ΕΛΛΗΝΙΚΟ ΜΕΣΟΓΕΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΠΟΛΥ-ΟΜΙΚΗ ΑΝΑΛΥΣΗ ΓΕΝΟΜΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΔΕΙΓΜΑΤΩΝ ΓΛΟΙΟΒΛΑΣΤΩΜΑΤΟΣ

ΟΙΚΟΝΟΜΟΥ ΝΙΚΟΛΑΟΣ

ΤΠ4845

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

ΤΣΙΚΝΑΚΗΣ ΜΑΝΩΛΗΣ

ΗΡΑΚΛΕΙΟ

ΔΕΚΕΜΒΡΙΟΣ 2022

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους καθηγητές μου, κ. Τσικνάκη Μανώλη ως επιβλέπων καθηγητή για την συνεργασία μας στην διάρκεια υλοποίησης της πτυχιακής μου και τον κ. Κουμάκη Λευτέρη για τη μεγάλη βοήθεια που μου προσέφερε όπου και όποτε την χρειαζόμουν, τόσο στο θεωρητικό αλλά και πρακτικό επίπεδο καθώς επίσης για το ενδιαφέρον που μου καλλιέργησε σχετικά με τον τομέα της βιοπληροφορικής και της βιοϊατρικής γενικότερα. Τέλος, θα ήθελα επίσης να ευχαριστήσω όλους όσους με βοήθησαν και με στήριξαν στην υλοποίηση της πτυχιακής εργασίας και των σπουδών μου γενικότερα.

Περίληψη

Η ομαδοποίηση μεμονωμένων ομικών συνόλων δεδομένων έχει αποδειχθεί ανεκτίμητη για τη βιολογική και ιατρική έρευνα. Το μειούμενο κόστος και η ανάπτυξη μεθόδων αλληλούχισης νέας γενιάς (NGS) επιτρέπουν πλέον τη μέτρηση πολύ-ομικών δεδομένων. Η ομαδοποίηση πολλαπλών ομικών δεδομένων έχει τη δυνατότητα να αποκαλύψει περαιτέρω γνώσεις σε επίπεδο συστήματος, αλλά εγείρει υπολογιστικές και βιολογικές προκλήσεις.

Η παρούσα πτυχιακή εξετάζει αλγορίθμους για ομαδοποίηση πολλαπλών ομικών δεδομένων καθώς και μεθοδολογίες από την περιοχή της μηχανικής μάθησης για την κοινή ομαδοποίηση πολλαπλών τύπων δεδομένων. Το πρώτο μέρος της εν λόγω μελέτης ασχολείται με την περιγραφή του γλοιοβλαστώματος ως καρκινικός τύπος και περιλαμβάνει έννοιες από το πεδίο της βιολογίας οι οποίες κρίνονται σημαντικές. Στη συνέχεια, γίνεται αναφορά στο πεδίο της μηχανικής μάθησης αναλύοντας κατηγορίες αλγορίθμων όπως για παράδειγμα επιτηρούμενη μάθηση, μη επιτηρούμενη μάθηση, ενισχυτική μάθηση και παρουσιάζει εφαρμογές της μηχανικής μάθησης στην σημερινή εποχή.

Πέρα των προαναφερθέντων εννοιών, γίνεται παρουσίαση των αποτελεσμάτων των αλγορίθμων και των τεχνικών που χρησιμοποιήθηκαν όπως για παράδειγμα η χρήση του αλγορίθμου SMOTE και τα τελικά συμπεράσματα που προέκυψαν.

Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι αλγόριθμοι μηχανικής μάθησης SVM και Decision Tree σε ομικά δεδομένα, Gene expression, DNA Methylation, miRNA και κλινικά δεδομένα που αντλήθηκαν από τη γενομική βάση TCGA και αφορούν ασθενείς με γλοιοβλάστωμα. Έγινε η απαραίτητη προ-επεξεργασία των δεδομένων και οι ασθενείς κατηγοριοποιήθηκαν βάση της κατάστασης τους (εν ζωή ή αποθανών). Να σημειωθεί ότι έγινε χρήση του αλγορίθμου υπερδειγματοληψίας SMOTE καθώς τα των δεδομένων ήταν ανισόρροπο. Εφαρμόστηκαν τα μοντέλα σε κάθε ομικό σύνολο ξεχωριστά και στην ενοποιημένη αναπαράσταση που προέκυψε από την χρήση της στρατηγικής για πολύ-ομικα δεδομένα, early integration. Προέκυψε ότι, η εν λόγω στρατηγική έδωσε καλύτερα αποτελέσματα συγκριτικά με την ανάλυση κάθε ομικού επιπέδου ξεχωριστά δίνοντας ως αποτέλεσμα στο δέντρο απόφασης χαρακτηριστικά από όλα τα ομικά επίπεδα. Ειδικότερα, ο αλγόριθμος SVM παρουσίασε ακρίβεια

88.23% και ο Decision Tree 72,54%. Τέλος, η εν λόγω στρατηγική χρησιμοποιήθηκε και σε άλλα καρκινικά δεδομένα, ώστε να φανεί η αποτελεσματικότητα της και σε άλλα δεδομένα πέρα του γλοιοβλαστώματος.

Λέξεις κλειδιά: Βιολογία, Γονίδιο, Μηχανική Μάθηση, DNA, RNA, Μεθυλίωση DNA, Γλοιοβλάστωμα, Καρκίνος, Βιοπληροφορική, Πολυ-ομικά δεδομένα, Πολύ-ομική ανάλυση

Abstract

Clustering of individual omic datasets has proven invaluable for biological and medical research. Decreasing costs and the development of next-generation sequencing (NGS) methods now have the measurement of multi-omic data. Clustering multiple omics data has the potential to reveal further systems-level insights, but raises computational and biological challenges.

This thesis examines algorithms for clustering multiple omic data as well as methodologies from the area of machine learning for joint clustering of multiple data types. The first part of this study deals with the description of glioblastoma as a type of cancer and includes concepts from the field of biology which are considered important. Then, a reference is made to the field of machine learning by analyzing categories of algorithms such as supervised learning, unsupervised learning, reinforcement learning and applications of machine learning in today's era.

In addition to the aforementioned concepts, the results of the algorithms and techniques used are presented, such as the use of the SMOTE algorithm and the final conclusions obtained.

More specifically, SVM and Decision Tree machine learning algorithms were used on omic data, Gene expression, DNA Methylation, miRNA and clinical data extracted from the TCGA genomic database and related to patients with glioblastoma. The necessary pre-processing of the data was done and the patients were categorized based on their status (alive or deceased). Note that, the SMOTE oversampling algorithm was used as the data was imbalanced. The models were applied to each omic set separately and to the unified representation resulting from the use of the strategy for multi-omic data, early integration. It turned out, that this strategy gave better results compared to analyzing each omic level separately, resulting in a decision tree with features from all omic levels. In particular, the SVM algorithm presented an accuracy of 88.23% and the Decision Tree 72.54%. Finally, this strategy was also used in other cancer data to show its effectiveness in other data beyond glioblastoma.

Keywords: Biology, Gene, Machine Learning, DNA, RNA, DNA Methylation, Glioblastoma, Cancer, Bioinformatics, Multi-omics data, Multi-omics analysis

Περιεχόμενα

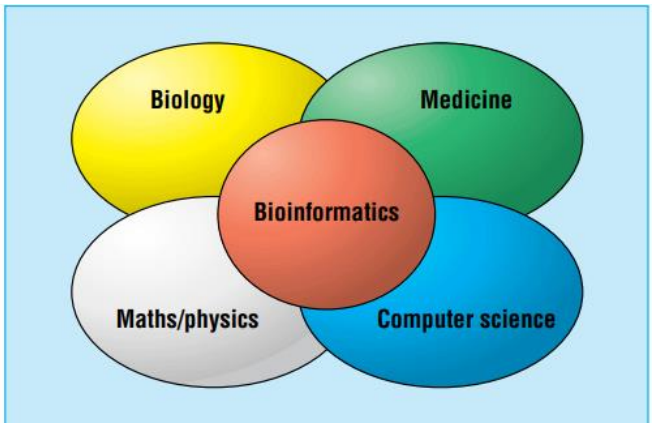
Ευρετήριο εικόνων

Ευρετήριο πινάκων

1. Εισαγωγή

1.1 Τι είναι βιοπληροφορική;

Ο όρος βιοπληροφορική ορίζεται ως η εφαρμογή εργαλείων υπολογισμού και ανάλυσης για την κατανόηση και διαχείριση βιολογικών δεδομένων. Το πεδίο της βιοπληροφορικής περιλαμβάνει την συνεργασία κι άλλων επιστημονικών πεδίων όπως των μαθηματικών, της επιστήμης υπολογιστών, της βιολογίας και της φυσικής.



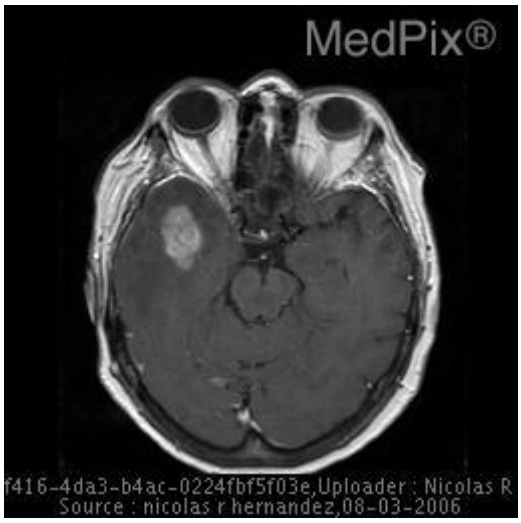
Εικόνα 1: Εικόνα που δείχνει τα πεδία από τα οποία αποτελείται η βιοπληροφορική. Πηγή: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1122955/

Η βιοπληροφορική υπάρχει εδώ και αρκετά χρόνια με τον όρο αυτόν να πρωτοδιατυπώνεται στις αρχές της δεκαετίας 1960 όπου αρχικά είχε εφαρμογή σε υπολογιστικές μεθόδους για την ανάλυση της αλληλουχίας των πρωτεϊνών [1]. Στις αρχές του 21^{ου} αιώνα, ολοκληρώθηκε η χαρτογράφηση του ανθρώπινου γονιδιώματος (Human Genome Project) με τον τομέα της βιοπληροφορικής να διαδραματίζει σημαντικό ρόλο στην επίτευξη του έργου. Σήμερα, η βιοπληροφορική ασχολείται με την έκφραση των γονιδίων, την ανάλυση ακολουθιών γονιδίων, τη λειτουργία και δομή των πρωτεϊνών κ.α.. Αυτό δίνει στους επιστήμονες την δυνατότητα να εξελίξουν αποτελεσματικές θεραπείες γονιδίων (gene therapies) και φάρμακα καταπολεμώντας ασθένειες συμπεριλαμβανομένου και του καρκίνου [2]. Παρακάτω θα ερευνηθεί αναλυτικότερα ο καρκίνους του γλοιοβλαστώματος μέσα από τις εφαρμογές της βιοπληροφορικής.

1.2 Γλοιοβλάστωμα

Πριν προχωρήσουμε στην κατανόηση του γλοιοβλαστώματος πρέπει να γίνει αναφορά στην ευρύτερη κατηγορία που ανήκει. Πιο συγκεκριμένα γίνεται λόγος για τα **γλοιώματα** (Gliomas), που συνιστούν σοβαρούς πρωτοπαθείς όγκους του εγκεφάλου και προέρχονται από τα γλοιο-κύτταρα (glial cells). Χωρίζονται στις εξής κατηγορίες: **αστροκυττώματα**, **ολιγοδεντρογλοιώματα** και **ολιγοαστροκυττώματα**. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, η πιο συχνή μορφή γλοιωμάτων ως **4ο στάδιο** καρκίνου, είναι το γλοιοβλάστωμα [3].

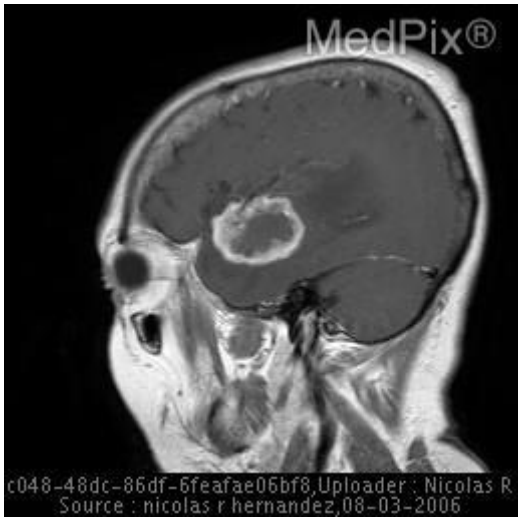
Ειδικότερα το γλοιοβλάστωμα (GBM), αποτελεί τον πιο κοινό πρωτοπαθή και θανατηφόρο καρκίνο του κεντρικού νευρικού συστήματος. Χωρίζεται σε 4 μοριακούς τύπους: τον **κλασσικό**, τον **μεσεγχυματικό**, τον **νευρωνικό** και τέλος τον **προνευρικό**[4].



Εικόνα 2: Όγκος γλοιοβλαστώματος 4εκατοστών σε γυναίκα ασθενή 76 ετών. Πηγή: https://medpix.nlm.nih.gov/case?id=0c6fc001-3971-40b1-857a-6769b015e7f8

Αυτός ο τύπος καρκίνου σε σχέση με άλλους καρκίνους εγκεφάλου, σπάνια παρουσιάζει μετάσταση [5]. Έρευνες έχουν δείξει ότι προσβάλλει το 3:100000 πληθυσμού στις ΗΠΑ κάθε χρόνο με τους άνδρες ηλικίας 64 χρονών να διαγιγνώσκονται με γλοιοβλάστωμα 1.5 φορές περισσότερο από ότι τις γυναίκες. Ανάμεσα στα συνήθη συμπτώματα της εν λόγω νόσου σημειώνονται **πονοκέφαλοι απώλεια μνήμης**, **σύγχυση, νευρικές διαταραχές** ή **επιληπτικές κρίσεις** [6].

Η εμφάνιση του είναι πιθανή σε οποιαδήποτε ηλικία αλλά κυρίως παρατηρείται στις ηλικίες 55-60, με το προσδόκιμο ζωής των ασθενών που διαγιγνώσκονται με γλοιοβλάστωμα να ανέρχεται στους 14-15 μήνες μετά την διάγνωση [7].



Εικόνα 3: Όγκος γλοιοβλαστώματος 4εκατοστών σε γυναίκα ασθενή 76 ετών. Πηγή: https://medpix.nlm.nih.gov/case?id=f4c16b27-4936-497f-9db5-80a97a541c53

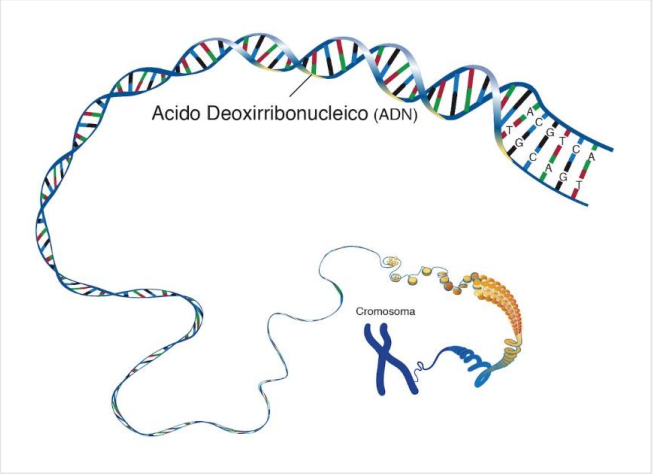
Το γλοιοβλάστωμα χωρίζεται σε δυο κλινικές κατηγορίες: **De novo** και το **δευτερογενές γλοιοβλάστωμα** (secondary glioblastomas).

Αναλυτικότερα, το de novo γλοιοβλάστωμα είναι ο πιο συχνός τρόπος εμφάνισης του καρκίνου χωρίς να υπάρχει προγενέστερη μορφή όγκου. Οι ασθενείς αυτής της κατηγορίας είναι συνήθως ηλικιωμένοι και έχουν ενδείξεις ανεβασμένου EGFR, την ύπαρξη φωσφατάσης (phosphatase) και τενσίνης (tensin) διαγραμμένες στο χρωμόσωμα 10. Από την άλλη πλευρά το δευτερογενές γλοιοβλάστωμα εμφανίζεται μετά από προηγούμενη διάγνωση μικρότερου βαθμού όγκου. Παρατηρείται ότι οι μεταλλάξεις TP53 και RB είναι πιο συχνές στην εμφάνιση δευτερογενούς γλοιοβλαστώματος [5].

1.3 DNA

Το DNA ή αλλιώς δεοξυριβονουκλεϊκό οξύ, είναι ένα σύνθετο κληρονομικό μόριο που υπάρχει στους ανθρώπους αλλά και σχεδόν σε κάθε οργανισμό περιέχοντας όλη την πληροφορία που είναι απαραίτητη για τη δημιουργία και ύπαρξη της ζωής [8]. Βρίσκεται αποθηκευμένο στον πυρήνα κάθε κυττάρου (πυρηνικό DNA) οργανωμένο

σε μορφές που ονομάζονται χρωμοσώματα. Πιο συγκεκριμένα, στα ευκαρυωτικά κύτταρα βρίσκεται εντός του πυρήνα ενώ στα προκαρυωτικά ελεύθερο στο κυτταρόπλασμα καθώς τα κύτταρα αυτά δεν διαθέτουν πυρήνα. Αποτελείται από δύο σκέλη τα οποία δημιουργούν ένα σχήμα γνωστό ως διπλή έλικα. Συνθέτεται από τέσσερις βάσεις, την αδενίνη (A), τη θυμίνη (T), τη γουανίνη (G) και την κυτοσίνη (C) όπου η αδενίνη ενώνεται η με την θυμίνη (A-T) και η γουανίνη ενώνεται με την κυτοσίνη (G-C) [9]. Υπολογίζεται ότι το ανθρώπινο γονιδίωμα αποτελείται περίπου από 3 δισεκατομμύρια βάσεις οι οποίες είναι στην συντριπτική πλειοψηφία τους ίδιες σε όλους τους ανθρώπους [10].



Εικόνα 4: Η δομή του DNA. Πηγή: https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid

1.3.1 Μεθυλίωση DNA

Η μεθυλίωση DNA (DNA methylation) είναι ένας επιγενετικός μηχανισμός, υπεύθυνος για τη μεταφορά μεθυλίου (CH3) στη θέση 5 της κυτοσίνης [11]. Πιο συγκεκριμένα, η μεθυλίωση του DNA επιτρέπει στα κύτταρα να ελέγχουν την έκφραση των γονιδίων. Ο μηχανισμός αυτός είναι μείζονος σημασίας σε αρκετές κυτταρικές λειτουργίες όπως στην ανάπτυξη του εμβρύου, στην αδρανοποίησης του X χρωμοσώματος και στη διατήρησης της σταθερότητας των χρωμοσωμάτων [12]. Το ανθρώπινο γονιδίωμα περιλαμβάνει μεθυλιωμένες (methylated) και μη μεθυλιωμένες (unmethylated) περιοχές.

Υπολογίζεται ότι υπάρχουν περίπου 29 εκατομμύρια CpG περιοχές με το μεγαλύτερο ποσοστό να είναι μεθυλιωμένες. Το 7% των περιοχών αυτών βρίσκονται

σε κάποιες περιοχές που ονομάζονται νησίδες CpG (CpG islands) [13]. Ειδικότερα, τα CpG islands αποτελούν τμήματα μη μεθυλιωμένου DNA με πολύ μεγάλο αριθμό CpG δινουκλεοτιδίων (CpG dinucleotides) [14].

Επιπλέον, η μεθυλίωση του DNA επιτυγχάνεται από μια ομάδα ενζύμων που ονομάζεται μεθυλοτρανσφεράσες DNA (DNA methyltransferases) (DNMT). Οι μεθυλοτρανσφεράσες DNA διακρίνονται σε: DNMTs: DNMT1, DNMT2, DNMT3A, DNMT3B, DNMT3C και DNMT3L [15][13].

Όπως αναφέρθηκε, η μεθυλίωση του DNA είναι μία λειτουργία απαραίτητη για την ανάπτυξη και την φυσιολογική λειτουργία ενός κυττάρου. Παρόλα αυτά παρατηρούνται ανωμαλίες οι οποίες μπορεί να οδηγήσουν σε διάφορες ασθένειες μέσα στις οποίες συγκαταλέγεται και ο καρκίνος. Τα καρκινικά κύτταρα παρουσιάζουν διαφορετική μεθυλίωση σε σχέση με τα φυσιολογικά. Πιο συγκεκριμένα παρατηρούνται περιπτώσεις υπομεθυλίωσης (hypomethylation) και υπερμεθυλίωσης (hypermethylation) [16]. Η υπομεθυλίωση αφορά συνήθως επαναλαμβανόμενες ακολουθίες DNA (repeated DNA sequences) και η υπερμεθυλίωση περιλαμβάνει νησίδες CpG. Πιο συγκεκριμένα, τις περισσότερες φορές συναντάται η περίπτωση της υπερμεθυλίωσης αντί της υπομεθυλίωσης σε ένα καρκίνο. Παρά τους μηχανισμούς που αποτρέπουν την υπερμεθυλίωση, έχουν βρεθεί υπερμεθυλιωμένα αρκετά γονίδια όπως τα $p16^{INK4a}$, $p15^{INK4a}$, BRCA1, MGMT που είναι υπεύθυνα για την ρύθμιση του κυτταρικού κύκλου και τις διορθώσεις του DNA αντίστοιχα [17].

Η περίπτωση της υπομεθυλίωσης, παρατηρείται στην καρκινογένεση αλλά και στην εξέλιξη του όγκου [18]. Επίσης, σημειώνεται σε αρκετούς καρκίνους συμπεριλαμβανομένου και του εγκεφάλου παρουσιάζοντας αυξημένη επιθετικότητα [17].

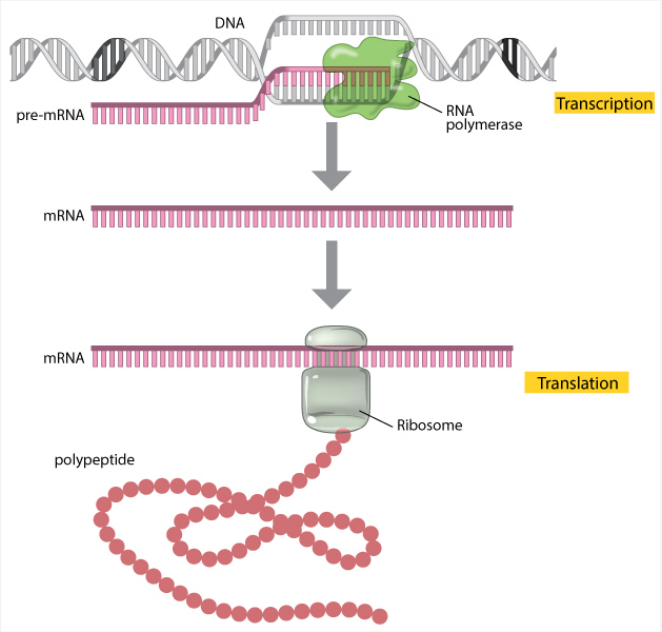
1.4 RNA

Το RNA ή αλλιώς ριβονουκλεϊκό οξύ, είναι ένα μονόκλωνο μόριο που μοιάζει με το DNA και είναι φτιαγμένο από ριβόζη και φωσφορικές ομάδες. Αποτελείται από τέσσερις βάσεις , την αδενίνη (A), την ουρακίλη (U), την κυτοσίνη (C) και τη γουανίνη (G). Κατηγοριοποιείται σε διάφορους τύπους , το αγγελιοφόρο RNA (mRNA), το μεταφορικό RNA (tRNA) και το ριβοσωμικό RNA (rRNA) [19]. Στην παρούσα

πτυχιακή θα μας απασχολήσει το mRNA, καθώς επίσης και το miRNA όροι οι οποίοι θα αναλυθούν στην συνέχεια. Το RNA χρησιμοποιείται στο κύτταρο για την σύνθεση των πρωτεϊνών. Αξίζει να σημειωθεί, ότι σε κάποιους ιούς το RNA «κουβαλάει» τον γενετικό κώδικα του αντικαθιστώντας το DNA [20].

1.4.1 Τι είναι το mRNA

Το αγγελιαφόρο ριβονουκλεϊκό οξύ ή αλλιώς mRNA είναι ένα μονόκλωνο μόριο RNA που αντιστοιχεί σε ένα κλώνο DNA ενός γονιδίου. Το mRNA μπορεί και εξέρχεται του πυρήνα του κυττάρου κατευθυνόμενο προς το κυτταρόπλασμα όπου γίνεται η παραγωγή των πρωτεϊνών. Μέσα στο κυτταρόπλασμα, όπου βρίσκονται και τα ριβοσώματα , ένα από τα οποία κινείται κατά μήκος του mRNA διαβάζοντας έτσι την ακολουθία των βάσεων και χρησιμοποιεί τον γενετικό κώδικα για να μεταφράσει κάθε τριπλέτα βάσεων (three-base triplet) ή κωδικόνιο (condon) στο αντίστοιχο αμινοξύ του [21].



Εικόνα 5: Από το DNA στο RNA και στην πρωτεΐνη. Πηγή:"https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/"

Λάθη κατά την μετάφραση του mRNA μπορεί να προκαλέσουν μείωση των επιπέδων των λειτουργικών πρωτεϊνών με αποτέλεσμα την αύξηση των επιβλαβών μορίων

οδηγώντας στην πρόκληση σπάνιων γενετικών διαταραχών [22]. Σύμφωνα με μελέτες ο ρόλος του mRNA στην θεραπεία ασθενειών όπως ο καρκίνος διαφόρων τύπων είναι πολύ σημαντικός και γεννά ελπίδες για τους καρκινοπαθείς [23][24]. Επιπρόσθετα, η χρήση mRNA εμβολίων για την αντιμετώπιση του καρκίνου αλλά και άλλων ασθενειών, παρουσιάζει εξαιρετικό ενδιαφέρον από την επιστημονική κι ιατρική κοινότητα. Ένα από τα πλεονεκτήματα των συγκεκριμένων εμβολίων είναι η ικανότητα τους μέσω των CD8+T κυττάρων να παρέχουν ανοσία, γεγονός που τα καθιστά εξαιρετικά ικανά στην καταπολέμηση των όγκων. Τέλος, να σημειωθεί ότι οι ασθενείς που λαμβάνουν mRNA εμβόλια, δεν παρουσιάζουν σοβαρές παρενέργειες σε σχέση με εμβόλια DNA [25].

1.4.2 Τι είναι το miRNA

Το MicroRNA (miRNA) είναι μικρά μόρια (non-coding) RNA με μέγεθος περίπου 22 νουκλεοτίδια που εμπλέκονται στην έκφραση των γονιδίων, τη μεταγραφική καταστολή και βρίσκεται σε όλα τα ευκαρυωτικά κύτταρα [26][27]. Υπολογίζεται ότι το ανθρώπινο γονιδίωμα διαθέτει περισσότερα από 1000 miRNA γονίδια που διαδραματίζουν σημαντικό ρόλο στην διαχείριση/λειτουργία του κυττάρου [28] και βρίσκονται οργανωμένα σε ομάδες. Παράλληλα, διαθέτουν τις ίδιες μεταγραφικές ρυθμιστικές μονάδες και εκφράζονται ανεξάρτητα. Παρόλα αυτά, το miRNA εντοπίζεται στα αίτια εκδήλωσης αρκετών ασθενειών, μεταξύ των οποίων, αυτοάνοσα νοσήματα, έμφραγμα του μυοκαρδίου αλλά και καρκίνο [26].

Σύμφωνα με τα προηγούμενα στοιχεία, η έκφραση του miRNA μπορεί να προκαλέσει καρκίνο. Πολλές έρευνες έχουν καταλήξει στο γεγονός ότι συγκεκριμένες ομάδες/οικογένειες miRNA είναι αρκετές φορές απορρυθμισμένες σε ορισμένους τύπους κακοηθειών, γεγονός εξαιρετικά χρήσιμο στη διάγνωση, την εξέλιξη αλλά και στη θεραπεία του ασθενή. Αξίζει να σημειωθεί ότι η απορρύθμιση αυτή δεν είναι τυχαίο φαινόμενο [29].

Η ανοδική (up-regulation) ή πτωτική ρύθμιση (down-regulation) των miRNA εμφανίζεται σε πολλούς καρκίνους. Παραδείγματος χάριν, υπερεκφρασμένα miRNA

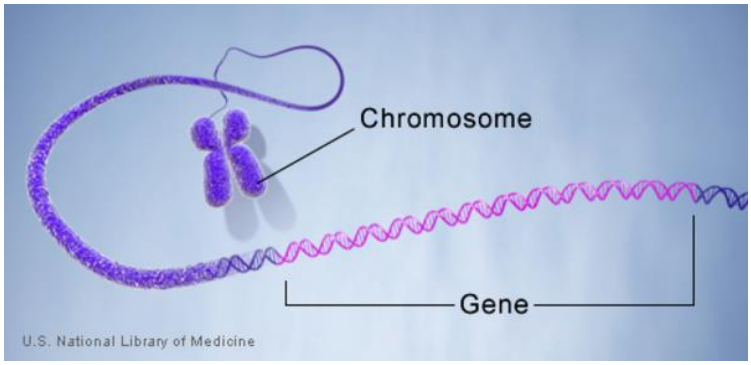
μπορεί να λειτουργήσουν ως ογκογονίδια (oncogenes) ή/και ως ρυθμιστής σε διεργασίες του κυττάρου [26].

Τα miRNA έχει παρατηρηθεί ότι διαδραματίζουν σημαντικό ρόλο στη παθογένεια του γλοιοβλαστώματος [30]. Πιο συγκεκριμένα, σε μια ανασκόπηση των miRNA στο γλοιοβλάστωμα βρέθηκαν 253 υπερεκφρασμένα, 95 υποεκφρασμένα και 17 είναι υπό αμφισβήτηση ρυθμισμένα. Τα προς τα πάνω ρυθμισμένα (upregulated) miRNA μπορούν να λειτουργήσουν ως ογκογονίδια, αδρανοποιώντας τα ογκοκατασταλτικά γονίδια ενώ τα προς τα κάτω ρυθμισμένα (downregulated) μπορεί να δράσουν ως καταστολείς του όγκου. Τα υπορυθμισμένα miRNA αναγνωρίστηκαν επίσης ως πιθανά για θεραπευτικούς σκοπούς. Ειδικότερα, κάποια miRNA που εκφράζονται στην παθογένεια του γλοιοβλαστώματος είναι τα miRNA-21, miRNA-221/22, miRNA-181a & miRNA-181b, miR-7. Για παράδειγμα, το miRNA-21 έχει βρεθεί υπερεκφρασμένο σε γλοιοβλαστώματα αλλά και μικρότερου βαθμού αστροκυτώματα. Το εν λόγω miRNA έχει σημαντικό ρόλο σε κακοήθεις διεργασίες, στοχεύοντας γονίδια όπως το PTEN, το Tap63, το TIMP3 κ.α.[31][32].

1.5 Τι είναι το γονιδίωμα και τα γονίδια;

Το γονιδίωμα είναι η συλλογή όλης της γενετικής πληροφορίας (DNA) ενός οργανισμού, πληροφορία απαραίτητη για την ζωή και ανάπτυξη του. Βρίσκεται σε κάθε κύτταρο του οργανισμού, φυλασσόμενο εντός του πυρήνα και χωρίζεται σε μικρότερα τμήματα που ονομάζονται γονίδια [33].

Τα γονίδια είναι τμήματα DNA που περιέχουν την πληροφορία για τη σύνθεση μιας πρωτεΐνης ή ενός συνόλου πρωτεϊνών. Υπολογίζεται ότι στο ανθρώπινο γονιδίωμα υπάρχουν 20.000 – 25.000 γονίδια όπου κάθε ένα από αυτά είναι υπεύθυνο για την κωδικοποίηση τριών περίπου πρωτεϊνών. Τα γονίδια εντοπίζονται εντός του πυρήνα του κυττάρου σε 23 ζεύγη χρωμοσωμάτων και με την βοήθεια αγγελιοφόρων μορίων και ενζύμων, διευθύνουν την παραγωγή των πρωτεϊνών [34].



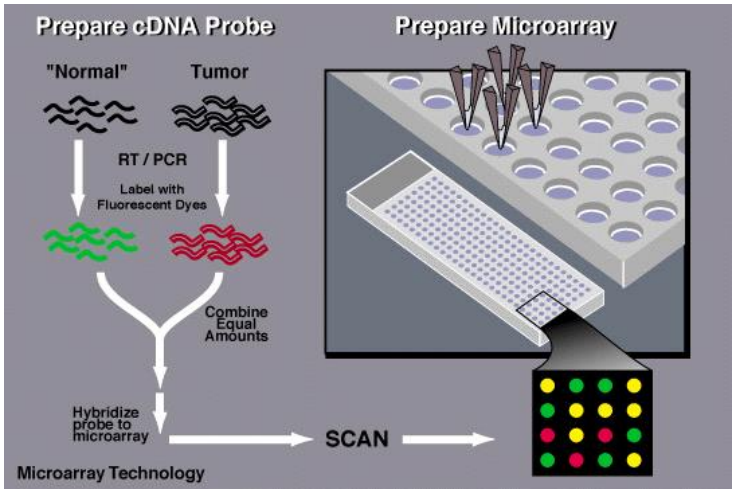
Εικόνα 6: Κάθε χρωμόσωμα περιέχει πολλά γονίδια. Πηγή: https://medlineplus.gov/genetics/understanding/basics/gene/#:~:text=A%20gene%20is%20the%20basic,more%20than%202%20million%20bases.

1.6 Τι είναι η έκφραση των γονιδίων;

Όπως αναφέρθηκε παραπάνω τα γονίδια κωδικοποιούν πρωτεΐνες οι οποίες με την σειρά τους υπαγορεύουν την λειτουργία του κυττάρου. Επομένως, οι λειτουργίες ενός κυττάρου καθορίζονται από την έκφραση των γονιδίων του. Η διαδικασία της έκφρασης των γονιδίων χωρίζεται σε δυο βήματα/κατηγορίες, την **μεταγραφή** (transcription) και τη **μετάφραση** (translation). Κατά τη διάρκεια της μεταγραφής από το DNA δημιουργείται το RNA και κατά τη μετάφραση το RNA μεταφράζεται σε πρωτεΐνη. Ο έλεγχος αυτών των διεργασιών καθορίζει σημαντικά ποιες πρωτεΐνες υπάρχουν στο κύτταρο και σε ποιες ποσότητες [35]. Οι πιο διαδεδομένοι τρόποι ανάλυσης της γονιδιακής έκφρασης είναι οι **μικροσυστοιχίες** DNA (DNA microarrays) και η **αλληλούχιση** RNA (RNA sequencing).

1.6.1 Μικροσυστοιχίες DNA

Η ανάλυση με μικροσυστοιχίες είναι ένα εργαστηριακό εργαλείο που χρησιμοποιείται από τους επιστήμονες για την ανίχνευση της έκφρασης χιλιάδων γονιδίων την ίδια χρονική στιγμή. Είναι ένα πλακίδιο (ή αλλιώς DNA chip) το οποίο διαθέτει χιλιάδες μικροσκοπικά σημεία που ονομάζονται ανιχνευτές (probes) σε συγκεκριμένες θέσεις και ο κάθε ανιχνευτής αντιπροσωπεύει ένα γονίδιο. Στις θέσεις αυτές τοποθετείται το δείγμα προς ανάλυση έπειτα από μια επεξεργασία που έχει υποστεί [36].



Εικόνα 7: Η γονιδιακή έκφραση με την χρήση μικροσυστοιχιών DNA. Πηγή:https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology

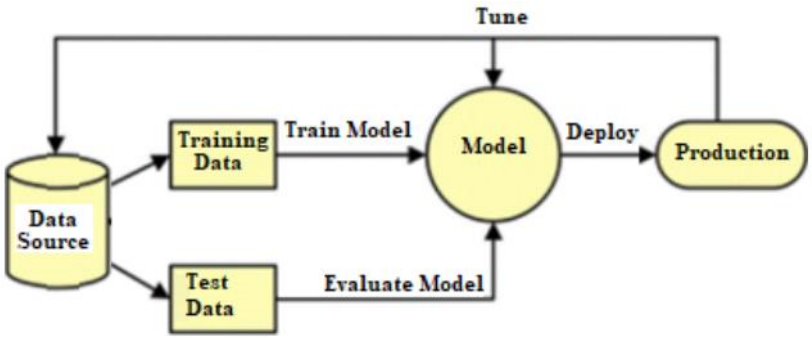
1.6.2 Αλληλούχιση RNA

Η μέθοδος αλληλούχισης RNA συγκριτικά με άλλες μεθόδους όπως αυτή της χρήσης μικροσυστοιχιών, παρέχει πολύ μεγαλύτερη κάλυψη και υψηλότερη ανάλυση της δυναμικής φύσης του μεταγραφώματος. Προκειμένου να παρέχει πληροφορίες που αφορά την μεταγραφή ενός κυττάρου, χρησιμοποιεί τις δυνατότητες μεθόδων αλληλουχίας υψηλής απόδοσης. Τέλος, αξίζει να σημειωθεί ότι εκτός των στοιχείων που προκύπτουν από την έκφραση των γονιδίων, η παραγωγή των δεδομένων απ' τη μέθοδο αυτή, δίνει την δυνατότητα στους ερευνητές να ανακαλύψουν νέες μεταγραφές, να προβούν σε ταυτοποίηση συνδεδεμένων γονιδίων αλλά και στην ανίχνευση ειδικής αλληλόμορφου έκφρασης [37].

2. Τι είναι η μηχανική μάθηση;

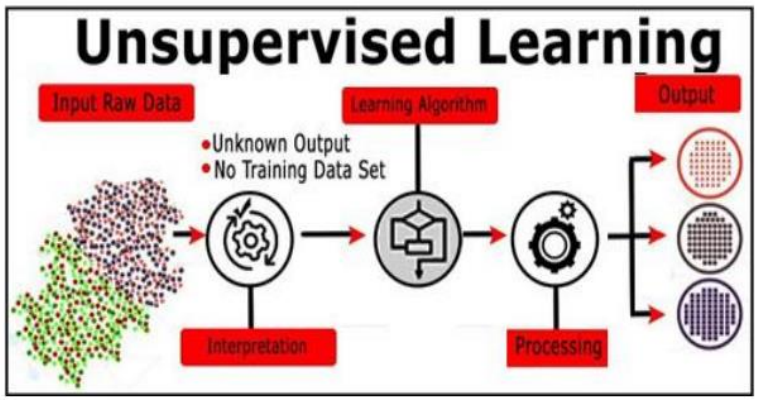
Η μηχανική μάθηση είναι ένας κλάδος της επιστήμης υπολογιστών που επιδιώκει να δώσει την δυνατότητα στους υπολογιστές στο να «μάθουν» χωρίς να είναι απευθείας προγραμματισμένοι. Κάθε φορά που ο υπολογιστής «μαθαίνει», αποκτά εμπειρία καταφέρνοντας έτσι να αυτοβελτιώνεται [38] με τη διαδικασία αυτή να επιτυγχάνεται με τη χρήση αλγορίθμων και στατιστικών μοντέλων. Επειδή κάθε πρόβλημα που καλείται να λύσει η μηχανική μάθηση είναι διαφορετικό, κάθε φορά χρησιμοποιείται και ο κατάλληλος αλγόριθμος που χαρακτηρίζεται ως βέλτιστος για την επίλυση του συγκεκριμένου προβλήματος. Τρεις μεγάλες κατηγορίες αλγορίθμων μηχανικής μάθησης είναι η **επιτηρούμενη μάθηση** (Supervised Learning), η **μη επιτηρούμενη μάθηση** (Unsupervised Learning) και η **ενισχυτική μάθηση** (Reinforcement Learning) [39]. Πιο αναλυτικά:

Επιτηρούμενη μάθηση: Ο αλγόριθμος δέχεται μία σειρά από παραδείγματα τα οποία έχουν ετικέτες (labels) χρησιμοποιώντας τα ως δεδομένα εκπαίδευσης (training data). Στη συνέχεια, πραγματοποιεί προβλέψεις για όλα τα άγνωστα δεδομένα που λαμβάνει (testing data).



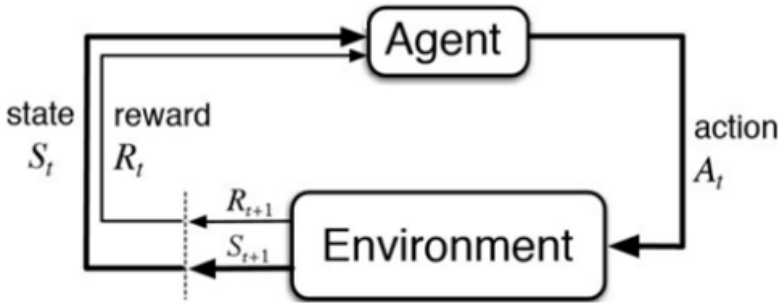
Εικόνα 8: Επιτηρούμενη μάθηση. Πηγή: Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386. (Ανακτήθηκε 15/4/2022)

Μη επιτηρούμενη μάθηση: Ο αλγόριθμος δέχεται μια σειρά από δεδομένα τα οποία σε σχέση με τον επιτηρούμενης μάθησης είναι χωρίς ετικέτες (unlabeled) χρησιμοποιώντας τα ως δεδομένα εκπαίδευσης (training data). Στην συνέχεια, πραγματοποιεί προβλέψεις για όλα τα άγνωστα δεδομένα που λαμβάνει (testing data).



Εικόνα 9:Μη επιτηρούμενη μάθηση. Πηγή: Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386. (Ανακτήθηκε 15/4/2022)

Ενισχυτική μάθηση: Ο αλγόριθμος ενισχυτικής μάθησης προκειμένου να συλλέξει δεδομένα, αλληλεπιδρά με το περιβάλλον και κάποιες φορές μπορεί να το επηρεάσει λαμβάνοντας την ίδια χρονική στιγμή μια επιβράβευση γι' την πράξη του. Στόχος του αλγορίθμου είναι να μεγιστοποιήσει τις ανταμοιβές που λαμβάνει από τις αλληλεπιδράσεις του με το περιβάλλον. Επιπροσθέτως να σημειωθεί ότι, αν ο αλγόριθμος κατά τη διάρκεια εκτέλεσης του πάψει να λαμβάνει πια κάποια επιβράβευση από το περιβάλλον, τότε βρίσκεται αντιμέτωπος με την απόφαση της εξερεύνησης (exploration) ή εκμετάλλευσης (exploitation). Αυτό σημαίνει ότι πρέπει είτε να συνεχίσει να εξερευνά άγνωστα προς αυτόν δεδομένα για να αποκτήσει περισσότερες πληροφορίες ή να εκμεταλλευτεί τα ήδη υπάρχοντα δεδομένα που έχει συλλέξει [40].



Εικόνα 10: Ενισχυτική μάθηση. Πηγή: Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

2.1 Εφαρμογές της μηχανικής μάθησης σήμερα

Η μηχανική μάθηση στην σημερινή εποχή έχει μεγάλη επίδραση στην καθημερινότητα μας . Χρησιμοποιείται για την πρόβλεψη του καιρού, από εταιρίες γι' να τις βοηθήσει να αυξήσουν την παραγωγικότητα τους, από την ιατρική κοινότητα για να εντοπιστούν ασθένειες και σε πολλούς άλλους τομείς [41]. Για παράδειγμα, τα τελευταία χρόνια η μηχανική μάθηση έχει προσφέρει σημαντική βοήθεια στην πρόβλεψη και πρόγνωση του καρκίνου βοηθώντας έτσι τους ερευνητές να αναπτύξουν νέες στοχευμένες θεραπείες για την καταπολέμηση του. Αυτό γίνεται μέσω της ανάλυσης πολύπλοκων γενομικών και κλινικών δεδομένων τα οποία έχουν συλλεχθεί από ασθενείς [42]. Στην παρούσα πτυχιακή, χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης θα μπορέσουμε να αναλύσουμε γενομικά δεδομένα για την κατηγοριοποίηση του γλοιοβλαστώματος.

3. Υπάρχουσες μεθοδολογίες για την ομαδοποίηση πολλαπλών ομικών δεδομένων

Συνήθως, για την ανάλυση πολλαπλών ομικών δεδομένων, συγκεντρώνεται και συνδυάζεται όλη η διαθέσιμη πληροφορία από κάθε ομικό επίπεδο (DNA methylation, miRNA, gene expression κ.α) ώστε να γίνει χρήση όλου του εύρους της. Αυτό γίνεται με αλγορίθμους μηχανικής μάθησης παράγοντας έτσι βιοδείκτες διάγνωσης και ομαδοποίησης. Παρακάτω γίνεται αναφορά σε ορισμένες στρατηγικές που αφορούν την ενσωμάτωση δεδομένων, συγκεκριμένα στις **early integration**, **mixed integration** και **intermediate integration**.

3.1 Early integration

Η στρατηγική αυτή βασίζεται στην **συνένωση κάθε συνόλου δεδομένων** (dataset) σε έναν **μεγάλο ενιαίο πίνακα**. Αυτό έχει ως αποτέλεσμα την αύξηση των μεταβλητών (columns) με σταθερό όμως αριθμό παρατηρήσεων (rows). Λόγω του ότι κάθε σύνολο δεδομένων είναι διαφορετικού μεγέθους, μπορεί να προκληθεί ανισόρροπη μάθηση (imbalanced learning) με το μοντέλο να εκπαιδεύεται πιο πολλή ώρα σε δεδομένα με μεγαλύτερες τιμές αδιαφορώντας για τα υπόλοιπα. Επίσης, η στρατηγική αυτή παραβλέπει τη συγκεκριμένη κατανομή δεδομένων κάθε ομικού επιπέδου «μπερδεύοντας» έτσι τους αλγορίθμους μηχανικής μάθησης, ωθώντας τους να βρίσκουν άσχετα μοτίβα συσχετίσεων που στην πραγματικότητα αφορούν χαρακτηριστικά του ίδιου ομικού επίπεδου.

Παρόλα αυτά, η στρατηγική του early integration, χρησιμοποιείται συχνά καθώς είναι απλή, εύκολη στην υλοποίηση αλλά κυρίως το «δυνατό της χαρακτηριστικό» είναι ότι μπορεί να συνδυάσει δεδομένα από κάθε ομικό επίπεδο επιτρέποντας στους αλγορίθμους μηχανικής μάθησης να ανιχνεύσουν συσχετίσεις μεταξύ των διαφορετικών επιπέδων. Αξίζει να σημειωθεί ότι τα μειονεκτήματα της εν λόγω στρατηγικής, δεν είναι γνωστό σε τι βαθμό επηρεάζουν την ανάλυση και υπάρχει πιθανότητα η απόδοση κάποιων μοντέλων μηχανικής μάθησης να μην σημειώνουν πτώση.

Τέλος, λόγω της πολυπλοκότητα που προκύπτει ως αποτέλεσμα της συνένωσης των συνόλων δεδομένων σε έναν μεγάλο ενιαίο πίνακα, απαιτείται συχνά να γίνει μείωση του αριθμού των μεταβλητών με τη χρήση διαφόρων μεθόδων όπως για παράδειγμα, επιλογή χαρακτηριστικών (Feature Selection).

3.2 Mixed integration

Στη στρατηγική του mixed integration δεν συναντάμε τις μειονεκτήματα της προηγούμενης στρατηγικής. Αντιθέτως, κάθε σύνολο δεδομένων, μεμονωμένα, μετατρέπεται σε μια πιο απλή αναπαράσταση. Η νέα αυτή αναπαράσταση μπορεί να είναι μικρότερη σε διαστάσεις, με λιγότερο θόρυβο στα δεδομένα γεγονός που κάνει την ανάλυση ευκολότερη. Ως γνωστόν, οι ετερογένειες μεταξύ των συνόλων δεδομένων των ομικών επιπέδων, παραδείγματος χάρη διαφορές στα μεγέθη ή διαφορετικοί τύποι δεδομένων , συχνά προκαλούν προβλήματα, που με τη νέα όμως αναπαράσταση αυτά καταργούνται. Έτσι, η συνδυασμένη αναπαράσταση μπορεί να αναλυθεί με κλασσικά μοντέλα μηχανικής μάθησης.

3.3 Intermediate integration

Ως intermediate integration ορίζεται η στρατηγική που κάνει χρήση οποιοδήποτε μεθόδων οι οποίες έχουν ως στόχο την συνένωση των ομικών συνόλων δεδομένων χωρίς να έχει προηγηθεί κάποιος μετασχηματισμός αλλά και χωρίς να είναι μια απλή συνένωση. Αυτό που συνήθως δημιουργείται είναι μια κοινή αναπαράσταση για όλα τα ομικά δεδομένα και σε κάποια δεδομένα ξεχωριστά όπου μπορεί να γίνει ανάλυση. Με τις μεθοδολογίες αυτές μειώνεται το μέγεθος αλλά και η πολυπλοκότητα των ομικών συνόλων δεδομένων. Τις περισσότερες φορές έχει προηγηθεί η μέθοδος της επιλογής χαρακτηριστικών και προ επεξεργασία, καθώς η ετερογένεια μεταξύ του κάθε συνόλου δεδομένων μπορεί να τα αποτρέψει από το να λειτουργούν ορθά. Εν κατακλείδι, προς το παρόν έχουν αναπτυχθεί λίγες μέθοδοι οι οποίες βρίσκουν κοινά μοτίβα μεταξύ ορισμένων ομικών δεδομένων όχι όμως όλων. Ενδεικτικά, τέτοιες μέθοδοι είναι οι SLIDE, NMF (Non-negative Matrix Factorization) και τέλος η CCA (Canonical Correlation Analysis) η οποία χρησιμοποιείται μόνο για δύο σύνολα δεδομένων [43].

4. Ανάλυση του προβλήματος

Η εξέλιξη της τεχνολογίας έχει φέρει την επανάσταση στον χώρο της βιολογίας και της ιατρικής έρευνας [44]. Πλέον παράγονται όλο και περισσότερα βιολογικά δεδομένα που φαίνονται πολύ υποσχόμενα για την θεραπεία ασθενειών και την εξατομικευμένη ιατρική. Κλινικά και ομικά δεδομένα μπορούν να ανακτηθούν από γενομικές βάσεις όπως για παράδειγμα το TCGA, το ICGC, το CCLE κ.α. [45]. Κάθε τύπος ομικών δεδομένων δίνει πληροφορίες για ένα συγκεκριμένο βιολογικό «επίπεδο» όπως για παράδειγμα genomics, epigenomics, transcriptomics, proteomics, metabolomic δίνοντας μια ολοκληρωμένη εικόνα ενός βιολογικού συστήματος ή ατόμου.

Στο παρελθόν η ανάλυση των παραπάνω επιπέδων μεμονωμένα, χρησιμοποιούταν για την ανακάλυψη της πρόκλησης ασθενειών βοηθώντας να βρεθεί μια θεραπεία. Σήμερα όμως οι ερευνητές συνειδητοποίησαν ότι η ανάλυση αυτή είναι πολύ απλοϊκή καθώς οι περισσότερες ασθένειες επηρεάζουν πολύπλοκα μοριακά μονοπάτια (molecular pathways) όπου διαφορετικά βιολογικά επίπεδα αλληλοεπιδρούν μεταξύ τους. Εδώ έρχεται να δώσει λύση η πολύ-ομική ανάλυση δεδομένων.

Η πολυ-ομική ανάλυση είναι ο συνδυασμός των παραπάνω ομικών δεδομένων στη φάση της ανάλυσης, δημιουργώντας ένα μεγαλύτερο σετ δεδομένων (πολύ-ομικά δεδομένα) παράγοντας μια πιο ολοκληρωμένη εικόνα για έναν φαινότυπο [46]. Έτσι, οι ερευνητές έχουν μια καλύτερη κατανόηση της ροής των πληροφοριών από την αρχική αιτία εκδήλωσης της ασθένειας έως τις λειτουργικές συνέπειες [44].

5. Προτεινόμενη λύση

Στη περίπτωση μας, διαθέτουμε δεδομένα από ασθενείς με γλοιοβλάστωμα όπου με την χρήση αλγορίθμων και μεθόδους μηχανικής μάθησης θα προσπαθήσουμε να κατηγοριοποιήσουμε τα δεδομένα και τέλος θα συγκρίνουμε την απόδοση των μοντέλων αυτών ώστε να αξιολογήσουμε ποιος αλγόριθμος ήταν πιο αποτελεσματικός στην κατηγοριοποίηση των δεδομένων που χρησιμοποιήσαμε. Πιο συγκεκριμένα επιλέγονται οι αλγόριθμοι **SVM** (Support Vector Machine) & **Decision Tree**. Επίσης,

κατά την διάρκεια της ανάλυσης, θα δούμε πως θα αντιμετωπίσουμε προβλήματα που τυχόν προκύψουν.

6. Υλοποίηση

6.1 Επιλογή δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν αντλήθηκαν από το TCGA (The Cancer Genome Atlas Program), το οποίο είναι η μεγαλύτερη συλλογή γενομικών δεδομένων. Διαθέτει δεδομένα για 33 διαφορετικούς τύπους καρκίνου από περισσότερους από 11.000 ασθενείς. Περιλαμβάνει DNA sequencing, RNA sequencing, DNA methylation, Gene expressions κ.α. Ακόμη, παρέχει κλινικά δεδομένα για τον κάθε ασθενή όπως για παράδειγμα την ηλικία την εθνικότητα του, φάρμακα που έχει λάβει ή θεραπεία. Τα περισσότερα δεδομένα μεγέθους πάνω από 2.5 petabytes έχουν ελεύθερη πρόσβαση και έτσι ώστε ο καθένας απ' την ερευνητική κοινότητα μπορεί να τα χρησιμοποιήσει [47]. Κατεβάσαμε δεδομένα που προέρχονται από ασθενείς με γλοιοβλάστωμα και περιλαμβάνουν σύνολα δεδομένων που αφορούν DNA methylation, miRNA, Gene expression καθώς επίσης δεδομένα επιβίωσης (survival) και κλινικά (clinical) δεδομένα για τον κάθε ασθενή. Ενδεικτικά εδώ βλέπουμε τα δεδομένα από gene expression και clinical πριν την οποιαδήποτε επεξεργασία, στην αρχική τους δηλαδή μορφή.

	TCGA.02.0001.01	TCGA.02.0003.01	TCGA.02.0004.01	TCGA.02.0007.01	TCGA.02.0009.01	TCGA.02.0010.01	TCGA.02.0011.01	TCGA.02.0014.01	TCGA.0
AACS	6.500551	6.539245	7.377848	7.186891	7.675038	7.996010	8.355122	6.840142	
FSTL1	8.729663	9.794400	12.059550	4.945053	10.840095	8.931571	4.240622	7.738483	
ELMO2	5.511362	6.213981	7.051738	5.230444	6.620676	7.552416	6.707334	7.262258	
CREB3L1	4.882953	4.836276	6.112444	5.818606	5.333213	6.087341	4.865492	4.524546	
RPS11	10.984784	10.811245	10.436374	10.477304	10.637267	11.001533	10.685883	10.661348	
...	...	...	...	...	...	...	...	...	
RPS27	12.911917	13.419446	13.229641	13.575052	13.286953	13.480339	13.474810	13.255097	
SNRPD2	11.872288	11.109714	11.320978	11.665592	11.480397	10.951578	11.542365	11.690116	
SLC39A6	6.867686	8.147826	8.519248	9.189867	8.556436	9.209863	8.437613	10.204783	
CTSC	10.354085	11.367438	10.379502	11.478158	10.404706	8.417933	10.242709	9.188361	
AQP7	4.709882	4.196368	3.733875	4.323108	4.639090	4.167286	4.179506	4.058945	

12042 rows × 538 columns

Εικόνα 11: Σύνολο δεδομένων που αφορά Gene Expression

	sampleID	CDE_DxAge	CDE_alk_chemoradiation_standard	CDE_chemo_adjuvant_alk	CDE_chemo_adjuvant_tmz	CDE_chemo_alk	CDE_chemo_alk_days	CD
0	TCGA-02-0001-01	44.30	False	False	False	False	0.0	
1	TCGA-02-0002-01	NaN	NaN	NaN	NaN	NaN	NaN	
2	TCGA-02-0003-01	50.21	False	False	False	False	0.0	
3	TCGA-02-0004-01	59.18	True	True	True	True	110.0	
4	TCGA-02-0006-01	56.17	False	True	True	True	61.0	
...	...	...	...	...	...	...	...	
624	TCGA-87-5896-01	50.35	False	False	False	False	0.0	
625	TCGA-OX-A56R-01	NaN	NaN	NaN	NaN	NaN	NaN	
626	TCGA-RR-A6KA-01	NaN	NaN	NaN	NaN	NaN	NaN	
627	TCGA-RR-A6KB-01	NaN	NaN	NaN	NaN	NaN	NaN	
628	TCGA-RR-A6KC-01	NaN	NaN	NaN	NaN	NaN	NaN	

629 rows × 138 columns

Εικόνα 12: Σύνολο δεδομένων που αφορά τα κλινικά δεδομένα

Στην προκειμένη περίπτωση, το αρχείο Gene expression περιέχει 12042 γραμμές οι οποίες αναπαριστούν τα γονίδια και 538 στήλες οι οποίες περιλαμβάνουν κάθε ασθενή. Αντίστοιχα, για το αρχείο των κλινικών δεδομένων, έχουμε 629 γραμμές όπου βρίσκονται οι ασθενείς και 138 στήλες όπου βρίσκονται οι τιμές για κάθε κλινικό δεδομένο, όπως για παράδειγμα, η ηλικία του ασθενούς, συγκεκριμένες θεραπείες που έχει λάβει ή όχι μέρος κ.α.

6.2 Επεξεργασία δεδομένων

Για την υλοποίηση της πτυχιακής εργασίας όσο αφορά το υπολογιστικό μέρος, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python καθώς επίσης και ένα σύνολο από βιβλιοθήκες της, οι οποίες είναι ιδιαίτερα χρήσιμες για την επεξεργασία των δεδομένων. Μερικές από τις κύριες βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

- **Matplotlib**: Είναι μια βιβλιοθήκη της Python που χρησιμοποιείται για τη δημιουργία δισδιάστατων (2D) γραφικών πινάκων με εύκολο τρόπο και παρουσιάζει αρκετές ομοιότητες με το εργαλείο MATLAB. Προκειμένου να μπορεί να επεξεργαστεί μεγάλου μεγέθους πίνακες, κάνει χρήση της βιβλιοθήκης NumPy [48].

- **NumPy**: Η NumPy αποτελεί τη βασική βιβλιοθήκη της Python για επιστημονικούς υπολογισμούς. Χρησιμοποιείται για στατιστικές πράξεις, γρήγορη επεξεργασία σε πίνακες, όπως μαθηματικές και λογικές πράξεις, στον μετασχηματισμό Fourier κ.α. Επίσης, πολλές άλλες βιβλιοθήκες της Python κάνουν χρήση της NumPy (πχ Matplotlib) [49].

- **Scikit-learn**: Η Scikit-learn είναι μια βιβλιοθήκη μηχανικής μάθησης της Python. Περιλαμβάνει αλγορίθμους μηχανικής μάθησης για προβλήματα επιτηρούμενης (supervised) και μη επιτηρούμενης (unsupervised) μάθησης. Η χρήση της βιβλιοθήκης μπορεί να γίνει και από άτομα που δεν είναι τόσο εξοικειωμένα, χρησιμοποιώντας την γλώσσα προγραμματισμού Python [50].

- **Pandas**: Είναι και αυτή μια βιβλιοθήκη της Python για ανάλυση δεδομένων παρέχοντας γρήγορη και ευέλικτη επεξεργασία δεδομένων. Είναι σχεδιασμένη να δουλεύει τόσο με σχεσιακά δεδομένα (πχ SQL) όσο και με δεδομένα που διαθέτουν ετικέτες (πχ datasets). Μια λειτουργία όπου θα χρησιμοποιηθεί και στην παρούσα πτυχιακή, είναι η μετατροπή δομών δεδομένων με διαφορετική αρίθμηση (differently-indexed) σε Dataframe αντικείμενα [51].

Αρχικά, βρήκαμε ενδιαφέρον να εστιάσουμε στους ασθενείς (από το αρχείο κλινικών δεδομένων) οι οποίοι έζησαν **πάνω από 100 ημέρες** το οποίο είναι η στήλη **days_to_last_followup** και να τους κατηγοριοποιήσουμε βάση του χαρακτηριστικού **CDE_vital_status**, το οποίο φανερώνει αν ο ασθενής είναι ή όχι στη ζωή. Το χαρακτηριστικό αυτό παρουσιάζει δύο τιμές (κλάσεις), **DECEASED** (αποθανών) και **LIVING** (εν ζωή). Στη συνέχεια, πηγαίνουμε στο σύνολο δεδομένων του clinical αρχείου και παρατηρούμε ότι αυτό περιλαμβάνει πολλές τιμές NaN. Από όλα τα δεδομένα ξεχωρίζουμε αυτά του ενδιαφέροντος μας (days_to_last_followup, CDE_vital_status) και κρατάμε μόνο τους ασθενείς που δεν έχουν στα παραπάνω χαρακτηριστικά NaN τιμές. Επίσης, γίνεται επεξεργασία στον αριθμό αναγνωριστικού κάθε ασθενή (sampleID) ώστε αυτό να είναι στη συνέχεια συμβατό με τα υπόλοιπα σύνολα δεδομένων που περιέχουν το sampleID σε λίγο διαφορετική σύνταξη. Τέλος

προκύπτει μια αναπαράσταση που περιλαμβάνει 447 γραμμές και 3 στήλες όπως φαίνεται στο παρακάτω στιγμιότυπο.

	sampleID	CDE_vital_status	days_to_last_followup
0	TCGA.02.0001.01	DECEASED	279.0
1	TCGA.02.0003.01	DECEASED	144.0
2	TCGA.02.0004.01	DECEASED	345.0
3	TCGA.02.0006.01	DECEASED	558.0
4	TCGA.02.0007.01	DECEASED	705.0
...	...	...	...
442	TCGA.74.6573.01	DECEASED	105.0
443	TCGA.74.6573.11	DECEASED	105.0
444	TCGA.76.4926.01	DECEASED	138.0
445	TCGA.76.4927.01	DECEASED	535.0
446	TCGA.87.5896.01	LIVING	800.0

447 rows × 3 columns

Εικόνα 13: Νέα μορφή clinical συνόλου δεδομένων.

Αφού έχει ολοκληρωθεί με επιτυχία η επεξεργασία του παραπάνω συνόλου δεδομένων, έγινε επεξεργασία των ομικών συνόλων δεδομένων. Στόχος της επεξεργασίας αυτής είναι η εύρεση των κοινών ασθενών κάθε ομικού επιπέδου και κλινικού αρχείου κλινικών δεδομένων. Αρχικά, έγινε μετατόπιση των στηλών και των γραμμών όλων των αρχείων (εκτός του clinical αρχείου), δηλαδή οι γραμμές έγιναν στήλες και οι στήλες γραμμές. Αυτό είναι απαραίτητο βήμα για τη συνέχεια όπου τα δεδομένα θα χρησιμοποιηθούν από τα μοντέλα μηχανικής μάθησης. Έπειτα, πραγματοποιήθηκε σύγκριση κάθε αρχείου όπου κρατήσαμε μόνο τους κοινούς ασθενείς μεταξύ των αυτών. Έτσι δημιουργήθηκαν τρία νέα σύνολα δεδομένων. Επιπροσθέτως, σε κάθε νέα αναπαράσταση προστέθηκε στο τέλος μία νέα στήλη η οποία αφορά την το χαρακτηριστικό CDE_vital_status, γνωρίζοντας έτσι σε κάθε ομικό σύνολο δεδομένων την κατάσταση του κάθε ασθενή.

sampleID	AACS	FSTL1	ELMO2	CREB3L1	RPS11	PNMA1	MMP2	SAMD4A	SMARCD3	A4GNT	...	DHRS2	RAB8A	SGEF	P
TCGA.02.0001.01	6.500551	8.729663	5.511362	4.882953	10.984784	7.535193	8.674010	5.032552	4.710970	5.108478	...	7.153611	9.048851	4.366050	6.(
TCGA.02.0003.01	6.539245	9.794400	6.213981	4.836276	10.811245	6.997933	9.348590	5.026961	5.327734	4.348606	...	4.006913	8.864498	4.385314	6.(
TCGA.02.0007.01	7.186891	4.945053	5.230444	5.818606	10.477304	8.356117	4.429521	5.175938	4.440470	4.824183	...	4.290511	8.483590	4.798488	5.(
TCGA.02.0009.01	7.675038	10.840095	6.620676	5.333213	10.637267	6.942901	9.452231	5.164914	4.952207	4.204604	...	5.410476	9.239238	4.343192	6.(
TCGA.02.0010.01	7.996010	8.931571	7.552416	6.087341	11.001533	8.044375	4.501725	4.970135	8.638965	4.729682	...	4.388377	9.367783	4.219649	5.(
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
TCGA.41.2572.01	7.564477	10.626080	7.830206	4.874744	10.974133	9.512059	8.568646	5.998493	8.806333	4.407308	...	4.085567	8.949573	6.621713	5.(
TCGA.41.2573.01	8.352362	9.278117	8.222441	4.438118	10.934162	10.417580	7.775023	5.161021	9.229107	4.230057	...	4.084781	8.700344	5.308027	7.(
TCGA.41.2575.01	7.470726	9.022293	7.268008	4.886906	11.152242	9.087012	8.191436	5.380936	9.408020	4.376040	...	4.132281	8.511591	5.876536	6.(
TCGA.41.3393.01	7.147460	9.759453	7.701716	4.831633	10.938819	9.655464	8.573679	6.090131	8.512957	4.155312	...	4.044985	9.273159	6.305258	5.(
TCGA.41.3915.01	6.700266	11.117519	6.831318	4.405531	11.047735	9.554216	7.930944	5.740704	7.871436	4.428249	...	3.862120	8.889573	4.928122	6.(

230 rows × 12043 columns

Εικόνα 14: Σύνολο δεδομένων Gene expression με τους κοινούς ασθενείς

sampleID	cg00000292	cg00002426	cg00005847	cg00015770	cg00027083	cg00029931	cg00030047	cg00041575	cg00043004	cg00056767	...	cg2762261
TCGA.02.0001.01	0.826063	0.178659	0.381919	0.045782	0.025546	0.073651	0.326046	0.640585	0.637638	0.030499	...	0.61574
TCGA.02.0003.01	0.696610	0.555948	0.851630	0.035110	0.879701	0.063262	0.819071	0.549288	0.840612	0.082296	...	0.55922
TCGA.02.0007.01	0.826712	0.088165	0.404206	0.218564	0.019343	0.393609	0.923292	0.770041	0.888077	0.173058	...	0.60895
TCGA.02.0009.01	0.836044	0.194659	0.635270	0.367296	0.882392	0.541412	0.849001	0.709482	0.407752	0.078154	...	0.39827
TCGA.02.0010.01	0.687137	0.912093	0.876321	0.114126	0.019867	0.025253	0.495842	0.696550	0.868179	0.932783	...	0.02395
...	...	...	...	...	...	...	...	...	...	...	...	...
TCGA.41.2572.01	0.597697	0.131811	0.577767	0.722128	0.786483	0.020554	0.824263	0.734042	0.389067	0.101514	...	0.37880
TCGA.41.2573.01	0.400045	0.100287	0.470516	0.475888	0.623773	0.019328	0.392965	0.329292	0.255324	0.028058	...	0.18572
TCGA.41.2575.01	0.064723	0.068986	0.614811	0.391715	0.152668	0.026295	0.826769	0.606009	0.291700	0.027509	...	0.09372
TCGA.41.3393.01	0.704422	0.286588	0.615784	0.112660	0.486311	0.408998	0.727590	0.716248	0.044905	0.034523	...	0.46250
TCGA.41.3915.01	0.475091	0.192033	0.531287	0.117995	0.140576	0.020317	0.452063	0.456375	0.715601	0.180295	...	0.35021

230 rows × 5001 columns

Εικόνα 15: Σύνολο δεδομένων DNA Methylation με τους κοινούς ασθενείς

sampleID	ebv-miR-BART1-3p	ebv-miR-BART1-5p	ebv-miR-BART10	ebv-miR-BART11-3p	ebv-miR-BART11-5p	ebv-miR-BART12	ebv-miR-BART13	ebv-miR-BART14-3p	ebv-miR-BART14-5p	ebv-miR-BART15	...	kshv-miR-K12-4-3p	kshv-miR-K12-4-5p	kshv-miR-K12-5	ksh mi K12-
TCGA.02.0001.01	5.855126	5.799428	5.862059	5.608860	5.812956	5.932238	7.417908	5.769454	5.789184	5.946758	...	5.745194	5.811312	5.736368	5.7280
TCGA.02.0003.01	5.801614	5.790478	5.818763	5.613089	5.768700	5.870240	7.314205	5.747051	5.795356	5.853679	...	5.765822	5.800511	5.721188	6.3796
TCGA.02.0007.01	5.818828	5.800582	5.818181	5.585730	5.785541	5.897554	7.347849	5.804559	5.803284	5.802757	...	5.723913	5.850117	5.724573	5.7638
TCGA.02.0009.01	5.766792	5.812545	5.888331	5.948601	5.803021	5.935245	7.933744	5.756250	5.754840	5.839917	...	5.775323	5.835817	5.725099	5.7596
TCGA.02.0010.01	5.830012	5.762413	5.805194	5.976939	5.766500	5.866709	7.811017	5.791926	5.744780	5.791533	...	5.775821	5.780900	5.744699	5.7493
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
TCGA.41.2572.01	5.876953	5.856112	5.905694	6.153589	5.866547	5.869405	10.290012	5.807951	5.995493	5.832601	...	5.875839	5.853907	5.827569	5.8336
TCGA.41.2573.01	5.768540	5.782876	5.821910	5.952710	5.750902	5.845650	7.598841	5.780203	5.776816	5.821024	...	5.776287	5.832648	5.734893	5.7418
TCGA.41.2575.01	5.760604	5.740049	5.809435	6.067070	5.732453	5.799317	7.897456	5.773762	5.781377	5.723617	...	5.775477	5.826861	5.734939	5.7615
TCGA.41.3393.01	5.744980	5.712206	5.770991	6.009412	5.729474	5.784114	8.206138	5.777640	5.823528	5.750772	...	5.764739	5.896441	5.700280	5.7478
TCGA.41.3915.01	5.762561	5.749826	5.791805	5.977554	5.730791	5.819341	7.867461	5.774027	5.761531	5.748902	...	5.795322	5.824176	5.773722	5.7511

230 rows × 535 columns

Εικόνα 16: Σύνολο δεδομένων miRNA με τους κοινούς ασθενείς

Βλέπουμε ότι σε κάθε σύνολο δεδομένων έχει μειωθεί ο αριθμός των ασθενών (γραμμών) γιατί έμειναν μόνο οι κοινοί ασθενείς μεταξύ των αρχείων.

Στη συνέχεια, παρατηρούμε από τα στατιστικά στοιχεία ότι αρκετά χαρακτηριστικά παρουσιάζουν διακυμάνσεις στο εύρος τιμών τους. Για παράδειγμα, στο σύνολο δεδομένων Gene expression το γονίδιο FSTL1 αποτελείται από τιμές που κυμαίνονται μεταξύ 3.66 και 12.05 συγκριτικά με το γονίδιο ELMO2 που έχει τιμές από 4.93 – 8.91.

	AACS	FSTL1	ELMO2	CREB3L1	RPS11	PNMA1	MMP2	SAMD4A	SMARCD3	A4GNT	...	KIAA0802	DHRS2
count	538.000000	538.000000	538.000000	538.000000	538.000000	538.000000	538.000000	538.000000	538.000000	538.000000	...	538.000000	538.000000
mean	6.857476	9.698384	7.033371	4.798021	10.717471	9.540024	7.712223	5.563474	8.322169	4.381581	...	6.626050	4.238019
std	0.580120	1.057241	0.685507	0.498979	0.393487	0.821276	1.175252	0.379776	1.036997	0.197806	...	0.815168	0.566826
min	5.536622	3.662355	4.930723	3.931335	9.643697	5.443112	3.793852	4.827805	3.674979	3.892355	...	4.677598	3.651946
25%	6.486874	9.152143	6.675970	4.422475	10.451898	9.213701	7.095800	5.286227	7.972591	4.230066	...	6.044459	4.008727
50%	6.829574	9.835648	7.057627	4.737883	10.653345	9.663865	7.779782	5.557519	8.500311	4.369468	...	6.627615	4.120118
75%	7.155040	10.396886	7.499242	5.036134	10.913628	10.067844	8.441331	5.777569	8.981484	4.499589	...	7.107895	4.260605
max	9.394954	12.059550	8.919199	8.120004	12.143343	11.417741	11.248461	7.825115	10.265944	5.231504	...	9.927499	10.938656

8 rows × 12042 columns

Εικόνα 17: Τα στατιστικά δεδομένα πριν την εφαρμογή ομαλοποίησης

Αυτή η διακύμανση τιμών κάποιες φορές μπορεί να επηρεάσει την απόδοση του μοντέλου μηχανικής μάθησης αλλά και την ικανότητα του στην φάση της εκπαίδευσης των δεδομένων. Ο συνηθισμένος τρόπος αντιμετώπισης της κατάστασης αυτής είναι η **ομαλοποίηση** (Normalization).

Κατά τη διαδικασία της ομαλοποίησης όλες οι τιμές τροποποιούνται και λαμβάνουν συνήθως εύρος από **0–1** για όλα τα χαρακτηριστικά του συνόλου δεδομένων. Πιο συγκεκριμένα, αυτή η διαδικασία χρησιμοποιεί τον αλγόριθμο:

$$X' = X(X - Xmin)/(Xmax - Xmin)$$

Όπου Xmax και Xmin είναι οι μέγιστες και ελάχιστες τιμές κάθε χαρακτηριστικού αντίστοιχα [52].

Με τη χρήση της συνάρτησης **MinMaxScaler** της python εφαρμόζουμε την παραπάνω διαδικασία σε κάθε σύνολο δεδομένων. Όπως βλέπουμε για το σύνολο δεδομένων των γονιδίων οι τιμές έχουν πλέον εύρος από 0-1. Αντίστοιχα γίνεται και στα υπόλοιπα σύνολα.

	AACS	FSTL1	ELMO2	CREB3L1	RPS11	PNMA1	MMP2	SAMD4A	SMARCD3	A4GNT	...	KIAA0802	DHRS2	23
count	230.000000	230.000000	230.000000	230.000000	230.000000	230.000000	230.000000	230.000000	230.000000	230.000000	...	230.000000	230.000000	
mean	0.368888	0.740680	0.618194	0.194603	0.358460	0.678348	0.588200	0.379900	0.696455	0.319337	...	0.374688	0.075382	
std	0.130531	0.135118	0.194964	0.125225	0.190049	0.167652	0.173836	0.194197	0.182276	0.167203	...	0.161551	0.097937	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
25%	0.289473	0.679618	0.507440	0.101155	0.230263	0.603674	0.502323	0.225978	0.623180	0.204820	...	0.259979	0.036849	
50%	0.373759	0.755946	0.631930	0.179011	0.321386	0.711946	0.598653	0.369299	0.733030	0.311800	...	0.372442	0.053821	
75%	0.434407	0.826870	0.750269	0.253213	0.449348	0.790654	0.697849	0.492338	0.808007	0.400245	...	0.467614	0.070622	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	

8 rows × 12042 columns

Εικόνα 18: Τα στατιστικά δεδομένα μετά την εφαρμογή ομαλοποίησης

Στο κεφάλαιο 3 αναλύσαμε στρατηγικές που αφορούν την ενσωμάτωση (integration) δεδομένων. Στην εν λόγω μελέτη, εφαρμόσαμε την στρατηγική early integration. Αρχικά, έχουμε ήδη προ-επεξεργαστεί τα δεδομένα με τις παραπάνω μεθοδολογίες.

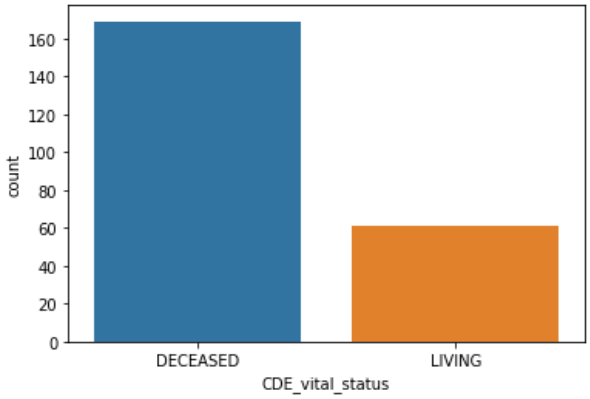
Στη συνέχεια, προχωρήσαμε στην συνένωση των τριών συνόλων δεδομένων DNA methylation, miRNA και Gene expression σε ένα νέο καινούργιο πίνακα που περιέχει όλα τα σύνολα δεδομένων κρατώντας μόνο τους κοινούς ασθενείς μεταξύ τους. Έτσι, έχουμε καταλήξει σε μια αναπαράσταση με 230 κοινούς ασθενείς (rows) και 17577 χαρακτηριστικά (columns).

sampleID	AACS	FSTL1	ELMO2	CREB3L1	RPS11	PNMA1	MMP2	SAMD4A	SMARCD3	A4GNT	...	kshv-miR-K12-4-3p	kshv-miR-K12-4-5p	kshv-miR-K12-5	ks m K1
TCGA.02.0001.01	0.249831	0.614165	0.149741	0.227189	0.451030	0.292097	0.718822	0.092452	0.160680	1.000000	...	0.086404	0.136568	0.166382	0.082
TCGA.02.0003.01	0.259859	0.743213	0.370099	0.216045	0.368801	0.182436	0.818184	0.089434	0.256339	0.315462	...	0.119229	0.120196	0.133401	1.000
TCGA.02.0007.01	0.427716	0.155465	0.061639	0.450566	0.210567	0.459657	0.093631	0.169854	0.118726	0.743891	...	0.052540	0.195386	0.140754	0.133
TCGA.02.0009.01	0.554233	0.869953	0.497648	0.334683	0.286364	0.171204	0.833450	0.163903	0.198096	0.185737	...	0.134347	0.173711	0.141897	0.127
TCGA.02.0010.01	0.637423	0.638637	0.789864	0.514724	0.458967	0.396027	0.104266	0.058759	0.769906	0.658758	...	0.135140	0.090472	0.184482	0.112
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
TCGA.41.2572.01	0.525578	0.844014	0.876986	0.225229	0.445983	0.695598	0.703303	0.613876	0.795864	0.368345	...	0.294295	0.201132	0.364532	0.231
TCGA.41.2573.01	0.729782	0.680639	1.000000	0.120989	0.427044	0.880425	0.586406	0.161801	0.861436	0.208667	...	0.135882	0.168908	0.163176	0.102
TCGA.41.2575.01	0.501280	0.649632	0.700667	0.228132	0.530378	0.608841	0.647741	0.280514	0.889185	0.340176	...	0.134593	0.160137	0.163276	0.129
TCGA.41.3393.01	0.417496	0.738977	0.836688	0.214937	0.429251	0.724868	0.704044	0.663343	0.750362	0.141332	...	0.117506	0.265601	0.087973	0.110
TCGA.41.3915.01	0.301592	0.903577	0.563711	0.113209	0.480859	0.704203	0.609372	0.474719	0.650863	0.387209	...	0.166172	0.156067	0.247540	0.115

230 rows × 17577 columns

Εικόνα 19: Ενοποιημένο σύνολο δεδομένων

Μετά την ενοποίηση των δεδομένων παρατηρήθηκε ότι στην κλάση DECEASED (169 δείγματα) υπάρχουν περισσότερα δείγματα από την κατηγορία LIVING (61 δείγματα) δημιουργώντας έτσι ένα ανισόρροπο σετ δεδομένων (Imbalanced dataset).



Εικόνα 20: Ανισόρροπο ενοποιημένο σύνολο δεδομένων

Αυτό έχει ως αποτέλεσμα τα μοντέλα μηχανικής μάθησης να μην μπορούν να προβλέψουν σωστά την κλάση με τα λιγότερα δείγματα. Να σημειωθεί ότι το φαινόμενο αυτό υφίσταται και στο κάθε σύνολο δεδομένων μεμονωμένα όπου έχουμε:

Για Gene expression σύνολο δεδομένων:

```
DECEASED    169
LIVING       61
```

Εικόνα 21: Ανισόρροπο σύνολο δεδομένων για gene expression

Για DNA methylation σύνολο δεδομένων:

```
DECEASED    169
LIVING       61
```

Εικόνα 22: Ανισόρροπο σύνολο δεδομένων για DNA methylation

Για miRNA σύνολο δεδομένων:

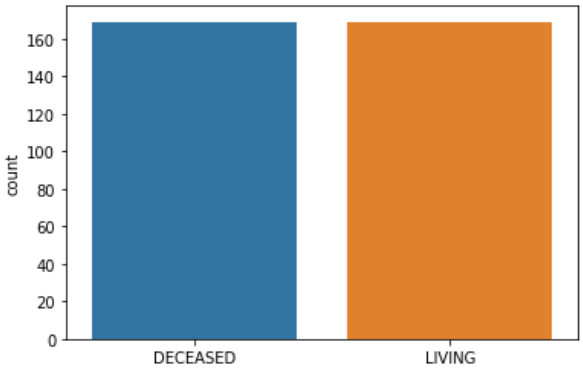
```
DECEASED    169
LIVING       61
```

Εικόνα 23: Ανισόρροπο σύνολο δεδομένων για miRNA

Εδώ έρχεται να δώσει λύση ο αλγόριθμος **SMOTE** (Synthetic Minority Oversampling Technique). Ο εν λόγω αλγόριθμος **αυξάνει το μέγεθος** (oversampling) της μειοψηφούσας κλάσης δημιουργώντας **νέα συνθετικά δείγματα** αντί να αντιγράφει από τα ήδη υπάρχοντα τα οποία δεν παρέχουν πρόσθετες πληροφορίες στο μοντέλο. Ο αλγόριθμος SMOTE λειτουργεί επιλέγοντας δείγματα τα οποία είναι κοντά

στον χώρο των χαρακτηριστικών (feature space) «χαράζοντας» μια γραμμή ανάμεσα στα δείγματα στον χώρο των χαρακτηριστικών και τέλος δημιουργεί ένα νέο δείγμα κατά το μήκος αυτής της γραμμής. Αρχικά, επιλέγει τυχαία ένα δείγμα (έστω a) από την μειοψηφούσα κλάση και βρίσκει τους k πιο κοντινούς γείτονες του (συνήθως k=5). Ένας τυχαίος γείτονας (έστω b) επιλέγεται και δημιουργείται ένα συνθετικό δείγμα σε τυχαίο σημείο ανάμεσα στα δύο δείγματα στον χώρο των χαρακτηριστικών [53][54].

Με την εφαρμογή λοιπόν του SMOTE αλγορίθμου στο ενοποιημένο σετ δεδομένων σχηματίστηκε η εξής εικόνα όπου η μειοψηφούσα κλάση γίνεται ίση με την πλειοψηφούσα:



Εικόνα 24: Ανισόρροπο ενοποιημένο σύνολο δεδομένων με χρήση SMOTE

Πριν τα δεδομένα εισαχθούν στους αλγορίθμους, πρέπει να περάσουν μια ακόμη διαδικασία.

Αρχικά, για κάθε ανάλυση χωρίσαμε τα σύνολα δεδομένων σε **δεδομένα εκπαίδευσης** (Train set) **και δεδομένα δοκιμής** (Test set) με τη χρήση της συνάρτησης **train_test_split** της Scikit-Learn. Αυτό είναι μια αναγκαία διαδικασία που πρέπει να γίνει καθώς οι αλγόριθμοι μηχανικής μάθησης πρέπει να εκπαιδευτούν με δεδομένα και έπειτα να δοκιμαστούν σε δεδομένα που δεν έχουν εκπαιδευτεί, έτσι ώστε να φανεί η αποτελεσματικότητα τους. Το σύνολο των δεδομένων δοκιμής θα πρέπει να είναι αρκετά μεγάλο ώστε να αποφέρει σημαντικά στατιστικά αποτελέσματα και να αντιπροσωπεύει το σύνολο δεδομένων, δηλαδή να έχει διαφορετικά χαρακτηριστικά από τα δεδομένα εκπαίδευσης. Σκοπός είναι να αναπτυχθεί ένα μοντέλο που θα είναι αποτελεσματικό σε δεδομένα που δεν έχει εκπαιδευτεί όπου αυτό επιτυγχάνεται με τη διαδικασία που προαναφέρθηκε [55].

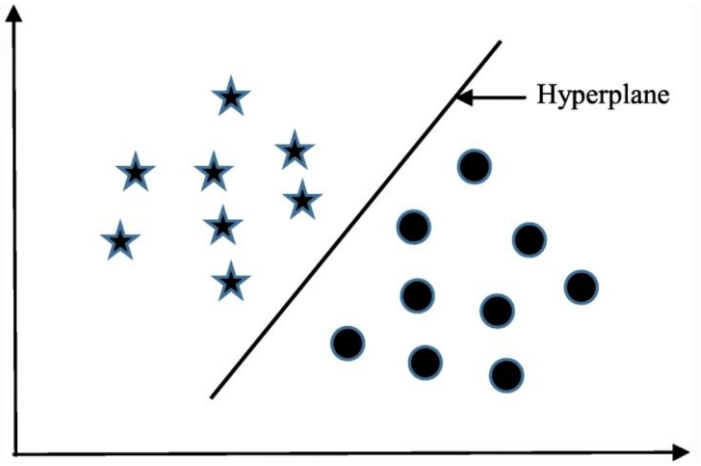
Να σημειωθεί, ότι όταν ένα μοντέλο είναι εκπαιδευμένο πάρα πολύ καλά στα δεδομένα εκπαίδευσης, αποτυγχάνει να κάνεις σωστές προβλέψεις σε νέα δεδομένα. Αυτό το φαινόμενο ονομάζεται **υπερπροσαρμογή** (Overfitting) και πρέπει να αποφεύγεται [56].

Στο σημείο αυτό έχουμε ολοκληρώσει την προ-επεξεργασία των δεδομένων που απαιτούνται για τη συνέχεια της εν λόγω πτυχιακής. Παρακάτω θα προχωρήσουμε στην κατασκευή των μοντέλων μηχανικής μάθησης. Θα έχει ιδιαίτερο ενδιαφέρον να προσέξουμε το πως θα συμπεριφερθούν τα μοντέλα, δηλαδή πόσο «καλά» θα προβλέψουν χρησιμοποιώντας δεδομένα που έχουν και δεδομένα που δεν έχουν υποστεί εφαρμογή του SMOTE αλγορίθμου.

6.3 Κατασκευή των μοντέλων μηχανικής μάθησης

Για τη κατηγοριοποίηση των δεδομένων επιλέχτηκαν οι γνωστοί αλγόριθμοι **SVM** και **Decision Tree**. Οι εν λόγω αλγόριθμοι, χρησιμοποιούνται συχνά στην ανάλυση βιολογικών δεδομένων [57].

Ο SVM (Support Vector Machine) είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται σε προβλήματα επιτηρούμενης μάθησης και μπορεί να ταξινομήσει **γραμμικά** (linear) και **μη γραμμικά** (non-linear) δεδομένα. Στην συνέχεια μπορεί από μόνος του να αναγνωρίσει το υπερπλάνο (hyperplane) το οποίο χωρίζει τα δεδομένα σε δυο κλάσεις μεγιστοποιώντας την οριακή απόσταση μεταξύ των δύο κλάσεων και ελαχιστοποιώντας τα σφάλματα που προκύπτουν κατά τη διάρκεια της ταξινόμησης. Τέλος, να σημειωθεί ότι για να πραγματοποιηθεί μια ταξινόμηση πρέπει να βρεθεί το υπερπλάνο που διαχωρίζει τις δύο κλάσεις κατά το μέγιστο περιθώριο [58][59].



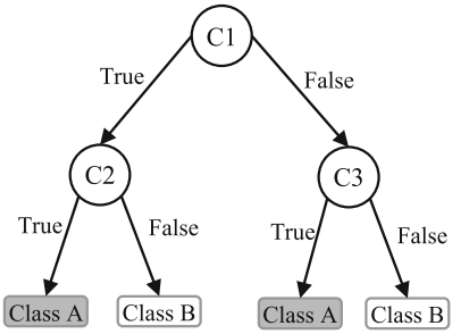
Εικόνα 25: Παράδειγμα υπερπλάνου SVM. Πηγή: https://link.springer.com/article/10.1186/s12911-019-1004-8 (Ανακτήθηκε 27/4/2022)

Πλεονεκτήματα του αλγορίθμου αυτού είναι η αποτελεσματικότητα σε χώρους μεγάλων διαστάσεων δηλαδή μπορεί να αντιμετωπίσει μεγάλο όγκο δεδομένων, αποτελεσματικότητα σε δεδομένα όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων. Επιπροσθέτως, χρησιμοποιεί υποσύνολα (subsets) των δεδομένων εκπαίδευσης για τη συνάρτηση απόφασης (decision function) γεγονός που τον καθιστά αποτελεσματικό στην διαχείριση της μνήμης. Τέλος, στα πλεονεκτήματα περιλαμβάνεται και η ικανότητα να χρησιμοποιεί διαφορετικούς πυρήνες (Kernels) για την συνάρτηση απόφασης.

Στα μειονεκτήματα, ο εν λόγω αλγόριθμος, όταν ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος από τον αριθμό των δειγμάτων τότε μπορεί να υπάρξει το πρόβλημα της υπερπροσαρμογής (Overfitting). Προκειμένου να αποφευχθεί αυτό, πρέπει να χρησιμοποιηθούν διαφορετικοί πυρήνες. Επίσης, δεν παρέχει άμεσα τις εκτιμήσεις των πιθανοτήτων, αυτές υπολογίζονται βάση μιας ακριβής πενταπλάσιας διασταυρούμενης επικύρωσης (5-fold cross-validation) [60].

Ο Decision Tree αποτελεί έναν από τους παλαιότερους αλγορίθμους μηχανικής μάθησης. Ταξινομεί τα δεδομένα βάση των τιμών των χαρακτηριστικών οπτικοποιώντας τα σε **δενδροειδή μορφή**. Κάθε κόμβος αναπαριστά ένα χαρακτηριστικό (feature) για ταξινόμηση ενώ κάθε κλαδί αναπαριστά μια τιμή που ο κόμβος μπορεί να υποθέσει. Η ταξινόμηση αρχίζει από τον κόμβο ρίζα (root node). Τα φύλλα ή τερματικοί κόμβοι αντιστοιχούν στα αποτελέσματα της απόφασης του δέντρου απόφασης. Επίσης, να σημειωθεί, ότι χρησιμοποιείται αρκετά από ιατρικά

διαγνωστικά πρωτόκολλα καθώς εύκολα μπορεί κανείς να ερμηνεύσει τα αποτελέσματα του αλλά και να μάθει να το χειρίζεται. Τέλος, οι ταξινομητές (classifiers) με τη χρήση μετα-κλαδέματος (post-pruning) πραγματοποιούν αξιολόγηση της απόδοσης χρησιμοποιώντας ένα σετ επικύρωσης (validation set) κατά τη διάρκεια του κλαδέματος [58][59].



Εικόνα 26: Παράδειγμα λειτουργίας Decision Tree. Πηγή: https://link.springer.com/article/10.1186/s12911-019-1004-8 (Ανακτήθηκε 27/4/2022)

Χωρίσαμε τα σύνολα δεδομένα σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής πριν τα εφαρμόσουμε στους αλγορίθμους που αναφερθήκαμε παραπάνω. Πιο συγκεκριμένα στο παρακάτω στιγμιότυπο παρουσιάζονται τα δεδομένα που αφορούν το ενοποιημένο σύνολο δεδομένων.

```
X_train: (161, 17576)
X_test: (69, 17576)
y_train: (161,)
y_test: (69,)
```

Εικόνα 27 Δεδομένα εκπαίδευσης και δοκιμής του ενοποιημένου συνόλου δεδομένων

Από την παραπάνω εικόνα βλέπουμε ότι τα δεδομένα εκπαίδευσης αποτελούνται από 161 δείγματα, αντιπροσωπεύοντας το 70% του συνολικού αριθμού των δειγμάτων και τα δεδομένα δοκιμής από 69 δείγματα τα οποία αφορούν το υπόλοιπο 30%.

Στη συνέχεια, προχωρήσαμε στην εφαρμογή των συνόλων δεδομένων στους αλγορίθμους SVM & Decision Tree. Να σημειωθεί ότι στην προκειμένη φάση, δοκιμάζουμε τα μοντέλα **χωρίς** την χρήση του αλγορίθμου SMOTE.

Για το σύνολο δεδομένων Gene expression έχουμε:

```
SVM Accuracy: 0.7101449275362319
DecisionTree Accuracy: 0.6231884057971014
```

Εικόνα 28: Εφαρμογή SVM & DT στο σύνολο δεδομένων Gene expression

Για το σύνολο δεδομένων DNA Methylation έχουμε:

```
SVM Accuracy: 0.782608695652174
DecisionTree Accuracy: 0.6086956521739131
```

Εικόνα 29: Εφαρμογή SVM & DT στο σύνολο δεδομένων DNA Methylation

Για το σύνολο δεδομένων miRNA έχουμε:

```
SVM Accuracy: 0.5652173913043478
DecisionTree Accuracy: 0.5362318840579711
```

Εικόνα 30: Εφαρμογή SVM & DT στο σύνολο δεδομένων miRNA

Για το ενοποιημένο σύνολο δεδομένων έχουμε:

```
SVM Accuracy: 0.6956521739130435
DecisionTree Accuracy: 0.6086956521739131
```

Εικόνα 31: Εφαρμογή SVM & DT στο ενοποιημένο σύνολο δεδομένων

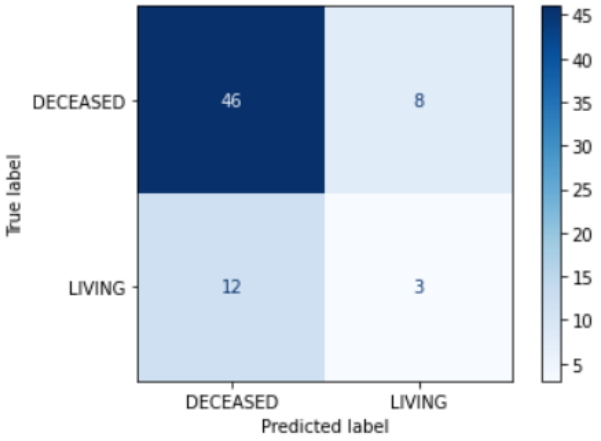
Στα παραπάνω αποτελέσματα παρατηρήθηκε ότι παρόλο το ποσοστό της ακρίβειας που επιτεύχθηκε από τους αλγορίθμους, τα μοντέλα πρόβλεπαν τα περισσότερα δεδομένα της κλάσης LIVING ως DECEASED. Αυτό συμβαίνει διότι τα δείγματα της κλάσης DECEASED είναι πολύ περισσότερα από της κλάσης LIVING δημιουργώντας το φαινόμενο του μη ισορροπημένου συνόλου δεδομένων που αναφερθήκαμε παραπάνω. Αυτό το διαπιστώνουμε από τον **πίνακα σύγχυσης** (confusion matrix).

Ο εν λόγω πίνακας, χρησιμοποιείται σε προβλήματα ταξινόμησης και δείχνει που έχουν γίνει **λάθη** στο μοντέλο που φτιάξαμε. Οι γραμμές (rows) αντιπροσωπεύουν τις πραγματικές κλάσεις που θα έπρεπε να είναι τα αποτελέσματα. Οι στήλες (rows) αντιπροσωπεύουν τις προβλέψεις που έχει κάνει το μοντέλο. Έτσι, χρησιμοποιώντας

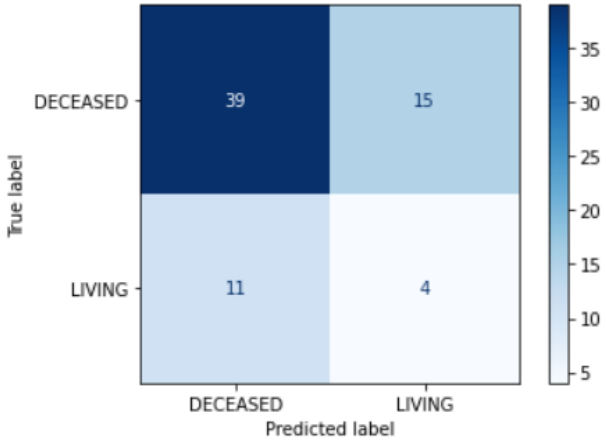
τον πίνακα σύγχυσης μπορούμε να δούμε πολύ εύκολα ποιες προβλέψεις είναι λανθασμένες [61].

Παρακάτω βλέπουμε τους πίνακες σύγχυσης που προέκυψαν για κάθε σύνολο δεδομένων ξεχωριστά αλλά και από το ενοποιημένο σύνολο.

Για το σύνολο δεδομένων Gene Expression έχουμε:

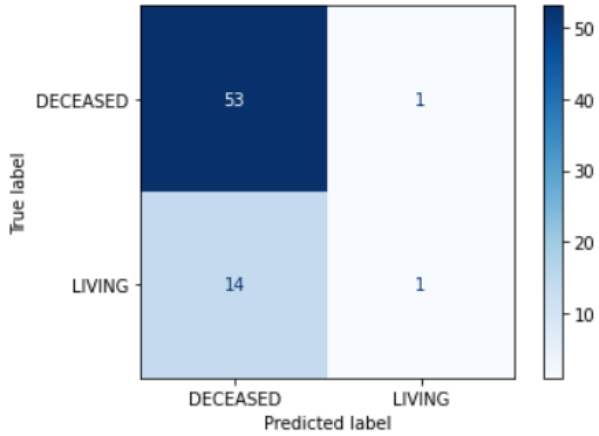


Εικόνα 32: Πίνακας σύγχυσης του συνόλου δεδομένων Gene Expression του αλγορίθμου SVM

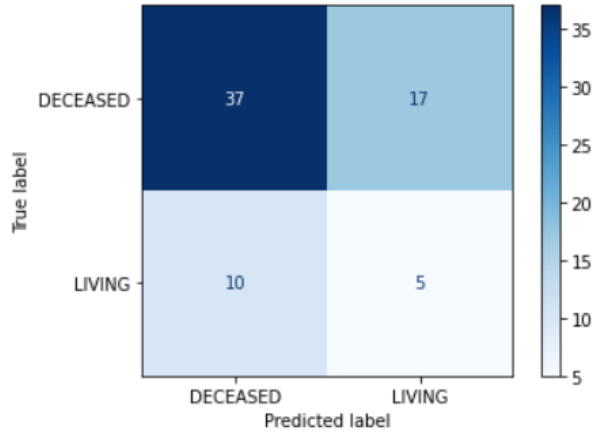


Εικόνα 33: Πίνακας σύγχυσης του συνόλου δεδομένων Gene Expression του αλγορίθμου Decision Tree

Για το σύνολο δεδομένων DNA Methylation έχουμε:

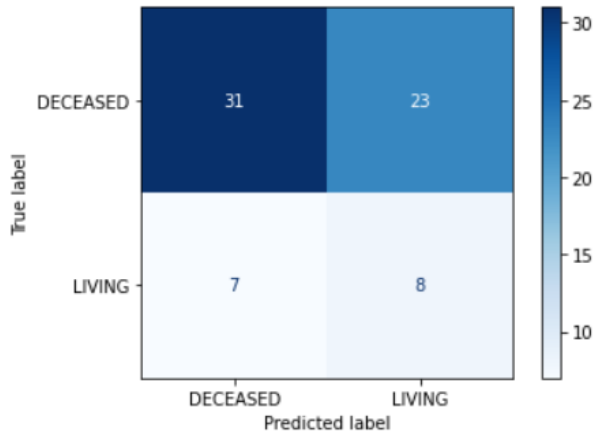


Εικόνα 34: Πίνακας σύγχυσης του συνόλου δεδομένων DNA Methylation του αλγορίθμου SVM

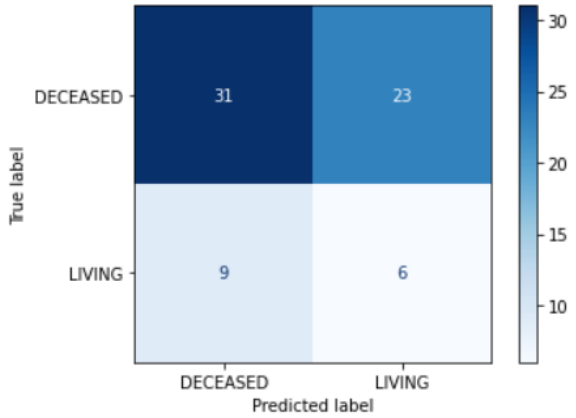


Εικόνα 35: Πίνακας σύγχυσης του συνόλου δεδομένων DNA Methylation του αλγορίθμου Decision Tree

Για το σύνολο δεδομένων miRNA έχουμε:

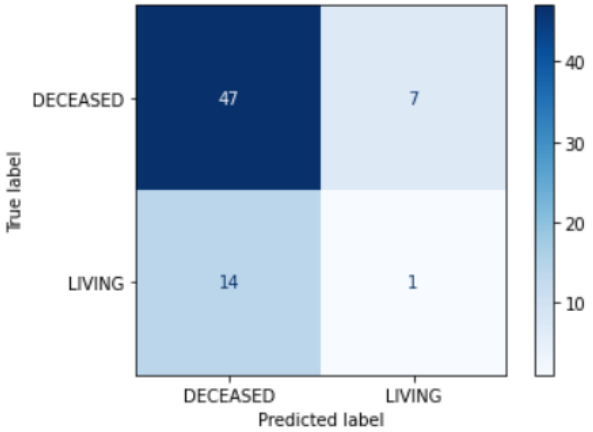


Εικόνα 36: Πίνακας σύγχυσης του συνόλου δεδομένων miRNA του αλγορίθμου SVM

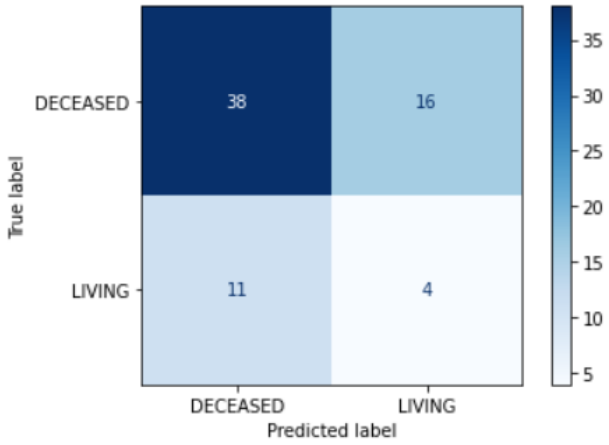


Εικόνα 37: Πίνακας σύγχυσης του συνόλου δεδομένων miRNA του αλγορίθμου Decision Tree

Για το ενοποιημένο σύνολο δεδομένων έχουμε:



Εικόνα 38: Πίνακας σύγχυσης του ενοποιημένου συνόλου δεδομένων του αλγορίθμου SVM



Εικόνα 39: Πίνακας σύγχυσης του ενοποιημένου συνόλου δεδομένων του αλγορίθμου Decision Tree

Από τα παραπάνω στοιχεία παρατηρούμε ότι ο αλγόριθμος SVM διαχειρίζεται πιο καλά τα δεδομένα σε σχέση με τον Decision Tree σημειώνοντας λιγότερα λάθη στις προβλέψεις του. Επίσης, όπως αναφερθήκαμε και παραπάνω, βλέπουμε ότι και τα δυο

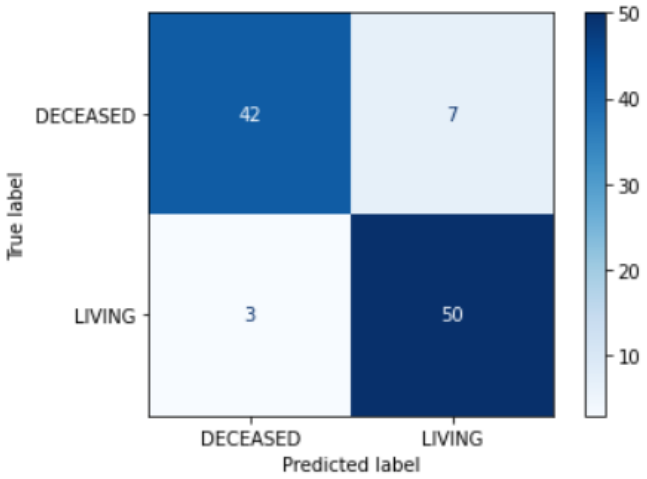
μοντέλα, αδυνατούν να κάνουν τη καλή πρόβλεψη της κλάσης LIVING, θεωρώντας την ως DECEASED λόγω του μη ισορροπημένου συνόλου δεδομένων.

Λόγω του συγκεκριμένου γεγονότος, έγινε η χρήση του αλγορίθμου SMOTE (όπως αναλύθηκε και στο κεφάλαιο 6.2). Με αυτόν τον τρόπο επιδιώκουμε την αύξηση των χαρακτηριστικών της μειοψηφούσας κλάσης επιδιώκοντας οι αλγόριθμοι μηχανικής μάθησης να «βοηθηθούν» ώστε να μπορέσουν να έχουν μεγαλύτερη ακρίβεια στις προβλέψεις τους. Πιο συγκεκριμένα, εφαρμόζουμε τον αλγόριθμο SMOTE σε κάθε σύνολο δεδομένων, γίνεται διαχωρισμός σε δεδομένα εκπαίδευσης και δοκιμής και στη συνέχεια γίνεται η εκπαίδευση των μοντέλων. Παρακάτω παρατηρούμε τους πίνακες σύγχυσης αλλά και το ποσοστό ακρίβειας το πως αυτό διαμορφώνεται με τη χρήση του εν λόγω αλγορίθμου (SMOTE).

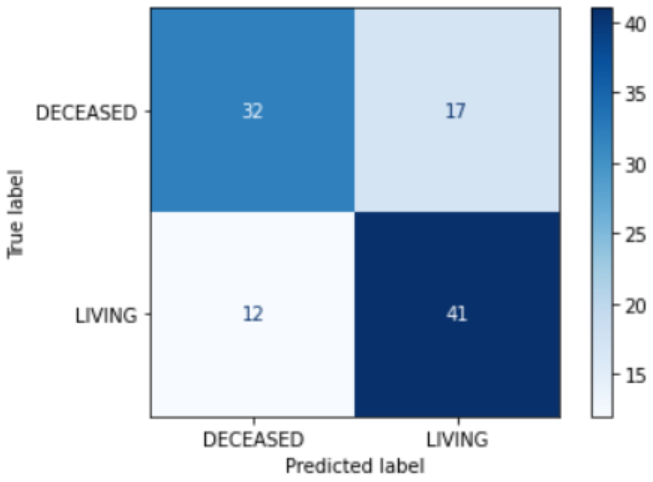
Για το σύνολο δεδομένων Gene Expression:

```
SVM Accuracy: 0.9019607843137255
DecisionTree Accuracy: 0.7156862745098039
```

Εικόνα 40: Εφαρμογή SVM & DT στο σύνολο δεδομένων Gene expression με τη χρήση SMOTE



Εικόνα 41: Πίνακας σύγχυσης του συνόλου δεδομένων Gene Expression του αλγορίθμου SVM με τη χρήση SMOTE

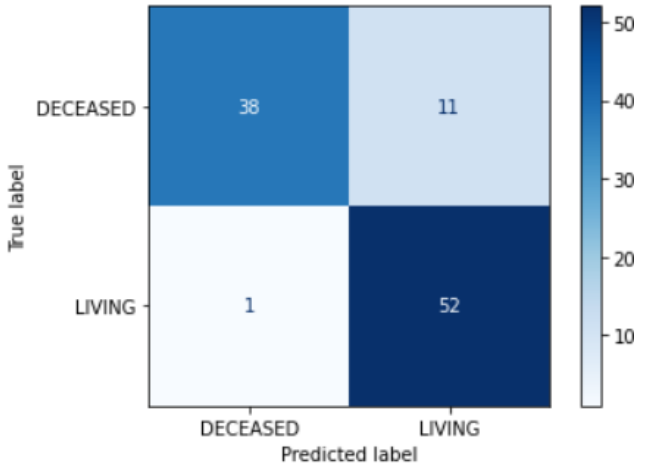


Εικόνα 42: Πίνακας σύγχυσης του συνόλου δεδομένων Gene Expression του αλγορίθμου Decision Tree με τη χρήση SMOTE

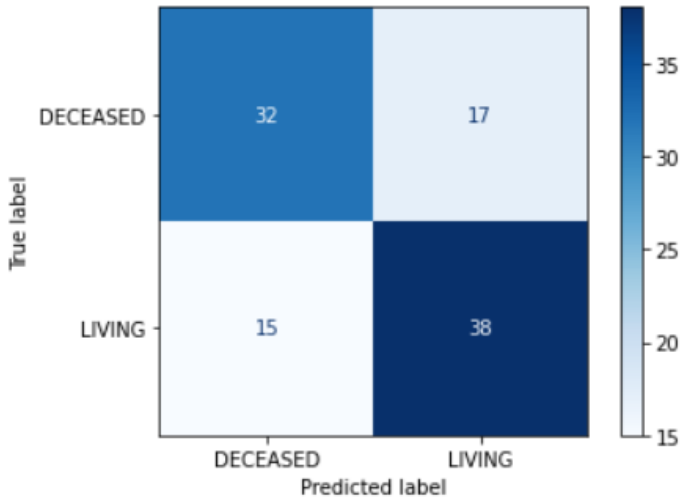
Για το σύνολο δεδομένων DNA Methylation έχουμε:

```
SVM Accuracy: 0.8823529411764706
DecisionTree Accuracy: 0.6862745098039216
```

Εικόνα 43: Εφαρμογή SVM & DT στο σύνολο δεδομένων DNA Methylation με τη χρήση SMOTE



Εικόνα 44: Πίνακας σύγχυσης του συνόλου δεδομένων DNA Methylation του αλγορίθμου SVM με τη χρήση SMOTE

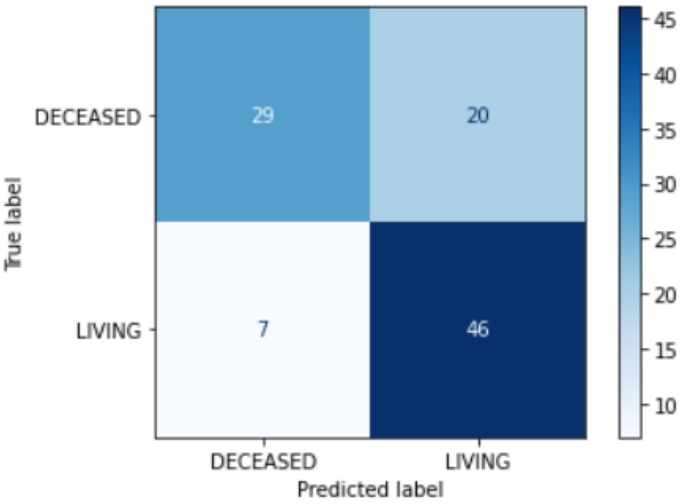


Εικόνα 45: Πίνακας σύγχυσης του συνόλου δεδομένων DNA Methylation του αλγορίθμου Decision Tree με τη χρήση SMOTE

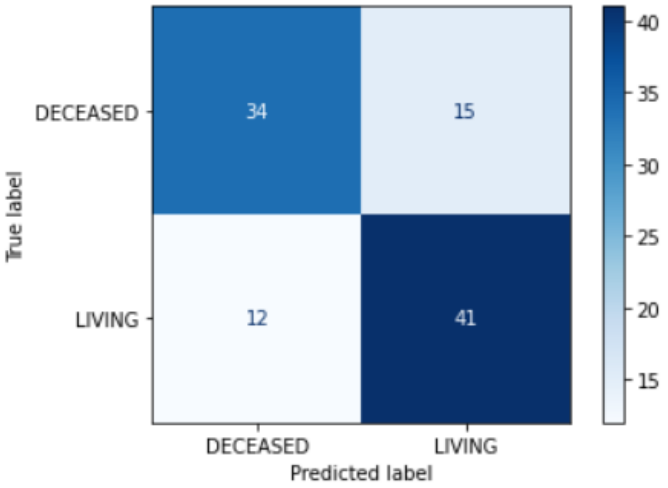
Για το σύνολο δεδομένων miRNA έχουμε:

```
SVM Accuracy: 0.7352941176470589
DecisionTree Accuracy: 0.7352941176470589
```

Εικόνα 46: Εφαρμογή SVM & DT στο σύνολο δεδομένων miRNA με τη χρήση SMOTE



Εικόνα 47: Πίνακας σύγχυσης του συνόλου δεδομένων miRNA του αλγορίθμου SVM με τη χρήση SMOTE

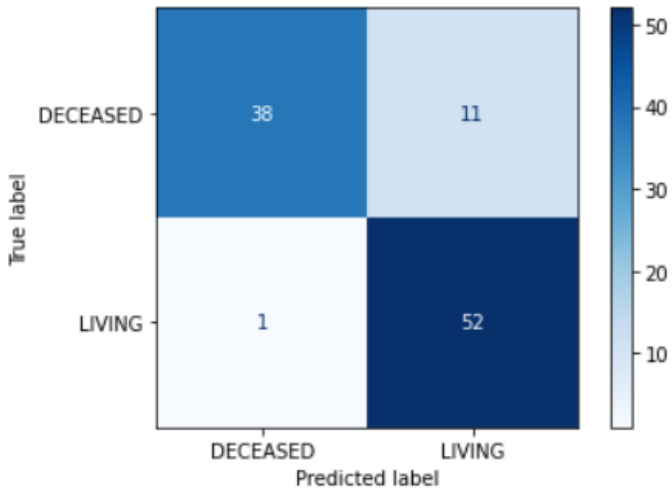


Εικόνα 48: Πίνακας σύγχυσης του συνόλου δεδομένων miRNA του αλγορίθμου Decision Tree με τη χρήση SMOTE

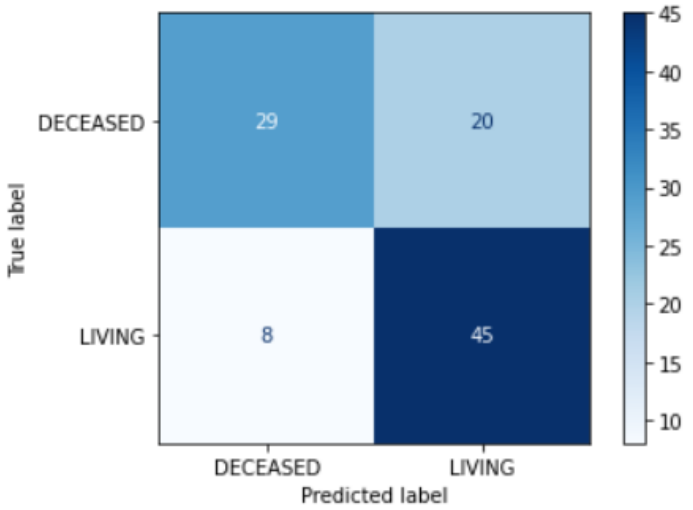
Για το ενοποιημένο σύνολο δεδομένων έχουμε:

```
SVM Accuracy: 0.8823529411764706
DecisionTree Accuracy: 0.7254901960784313
```

Εικόνα 49: Εφαρμογή SVM & DT στο ενοποιημένο σύνολο δεδομένων με τη χρήση SMOTE



Εικόνα 50: Πίνακας σύγχυσης του ενοποιημένου συνόλου δεδομένων του αλγορίθμου SVM με τη χρήση SMOTE



Εικόνα 51: Πίνακας σύγχυσης του ενοποιημένου συνόλου δεδομένων του αλγορίθμου Decision Tree με τη χρήση SMOTE

Βλέπουμε ότι η ακρίβεια της πρόβλεψης σε κάθε σύνολο δεδομένων έχει αυξηθεί αρκετά τόσο για τον αλγόριθμο SVM όσο και για τον αλγόριθμο Decision Tree (συγκριτικά χωρίς την εφαρμογή SMOTE). Επίσης, παρατηρούμε ότι στο σύνολο δεδομένων Gene Expression και στο ενοποιημένο σύνολο, ο αλγόριθμος SVM έχει πολύ λιγότερα λάθη σε σχέση με τον Decision Tree καθώς επίσης και ο αριθμός λαθών που προκύπτει είναι μικρότερος από τον αριθμό λαθών χωρίς την εφαρμογή του αλγορίθμου SMOTE.

Επιπλέον, παρατηρούμε ότι το σύνολο δεδομένων miRNA και στις δυο περιπτώσεις (με και χωρίς SMOTE) παρουσιάζει μικρό ποσοστό ακρίβειας συγκριτικά με τα υπόλοιπα σύνολα δεδομένων. Ακόμη, βλέπουμε ότι με τη χρήση ή όχι του αλγορίθμου SMOTE, ο αλγόριθμος SVM έχει λιγότερα λάθη από το Decision Tree, δηλαδή παρουσιάζει καλύτερες προβλέψεις στα δεδομένα του γλοιοβλαστώματος που διαχειριζόμαστε.

7. Αποτελέσματα της στρατηγικής Early integration σε άλλα καρκινικά δεδομένα

Δοκιμάσαμε την αποτελεσματικότητα της στρατηγικής early integration και σε άλλα καρκινικά δεδομένα πλην του γλοιοβλαστώματος. Πιο συγκεκριμένα, αντλήσαμε από το TCGA δεδομένα από τους ακόλουθους καρκίνους:

- **AML** (Οξεία μυελογενής λευχαιμία)
- **BIC** (Καρκίνος του στήθους)
- **COAD** (Αδενοκαρκίνωμα του παχέος εντέρου)
- **KIRC** (Συμβατικό νεφρικό καρκίνωμα)
- **LIHC** (Ηπατοκυτταρικό καρκίνωμα ήπατος)
- **LUSC** (ακανθοκυτταρικό καρκίνωμα του πνεύμονα)
- **SKCM** (Δερματικό μελάνωμα του δέρματος)
- **OV** (Καρκίνος των ωοθηκών)
- **SARC** (Σάρκωμα)

Στη συνέχεια ακολουθήσαμε σχεδόν τις ίδιες μεθοδολογίες που χρησιμοποιήσαμε για τον καρκίνο του γλοιοβλαστώματος. Έγινε επεξεργασία όλων των δεδομένων για να τα έχουμε στην μορφή που χρειάζεται και από το αρχείο με τα δεδομένα επιβίωσης του ασθενή κρατήσαμε την κατάσταση του ασθενή , αποθανών ή εν ζωή. Βρήκαμε τους κοινούς ασθενείς μεταξύ των ομικών συνόλων δεδομένων και του αρχείου επιβίωσης και στη συνέχεια πραγματοποιήθηκε ομαλοποίηση των δεδομένων στα νέα σύνολα δεδομένων που προέκυψαν. Εφαρμόζουμε την στρατηγική early integration και γίνεται εκπαίδευση των μοντέλων μηχανικής μάθησης βάση του χαρακτηριστικού της κατάστασης του ασθενή. Ακόμη, κάναμε χρήση του αλγορίθμου επιλογής χαρακτηριστικών (Feature Selection Algorithm) **ANOVA** (Analysis of Variance) για να έχουμε πιο ακριβής πρόβλεψη. Η αρχή λειτουργίας του είναι, η σύγκριση πολλών μέσων τιμών του συνόλου δεδομένων οπτικοποιώντας την όποια σημαντική διαφορά προκύπτει μεταξύ των μέσων τιμών πολλών ομάδων (classes). Ο εν λόγω αλγόριθμος χρησιμοποιήθηκε για την αφαίρεση των μη επιθυμητών και μη σχετικών χαρακτηριστικών από τα δεδομένα διατηρώντας μόνο τα πιο σχετικά [62].

Στο ακόλουθο στιγμιότυπο βλέπουμε το σύνολο δεδομένων Gene Expression που έχει προκύψει από το τη χρήση του αλγορίθμου ANOVA να έχει 290 γραμμές (ασθενείς) και 11 στήλες όπου η τελευταία φανερώνει την κατάσταση του κάθε ασθενή (εν ζωή ή αποθανών). Οι υπόλοιπες στήλες είναι τα χαρακτηριστικά που επιλέχθηκαν από τον αλγόριθμο.

	PPL.5493	CD40LG.959	PLA2G2D.26279	SCNN1A.6337	TAP1.6890	COMMD5.28991	CMTM4.146223	CXCL11.6373	WARS.7453	EMP1.2012	Dea
TCGA.04.1348.01	0.063338	0.202911	0.061020	0.369130	0.545719	0.119194	0.010535	0.218349	0.315023	0.028874	
TCGA.04.1362.01	0.361276	0.045787	0.001381	0.498019	0.088334	0.298812	0.198963	0.079441	0.097644	0.513713	
TCGA.04.1364.01	0.098487	0.000000	0.000000	0.117451	0.086145	0.254391	0.131393	0.004423	0.022478	0.000975	
TCGA.04.1365.01	0.127952	0.174340	0.067630	0.222170	0.460217	0.235073	0.078672	0.425727	0.181159	0.060271	
TCGA.04.1514.01	0.192057	0.007241	0.001441	0.156806	0.032505	0.004947	0.431469	0.006019	0.105310	0.100393	
...	...	...	...	...	...	...	...	...	...	...	
TCGA.61.2104.01	0.378977	0.165107	0.019833	0.211851	0.563748	0.199416	0.287254	0.438226	0.054800	0.108908	
TCGA.61.2109.01	0.203871	0.047974	0.000398	0.355129	0.282506	0.127815	0.409424	0.026860	0.056889	0.102089	
TCGA.61.2110.01	0.346331	0.095871	0.000397	0.467074	0.037244	0.013731	0.360306	0.000389	0.048827	0.058987	
TCGA.61.2111.01	0.289938	0.166133	0.236944	0.401700	0.118522	0.055764	0.120497	0.004777	0.108886	0.016465	
TCGA.61.2113.01	0.291320	0.105478	0.013994	0.276723	0.223346	0.191957	0.250409	0.182658	0.538927	0.301495	

290 rows × 11 columns

Εικόνα 52: Εφαρμογή του αλγορίθμου ANOVA στο σύνολο δεδομένων Gene Expression

Στην παρακάτω εικόνα παρουσιάζεται το ενοποιημένο σύνολο δεδομένων το οποίο αποτελείται αντίστοιχα από 290 γραμμές (κοινούς ασθενείς) και 31 στήλες καθώς έχουν συνενωθεί τα τρια ομικά σύνολα δεδομένων (Gene expression, DNA methylation, miRNA) σύμφωνα με την στρατηγική Early integration. Τέλος, η τελευταία στήλη αντιπροσωπεύει την κατάσταση του εκάστοτε ασθενή.

	PPL.5493	CD40LG.959	PLA2G2D.26279	SCNN1A.6337	TAP1.6890	COMMD5.28991	CMTM4.146223	CXCL11.6373	WARS.7453	EMP1.2012	...
TCGA.04.1348.01	0.063338	0.202911	0.061020	0.369130	0.545719	0.119194	0.010535	0.218349	0.315023	0.028874	...
TCGA.04.1362.01	0.361276	0.045787	0.001381	0.498019	0.088334	0.298812	0.198963	0.079441	0.097644	0.513713	...
TCGA.04.1364.01	0.098487	0.000000	0.000000	0.117451	0.086145	0.254391	0.131393	0.004423	0.022478	0.000975	...
TCGA.04.1365.01	0.127952	0.174340	0.067630	0.222170	0.460217	0.235073	0.078672	0.425727	0.181159	0.060271	...
TCGA.04.1514.01	0.192057	0.007241	0.001441	0.156806	0.032505	0.004947	0.431469	0.006019	0.105310	0.100393	...
...	...	...	...	...	...	...	...	...	...	...	...
TCGA.61.2104.01	0.378977	0.165107	0.019833	0.211851	0.563748	0.199416	0.287254	0.438226	0.054800	0.108908	...
TCGA.61.2109.01	0.203871	0.047974	0.000398	0.355129	0.282506	0.127815	0.409424	0.026860	0.056889	0.102089	...
TCGA.61.2110.01	0.346331	0.095871	0.000397	0.467074	0.037244	0.013731	0.360306	0.000389	0.048827	0.058987	...
TCGA.61.2111.01	0.289938	0.166133	0.236944	0.401700	0.118522	0.055764	0.120497	0.004777	0.108886	0.016465	...
TCGA.61.2113.01	0.291320	0.105478	0.013994	0.276723	0.223346	0.191957	0.250409	0.182658	0.538927	0.301495	...

290 rows × 31 columns

Εικόνα 53: Εφαρμογή του αλγορίθμου ANOVA στο ενοποιημένο σύνολο δεδομένων

Εν συνεχεία, εκπαιδεύτηκαν τα μοντέλα μηχανικής μάθησης πριν και μετά τη χρήση του παραπάνω αλγορίθμου. Παρακάτω παρουσιάζονται οι αποδόσεις των

μοντέλων SVM και Decision Tree στα καρκινικά ομικά δεδομένα ξεχωριστά και στο ενοποιημένο σύνολο που προκύπτει κάθε φορά από αυτά.

Αποτελέσματα πριν την χρήση του αλγορίθμου ANOVA

Cancer	SVM (Gene Expression)	Decision Tree (Gene Expression)	SVM (DNA Methylation)	Decision Tree (DNA Methylation)	SVM (miRNA)	Decision Tree (miRNA)	SVM (Ενοποιημένο σύνολο)	Decision Tree (Ενοποιημένο σύνολο)
LIHC	63.93%	60.65%	59.83%	54.91%	59.83%	61.47%	63.93%	54.91%
BIC	88.70%	82.25%	88.17%	76.34%	90.32%	79.56%	89.24%	80.64%
OV	49.42%	48.27%	58.62%	56.32%	56.32%	56.32%	55.17%	51.72%
SARC	59.49%	64.55%	63.29%	56.96%	55.69%	58.22%	64.55%	58.22%
KIRC	69.35%	61.29%	72.58%	70.96%	66.12%	64.51%	74.19%	54.83%
LUSC	47.05%	53.92%	52.94%	55.88%	51.96%	50%	49.01%	45.09%
COAD	75.38%	69.23%	72.30%	67.69%	73.84%	64.61%	72.30%	70.69%
AML	64.58%	39.58%	60.41%	50%	56.25%	58.33%	64.58%	54.16%
SKCM	58.33%	50.75%	54.54%	53.78%	53.03%	56.06%	56.06%	51.51%

Πίνακας 1: Αποτελέσματα πριν την χρήση του αλγορίθμου ANOVA

Cancer	SVM (Gene Expression)	Decision Tree (Gene Expression)	SVM (DNA Methylation)	Decision Tree (DNA Methylation)	SVM (miRNA)	Decision Tree (miRNA)	SVM (Ενοποιημένο σύνολο)	Decision Tree (Ενοποιημένο σύνολο)
LIHC	**72.95%**	**67.21%**	**62.29%**	52.45%	**71.31%**	54.91%	**70.49%**	52.45%
BIC	**91.93%**	**83.87%**	**91.93%**	**78.49%**	**91.93%**	**80.10%**	**91.93%**	**84.40%**
OV	**59.77%**	**55.17%**	**68.96%**	**62.06%**	**66.66%**	**64.36%**	**74.71%**	**66.66%**
SARC	**63.29%**	**65.82%**	**65.82%**	53.16%	**67.08%**	**62.02%**	62.02%	**62.02%**
KIRC	**79.03%**	**72.58%**	**82.25%**	58.06%	**85.48%**	**66.12%**	**82.25%**	**75.80%**
LUSC	**63.72%**	**61.76%**	**60.78%**	55.88%	**60.78%**	**50.98%**	**61.76%**	**52.94%**
COAD	**76.92%**	66.15%	**75.38%**	**72.30%**	**78.46%**	63.07%	**76.92%**	56.92%
AML	**75%**	**68.75%**	**81.25%**	**75%**	**72.91%**	**62.5%**	**77.08%**	**60.41%**
SKCM	**67.42%**	46.96%	**61.36%**	**55.30 %**	**61.36%**	51.51%	**66.66%**	**61.36%**

Πίνακας 2: Αποτελέσματα μετά την χρήση του αλγορίθμου ANOVA

Παρατηρούμε ότι στα περισσότερα ομικά επίπεδα και σχεδόν σε όλα τα ενοποιημένα σύνολα, ότι η ακρίβεια πρόβλεψης των μοντέλων έχει βελτιωθεί ακόμη περισσότερο χρησιμοποιώντας τον αλγόριθμο ANOVA.

Έτσι μπορούμε να συμπεράνουμε ότι η με την στρατηγική Early integration πράγματι έχουμε καλύτερη πρόβλεψη στην κατηγοριοποίηση των δειγμάτων συγκριτικά με την μεμονωμένη ανάλυση κάθε ομικού επιπέδου.

8. Συμπεράσματα

Το γλοιοβλάστωμα αποτελεί τον πιο θανατηφόρο καρκίνο του εγκεφάλου (4ο στάδιο) με πολύ μικρά ποσοστά επιβίωσης. Στα δείγματα των ασθενών που αναλύσαμε από το TCGA επιλέξαμε τους ασθενείς που έζησαν πάνω από 100 ημέρες και τους ταξινομήσαμε βάση την κατάσταση LIVING (εν ζωή) και DECEASED (αποθανών). Στη συνέχεια, με την χρήση αλγορίθμων μηχανικής μάθησης, συγκεκριμένα τους Decision Tree και SVM και τη χρήση της στρατηγικής Early integration προχωρήσαμε στην εκπαίδευση των μοντέλων χρησιμοποιώντας τα ομικά δεδομένα Gene Expression, DNA Methylation, miRNA και το ενοποιημένο σύνολο.

Καταλήξαμε ότι ο αλγόριθμος SVM μας έδωσε (σχεδόν κάθε φορά) καλύτερα αποτελέσματα από τον Decision Tree. Να σημειωθεί ότι τα μοντέλα χρησιμοποιήθηκαν με και χωρίς τη χρήση του αλγορίθμου SMOTE καθώς η κλάση LIVING δημιουργούσε ένα ανισόρροπο σύνολο δεδομένων.

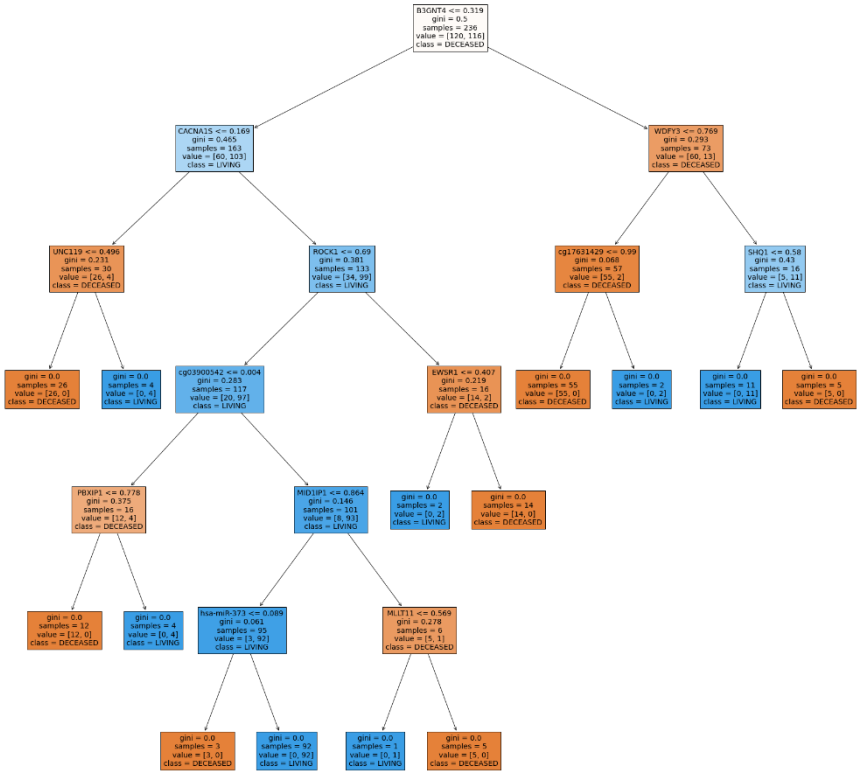
Στο σύνολο δεδομένων Gene expression ο αλγόριθμος SVM είχε πολύ καλύτερη κατηγοριοποίηση συγκριτικά με τα αποτελέσματα του Decision Tree όπου εμφάνισε μεγαλύτερο αριθμό λαθών επιτυγχάνοντας 90.19% και 71.57% αντίστοιχα.

Στο σύνολο δεδομένων DNA Methylation ο αλγόριθμος SVM έδωσε και εδώ καλύτερη κατηγοριοποίηση στα δεδομένα που διαχειριζόμαστε συγκριτικά με τον Decision Tree. Η ακρίβεια της πρόβλεψης είναι 88.23% και 68.62% αντίστοιχα. Παρατηρήθηκε επίσης ότι, ο αλγόριθμος Decision Tree δεν έκανε τόσο καλή πρόβλεψη της κλάσης LIVING παρόλο που έχει εφαρμοστεί στα δεδομένα ο αλγόριθμος SMOTE.

Στο σύνολο δεδομένων miRNA οι δύο αλγόριθμοι είχαν ίδια ακρίβεια πρόβλεψης 73.52% αλλά ο αλγόριθμος Decision Tree σημείωσε και εδώ αρκετά μεγαλύτερο αριθμό λανθασμένων προβλέψεων για την κλάση LIVING. Και τα δυο μοντέλα παρουσίασαν επίσης μεγάλο αριθμό λαθών και για την κλάση DECEASED συγκριτικά με τα υπόλοιπα σύνολα δεδομένων.

Τέλος, για το ενοποιημένο σύνολο δεδομένων που προκύπτει από την εφαρμογή της στρατηγικής early integration, επιτεύχθηκε καλύτερη κατηγοριοποίηση των δεδομένων με ακρίβεια 88.23% με την εφαρμογή του SVM και 72.54% με τη χρήση του Decision Tree.

Από το ενοποιημένο σύνολο δεδομένων με τη χρήση Decision Tree και του αλγορίθμου SMOTE προέκυψε το παρακάτω δέντρο απόφασης.



Εικόνα 54: Δέντρο απόφασης ενοποιημένου συνόλου δεδομένων

Από το παραπάνω δέντρο απόφασης διαπιστώνουμε ότι περιλαμβάνονται χαρακτηριστικά (features) από κάθε ομικό επίπεδο με την πλειοψηφία αυτών να είναι εκφράσεις γονιδίων. Παρακάτω, θα δούμε πως μερικά από τα εν λόγω χαρακτηριστικά σχετίζονται με την ασθένεια του γλοιοβλαστώματος. Πιο συγκεκριμένα, θα δούμε πως τα γονίδια MLLT11, PBXIP1, ROCK1 και το miRNA miRNA-373 σχετίζονται με το γλοιοβλάστωμα.

- **MLLT11**: Το MLLT11 κατέχει καθοριστικό ρόλο στην εξέλιξη του γλοιώματος και έχει την μπορεί να αποτελέσει νέο προγνωστικό δείκτη για το γλοίωμα. Το εν λόγω γονίδιο εντοπίζεται στο χρωμόσωμα 1q21 και κωδικοποιεί μια πρωτεΐνη που αποτελείται από 270 αμινοξέα. Η έκφραση του MLLT11 αυξήθηκε σε ποικίλες αιματολογικές διαταραχές όπως την λεμφοκυτταρική λευχαιμία. Σήμερα, έχει φανεί ότι υπάρχει συσχέτιση ανάμεσα στο επίπεδο έκφρασης του γονιδίου και στην κυτταρική διαφοροποίηση. Πιο συγκεκριμένα, στα γλοιώματα υπήρξαν ευδιάκριτες

διαφορές στην έκφραση του MLLT11 ανάμεσα σε διαφορετικούς βαθμούς γλοιωμάτων. Η έκφραση του γονιδίου βρέθηκε μειωμένη σε γλοιώματα υψηλού βαθμού και υψηλή στους υποτύπους προνευρικό και νευρικό. Ακόμη, μελέτες έδειξαν ότι το MLLT11, έχει σημαντικό ρόλο στην νευρωνική διαφοροποίηση και συντήρηση κατά την διάρκεια της εμβρυικής ανάπτυξης [63].

- **PBXIP1**: Η πρωτεΐνη που παράγεται από το PBXIP1 έχει βρεθεί υπερεκφρασμένη στο αστροκύτωμα , στο γλοιοβλάστωμα και σε ιστούς επενδυμώματος έδειξε ότι αποτελεί ένα νέο δείκτη προγονικών κυττάρων αστροκυττάρων κατά τη διάρκεια της ανάπτυξης του εγκεφάλου φανερώνοντας στοιχεία ότι οδηγεί την μετανάστευση και τον πολλαπλασιασμό των κυττάρων γλοιοβλαστώματος. Με λίγα λόγια το PBXIP1 είναι σημαντικό για την ανάπτυξη των κυττάρων γλοιώματος και αποτελεί πιθανό υποψήφιο στόχο για γονιδιακή αντικαρκινική θεραπεία [64].

- **ROCK1**: Το ROCK1 επηρεάζει την κυτταρική εισβολή και μετανάστευση αλλάζοντας την κατάσταση του κυτταροσκελετού. Τα τελευταία χρόνια, έχει βρεθεί υπερεκφρασμένο σε διαφορετικά είδη καρκίνου. Σε τελευταίες μελέτες φαίνεται ότι το ROCK1 εμφανίστηκε υπερεκφρασμένο σε πολλούς όγκους, συμπεριλαμβανομένων των, καρκίνο του πνεύμονα, καρκίνο του προστάτη, του γαστρικού καρκίνου κ.α. Η υπερέκφραση αυτή συσχετίζεται με την μετάσταση και την πρόγνωση του αντίστοιχου καρκίνου. Το θετικό ποσοστό έκφρασης (positive expression rate) του ROCK1 mRNA στο γλοίωμα ήταν σημαντικά υψηλότερο από τον παρακείμενο φυσιολογικό ιστό. Έτσι σύμφωνα με το ρόλο του ROCK1 mRNA σε άλλους καρκινικούς ιστούς και την υψηλή έκφραση του στο γλοίωμα, πιθανόν έχει σημαντικό ρόλο στην καρκινογένεση και εξέλιξη του καρκίνου. Τέλος, το θετικό ποσοστό έκφρασης του ROCK1 mRNA σε υψηλού βαθμού κακοήθη γλοιώματα εμφανίστηκε εντυπωσιακά υψηλότερο από ότι σε μικρού βαθμού κακοήθη γλοιώματα δείχνοντας έτσι ότι η υψηλή αυτή έκφραση του ROCK1 συνδέεται με το επίπεδο της κακοήθειας του γλοιώματος [65].

- **miRNA-373**: Τα microRNA (miRNA) εμπλέκονται σχεδόν σε κάθε κυτταρική διαδικασία όπως ο κυτταρικός πολλαπλασιασμός η απόπτωση, η γήρανση, η διαφοροποίηση κ.α. Απορυθμίσεις των miRNA έχουν παρατηρηθεί σε πολλές ασθένειες, ιδίως στον καρκίνο όπου λειτουργούν ως ογκογονίδια ή ογκοκατασταλτικά [66]. Σημαντικές διαφορές έχουν παρατηρηθεί στην έκφραση κάποιων miRNA ανάμεσα σε καρκινικούς ιστούς γλοιοβλαστώματος και φυσιολογικούς ιστούς εγκεφάλου, γεγονός που κάνει τα εν λόγω microRNA εν δυνάμει στόχους για τη διάγνωση και θεραπεία του γλοιοβλαστώματος. Ειδικότερα, το miRNA-373, έχει αναφερθεί ότι αναστέλλει την ικανότητα μετανάστευσης και εισβολής των κυττάρων γλοιοβλαστώματος με το να υπορρυθμίζει την έκφραση των γονιδίων που κωδικοποιούν το CD44 και μετασχηματίζοντας τον αυξητικό παράγοντα βήτα υποδοχέα δυο (TGFBR2). Αξίζει να σημειωθεί ότι το παρατηρήθηκε μεγαλύτερη έκφραση του γονιδίου CD44 στους ιστούς γλοιοβλαστώματος και συσχετίστηκε με τη πρόγνωση των ασθενών του καρκίνου αυτού. Τέλος, βρέθηκε ότι το η έκφραση των miRNA-373 και miRNA-520 ενδεχομένως θα μπορούσαν να αποτελέσουν πολύτιμους βιοδείκτες στη διάγνωση και πρόγνωση του γλοιοβλαστώματος καθώς επίσης κ το ότι αυτά τα miRNA καταστέλλουν την έκφραση του CD44 σε κύτταρα γλοιώματος [67].

Αξίζει να σημειώσουμε στο σημείο αυτό ότι η πολύ-ομική ανάλυση δεδομένων μπορεί να καταστεί δαπανηρή όπως και η πρόσβαση σε πολλά αναλυτικά όργανα και εξειδικευμένο επιστημονικό προσωπικό απαιτεί υψηλή χρηματοδότηση της έρευνας. Για παράδειγμα, μια σημαντική μελέτη πάνω σε ανθρώπινο ερευνητικό πληθυσμό μπορεί να ξεπεράσει σε κόστος τα 500.000 δολάρια. Η έλλειψη επαρκούς χρηματοδότησης σε πολύ-ομικές μελέτες, αποτελεί συχνά περιοριστικό παράγοντα [68]. Με την μείωση του χρόνου και του κόστους που απαιτείται για την παραγωγή ομικών συνόλων δεδομένων, η ενοποίηση των ομικών δεδομένων έχει δημιουργήσει μεγάλες ευκαιρίες και προκλήσεις για τους επιστήμονες στους χώρους της βιολογίας, βιοστατιστικής και βιομαθηματικών [69].

9. Μελλοντικές επεκτάσεις

Πιθανές μελλοντικές επεκτάσεις της πτυχιακής εργασίας θα μπορούσε να ήταν η επιλογή διαφορετικού χαρακτηριστικού για την κατηγοριοποίηση των ασθενών από το κλινικό αρχείο, αντί της κατηγοριοποίησης που χρησιμοποιήσαμε στην παρούσα φάση επιλέγοντας τους ασθενείς όπου έζησαν πάνω από 100 ημέρες κι τη ζωτική τους κατάσταση, αν ήταν αποθανών ή εν ζωή. Επίσης, θα ήταν ενδιαφέρον αντί τους αλγορίθμους μηχανικής μάθησης SVM και Decision Tree, να κατασκευαστεί ένα μοντέλο νευρωνικού δικτύου ώστε να δούμε αν αυτό παρουσιάζει μεγαλύτερη ακρίβεια πρόβλεψης και νέα αποτελέσματα στα δεδομένα γλοιοβλαστώματος που διαχειριστήκαμε.

Βιβλιογραφία

[1] Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063 (Ανακτήθηκε 13/4/2022)

[2] Bayat A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ (Clinical research ed.)*, *324*(7344), 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018 (Ανακτήθηκε 13/4/2022)

[3] Bo, L., Wei, B., Li, C., Wang, Z., Gao, Z., & Miao, Z. (2017). Identification of potential key genes associated with glioblastoma based on the gene expression profile. Oncology letters, 14(2), 2045–2052. https://doi.org/10.3892/ol.2017.6460

[4] YIN W., TANG G., ZHOU Q., CAO Y., LI H., FU X., WU Z AND JIANG X., (2019). Expression Profile Analysis Identifies a Novel Five-Gene Signature to Improve Prognosis Prediction of Glioblastoma. Frontiers, Frontiers in Genetics | Computational Genomics. https://www.frontiersin.org/article/10.3389/fgene.2019.00419

[5] Rich, J. N., Hans, C., Jones, B., Iversen, E. S., McLendon, R. E., Rasheed, B. K., Dobra, A., Dressman, H. K., Bigner, D. D., Nevins, J. R., & West, M. (2005). Gene expression profiling and genetic markers in glioblastoma survival. Cancer research, 65(10), 4051–4058. https://doi.org/10.1158/0008-5472.CAN-04-3936

[6] Alifieris, C., & Trafalis, D. T. (2015). Glioblastoma multiforme: Pathogenesis and treatment. Pharmacology & therapeutics, 152, 63–82. https://doi.org/10.1016/j.pharmthera.2015.05.005

[7] Hanif, F., Muzaffar, K., Perveen, K., Malhi, S. M., & Simjee, S. (2017). Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. Asian Pacific journal of cancer prevention : APJCP, 18(1), 3–9. https://doi.org/10.22034/APJCP.2017.18.1.3

[8] Nature Scitable | "Introduction: What is DNA?". Ανακτήθηκε 13/4/2022 από: https://www.nature.com/scitable/topicpage/introduction-what-is-dna-6579978/

[9] Nature Scitable | "DNA Is a Structure That Encodes Biological Information", Ανακτήθηκε 13/4/2022 από: https://www.nature.com/scitable/topicpage/DNA-Is-a-Structure-that-Encodes-Information-6493050/

[10] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US); [updated 2020 Jun 24]. What is DNA?; [updated 2020 Jun 18; reviewed 2018

Jun 01; cited 2022 Apr 13]; [about 1 p.]. Ανακτήθηκε 13/04/2022 από: https://medlineplus.gov/genetics/understanding/basics/dna/

[11] Moore, L., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **38**, 23-38 (2013). https://doi.org/10.1038/npp.2012.112

[12] Phillips, T. (2008) The role of methylation in gene expression. Nature Education 1(1):116. Ανακτήθηκε 30/4/2022 από https://www.nature.com/scitable/topicpage/the-role-of-methylation-in-gene-expression-1070/

[13] Kim, M., Costello, J. DNA methylation: an epigenetic mark of cellular memory. *Exp Mol Med* **49**, e322 (2017). https://doi.org/10.1038/emm.2017.10

[14] Antequera, F., Bird, A. (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Current Biology*, Volume 9, Issue 17, R661-R667. https://doi.org/10.1016/S0960-9822(99)80418-7.

[15] Jin, B., Li, Y., & Robertson, K. D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy?. Genes & cancer, 2(6), 607–617. https://doi.org/10.1177/1947601910393957

[16] Kulis, M., Esteller, M. (2010) DNA Methylation and Cancer. *Advances in Genetics*, Volume 70, 27-55. https://doi.org/10.1016/B978-0-12-380866-0.60002-2

[17] Das, P. M., & Singal, R. (2004). DNA methylation and cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 22(22), 4632–4642. https://doi.org/10.1200/JCO.2004.07.151

[18] Ehrlich M. (2009). DNA hypomethylation in cancer cells. Epigenomics, 1(2), 239–259. https://doi.org/10.2217/epi.09.33

[19] National Human Genome Research Institute. Ribonucleic Acid (RNA). Ανακτήθηκε 14/4/2022 από: https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid

[20] Chatterjee, K. and Wan, . Yao (2022, March 17). RNA. Encyclopedia Britannica. https://www.britannica.com/science/RNA

[21] National Human Genome Research Institute. Messenger RNA (mRNA). Ανακτήθηκε 14/4/2022 από: https://www.genome.gov/genetics-glossary/messenger-rna

[22] Kapur, M., Ackerman, S. (2018). mRNA Translation Gone Awry: Translation Fidelity and Neurological Disease, *Trends in Genetics,* Volume 34, Issue 3, 218-231. https://doi.org/10.1016/j.tig.2017.12.007

[23] Rausch, S., Schwentner, C., Stenzl, A., & Bedke, J. (2014). mRNA vaccine CV9103 and CV9104 for the treatment of prostate cancer. Human vaccines & immunotherapeutics, 10(11), 3146–3152. https://doi.org/10.4161/hv.29553

[24] Pal, I., Safari, M., Jovanovic, M., Bates, S. E., & Deng, C. (2019). Targeting Translation of mRNA as a Therapeutic Strategy in Cancer. Current hematologic malignancy reports, 14(4), 219–227. https://doi.org/10.1007/s11899-019-00530-y

[25] Bidram, M., Zhao, Y., Shebardina, N. G., Baldin, A. V., Bazhin, A. V., Ganjalikhany, M. R., Zamyatnin, A. A., Jr, & Ganjalikhani-Hakemi, M. (2021). mRNA-Based Cancer Vaccines: A Therapeutic Strategy for the Treatment of Melanoma Patients. Vaccines, 9(10), 1060. https://doi.org/10.3390/vaccines9101060

[26] Ardekani, A. M., & Naeini, M. M. (2010). The Role of MicroRNAs in Human Diseases. Avicenna journal of medical biotechnology, 2(4), 161–179. Ανακτήθηκε 5/4/2022 από: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3558168/

[27] Smirnova, L., Gräfe, A., Seiler, A., Schumacher, S., Nitsch, R., & Wulczyn, F. G. (2005). Regulation of miRNA expression during neural cell specification. The European journal of neuroscience, 21(6), 1469–1477. https://doi.org/10.1111/j.1460-9568.2005.03978.x

[28] Ji, B. Y., You, Z. H., Cheng, L., Zhou, J. R., Alghazzawi, D., & Li, L. P. (2020). Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. Scientific reports, 10(1), 6658. https://doi.org/10.1038/s41598-020-63735-9

[29] Di Leva, G., & Croce, C. M. (2013). miRNA profiling of cancer. Current opinion in genetics & development, 23(1), 3–11. https://doi.org/10.1016/j.gde.2013.01.004

[30] Chen, M., Medarova, Z., & Moore, A. (2021). Role of microRNAs in glioblastoma. Oncotarget, 12(17), 1707–1723. https://doi.org/10.18632/oncotarget.28039

[31] Buruiană, A., Florian, Ș. I., Florian, A. I., Timiș, T. L., Mihu, C. M., Miclăuș, M., Oșan, S., Hrapșa, I., Cataniciu, R. C., Farcaș, M., & Șușman, S. (2020). The Roles of miRNA in Glioblastoma Tumor Cell Communication: Diplomatic and

Aggressive Negotiations. International journal of molecular sciences, 21(6), 1950. https://doi.org/10.3390/ijms21061950

[32] Novakova, J., Slaby, O., Vyzula, R., & Michalek, J. (2009). MicroRNA involvement in glioblastoma pathogenesis. Biochemical and biophysical research communications, 386(1), 1–5. https://doi.org/10.1016/j.bbrc.2009.06.034

[33] Nature Scitable. Genome. Ανακτήθηκε 14/4/2022 από: https://www.nature.com/scitable/definition/genome-43

[34] National Human Genome Research Institute. A Brief Guide to Genomics. Ανακτήθηκε 14/4/2022 από: https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

[35] Nature Scitable. Gene Expression. Ανακτήθηκε 14/4/2022 από: https://www.nature.com/scitable/topicpage/gene-expression-14121669/

[36] Nature Scitable. Microarray. Ανακτήθηκε 14/4/2022 από: https://www.nature.com/scitable/definition/microarray-202/

[37] Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. Cold Spring Harbor protocols, 2015(11), 951–969. https://doi.org/10.1101/pdb.top084970

[38] Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A Primer for the epidemiologist. OUP Academic. Ανακτήθηκε 15/4/2022 από: https://academic.oup.com/aje/article/188/12/2222/5567515

[39] Mahesh, B. (2018). Machine learning algorithms - a review Ανακτήθηκε 15/4/2022, από: https://www.researchgate.net/profile/Batta Mahesh/publication/344717762_Machine_Learning_Algorithms_-A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf

[40] Mohri, M., Rostamizadeh, A., Talwalkar, A,. *Foundations of Machine Learning.* Ανακτήθηκε 15/4/2022 από: https://books.google.gr/books?id=maz6AQAAQBAJ&printsec=frontcover#v=onepage&q&f=false

[41] Harrington, P. *Machine Learning in action.* Ανακτήθηκε 15/4/2022 από: https://books.google.gr/books/about/Machine_Learning_in_Action.html?id=XTozEAAAQBAJ&printsec=frontcover&source=kp_read_button&hl=en&redir_esc=y#v=onepage&q&f=false

[42] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

[43] Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, *19*, 3735-3746. Ανακτήθηκε 8/5/2022 από: https://www.sciencedirect.com/science/article/pii/S2001037021002683

[44] Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, *18*(1), 1-15. Ανακτήθηκε 8/5/2022 από: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1215-1

[45] Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, *14*, 1177932219899051. Ανακτήθηκε 8/5/2022 από: https://journals.sagepub.com/doi/full/10.1177/1177932219899051

[46] Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, *19*, 3735-3746. Ανακτήθηκε 8/5/2022 από: https://www.sciencedirect.com/science/article/pii/S2001037021002683

[47] What's in the Cancer Genome Atlas? Ανακτήθηκε 3/5/2022 από: https://www.sevenbridges.com/tcga/

[48] Matplotlib Release 2.0.2 Documentation ανακτήθηκε 23/4/2022 από: https://matplotlib.org/2.0.2/Matplotlib.pdf

[49] NumPy User Guide Release 1.23.0 Ανακτήθηκε 23/4/2022 από: https://numpy.org/doc/stable/numpy-user.pdf

[50] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830. Ανακτήθηκε 23/4/2022 από: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

[51] Package overview. Ανακτήθηκε 23/4/2022 από https://pandas.pydata.org/pandas-docs/stable/getting_started/overview.html

[52] Google Developers, Normalization. Ανακτήθηκε 23/4/2022 από: https://developers.google.com/machine-learning/data-prep/transform/normalization

[53] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. Ανακτήθηκε 27/4/2022 από: https://www.jair.org/index.php/jair/article/view/10302

[54] SMOTE for Imbalanced Classification with Python. Ανακτήθηκε 27/4/2022 από: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[55] Google Developers, Training and Test Sets: Splitting Data. Ανακτήθηκε 27/4/2022 από: https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

[56] Google Developer, Overfitting. Ανακτήθηκε 27/4/2022 από: https://developers.google.com/machine-learning/glossary#overfitting

[57] Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*, *7*(1), 86-112. Ανακτήθηκε 16/6/2022 από: https://academic.oup.com/bib/article/7/1/86/264025?login=false

[58] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), 1-16. Ανακτήθηκε 28/4/2022 από: https://link.springer.com/article/10.1186/s12911-019-1004-8

[59] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138. Ανακτήθηκε 28/4/2022 από: https://www.researchgate.net/profile/J-E-T-Akinsola/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison/links/596481dd0f7e9b819497e265/Supervised-Machine-Learning-Algorithms-Classification-and-Comparison.pdf

[60] Scikit-Learn, 1.4 Support Vector Machines. Ανακτήθηκε 28/4/2022 από: https://scikit-learn.org/stable/modules/svm.html#support-vector-machines

[61] W3schools, Machine Learning - Confusion Matrix. Ανάκτηση 28/4/2022 από: https://www.w3schools.com/python/python_ml_confusion_matrix.asp

[62] Kumar, M., Rath, N. K., Swain, A., & Rath, S. K. (2015). Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor. Procedia Computer Science, 54, 301-310. https://doi.org/10.1016/j.procs.2015.06.035

[63] Li, X., Chen, L., Xiong, Z., Zhao, H., Teng, C., Liu, H., ... & Wanggou, S. (2022). Identification of novel prognostic biomarker, MLLT11, reveals its relationship to immune checkpoint markers in glioma. Frontiers in Oncology, 4122. https://doi.org/10.3389/fonc.2022.889351

[64] van Vuurden, D. G., Aronica, E., Hulleman, E., Wedekind, L. E., Biesmans, D., Malekzadeh, A., ... & Van Der Stoop, P. P. (2014). Pre-B-cell leukemia homeobox interacting protein 1 is overexpressed in astrocytoma and promotes tumor cell growth and migration. Neuro-oncology, 16(7), 946-959. https://doi.org/10.1093/neuonc/not308

[65] Zhang, P., Lu, Y., Liu, X. Y., & Zhou, Y. H. (2015). Knockdown of Rho-associated protein kinase 1 suppresses proliferation and invasion of glioma cells. Tumor Biology, 36(1), 421-428. https://doi.org/10.1007/s13277-014-2673-7

[66] Wei, F., Wang, Q., Su, Q., Huang, H., Luan, J., Xu, X., & Wang, J. (2016). miR-373 inhibits glioma cell U251 migration and invasion by down-regulating CD44 and TGFBR2. Cellular and molecular neurobiology, 36(8), 1389-1397. https://doi.org/10.1007/s10571-016-0338-3

[67] Feng, S., Wang, K., Shao, Z., Lin, Q., Li, B., & Liu, P. (2022). MiR-373/miR-520s-CD44 Axis Significantly Inhibits the Growth and Invasion of Human Glioblastoma Cells. Archives of Medical Research, 53(6), 550-561. https://doi.org/10.1016/j.arcmed.2022.08.003

[68] Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., & Wishart, D. (2019). Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites, 9(4), 76. https://doi.org/10.3390/metabo9040076

[69] Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. Journal of molecular endocrinology, 62(1), R21-R45. https://doi.org/10.3390/metabo9040076