Ελληνικό Ανοικτό Πανεπιστήμιο

Πληροφοριακά Συστήματα

Διπλωματική εργασία

Πολυπαραγοντική μονοκυτταρική ενσωμάτωση με χρήση Μηχανικής Μάθησης

Καραθάνου Θεοδώρα

Επιβλέπων καθηγητής: Ρεφανίδης Ιωάννης

Θεσσαλονίκη, Ιούνιος 2023

Πολυπαραγοντική μονοκυτταρική ενσωμάτωση με χρήση Μηχανικής Μάθησης

Καραθάνου Θεοδώρα

Επιτροπή Επίβλεψης Διπλωματικής Εργασίας

Επιβλέπων καθηγητής:	Συνεπιβλέπων καθηγητής:
Ρεφανίδης Ιωάννης	Δημήτρης Καλλές
Μέλος ΣΕΠ του ΕΑΠ	Μέλος ΣΕΠ του ΕΑΠ

Θεσσαλονίκη, Σεπτέμβριος 2023

Ευχαριστίες

Πρώτα και κύρια θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου για την συγκεκριμένη διπλωματική εργασία, κ. Ιωάννη Ρεφανίδη αρχικά για την ευκαιρία που μου έδωσε να πραγματοποιήσω την ιδέα μου με τα εργαλεία που ήθελα, αλλά και για την στήριξη του καθ' όλη την διάρκεια εκπόνησης της εργασίας, της βοήθειας που μου έδωσε όταν χρειάστηκε και την συνολική εμπιστοσύνη που μου έδειξε. Μαζί με τον συνεπιβλέποντα καθηγητή κ. Δημήτρη Καλλέ, τους ευχαριστώ για την επίβλεψη της συγκεκριμένης διπλωματικής εργασίας.

Ακόμη θα ήθελα να ευχαριστήσω και όλους τους καθηγητές που είχα στις ΘΕ του μεταπτυχιακού.

Τέλος, ευχαριστώ τους γονείς μου, την γιαγιά μου και τον φίλο και συνάδελφό μου Γιώργο, για το κουράγιο και τη στήριξη που μου προσέφεραν κατά τη διάρκεια του μεταπτυχιακού καθώς και για την υπομονή τους.

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει μια αναδυόμενη περιοχή στον τομέα της Βιολογίας που αφορά τη μελέτη μεμονωμένων κυττάρων χρησιμοποιώντας αλγορίθμους Μηχανικής Μάθησης. Αυτή η προσέγγιση έχει επιτρέψει τη συστηματική διερεύνηση της κυτταρικής ανομοιογένειας σε διάφορους ιστούς και πληθυσμούς κυττάρων, προσφέροντας νέες εισαγωγές σχετικά με τη σύνθεση, τη δυναμική και τους ρυθμιστικούς κανόνες των κυτταρικών καταστάσεων κατά τη διάρκεια της ανάπτυξης του ανθρώπινου οργανισμού και των ασθενειών του.

Η ανάλυση των μεμονωμένων κυττάρων προσφέρει σημαντικά οφέλη στους επιστήμονες της Βιολογίας, καθώς αποκαλύπτει την κυτταρική ετερογένεια, αποκωδικοποιεί μηχανισμούς ασθενειών, διευκολύνει την ανακάλυψη φαρμάκων και αναλύει θεμελιώδη βιολογικά ερωτήματα. Επιπλέον, αυτή η προσέγγιση επιτρέπει την παρακολούθηση φαρμάκων και την ανακάλυψη νέων θεραπευτικών επιλογών για διάφορες ασθένειες.

Για την ανάλυση αυτής της μεθοδολογίας, χρησιμοποιούνται ποικίλες τεχνικές, όπως κυτταρομετρία ροής, Single-Cell RNA Sequencing, Single-Cell Proteomics, Single-Cell Epigenomics, και κυτταρομετρία μάζας. Η ανάλυση και η ερμηνεία των πολύπλοκων δεδομένων απαιτούν εξελιγμένες υπολογιστικές μεθόδους και εργαλεία βιοπληροφορικής.

Στην παρούσα εργασία αναπτύσσονται τρία μοντέλα μηχανικής μάθησης που προβλέπουν τη συσχέτιση DNA, RNA και πρωτεϊνών σε μεμονωμένα κύτταρα, χρησιμοποιώντας τεχνολογίες Cite-Seq και Multiome. Αυτή η διπλωματική εργασία συνεισφέρει στην κατανόηση της κυτταρικής βιολογίας και των ασθενειών και παρέχει εργαλεία για την εξατομίκευση της ιατρικής πρακτικής.

Συνολικά, αυτή η διπλωματική εργασία προωθεί την κατανόηση της κυτταρικής ποικιλομορφίας και λειτουργίας, και ενισχύει τις προοπτικές για τη διάγνωση και θεραπεία ασθενειών.

Λέξεις - Κλειδιά

Μεμονωμένα κύτταρα, Μηχανική μάθηση, γονιδιακή έκφραση, ανάλυση δεδομένων

Multimodal single-cell analysis by using Machine Learning algorithms.

Karathanou Theodora

Abstract

This master's thesis explores an emerging area in the field of Biology, focusing on the study of individual cells using Machine Learning algorithms. This approach has enabled the systematic investigation of cellular heterogeneity across various tissues and cell populations, providing novel insights into the composition, dynamics, and regulatory rules governing cellular states during human development and disease.

Analyzing individual cells offers significant benefits to biologists by unveiling cellular heterogeneity, decoding disease mechanisms, facilitating drug discovery, and addressing fundamental biological questions. Furthermore, this approach allows drug monitoring and the discovery of new therapeutic options for various diseases.

To conduct this analysis, a variety of techniques are employed, such as flow cytometry, Single-Cell RNA Sequencing, Single-Cell Proteomics, Single-Cell Epigenomics, and mass cytometry. Analyzing and interpreting complex data require advanced computational methods and bioinformatics tools.

In this thesis, three machine learning models are developed to predict the correlation between DNA, RNA, and proteins in individual cells, using Cite-Seq and Multiome technologies. This work contributes to a deeper understanding of cellular biology and diseases and provides tools for personalized medicine.

Overall, this thesis advances our understanding of cellular diversity, function, and disease processes, enhancing prospects for diagnosis, therapy, and personalized medicine.

Key - Words

Single cell, Machine learning, cellular heterogeneity, data analysis

Περιεχόμενα

1. ΕΙΣΑΓΩΓΗ

Η διπλωματική εργασία αφορά μια όλο και περισσότερο αναδυόμενη περιοχή στο πεδίο της Βιολογίας, στην μελέτη σε επίπεδο ενός μεμονωμένου κυττάρου χρησιμοποιώντας αλγορίθμους Μηχανικής μάθησης. Η ανάλυση που βασίζεται στην μελέτη ενός κυττάρου αποτελεί μια ταχέως εξελισσόμενη προσέγγιση στην Βιολογία καθώς επέτρεψε τη συστηματική διερεύνηση της κυτταρικής ανομοιογένειας σε ένα ευρύ φάσμα ιστών και πληθυσμών κυττάρων, δίνοντας νέες γνώσεις για τη σύνθεση, τη δυναμική και τους ρυθμιστικούς κανόνες των κυτταρικών καταστάσεων κατά την διάρκεια της ανάπτυξης του ανθρώπινου οργανισμού καθώς και των ασθενειών του. Επιπλέον, η ανάπτυξη εργαλείων που βασίζονται στην μελέτη ενός κυττάρου στην ανακάλυψη και παράδοση φαρμάκων σε συνδυασμό με την είσοδο πλατφόρμων για την παρακολούθηση των φαρμάκων καθιστά πλέον δυνατή την ανακάλυψη νέων θεραπευτικών επιλογών για διάφορες ασθένειες.

Η ανάλυση της συγκεκριμένης μεθοδολογίας ουσιαστικά αναφέρεται στην μελέτη και τον χαρακτηρισμό μεμονωμένων κυττάρων σε ένα ετερογενή πληθυσμό. Η διαφοροποίηση του τρόπου αυτού μελέτης σχετικά με τις παραδοσιακές μελέτες είναι ότι οι παραδοσιακές μελέτες στην ουσία παρείχαν μια μέση μέτρηση των γενετικών και βιοχημικών ιδιοτήτων από ένα πλήθος κυττάρων. Με τον τρόπο αυτό συσκότιζαν τις επιμέρους διακυμάνσεις εντός του πληθυσμού. Σε συνδυασμό με τις τεχνολογικές εξελίξεις η ανάλυση σε ένα μεμονωμένο κύτταρο έχει γίνει όλο και πιο εφικτή και έχει φέρει “επανάσταση” στην κατανόησή μας για την κυτταρική ποικιλομορφία, ετερογένεια και λειτουργία.

Εμβαθύνοντας περαιτέρω στην σημαντικότητα της μεθόδου αυτής θα μπορούσαμε να ξεχωρίσουμε τους παρακάτω λόγους για τους οποίους οι επιστήμονες την χαρακτηρίζουν μια νέα εποχή στην Βιολογία[2]:

- Κυτταρική ετερογένεια: Τα κύτταρα μέσα στον ίδιο ιστό ή στο ίδιο όργανο μπορεί να παρουσιάζουν σημαντικές διαφορές. Διαφορές ως προς τη γονιδιακή έκφραση, τα επίπεδα πρωτεϊνών, τις επιγενετικές τροποποιήσεις και άλλα μοριακά χαρακτηριστικά. Οι παραδοσιακές μέθοδοι μαζικής ανάλυσης κυττάρων μπορεί να μέσες μετρήσεις του

πως είναι τα κύτταρα. Η ανάλυση όμως σε επίπεδο ενός κυττάρου μας επιτρέπει να αποκαλύψουμε και να κατανοήσουμε αυτή την ποικιλομορφία, αποκαλύπτοντας διαφορετικού τύπους κυττάρων, τους υποπληθυσμούς και τα σπάνια κύτταρα που μπορεί να χαθούν σε μαζικές μετρήσεις.[2]

- Αποκάλυψη κυτταρικής ανάπτυξης και διαφοροποίησης: Η ανάλυση ενός κυττάρου επιτρέπει τη μελέτη των διαδικασιών κυτταρικής ανάπτυξης και διαφοροποίησης. Μελετώντας τα μοτίβα των γονιδίων σε μεμονωμένα κύτταρα σε διαφορετικές χρονικ ές στιγμές, οι επιστήμονες μπορούν μπορούν να εντοπίσουν μοριακές υπογραφές που σχετίζονται με αποφάσεις κυτταρικής μοίρας, δέσμευση γενεαλογίας και αναπτυξιακές τροχιές.[2]
- Μηχανισμοί ασθενειών και βιοδείκτες: Πολλές ασθένειες, όπως ο καρκίνος, παρουσιάζουν ενδοκαρκινική ετερογένεια, όπου διαφορετικά κύτταρα μέσα σε έναν όγκο διαθέτουν διακριτά μοριακά προφίλ. Η ανάλυση των μεμονωμένων κυττάρων βοηθά στον εντοπισμό και τον χαρακτηρισμό αυτών των υποπληθυσμών, παρέχοντας πληροφορίες σχετικά με τους μηχανισμούς της νόσου, την αντίσταση στη θεραπεία και την ανακάλυψη πιθανών βιοδεικτών για έγκαιρη ανίχνευση ή εξατομικευμένες θεραπείες.[2]
- Ανακάλυψη και ανάπτυξη φαρμάκων: Η ανάλυση ενός κυττάρου μπορεί να βοηθήσει τους επιστήμπονες να ανακαλύωουν και να αναπτύξουν φάρμακα. Η εξέταση διαφορετικών τύπων κυττάρων και υποπληθυσμών τους στο σώμα ή σε ένα όγκο βοηθάει στον εντοπισμό των υπευθύνων για την εξέλιξη μιας νόσου ή των θεραπευτικών αποκρίσεων. Η κατανόηση της πολικιλομορφία των κυττάρων εντός ενός ιστού ή όγκου μπορεί να οδηγήσει την ανάπτυξη στοχευμένων θεραπειών και να διευκολύνει τον εντοπισμό νέων φαρμάκων.[2]
- Δίνει απάντηση σε θεμελιώδη βιολογικά ερωτήματα: Η εμβάθυνση στην μελέτη της κυτταρικής Βιολογίας βοηθά στην μελέτη και άλλων τομέων της βιολογίας όπως της αναπτυξιακής, της ανοσολογίας, της νευροβιολογίας καθώς και της γεννετικής ιατρικής. Η εμβάθυνση αυτή έρχεται σαν αποτέλεσμα της μελέτης σε επίπεδο ενός κυτταρου που επιτρέπει στους ερευνητές να διαρευνήσουν κυτταρικές διεργασίες, μονοπάτια σηματοδότησης, κυτταρικές αλληλεπιδράσεις καθώς και περιβαλλοντικά ερεθίσματα που αντιμετωπίζει το κάθε κύτταρο. [2]

Συνολικά, η ανάλυση ενός κυττάρου παρέχει μια πιο λεπτομερή και ολοκληρωμένη εικόνα της κυτταρικής ετερογένειας, της ανάπτυξης, των διαδικασιών ασθενειών και των θεμελιωδών βιολογικών μηχανισμών, οδηγώντας σε ανακαλύψεις σε διάφορους τομείς έρευνας και πιθανές εφαρμογές στη διάγνωση, τη θεραπευτική και την εξατομικευμένη ιατρική. Για τον σκοπό αυτό χρησιμοποιούνται μια σειρά τεχνικών με τα πλεονεκτήματα και τους περιορισμούς τους[26]:

1) Κυτταρομετρία ροής(Flow Cytometry): μια ευρέως χρησιμοποιούμενη τεχνική που επιτρέπει την ανάλυση μεμονωμένων κυττάρων με βάση τις φυσικές και χημικές τους ιδιότητες. Τα κύτταρα επισημαίνονται με δείκτες φθορισμού και περνούν μέσα από ένα κυτταρόμετρο ροής, όπου εξετάζονται ξεχωριστά και ταξινομούνται με βάση τα χαρακτηριστικά φθορισμού τους.[26]

2) Single-Cell RNA Sequencing (scRNA-seq): η τεχνική αυτή αποκαλύπτει το προφίλ της γονιδιακής έκφρασης σε μεμονωμένα κύτταρα, παρέχοντας πληροφορίες σχετικά με την κυτταρική ετερογένεια και προσδιορίζοντας κυτταρικούς τύπους ή υποπληθυσμούς. Αυτή η τεχνική περιλαμβάνει την απομόνωση μεμονωμένων κυττάρων, την αντίστροφη μεταγραφή του RNA τους σε συμπληρωματικό DNA (cDNA), την ενίσχυση του cDNA και την αλληλουχία του για τον προσδιορισμό του προφίλ της γονιδιακής έκφρασης.[26], [21]

3) Single-Cell Proteomics: στοχεύει στον χαρακτηρισμό της πρωτεϊνικής έκφρασης σε μεμονωμένα κύτταρα. Αυτή η τεχνική επιτρέπει την ταυτοποίηση και την ποσοτικοποίηση των πρωτεϊνών σε επίπεδο ενός κυττάρου και παρέχει πολύτιμες πληροφορίες σχετικά με την κυτταρική λειτουργία και τις οδούς σηματοδότησης.[4],[26]

4) Single-Cell Epigenomics: Οι επιγενετικές τροποποιήσεις διαδραματίζουν κρίσιμο ρόλο στη ρύθμιση της γονιδιακής έκφρασης. Οι τεχνικές επιγονιδιωματικής ενός κυττάρου, όπως η αλληλουχία διθειώδους ενός κυττάρου ή η ATAC-seq σε ένα κύτταρο (Assay for Transposase-Accessible Chromatin using sequencing), επιτρέπουν την ανάλυση μοτίβων μεθυλίωσης DNA ή προσβασιμότητας χρωματίνης σε μεμονωμένα κύτταρα, αντίστοιχα.[26],[2],[4]

5) Κυτταρομετρία μάζας(CyTOF): Η κυτταρομετρία μάζας συνδυάζει την κυτταρομετρία ροής με τη φασματομετρία μάζας, επιτρέποντας την ταυτόχρονη μέτρηση πολλαπλών παραμέτρων σε μεμονωμένα κύτταρα. Αντί για φθορίζουσες ετικέτες, τα αντισώματα επισημαίνο-νται

με ισότοπα βαρέων μετάλλων, τα οποία ανιχνεύονται και ποσοτικοποιούνται χρησιμοποιώντας φασματομετρία μάζας.[26], [4]

Όλες οι παραπάνω τεχνικές αντικειμενικά παράγουν έναν μεγάλο όγκο δεδομένων που περιλαμβάνουν πληροφορίες σχετικές με τα επίπεδα γονιδιακής έκφρασης, τις μεταλλάξεις του DNA, την αφθονία των πρωτεϊνών καθώς και τις επιγενετικές τροποποιήσεις. Η ανάλυση και η ερμηνεία των δεδομένων αυτών απαιτεί εξελιγμένες υπολογιστικές μεθόδους και εργαλεία βιοπληροφορικής. Οι ερευνητές χρησιμοποιούν τεχνικές ανάλυσης δεδομένων για να επεξεργαστούν, να οπτικοποιήσουν και να ερμηνεύσουν τα πολύπλοκα σύνολα δεδομένων που δημιουργούνται από πειράματα ενός κυττάρου. Πιο συγκεκριμένα τα εργαλεία της μηχανικής μάθησης έδωσαν την δυνατότητα στους επιστήμονες να επεξεργάζονται πιο γρήγορα και αποτελεσματικά τα πολύπλοκα αυτά δεδομένα και να βγάζουν συμπεράσματα. Τα εργαλεία μηχανικής μάθησης μπορούν να χρησιμοποιηθούν[23],[24] :

- Στην προεπεξεργασία δεδομένων: Η μελέτη σε επίπεδο ενός κυττάρου έχει σαν αποτέλεσμα την παραγωγή πολύ μεγάλου όγκου δεδομένων που χαρακτηρίζονται απο αρκετές πολυπλοκότητες όπως το ότι παράγονται σύνολα δεδομένων υψηλών διαστάσεων με πολλές μεταβλητές όπως τα επίπεδα γονιδιακής έκφρασης καθώς και την αφθονία των πρωτεινών σε μεμονωμένα κύτταρα. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για εργασίες προεπεξεργασίας δεδομένων όπως κανονικοποίηση, επιλογή χαρακτηριστικών και μείωση διαστάσεων. Τεχνικές όπως η PCA ή η t-SNE μπορούν να βοηθήσουν στην οπτικοποίηση και τη συμπίεση των δεδομένων διατηρώντας παράλληλα σημαντική βιολογική παραλλαγή.
- Ομαδοποίηση και ταξινόμηση κυττάρων: Οι αλγόριθμοι μηχανικής μάθησης, ιδιαίτερα οι τεχνικές μάθησης χωρίς επίβλεψη, όπως η ομαδοποίηση, μπορούν να χρησιμοποιηθούν για τον προσδιορισμό διακριτών πληθυσμών κυττάρων ή υποτύπων εντός του συνόλου δεδομένων ενός κυττάρου. Οι αλγόριθμοι ομαδοποίησης όπως το k-means, η ιεραρχική ομαδοποίηση ή η ομαδοποίηση με βάση την πυκνότητα μπορούν αυτόματα να ομαδοποιήσουν τα κύτταρα με βάση τα μοριακά τους προφίλ. Επιπλέον, μέθοδοι εποπτευόμενης μάθησης μπορούν να εφαρμοστούν για την ταξινόμηση των κυττάρων σε προκαθορισμένες κατηγορίες, όπως οι τύποι κυττάρων ή οι καταστάσεις ασθενειών, με βάση επισημασμένα δεδομένα εκπαίδευσης.

- Ανάλυση τροχιάς και συμπέρασμα κυτταρικής γενεαλογίας: Η ανάλυση ενός κυττάρου συχνά περιλαμβάνει μελέτη της κυτταρικής ανάπτυξης, διαφοροποίησης και σχέσεων γενεαλογίας. Οι αλγόριθμοι μηχανικής μάθησης, όπως η ανάλυση ψευδοχρόνου ή οι χάρτες διάχυσης, μπορούν να συναγάγουν κυτταρικές τροχιές και ιεραρχίες γενεαλογίας από τα δεδομένα ενός κελιού. Αυτοί οι αλγόριθμοι ανακατασκευάζουν τη χρονική σειρά των κυττάρων και παρέχουν πληροφορίες για την εξέλιξη των κυτταρικών καταστάσεων κατά την ανάπτυξη ή την εξέλιξη της νόσου.
- Πρόβλεψη μοντελοποίησης και ανακάλυψη βιοδεικτών: Τα μοντέλα μηχανικής μάθησης μπορούν να εκπαιδευτούν ώστε να προβλέπουν διάφορα αποτελέσματα ή ιδιότητες με βάση δεδομένα ενός κυττάρου. Για παράδειγμα, ενσωματώνοντας δεδομένα μεταγραφομικής μονοκυττάρων με κλινικές πληροφορίες, αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την πρόβλεψη των αποτελεσμάτων των ασθενών, τις απαντήσεις στη θεραπεία ή τον εντοπισμό προγνωστικών βιοδεικτών. Αυτά τα μοντέλα μπορούν να βοηθήσουν σε εξατομικευμένες ιατρικές προσεγγίσεις και να καθοδηγήσουν τη λήψη θε-ραπευτικών αποφάσεων.
- Ανάλυση δικτύου και συμπέρασμα διαδρομής: Η ανάλυση ενός κυττάρου συχνά δημιουργεί δεδομένα που μπορούν να αναπαρασταθούν ως δίκτυα αλληλεπίδρασης, όπως δίκτυα ρύθμισης γονιδίων ή δίκτυα αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να βοηθήσουν στην ανάλυση αυτών των δικτύων, στον εντοπισμό βασικών ρυθμιστικών ή σηματοδοτικών μονοπατιών και στη συναγωγή αιτιωδών σχέσεων μεταξύ γονιδίων ή πρωτεϊνών.

Στην παρούσα διπλωματική εργασία αναπτύσσονται τρία μοντέλα μηχανικής μάθησης που προβλέπουν την συσχέτιση DNA, RNA και πρωτεϊνών σε μεμονωμένα κύτταρα. Συγκεκριμένα με βάση τις τεχνολογίες Cite-Seq (τεχνολογία που αναφέρεται στην μελέτη της αλληλουχίας του RNA και την ανάλυση της πρωτεϊνικής επιφάνειας) και Multiome (τεχνολογία που ταυτόχρονα υποδεικνύει το Mrna και την προσβασιμότητα της χρωματίνης στο κύτταρο) που χρησιμοποιήθηκαν για την παραγωγή των δεδομένων αναπτύσσονται τρία μοντέλα Catboost, MLP και KNN για κάθε μια από τις δύο αυτές τεχνολογίες που βοηθούν στο να προβλεφθεί με συγκεκριμένα δεδομένα σε διάστημα 10 ημερών η γονιδιακή έκφραση και τα πρωτεϊνικά επίπεδα σε μεμονωμένα κύτταρα.

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Κύτταρο

Τα κύτταρα έχουν όλο τον εξοπλισμό και την τεχνογνωσία που απαιτούνται για την εκτέλεση των λειτουργιών της ζωής. Ένα κύτταρο μπορεί να φάει, να αναπτυχθεί και να κινηθεί. Μπορεί να εκτελέσει την απαραίτητη συντήρηση, να ανακυκλώσει εξαρτήματα και να απορρίψει τα απόβλητα. Μπορεί να προσαρμοστεί στις αλλαγές του περιβάλλοντός του και μπορεί ακόμη και να αναπαραχθεί. Πολλά έμβια όντα αποτελούνται από τεράστιο αριθμό κυττάρων που συνεργάζονται μεταξύ τους. Άλλες μορφές ζωής, ωστόσο, αποτελούνται από ένα μόνο κύτταρο, όπως τα πολλά είδη βακτηρίων και πρωτόζωων είναι δηλαδή μονοκύτταροι οργανισμοί . Τα κύτταρα, είτε ζουν μόνα τους είτε ως μέρος ενός πολυκύτταρου οργανισμού, είναι συνήθως πολύ μικρά για να παρατηρηθούν χωρίς μικροσκόπιο φωτός. Τα κύτταρα μοιράζονται πολλά κοινά χαρακτηριστικά, αλλά μπορούν να φαίνονται εξαιρετικά διαφορετικά. Στην πραγματικότητα, τα κύτταρα έχουν προσαρμοστεί εδώ και δισεκατομμύρια χρόνια σε ένα ευρύ φάσμα περιβαλλόντων και λειτουργικών ρόλων. Τα νευρικά κύτταρα, για παράδειγμα, έχουν μακριές, λεπτές επεκτάσεις που μπορούν να φτάσουν για μέτρα και να χρησιμεύσουν για τη γρήγορη μετάδοση σημάτων. Τα στενά τοποθετημένα, σχήματος τούβλου, φυτικά κύτταρα έχουν ένα άκαμπτο εξωτερικό στρώμα που βοηθά στην παροχή της δομικής υποστήριξης που χρειάζονται τα δέντρα και άλλα φυτά. Τα μακριά, κωνικά μυϊκά κύτταρα έχουν μια εγγενή ελαστικότητα που τους επιτρέπει να αλλάζουν μήκος μέσα στους δικέφαλους και χαλαρωτικούς δικέφαλους.

Ωστόσο, όσο διαφορετικά και αν είναι αυτά τα κύτταρα, όλα βασίζονται στις ίδιες βασικές στρατηγικές για να κρατήσουν το εξωτερικό έξω, να επιτρέψουν στις απαραίτητες ουσίες να εισέλθουν και να επιτρέψουν σε άλλους να φύγουν, να διατηρήσουν την υγεία τους και να αναπαραχθούν. Στην πραγματικότητα, αυτά τα χαρακτηριστικά είναι ακριβώς αυτά που κάνουν ένα κύτταρο, κύτταρο.

Τα κύτταρα θεωρούνται οι βασικές μονάδες της ζωής εν μέρει επειδή έρχονται σε διακριτά και εύκολα αναγνωρίσιμα πακέτα. Αυτό συμβαίνει επειδή όλα τα κύτταρα περιβάλλονται από μια δομή που ονομάζεται κυτταρική μεμβράνη - η οποία, όπως και οι τοίχοι ενός σπιτιού,

χρησιμεύει ως σαφές όριο μεταξύ του εσωτερικού και του εξωτερικού περιβάλλοντος του κυττάρου. Οι κυτταρικές μεμβράνες βασίζονται σε ένα πλαίσιο μορίων με βάση το λίπος που ονομάζονται φωσφολιπίδια, τα οποία εμποδίζουν φυσικά τις υδρόφιλες ή υδρόφιλες ουσίες να εισέλθουν ή να διαφύγουν από το κύτταρο. Αυτές οι μεμβράνες είναι επίσης γεμάτες με πρωτεΐνες που εξυπηρετούν διάφορες λειτουργίες. Ορισμένες από αυτές τις πρωτεΐνες δρουν ως φύλακες, καθορίζοντας ποιες ουσίες μπορούν και δεν μπορούν να διασχίσουν τη μεμβράνη. Άλλες λειτουργούν ως δείκτες, προσδιορίζοντας το κύτταρο ως μέρος του ίδιου οργανισμού ή ως ξένο. Ακόμα άλλες λειτουργούν σαν συνδετήρες, συνδέοντας τα κύτταρα μεταξύ τους, ώστε να μπορούν να λειτουργήσουν ως μονάδα. Ωστόσο, άλλες μεμβρανικές πρωτεΐνες χρησιμεύουν για την επικοινωνία, στέλνοντας και λαμβάνοντας σήματα από γειτονικά κύτταρα και το περιβάλλον - είτε φιλικά είτε ανησυχητικά.

Μέσα σε αυτή τη μεμβράνη, το εσωτερικό περιβάλλον ενός κυττάρου βασίζεται στο νερό και ονομάζεται κυτταρόπλασμα. Αυτό το υγρό περιβάλλον είναι γεμάτο κυτταρικούς μηχανισμούς και δομικά στοιχεία. Στην πραγματικότητα, οι συγκεντρώσεις πρωτεϊνών μέσα σε ένα κύτταρο ξεπερνούν κατά πολύ εκείνες στο εξωτερικό - είτε το εξωτερικό είναι ωκεάνιο νερό (όπως στην περίπτωση ενός μονοκύτταρου φύκου) είτε ορός αίματος (όπως στην περίπτωση ενός ερυθρού αιμοσφαιρίου).[15] Το κυτταρόπλασμα ενός κυττάρου φιλοξενεί όπως προαναφέρθηκε πολλά λειτουργικά και δομικά στοιχεία. Αυτά τα στοιχεία υπάρχουν με τη μορφή μορίων και οργανιδίων με τις σημαντικότερες κατηγορίες ενδοκυτταρικών οργανικών μορίων να είναι τα νουκλεϊνικά οξέα, οι πρωτεΐνες, οι υδατάνθρακες και τα λιπίδια, τα οποία είναι απαραίτητα για τις λειτουργίες του κυττάρου. Πιο συγκεκριμένα:

- Τα νουκλεϊνικά οξέα-> τα μόρια που περιέχουν και βοηθούν στην έκφραση του γενετικού κώδικα ενός κυττάρου. Υπάρχουν δύο μεγάλες κατηγορίες νουκλεϊνικών οξέων: δεσοξυριβονουκλεϊνικό οξύ (DNA) και ριβονουκλεϊνικό οξύ (RNA). Το DNA είναι το μόριο που περιέχει όλες τις πληροφορίες που απαιτούνται για την κατασκευή και τη διατήρηση του κυττάρου ενώ το RNA έχει διάφορους ρόλους που σχετίζονται με την έκφραση των πληροφοριών που αποθηκεύονται στο DNA.[15]
- Οι πρωτεΐνες -> οι ουσίες που κατασκευάζονται από αλυσίδες μικρότερων μορίων που ονομάζονται αμινοξέα και εξυπηρετούν μια ποικιλία λειτουργιών στο κύτταρο, τόσο καταλυτικές όσο και δομικές. Για παράδειγμα, οι πρωτεΐνες που ονομάζονται ένζυμα μετατρέπουν κυτταρικά μόρια (είτε πρωτεΐνες, υδατάνθρακες, λιπίδια ή νουκλεϊνικά

οξέα) σε άλλες μορφές που θα μπορούσαν να βοηθήσουν ένα κύτταρο να καλύψει τις ενεργειακές του ανάγκες, να χτίσει δομές υποστήριξης ή να αντλήσει απόβλητα.[14]

- Οι υδατάνθρακες-> μαζί με τα άμυλα και τα σάκχαρα στα κύτταρα, είναι ένας άλλος σημαντικός τύπος οργανικού μορίου. Οι απλοί υδατάνθρακες χρησιμοποιούνται για τις άμεσες ενεργειακές απαιτήσεις του κυττάρου, ενώ οι σύνθετοι υδατάνθρακες χρησιμεύουν ως ενδοκυτταρικές αποθήκες ενέργειας. Οι σύνθετοι υδατάνθρακες βρίσκονται επίσης στην επιφάνεια ενός κυττάρου, όπου παίζουν κρίσιμο ρόλο στην αναγνώριση των κυττάρων.
- Τα λιπίδια -> αλλιώς ονομάζονται τα μόρια λίπους είναι συστατικά των κυτταρικών μεμβρανών - τόσο της μεμβράνης πλάσματος όσο και διαφόρων ενδοκυτταρικών μεμβρανών. Εμπλέκονται επίσης στην αποθήκευση ενέργειας, καθώς και στην αναμετάδοση σημάτων εντός των κυττάρων και από την κυκλοφορία του αίματος στο εσωτερικό ενός κυττάρου.

Ορισμένα κύτταρα διαθέτουν επίσης τακτικές διατάξεις μορίων που ονομάζονται οργανίδια. Οι δομές αυτές χωρίζονται από το υπόλοιπο εσωτερικό ενός κυττάρου από τη δική τους ενδοκυτταρική μεμβράνη. Τα οργανίδια περιέχουν τεχνικό εξοπλισμό που απαιτείται για συγκεκριμένες εργασίες εντός του κυττάρου. Ένα παράδειγμα είναι το μιτοχόνδριο - κοινώς γνωστό ως "εργοστάσιο παραγωγής ενέργειας" του κυττάρου - το οποίο είναι το οργανίδιο που συγκρατεί και συντηρεί τον εξοπλισμό που εμπλέκεται στις χημικές αντιδράσεις παραγωγής ενέργειας.

2.2 Το κεντρικό δόγμα της Βιολογίας

Τα νουκλεϊνικά οξέα (DNA και RNA) είναι τα βασικά κύτταρα μέσω των οποίων επιτυγχάνεται η μεταφορά του γενετικού υλικού από γενιά σε γενιά. Μορφολογικά είναι σαν μακριές αλυσίδες που περιέχουν και μεταφέρουν πληροφορίες τέτοιες που να μπορούν να μεταβιβάζονται από γενιά σε γενιά. Αποτελούν μακρομόρια που συντίθεται από έναν μεγάλο αριθμό συνδεδεμένων μεταξύ τους νουκλεοτιδίων καθένα από τα οποία αποτελούνται από : 1) ένα σάκχαρο, 2)μια φωσφορική ομάδα και 3)μια βάση. Η αλληλουχία των βάσεων κατά μήκος της αλυσίδας είναι αυτή που μεταφέρει τη γενετική πληροφορία. Το μόριο DNA μορφολογικά μοιάζει με μια διπλή έλικα, ελικοειδούς δομής που αποτελείται από δύο συμπληρωματικές αλυσίδες νουκλεϊκών οξέων.[11] Το γονίδιο αποτελεί ένα τμήμα DNA το οποίο περιέχει

πληροφορίες για την σύνθεση πρωτεϊνών ή ενός μορίου RNA. Τα γονίδια είναι ουσιαστικά οι ομάδες στις οποίες αποθηκεύονται όλες οι γενετικές πληροφορίες και καθορίζουν τα χαρακτηριστικά ενός οργανισμού. [12]

Κάποια γονίδια είναι αυτά που ορίζουν τα διάφορα είδη των πρωτεϊνών που παράγονται στα κύτταρά μας, ωστόσο το DNA δεν λειτουργεί άμεσα ως εκμάγειο για την σύνθεση των πρωτεϊνών. Αντί να χρησιμοποιηθεί απευθείας το DNA για την πρωτεϊνοσύνθεση , αντιγράφεται σε πολλαπλά αντίγραφα γονιδίων που ονομάζονται αγγελιαφόρα RNA (mRNAs).[12] Η διαδικασία αυτή ονομάζεται μεταγραφή και ακολουθείται από την μετάφραση όπου ουσιαστικά αποτελεί την σύνθεση των πρωτεϊνών που γίνεται με βάση τις οδηγίες που δίνονται από τα εκμαγεία των mRNAs. Οι δύο αυτές διαδικασίες συνθέτουν την διαδικασία της γονιδιακής έκφρασης κατά την οποία το γονίδιο ουσιαστικά τίθεται σε επεξεργασία και παράγει λειτουργικά για τον οργανισμό προϊόντα. [11] Η διαδικασία αυτή της επεξεργασίας πληροφοριών στο επίπεδο της γονιδιακής έκφρασης είναι αρκετά πολύπλοκη και εξηγήθηκε πρώτη φορά από τον Francis Crick το 1958.

Με βάση τον Crick το βασικό δόγμα της βιολογίας αποτυπώνεται όπως παρακάτω

DNA ---------→ RNA----------→ πρωτεΐνη

Που σημαίνει ότι ουσιαστικά το DNA μεταγράφεται σε RNA και το RNA μεταφράζεται σε πρωτεΐνη.

2.3 Τα βλαστοκύτταρα

Μια ειδική κατηγορία κυττάρων είναι τα γνωστά σε όλους μας βλαστοκύτταρα. Οι κύριες κατηγορίες των βλαστοκυττάρων με βάση την ηλικία τους είναι τα εμβρυικά και τα ενήλικα βλαστοκύτταρα. Ενώ με βάση την ικανότητα διαφοροποίησής τους διαχωρίζονται σε παντοδύναμα (διαφοροποιείται σε όλα τα είδη κυττάρων εμβρύου), πλειοδύναμα (διαφοροποιούνται σε σχεδόν 200 είδη διαφορετικών κυττάρων), πολυδύναμα (με δυνατότητα αυτό-ανανέωσης και περιορισμένη διαφοροποίηση), ολιγοδύναμα (με δυνατότητα να διαφοροποιηθούν σε λίγα είδη κυττάρων) και μονοδύναμα (διαφοροποιούνται σε έναν τύπο κυττάρου). [8],[17]

Οι ιδιότητες των βλαστοκυττάρων είναι ότι είναι αρχέγονα δηλαδή πριν καν δημιουργηθεί ένα ιστός προϋπάρχει το βλαστοκύτταρο που δίνει την σηματοδότηση για την δημιουργία του ιστού. Είναι επίσης αδιαφοροποίητα που σημαίνει ότι δεν υπάρχουν κάποιες συγκεκριμένες πρωτεΐνες που να τα χαρακτηρίζουν, δεν εμφανίζουν κάποια εξειδίκευση και δεν έχουν συγκεκριμένη κυτταρική παραλλαγή. Τέλος αποτελούν κύριες ιδιότητές τους ότι παρουσιάζουν τις δυνατότητες της αυτο-ανανέωσης και της διαφοροποίησης. [8],[17]

Αποτελούν μια ξεχωριστή κατηγορία κυττάρων γιατί είναι η «πρώτη ύλη» του σώματος και έχουν τις δυνατότητες να αυτό-αναπαράγονται και να διαφοροποιούνται σε άλλα κύτταρα με πιο ειδικές λειτουργίες. Μπορούν να διαφοροποιηθούν σε αιμοποιητικά κύτταρα, σε οστικά κύτταρα, εγκεφαλικά κύτταρα καθώς και σε καρδιακά κύτταρα κτλ.

2.3.1 Η αυτό-ανανέωση των βλαστοκυττάρων

Τα βλαστοκύτταρα έχουν την δυνατότητα απεριόριστου πολλαπλασιασμού και αναπαραγωγής, δίνουν συνεχείς κύκλους διαίρεσης όπου τουλάχιστον ένα θυγατρικό κύτταρο είναι αντίστοιχο του μητρικού.

Μπορούν δηλαδή να δημιουργήσουν αντίγραφα του εαυτού τους μέσω μιας διαδικασίας δυναμικής αλληλεπίδρασης των εγγενών τους πρωτεϊνών και των εξωτερικών σημάτων του περιβάλλοντός τους. Η γνώση του τρόπου με τον οποίο λειτουργεί αυτή η διαδικασία είναι σημαντική για το πώς μεγαλώνει και γερνάει το σώμα μας και για την καταπολέμηση ασθενειών όπως ο καρκίνος. Για να επιτευχθεί η διαδικασία της αυτό-ανανέωσης θα πρέπει από την μία το κύτταρο να μπει στον κυτταρικό κύκλο και να διαφοροποιηθεί και από την άλλη τουλάχιστον ένας από τους απογόνους του βλαστοκυττάρου να είναι αδιαφοροποίητο κύτταρο. Σε διαφορετική περίπτωση το κύτταρο οδηγείται σε κυτταρική εξάντληση και ενδεχόμενη δυσλειτουργία των ιστών. [17]

Πέρα από τα βλαστοκύτταρα , η αυτό-ανανέωση υφίσταται και σε άλλου τύπου κυττάρων όπως είναι τα προγονικά κύτταρα που μπορούν και αυτά να αυτό-ανανεωθούν. Η βασική διαφορά της μίας από την άλλη αυτό-ανανέωση είναι ότι στα προγονικά κύτταρα μιλάμε για βραχυπρόθεσμη αυτό-ανανέωση ενώ στα βλαστοκύτταρα για μακροπρόθεσμη. Δηλαδή στα βλαστοκύτταρα μια τέτοια διαδικασία μπορεί να υφίσταται καθ' όλη την διάρκεια ζωής του έμβιου όντος. Παρά το γεγονός ότι τα έμβια όντα έχουν μεγάλες διαφορές μεταξύ τους (πχ ο ανθρώπινος οργανισμός σε σχέση με τα έντομα) η μακροπρόθεσμη αυτό-ανανέωση

υποδεικνύει την διάκριση μεταξύ θανάτου και ζωής των περισσότερων πολυκύτταρων οργανισμών. Στις περιπτώσεις που τα βλαστοκύτταρα εξαντλούνται γρήγορα ή προκύψουν γενετικές ανωμαλίες που μειώσουν την δυναμική του πολλαπλασιασμού τους ενδεχόμενα προκύπτει πρόωρη γήρανση ή ατροφία των ιστών. Επίσης σε περίπτωση μεταλλάξεων που επιφέρουν συχνότερες διαιρέσεις βλαστοκυττάρων χωρίς ισορροπία στον βαθμό διαφοροποίησης μπορεί οι ιστοί να αναπτυχθούν ανώμαλα ακόμη και να έχουμε ενδείξεις καρκίνου. Για αυτό ακριβώς η αυτό-ανανέωση αποτελεί μια σημαντική διαδικασία για τον ανθρώπινο και όλους τους έμβιους οργανισμούς που υποδεικνύει σημαντικές λειτουργίες του.[1]

2.3.2 Η διαφοροποίηση των βλαστοκυττάρων

Ως διαφοροποίηση των βλαστοκυττάρων αναφέρεται η διαδικασία κατά την οποία μετατρέπονται σε πιο εξειδικευμένους τύπους κυττάρων. Η διαδικασία αυτή επιτυγχάνεται μέσα από μια διαδοχή αλλαγών στην μορφολογία των κυττάρων, στο δυναμικό της μεμβράνης των κυττάρων , στην μεταβολική δραστηριότητα και την ανταπόκριση των σημάτων. Τα τελευταία χρόνια οι επιστήμονες έχουν αρχίσει να κατανοούν ,μέσα από τις μελέτες που έχουν κάνει, τα σήματα που ενεργοποιούν κάθε βήμα της διαφοροποίησης. Τα σήματα αυτά περιλαμβάνουν παράγοντες που εκκρίνονται από άλλα κύτταρα, φυσική επαφή με γειτονικά κύτταρα καθώς και με ορισμένα μόρια στο μικροπεριβάλλον.[9] , [17]

Ακολουθώντας το κεντρικό δόγμα της βιολογίας που αναφέραμε και παραπάνω μπορούμε να εξετάσουμε τον τρόπο με τον οποίο γίνεται η διαφοροποίηση των βλαστοκυττάρων. Ξεκινώντας από την παραδοχή ότι όλα τα κύτταρα στον οργανισμό μας περιέχουν τις ίδιες γενετικές πληροφορίες που είναι αποθηκευμένες στο DNA μας εγείρεται το ερώτημα του πως διαμορφώνονται διάφορα εξειδικευμένα κύτταρα όπως ηπατικά κύτταρα, κύτταρα του δέρματος, κτλ. Η διαμόρφωση των εξειδικευμένων αυτών κυττάρων επιτυγχάνεται μέσω της διαδικασίας που ονομάζεται διαφορική έκφραση γονιδίων. Η διαδικασία αυτή ξεκινά όταν ένα ερέθισμα επιδρά σε ένα αδιαφοροποίητο κύτταρο δηλαδή σε ένα βλαστοκύτταρο. Το αποτέλεσμα της επίδρασης του ερεθίσματος αυτού μπορεί να δραστηριοποιήσει κάποια γονιδιώματα του DNA και να τα μετατρέψει σε μη ενεργά. Ταυτόχρονα τα γονιδιώματα που παραμένουν ενεργά είναι ικανά, ακολουθώντας το κεντρικό δόγμα της βιολογίας, να

μεταγραφούν για να παράγουν RNA. Το αγγελιαφόρο RNA (mRNA) εισέρχεται στο κυτταρόπλασμα αφήνοντας τον πυρήνα του κυττάρου και έπειτα προσκολλάται σε ένα ριβόσωμα που είτε είναι ελεύθερο είτε βρίσκεται στο ενδοπλασματικό δίκτυο. Την μεταγραφή, όπου ένα συγκεκριμένο τμήμα του DNA αντιγράφεται σε ένα νέο μόριο RNA για να μεταφερθεί, ακολουθεί η μετάφραση. Η πληροφορία που προέκυψε από την μεταγραφή «διαβάζεται» προκειμένου να μεταφραστεί και να παραχθεί ένα συγκεκριμένο πολυπεπτίδιο που κωδικοποιεί το γονίδιο. Η μεταγραφή και η μετάφραση συμβαίνουν επανειλημμένα έως ότου κωδικοποιηθούν όλα τα ενεργά γονίδια στο κύτταρο, παράγοντας διαφορετικές πρωτεΐνες. Οι πρωτεΐνες αυτές με τη σειρά τους είναι ικανές να αλλάξουν την δομή του κυττάρου με διαφορετικούς τρόπους επομένως είναι και σε θέση να καθορίσουν τη λειτουργία του. Επομένως κατά την διαδικασία της διαφοροποίησης τα αδιαφοροποίητα μη εξειδικευμένα βλαστικά κύτταρα επηρεάζονται από εξωτερικά ερεθίσματα που μεταβάλλουν την γονιδιακή τους έκφραση. Μετά τη μεταγραφή και την παραγωγή του mRNA η μετάφραση παράγει πρωτεΐνες που μπορούν να καθορίσουν την λειτουργία του κυττάρου και έτσι επιτυγχάνεται η διαφοροποίηση.[9] , [17]

Η μελέτη των βλαστοκυττάρων, των κατηγοριών τους και των ιδιοτήτων τους, μετατρέπεται σε ολοένα και πιο σημαντική στον κλάδο της Βιολογίας καθώς μέσω της μελέτης τους μπορούν να επικεντρώσουν στην μελέτη της ανθρώπινης βιολογίας και στην ανάπτυξη θεραπειών. Η καλύτερη μελέτη και κατανόηση των μοριακών σημάτων που ρυθμίζουν την κυτταρική διαίρεση , την εξειδίκευση και τη διαφοροποίηση στα βλαστικά κύτταρα μπορούν να δώσουν πληροφορίες σχετικά με το πως προκύπτουν ασθένειες, να προταθούν νέες θεραπευτικές στρατηγικές.

2.4 Βλαστικά κύτταρα του μυελού των οστών και η διαδικασία της αιμοποίησης

Μια βασική κατηγορία βλαστοκυττάρων είναι τα αιματοποιητικά βλαστικά κύτταρα τα οποία ευθύνονται για την διαδικασία της αιμοποίησης. Η αιμοποίηση είναι η διαδικασία μέσα από την οποία το σώμα μας παράγει τα αιμοποιητικά κύτταρα. Η δια βίου παραγωγή αιμοποιητικών κυττάρων εξαρτάται από τα αιμοποιητικά βλαστικά κύτταρα την ικανότητά τους να αυτό-

αναπαράγονται και να διαφοροποιούνται σε αιματικές αλληλουχίες.[9], [20] Τα αιμοποιητικά βλαστοκύτταρα αναπτύσσονται μέσα από μια περίπλοκη διαδικασία που εξελίσσεται κατά την διάρκεια του σχηματισμού και της ανάπτυξης ενός εμβρύου και τελικά αποικίζουν στον μυελό των οστών. [20]Μετά την γέννηση του εμβρύου στον μυελό των οστών δημιουργείται μια σταθερή κατάσταση στην οποία το μέγεθος της «δεξαμενής» των αιμοποιητικών βλαστοκυττάρων διατηρείται με βάση το ρυθμιστικό πλαίσιο που διέπει η διαδικασία της αυτό-ανανέωσης και διαφοροποίησης των βλαστοκυττάρων. Κάτι τέτοιο καθίσταται δυνατό καθώς στον μυελό των οστών περιλαμβάνονται εξειδικευμένες «φωλιές» (niches), τα σημεία εκείνα δηλαδή στα οποία τα βλαστικά κύτταρα διατηρούνται μέσω κυτταρικών διαιρέσεων και οι απόγονοί τους ακολουθούν την διαφοροποίηση των κυτταρικών σειρών.[20]

Η διαδικασία της αιμοποίησης είναι λοιπόν μια βιολογική διαδικασία μέσω της οποίας παράγεται διατηρείται και ανανεώνεται το αίμα στον ανθρώπινο οργανισμό. Είναι ένα πολύπλοκη και καλά ρυθμισμένη βιολογική διαδικασία η οποία εξαρτάται από την δράση διάφορων γενετικών παραγόντων και σημάτων στον οργανισμό.[3],[20]

Υπάρχουν εκατοντάδες χιλιάδες ειδών βλαστικών κυττάρων ωστόσο στην συγκεκριμένη διπλωματική εργασία ασχολούμαστε με δύο είδη βλαστοκυττάρων από τα οποία έχει προκύψει και ο όγκος δεδομένων που έχουμε επεξεργαστεί. Τα δύο αυτά είναι τα περιφερικά CD34+ αιμοποιητικά βλαστοκύτταρα (CD34+ hematopoietic stem cells) και τα προγονικά βλαστοκύτταρα(progenitor stem cells). Πιο συγκεκριμένα :

1) CD34+ αιμοποιητικά βλαστοκύτταρα-> είναι το συγκεκριμένο είδος βλαστοκυττάρων που βρίσκονται στο περιφερικό αίμα και διαδραματίζουν καθοριστικό ρόλο στην λειτουργία του ανοσοποιητικού συστήματος αλλά και σε θεραπευτικές επεμβάσεις. Η ονομασία CD34+ αποτελεί έναν δείκτη επιφάνειας του βλαστοκυττάρου που βοηθά στον εντοπισμό και την απομόνωσή τους. Τα συγκεκριμένα βλαστοκύτταρα συμμετέχουν στην παραγωγή διάφορων ανοσοκυττάρων όπως των λευκών και ερυθρών αιμοσφαιρίων καθώς και παίζουν πρωταγωνιστικό ρόλο σε ένα ισορροπημένο και λειτουργικό ανοσοποιητικό σύστημα που μπορεί να ανταποκριθεί σε μολύνσεις, να παράγει αντισώματα κτλ. Τα αιμοποιητικά αυτά βλαστικά κύτταρα μπορούν να μεταφερθούν από τον μυελό των οστών στο περιφερειακό αίμα σε περιπτώσεις που υπάρχει αντίδραση σε κάποιες θεραπείες ή κατά την διάρκεια κάποιων νοσημάτων. Η

μετακίνηση αυτή επιτρέπει την απομόνωση και συλλογή των βλαστοκυττάρων αυτών για διάφορες ιατρικές διαδικασίες όπως είναι η μεταμόσχευση αιμοποιητικών βλαστικών κυττάρων. [25]

2) Προγονικά βλαστοκύτταρα-> είναι ένας τύπος βλαστοκυττάρων που έπονται των εμβρυικών καθώς είναι μεταγενέστερα και πιο εξειδικευμένα ωστόσο δεν έχουν διαφοροποιηθεί πλήρως σε συγκεκριμένους τύπους κυττάρων. Είναι τα ενδιάμεσα βλαστοκύτταρα μεταξύ των βλαστικών και των πλήρως διαφοροποιημένων κυττάρων. Διαδραματίζουν καθοριστικό ρόλο στην ανάπτυξη και την αναπαραγωγή των ιστών αλλά και αντικαθιστούν χαμένα ή κατεστραμμένα κύτταρα σε συγκεκριμένους ιστούς ή όργανα. Έχουν την ικανότητα να διαιρούνται και αποτελούν την αφετηρία ανάπτυξης πολλών τύπων εξειδικευμένων κυττάρων εντός ενός ιστού αλλά σε αντίθεση με τα εμβρυικά μπορούν να δημιουργήσουν μόνο συγκεκριμένους τύπους κυττάρων. Συχνά συναντώνται σε διάφορους ιστούς στο σώμα όπως στον μυελό των οστών, στο αίμα, στον εγκέφαλο αλλά και στο δέρμα. [25]

2.5 Single cell analysis- η μελέτη σε επίπεδο μεμονωμένου κυττάρου

Πριν την μελέτη σε επίπεδο μεμονωμένου κυττάρου οι μελέτες των επιστημόνων επικέντρωναν στα αποτελέσματα των μέσων μετρήσεων συνόλων που προέρχονταν από εκατομμύρια κύτταρα μαζί. Ο τρόπος αυτός μελέτης δεν αρκούσε για την κατανόηση των κυτταρικών αλληλεπιδράσεων με τα διάφορα μόρια και οργανίδια ενώ στερούνταν και της αποτελεσματικής εξερεύνησης της κυτταρικής ετερογένειας, των χαρακτηριστικών και της δυναμικής σε γονιδιακό επίπεδο, σε πιο συγκεκριμένους κυτταρικούς πληθυσμούς. Σε αντίθεση με τον τρόπο αυτό , η ανάλυση σε ένα μεμονωμένο κύτταρο (single-cell analysis, SCA) μας παρέχει ξεκάθαρες πληροφορίες για κάθε συγκεκριμένο βιολογικό παράγοντα του κυττάρου που χρησιμεύουν στην κατανόηση της «συμπεριφοράς» των βλαστοκυττάρων ή ακόμη και την εξέλιξη ενός όγκου. Το μοριακό περιεχόμενο των κυττάρων που σχετίζεται με τον κυτταρικό τύπο και την κυτταρική κατάσταση τους χωρικούς και χρονικούς μετασχηματισμούς και το μικροπεριβάλλον μελετάται λεπτομερώς με την συγκεκριμένη ανάλυση. Η πληθώρα τεχνικών που χρησιμοποιούνται στην μελέτη ενός μεμονωμένου κυττάρου αποκαλύπτουν πληροφορίες σχετικά με τις λειτουργικές μεταλλάξεις αλλά και τις

παραλλαγές των κυττάρων που αντιγράφονται. Αποκαλύπτουν την γονιδιακή έκφραση που χρησιμοποιείται ευρέως στην ιατρική για την πρόβλεψη πολλών βιολογικών, βιοιατρικών και παθολογικών συνθηκών για πληθώρα αναλύσεων γύρω από ασθένειες, παίζουν σημαντικό ρόλο στην μελέτη των χαρακτηριστικών της κυτταρικής ετερογένειας, σύνθετων ασθενειών όπως ο καρκίνος, εξελίσσουν την δυναμική της διαφοροποίησης και της ποσοτικοποίησης της μεταγραφικής στοχαστικότητας. Επιπλέον οι τεχνικές αυτές βοηθούν στην μελέτη βιολογικών φαινομένων όπως είναι η ανάπλαση του ιστού, η ανοσολογική απόκριση ακόμη και την εμβρυική εξέλιξη. Αυτές είναι μερικές μόνο από τις δυνατότητες που μας έδωσαν οι τεχνικές της μεθόδου αυτής.[2], [22] Για τους σκοπούς της παρούσας εργασίας θα εστιάσουμε σε δύο τεχνικές που είναι και ευρέως χρησιμοποιούμενες τα τελευταία χρόνια αυτές των : 1) Single-Cell RNA Sequencing (scRNA-seq) και 2) Single-Cell Epigenomics.

- Single-Cell RNA Sequencing (scRNA-seq)

Η ανάλυση της αλληλούχισης RNA σε ένα μεμονωμένο κύτταρο ή αλλιώς η scRNA-seq τεχνική διερευνά το προφίλ της γονιδιακής έκφρασης σε επίπεδο κυττάρου. Απαντά σε ερωτήματα κυτταρικής ετερογένειας και ανάπτυξης πρώιμων εμβρύων. Σε ένα δεδομένο δείγμα η ανάλυσης αυτή βοηθά στην μέτρηση των μορίων RNA σε ένα κύτταρο. Οι πληροφορίες αυτές παρέχουν ένα στιγμιότυπο του μεταγραφώματος (transcriptome) ή αλλιώς του εύρους των αγγελιαφόρων RNA που εκφράζονται σε έναν οργανισμό. Παρά το γεγονός ότι τελικά η πρωτεΐνη είναι το τελικό προϊόν της έκφρασης ενός γονιδίου, η ανίχνευση των αγγελιαφόρων RNA δείχνει αν το γονίδιο είναι ή όχι ενεργοποιημένο και επομένως αν έχει την δυνατότητα να μεταφραστεί και έπειτα να εκφραστεί. Οι μεταγραφικές διαφορές που μπορεί να προκύψουν εντός του κυτταρικού πληθυσμού βοηθούν στον εντοπισμό υποπληθυσμών όπως είναι τα κακοήθη κύτταρα.[22]

1. Πρώτο βήμα είναι η απομόνωση ενός κυττάρου που υποβάλλονται σε λύση
2. Τα αγγελιαφόρα RNA των κυττάρων αυτών επισημαίνονται με μια σειρά νουκλεοτιδίων αδενίνης για να βοηθήσουν στην εξαγωγή του mRNA και να διατηρήσουν τη σταθερότητά τους κατά τη μετάφραση
3. Έπειτα εμπλουτίζονται με την χρήση poly[T]- εκκινητών που ενεργοποιούν την ουρά poly-A του mRNA

4. Ακολουθεί η αντίστροφη μεταγραφή σε από το μονόκλωνα mRNAs σε συμπληρωματικό DNA (cDNA).
5. Η ποσότητα cDNA ενισχύεται επιπλέον με την χρήση της συμβατικής PCR ή in vintro μεταγραφής
6. Τελικά οι ετικέτες γραμμωτού κώδικα και άλλες σύντομες αλληλουχίες που απαιτούνται από την πλατφόρμα αλληλούχισης προστίθενται στα μόρια του cDNA

Η μελέτη σε ένα μεμονωμένο κύτταρο δίνει την δυνατότητα να αξιολογηθούν οι διαφορές στην γονιδιακή έκφραση μεταξύ κυττάρων πολύ πιο αποτελεσματικά από την ανάλυση σε μαζικό επίπεδο. Ένα χαρακτηριστικό παράδειγμα είναι η δυνατότητα εύρεσης και χαρακτηρισμού εξωγενών κυττάρων (outlier cells) σε έναν πληθυσμό που επιδρά στην παραπέρα κατανόηση της αντίστασης στα φάρμακα και την υποτροπή στη θεραπεία του καρκίνου. Επομένως η scRNA-seq χρησιμοποιώντας αλληλουχίες επόμενης γενιάς βοηθά στην βαθύτερη κατανόηση της βιολογίας συνολικά.

- Single-Cell Epigenomics

Η ανάλυση αυτή επικεντρώνει στην μελέτη των ρυθμιστικών συστημάτων που ενεργοποιούν τις κληρονομικές αλλαγές στην γονιδιακή έκφραση ενός γονιδίου. Πραγματοποιούνται μέσω χημικών τροποποιήσεων του DNA και των πρωτεϊνών καθώς και αλλαγών στην προσβασιμότητα του DNA και της διαμόρφωσης της χρωματίνης (Μεθυλίωση DNA, τροποποιήσεις ιστόνης, συμπύκνωση χρωματίνης και οργάνωση πυρήνα). Ουσιαστικά διερευνά το πως τροποποιείται το γονιδίωμα με τέτοιον τρόπο ώστε η αλληλουχία του DNA να μην τροποποιείται και έτσι καθορίζονται οι κυτταρικοί φαινότυποι που ρυθμίζουν τη δυναμική της γονιδιακής έκφρασης. Η ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) ανάλυση πιο συγκεκριμένα χρησιμοποιείται για τον χαρακτηρισμό των καταστάσεων χρωματίνης σε δείγματα κυττάρων και ιστών, για τον προσδιορισμό περιοχών του γονιδιώματος που έχουν ανοικτές καταστάσεις χρωματίνης που γενικά σχετίζονται με θέσεις που υποβάλλονται σε ενεργή μεταγραφή. Σε σχέση με την scRNA-seq ανάλυση , η οποία παρέχει πληροφορίες σχετικά με τα γονίδια που εκφράζονται, η ATAC-Seq ανάλυση παρέχει μια εικόνα των δυνητικά ενεργών γονιδιακών διακοπτών και μεταγραφικών θέσεων σε όλο το γονιδίωμα. [22]

1. Αρχικά υφίσταται η απομόνωση των πυρήνων των κυττάρων. Χρειάζεται επίσης να αποικοδομηθεί η μεμβράνη του πλάσματος ενώ ταυτόχρονα να μένει ανέπαφη η μεμβράνη του πυρήνα. Η απομόνωση των πυρήνων των κυττάρων γίνεται με βάση κάποια πρωτόκολλα που περιλαμβάνουν την θερμοκρασία που πρέπει να έχουν τα παγωμένα κύτταρα και οι ιστοί , το ποσοστό βιωσιμότητας του δείγματος για να παραχθούν έγκυρες πληροφορίες κτλ.
2. Την απομόνωση ακολουθεί η διαδικασία της σήμανσης. Το ένζυμο Τ5 μαζί με δείκτες αλληλουχίας προστίθενται στο δείγμα.
3. Έπειτα το Tn5 εισέρχεται στους πυρήνες (με την πυρηνική μεμβράνη ακόμα άθικτη), κατακερματίζει το DNA σε ανοικτές περιοχές χρωματίνης και προσθέτει τους προσαρμογείς ευρετηρίασης. Για να εξασφαλιστεί η όσο το δυνατόν μεγαλύτερη επιτυχία σήμανσης, ο αριθμός πυρήνων αναλογίας και η ποσότητα Tn5 πρέπει να βελτιστοποιηθούν.
4. Στην συνέχεια γίνεται η επισήμανση κάθε κυττάρου ξεχωριστά με την χρήση μοναδικών barcodes που μαζί με τους μεμονωμένους πυρήνες τους ενθυλακώνονται είτε σε σφαιρίδια που περιέχουν ένα ειδικό gel είτε ειδικά φυσίγγια που περιέχουν λάδι.
5. Οι πυρήνες υπόκεινται σε αραίωση ώστε κάθε κάψουλα να περιέχει μόνο έναν πυρήνα. Μόλις ένας πυρήνας και ένα barcode ενθυλακωθούν, μια αντίδραση ενζύμων συνδέει τον γραμμωτό κώδικα με τους δείκτες αλληλουχίας.
6. Τέλος, οι γραμμωτοί κώδικες συνδέονται με θραύσματα DNA και δημιουργούνται βιβλιοθήκες που ενισχύονται με PCR πριν από την αλληλούχιση. Οι βιβλιοθήκες έχουν διπλό ευρετήριο και μπορούν να ταξινομηθούν με αναγνώσεις σε ζεύγη.

Το κύριο όφελος της ανάπτυξης της τεχνολογίας ATAC-Seq είναι ότι αυτή η μέθοδος επέτρεψε την αναγνώριση της ανοικτής χρωματίνης σε ετερογενή ή σύνθετα δείγματα ιστών και κυττάρων. Πολλά βιολογικά δείγματα, όπως όγκοι και ιστοί σε διαφορετικές αναπτυξιακές καταστάσεις, περιέχουν πολλαπλούς υποπληθυσμούς κυττάρων που δυνητικά έχουν διαφορετικά επιγονιδιωματικά προφίλ. Ουσιαστικά επιτρέπει την ταυτοποίηση τέτοιων κυτταρικών υποπληθυσμών και έτσι δίνει την ακριβέστερη περιγραφή της κατάστασης χρωματίνης σε αυτές τις δυναμικές διεργασίες.

2.6 Μηχανική Μάθηση

Μηχανική μάθηση είναι ένα πεδίο της τεχνητής νοημοσύνης που ασχολείται με τον σχεδιασμό και την ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μάθουν από τα δεδομένα και να προβλέπουν, αναγνωρίζουν πρότυπα και λαμβάνουν αποφάσεις χωρίς να χρειάζεται να προγραμματιστούν απευθείας.

Οι αλγόριθμοι μηχανικής μάθησης είναι σε θέση να εξάγουν σημαντικές πληροφορίες από τα δεδομένα, να ανακαλύπτουν πρότυπα και να κάνουν προβλέψεις. Αυτό μπορεί να χρησιμοποιηθεί σε πολλούς τομείς, όπως η αναγνώριση εικόνων, η αυτόματη μετάφραση, η ανάλυση κειμένου, η συσταδοποίηση δεδομένων, η πρόβλεψη τιμών και η ανίχνευση ανωμαλιών.

Ο κεντρικός στόχος της μηχανικής μάθησης είναι να δημιουργήσει μοντέλα που μπορούν να εκπαιδευτούν με δεδομένα και να βελτιώνονται με την εμπειρία. Αυτό τα καθιστά ευέλικτα και ικανά να προσαρμόζονται σε νέες καταστάσεις και δεδομένα. Μέσω της ανάλυσης και της επεξεργασίας των δεδομένων η μηχανική μάθηση ανοίγει τον δρόμο για την ανακάλυψη νέων πληροφοριών και την παραγωγή αξιόπιστων αποτελεσμάτων σε σύντομο χρονικό διάστημα.

Η τεχνολογία αυτή βασίζεται σε μεθόδους κατηγοριοποίησης, πρόβλεψης ή συσταδοποίησης παρατηρήσεων ανάλογα και με τα δεδομένα που τίθενται σε επεξεργασία. Το σύνολο παρατηρήσεων που δέχονται οι αλγόριθμοι αυτοί μπορεί να είναι συνεχή, κατηγορηματικά ή δυαδικά. Ανάλογα με το είδος των παρατηρήσεων η Μηχανική Μάθηση είναι είτε επιβλεπόμενη (Supervised learning) είτε μη επιβλεπόμενη (Unsupervised learning). Τα χαρακτηριστικά των δύο αυτών ειδών Μάθησης αποτυπώνονται στον παρακάτω πίνακα. [7]

Επιβλεπόμενη (Supervised learning)	**Μη επιβλεπόμενη (Unsupervised learning)**
• Το μοντέλο μαθαίνει από ετικετοποιημένα(labeled) δεδομένα εκπαίδευσης • Τα δεδομένα εκπαίδευσης αποτελούνται από χαρακτηριστικά εισόδου (ονομάζονται επίσης	• Το μοντέλο μαθαίνει από μη ετικετοποιημένα(unlabeled) δεδομένα • Τα δεδομένα εισαγωγής είναι διαθέσιμα χωρίς αντίστοιχες μεταβλητές ή ετικέτες-στόχους.

ανεξάρτητες μεταβλητές) και αντίστοιχες μεταβλητές-στόχους (ονομάζονται επίσης εξαρτημένες μεταβλητές ή ετικέτες) • Στόχος της είναι να εκπαιδευτεί το μοντέλο σε μια χαρτογράφηση μεταξύ των χαρακτηριστικών εισόδου και των μεταβλητών-στόχων, επιτρέποντας στο μοντέλο να κάνει προβλέψεις σε μη γνωστά δεδομένα • Παραδείγματα αλγορίθμων που χρησιμοποιούνται περιλαμβάνουν τη γραμμική παλινδρόμηση(linear regression), logistic regression, δέντρα αποφάσεων, random forests, μηχανές διανυσμάτων υποστήριξης (SVM) και νευρωνικά δίκτυα • Τεχνικές που περιλαμβάνονται είναι η κατηγοριοποίηση (classification), η παλινδρόμηση(regression), παλινδρόμηση με χρονοσειρά δεδομένων (regression with time series), sequence labeling κτλ.	• Στόχος της είναι η ανακάλυψη μοτίβων, δομών, ομαδοποιήσεων στα δεδομένα, εκτέλεση μείωσης διαστάσεων ή δημιουργία αντιπροσωπευτικών περιλήψεων των δεδομένων. • Παραδείγματα αλγορίθμων μάθησης χωρίς επίβλεψη είναι οι αλγόριθμοι ομαδοποίησης (k-means clustering, hierarchical clustering και DBSCAN), τεχνικές μείωσης διαστάσεων (όπως ανάλυση κύριων συνιστωσών (PCA) και t-SNE) και παραγωγικά μοντέλα (Gaussian mixture models , autoencoders) • Τεχνικές που περιλαμβάνονται είναι συσταδοποίηση(clustering), μείωση διαστάσεων (dimensionality reduction), ανίχνευση ανωμαλιών (anomaly detection)

Η μάθηση, όπως η νοημοσύνη, καλύπτει ένα τόσο ευρύ φάσμα διαδικασιών που είναι δύσκολο να καθοριστεί με ακρίβεια. Πολλές τεχνικές στη μηχανική μάθηση προέρχονται από τις προσπάθειες των ερευνητών αυτών, να κάνουν πιο ακριβείς τις θεωρίες τους για την εκμάθηση των ζώων και των ανθρώπων μέσω υπολογιστικών μοντέλων. Επίσης οι έννοιες και οι τεχνικές που διερευνώνται από τους ερευνητές στη μηχανική μάθηση μπορούν να φωτίσουν ορισμένες

πτυχές της βιολογικής μάθησης. Όσον αφορά τις μηχανές, μια μηχανή μαθαίνει κάθε φορά που αλλάζει τη δομή, το πρόγραμμα ή τις παρατήρησης της με τέτοιο τρόπο ώστε να βελτιώνεται η αναμενόμενη μελλοντική της απόδοση. Όταν η απόδοση μιας μηχανής αναγνώρισης ομιλίας βελτιώνεται μετά από ακρόαση αρκετών δειγμάτων της ομιλίας ενός ατόμου, αισθανόμαστε αρκετά δικαιολογημένη σε αυτή την περίπτωση να πούμε ότι η μηχανή έχει μάθει.[7]

Για τους σκοπούς της παρούσας εργασίας η ενότητα του θεωρητικού υπόβαθρου επικεντρώνει σε μια τεχνική Unsupervised την Συσταδοποίηση και σε δύο τεχνικές Supervised Learning τις Κατηγοριοποίηση και Παλινδρόμηση. Η επιλογή των τεχνικών αυτών γίνεται επειδή είναι οι πιο ευρέως γνωστές και χρησιμοποιούμενες στην Μηχανική Μάθηση αλλά και επειδή η διπλωματική εργασία βασίζεται σε τεχνικές Supervised Learning.

2.6.1 Συσταδοποίηση (Clustering)

Η συσταδοποίηση είναι μια τεχνική μη επιβλεπόμενης μάθησης στην οποία τα δεδομένα ομαδοποιούνται σε ομάδες (clusters) βάσει της ομοιότητάς τους. Στόχος της συσταδοποίησης είναι να βρεθούν μοτίβα ή δομές στα δεδομένα που μπορούν να ερμηνευθούν.

Ο αλγόριθμος συσταδοποίησης εξετάζει τις ομοιότητες μεταξύ των δεδομένων, όπως η απόσταση μεταξύ τους, και προσπαθεί να δημιουργήσει ομάδες δεδομένων που είναι πιο όμοια μεταξύ τους από ό,τι με τα υπόλοιπα δεδομένα.

Χρησιμοποιούνται σε πολλά πεδία και εφαρμόζονται σε ποικίλα προβλήματα. Μερικοί από τους δημοφιλείς αλγορίθμους συσταδοποίησης περιλαμβάνουν τον K-Means, τον Hierarchical Clustering, τον DBSCAN και τον Gaussian Mixture Models (GMM).

Οι αλγόριθμοι συσταδοποίησης μπορούν να χρησιμοποιηθούν για να ανακαλύψουν κρυμμένες κατηγορίες ή σχέσεις μεταξύ των δεδομένων και να βοηθήσουν στην πρόβλεψη ή στην εξαγωγή γνώσης από τα δεδομένα.[6]

2.6.2 Κατηγοριοποίηση (Classification)

Η επιβλεπομένη μάθηση (Supervised learning) αρχίζει συνήθως με ένα σύνολο παρατηρήσεων και με ορισμένες τις κλάσεις (labels) που έχουν ταξινομηθεί οι παρατηρήσεις αυτές. Η επιβλεπομένη μάθηση έχει σκοπό να εντοπίσει μοτίβα στα δεδομένα τα οποία μπορούν να εφαρμοστούν σε μια διαδικασία ανάλυσης δεδομένων.[19] Κάθε παρατήρηση από το σύνολο δεδομένων ανήκει σε μια κλάση που καθορίζει τη σημασία της μέσα στο σύνολο δεδομένων.

Για παράδειγμα, υπάρχουν αρκετά καρκινικά είδη στον ανθρώπινη φύση τα οποία διαφέρουν μεταξύ τους λόγο διαφορετικών χαρακτηριστικών και μοτίβων. Έτσι από μια συλλογή παρατηρήσεων συγκεκριμένων χαρακτηριστικών της κάθε κατηγορίας καρκίνου, ένας αλγόριθμος κατηγοριοποίησης μπορεί να κατηγοριοποιήσει νέες έγραφες βασισμένος στην ήδη υπάρχουσα πληροφορία που έχει εκπαιδευτεί.[19] Συνήθως στις μεθόδους κατηγοριοποιήσεις το σύνολο δεδομένων εισόδου χωρίζεται σε:

- Ένα σύνολο εκπαίδευσης (training set)
- Ένα σύνολο ελέγχου (test set).

Το σύνολο εκπαίδευσης χρησιμοποιείται για να κατασκευαστεί το μοντέλο, ενώ το σύνολο ελέγχου για να την επικύρωση του μοντέλου. Οι αλγόριθμοι επιβλεπομένης μάθησης εκπαιδεύονται χρησιμοποιώντας την υπάρχουσα πληροφορία που θα γίνει η ανάλυση, και έτσι η απόδοση των αλγορίθμων αξιολογεί τις παρατηρήσεις αυτές. Γενικά, τα πρότυπα που εντοπίζονται σε ένα υποσύνολο παρατηρήσεων δεν μπορούν να ανιχνευθούν σε μεγαλύτερο πληθυσμό παρατηρήσεων. Αν το μοντέλο είναι ικανό να αντιπροσωπεύει μόνο τα μοτίβα που υπάρχουν στο υποσύνολο εκπαίδευσης, δημιουργείτε ένα πρόβλημα που ονομάζεται Overfitting.[19] Το Overfitting σημαίνει ότι το μοντέλο είναι εκπαιδευμένο με ακρίβεια για τις παρατηρήσεις εκπαίδευσης, επομένως μπορεί να μην είναι ικανό να προβλεπτή σωστά για μεγάλα σύνολα άγνωστων παρατηρήσεων το οποίο είναι ασυσχέτιστο με τις παρατηρήσεις εκπαίδευσης .

2.6.3 Παλινδρόμηση (Regression)

Στην τεχνική αυτή ο στόχος δεδομένων είναι συνεχής και σκοπός της είναι η πρόβλεψη αριθμητικών αξιών ή συνεχόμενων εξόδων που βασίζονται σε features εισόδων.[19] Το μοντέλο παλινδρόμησης μαθαίνει τη σχέση μεταξύ των μεταβλητών εισόδου και της μεταβλητής συνεχούς στόχου για να κάνει προβλέψεις. Σε σχέση με την μη επιβλεπόμενη μάθηση ο στόχος της τεχνικής αυτής είναι η πρόβλεψη μιας συνεχούς μεταβλητής εξόδου ενώ στην μη επιβλεπόμενη μάθηση η εκπαίδευση δεν περιλαμβάνει συγκεκριμένες μεταβλητές-στόχους αλλά επικεντρώνει στην ανακάλυψη μοτίβων και δομών μην ετικετοποιημένων δεδομένων.

Ενώ και οι δύο τεχνικές classification και regression υπάγονται στην εποπτευόμενη μάθηση έχουν διαφορές -κλειδιά οι οποίες είναι οι εξής :

1) Στην regression τεχνική τα δεδομένα εξαγωγής (output type) όπως και οι μεταβλητές στόχοι (target variables) είναι συνεχείς μεταβλητές-στόχοι ενώ στην κατηγοριοποίηση προβλέπονται κατηγορικές ετικέτες κλάσης.
2) Οι αλγόριθμοι που χρησιμοποιούνται κατά βάση στην κατηγοριοποίηση είναι οι logistic regression, τα δέντρα αποφάσεων, random forests, support vector machines καθώς και naive Bayes. Στην παλινδρόμηση κάποιοι από τους πιο συνηθισμένους χρησιμοποιούμενους αλγόριθμους είναι η γραμμική παλινδρόμηση , polynomial regression, τα δέντρα αποφάσεων , τα νευρωνικά δίκτυα.
3) Ως προς τις μετρικές αξιολόγησης των μοντέλων η regression τεχνικής αξιολογείται με δείκτες όπως ο MSE (μέσο τετραγωνικό σφάλμα), MAE (μέσο απόλυτο σφάλμα) ή τον R-squared. Η αξιολόγηση στην κατηγοριοποίησης επιτυχγάνεται συνήθως με μετρικές ακρίβειας, ανάκλησης, F1-score, AUC-ROC[19]

2.6.4 Ο αλγόριθμος tSVD (truncated singular value decomposition)

Ο αλγόριθμος αυτός χρησιμοποιείται στην Μηχανική μάθηση τόσο σε περιπτώσεις unsupervised όσο και σε περιπτώσεις supervised learning και χρησιμεύει στην μείωση των διαστάσεων και την συμπίεση των δεδομένων. Μέσω της τεχνικής αυτής ένας πίνακας αποσυντίθεται στις μοναδικές τιμές και στα διανύσματα που περιέχει. Παρακάτω παρουσιάζεται μια περιγραφή του συγκεκριμένου αλγόριθμου:

1. Δεδομένου ενός αρχικού πίνακα X μεγέθους m x n, όπου m είναι ο αριθμός των γραμμών (samples) και n είναι ο αριθμός των στηλών (features).
2. Υπολογίζουμε την μοναδική τιμή συμπίεσης (SVD) του X, η οποία εκφράζει τον X ως το γινόμενο τριών πινάκων: X = UΣV^T.

   Όπου, U είναι ένας ορθογώνιος πίνακας μεγέθους m x m, που περιέχει τα αριστερά μοναδικά διανύσματα του X,

   Σ είναι ένας διαγώνιος πίνακας μεγέθους m x n, που περιέχει τις μοναδικές τιμές του X με φθίνουσα σειρά.

Και V^T είναι ο μετασχηματισμός του ορθογώνιου πίνακα V μεγέθους n x n, που περιέχει τα δεξιά μοναδικά διανύσματα του X.

3. Απο τον πίνακα Σ κρατάμε μόνο τις k μεγαλύτερες μοναδικές τιμές, όπου το k είναι μια παράμετρος που ορίζεται από τον χρήστη και καθορίζει την επιθυμητή διαστασιμότητα των μειωμένων δεδομένων.
4. Δημιουργούμε ένα νέο διαγώνιο πίνακα Σ_k αντικαθιστώντας όλες τις μοναδικές τιμές που βρίσκονται μετά την k-τη τιμή με μηδενικά.
5. Ανακατασκευάζουμε τον περικομμένο πίνακα X_k πολλαπλασιάζοντας τους περικομμένους πίνακες: X_k = UΣ_kV^T.
6. Ο πίνακας X_k που προκύπτει αντιπροσωπεύει τη μειωμένη αναπαράσταση των αρχικών δεδομένων X. Διατηρεί τις πιο σημαντικές πληροφορίες ενώ απορρίπτει τις λιγότερο σημαντικές συνιστώσες.

2.6.5 Ο αλγόριθμος CatBoost

Το CatBoost είναι ένας αλγόριθμος βελτιστοποίησης κλίσης που χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης. Το όνομά του προέρχεται από τη φράση "Categorical Boosting" καθώς είναι ιδιαίτερα αποτελεσματικό στην αντιμετώπιση κατηγορικών χαρακτηριστικών στα δεδομένα. Το CatBoost έχει σχεδιαστεί για να αντιμετωπίζει κατηγορικές μεταβλητές χωρίς την ανάγκη για έκτακτη προεπεξεργασία, κάνοντάς το ένα ισχυρό εργαλείο για προβλήματα ταξινόμησης που περιλαμβάνουν μείγμα αριθμητικών και κατηγορικών χαρακτηριστικών. Ο αλγόριθμος αυτός έχει τα εξής χαρατκηριστικά:

- Βελτιστοποίηση Κλίσης: Το CatBoost βασίζεται στον έννοια της βελτιστοποίησης κλίσης, που είναι μια μέθοδος σύνθεσης μοντέλων. Η σύνθεση μοντέλων συνδυάζει πολλά αδύναμα μοντέλα (ατομικά δέντρα απόφασης σε αυτήν την περίπτωση) για να δημιουργήσει ένα ισχυρό προβλεπτικό μοντέλο.
- Χειρισμός Κατηγορικών Χαρακτηριστικών: Το CatBoost χρησιμοποιεί μια ειδική μέθοδο για την αντιμετώπιση των κατηγορικών μεταβλητών κατά τη διάρκεια της διαδικασίας εκπαίδευσης. Χρησιμοποιεί μια παραλλαγή της βελτιστοποίησης κλίσης που εφαρμόζει έναν αλγόριθμο διαδοχικής βελτιστοποίησης στα κατηγορικά χαρακτηριστικά. Αυτό επιτρέπει στο CatBoost να μετατρέπει αυτόματα τις

κατηγορικές μεταβλητές σε αριθμητικές αναπαραστάσεις χωρίς την ανάγκη για χειροκίνητη κωδικοποίηση.

- Σημασία Κατηγορικών Χαρακτηριστικών: Το CatBoost παρέχει μια μοναδική υπολογιστική μέθοδο για την υπολογισμό της σημασίας των κατηγορικών χαρακτηριστικών. Λαμβάνει υπόψη τις εσωτερικές μετατροπές που εφαρμόζο-νται στις κατηγορικές μεταβλητές κατά τη διάρκεια της διαδικασίας εκπαίδευσης και παρέχει ενδείξεις για τη σημασία τους στις προβλέψεις του μοντέλου.
- Βελτιστοποίηση και Πρόληψη Υπερπροσαρμογής: Το CatBoost περιλαμβάνει διάφορες τεχνικές βελτιστοποίησης για την αποτροπή της υπερπροσαρμογής. Υποστηρίζει τη χρήση της L1 και L2 βαρύτητας, καθώς και την υπολογιστική σημασία των χαρακτηριστικών βασισμένη στην κλίση και την τυχαία μετάθεση για την αναγνώριση και απόρριψη θορυβώδων ή άσχετων χαρακτηριστικών.
- Δυνατότητα κλιμάκωσης: Το CatBoost έχει σχεδιαστεί για να αντιμετωπίζει μεγάλα σύνολα δεδομένων με αποδοτικό τρόπο. Εφαρμόζει παράλληλη υπολογιστική κατά τη διάρκεια της εκπαίδευσης, το οποίο του επιτρέπει να επεξεργάζεται τα δεδομένα με διανεμημένο τρόπο, καθιστώντας το κατάλληλο για σενάρια μεγάλων δεδομένων.

Συνοψίζοντας, ο CatBoost είναι ένας ισχυρός αλγόριθμος βελτιστοποίησης κλίσης που εξαίρει στη χειρισμό κατηγορικών χαρακτηριστικών, καθιστώντας τον ιδιαίτερα χρήσιμο για προβλήματα ταξινόμησης.

2.6.6 Ο αλγόριθμος MLP

Ο αλγόριθμος MLP (Multilayer Perceptron) είναι ένας τύπος τεχνητού νευρωνικού δικτύου που χρησιμοποιείται συνήθως για προβλήματα ταξινόμησης. Το MLP είναι ένα μοντέλο τροφοδοσίας μπροστά (feedforward neural network) που αποτελείται από πολλά επίπεδα συνδεδεμένων κόμβων, γνωστών ως νευρώνες. Κάθε νευρώνας στο δίκτυο είναι υπεύθυνος για την επεξεργασία και τη μετάδοση πληροφοριών στους νευρώνες των επόμενων επιπέδων. Παρακάτω είναι τα χαρακτηριστικά του :

- Αρχιτεκτονική: Το MLP αποτελείται από ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου. Το επίπεδο εισόδου λαμβάνει τα χαρακτηριστικά ή τα χαρακτηριστικά των εισόδων. Τα κρυφά επίπεδα πραγματοποιούν

υπολογισμούς και εφαρμόζουν μη γραμμικές μετασχηματισμούς στα δεδομένα εισόδου. Το επίπεδο εξόδου παράγει την τελική ταξινόμηση ή πρόβλεψη.

- Νευρώνες και Συναρτήσεις Ενεργοποίησης: Κάθε νευρώνας στο MLP συσχετίζεται με μια συνάρτηση ενεργοποίησης, η οποία εισάγει μη γραμμικότητα στο μοντέλο. Οι συνηθισμένες συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στο MLP περιλαμβάνουν τη σιγμοειδή συνάρτηση, την υπερβολική εφαπτομένη συνάρτηση και τη συνάρτηση γραμμικής οριοθέτησης (ReLU). Αυτές οι συναρτήσεις ενεργοποίησης επιτρέπουν στο δίκτυο να μάθει πολύπλοκες σχέσεις μεταξύ των δεδομένων εισόδου και των προβλέψεων εξόδου.
- Προώθηση μπροστά: Το MLP πραγματοποιεί προώθηση μπροστά, που περιλαμβάνει τη διέλευση των δεδομένων εισόδου μέσω του δικτύου από το επίπεδο εισόδου στο επίπεδο εξόδου. Οι τιμές από κάθε νευρώνα σε ένα επίπεδο χρησιμοποιούνται ως είσοδοι στους νευρώνες του επόμενου επιπέδου, ακολουθώντας έναν υπολογισμό με συνάρτηση ζυγισμένου αθροίσματος και τη συνάρτηση ενεργοποίησης. Αυτή η διαδικασία συνεχίζεται μέχρι το επίπεδο εξόδου να παράγει την τελική πρόβλεψη.
- Αντίστροφη διάδοση: Μετά το βήμα της προώθησης μπροστά, το MLP συγκρίνει την προβλεπόμενη έξοδο με την πραγματική έξοδο και υπολογίζει το σφάλμα. Στη συνέχεια, εφαρμόζει τον αλγόριθμο αντίστροφης διάδοσης (backpropagation) για να προσαρμόσει τις συνδέσεις μεταξύ των νευρώνων. Αυτή η επαναληπτική διαδικασία ελαχιστοποιεί το σφάλμα ανανεώνοντας τις συνδέσεις βάσει της κλίσης της συνάρτησης απώλειας ως προς τις συνδέσεις.
- Εκπαίδευση: Το MLP εκπαιδεύεται χρησιμοποιώντας ετικεταρισμένα δεδομένα, όπου τόσο τα χαρακτηριστικά εισόδου όσο και οι αντίστοιχες ετικέτες-στόχοι γνωρίζονται. Κατά τη διάρκεια της εκπαίδευσης, το δίκτυο προσαρμόζει τις συνδέσεις του για να ελαχιστοποιήσει τη διαφορά μεταξύ της προβλεπόμενης έξοδου και της πραγματικής έξοδου. Η διαδικασία εκπαίδευσης συνεχίζεται μέχρι το μοντέλο να φτάσει σε ένα ικανοποιητικό επίπεδο ακρίβειας ή σύγκλισης.
- Ταξινόμηση: Αφού εκπαιδευτεί το MLP, μπορεί να χρησιμοποιηθεί για την ταξινόμηση. Δεδομένου ενός νέου συνόλου χαρακτηριστικών εισόδου, το δίκτυο εφαρμόζει προώθηση μπροστά για να παράξει μια πρόβλεψη για την αντίστοιχη κατηγορία. Η κατηγορία με την υψηλότερη πιθανότητα ή την υψηλότερη τιμή ενεργοποίησης στο επίπεδο εξόδου θεωρείται ως η προβλεπόμενη κατηγορία.

Συνολικά, ο αλγόριθμος MLP είναι ένας τύπος νευρωνικού δικτύου που χρησιμοποιείται συχνά για προβλήματα ταξινόμησης. Αποτελείται από πολλά επίπεδα νευρώνων, χρησιμοποιεί μη γραμμικές συναρτήσεις ενεργοποίησης και χρησιμοποιεί την προώθηση μπροστά και τον αλγόριθμο αντίστροφης διάδοσης για εκπαίδευση και πρόβλεψη

2.6.7 Ο αλγόριθμος K-Nearest Neighbors (KNN)

Ο K-Nearest Neighbors (KNN) είναι ένας απλός και ευέλικτος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται τόσο για ταξινόμηση όσο και για προβλήματα παλινδρόμησης. Είναι ένας αλγόριθμος που βασίζεται σε παραδείγματα, ότι δεν κάνει δηλαδή καμία υπόθεση για την κατανομή των δεδομένων.

Η λειτουργία του βασίζεται στα εξής :

- Αρχικοποίηση: Γίνεται η επιλογή ενός θετικού ακεραίου που ως τιμή <K> που αντιπροσωπεύει τον αριθμό των κοντινότερων γειτόνων που θα ληφθούν υπόψην.

- Προεπεξεργασία δεδομένων : υπάρχει ένα σύνολο δεδομένων με labeled σημεία δεδομένων. Κάθε σημείο δεδομένων αποτελείται από συγκεκριμένες ιδιότητες και την αντίστοιχη κλάση. Ο KNN αλγόριθμος λειτουργεί σε ένα σύνολο δεδομένων που περιλαμβάνουν labeled δεδομένα όπου κάθε σημείο δεδομένων σχετίζεται με μια κλάση ή τιμη.

- Υπολογισμός απόστασης : Για ένα δεδομένο σημείο που έχει επιλεγεί για ταξινόμηση ή για να γίνει μια πρόβλεψη , υπολογίζεται η απόσταση μεταξύ του σημείου αυτού και όλων των υπόλοιπων σημείων δεδομένων μέσα στο σύνολο δεδομένων. Κάποιες συνηθισμένες μέθοδοι για να μετράται η απόσταση αυτή είναι η ευκλείδεια απόσταση, η απόσταση Manhattan ή η απόσταση Minkowski. Η επιλογή της μετρικής απόστασης μπορεί να επηρεάσει τα αποτελέσματα, επομένως είναι σημαντική η επιλογή μιας κατάλληλης μετρικής βάσει του πεδίου του προβλήματος.

- Επιλογή γειτόνων: Γίνεται η ταξινόμηση των αποστάσεων που υπολογίστηκαν στο προηγούμενο βήμα με αύξουσα σειρά και επιλέγονται τα «Κ» σημεία δεδομένων με τις μικρότερες αποστάσεις. Αυτοί είναι οι Κ κοντινότεροι γείτονες στο σημείο που επιθυμούμε να γίνει η πρόβλεψη ή η ταξινόμηση

- Μέση τιμή: στα προβλήματα παλινδρόμησης υπολογίζεται ο μέσος όρος των τιμών στόχου των Κ κοντινότερων γειτόνων. Αυτή η μέση τιμή γίνεται η προβλεπόμενη τιμή για το σημείο δεδομένων.

- Πρόβλεψη : αντιστοιχίζεται η προβλεπόμενη τιμή στο σημείο δεδομένων που ενδιαφέρει να ερευνηθεί.

- Επανάληψη : Η διαδικασία επεναλαμβάνεται για κάθε σημείο δεδομένων που επιδιώκουμε να γίνει πρόβλεψη η ταξινόμηση στο σύνολο των δεδομένων

Γενικότερα ο συγκεκριμένος αλγόριθμος δεν δημιουργεί ένα μοντέλο κατά την διάρκεια της εκπαίδευσης αλλά αποθηκεύει τα δεδομένα στη μνήμη. Μπορεί να αποτελέσει έναν υπολογιστικά δαπανηρό αλγόριθμο όταν χρησιμοποιηθεί σε μεγάλα σύνολα δεδομένων επειδή απαιτεί τον υπολογισμό μεταξύ του σημείου ενδιαφέροντος και όλων των άλλων σημείων δεδομένων. Χρησιμοποιείται συχνά ως ένας αλγόριθμος αναφοράς και μπορεί να λειτουργήσει καλά σε καταστάσεις όπου τα όρια απόφασης δεν είναι πολύπλοκα, αλλά ενδέχεται να μην αποδίδει καλά σε υψηλές διαστάσεις ή με μη ισορροπημένα σύνολα δεδομένων. Η σωστή προεπεξεργασία των δεδομένων και η κανονικοποίηση των χαρακτηριστικών μπορούν επίσης να είναι σημαντικές για την απόδοση του KNN.

3. ΠΕΡΙΓΡΑΦΗ ΠΡΟΒΛΗΜΑΤΟΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Ο στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η ανάπτυξη μοντέλων μηχανικής μάθησης που χρησιμοποιούνται στην πρόβλεψη του πώς συσχετίζονται οι μετρήσεις του DNA, του RNA και των πρωτεϊνών σε μεμονωμένα κύτταρα. Τα κύτταρα αυτά αναφέρονται σε βλαστοκύτταρα του μυελού των οστών καθώς εξελίσσονται σε πιο ώριμα αιμοποιητικά κύτταρα.

Συγκεκριμένα το δείγμα των δεδομένων μας περιλαμβάνει 300.000 στοιχεία CD34+ αιματοποιητικών βλαστικών κύτταρων καθώς και των προγονικών τους (HSPCs). Η υπόθεση αυτή εργασίας έχει γίνει με την δήλωση της συμμετοχής σε διαγωνισμό στην ιστοσελίδα Kaggle. Ο διαγωνισμός που έλαβα μέρος είχε τον τίτλο «Open Problems - Multimodal Single-Cell Integration» και περιελάβανε δείγμα 30gb δεδομένων. Τα δεδομένα αυτά παράχθηκαν από την Cellarity, μια εταιρεία παραγωγής φαρμάκων που επικεντρώνεται στα κύτταρα. Η συμμετοχή στον διαγωνισμό πραγματοποιήθηκε πριν από περίπου 10 μήνες όπου εκτελέστηκε και παραδόθηκε το τεχνικό μέρος της εργασίας που ακολουθεί παρακάτω. Για τον σκοπό της διπλωματικής εργασίας το τεχνικό μέρος ατόφιο όπως είχε δημιουργηθεί χρησιμοποιήθηκε για την εργασία η οποία εμπλουτίστηκε με το θεωρητικό υπόβαθρο ώστε σαν σύνολο να ανταποκρίνεται στις ανάγκες μιας διπλωματικής εργασίας.

Η ουσία της διπλωματικής εργασίας σχετίζεται με την ανάπτυξη μοντέλων εκπαιδευμένων στο παραπάνω σύνολο δεδομένων που προήλθαν από 4 δότες σε 5 διαφορετικά χρονικά σημεία. Με την δημιουργία και εφαρμογή αυτών των μοντέλων στην ζωή μπορεί να δοθεί μια ιατρική εικόνα της κατάστασης ενός μεμονωμένου κυττάρου ως προς την εξέλιξη των πρωτεϊνών σε μελλοντικές χρονικές στιγμές. Ένα τέτοιο εργαλείο μπορεί να βοηθήσει στην επιτάχυνση και την καινοτομία στις μεθόδους καταγραφής των γενετικών πληροφοριών σε διαφορετικά επίπεδα κυτταρικής κατάστασης όπως αυτή εξελίσσεται στον χρόνο. Αν μπορούμε να προβλέψουμε μια μελλοντική βιολογική εικόνα μπορούμε και να εμβαθύνουμε στην κατανόηση των κανόνων που διέπουν τις πολύπλοκες ρυθμιστικές διαδικασίες που διέπουν το ίδιο το κύτταρο.[16]

Όπως αναλύθηκε και παραπάνω οι δυνατότητες που έχει αναπτύξει η μελέτη σε επίπεδο ενός μεμονωμένου κυττάρου επανασχεδιάζουν δυναμικά το πως οι επιστήμονες αναλύουν το DNA, το RNA και οι πρωτεΐνες μέσα σε μεμονωμένα κύτταρα. Οι καινοτόμες αυτές τεχνολογίες παρέχουν απαράμιλλες ευκαιρίες για τη μελέτη της βιολογίας σε μεγάλη κλίμακα και ανάλυση. Αποτελέσματα αυτής της προσπάθειας είναι λεπτομερείς χάρτες της αρχικής ανθρώπινης εμβρυϊκής ανάπτυξης, η ανακάλυψη νέων κυτταρικών τύπων συνδεδεμένων με ασθένειες και η ανάπτυξη κατευθυνόμενων θεραπευτικών παρεμβάσεων. Επιπλέον, με τις πρόσφατες προόδους στις πειραματικές τεχνικές, είναι πλέον δυνατή η μέτρηση πολλαπλών γενομικών μεθόδων στο ίδιο κύτταρο.[16]

Παρά την αυξανόμενη διαθεσιμότητα μεγάλου όγκου δεδομένων που προκύπτουν από τις μελέτες αυτές, υπάρχει έλλειψη αποτελεσματικών μεθόδων ανάλυσης των δεδομένων αυτών. Τα δεδομένα αυτά χαρακτηρίζονται και ως πολυτροπικά (multimodal) ή πολύ-μεταβλητά γιατί όταν τα αναλύουμε θα πρέπει να παίρνουμε υπόψιν ότι αναφέρονται σε διαφορετικούς χώρους, καθώς και ταυτόχρονα σε ποικιλία κυτταρικών εικόνων. Παίρνοντας επίσης υπόψιν ότι τα δεδομένα που προκύπτουν δεν είναι κάποια στατιστικά στιγμιότυπα που αφορούν τα κύτταρα αλλά προκύπτουν μέσα από μια συνεχώς δυναμική βιολογική διαδικασία. Η πρόκληση λοιπόν στην μελέτη και ανάλυση τέτοιου είδους δεδομένων είναι η συσχέτιση της χρονικής δυναμικής με τις παράλληλες αλλαγές στην κατάσταση των ίδιων των κυττάρων.[16]

Το σύνολο δεδομένων για αυτόν τον διαγωνισμό περιλαμβάνει δεδομένα που συλλέχθηκαν από περιφερικά CD34+ αιματοποιητικά βλαστοκύτταρα και τα προγονικά του (HSPCs), που απομονώθηκαν από τέσσερις υγιείς δότες που ονομάστηκαν 13176, 31800, 32606 και 27678.

Οι μετρήσεις πραγματοποιήθηκαν πέντε φορές σε διάρκεια δέκα ημερών. Κατά την περίοδο αυτή τα κύτταρα καλλιεργούνταν στο μέσο καλλιέργειας StemSpan SFEM ενισχυμένο από CC100, ένα κοκτέιλ κυτοκινών που επεκτείνει τα κύτταρα σε περιβάλλον άνευ ορού. Το μέσο αυτό περιείχε και θρομβοποιητίνη (TPO), άλλαζε κάθε 2-3 μέρες , εγκιβωτίστηκε στους 37ºC και δεν χρησιμοποιήθηκαν επιπρόσθετα μέσα.

Από κάθε πλάκα καλλιέργειας σε κάθε χρονικό σημείο δειγματοληψίας, συλλέχθηκαν κύτταρα για μέτρηση με δύο μεθόδους . Η πρώτη είναι η τεχνολογία 10x Chromium Single Cell Multiome ATAC + Gene Expression (Multiome) και η δεύτερη είναι η τεχνολογία 10x

Genomics Single Cell Gene Expression με Feature Barcoding χρησιμοποιώντας το TotalSeq™-B Human Universal Cocktail, V1.0 (CITEseq).

Η τεχνολογία Multiome μετράει την προσβασιμότητα σε χρωματίνη και την έκφραση γονιδίων αν δηλαδή εκφράζονται ή όχι σε πρωτεΐνη ενώ η Citeseq μετρά την έκφραση των γονιδίων και τα επίπεδα πρωτεϊνών επιφανείας.

Τα δεδομένα ομαδοποιήθηκαν ως εξής:

-το σύνολο εκπαίδευσης (training set) χρησιμοποιεί δείγματα από του 13176, 31800 και 32606 δότες ενώ το σύνολο ελέγχου (test set) χρησιμοποιεί δείγματα μόνο από τον 27678 δότη

-για τα δείγματα της Multiome τεχνολογίας , το σύνολο εκπαίδευσης χρησιμοποιεί δείγματα από τις μέρες 2,3,4 και 7 και το σύνολο ελέγχου από τις μέρες 2,3 και 7.

-για τα δείγματα της Citeseq τεχνολογίας, το σύνολο εκπαίδευσης αλλά και το σύνολο ελέγχου χρησιμοποιεί δείγματα από τις μέρες 2,3 και 4.

Τα δεδομένα χωρίστηκαν στα εξής αρχεία :

1) metadata.csv(977mb)

   περιείχε 5 στήλες (cell_id που είναι μοναδικό αναγνωριστικό του κυττάρου, day-> αριθμός ημέρας, το νούμερο του δότη, ο τύπος του κυττάρου cell_type (hsc, hidden και other), τεχνολογία (multiome και citeseq).

   Το αρχείο αυτό περιγράφει την δειγματοληψία του κυττάρου.

   Ένα χαρακτηριστικό στιγμιότυπο που άνοιξε με την χρήση jupyter είναι το παρακάτω:

	cell_id	day	donor	cell_type	technology
0	c2150f55becb	2	27678	HSC	citeseq
1	65b7edf8a4da	2	27678	HSC	citeseq
2	c1b26cb1057b	2	27678	EryP	citeseq
3	917168fa6f83	2	27678	NeuP	citeseq
4	2b29feeca86d	2	27678	EryP	citeseq
...	...	...	...	...	...
281523	96a60b026659	10	31800	hidden	multiome
281524	d493e546991e	10	31800	hidden	multiome
281525	05666c99aa48	10	31800	hidden	multiome
281526	121f946642b5	10	31800	hidden	multiome
281527	b847ba21f59f	10	31800	hidden	multiome

281528 rows × 5 columns

2) evaluation_ids.csv(2,42gb)

Περιλαμβάνει 65,7 εκ γραμμές, 3 στήλες (row_id το αναγνωριστικό της γραμμής, cell_id, gene_id το αναγνωριστικό του γονιδίου)

Το αρχείο αυτό περιγράφει τα γονίδια που περιέχονται στο κάθε κύτταρο.

Ένα χαρακτηριστικό στιγμιότυπο που άνοιξε με την χρήση jupyter είναι το παρακάτω:

	row_id	cell_id	gene_id
0	0	c2150f55becb	CD86
1	1	c2150f55becb	CD274
2	2	c2150f55becb	CD270
3	3	c2150f55becb	CD155
4	4	c2150f55becb	CD112
...	...	...	...
65744175	65744175	2c53aa67933d	ENSG00000134419
65744176	65744176	2c53aa67933d	ENSG00000186862
65744177	65744177	2c53aa67933d	ENSG00000170959
65744178	65744178	2c53aa67933d	ENSG00000107874
65744179	65744179	2c53aa67933d	ENSG00000166012

65744180 rows × 3 columns

3) metadata_cite_day_2_donor_27678.csv(234kb)

Περιέχει ακριβώς τα στοιχεία που περιλαμβάνονται στο αρχείο metadata.csv μόνο donor 27678 την 2η μέρα

Ένα χαρακτηριστικό στιγμιότυπο που άνοιξε με την χρήση jupyter είναι το παρακάτω:

	cell_id	day	donor	cell_type	technology
0	83d6659a6a32	2	27678	NeuP	citeseq
1	d98594f13d2e	2	27678	NeuP	citeseq
2	5f93d8ffc72f	2	27678	NeuP	citeseq
3	7dfa2699d351	2	27678	EryP	citeseq
4	6d2533edd0e0	2	27678	HSC	citeseq
...	...	...	...	...	...
7011	be92120b3a00	2	27678	HSC	citeseq
7012	396d0c31d41c	2	27678	HSC	citeseq
7013	ef6bf272cdcf	2	27678	EryP	citeseq
7014	6339da0de3a0	2	27678	HSC	citeseq
7015	397bef68ded6	2	27678	HSC	citeseq

7016 rows × 5 columns

Τα υπόλοιπα αρχεία που περιλαμβάνονταν στην υπόθεση εργασίας μας ήταν της μορφής h5. Η χρήση της μορφής αυτής για την αποθήκευση στων δεδομένων υιοθετήθηκε γιατί βοηθάει στην οργάνωση και αποθήκευση μεγάλου όγκου και σύνθετων δεδομένων , έχει ιεραρχική δομή , χαρακτηρίζεται από ευελιξία αφού μπορούν να αποθηκευτούν διαφορετικοί τύποι δεδομένων, παρέχει την δυνατότητα συμπίεσης των δεδομένων κτλ.

Για το άνοιγμα των αρχείων αυτών στο jupyter υπάρχει η δυνατότητα κατεβάσματος και εισαγωγής της h5py library ώστε να υπάρξει δυνατότητα χρήσης και επεξεργασίας τους.

4) train_cite_inputs.h5(2.5 gb)

Οι γραμμές ανταποκρίνονται στα κύτταρα και οι στήλες στα γονίδια τα οποία είναι υπό την μορφή gene_name_gene ensemble_ids.

Η συσχέτιση των δύο αυτών στοιχείων με έναν αριθμό μας δείχνει την τιμή έκφρασης ενός γονιδίου σε ένα κύτταρο. Στα δεδομένα μας οι τιμές ξεκινούσαν από 0 που ουσιαστικά δεν υπάρχει συσχέτιση και έφταναν μέχρι 11.56.

5) train_cite_targets.h5(38,54 gb)

Οι γραμμές δείχνουν τα κύτταρα και οι στήλες τις πρωτεΐνες .

Οι μεταξύ τους τιμές αποτυπώνουν την τιμή της έκφρασης του επίπεδου πρωτεΐνης στην επιφάνεια του κυττάρου. Ξεκινούν από -26.3 και φτάνουν μέχρι 71.3. Οι αρνητικές τιμές ενώ δεν έχουν νόημα στην τεχνολογία cite seq, προκύπτουν επειδή τα δεδομένα έχουν περάσει και από dsb κανονικοποίηση όπου ο παράγοντας κλιμάκωσης για έναν συγκεκριμένο συνδυασμό γονιδίου και κυττάρου είναι μεγαλύτερος από την αντίστοιχη γονιδιακή έκφραση

6) test_cite_inputs.h5(1,7 gb)

Είναι το αρχείο που περιλαμβάνει τα δεδομένα που δεν θα χρησιμοποιηθούν κατά την εκπαίδευση του μοντέλου αλλά κατά την αξιολόγησή του.

7) test_cite_inputs_day_2_donor_27678.h5(307.96mb)

Είναι το αρχείο που περιλαμβάνει τα δεδομένα που δεν θα χρησιμοποιηθούν κατά την εκπαίδευση του μοντέλου αλλά κατά την αξιολόγησή του και συγκεκριμένα για τον δότη 27678 την 2^{η} μέρα.

8) train_multi_inputs.h5(11,33 gb)

Οι γραμμές είναι τα κύτταρα και οι στήλες ανταποκρίνονται στην τοποθεσία του γονιδιώματος .

Οι μεταξύ τους τιμές συσχέτισης δείχνουν το επίπεδο προσβασιμότητας του γονιδιώματος στην συγκεκριμένη τοποθεσία του γονιδίου σε ένα κύτταρο και κυμαίνονται από 0 μέχρι 17.4

9) train_multi_targets.h5(11,33 gb)

Οι γραμμές είναι τα κύτταρα και οι στήλες είναι τα γονίδια.

Οι μεταξύ τους τιμές συσχέτισης δείχνουν τα επίπεδα γονιδιακής έκφρασης RNA σε κάθε κύτταρο. Οι τιμές αυτές έχουν περάσει από κανονικοποίηση μεγέθους βιβλιοθήκης μια τεχνική που χρησιμοποιείται στην ανάλυση δεδομένων υψηλής εντατικότητας αλληλουχίας ώστε να ληφθούν υπόψιν οι διαφορές στο βάθος αλληλουχίας ή το μέγεθος της βιβλιοθήκης μεταξύ των δειγμάτων. Έχουν επίσης υποστεί και μετασχηματισμό με βάση την συνάρτηση log1p που αναφέρεται σε μια μετασχηματισμένη συνάρτηση που χρησιμοποιείται για την αντιμετώπιση περιπτώσεων που οι τιμές είναι κοντά στο 0 ή αρνητικές εξαλείφοντας το πρόβλημα των μη ορισμένων ή αρνητικών τιμών.

10) test_multi_inputs.h5(1,7 gb)

Είναι το αρχείο που περιλαμβάνει τα δεδομένα που δεν θα χρησιμοποιηθούν κατά την εκπαίδευση του μοντέλου αλλά κατά την αξιολόγησή του.

4. ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 ΜΕΘΟΔΟΛΟΓΙΑ

Η διπλωματική εργασία υλοποιήθηκε χρησιμοποιώντας την προγραμματιστική γλώσσα Python αξιοποιώντας το Jupyter ένα ανοιχτού κώδικα περιβάλλον που χρησιμοποιείται ευρέως για επιστημονικούς υπολογισμούς, ανάλυση δεδομένων και αναπαράσταση αποτελεσμάτων. Παρά τον μεγάλο όγκο των δεδομένων και της περιπλοκότητας τους δεν χρειάστηκε η αλλαγή περιβάλλοντος αλλά υλοποιήθηκε όλη η εργασία με αυτό το εργαλείο.

Ο κώδικας υλοποιήθηκε τμηματικά αρχικά για να ανοίξουν τα αρχεία που περιείχαν τα δεδομένα που αναφέρθηκαν παραπάνω. Για το άνοιγμα των h5 αρχείων κατέβηκε και εισάχθηκε στον Jupyter η βιβλιοθήκη h5py library ενώ για τα αρχεία csv δεν απαιτήθηκε κάποια διαφορετική βιβλιοθήκη. Έπειτα ακολούθησε η διαδικασία της προεργασίας των δεδομένων για τις δύο τεχνολογίες Citeseq και Multiome ξεχωριστά που θα αναφερθεί πιο αναλυτικά παρακάτω πως υλοποιήθηκε. Ακολούθησε η εκπαίδευση τριων μοντέλων Catboost , MLP και KNN για κάθε μια από τις δύο τεχνολογίες που έγινε μετά την όσο το δυνατόν πιο επιτυχή προεργασία του μεγάλου όγκου δεδομένων.

Καθ' όλη την διάρκεια της υλοποίησης της εργασίας έγινε χρήση και εισαγωγή των απαραίτητων βιβλιοθηκών (Numpy, Pandas , h5py, Scikit-learn). Σημαντικό σε αυτό το σημείο είναι να αναφερθεί ότι πως η βιβλιοθήκη Scikit-learn παρέχει μια ευρεία γκάμα εργαλείων , αλγορίθμων και λειτουργιών που βοηθούν στην εκπαίδευση, αξιολόγηση και εφαρμογή μοντέλων μηχανικής μάθησης με αποτελεσματικό τρόπο. Η διπλωματική εργασία βασίστηκε σε μεγάλο βαθμό στην χρήση των εργαλείων αυτών.

4.1.1. Η προεργασία των δεδομένων

Για την τεχνολογία Multiome

Εφαρμόστηκε τμήμα κώδικα που άνοιξε ξεχωριστά τα δύο αρχεία "train_multi_imputs.h5" και "train_multi_targets.h5" με την χρήση της βιβλιοθήκης h5py. Εμφάνισε τη λίστα των datasets που περιέχονταν στα δύο αρχεία καθώς και τα κλειδιά. Έπειτα αποθήκευσε την τιμή του κλειδιού "axis0" στην μεταβλητή "axis0", εμφάνισε το μήκος της και στην συνέχεια μετέτρεψε την μεταβλητή "axis0" σε πίνακα "axis0_arr" και εφάρμοσε τη λειτουργία decode για να

αποκωδικοποιήσει τις τιμές των στοιχείων του πίνακα σε αναγνώσιμη μορφή . Επιπλέον αποθήκευσε την τιμή του κλειδιού "axis1" στην μεταβλητή "axis1" και ακολούθησε την ίδια διαδικασία σε μετατροπή σε πίνακα "axis1_arr" , εμφάνισε το μήκος της και τέλος αποθήκευσε και εμφάνισε τις τιμές των κλειδιών "block0_items" και "block0_values" στις μεταβλητές "block0_items_arr" και "block0_values_arr". Το ίδιο τμήμα κώδικα αξιοποιήθηκε και για τα δύο αρχεία train_multi_imputs.h5 και train_multi_targets.h5 και βοήθησε στην κατανόηση και εξερεύνηση των δεδομένων που περιέχονται στα δύο αρχεία.

Στην συνέχεια αξιοποιήθηκε κώδικας που δημιούργησε ένα DataFrame με όνομα "df_train_target" χρησιμοποιώντας την βιβλιοθήκη pandas. Το DataFrame περιέχει τα δεδομένα από το "block0_values_arr" με τις στήλες που ορίζονται από το "axis0_arr" και τα δείγματα που ορίζονται από το "axis1_arr", όρισε επίσης τα ονόματα των στηλών του DataFrame στη μεταβλητή "target_columns". Αντίστοιχη διαδικασία ακολουθήθηκε και για το "df_train_input". Για να επιβεβαιωθεί ότι τα δύο αυτά DataFrames άρα και οι πίνακες είναι σε αντιστοιχία δόθηκε εντολή εκτύπωσης του μήκους των dataframes όπου εμφανίστηκε ο ίδιος αριθμός (5000) κάτι που επιβεβαίωσε ότι υπάρχει ομοιομορφία μεταξύ των εισόδων και των αντίστοιχων εξόδων που θα δώσουμε στα μοντέλα μας.

Για την περαιτέρω διαδικασία προεπεξεργασίας των δεδομένων πριν την εκπαίδευση των μοντέλων μηχανικής μάθησης χρησιμοποιήθηκε κομμάτι κώδικα που επιτέλεσε τις παρακάτω ενέργειες:

- Κανονικοποίηση των δεδομένων εισόδου (Input Normalization): Εφαρμόζεται η κανονικοποίηση L2 στα δεδομένα εισόδου df_train_input.. Η κανονικοποίηση L2 είναι ευρέως χρησιμοποιούμενη στην μηχανική μάθηση γιατί διατηρεί τις ιδιότητες κατανομής των δεδομένων και εξαλείφει την επίδραση της κλίμακας στους αλγόριθμους μάθησης. Κατά την χρήση της μεθόδου αυτής τα δεδομένα προσαρμόζονται ώστε η Ευκλείδια απόσταση από την αρχή των αξόνων (0,0) μέχρι το κάθε διάνυσμα να είναι 1. Τα δεδομένα διαιρούνται με την τετραγωνική ρίζα του αθροίσματος των τετραγώνων των τιμών σε κάθε διάσταση. Στην συγκεκριμένη περίπτωση η κανονικοποίηση έγινε ανά γραμμή (axis=1) και δημιουργληθηκε ένα νέο dataframe df_norm με κανονικοποιημένες τιμές

- Αφαίρεση της μέσης τιμής (Mean Subtraction): Από το df_norm αφαιρείται η μέση τιμή ανά στήλη κατά μήκος των γραμμών για να δημιουργηθεί το dataframe df_median. Η διαδικασία αυτή είναι χρήσιμη γιατί εξαλείφει την αλληλεξάρτηση των χαρακτηριστικών των δεδομένων, κανονικοποιεί την κλίμακα των χαρακτηριστικών ιδιαίτερα σε αλγόριθμους που είναι βασισμένοι στην απόσταση συνολικά είναι σημαντικό βήμα στην προεργασία δεδομένων γιατί μπορεί να βελτιώσει την απόδοση και της σταθερότητα των μοντέλων.
- Εφαρμόζεται ο αλγόριθμος TruncatedSVD για να μειωθεί η διάσταση των δεδομένων εισόδου σε 128 συνιστώσες. Το νέο dataframe μετά τη μείωση της διάστασης ονομάζεται df_tsvd. Έχει ειπωθεί παραπάνω η σημασία του αλγόριθμου αυτού.
- Ταξινόμηση κατά δείγμα (Sort by Index): Τα πλαίσια δεδομένων df_train_input ταξινομούνται κατά αύξουσα σειρά με βάση τον δείκτη index. Ο δείκτης αυτός στην συγκεκριμένη περίπτωση είναι ο df_median.
- Κανονικοποίηση εξόδου (Target Normalization): Εφαρμόζεται η ίδια διαδικασία κανονικοποίησης και μείωσης διαστάσεων στα δεδομένα εξόδου df_train_target.
- Αφαιρείται επίσης η μέση τιμή ανά στήλη κατά μήκος των γραμμών.
- Εφαρμόζεται ο αλγόριθμος TruncatedSVD όπως και παραπάνω
- Επαναταξινόμηση κατά δείγμα (Sort by Index): Τα dataframes df_train_input και df_train_target ταξινομούνται κατά αύξουσα σειρά με βάση τον δείκτη index που είναι ο df_median
- Συγχώνευση (Merge): Τα πλαίσια δεδομένων df_train_input και df_train_target συγχωνεύονται με βάση τον κοινό δείκτη (index) και αποθηκεύονται στον νέο πινακα δεδομένων df_merged.
- Επιτυγχάνεται η εξαγωγή ονομάτων στηλών εξόδου (Target Columns): Βρίσκονται οι στήλες στο df_train_target που περιέχουν το χαρακτηριστικό _t στο όνομά τους και αποθηκεύονται στη λίστα target_columns.
- Απομόνωση εισόδου και εξόδου (Input and Output Separation): Το πλαίσιο δεδομένων df_merged χωρίζεται σε δύο υποπλαίσια δεδομένων X και y, όπου το X περιλαμβάνει τις στήλες εισόδου και το y περιλαμβάνει τις στήλες εξόδου που βρίσκονται στη λίστα target_columns. Στον παρακάτω πίνακα αποτυπώνεται το παράδειγμα 5 γραμμών και 128 στηλών των πινάκων X και ψ όπου για το ίδιο κύτταρο σε κάθε γραμμή αντίστοιχα όπου στον πίνακα X βλέπουμε για κάθε κύτταρο την τιμή που επιπέδου

προσβασιμότητας ενός γονιδιώματος και στον πίνακα y για κάθε ίδιο κύτταρο τα επίπεδα γονιδιακής έκφρασης RNA.

	0_i	1_i	2_i	3_i	4_i	5_i	6_i	7_i	8_i	9_i	...	118_i	119_i	120_i	
0027af31b8e6	0.181153	-0.027401	0.024284	0.016994	-0.012219	0.002101	-0.000705	-0.020737	0.010130	-0.004257	...	-0.004447	0.000317	0.015327	0.0
00393c8f0613	0.238334	-0.009519	0.076242	-0.014547	-0.003825	0.025063	0.025620	0.009386	-0.016551	-0.013042	...	0.003929	-0.019417	0.022882	-0.0
0043276e1d8d	0.136304	-0.003836	0.005961	-0.019946	-0.031509	0.003577	0.010530	0.002389	-0.005332	-0.003003	...	0.009798	-0.012348	-0.032821	0.0
004717e0d2a6	0.305095	-0.011865	0.072126	-0.002247	0.028768	0.001748	0.005004	0.000620	-0.018646	0.006892	...	0.001628	-0.007830	-0.006169	0.0
0072e49abd3f	0.201238	-0.037947	0.068900	0.023056	0.015214	0.018078	-0.001881	0.001169	-0.017049	0.002197	...	0.016782	0.000302	-0.006824	-0.0

5 rows × 128 columns

```
y.head()
```

	0_t	1_t	2_t	3_t	4_t	5_t	6_t	7_t	8_t	9_t	...	118_t	119_t	120_t	
0027af31b8e6	0.411531	-0.046144	0.050472	0.071255	-0.049953	-0.001651	0.003802	-0.008149	0.016287	0.002196	...	0.010286	0.020526	-0.018358	0.01
00393c8f0613	0.416983	-0.093569	-0.027570	0.041462	0.092308	0.061098	-0.040787	-0.014937	0.012732	-0.000417	...	-0.000128	-0.002847	0.008114	0.00
0043276e1d8d	0.374682	0.009108	-0.012661	0.000910	0.014156	-0.040793	-0.021196	-0.049929	0.010495	-0.002096	...	0.026712	-0.007196	0.002464	0.01
004717e0d2a6	0.436726	-0.099591	-0.020397	0.041657	0.028528	-0.009814	-0.016125	0.009522	0.002834	-0.020781	...	-0.009645	-0.013474	-0.018086	0.00
0072e49abd3f	0.409511	-0.035211	-0.022421	0.090624	0.072089	0.059848	-0.001700	0.019652	0.019127	-0.007321	...	0.004286	-0.009124	-0.000365	0.00

5 rows × 128 columns

Τέλος το κομμάτι κώδικα εμφάνισε τις τιμές len(df_train_input) και len(df_train_target) που δείχνουν τον αριθμό των δειγμάτων στους πίνακες δεδομένων εισόδου και εξόδου αντίστοιχα. Οι δύο αριθμοί είναι ίσοι, που σημαίνει ότι ο αριθμός των δειγμάτων στους δύο πίνακες είναι ίδιος και επομένως μπορούν να χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης.

Σε αυτή την φάση ολοκληρώθηκε η διαδικασία της προεργασίας των δεδομένων για την τεχνολογία Multiome.

Για την τεχνολογία CiteSeq

Η διαδικασία που ακολουθήθηκε για την προεργασία των δεδομένων σε αυτή τη τεχνολογία δεν διαφέρει σε μεγάλο βαθμό από αυτή που περιεγράφηκε παραπάνω για την τεχνολογία Multiome. Για τα δεδομένα εισόδου χρησιμοποιήθηκε το αρχείο “train_cite_imputs.h5” και για τα δεδομένα εξόδου το αρχείο “train_cite_targets.h5”. Τα αρχεία άνοιξαν και διαβάστηκαν

όπως περιεγράφηκε παραπάνω με την χρήση της βιβλιοθήκης h5py. Έπειτα δημιουργήθηκαν τα αντίστοιχα dataframes.

Για το dataframe "df_train_input" που αποτυπώνει τις τιμές εισόδου, ένα δείγμα των δεδομένων φαίνεται παρακάτω :

	ENSG00000094914_AAAS	ENSG00000081760_AACS
45006fe3e4c8	0.000000	0.000000
d02759a80ba2	0.000000	0.000000
c016c6b0efa5	0.000000	3.847321
ba7f733a4f75	3.436846	3.436846
fbcf2443ffb2	0.000000	4.196826
...	...	...
1a327f97d933	0.000000	0.000000
b5f0bc8be222	0.000000	0.000000
d5a8f9ad65db	4.742704	0.000000
671d60cb0b14	0.000000	0.000000
cc270cbee870	3.481027	0.000000

	ENSG00000109576_AADAT	ENSG00000103591_AAGAB
45006fe3e4c8	0.000000	0.000000
d02759a80ba2	0.000000	0.000000
c016c6b0efa5	3.847321	0.000000
ba7f733a4f75	0.000000	0.000000
fbcf2443ffb2	0.000000	0.000000

Όπου οι γραμμές είναι τα κύτταρα και οι στήλες τα γονιδιώματα με την μορφή gene_name_gene ensemble_ids και οι μεταξύ τους τιμές δείχνουν την έκφραση ενός γονιδιώματος σε ένα κύτταρο

Για το dataframe "df_train_target" που αποτυπώνει τις τιμές εξόδου, ένα δείγμα των δεδομένων φαίνεται παρακάτω :

	CD86	CD274	CD270	CD155	CD112	CD47
45006fe3e4c8	1.167804	0.622530	0.106959	0.324989	3.331674	6.426002
d02759a80ba2	0.818970	0.506009	1.078682	6.848758	3.524885	5.279456
c016c6b0efa5	-0.356703	-0.422261	-0.824493	1.137495	0.518924	7.221962
ba7f733a4f75	-1.201507	0.149115	2.022468	6.021595	7.258670	2.792436
fbcf2443ffb2	-0.100404	0.697461	0.625836	-0.298404	1.369898	3.254521

	CD48	CD40	CD154	CD52	...	CD94	\
45006fe3e4c8	1.480766	-0.728392	-0.468851	-0.073285	...	-0.448390	
d02759a80ba2	4.930438	2.069372	0.333652	-0.468088	...	0.323613	
c016c6b0efa5	-0.375034	1.738071	0.142919	-0.971460	...	1.348692	
ba7f733a4f75	21.708519	-0.137913	1.649969	-0.754680	...	1.504426	
fbcf2443ffb2	-1.659380	0.643531	0.902710	1.291877	...	0.777023	

Όπου οι γραμμές δείχνουν τα κύτταρα και οι στήλες τις πρωτεΐνες. Οι μεταξύ τους τιμές αποτυπώνουν την τιμή της έκφρασης του επίπεδου πρωτεΐνης στην επιφάνεια του κυττάρου.

Στην συνέχεια χρησιμοποιήθηκε κομμάτι κώδικα αντίστοιχου με την τεχνολογία multiome αξιοποιώντας τους ίδιους αλγόριθμους για να καταλήξουμε σε δύο πίνακες X και y όπως αποτυπώνονται παρακάτω:

```
In [18]: X.head()
```

Out[18]:

	0_i	1_i	2_i	3_i	4_i	5_i	6_i	7_i	8_i	9_i	...	118_i	119_i	120_i	
0005b45cd15b	0.336289	-0.075449	-0.031959	-0.037601	-0.056768	0.061093	0.042131	-0.034148	-0.029560	0.003118	...	-0.011114	-0.010847	-0.003503	-0.
00207000d28a	0.347569	-0.012736	0.084567	0.116702	-0.021714	-0.048811	-0.057553	-0.075695	0.086429	-0.017207	...	0.019945	0.007140	-0.002688	-0.
003573cc08c6	0.344860	-0.104732	0.015988	0.064522	-0.042115	-0.043898	-0.005488	0.010995	-0.057977	0.021604	...	0.006630	-0.007997	0.001530	-0.
003f622bda01	0.321715	-0.059371	0.023251	0.055557	0.003123	-0.031986	-0.038045	0.034664	-0.020813	-0.053628	...	0.016052	-0.005806	0.001473	-0.
003fe3679efa	0.315571	0.135698	0.078503	0.001585	-0.003489	-0.045698	0.061781	0.013933	0.096001	0.002364	...	-0.015296	-0.002111	0.016960	-0.

5 rows × 128 columns

```
In [19]: y.head()
```

Out[19]:

	0_t	1_t	2_t	3_t	4_t	5_t	6_t	7_t	8_t	9_t	...	118_t	119_t	120_t	
0005b45cd15b	-0.063050	0.042113	-0.070843	0.161716	-0.041039	0.018276	-0.068818	-0.069538	-0.073775	-0.017696	...	0.013379	-0.004716	0.003785	-0.
00207000d28a	0.228945	0.172125	0.063207	0.096174	-0.087438	0.046270	-0.100316	-0.030770	0.065108	0.065382	...	0.015839	-0.020271	-0.014295	0.
003573cc08c6	0.624723	0.089594	0.083770	0.000583	0.090219	-0.082379	-0.072608	-0.058128	0.049151	-0.114025	...	0.015107	0.019738	-0.017817	-0.
003f622bda01	0.268530	-0.227379	-0.005710	0.032999	-0.117642	0.103103	-0.023070	-0.065778	0.013965	0.058114	...	-0.005214	0.010839	-0.021877	0.
003fe3679efa	-0.011144	-0.028758	-0.016312	0.052103	0.034297	-0.072243	-0.082869	-0.082715	0.188898	0.083966	...	-0.025154	-0.001049	0.001723	0.

5 rows × 128 columns

Όπου για τα ίδια κύτταρα από το πίνακα X παίρνουμε την έκφραση ενός γονιδιώματος στο κύτταρο και από τον πίνακα y την έκφραση πρωτεΐνης στην επιφάνεια του κυττάρου.

Τέλος το κομμάτι κώδικα εμφάνισε τις τιμές len(df_train_input) και len(df_train_target) που ήταν ίσες επομένως ο αριθμός των δειγμάτων στους δύο πίνακες είναι ίδιος και συνεπώς μπορούν να χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης.

Σε αυτή την φάση ολοκληρώθηκε η διαδικασία της προεργασίας των δεδομένων για την τεχνολογία Citeseq.

Με την ολοκλήρωση της προεργασίας των δεδομένων και για τις δύο τεχνολογίες περάσαμε στην χρησιμοποίηση των μοντέλων CatBoost και MLP για τα οποία έχει γίνει εκτενής αναφορά στο θεωρητικό υπόβαθρο παραπάνω. Από προγραμματιστικής άποψης και για τις δύο τεχνολογίες χρησιμοποιήθηκαν οι ίδιες προγραμματιστικές εντολές για να εφαρμοστούν τα μοντέλα μηχανικής μάθησης.

4.1.2 Τμήματα κώδικα για το μοντέλο MLP

Στο 1ο μέρος ακολουθήθηκαν τα εξής βήματα:

-Εισάγονται οι απαραίτητες βιβλιοθήκες, όπως ο MLPRegressor από το scikit-learn, καθώς και συναφείς μετρικές και στατιστικές συναρτήσεις.

-Πραγματοποιείται διαχωρισμός των δεδομένων εισόδου (X) και των δεδομένων εξόδου (y) σε σύνολα εκπαίδευσης (train) και ελέγχου (test), με αναλογία 80:20.

-Ορίζεται το πλέγμα παραμέτρων (param_grid) που πρέπει να εξεταστούν στο Grid Search. Η Grid Search είναι μια μέθοδος υπερπαραμετροποίησης που επιτρέπει την αυτόματη εύρεση των βέλτιστων παραμέτρων ενός μοντέλου. Στον κώδικα, ορίζονται τα πιθανά σύνολα τιμών για τις παραμέτρους του MLPRegressor και εκτελείται η Grid Search για την αξιολόγηση και επιλογή των βέλτιστων παραμέτρων. Ορίζονται πιθανές τιμές για το μέγεθος των κρυφών επιπέδων, τη συνάρτηση ενεργοποίησης, τον αλγόριθμο επίλυσης, τον αριθμό των επαναλήψεων (max_iter) και το παράγοντα ποινής (alpha).

-Δημιουργείται ένα αντικείμενο MLPRegressor ως μοντέλο για το Grid Search.

-Ορίζεται ένα αντικείμενο GridSearchCV για το Grid Search, με το μοντέλο MLPRegressor και το πλέγμα παραμέτρων που ορίστηκε προηγουμένως. Το cv=5 υποδηλώνει ότι η διαδικασία του Grid Search θα εκτελεστεί με διασπορά 5-fold cross-validation.

-Εκτελείται η διαδικασία του Grid Search, κατά την οποία αξιολογούνται όλοι οι συνδυασμοί των παραμέτρων στα δεδομένα εκπαίδευσης.

-Εκτυπώνονται οι καλύτερες παράμετροι (best params) που επιτεύχθηκαν και ο καλύτερος βαθμός R2 (best r2 score) που επιτεύχθηκε από το Grid Search.

Στην συνέχεια για το 2ο μέρος ακολουθήθηκαν τα παρακάτω βήματα:

-Εισήχθησαν οι απαραίτητες βιβλιοθήκες, όπως ο MLPRegressor από το scikit-learn, καθώς και συναφείς μετρικές και στατιστικές συναρτήσεις.

-Πραγματοποιήθηκε διαχωρισμός των δεδομένων εισόδου (X) και των δεδομένων εξόδου (y) σε σύνολα εκπαίδευσης (train) και ελέγχου (test), με αναλογία 80:20.

-Δημιουργήθηκε ένα MLPRegressor μοντέλο με τις καλύτερες παραμέτρους που βρέθηκαν από το Grid Search.

-Το μοντέλο εκπαιδεύτηκε στα δεδομένα εκπαίδευσης (X_train, y_train).

-Προβλέπονται οι τιμές των δεδομένων εξόδου για τα δεδομένα ελέγχου (X_test) χρησιμοποιώντας το εκπαιδευμένο μοντέλο.

-Υπολογίζονται οι μετρικές mean absolute error (MAE) και mean squared error (MSE) ανάμεσα στις πραγματικές τιμές (y_test) και τις προβλεπόμενες τιμές (mlp_y_preds).

-Υπολογίζεται η μέση απώλεια συσχέτισης (mean correlation loss) για κάθε στήλη των δεδομένων εξόδου, με βάση τον συντελεστή συσχέτισης Pearson ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες τιμές.

Η εκτύπωση των αποτελεσμάτων παρέχει πληροφορίες για την απόδοση του μοντέλου σε σχέση με τις μετρικές αξιολόγησης MAE, MSE και τη μέση απώλεια συσχέτισης.

4.1.3 Τμήματα κώδικα για το μοντέλο CatBoost

Ο κώδικας ξεκινά Grid search για το CatBoost Regressor μοντέλο. Οι ενέργειες που πραγματοποιούνται είναι οι εξής:

-Εισάγονται οι απαραίτητες βιβλιοθήκες, όπως ο CatBoostRegressor από το CatBoost και άλλες σχετικές βιβλιοθήκες.

-Πραγματοποιείται διαχωρισμός των δεδομένων εισόδου (X) και των δεδομένων εξόδου (y) σε σύνολα εκπαίδευσης (train) και ελέγχου (test), με αναλογία 80:20.

-Τα δεδομένα εκπαίδευσης (X_train, y_train) φορτώνονται σε ένα αντικείμενο Pool του CatBoost.

-Ορίζεται το πλέγμα υπερπαραμέτρων που θα αναζητηθεί, περιλαμβάνοντας παράμετρους όπως ο αριθμός των επαναλήψεων (iterations), η βάθος του δέντρου (depth), ο ρυθμός μάθησης (learning_rate), η κανονικοποίηση L2 (l2_leaf_reg) και ο αριθμός ορίων (border_count).

-Δημιουργείται ένα CatBoostRegressor μοντέλο με τις βέλτιστες παραμέτρους που προκύπτουν από την Grid search αναζήτηση.

-Πραγματοποιείται η grid search με χρήση της διασταυρούμενης επικύρωσης (cross-validation) και χρησιμοποιώντας τον αρνητικό μέσο τετραγωνικό όρο σφάλματος (neg_mean_squared_error) ως μέτρο αξιολόγησης.

-Εκτυπώνονται οι καλύτερες παράμετροι που βρέθηκαν από την grid search καθώς και ο βέλτιστος αρνητικός μέσος τετραγωνικός όρος σφάλματος (best negative mean squared error).

Αφού βρεθούν οι βέλτιστες παράμετροι η εκτέλεση του κώδικα συνεχίζεται ως εξής:

Ο κώδικας εκτελεί το μοντέλο CatBoostRegressor με τις βέλτιστες παραμέτρους που βρέθηκαν από την αναζήτηση πλέγματος. Οι ενέργειες που πραγματοποιούνται είναι οι εξής:

-Πραγματοποιείται διαχωρισμός των δεδομένων εισόδου (X) και των δεδομένων εξόδου (y) σε σύνολα εκπαίδευσης (train) και ελέγχου (test), με αναλογία 80:20.

-Δημιουργείται ένα CatBoostRegressor μοντέλο με τις βέλτιστες παραμέτρους που προέκυψαν από την grid search με προκαθορισμένες παραμέτρους iterations=1000, depth=8, rate=0.1, L2_leaf_reg=5, border count= 128

-Εκπαιδεύεται το μοντέλο στα δεδομένα εκπαίδευσης (X_train, y_train) με τη χρήση της μεθόδου fit. Η μέθοδος fit είναι μια ευρέως χρησιμοποιούμενη μέθοδος κατά την διάρκεια εκπαίδευσης των μοντέλων που χρησιμοποιείται για να βρει τις βέλτιστες παραμέτρους. Περιλαμβάνει μια επαναληπτική διαδικασία βελτιστοποίησης με στόχο την εύρεση παραμέτρων που ελαχιστοποιούν την διαφορά μεταξύ των προβλεπόμενων εξόδων και των πραγματικών τιμών στόχου των δεδομένων.

-Πραγματοποιείται πρόβλεψη στα δεδομένα ελέγχου (X_test) με τη χρήση της μεθόδου predict. Αφού έχει χρησιμοποιηθεί η μέθοδος fit έπειτα καλείται η μέθοδος predict ώστε να γίνουν οι προβλέψεις στα νέο εισερχόμενα δεδομένα. Δέχεται σαν είσοδο τα δεδομένα αυτά και επιστρέφει τις προβλεπόμενες τιμές.

-Υπολογίζονται οι μετρικές απόδοσης του μοντέλου, όπως η απόλυτη μέση απόκλιση σφάλματος (mae) και ο μέσος τετραγωνικός όρος σφάλματος (mse) μεταξύ των πραγματικών τιμών y_test και των προβλεπόμενων τιμών y_pred

-Υπολογίζεται η μέση απώλεια συσχέτισης (Mean Correlation Loss) για κάθε στήλη του y_test και y_pred. Η απώλεια συσχέτισης λογίζεται ως αφαίρεση της συσχέτισης μεταξύ των

πραγματικών τιμών και των προβλεπόμενων τιμών για κάθε στήλη. Οι απώλειες συσχέτισης αποθηκεύονται στη λίστα corr_lost.

-Εκτυπώνονται στην οθόνη οι mae, mse και mean correlation loss.

4.1.3 Τμήματα κώδικα για το μοντέλο K-Nearest Neighbors (KNN)

Για την υλοποίηση του συγκεκριμένου μοντέλου πραγμτοποιούνται οι παρακάτω ενέργειες :

- Γίνεται η εισαγωγή των απαραίτητων βιβλιοθηκών όπως numpy για αριθμητικές λειτουργίες, το train_test_split για τη διαίρεση του συνόλου δεδομένων, το KNeighborsRegressor για την πραγματοποίηση παλινδρόμησης με KNN και μερικές μετρικές από τη βιβλιοθήκη sklearn.metrics (mean_squared_error, r2_score) για τον έλεγχο της απόδοσης του μοντέλου.

- Δημιουργείται ένα αντικείμενο regressor του KNN (knn_regressor) με 8 γείτονες.

- Εκπαιδεύεται το regressor του KNN στα δεδομένα εκπαίδευσης (X_train και y_train) χρησιμοποιώντας τη μέθοδο fit.

- Χρησιμοποιεί το εκπαιδευμένο regressor του KNN για να πραγματοποιήσει προβλέψεις στα δεδομένα ελέγχου (X_test) και αποθηκεύει τις προβλέψεις στη μεταβλητή y_pred_knn.

- Υπολογίζει τις μετρικές παλινδρόμησης όπως και στα προηγούμενα μοντέλα ώστε να διαπιστωθεί η αποτελεσματικότητα του μοντέλου

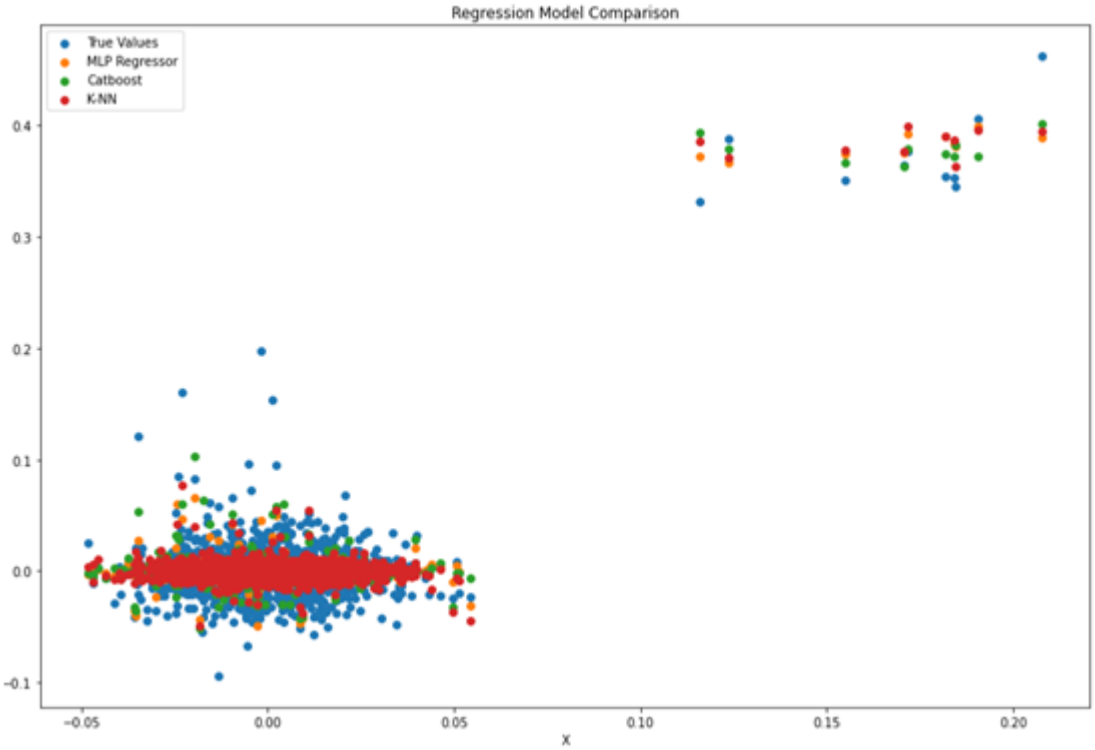
-Εκτυπώνονται στην οθόνη οι mae, mse και mean correlation loss.

4.2ΑΠΟΤΕΛΕΣΜΑΤΑ

Αφού έχουν τρέξει τα τρία μοντέλα για τις δύο διαφορετικές τεχνολογίες μπορούμε να δούμε τα αποτελέσματα που παράγονται για ένα δείγμα δεδομένων.

Για την τεχνολογία Multiome

Σε δείγμα 1000 δεδομένων η σύγκριση μεταξύ των μοντέλων Μηχανικής Μάθησης αποτυπώνεται στο παρακάτω διάγραμμα.

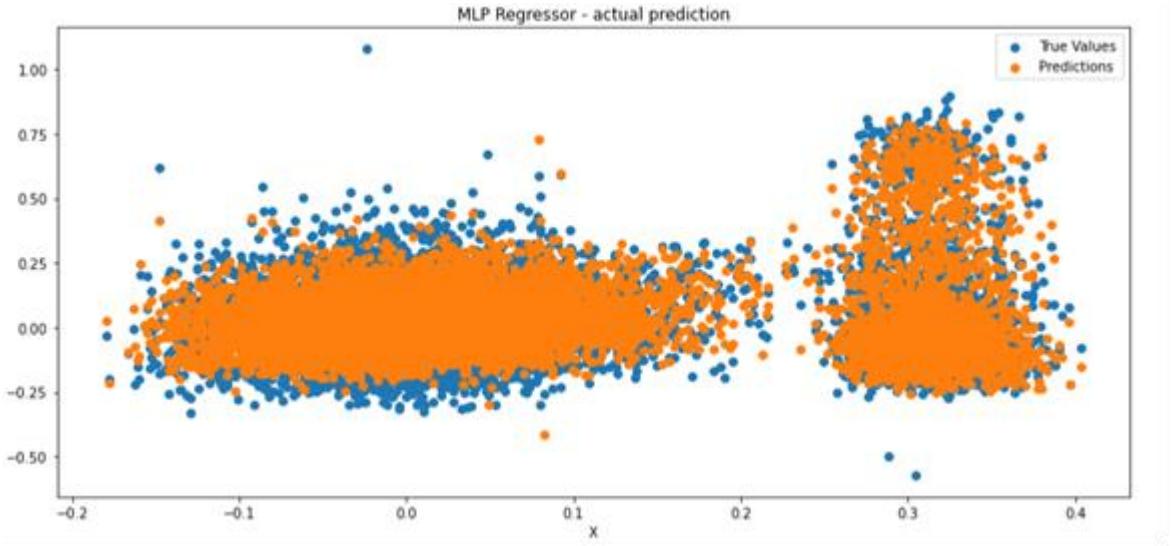


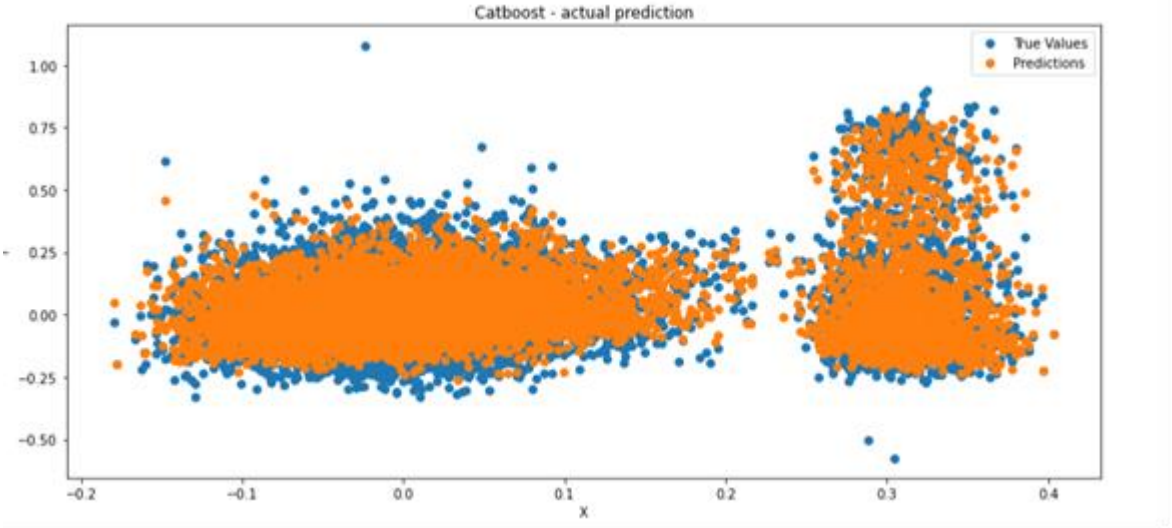
Ενώ όσον αφορά τα αριθμητικά αποτελέσματα των δεικτών απόδοσης για 5000 δεδομένα αποτυπώνονται στον ακόλουθο πίνακα:

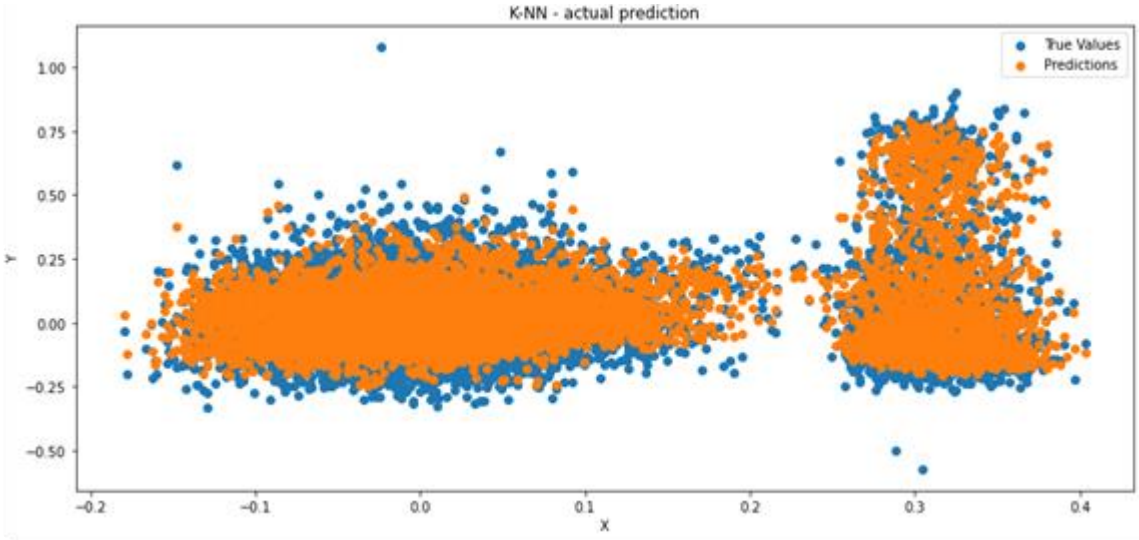
Μοντέλο	Δείκτες Απόδοσης		
	mae	mse	Mean Colleration
MLP	0.012	0.0003	0.94
Catboost	0.012	0.0002	0.9
KNN	0.013	0.0003	0.91

Για την τεχνολογία CITESEQ

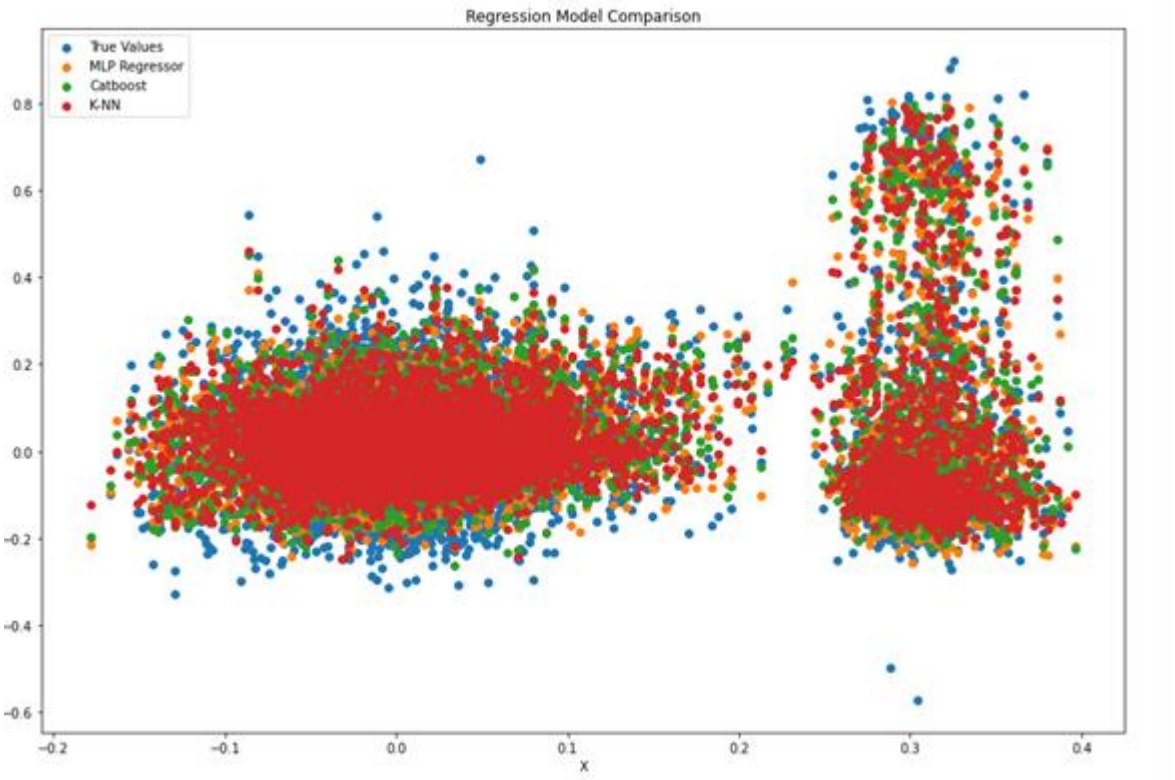
Σε δείγμα 2000 δεδομένων αποτυπώνεται παρακάτω η σύγκριση μεταξύ των πραγματικών τιμών που εισήχθησαν στα μοντέλα και των προβλεπόμενων τιμών:


MLP Regressor - actual prediction
True Values
Predictions
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
X


Catboost - actual prediction
True Values
Predictions
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
X


K-NN - actual prediction
True Values
Predictions
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Y
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
X

Ενώ τα συγκριτικά αποτελέσματα μεταξύ των μοντέλων αποτυπώνονται στο παρακάτω διάγραμμα:



Σε σχέση με τα αριθμητικά αποτελέσματα των δεικτών απόδοσης των μοντέλων για δείγμα 10000 δεδομένων εξάγονται τα παρακάτω αποτελέσματα:

Μοντέλο	Δείκτες Απόδοσης		
	mae	mse	Mean Colleration
MLP	0.024	0.0012	0.88
Catboost	0.024	0.0012	0.86
KNN	0.025	0.0013	0.88

Και οι 3 παραπάνω δείκτες απόδοσης των μοντέλων Μηχανικής Μάθησης σε επιβλεπόμενη μάθηση με την τεχνική της παλινδρόμησης υποδεικνύουν με διαφορετικούς τρόπους την διαφορά ανάμεσα στις προβλεπόμενες και τις πραγματικές τιμές. Στον παρακάτω πίνακα ομαδοποιούνται τα χαρακτηριστικά των τριών αυτών δεικτών Mean Absolute Error (MAE), Mean Squared Error (MSE) και Mean Colleration Loss.

MAE	MSE	Mean Colleration Loss
• Μετρά την μέση απόλυτη διαφορά ανάμεσα στις προβλεπόμενες και τις πραγματικές τιμές • Δείχνει τον μέσο όρο του μεγέθους των σφαλμάτων ανάμεσα στις προβλεπόμενες τιμές και τις πραγματικές τιμές. • Παρέχει έναν απλό μέτρο για την ακρίβεια του μοντέλου, όπου χαμηλότερες τιμές του MAE υποδηλώνουν καλύτερη απόδοση • είναι λιγότερο ευαίσθητο στις ακραίες τιμές • δεν παρέχει πληροφορίες σχετικά με την κατεύθυνση των σφαλμάτων, αλλά μόνο το μέγεθός τους.	• Μετρά τον μέσο τετραγωνικό όρο των διαφορών ανάμεσα στις προβλεπόμενες τιμές και τις πραγματικές τιμές. • Δείχνει το μέσο τετραγωνικό μέγεθος των σφαλμάτων ανάμεσα στις προβλεπόμενες τιμές και τις πραγματικές τιμές • μετριέται σε μονάδες της τετραγωνικής μονάδας της μεταβλητής πρόβλεψης. • Επιδέχεται μεγέθυνση από τα μεγάλα σφάλματα και είναι ευαίσθητο σε ακραίες τιμές	• μετρά τη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών υπολογίζοντας το συντελεστή συσχέτισης μεταξύ τους • Ο συντελεστής συσχέτισης μετρά τη γραμμική σχέση μεταξύ δύο μεταβλητών, κυμαίνεται από -1 (τέλεια αρνητική συσχέτιση) έως 1 (τέλεια θετική συσχέτιση) • Μια απώλεια συσχέτισης 0 υποδηλώνει μια τέλεια συσχέτιση μεταξύ των προβλεπόμενων και των πραγματικών τιμών • Αντικατοπτρίζει την ικανότητα του μοντέλου να αντιληφθεί τα

		υποκείμενα πρότυπα και σχέσεις στα δεδομένα

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι δείκτες που αναφέρθηκαν αναλυτικά παραπάνω αναδεικνύουν με διαφορετικούς τρόπους την αποτελεσματικότητα των μοντέλων μηχανικής μάθησης με την τεχνική της παλινδρόμησης.

Με βάση τις μετρήσεις που προέκυψαν και παίρνοντας σαν δεδομένο ότι για την προεργασία των δεδομένων τόσο εισόδου όσο και ελέγχου χρησιμοποιήθηκαν και για τις δύο τεχνολογίες και για τα τρία μοντέλα οι ίδιες διαδικασίες μπορούμε να διεξάγουμε το συμπέρασμα ότι τα μοντέλα μηχανικής μάθησης για δεδομένο δείγμα δεδομένων σημείωσαν επιτυχία ως προς την απόδοσή τους.

Στην τεχνολογία Multiome οι μετρήσεις των δεικτών απόδοσης (MAE, MSE, Mean Correlation Loss) είναι παρόμοιες και στα τρια μοντέλα μηχανικής μάθησης με τον MAE να βρίσκεται κοντά 0.013 που σημαίνει ότι οι προβλέψεις του μοντέλου αποκλίνουν από τις πραγματικές τιμές κατά 0.013 μονάδες που σημαίνει ότι βρίσκονται αρκετά κοντά , τον MSE να είναι ίδιος και στα τρία μοντέλα 0.003 που υποδεικνύει ότι η μέση τετραγωνική ρίζα των σφαλμάτων είναι αρκετά μικρή και άρα διαπιστώνεται μικρή διακύμανση μεταξύ προβλεπόμενων και πραγματικών τιμών. Τέλος όσον αφορά τον δείκτη Mean Correlation Loss που κινείται σε επίπεδα 0.9 -0.94 για τα τρία μοντέλα βρίσκεται δηλαδή κοντά το 1 υποδεικνύει ότι η συσχέτιση μεταξύ των πραγματικών και προβλεπόμενων τιμών είναι κοντά στην τέλεια θετική συσχέτιση.

Στην τεχνολογία CiteSeq οι μετρήσεις των δεικτών απόδοσης (MAE, MSE, Mean Correlation Loss) επίσης δεν διαφέρουν σημαντικά μεταξύ των 3 μοντέλων.

Σε περίπτωση επιλογής μεγαλύτερου δείγματος δεδομένων (εδώ επιλέχθηκε αυτό το πλήθος λόγω χωρητικότητας της μνήμης RAM) η πιθανότητα διαφοροποίησης μεταξύ των δεικτών απόδοσης μπορεί να διαφέρει σημαντικά ωστόσο δεν αναιρείται το συμπέρασμα ότι γενικότερα κινούνται σε παρόμοιες τιμές τα επιλεγμένα μοντέλα.

Συνοψίζοντας και τα τρία μοντέλα κατάφεραν σε μεγάλο βαθμό να εκπαιδευτούν και να ανταποκριθούν σε δεδομένα τόσο υψηλών απαιτήσεων. Το σημαντικότερο ρόλο σε αυτό έπαιξε το βάρος που δόθηκε στην προεργασία των δεδομένων γιατί η ανάλυση των single cell

δεδομένων χαρακτηρίζονται από τεχνικό «θόρυβο», τεχνικές διακυμάνσεις που προκαλούνται από τις διεργασίες πειράματος και αλληλουχίας, σφάλματα μέτρησης, δεδομένα κυττάρων που μπορεί να χαρακτηρίζονται από χαμηλή ποιότητα και απαιτείται το φιλτράρισμά τους, ενσωματωμένες αποκλίσεις. Τα συγκεκριμένα δεδομένα είναι επίσης υψηλών διαστάσεων με έναν μεγάλο αριθμό δεδομένων που αναφέρονται σε γονίδια και χαρακτηριστικά τους. Επομένως είναι απαραίτητη η μείωση του αριθμού των διαστάσεων για την πιο αποτελεσματική προεπεξεργασία τους που επιτυγχάνεται με την κατάλληλη χρήση αλγορίθμων μείωσης διαστάσεων.

Τέλος η χρήση περισσότερων μοντέλων μηχανικής μάθησης που θα μπορούσαν να εφαρμοστούν στην συγκεκριμένη περίπτωση καθώς και ένα πιθανά μεγαλύτερο δείγμα δεδομένων και η σύγκριση των μετρικών απόδοσης μεταξύ τους θα μπορούσαν να οδηγήσουν σε ασφαλέστερα και περισσότερα συμπεράσματα δεδομένου ότι το κύριο βάρος χρειάζεται να δίνεται στην προεπεξεργασία των δεδομένων ώστε με πιο εύκολα διαχειρίσιμα δεδομένα να απλοποιείται η χρήση και λειτουργία των μοντέλων. Σε κάθε περίπτωση η χρήση μοντέλων μηχανικής μάθησης μπορεί να απελευθερώσει και πολλαπλασιάσει τις δυνατότητες μελέτης και έρευνας του μεγάλου όγκου των δεδομένων που έχουν προκύψει τα τελευταία χρόνια από τις μελέτες σε μεμονωμένα κύτταρα. Από τα αποτελέσματα της συγκεκριμένης διπλωματικής αναδεικνύονται οι τεράστιες δυνατότητες της Μηχανική Μάθησης στο πλαίσιο της βιολογίας και επιβεβαιώνουν την πεποίθηση ότι οι επιστήμονες μπορούν μελλοντικά να αξιοποιήσουν τις τεχνολογίες αυτές για την περαιτέρω εξέλιξη και ανάπτυξη της επιστήμης της Βιολογίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1) Fuchs, E., & Chen, T. (2012). A matter of life and death: Self-renewal in Stem Cells. *EMBO Reports*, *14*(1), 39–48. Ανακτήθηκε από: https://doi.org/10.1038/embor.2012.197
2) Jackson, C. A., & Vogel, C. (2022). New horizons in the Stormy Sea of multimodal single-cell data integration. *Molecular Cell*, *82*(2), 248–259. Ανακτήθηκε από: https://doi.org/10.1016/j.molcel.2021.12.012
3) Konturek-Ciesla, A., & Bryder, D. (2022). Stem Cells, hematopoiesis and lineage tracing: Transplantation-centric views and beyond. *Frontiers in Cell and Developmental Biology*, *10*. Ανακτήθηκε από: https://doi.org/10.3389/fcell.2022.903528
4) Källberg, J., Xiao, W., Van Assche, D., Baret, J.-C., & Taly, V. (2022). Frontiers in single cell analysis: Multimodal Technologies and their clinical perspectives. *Lab on a Chip*, *22*(13), 2403–2422. Ανακτήθηκε από: https://doi.org/10.1039/d2lc00220e
5) Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., Liu, Q., Russkikh, N., Ryazantsev, G., Ohler, U., Pisco, A. O., Bloom, J., Krishnaswamy, S., & Theis, F. J. (2022). *Multimodal Single Cell Data Integration Challenge: Results and Lessons Learned*. Ανακτήθηκε από: https://doi.org/10.1101/2022.04.11.487796
6) Lin, X., Tian, T., Wei, Z., & Hakonarson, H. (2022). Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature Communications*, *13*(1). Ανακτήθηκε από: https://doi.org/10.1038/s41467-022-35031-9
7) Machine learning course I stanford online. (n.d.-a). Ανακτήθηκε από: https://online.stanford.edu/courses/cs229-machine-learning
8) Mikkola, H. K., & Orkin, S. H. (2006). The journey of developing hematopoietic stem cells. *Development*, *133*(19), 3733–3744. Ανακτήθηκε από: https://doi.org/10.1242/dev.02568
9) MyTutor. (n.d.). *How do cells become specialized?*. MyTutor. Ανακτήθηκε από: https://www.mytutor.co.uk/answers/59517/A-Level/Biology/How-do-cells-become-specialized/
10) Nature Publishing Group. (n.d.-a). *2.1 Information Transfer in Cells Requires Many Proteins and Nucleic Acids*. Nature news. Ανακτήθηκε από: https://www.nature.com/scitable/ebooks/essentials-of-cellbiology14749010/122996756/
11) Nature Publishing Group. (n.d.-b). *2.2 DNA Is Extensively Compacted with Proteins Chromosomes*. Nature news. Ανακτήθηκε από: https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122996796/
12) Nature Publishing Group. (n.d.-c). *2.3 Differential Control of Transcription and Translation Underlies Changes in Cell Function*. Nature news. Ανακτήθηκε από: https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122996928/

13) Nature Publishing Group. (n.d.-d). *2.4 The Functions of Proteins Are Determined by Their Three-Dimensional Structures*. Nature news. Ανακτήθηκε από: https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122996920/
14) Nature Publishing Group. (n.d.-e). *2.5 Proteins Are Responsible for a Diverse Range of Structural and Catalytic Functions in Cells*. Nature news. Ανακτήθηκε από: https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122996980/
15) Nature Publishing Group. (n.d.-f). Nature news. Ανακτήθηκε από: https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122996756/
16) *Open problems - multimodal single-cell integration*. Kaggle. (n.d.). Ανακτήθηκε από: https://www.kaggle.com/competitions/open-problems-multimodal/overview
17) Stem Cell Quick Guide: Stem cell basics - UC davis office of research. (n.d.-b). Ανακτήθηκε από:https://research.ucdavis.edu/wp-content/uploads/Stem-Cell-Research-Quick-Guide.pdf
18) Stuart, T., Srivastava, A., Lareau, C., & Satija, R. (2020). *Multimodal Single-Cell Chromatin Analysis with Signac*. Ανακτήθηκε από: https://doi.org/10.1101/2020.11.09.373613
19) *Supervised machine learning: Regression and classification*. Coursera. (n.d.). Ανακτήθηκε από: https://www.coursera.org/learn/machine-learning
20) Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A. D., Huber, W., Trumpp, A., Essers, M. A., & Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, *19*(4), 271–281. Ανακτήθηκε από: https://doi.org/10.1038/ncb3493
21) *What is single-cell sequencing and why is it important?*. What is single-cell sequencing and why is it important? | Babraham Institute. (n.d.). Ανακτήθηκε από: https://www.babraham.ac.uk/blog/single-cell-sequencing
22) Xu, Y., & McCord, R. P. (2022). Diagonal Integration of multimodal single-cell data: Potential Pitfalls and paths forward. *Nature Communications*, *13*(1). Ανακτήθηκε από: https://doi.org/10.1038/s41467-022-31104-x
23) YouTube. (2021, September 14). *Mia: Multimodal Single-cell data, open benchmarks, and a neurips 2021 competition*. YouTube. Ανακτήθηκε από: https://www.youtube.com/watch?v=ZXDILOyiy7A
24) YouTube. (2022, December 8). *2022 multimodal single-cell integration challenge workshop*. YouTube. Ανακτήθηκε από: https://www.youtube.com/watch?v=WmkbyGMPHOE
25) Zhang, Y., Huang, Y., Hu, L., & Cheng, T. (2022). New insights into human hematopoietic stem and progenitor cells via single-cell omics. *Stem Cell Reviews and Reports*, *18*(4), 1322–1336. Ανακτήθηκε από: https://doi.org/10.1007/s12015-022-10330-2

26) Zhu, C., Preissl, S., & Ren, B. (2020). Single-cell multimodal omics: The power of many. *Nature Methods*, *17*(1), 11–14. Ανακτήθηκε από: https://doi.org/10.1038/s41592-019-0691-5