Την κυριαρχία του επίσης χάνει το στατιστικό τεστ στην περίπτωση του πέμπτου σετ δεδομένων, όπου την μέγιστη ακρίβεια παρουσιάζει ένα μοντέλο της GMM μεθόδου με μικρή διαφορά.

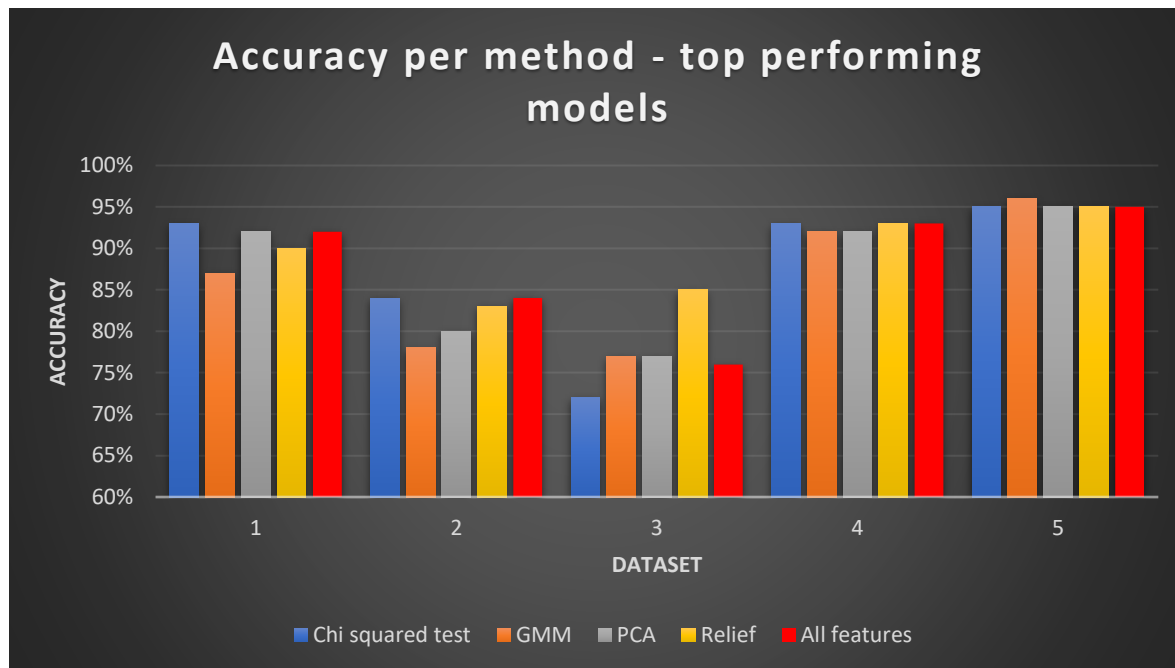


Αξίζει να αναφερθεί πως το τρίτο σετ δεδομένων έχει σημειώσει τις μικρότερες ακρίβειες, γεγονός που μπορεί να το χαρακτηρίσει ως το 'δύσκολο' ανάμεσα στα σετ. Σε αυτό το δύσκολο σετ οι πιο περίπλοκες μέθοδοι καταφέρνουν να δώσουν μία καλύτερη λύση στο πρόβλημα της ταξινόμησης, σε σχέση με το στατιστικό τεστ αλλά και με την χρήση όλων των SNPs. Ενώ συγκριτικά με όλες τις μεθόδους το στατιστικό τεστ επικρατεί πλην του τρίτου σετ δεδομένων, και η αντίστοιχη μέγιστη ακρίβεια πάντα είναι στα ίδια επίπεδα με την ακρίβεια της χρήσης όλων των χαρακτηριστικών ξανά εκτός από το τρίτο σετ δεδομένων. Το τρίτο σετ δεδομένων επιβεβαιώνει την δυσκολία του.

Η μέθοδος PCA αν και αρκετά χρήσιμη σε διάφορες καταστάσεις και προβλήματα σχετιζόμενα με την Μηχανική Μάθηση, δεν φάνηκε χρήσιμη στα συγκεκριμένα πειράματα. Η μέθοδος Relief που δρα με στόχο την εξερεύνηση των συσχετίσεων μεταξύ των SNPs, πέτυχε την καλύτερη λύση στο δύσκολο σετ δεδομένων και δείχνει πως πιθανώς η ανακάλυψη των συσχετισμών μεταξύ των SNPs να μπορεί να επιφέρει μεγαλύτερη προβλεπτική δύναμη από τον προσδιορισμό της σημαντικότητας του κάθε SNP ξεχωριστά. Τέλος, η GMM μέθοδος δείχνει μία αστάθεια στα αποτελέσματα αναλόγως του σετ δεδομένων, παρόλο που τα σετ αυτά αναφέρονται στο κοινό χαρακτηριστικό του χρώματος των ματιών.



Εικόνα 41: Συγκριτική ακρίβεια ανά μέθοδο



Εικόνα 42: Μέγιστη ακρίβεια ανά μέθοδο

### Επιλογή Χαρακτηριστικών SNPs

Στη συνέχεια έγινε η προσπάθεια να συγκριθούν τα σημαντικά εκείνα γονίδια που έχουν αναφερθεί στην βιβλιογραφία για το χρώμα των ματιών, σε σχέση με αυτά που θεωρούν ως σημαντικά οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν στα προηγούμενα πειράματα. Εξετάζοντας τα σετ δεδομένων, παρατηρήθηκαν 3 διαφορετικά γονίδια που έχουν αναφερθεί στην βιβλιογραφία ως σημαντικά για το συγκεκριμένο φαινότυπο. Τα πειράματα σχετίζονται με το αν άλλες μέθοδοι εκτός των GWAS μπορούν να εντοπίσουν σημαντικά γονίδια και αν οι εντοπισμοί αυτοί έχουν κοινό τόπο.

Η μέθοδος GMM είναι η μόνη μέθοδος από τις εξεταζόμενες μεθόδους όπου από τα πρώτα κιάλας 10 χαρακτηριστικά καταφέρνει να εντοπίσει και τα 3 γονίδια που

χαρακτηρίζονται ως σημαντικά. Συγκριτικά με την μέθοδο Relief και με εξαίρεση σε ένα σετ δεδομένων, η μέθοδος GMM παρότι έχει αντιληφθεί τα σημαντικά γονίδια δεν καταφέρνει να ξεπεράσει σε ακρίβεια στα πειράματα την μέθοδο Relief. Η μέθοδος Relief επικρατεί της μεθόδου GMM δίχως να χρησιμοποιεί τα σημαντικά αυτά γονίδια, παρά μόνο ένα από αυτά. Μέχρι και στα 1000 χαρακτηριστικά όπου όλες οι υπόλοιπες μέθοδοι έχουν συμπεριλάβει το σύνολο των σημαντικών γονιδίων, η μέθοδος Relief σταθερά αφήνει εκτός επιλογής τα δυο από τα τρία σημαντικά γονίδια. Κατά τα άλλα δεν εντοπίζεται κάποια ιδιαίτερη διαφορά μεταξύ των ταξινομητών RF και SVM.

Σε συνδυασμό με τα προηγούμενα συμπεράσματα, παρατηρείται πως ενώ η ακρίβεια που σημειώνεται από κάθε μέθοδο είναι πολλές φορές παρόμοια ή αρκετά κοντά, ο τρόπος με τον οποίο λειτουργούν και παράγουν τις ακρίβειες αυτές διαφέρει. Στα συγκεκριμένα πειράματα εξετάστηκαν οι ταξινομητές SVM και RF, καθώς και δυο μέθοδοι επιλογής χαρακτηριστικών. Επίσης έγινε σύγκριση πάνω στις ακρίβειες που σημείωσαν στο πρόβλημα της δυαδικής ταξινόμησης του φαινοτύπου. Αξίζει να σημειωθεί πως και στην περίπτωση των μεθόδων επιλογής χαρακτηριστικών, μετά την επιλογή αυτή η τελική ταξινόμηση πραγματοποιήθηκε με τους προαναφερμένους ταξινομητές Μηχανικής μάθησης.

Η βιβλιογραφία αναφέρεται συνήθως στα γονίδια εκείνα που προκύπτουν ως καθοριστικά για κάποιο χαρακτηριστικό με βάση τις GWAS. Στην περίπτωση της Μηχανικής μάθησης και των περίπλοκων μοντέλων δεν είναι αναμενόμενο να επιβεβαιωθεί η σημαντικότητα αυτών των γονιδίων. Κάθε μοντέλο λειτουργεί διαφορετικά και ανακαλύπτει τα δικά του πρότυπα μέσα στα δεδομένα. Οι διαφορές που μπορούν να προκύψουν εμφανίζονται όχι μόνο μεταξύ PRS μεθόδων και μεθόδων Μηχανικής μάθησης, αλλά και εσωτερικά μεταξύ των μεθόδων της Μηχανικής Μάθησης.

Συνεπώς επιβεβαιώνεται πως η διαφορετικότητα της κάθε μεθόδου επηρεάζει τον τρόπο με τον οποίο χαρακτηρίζονται τα διάφορα γονίδια ως καθοριστικά για κάποιο χαρακτηριστικό. Υπάρχει η δυνατότητα να προκύψουν παρόμοια αποτελέσματα δίνοντας βάρος σε διαφορετικά SNPs και αυτό εξηγείται από την εξάρτηση των προς μελέτη χαρακτηριστικών όχι μόνο από τα SNPs αλλά και από τις αλληλεπιδράσεις μεταξύ τους. Δηλαδή ενδέχεται η αλληλεπίδραση 2 διαφορετικών SNPs να δίνει περισσότερη πληροφορία για την πρόβλεψη από την ύπαρξη μόνο και μόνο ενός σημαντικού SNP, και αυτό το στοιχείο μπορεί να εντοπισθεί από τις μεθόδους της Μηχανικής Μάθησης.

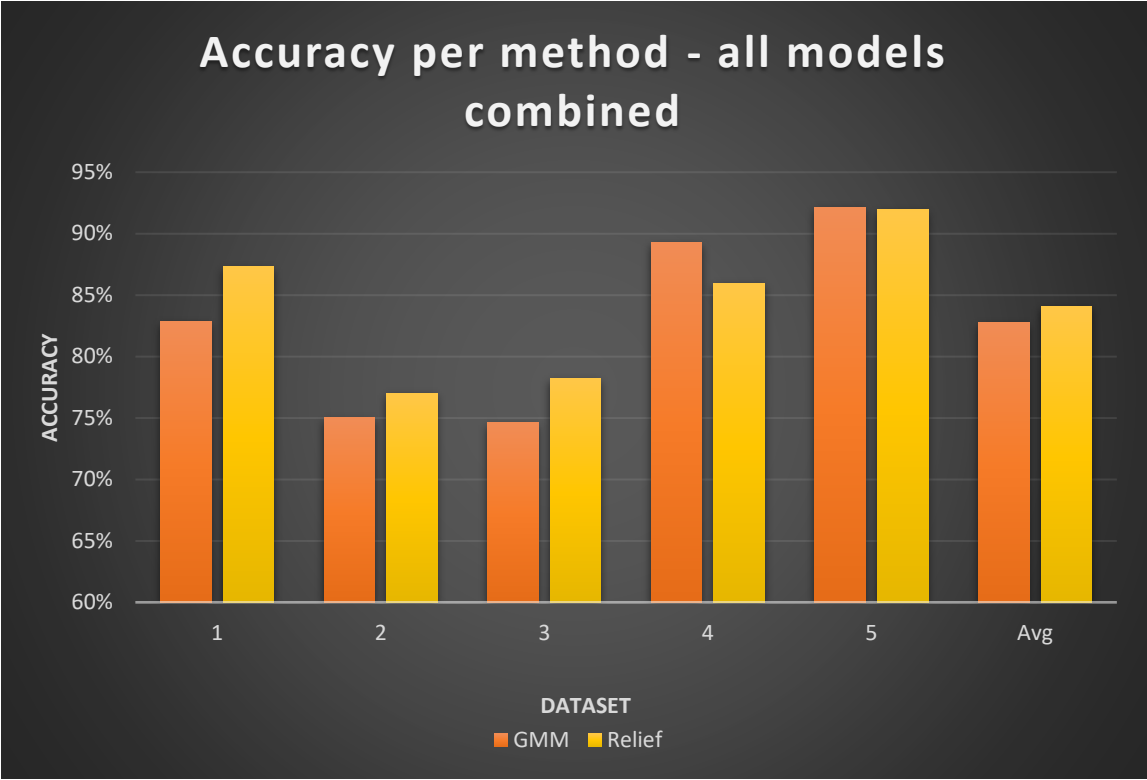
Η ερευνητική κοινότητα δείχνει όλο και περισσότερο την ύπαρξη αυτών των αλληλεπιδράσεων και οι μέθοδοι μηχανικής μάθησης είναι ικανοί να τις

μοντελοποιήσουν. Ακόμα βρίσκονται σε εξέλιξη και μέθοδοι PRS που θεωρούν ως δεδομένα συγκεκριμένα pathways, δηλαδή μονοπάτια από ομάδες SNPs που με την αλληλεπίδραση τους επηρεάζουν το αποτέλεσμα περισσότερο από ένα μοναδικό SNP.

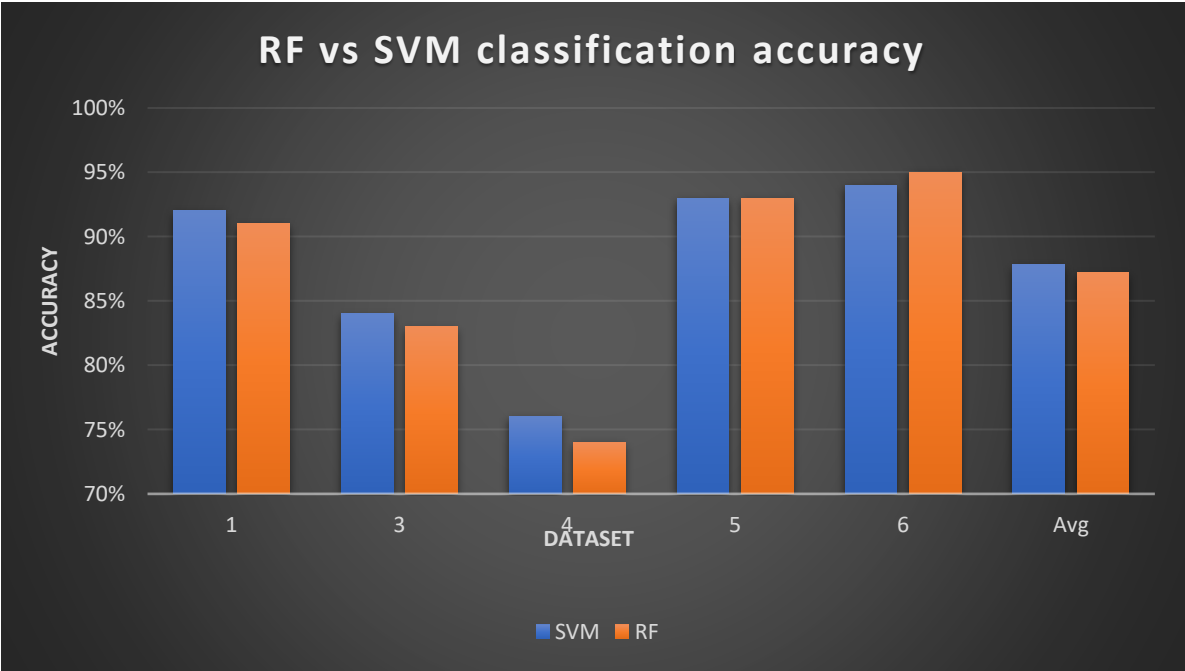
Έχοντας υπόψιν ότι τα αποτελέσματα αυτά αναφέρονται σε σετ δεδομένων που παρότι δεν είναι ίδια αναφέρονται σε ένα κοινό χαρακτηριστικό, μπορεί κανείς να αντιληφθεί πόσο πολύπλοκο είναι το πρόβλημα της πρόβλεψης του φαινοτύπου μέσω του γενοτύπου. Η πολυπλοκότητα αυτή φαίνεται από το πόσο διαφορετικά προσεγγίζει κάθε αλγόριθμος το πρόβλημα, από τα μη σταθερά αποτελέσματα σε παρόμοια σετ δεδομένων αλλά και από την γενικότερη δυσκολία της πρόσβασης σε μεγάλο και καθαρό όγκο δεδομένων.



Εικόνα 43: Σημαντικά SNPs ανά μέθοδο, συγκριτικά με την βιβλιογραφία



Εικόνα 44: Σύγκριση ακρίβειας στις προβλέψεις, GMM vs Relief



Εικόνα 45: Σύγκριση ακρίβειας στις προβλέψεις, RF vs SVM

Η μέθοδος PRS είναι μία μέθοδος που αξιολογεί την τάση ενός ατόμου προς το χαρακτηριστικό προς εξέταση με βάση τα γενετικά του δεδομένα. Παράγεται ένα σκορ το οποίο αντικατοπτρίζει την τάση αυτή. Στην περίπτωση της ταξινόμησης σε κατηγορίες case-control, δηλαδή σε έχοντες και μη έχοντες το προς μελέτη χαρακτηριστικό, παράγεται απλά μια πρόβλεψη της κατηγορίας και όχι κάποιο σκορ. Προκειμένου να συγκριθούν αυτές οι 2 προσεγγίσεις, έγιναν κάποιες επεξεργασίες στις μεθόδους ώστε να έχουν κοινό σημείο αναφοράς.

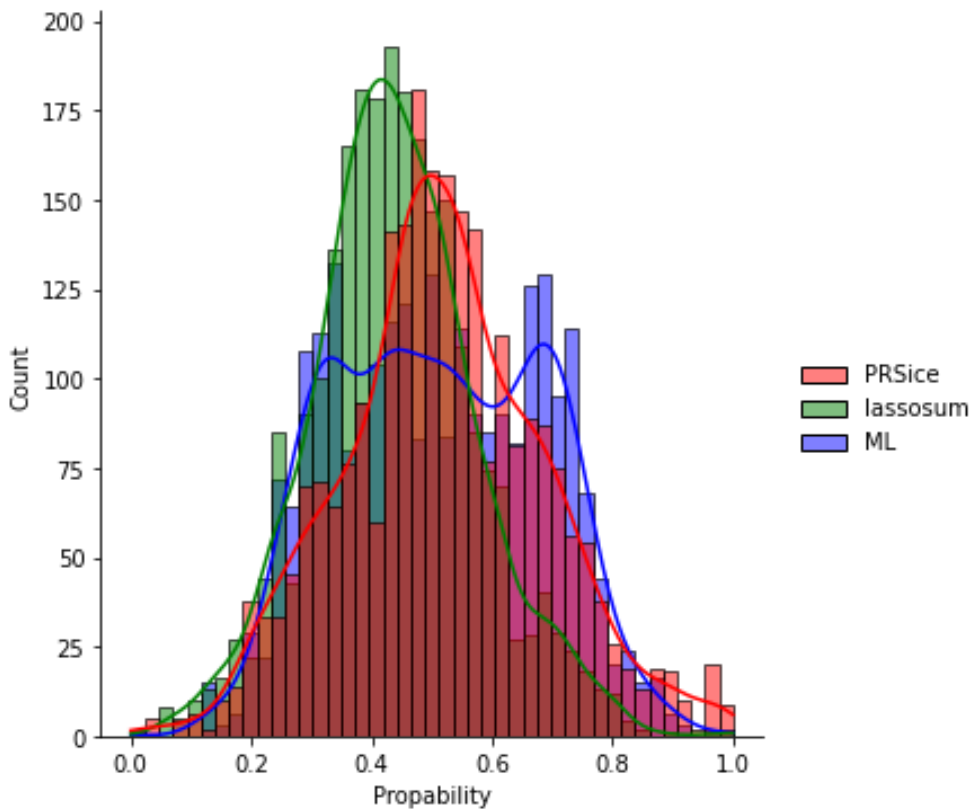
Παρατηρείται πως η διαφορά στην αποτελεσματικότητα είναι αρκετά μεγάλη ανάμεσα στις PRS μεθόδους και την εκδοχή της Μηχανικής Μάθησης. Αρχικά, τα διαφορετικά στρώματα που χρησιμοποιήθηκαν στην εκδοχή αυτή και οι περισσότερες παράμετροι προς εκπαίδευση τελικώς δεν αύξησαν την ακρίβεια των προβλέψεων, δείχνοντας πως ότι πρότυπα υπήρχαν στα δεδομένα έχουν αφομοιωθεί από το πρώτο και ελαφρύτερο κιάλας δίκτυο. Λαμβάνοντας υπόψιν τον τρόπο λειτουργίας των μεθόδων, δεν θα είχε νόημα να γίνει κάποια άλλη σύγκριση σε μεγαλύτερο αριθμό SNPs καθώς δεν θα είχε νόημα η σύγκριση με την PRS μέθοδο που δουλεύει σε υποσύνολα αυτών. Πιθανώς με περισσότερα SNPs τα δίκτυα να έβρισκαν περισσότερα πρότυπα και να πετύχαιναν ακόμα καλύτερες ακρίβειες.

Στα 10000 δείγματα προς εκπαίδευση και δυνητικά παραπάνω, δεδομένου ότι αυτά αποτελούν συνθετικά δεδομένα, είναι θεμιτό να χρησιμοποιηθεί ένας μεγαλύτερος χώρος χαρακτηριστικών χωρίς τον κίνδυνο του overfitting. Τα συνθετικά δεδομένα όμως παραμένουν συνθετικά, και η αυξημένη ακρίβεια στην πρόβλεψη δεν εγγυάται ένα μοντέλο που θα γενικεύει αποτελεσματικά στα μη-γνωστά και πραγματικά δεδομένα.

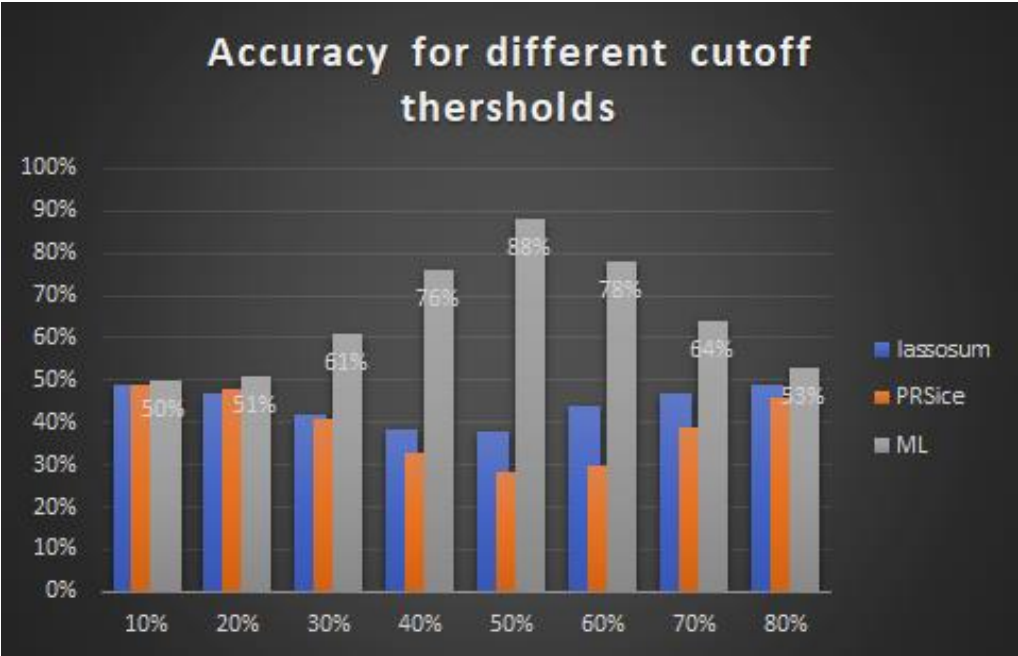
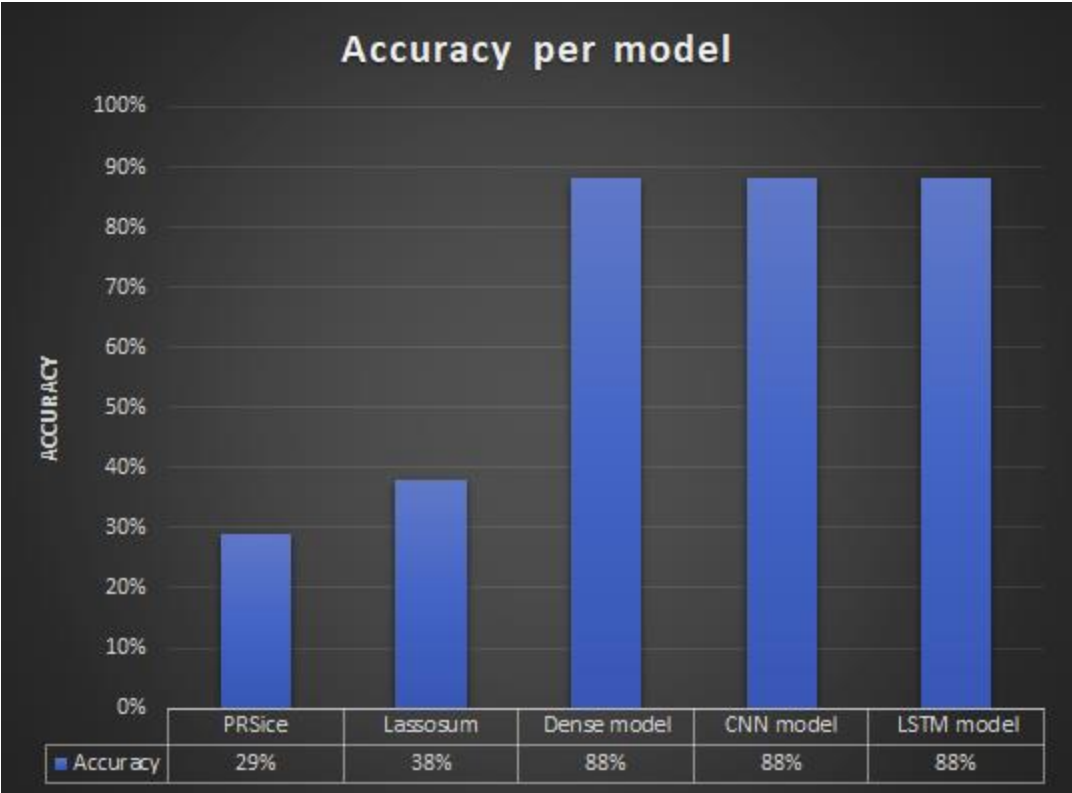
Γενικότερα οι PRS μέθοδοι είναι σχεδιασμένες έτσι ώστε να παράγουν ένα σκορ, πάνω στην λογική της πρόληψης. Τα γενετικά δεδομένα είναι συνήθως λίγα ως ευαίσθητα και αν υπάρχει τελικά πρόσβαση σε μεγάλο όγκο δεδομένων, αυτή αφορά συγκεκριμένα χαρακτηριστικά-ασθένειες. Οπότε, αυτή η λογική του σκορ αντικατοπτρίζει και όλες τις ασάφειες ή τα σφάλματα που μπορεί το σκορ να εμπεριέχει. Τα αποτελέσματα των μεθόδων PRS κατανέμονται ομαλά γύρω από μία τιμή περίπου στην μέση του διαστήματος. Παρατηρείται ότι η lasso εκδοχή είναι αποτελεσματικότερη της PRSice στο συγκεκριμένο πείραμα.

Μεταλλάσσοντας το ζητούμενο σε μία ξεκάθαρη πρόβλεψη ταξινόμησης, φαίνονται οι αδυναμίες της μεθόδου. Τα σκορ αυτά είναι εύκολο να παραχθούν και δεν απαιτούν κάποιο μεγάλο υπολογιστικό κόστος. Από την άλλη πλευρά το μοντέλο παλινδρόμησης που χρησιμοποιούν υποθέτει ότι το κάθε SNP συνεισφέρει ατομικά και μόνο στο μοντέλο και ότι τα δεδομένα που διαχειρίζεται είναι ασυσχέτιστα και ακολουθούν την κανονική κατανομή. Αυτοί οι ισχυρισμοί μπορεί και να μην ισχύουν στις γενετικές δομές, οδηγώντας σε μη αντικειμενικά συμπεράσματα. Γι' αυτό τον λόγο έχουν την έκφραση ενός σκορ και αντιμετωπίζονται μόνο ως υποστηρικτικά βοηθήματα από τους ειδικούς.

Στην περίπτωση των δικτύων της Μηχανικής μάθησης και του προβλήματος της ταξινόμησης, τα αποτελέσματα δείχνουν το πόσο αποτελεσματικά είναι τα δίκτυα αυτά στην μοντελοποίηση του προβλήματος σε πολυδιάστατα δεδομένα όπως το γονότυπο. Τα μοντέλα αυτά είναι σε θέση να αντιληφθούν όλους τους πολύπλοκους συσχετισμούς που παρουσιάζονται στα δεδομένα χωρίς συμβάσεις και αντίστοιχες υποθέσεις με τις μεθόδους PRS. Η κατανομή τους φαίνεται να έχει δυο κορυφές διαχωρίζοντας τα case-controls.



Εικόνα 46: Κατανομή των προβλέψεων





Πολύ σημαντικό στοιχείο είναι ότι η κυριαρχία της ακρίβειας της μεθόδου της Μηχανικής μάθησης δεν εξαρτάται από το κατώφλι του διαχωρισμού της πιθανότητας σε case-control, όπως φαίνεται στο διάγραμμα που ακολουθεί. Θα μπορούσε κάποιος να υποθέσει ότι τα PRS σκορ είναι αποτελεσματικά αν βγει από την εξίσωση το κατώφλι του διαχωρισμού, αλλά στην πραγματικότητα και με διαφορετικά κατώφλια διαχωρισμού ώστε να προσαρμοστούν καλύτερα στις κατανομές των δυο μεθόδων PRS τα αποτελέσματα δείχνουν ότι η μέθοδος ML ανταποκρίνεται καλύτερα.

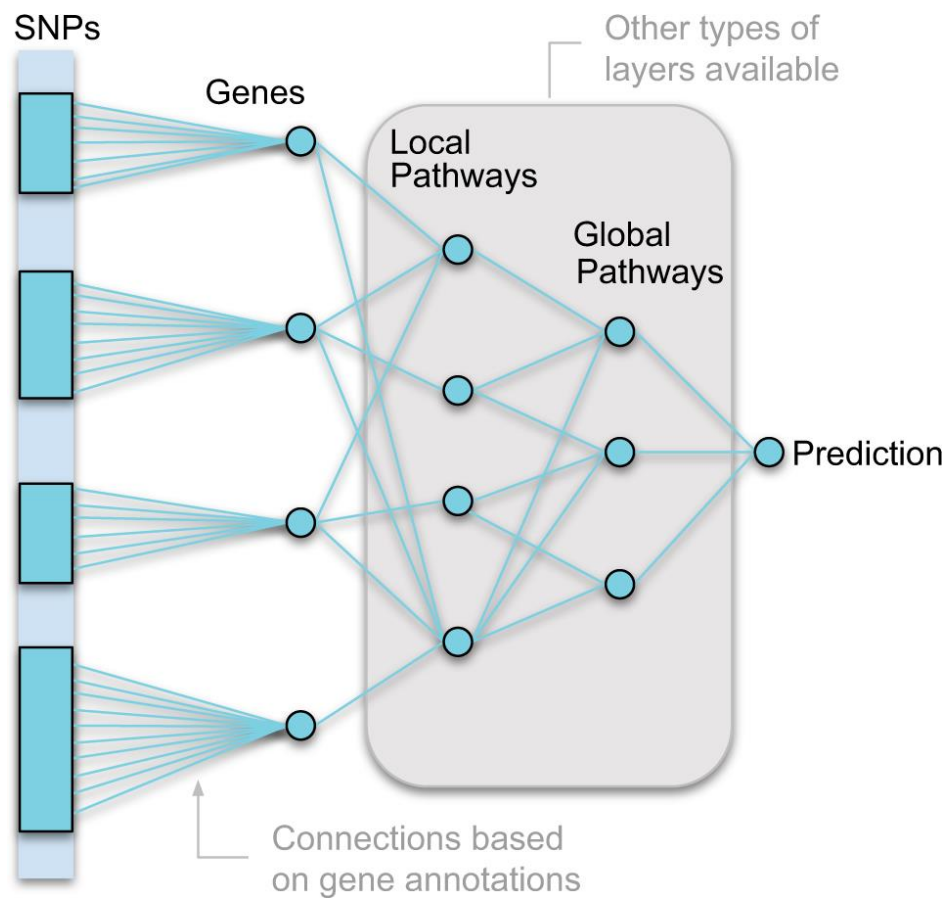
Ο κίνδυνος στην περίπτωση της Μηχανικής Μάθησης είναι το κατά πόσο αποτελεσματικά μπορούν να γενικεύσουν αυτά τα μοντέλα σε πραγματικά δεδομένα. Σίγουρα η προβλεπτική ικανότητα είναι τεράστια, αλλά όπως και σε εφαρμογές παρόμοιων δικτύων σε άλλους επιστημονικούς κλάδους απαιτούνται πάρα πολλά δεδομένα ώστε τα δίκτυα αυτά να εκπαιδευτούν σωστά και να μην προσαρμόζονται στα δεδομένα εκπαίδευσης. Τέλος ένα άλλο πλεονέκτημα των μοντέλων βαθιάς μάθησης είναι ότι επιφέρουν αυτά τα αποτελέσματα παρακάμπτοντας διάφορα βήματα επεξεργασίας, συνεισφέροντας σε μία end-to-end λύση σε σχέση με τις PRS μεθόδους.

Οπότε στο σημείο αυτό αξίζει να αναφερθεί ένα άλλο πλεονέκτημα της μεθόδου ML. Εκτός από τις αυξημένες δυνατότητες στην ακρίβεια παρατηρούνται κι άλλες δυνατότητες που μπορούν να αξιοποιηθούν. Αρχικά μια ML λύση μπορεί να χρησιμοποιηθεί σαν μια end-to-end λύση αφού μπορεί να χειριστεί αυτομάτως κάποιες διαδικασίες όπως για παράδειγμα την ύπαρξη διπλότυπων. Ένα νευρωνικό δίκτυο δεν έχει αυστηρή απαίτηση την αφαίρεση των διπλότυπων SNPs όπως μία PRS μέθοδος, καθώς η ίδια η εκπαίδευση του θα το κάνει να ξεχωρίζει το πόσο σημαντικό είναι το κάθε SNP. Μετέπειτα, όροι και τεστ όπως ο HWE, LD και MAF μπορούν να παρακαμφθούν για τον ίδιο ακριβώς λόγο. Τέλος στο πινακάκι από κάτω φαίνονται αυτές οι διαφορές μαζί με τις παραδοχές που συνοδεύουν τις PRS μεθόδους ενώ δεν απαιτούνται από τις ML μεθόδους.

Τέλος, από την φύση της η PRS μέθοδος δεν μπορεί να μοντελοποιήσει τις μη γραμμικές σχέσεις μεταξύ των SNPs. Σε αντίθεση, οι τεχνικές μηχανικής μάθησης έχουν αυτή τη δυνατότητα. Με βάση τις πρόσφατες ερευνητικές αναφορές, αυτοί οι συσχετισμοί υπάρχουν στην γενετική δομή και φαίνεται να επηρεάζουν αρκετά τα χαρακτηριστικά του ατόμου. Η έννοια των μονοπατιών (pathways) των SNPs, αναφέρεται στο γεγονός ότι συγκεκριμένοι συνδιασμοί SNPs μπορούν να συνεισφέρουν παραπάνω από ότι άλλοι συνδιασμοί SNPs.

Critical steps-assumptions	PRS	ML
QC (duplicates, sample overlap)	✓	✗
Minor allele frequency	✓	✗
Hardy-Weinberg Equilibrium	✓	✗
Linkage Disequilibrium	✓	✗
P-value threshold dependance	✓	✗
Linear relationship of SNPs	✓	✗
Independence of SNPs	✓	✗
Normal distrubution of underlying data	✓	✗

Εικόνα 47: Διαφορές ML-PRS



Εικόνα 48: Μοντελοποίηση interactions

## Επεξηγησιμότητα των μεθόδων

Ξεφεύγοντας από την ανάλυση των αποτελεσμάτων και εξετάζοντας πρακτικά τις 2 αυτές μεθόδους, στην περίπτωση σοβαρών θεμάτων υγείας ο καθένας θα προτιμούσε να βοηθηθεί από κάποια μέθοδο PRS σε συνδυασμό με την γνώμη των ειδικών. Η αποδεδειγμένα μεγαλύτερη προβλεπτική ικανότητα των μοντέλων βαθιάς μάθησης απαιτεί έναν μεγάλο όγκο δεδομένων που δεν βρίσκεται εύκολα και δεν εγγυάται κανείς για την ποιότητα τους. Μπορεί κανείς να εμπιστευτεί πλήρως τις προβλέψεις αυτές (κίνδυνος *overfitting*) καθώς και την ορθή εκπαίδευση τους? Αν θεωρηθεί ότι τα απαιτούμενα δεδομένα είναι διαθέσιμα και όλες οι δυσκολίες που μπορεί να προκύπτουν για τα μοντέλα βαθιάς μάθησης έχουν αντιμετωπισθεί, θα ήταν οι προβλέψεις έμπιστες?

Στα θέματα υγείας – σε σύγκριση με θέματα πρόβλεψης απλών χαρακτηριστικών – παίζει κρίσιμο ρόλο η επεξηγησιμότητα των προβλέψεων. Αυτό συμβαίνει διότι δεν μπορεί κανείς να εμπιστευτεί σε ένα ‘μαύρο κουτί’ αποφάσεις που θα κοστίσουν την ζωή του. Τα αναδραστικά στρώματα, τα συνελικτικά στρώματα και νευρωνικά δίκτυα γενικότερα δεν είναι εύκολα επεξηγήσιμα, ακόμα και από τους ειδικούς. Η απλότητα των PRS μεθόδων χαρίζει αυτή την απαραίτητη επεξηγησιμότητα και εμπνέει περισσότερη εμπιστοσύνη. Παρόλο που η προβλεπτική ικανότητα της είναι μικρότερη, δεν βασίζεται στην ύπαρξη πολλών δεδομένων για την αντικειμενικότητα της και ακρίβεια της αρκεί για έναν υποστηρικτικό παράγοντα στις αποφάσεις των ειδικών.

## ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

Η παρούσα εργασία είχε σαν σκοπό να εξερευνήσει τις δυνατότητες της Μηχανικής-Βαθιάς Μάθησης πάνω στα γενετικά δεδομένα και σε σύγκριση με τις ήδη καθιερωμένες μεθόδους πρόβλεψης χαρακτηριστικών μέσα από αυτά. Αρχικά έπρεπε να αναζητηθούν δεδομένα προς εκπαίδευση. Τα δεδομένα αυτά προήλθαν από την πύλη OpenSNP, αφού πρώτα πέρασαν αρκετά στάδια προεπεξεργασίας προκειμένου να εξασφαλιστεί η καταλληλότητα τους και να κατασκευαστούν τα σετ δεδομένων. Τα μόνα σετ δεδομένων που περιείχαν αρκετά δείγματα αφορούσαν το χρώμα των ματιών σαν βασικό χαρακτηριστικό. Έτσι δημιουργήθηκαν δυαδικά σετ δεδομένων με αυτό το χαρακτηριστικό.

Στη συνέχεια πραγματοποιήθηκαν διάφορα πειράματα που είχαν να κάνουν με την εφαρμογή της Μηχανικής-Βαθιάς Μάθησης στα γενετικά αυτά δεδομένα. Αρχικά προσδιορίστηκε ότι το προσθετικό μοντέλο κωδικοποίησης είναι το καταλληλότερο μοντέλο για την αριθμητική κωδικοποίηση των δεδομένων. Έπειτα εξετάστηκαν διάφοροι μέθοδοι μείωσης διαστάσεων-επιλογής χαρακτηριστικών και η συμπεριφορά τους, σε αντιστοιχία με το  $p$ -value thresholding στις μεθόδους PRS. Επιπρόσθετα συγκρίθηκαν τα πιο καθοριστικά γονίδια σχετικά με το χρώμα των ματιών που αναφέρει η βιβλιογραφία σε σχέση με αυτά που βρήκαν ως σημαντικότερα οι αλγόριθμοι Μηχανικής Μάθησης. Καταλήγοντας, με μια προσαρμογή των μεθόδων πάνω στο πρόβλημα της δυαδικής ταξινόμησης, συγκρίθηκαν οι μέθοδοι PRS με τα δίκτυα Βαθιάς μάθησης.

Σαν μελλοντικοί στόχοι και επεκτάσεις αυτής της εργασίας μπορούν να θεωρηθούν οι εξής:

- Η χρήση μεγαλύτερων σε αριθμό δειγμάτων και πληρέστερων από την άποψη των SNPs ( imputed ) σετ δεδομένων, καθώς και δεδομένων που αναφέρονται σε άλλα χαρακτηριστικά πέρα από το χρώμα των ματιών
- Η χρήση πιο εξελιγμένων δικτύων βαθιάς μάθησης, όπως attention-based δίκτυα ή Graph neural networks που θα μπορούσαν να μοντελοποιήσουν τις αλληλεπιδράσεις μεταξύ των SNPs. Για την εφαρμογή όλων αυτών, απαιτούνται δεδομένα
- Χρήση unsupervised τεχνικών ώστε από μία αποθήκη γενετικών δεδομένων όπως το OpenSNP να μπορεί κάποιος να εκμεταλλευτεί όλα τα γενετικά δεδομένα χωρίς το

συσχετισμένο φαινότυπο και να πραγματοποιήσει το fine-tuning του δικτύου με τα λίγα διαθέσιμα δεδομένα πετυχαίνοντας πιθανώς καλύτερες ακρίβειες

- Ανάπτυξη καινοτόμων λύσεων πάνω στο κομμάτι της επεξηγησιμότητα, ώστε τα μοντέλα Μηχανικής μάθησης να κατακτήσουν την εμπιστοσύνη της ιατρικής κοινότητας.

## ΑΝΑΦΟΡΕΣ

1. Basic tutorial for polygenic risk score analyses. [https:// choishingwan.github.io/PRS-Tutorial/](https://choishingwan.github.io/PRS-Tutorial/).
2. Hapgen version 2. [https://mathgen.stats.ox.ac.uk/genetics\\_software/hapgen/hapgen2.html](https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html). (Accessed on 03/20/2021).
3. Human genetic variation - wikipedia. [https://en.wikipedia.org/wiki/Human\\_genetic\\_variation](https://en.wikipedia.org/wiki/Human_genetic_variation).
4. Plink - basic tutorial for polygenic risk score analyses. [https:// choishingwan.github.io/PRS-Tutorial/plink/](https://choishingwan.github.io/PRS-Tutorial/plink/).
5. Polygenic risk: What's the score? <https://www.nature.com/articles/d42473-019-00270-w>.
6. Nihad A.M Al-Rashedi, Amar Mousa Mandal, and Laith AH ALObaidi. Eye color prediction using the IrisPlex system: a limited pilot study in the iraqi population. *Egyptian Journal of Forensic Sciences*, 10(1), August 2020.
7. Jahad Alghamdi, Manal Amoudi, Ahmad Ch. Kassab, Mansour Al Mufarrej, and Saleh Al Ghamdi. Eye color prediction using single nucleotide polymorphisms in saudi population. *Saudi Journal of Biological Sciences*, 26(7):1607–1612, November 2019.
8. Emily Baker and Valentina Escott-Price. Polygenic risk scores in alzheimer's disease: Current applications and future directions. *Frontiers in Digital Health*, 2, August 2020.
9. David J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, October 2006.
10. Hyo-Jeong Ban, Jee Yeon Heo, Kyung-Soo Oh, and Keun-Joon Park. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. *BMC Genetics*, 11(1):26, 2010.
11. Luis B Barreiro, Guillaume Laval, H el ene Quach, Etienne Patin, and Llu s Quintana-Murci. Natural selection has driven population differentiation in modern humans. *Nature Genetics*, 40(3):340–345, February 2008.
12. Statistical analysis for genome-wide association study. *Journal of Biomedical Research*, July 2015.
13. Ver nica Bol n-Canedo and Amparo Alonso-Betanzos. Ensembles for feature selection: A review and future trends. *Information Fusion*, 52:1–12, December 2019.
14. Marine S. O. Briec, Charles D. Waters, Daniel P. Drinan, and Kerry A. Naish. A practical introduction to random forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, 18(4):755–766, March 2018.
15. William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822, December 2012.
16. Gina M. Dembinski and Christine J. Picard. Evaluation of the IrisPlex DNA-based eye color prediction assay in a united states population. *Forensic Science International: Genetics*, 9:111–117, March 2014.
17. Alexandre Drouin, Ga el Letarte, Fr d ric Raymond, Mario Marchand, Jacques Corbeil, and Fran ois Laviolette. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific Reports*, 9(1), March 2019.

18. Theodoros Evgeniou and Massimiliano Pontil. Support vector machines: Theory and applications. In *Machine Learning and Its Applications*, pages 249–257. Springer Berlin Heidelberg, 2001.
19. João Fadista, Alisa K Manning, Jose C Florez, and Leif Groop. The (in)famous GWAS p-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8):1202–1205, January 2016.
20. Nastasiya F. Grinberg, Oghenejokpeme I. Orhobor, and Ross D. King. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning*, 109(2):251–277, October 2019.
21. Enzo Grossi and Massimo Buscema. Introduction to artificial neural networks. *European Journal of Gastroenterology & Hepatology*, 19(12):1046–1054, December 2007.
22. Giorgio Guzzetta, Giuseppe Jurman, and Cesare Furlanello. A machine learning pipeline for quantitative phenotype prediction from genotype data. *BMC Bioinformatics*, 11(S8), October 2010.
23. Daniel Sik Wai Ho, William Schierding, Melissa Wake, Richard Saffery, and Justin O’Sullivan. Machine learning SNP based prediction for precision medicine. *Frontiers in Genetics*, 10, March 2019.
24. Lefteris Koumakis. Deep learning models in genomics are we there yet? *Computational and Structural Biotechnology Journal*, 18:1466–1473, 2020.
25. OpenSNP. <https://opensnp.org/>.
26. Silke Szymczak, Joanna M. Biernacka, Heather J. Cordell, Oscar González-Recio, Inke R. König, Heping Zhang, and Yan V. Sun. Machine learning in genome-wide association studies. *Genetic Epidemiology*, 33(S1):S51–S57, 2009.
27. Yannian Wang, Shanshan Liu, Ruoxi Chen, Zhongning Chen, Jinlei Yuan, and Quanzhong Li. A novel classification indicator of type 1 and type 2 diabetes in china. *Scientific Reports*, 7(1), December 2017.