ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μείωση διάστασης για την οπτικοποίηση γονιδιακών δεδομένων μεγάλου όγκου

Δάλλας Ιωάννης

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Τασουλής Σωτήριος
Επίκουρος Καθηγητής

Λαμία, 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μείωση διάστασης για την οπτικοποίηση γονιδιακών δεδομένων μεγάλου όγκου

Δάλλας Ιωάννης

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Τασουλής Σωτήριος
Επίκουρος Καθηγητής

Λαμία, 2021

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Μείωση Διάστασης για την οπτικοποίηση γονιδιακών δεδομένων μεγάλου όγκου

Δάλλας Ιωάννης

Τριμελής Επιτροπή:

Τασουλής Σωτήριος, Επίκουρος Καθηγητής (επιβλέπων)

Πλαγιανάκος Βασίλειος, Καθηγητής

Μπάγκος Παντελεήμων, Καθηγητής

Ευχαριστίες

Η παρούσα εργασία αποτελεί πτυχιακή εργασία για τις σπουδές μου στο τμήμα Πληροφορικής με εφαρμογές στην Βιοϊατρική του Πανεπιστημίου Θεσσαλίας. Θα ήθελα στην παρούσα ενότητα να εκφράσω τις ειλικρινείς μου ευχαριστίες στον επιβλέπων καθηγητή κύριο Τασουλή Σωτήριο για την ευκαιρία που μου έδωσε να συνεργαστούμε, για την εμπιστοσύνη και την συνολική καθοδήγηση που παρείχε στην εκπόνηση της πτυχιακής μου εργασίας. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου για την ηθική συμπαράσταση, την ενθάρρυνση και την οικονομική υποστήριξη που παρείχε για όλο τα χρόνια των σπουδών μου.

Πίνακας περιεχομένων

Περίληψη

Η ανάπτυξη σύγχρονων μεθόδων ποσοτικοποίησης της γονιδιακής έκφρασης έχει οδηγήσει στην διαδικασία ανάπτυξης τεχνικών με τις οποίες καθίσταται δυνατό να αναλυθούν και εξαχθούν σημαντικές πληροφορίες για τις επιμέρους ιδιότητες ποικίλων βιολογικών διεργασιών. Η συγκέντρωση βιολογικών και ιατρικών δεδομένων περιλαμβάνει ένα τεράστιο σύνολο χαρακτηριστικών και ιδιοτήτων, διογκώνοντας το εύρος διαστάσεων, με αποτέλεσμα την δημιουργία προβλημάτων στις τεχνικές ανάλυσης και στον τρόπο με τον οποίο ερευνούμε την ροή των δεδομένων στον χώρο. Η παρούσα πτυχιακή εργασία πραγματεύεται τον ρόλο της μείωσης διάστασης σε δεδομένα μεγάλης κλίμακας και την οπτικοποίηση τους σε δισδιάστατη αναπαράσταση. Με την ταυτόχρονη παρουσίαση και εφαρμογή τεχνικών μείωσης διάστασης και οπτικοποίησης γίνεται αξιολόγηση των αλγοριθμικών μεθόδων για την επικύρωση της συνάφειας των αποτελεσμάτων. Η τελική αποτίμηση της σημασίας της μείωσης διάστασης δεδομένων μεγάλου όγκου αφορά την διερευνητική ανάλυση απλοϊκών μεθόδων κατηγοριοποίησης, με σκοπό να αναδείξει την εγκυρότητα των υποθέσεων και των συμπερασμάτων στα οποία προβήκαμε απ' την διαδικασία οπτικοποίησης.

Λέξεις – Κλειδιά : Μείωση διάστασης, οπτικοποίηση, αλγόριθμοι, βιολογικά δεδομένα, κατηγοριοποίηση, αξιολόγηση.

ABSTRACT

Recent sequencing techniques have developed new methods about the quantification of gene expression profiles. Biological and biomedical datasets are characterized by the vast amount of data, in which the set of attributes in a mathematical point of view are perceived as dimensions. The vast amount of dimensions are raising difficulties in the matter of data analysis and reducing algorithmic efficiency. Dimensionality reduction poses to be a powerful tool as it can summarize the high dimensionality into an intrinsic subspace. In addition, the visualization of high dimensional datasets in a two dimensional representation can unleash important insights of our data, helping us explore and capture the total variance with less possible information loss. The paper presents a set of dimensionality reduction algorithms, with which attempt to explore and analyze genomic data for two separate datasets. Our thesis concludes by evaluating our results using internal clustering criteria and attempting to validate the analysis and the interpretation of visualization representations using classification techniques.

Keywords: high dimensionality, algorithms, visualization, dimensionality reduction, evaluation, classification, genomic data, variance

Λίστα Πινάκων

Λίστα Εικόνων

Κεφάλαιο 1ο

Εισαγωγή

Από τις αρχές του 2000 και ύστερα, η ραγδαία εξέλιξη της τεχνολογίας, τόσο σε επίπεδο λογισμικού (*software*), αλλά και σε επίπεδο υλικού (*hardware*), ενθάρρυνε την ανάπτυξη του τομέα της τεχνητής νοημοσύνης και των μεθόδων της. Η ιδέα πίσω από την τεχνητή νοημοσύνη δεν είναι καινούργια, στην πραγματικότητα για να καταλήξουμε στην σημερινή εξελιγμένη και εφαρμοσμένη μορφή της χρειάστηκαν αρκετά στάδια, έως ότου ο *John McCarthy* εισάγει τον τομέα με τον όρο «*επιστήμη και μεθοδολογία της δημιουργίας νοημόνων μηχανών*» [1]. Η τεχνητή νοημοσύνη αποτελεί σημείο τομής πολλών επιστημών (όπως των μαθηματικών, της πληροφορικής κ.α.) με στοιχεία συλλογιστικής, αυτόματης μάθησης και προσαρμογής στα περιβαλλοντικά ερεθίσματα [1]. Στόχος είναι η ανάπτυξη ορθολογικών συστημάτων που «σκέφτονται και ενεργούν» σαν τον άνθρωπο [1], καταφέρνοντας να μιμηθεί πολυσύνθετες λειτουργίες για την επίλυση σύνθετων προβλημάτων με υψηλή ακρίβεια και απόδοση σε σύντομο χρονικό διάστημα. Η εξέλιξη και ανάπτυξη της τεχνητής νοημοσύνης δεν έγινε από την μία μέρα στην άλλη, αλλά απαιτήθηκε μία μακροχρόνια μετάβαση από τις θεμελιώδεις θεωρητικές προσεγγίσεις, που έχουν τις ρίζες τους στο 1943, έως την σημερινή εποχή [1]. Όπως προαναφέρθηκε, η τεχνητή νοημοσύνη οφείλει την δημιουργία της στην συνδρομή πολλών επιστημών ποικίλων τομέων των οποίων η εξέλιξη όλα αυτά τα χρόνια συνεπάγεται και ανάπτυξη των τεχνικών της τεχνητής νοημοσύνης. Οι συνεχείς προκλήσεις που παρουσιάζονται στην σημερινή εποχή συνοδεύονται από την υψηλή συσσώρευση δεδομένων μεγάλης κλίμακας ανεξαρτήτως του τομέα τον οποίο ερευνούμε. Η σύγχρονη εποχή χαρακτηρίζεται από αυτό το εκθετικό ρυθμό συγκέντρωσης δεδομένων (*big data*), δημιουργώντας ωστόσο εμπόδια στο τρόπο με το οποίο τα αναλύουμε και περιπλέκοντας την υπολογιστική και μαθηματική πολυπλοκότητα των τεχνικών της τεχνητής νοημοσύνης [2]. Η υψηλή διαστατικότητα από την οποία χαρακτηρίζονται τα δεδομένα απαιτούν την ανάπτυξη μεθόδων και αλγορίθμων, με τις οποίες καθίσταται δυνατό να συρρικνωθεί ο τεράστιος όγκος πληροφορίας, χωρίς ωστόσο να χαθεί το μεγαλύτερο ποσοστό της.

1.1 Είδη τεχνητής νοημοσύνης

Οι εφαρμογές της τεχνητής νοημοσύνης βασιζόμενες στην εμπειρία και στην μάθηση επιτρέπουν την εξαγωγή πολύπλοκων μοτίβων και χαρακτηριστικών πάνω στα δεδομένα (*data mining*), την αυτοματοποίηση διεργασιών και ανάπτυξη λογισμικών λήψης αποφάσεων. Ο τρόπος με τον οποίο εκπαιδεύεται και αναπτύσσεται μία εφαρμογή περιλαμβάνει διάφορα θεωρητικά υπόβαθρα, τα οποία ποικίλλουν και άλλοτε απαιτούν συνδυασμό των επιμέρους χαρακτηριστικών τους. Τα υποεπίπεδα της τεχνητής νοημοσύνης, ενδέχεται να διαφέρουν στις μεθόδους, τις μαθηματικές τους έννοιες αλλά και στο είδος της τεχνολογικής ισχύος που απαιτούν. Τα υποεπίπεδα δύνανται να διακριθούν σε [3][4]:

- *Μηχανική Μάθηση:* Αυτοματοποιεί την κατασκευή αναλυτικών μοντέλων συνδυάζοντας στατιστική, μεθόδους μαθηματικής ανάλυσης και γραμμική άλγεβρας. Στα πιο βαθιά υποστρώματα της μηχανικής μάθησης εμφανίζονται οι θεωρίες των τεχνητών νευρωνικών δικτύων και της βαθιάς μάθησης(*Deep Learning*) [3].
- *Γνωστική Υπολογιστική:* Η γνωστική υπολογιστική είναι μία μορφή τεχνητής νοημοσύνης με στόχο την δημιουργία ευφυών συστημάτων, τα οποία προσομοιώνουν ανθρώπινες αλληλεπιδράσεις ερμηνεύοντας εικόνες και ήχο, ανταποκρινόμενα με ορθολογικό τρόπο στα ερεθίσματα [3].
- *Υπολογιστική Όραση:* Επίπεδο που περιλαμβάνει την ανάλυση εικόνων και βίντεο, με σκοπό την αναγνώριση αντικειμένων-στόχων [3].
- *Επεξεργασία Φυσικής Γλώσσας:* Σύνολο τεχνικών και μεθόδων που επεξεργάζονται την ομιλούμενη γλώσσα επιχειρώντας να δημιουργήσουν ένα σημασιολογικό χάρτη για την κατανόηση, ερμηνεία και κατ' επέκταση απόκριση, σε ανθρώπινα επικοινωνιακά πρότυπα [3].

Είναι πασιφανές ότι το σύνολο των εφαρμογών της τεχνητής νοημοσύνης διεισδύει σε διάφορους επιστημονικούς κλάδους, καλούμενο να επιλύσει προβλήματα και νέες προκλήσεις με αποδοτικό τρόπο, εκεί που ξεπερνούνται τα όρια των ανθρώπινων δυνατοτήτων [4][5]. Από την πρόβλεψη δεικτών οικονομίας και την ανίχνευση περιπτώσεων οικονομικής απάτης μέχρι την ιατρική (ανίχνευση όγκων σε ιατρικές εικόνες, γονιδιωματική ανάλυση), τα ηλεκτρονικά παιχνίδια, τα αυτόνομα

αυτοκίνητα και την αγροκαλλιέργεια (χρήση ρομποτικών μηχανημάτων στην αποδοτική καλλιέργεια), η τεχνητή νοημοσύνη καλύπτει ένα ευρύ φάσμα διεργασιών, ενισχύοντας την αποδοτικότητα [5].

Παρ' όλα αυτά, η συσσώρευση δεδομένων για την παροχή εμπειρικής γνώσης στις τεχνικές τεχνητής νοημοσύνης δημιουργεί ένα τεράστιο όγκο στοιχειοσυνόλων, δυσχεραίνοντας την αποδοτικότητα των αλγοριθμικών μεθόδων και την προσπάθεια των χειριστών να ερμηνεύσουν κατάλληλα τα μοτίβα που υπάρχουν και να επιλύσουν προβλήματα [6]. Συνήθως, οι επιστήμονες διαχειρίζονται μία ευρεία κλίμακα δεδομένων, έχοντας ένα τεράστιο αριθμό συνιστωσών που καλούνται να λάβουν υπόψιν. Γι' αυτό το λόγο προβαίνουν σε τακτικές με τις οποίες μετασχηματίζονται τα δεδομένα με τέτοιο τρόπο, έτσι ώστε δύνανται να γίνουν διακριτά τα χαρακτηριστικά που αναζητούμε, καθιστώντας εφικτή και αποτελεσματική την εξαγωγή αποφάσεων για την επίλυση μίας πρόκλησης.

1.2 Γονιδιωματικά δεδομένα και Αλληλούχιση Μονοκύτταρου RNA

Τα τελευταία χρόνια ο εκσυγχρονισμός και η βελτίωση των τεχνικών χαρτογράφησης μαζικών πληροφοριών για το ανθρώπινο γονιδίωμα, παρέχει την δυνατότητα δημιουργίας βάσεων δεδομένων τεραστίων διαστάσεων, με τις οποίες αποθηκεύονται εξειδικευμένα βιολογικά δεδομένα [7][8]. Τα γονιδιωματικά δεδομένα (*Genomic Data*) ορίζονται ως ένα σύνολο πληροφοριών, που προκύπτουν έπειτα από τεχνικές αλληλούχισης βάσεων *DNA* και *RNA* [8], και μας διαθέτουν πληροφορίες οι οποίες αφορούν τα γονίδια που περιέχονται στο ανθρώπινο (ή άλλου είδους) γονιδίωμα και το ρόλο που διαδραματίζουν στη δόμηση ενός ιστού, στην κωδικοποίηση συγκεκριμένων πρωτεϊνών και τον εντοπισμό διακυμάνσεων της μορφής του *DNA* στο σύνολο του γονιδιώματος, σε ένα δείγμα ή σε ένα πληθυσμό [7][8].

Είναι ξεκάθαρο ότι νέες προκλήσεις έρχονται να δοκιμάσουν τις δυνατότητες της τεχνητής νοημοσύνης στο τομέα της βιολογικής αλληλούχισης *DNA* και *RNA* [6][9]. Η συσσώρευση γονιδιακών μετρήσεων ενδέχεται να αποκαλύπτει ιδιαίτερα σημαντικές πληροφορίες και μοτίβα, οι οποίες σε συνδυασμό με τις τεχνικές της

τεχνητής νοημοσύνης επιχειρούν να εξάγουν συμπεράσματα σε ποικίλους τομείς, όπως η ανοσοαπόκριση σε φάρμακα, η έκφραση συγκεκριμένων γονιδίων και η κωδικοποίηση πρωτεϊνών ανά ιστό, η εμφάνιση ασθενειών, ή ακόμα και η ανάδειξη πληροφοριών για εθνολογικές ποικιλομορφίες [7]. Ιδιαίτερα σημαντική αποδείχτηκε η ανάπτυξη λογισμικών αλληλούχισης γονιδιωμάτων από διάφορους ιστούς και η συγκέντρωση τους σε ανοικτές βάσεις δεδομένων ποικίλων κατηγοριών, ανάλογα την βιολογική ερευνητική προσέγγιση [9]. Ενδεικτικά, κάποιες από αυτές συνιστούν οι *Pfam* (για πρωτεϊνικά δεδομένα), η *PDB* (για δεδομένα σύστασης πρωτεϊνικών δομών) και η βάση δεδομένων *GEO*, για την ρύθμιση γονιδιακής έκφρασης σε διάφορα βιολογικά και ιατρικά ζητήματα [8][9].

Συγκεκριμένα, τα τελευταία χρόνια μία ευρέως διαδεδομένη τεχνική αλληλούχισης συνιστά η *μονοκύτταρη αλληλούχιση RNA(single cell RNA sequencing)* [8]. Η μέθοδος *RNA sequencing* επιδιώκει να εντοπίσει και να ποσοτικοποιήσει την έκφραση του αγγελιοφόρου *RNA* (*mRNA*) ενός βιολογικού δείγματος σε διάφορους κυτταρικούς τύπους [10]. Ένα κύριο πλεονέκτημα της συγκεκριμένης ανάλυσης αποτελεί η ταυτόχρονη ανάπτυξη ανεξάρτητων κυτταρικών πληθυσμών τους οποίους μπορούμε να συγκρίνουμε με ασφάλεια μεταξύ τους, στοχεύοντας στον εντοπισμό κυτταρικής ετερογένειας [10]. Η ποσοτικοποίηση της γονιδιακής έκφρασης μπορεί να αποκαλύψει ενδιαφέρουσες ενδείξεις για το πως τα γονίδια διαφοροποιούνται ανά κυτταρικό τύπο μεταξύ τους, για να επιτελέσουν μία διεργασία, όπως την κωδικοποίηση μίας πρωτεΐνης, την εκδήλωση μίας ασθένειας, ή ακόμα και την κυτταρική διαφοροποίηση κατά την δημιουργία ιστών [10]. Ένα αποτέλεσμα, που προκύπτει απ' την συγκεκριμένη αλληλούχιση, αποτελεί παράδειγμα ενός ιδιαίτερα ογκώδους συνόλου δεδομένων, το οποίο ενδέχεται να περιλαμβάνει μετρήσεις έως και 25.000 γονιδίων [11]. Ακόμα και μετά την διαδικασία επεξεργασίας των αρχικών μετρήσεων και το φιλτράρισμα ελάχιστης σημασίας γονιδίων, το πλήθος τους παραμένει αρκετά υψηλό, διατηρώντας ακόμα αρκετά θόρυβο και δυσχεραίνοντας την ανάλυση και μελέτη ιδιαίτερων χαρακτηριστικών [11]. Η διάσταση των δεδομένων παραμένει αρκετά υψηλή, απαιτώντας την ανάπτυξη τακτικών που θα καταφέρουν να τα μετασχηματίσουν σε μία καινούργια βάση, ερμηνεύσιμη και οπτικοποιημένη. Η επιλογή χαρακτηριστικών *(feature selection*), η μείωση διάστασης *(dimensionality reduction)* και η οπτικοποίηση (*visualization*) εντάσσονται στα βήματα της ανάλυσης *RNA-sequencing*, με τις οποίες λαμβάνουμε μία πρώτη γνώση για τον τρόπο με τον οποίο διασπείρονται τα στοιχεία στο χώρο [11].

Κεφάλαιο 2ο

Υψηλή Διαστατικότητα: Ζητήματα και Λύσεις

Στο παρόν κεφάλαιο διερευνούμε το ζήτημα της υψηλής διαστατικότητας στα βιολογικά δεδομένα του *RNA sequencing*, θέτοντας ως αντικείμενο μελέτης τρόπους μείωσης διάστασης και μετασχηματισμού των στοιχείων σε νέα βάση, επιδιώκοντας να εισαχθούν νέες έννοιες και προβληματισμοί.

2.1 Διαστατικότητα στην γονιδιακή έκφραση

Εννοιολογικά, είναι δυνατό να θέσουμε ως διάσταση το πλήθος των ιδιοτήτων-χαρακτηριστικών ενός συνόλου δεδομένων [12][13]. Στην περίπτωση μας, θα μπορούσαμε να υποθέσουμε την ύπαρξη μίας μήτρας δεδομένων X της μορφής nxD, όπου n αριθμός των δειγμάτων-εγγραφών, και κάθε πλειάδα x_i διαθέτει ένα σύνολο μεταβλητών D, ή αλλιώς D διαστάσεις.

$x_{n\times D}$	X_1	X_2	X_3	...	X_v
x_1	A_{11}	A_{12}	A_{13}	…	A_{1v}
x_2	A_{21}	A_{22}	A_{23}	…	A_{2v}
x_3	A_{31}	A_{23}	A_{33}	…	A_{3v}
…	…	…	…	…	…
x_n	A_{n1}	A_{n2}	A_{n3}	…	A_{nv}

Πίνακας 1 Παράδειγμα παρουσίασης ενός πίνακα χαρακτηριστικών σε ένα δειγματικό χώρο

Στην προκειμένη περίπτωση αντιμετωπίζουμε ως χώρο χαρακτηριστικών γονίδια τα οποία περιέχουν μετρήσεις για κάποιο συγκεκριμένο βιολογικό πρόβλημα. Στα βιολογικά δεδομένα, το πλήθος των γονιδίων συνήθως ξεπερνάει το πλήθος δειγμάτων [10][11]. Το σύνολο των γονιδίων, το οποίο καλούμαστε να αναλύσουμε, συνήθως υπερβαίνει τα 20.000 και σπάνια είναι λιγότερο από 10.000 [11]. Συνεπώς, η ανάπτυξη μεθόδων πρέπει να γίνει σε ένα τεράστιο εύρος διαστάσεων, με το οποίο

δύσκολα μπορούμε να αντιληφθούμε την διασπορά των σημείων στο χώρο [12]. Η συγκέντρωση υψηλού αριθμού διαστάσεων προκαλεί τη περιβόητη ***«κατάρα της διαστατικότητας»***, ένα φαινόμενο με το οποίο περιγράφεται ένα σύνολο προβλημάτων, που εμφανίζονται κατά την διαδικασία εκπαίδευσης, οργάνωσης και ανάλυσης των δεδομένων [12][13][14]. Η αύξηση της χρονικής πολυπλοκότητας και η επιβάρυνση ορισμένων αλγορίθμων, οι οποίοι θεωρούνται μη αποδοτικοί για μεγάλο όγκο εισόδου, είναι μερικές απ' τις συνέπειες του φαινομένου [12]. Τα κυριότερα προβλήματα, αντιστοιχούν στην διεύρυνση της αραίωσης των σημείων και στο γεγονός ότι οι κατά ζεύγη αποστάσεις των σημείων συγκλίνουν μεταξύ τους αποκρύπτοντας την πραγματική διασπορά [13]. Στην πρώτη περίπτωση, η εκπαίδευση των δεδομένων δεν λαμβάνει όλους του πιθανούς συνδυασμούς, αλλά τους πιο συχνούς με συνέπεια, ενώ στην φάση εκπαίδευσης το αποτέλεσμα να εμφανίζεται καλό, η δοκιμασία σε πραγματικό χρόνο να μην αποδίδει με απόλυτη ακρίβεια, καθώς σε νέες περιπτώσεις, που αντιστοιχούν σε πιο σπάνιους συνδυασμούς, το εμπειρικό σύστημα αποτυγχάνει να αποκρυπτογραφήσει την υφή των δεδομένων [13][14]. Στην δεύτερη περίπτωση, η σύγκλιση των σημείων μεταξύ τους αποτυπώνει τον διανυσματικό χώρο πιο ομογενοποιημένο από όσο είναι στην πραγματικότητα [6][13][14]. Από την στιγμή που όλες οι αποστάσεις «εμφανίζονται» ως ίσες μεταξύ τους, η έννοια της ομοιότητας καταλύεται, καθώς δεν δύναται να διακριθούν τα σημεία μεταξύ τους [13]. Συνεπώς, η διάκριση ομάδων για την εξαγωγή κοινών ιδιοτήτων είναι μη αποδοτική και δύσκολα ερμηνεύσιμη. Ταυτόχρονα, αλγόριθμοι οι οποίοι βασίζονται στις κατά ζεύγη αποστάσεις τείνουν να μην αποδίδουν σωστά, όπως ο *K-Nearest Neighbors*, ή ο *K-means* και γενικά μία πληθώρα τεχνικών ομαδοποίησης δεδομένων [13]. Από στατιστικής άποψης, αν ένα σύνολο δεδομένων διέθετε ένα τεράστιο πλήθος διαστάσεων, τότε ο απαιτούμενος αριθμός δειγμάτων προκειμένου να αποφευχθούν τα παραπάνω ζητήματα, θα αυξάνονταν εκθετικά [12]. Όπως προαναφέραμε, μία βάση δεδομένων *RNA sequencing* περιέχει χιλιάδες διαστάσεις, γεγονός που θα ισοδυναμούσε με υπερβολικά μεγάλη επιπλέον δειγματοληψία, κάτι που είναι πρακτικά αδύνατο.

2.2 Μείωση Διάστασης δεδομένων μεγάλου όγκου

Η μείωση διάστασης αποτελεί μία μη επιβλεπόμενη τεχνική μάθησης, με την οποία διαχειριζόμαστε πολυδιάστατα δεδομένα μετατρέποντας τον υψηλό διανυσματικό χώρο σε μία χαμηλότερη αναπαράσταση, διατηρώντας «ατόφιο» το μεγαλύτερο ποσοστό πληροφορίας [15]. Κατά την διαδικασία αυτή θεωρούμε ότι για ένα τεράστιο αριθμό διαστάσεων D υπάρχει ένας μικρότερος χώρος d $(d \ll D)$, ο οποίος δύναται να αναπαραστήσει τις πραγματικές αποστάσεις των σημείων, όπως αποτυπώνονται στην πολυεπίπεδη τοπολογία, διατηρώντας την αρχική γεωμετρική σύσταση [15]. Ένα βασικό πλεονέκτημα της μείωσης διαστάσεων συνιστά η δυνατότητα της οπτικοποίησης της διασποράς των σημείων σε μία δισδιάστατη ή τρισδιάστατη απεικόνιση, παρέχοντας την δυνατότητα να εντοπιστούν πιθανές διακυμάνσεις και ετερογένεια μεταξύ συγκεκριμένων πληθυσμιακών ομάδων στα δεδομένα [11][15]. Επιπρόσθετα, ο μετασχηματισμός από ένα υψηλά διανυσματικό χώρο σε ένα μικρότερο, στοχεύοντας στην λιγότερη απώλεια πληροφορίας, δημιουργεί μία σύνοψη των αρχικών στοιχείων, η οποία διοχετεύεται σε άλλες αλγοριθμικές μεθόδους, κατά πάσα πιθανότητα ομαδοποίησης ή κατηγοριοποίησης, αυξάνοντας την απόδοση τους και αντιμετωπίζοντας το φαινόμενο της *«κατάρας της διαστατικότητας» [11]*. Για την εφαρμογή της διαδικασίας μείωσης διάστασης έχει αναπτυχθεί μία σειρά από αλγορίθμους, με διαφορετικές ιδιαιτερότητες, πλεονεκτήματα και περιορισμούς [15]. Η έρευνα για τις εφαρμογές των διάφορων αλγοριθμικών μεθόδων έχει μελετηθεί αναλυτικά στην παρούσα αναφορά [15] και έτσι βασιζόμενοι στην βιβλιογραφία μπορούμε να διακρίνουμε τις περιπτώσεις όπου:

- Οι τεχνικές καταφέρνουν και εντοπίζουν μία και μόνο αναπαράσταση της πολυδιάστατης βάσης, δηλαδή το σύνολο των υποχώρων που προκύπτουν είναι «βελτιστοποίημενο» - σε ένα μικρότερο διανυσματικό χώρο (*convex*) και αυτές με την σειρά τους διαχωρίζονται σε [15]:
  - Αλγορίθμους που εφαρμόζουν ιδιοδιάσπαση τιμών σε ένα μητρώο με γραμμική συσχέτιση μεταξύ στοιχειοσυνόλων, το οποίο είτε περιέχει τις συνδιακυμάνσεις μεταξύ των διαστάσεων, είτε τις κατά ζεύγη αποστάσεις μεταξύ σημείων (*Full Spectral*). Παραδείγματα αποτελούν ο *PCA* και ο *Isomap* [15].
  - Αλγορίθμους που εφαρμόζουν ιδιοδιάσπαση τιμών σε ένα γενικευμένο πρόβλημα ιδιοδιάσπασης, δηλαδή δεν είναι εφικτό

να βρεθούν ιδιοδιανύσματα γραμμικά ανεξάρτητα, ώστε να αποτυπώνουν την νέα βάση (*Sparse Spectral*) [16]. Παραδείγματα συνιστούν ο *Laplacian Eigenmaps* και ο *Hessian LLE* [15]

- Οι τεχνικές αναζητούν και εντοπίζουν τον βέλτιστο απ' τους πολλούς μετασχηματισμούς βάσεων που έχουν εφαρμόσει (*non-convex*) μέσω συνάρτησης κόστους. Παράδειγμα αποτελεί ο *Sammon Mapping* [15].

Επιπρόσθετα σε μία πιο γενικευμένη προσέγγιση θα μπορούσαμε να κατηγοριοποιήσουμε τις αλγοριθμικές μεθόδους σ' αυτές που επιδιώκουν να συντηρήσουν άθικτη την δομή πληροφορίας στο σύνολο το δεδομένων και σ' αυτές που συντηρούν τις συσχετίσεις των σημείων σε τοπικό εύρος, «θυσιάζοντας» κατά κάποιο τρόπο την συνολική αρχική δομή [6].

Κεφάλαιο 3ο

Θεωρητικό Υπόβαθρο Αλγοριθμικών Τεχνικών

Στα παραπάνω κεφάλαια διαπιστώθηκε ότι η συσσώρευση μεγάλου όγκου δεδομένων συνοδεύεται με συγκεκριμένα προβλήματα μαθηματικής φύσεως και μη που δυσχεραίνουν το έργο της ανάλυσης και της απόδοσης των μεθόδων εξαγωγής μοτίβων και συμπερασμάτων. Η μείωση διάστασης, λοιπόν, καλείται να δώσει λύσεις στο θέμα υψηλής διαστατικότητας, αναπτύσσοντας συγκεκριμένους αλγορίθμους, οι οποίοι διατηρούν το χρήσιμο κομμάτι πληροφορίας ατόφιο. Στο παρόν κεφάλαιο, παρουσιάζουμε το βασικό (και όχι σε βάθος) θεωρητικό υπόβαθρο αλγορίθμων μείωσης διάστασης. Η επιλογή των αλγορίθμων έγινε εξαιτίας της διασημότητας και πολυχρησιμότητας τους (Ανάλυση Κύριων Συνιστωσών-*PCA*), των συχνών εφαρμογών σε δεδομένα *single cell RNA sequencing* (*t-SNE, UMAP*) και βάσει βιβλιογραφικής έρευνας (*RGt-SNE, RPEV*).

3.1 Ανάλυση Κύριων Συνιστωσών

Η ανάλυση κύριων (ή πρωτευόντων) συνιστωσών (***Principal Component Analysis, PCA***) συνιστά μία από τις πιο ευρέως διαδεδομένες τεχνικές στην μείωση διάστασης δεδομένων μεγάλου όγκου [17]. Πρόκειται για μία από τις πρώτες τεχνικές που εφαρμόζονται σε δεδομένα μεγάλης κλίμακας, καθώς ένα απ' τα κύρια πλεονεκτήματα της είναι ότι ο μετασχηματισμός τους σε μία νέα βάση προσδίδει υψηλή δυνατότητα ερμηνείας διατηρώντας όσο το δυνατόν ελάχιστη απώλεια πληροφορίας [18].

Ανακαλώντας την υποενότητα *2.2*, η *PCA* ανήκει στην πρώτη κατηγορία αλγορίθμων μείωσης διάστασης όπως ορίζουν οι συγγραφείς στο έργο [15] και αποτελεί ένα από τα πιο χαρακτηριστικά παραδείγματα μεθόδων που αποσκοπούν στην συμπύκνωση της μέγιστης συνολικής διακύμανσης των δεδομένων με τη δυνατότερη

ελάχιστη απώλεια [6]. Η βασική μεθοδολογία της *PCA*, σχετίζεται με την αναζήτηση του γραμμικού συνδυασμού των πολυδιάστατων δεδομένων, με ταυτόχρονη διατήρηση της μέγιστης διακύμανσης [18]. Η μείωση των διαστάσεων γίνεται ουσιαστικά μέσω της γεωμετρικής προβολής των σημείων σε μικρότερες διαστάσεις, τις κύριες συνιστώσες [19], οι οποίες μπορεί να αντιστοιχούν είτε στο πλήθος των δειγμάτων (γραμμών) είτε στο πλήθος των ιδιοτήτων-χαρακτηριστικών (στήλες), αναλόγως με το ποιο είναι πιο μικρό [18][19]. Με τον όρο κύρια συνιστώσα, αναζητούμε την ευθεία η οποία βελτιστοποιεί την διακύμανση, ή αλλιώς ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα [20]. Αναζητούμε, δηλαδή, τις αποστάσεις των σημείων από την ευθεία, των οποίων το άθροισμα των τετραγώνων τους είναι μέγιστο [20]. Η πρώτη κύρια συνιστώσα είναι αυτή που αποτυπώνει την μέγιστη προβαλλόμενη διακύμανση των στοιχείων και στην συνέχεια έπονται οι ακόλουθες κύριες συνιστώσες σε φθίνουσα σειρά. Κάθε κύρια συνιστώσα είναι ασυσχέτιστη με τις προηγούμενες της και γι' αυτό κάθε συνιστώσα είναι κάθετη στην προηγούμενη της [18][19]. Το γεγονός ότι κάθε κύρια συνιστώσα είναι ασυσχέτιστη με τις προηγούμενες της μας υποδεικνύει ότι διαφορετικές μεταβλητές επηρεάζουν την συνολική διακύμανση των σημείων στην εύρεση της βέλτιστης γραμμής [17][18][19]. Παραδείγματος χάριν, σε βιολογικά δεδομένα, όπου γονίδια συνιστούν τις ιδιότητες που μελετάμε, πιθανότατα τα σημεία να διασπείρονται με μεγαλύτερη εμβέλεια εξαιτίας του γονιδίου Α στην πρώτη κύρια συνιστώσα, ενώ αντίθετα για την δεύτερη κύρια συνιστώσα το γονίδιο Β ενδέχεται να είναι πιο σημαντικό στον τρόπο που προβάλλονται τα σημεία [18]. Κάθε κύρια συνιστώσα εξηγεί ένα ποσοστό διακύμανσης και αναζητείται το κατάλληλο πλήθος (*r*-διάστατη βάση), ώστε να προβούμε σε μείωση διαστάσεων με όσο το δυνατόν λιγότερη απώλεια πληροφορίας [18][19].

Η βασική μεθοδολογία της *PCA* περιλαμβάνει βήματα που αφορούν τεχνικές γραμμικής άλγεβρας. Σε μία υποθετική πολυδιάστατη μήτρα δεδομένων $n \times d$, αναζητούμε μία *r*-διάστατη βάση, που αποτυπώνει την μέγιστη συνολική διακύμανση. Αρχικά, κεντράρουμε τα δεδομένα, έτσι ώστε να έχουν μέσο $\mu=0$. Στην συνέχεια, κατασκευάζεται από τα κεντραρισμένα δεδομένα η μήτρα συνδιακύμανσης, η οποία περιγράφει όλα τα ζευγάρια των μετρήσεων που διαθέτουμε, ποσοτικοποιώντας ουσιαστικά τις συσχετίσεις τους [18][20]. Η χρησιμότητα της μήτρας συνδιακύμανσης αποτυπώνεται στο γεγονός ότι οι ιδιοτιμές της, δηλαδή οι τιμές που μηδενίζουν το χαρακτηριστικό πολυώνυμο, συνιστούν το ποσό της συνολικής διακύμανσης που περιγράφει κάθε συνιστώσα [20]. Η μεγαλύτερη ιδιοτιμή μεγιστοποιεί την

προβαλλόμενη διακύμανση [20]. Αντίστοιχα, τα ιδιοδιανύσματα που προκύπτουν αποτελούν την μειωμένη βάση, δηλαδή την ευθεία που ελαχιστοποιεί το τετράγωνο σφάλματος [20]. Τα τετράγωνα των αποστάσεων των δεδομένων από την βέλτιστη ευθεία που αναζητάμε σε σχέση με τα κεντραρισμένα δεδομένα, ουσιαστικά αντιστοιχούν στις ιδιοτιμές της μήτρας συνδιακύμανσης και τα ιδιοδιανύσματα στις κατευθύνσεις, δηλαδή τις κύριες συνιστώσες [20]. Οι τετραγωνικές ρίζες των ιδιοτιμών καθορίζουν τα μήκη των ημιαξόνων [20]. Για τον προσδιορισμό της βέλτιστης δυνατής διάστασης συνήθως θεωρούμε ως επαρκές ένα κατώφλι της τάξης 80-90% της συνολικής διακύμανσης [20]. Ο τελικός μετασχηματισμός γίνεται με τον πολλαπλασιασμό των αρχικών δεδομένων με την ορθογώνια μήτρα ιδιοδιανυσμάτων στον αριθμό διαστάσεων που έχουμε επιλέξει [20].

Αλγόριθμος 1 Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis) *[20]*

Είσοδος	$PCA\ (\ D_{n\times d}\ )$	➢ Εισαγωγή πολυδιάστατου πίνακα *D*.
1 :	$\mu\ =\ \frac{1}{n}\sum_{i=1}^{n} x_i$	➢ Εύρεση μέσης τιμής πίνακα *D*.
2 :	$Z =\ D - 1.\mu^T$	➢ Κεντράρισμα τιμών πίνακα *D* μέσω εύρεσης μέσης τιμής.
3 :	$S = \frac{1}{n}\ (\ Z^T Z\ )$	➢ Υπολογισμός μήτρας συνδιακύμανσης απ' τα κεντραρισμένα δεδομένα.
4 :	$(\lambda_1\ ,\ \lambda_2, \dots\ ,\lambda_d) = eigenvalues\ (S)$	➢ Εύρεση ιδιοιτιμών από την μήτρα συνδιακύμανσης.
5 :	$U = (\ u_1\ , u_2\ , \dots\ u_d) = eigenvectors\ (S)$	➢ Εύρεση ιδιοδιανυσμάτων ή αλλιώς κυρίων συνυστωσών από την μήτρα συνδιακύμανσης.
6 :	$f(r) = \frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$	➢ Επιλογή βέλτιστης μείωσης διάστασης
7 :	$A = \{\ a_i\ \vert \alpha_\iota = U_r^T \alpha\ x_i\ \}$	➢ Μετασχηματισμός σε μειωμένη βάση $A\ n\ \times r$

3.2 *t (κατανομή) - Στοχαστική Αφομοίωση Γειτονικών Σημείων*

Η ***t -Στοχαστική Αφομοίωση Γειτονικών σημείων (t – Stochastic Neighbor Embedding, t-SNE)*** είναι μία από τις πιο σύγχρονες και ευρέως διαδεδομένες τεχνικές για μείωση υψηλών διαστάσεων, όντας ιδιαίτερα εύχρηστο εργαλείο για βιολογικά δεδομένα γονιδιακής έκφρασης [11]. Ο αλγόριθμος αναπτύχθηκε από τον *Laurens van der Maaten* το 2008 [21] και έκτοτε αποτέλεσε ένα χρήσιμο εργαλείο για την διαχείριση δεδομένων μεγάλου όγκου. Όντας ένας αλγόριθμος που ανήκει στην κατηγορία του μη γραμμικού προγραμματισμού, ο *t-SNE* συνιστά ένα χρήσιμο εργαλείο για την διαδικασία οπτικοποίησης πολυδιάστατων δεδομένων σε μία δισδιάστατη απεικόνιση, παρουσιάζοντας ενδεχόμενες διαφοροποιήσεις σε πληθώρα πληθυσμιακών ομάδων πάνω στα δεδομένα μας.

Ο *t-SNE* αποτελεί ένα συνδυασμό των *Stochastic Neighbor Embedding (SNE),* τεχνική που αποτελεί το θεμέλιο και την χρήση της *Student-t* κατανομής για την κατάρτιση μίας μήτρας, που αντιπροσωπεύει την κατά ζεύγη ομοιότητα μεταξύ των σημείων στον υποχώρο προβολής [21][22]. Ο *t-SNE* υποθέτει ότι δεδομένης μιας υψηλής διάστασης *D* για ένα σύνολο δεδομένων *X*, αν δύο σημεία xi και xj είναι πολύ κοντά μεταξύ τους, τότε και σε μία μειωμένη αναπαράσταση d (όπου $d << D$), τα ομόλογα σημεία yi και yj απέχουν και αυτά ελάχιστα μεταξύ τους. Για να επιτευχθεί αυτό, αρχικά υπολογίζονται οι ομοιότητες για το αρχικό σετ δεδομένων X μέσω Γκαουσιανής συνάρτησης πυκνότητας πιθανοτήτων, ώστε να δημιουργηθούν οι υπό συνθήκη πιθανότητες, οι οποίες αναπαριστούν τον δείκτη ομοιότητας [21][22]. Πρακτικά ερμηνεύεται πόσο πιθανό είναι το σημείο xi να επιλέξει ως γείτονα του το σημείο $xj \rightarrow (pj|i)$ [21][22].

Εν αντιθέσει, με την βασική εκδοχή του *SNE* χρησιμοποιείται μία συμμετρική εκδοχή του, καθώς υπολογίζεται η από κοινού πιθανότητα για τα σημεία i, j, ώστε όλα τα σημεία να έχουν μία σημαντική συνεισφορά στην συνάρτηση κόστους, ακόμα και αν είναι ακραίο σημείο (*outlier*) [21][22].

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2} \quad (2)$$

Αντίστοιχα, στον χώρο (συνήθως δισδιάστατο) τον οποίο προσπαθούμε να προβάλουμε τα σημεία, χρησιμοποιούμε την *Student-t* κατανομή η οποία έχει μικρότερη κορυφή και πιο μεγάλο εύρος ακραίων σημείων (σε σχέση με την Γκαουσιανή) [21][22][23] . Ο λόγος χρήσης της συγκεκριμένης κατανομής γίνεται, ώστε να υπάρχει μία πιο ακριβής αντιστοιχία μεσαίας κλίμακας τιμών στον πολυδιάστατο χώρο και υψηλών τιμών στην συμπυκνωμένη αναπαράσταση, προκειμένου να μην υπάρξει συγκέντρωση σημείων μεταξύ τους, που στην πραγματικότητα δεν έχουν υψηλή ομοιότητα [21][22].

$$qij = \frac{\left(1 + \left|\left|yi - yj\right|\right|^2\right)^{-1}}{\sum_{k \neq l}(1 + ||\mathrm{yk} - \mathrm{yl}||^2)^{-1}} \quad (3)$$

Αφού καταρτιστούν οι πιθανοτικοί δείκτες ομοιότητας για τους δύο διανυσματικούς χώρους, ορίζουμε ως συνάρτηση κόστους των πιθανοτικών κατανομών την απόκλιση των *Kullback-Leibler*, ένα μέτρο το οποίο συγκρίνει πόσο κοντά είναι οι εκτιμήσεις των δύο κατανομών [21][22]. Όσο πιο μικρή είναι η τιμή απόκλισης, τόσο πιο πιστή είναι δισδιάστατη αναπαράσταση. Έτσι, χρησιμοποιώντας επαναληπτική βελτιστοποίηση πρώτης τάξης αναζητούμε τοπικό ελάχιστο μέσω διαφόρισης, που ελαχιστοποιεί την συνάρτηση κόστους και καταλήγουμε [21][22] στην εξίσωση:

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) (\mathbf{y}_i - \mathbf{y}_j)(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1} \quad (4)$$

Ο *t-SNE* είναι στοχαστικός αλγόριθμος και επομένως κάθε φορά που τον εφαρμόζουμε το αποτέλεσμα διαφέρει [24]. Επιπλέον, περιέχει ένα μεγάλο σύνολο παραμέτρων, οι οποίες επιτελούν διαφορετικές διεργασίες, συγκεκριμένα βελτιστοποίησης (*ρυθμός μάθησης-learning rate, ορμή-momentum, βαθμός μέγιστης νόρμας-max_grad norm*), επιτάχυνσης (*Burnes-Hut*) και βασικής δόμησης (*παράγοντας σύγχυσης-perplexity*, *υπερβολή-exaggeration*) [21][22][24]. Η παραμετροποίηση αποτελεί σύνθετη διαδικασία με άμεση επιρροή στην απόδοση του αλγορίθμου. Παρ' όλα αυτά άξια αναφοράς θεωρείται αυτήν την στιγμή, ίσως η πιο σημαντική παράμετρος (υπερπαράμετρος) του *perplexity*. Μπορεί να θεωρηθεί ως μία παράμετρος αναλογική με αυτή τον *k*-κοντινότερων γειτόνων, που ορίζουν το εύρος τιμών της

τυπικής απόκλισης της Γκαουσιανής κατανομής και χρησιμεύει στην διατήρηση των αποστάσεων των σημείων για διαφοροποιημένες βάσει πυκνότητας περιοχές στο χώρο [21][22][24].

Ορίζεται ως $Perplexity\ (Pi) = 2^{H(Pi)}$, όπου $H(pi)$ είναι η εντροπία του *Shannon* για την κατανομή Pi. Όσο αυξάνεται η τυπική απόκλιση σi , αυξάνεται και η εντροπία .

$$H(\mathrm{p}_i) = -\sum_i p_{j|i} \log_2 (p_{j|i})$$

Αλγόριθμος 2 ***Ψευδοκώδικας t-Distributed Stochastic Neighbor Embedding [21]***

1: ***Δεδομένα: Σύνολο Δεδομένων*** $\boldsymbol{X = \{x_1, x_2, \dots, x_n\}}$

2: *Θέσε παραμέτρους συνάρτησης κόστους :* perplexity Per p,

3: *Θέσε παραμέτρους βελτιστοποίησης: πλήθος επαναλήψεων T, βαθμός μάθησης η, momentum α(t)*

4: ***Αποτέλεσμα****: χαμηλής-διάστασης αναπαράσταση δεδομένων* $Y^{(t)} = \{y_1, y_2, \dots, y_n\}$

5: ***Ξεκίνα***

6: *Υπολόγισε τις κατά ζεύγη σχέσεις* $Pj|i$ *με perplexity Per p (Εξίσωση 1)*

7: *Θέτω* $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}$

8: *Δοκίμασε την αρχική λύση* $Y^{(t)} = \{y_1, y_2, \dots, y_n\}$ *από* $N(0, 10^{-4}I)$

9: ***Για*** *t=1 μέχρι T* ***κάνε:***

10: *Υπολόγισε τις κατά ζεύγη σχέσεις στον δοθέντα υποχώρο* qij *(Εξίσωση 3)*

11: *Διαφόρισε* $\frac{\delta C}{\delta y}$ (*Εξίσωση 4*)

12: *Θέσε* $Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$

13: ***Τέλος επανάληψης***

14: ***Τέλος***

3.3 Ομοιόμορφη Προσέγγιση και Προβολή Πολύπτυχου Μορφώματος

Η ***Ομοιόμορφη Προσέγγιση και Προβολή Πολύπτυχου μορφώματος*** (***Uniform Manifold Approximation and Projection, UMAP***) είναι ένα από τους πιο σύγχρονους αλγόριθμους της λίστας, ο οποίος αναπτύχθηκε από τους *McInnes*, και *Healy* το 2018 [25] και έκτοτε εξελίχθηκε σε μία από τις πιο διαδεδομένες μεθόδους για την μείωση διάστασης σε πολυδιάστατα δεδομένα και ιδιαίτερα σε επίπεδο γονιδιακής έκφρασης [11]. Ο *UMAP* αποτελεί ένα εξαιρετικά εύχρηστο εργαλείο, με χρησιμότητα τόσο σε επίπεδο προβολής σε χαμηλή διάσταση και εν συνεχεία διοχέτευσης σε τεχνικές κατηγοριοποίησης ή ομαδοποίησης, άλλα και σε επίπεδο οπτικοποίησης πολυδιάστατων δεδομένων σε δισδιάστατη αναπαράσταση [26]. Ο συγκεκριμένος αλγόριθμος θεμελιώνεται από ένα ισχυρό μαθηματικό υπόβαθρο, το οποίο έχει τις βασικές του αρχές στην εκμετάλλευση των γεωμετρικών ιδιοτήτων του πολυδιάστατου χώρου και της τοπολογικής του ανάλυσης [27]. Ανήκοντας στην κατηγορία των μεθόδων μη γραμμικού προγραμματισμού, όπως και ο *t-SNE*, οργανώνει ένα γράφο βαρών, που αναπαριστά το «πόσο κοντά» ή πόσο όμοια είναι ένα σύνολο σημείων μεταξύ τους [27].

Η κεντρική ιδέα βασίζεται στην αναζήτηση μίας τοπικής συσχέτισης των σημείων, με σεβασμό στις γεωμετρικές τους ιδιότητες στον χώρο υψηλής διάστασης και ακολούθως στην μετατροπή σε μία ακανόνιστη-ασαφή(*fuzzy*) τοπολογική αναπαράσταση [25]. Ύστερα, δεδομένου ενός μειωμένου διανυσματικού χώρου επαναλαμβάνεται η ίδια διαδικασία και αναζητείται ο βελτιστοποιημένος χώρος που ελαχιστοποιεί την συνάρτηση κόστους, η οποία υπολογίζεται μέσω του μέτρου της εγκάρσιας εντροπίας (*Cross – Entropy*) [25]. Η διαδικασία της τοπολογικής αναζήτησης ξεκινάει, αφού διατυπωθούν οι εξής υποθέσεις [25][26]:

- Τα δεδομένα μας ακολουθούν ομοιόμορφη κατανομή (αν και στα πραγματικά δεδομένα είναι εξαιρετικά δύσκολο).
- Το μετρικό σύστημα ακολουθεί τις ιδιότητες της γεωμετρίας *Riemannian* για ένα τοπικά συνδεδεμένο σύμπλοκο μεταξύ των σημείων.

Λαμβάνοντας υπόψιν τις παραπάνω υποθέσεις, δημιουργούνται διακριτές μετρικές αποστάσεων για κάθε τοπικό χώρο που δημιουργείται, οι οποίες πρέπει να

συνδυαστούν σε μία συνολική δομή που θα αναπαριστά το πολυδιάστατο επίπεδο [25][26].

Η λογική για να αναπαραστήσουμε αυτό το χώρο ακολουθεί την ευριστική αναζήτηση απλοϊκών αντικειμένων [25], που δημιουργούνται με την σύνδεση σημείων (δηλαδή δύο σημεία δημιουργούν μία γραμμή, τρία σημεία ένα τρίγωνο κ.ο.κ.) και κατ' επέκταση το συνδυασμό τους για την δημιουργία ένα συνολικού συμπλόκου *Čech* [27]. Κάθε σημείο αντιμετωπίζεται ως ξεχωριστό δείγμα, από το οποίο ξεκινάει μία ακτίνα προς όλες τις κατευθύνσεις δημιουργώντας ένα πεδίο που μοιάζει με μπάλα [26] και επεκτείνεται έως ότου επικαλυφθεί με το πεδίο ενός άλλου σημείου [27]. Τα σημεία αυτά στην συνέχεια συνδέονται δημιουργώντας σύμπλοκα διάφορων διαστάσεων [27]. Η τοπολογική αναπαράσταση ανά περιοχή φαίνεται να ερμηνεύεται κατάλληλα από την πιο απλοϊκή μορφή (σημείο και γραμμή) ενός συμπλόκου κατασκευάζοντας το γράφο ομοιότητας με αποδοτικό και υπολογιστικά επικερδές τρόπο [27]. Παρ' όλα αυτά είναι προφανές ότι στην πραγματικότητα τα δεδομένα σε υψηλές διαστάσεις δεν καθιστούν σαφή την έννοια της απόστασης μεταξύ των σημείων *(«κατάρα της διαστατικότητας»),* δυσχεραίνοντας το ζήτημα επιλογής του εύρους της ακτίνας προέκτασης [27]. Μία πολύ μικρή τιμή κατασκευάζει ένα μεγάλο πλήθος συνδεδεμένων συμπλόκων, ενώ μία μεγάλη τιμή δεν απλοποιεί την ανάκτηση της τοπολογικής αναπαράστασης [26]. Η προσαρμογή σε πραγματικού επιπέδου δυσκολίας δεδομένα, αντιμετωπίζεται μέσω της χρήσης ποικίλων τιμών για την ακτίνα κάθε σημείου βασιζόμενο στην απόσταση του από τον *k*-κοντινότερο γείτονα του [27]. Ταυτόχρονα, τίθεται ως προϋπόθεση ότι ακόμα και τα πιο απομακρυσμένα σημεία συνδέονται τουλάχιστον με το πιο κοντινό του [27]. Αυτό που καθιστά την τοπολογική αναπαράσταση ως «ακανόνιστο» σύμπλεγμα είναι η μετατροπή της συνεκτικότητας των σημείων σε μία μετρήσιμη πιθανοτική ποσότητα, με την οποία αναγνωρίζεται το κατά πόσο είναι πιθανό δύο σημεία να είναι κοντά (χαμηλή πιθανότητα ισοδυναμούν μακρινά σημεία) [27].

$P_{i|j} = e^{-\frac{d(x_i,x_j)-\rho_i}{\sigma i}}$, όπου ρ_i εκφράζει την απόσταση ενός σημείου i από τον πρώτο κοντινότερο γείτονα [28].

Αν και ορίστηκε μία αρκετά ισχυρή μαθηματική θεώρηση, από την στιγμή που κάθε σημείο θεωρήθηκε ξεχωριστή οντότητα αναζήτησης συνεκτικότητας, η έννοια αντίληψης της απόστασης διαφέρει. Παραδείγματος χάριν, ένα σημείο «α»

αντιλαμβάνεται με ένα συγκεκριμένο τρόπο την απόσταση από ένα σημείο «β», ενώ από την οπτική του «β» η απόσταση διαφέρει [27]. Για να επιλυθεί το συγκεκριμένο ζήτημα, επιλέγεται η συνδυαστική επιλογή των δύο τιμών-βαρών, ή από γεωμετρικής απόψεως η κατά κάποιον τρόπο ένωση των δύο διαφορετικών τοπολογικών συμπλόκων [27].

Τελικά, όλα καταλήγουν σε ένα γράφο βαρών, με τις ακμές-βάρη να αντικατοπτρίζουν την πιθανότητα δύο σημεία να συνδέονται στο υψηλά διαστατικό χώρο.

Στην συνέχεια αναζητείται μία νέα τοπολογική προσέγγιση στο μειωμένο διαστατικά χώρο που αναζητούμε, ακολουθώντας την προηγούμενη διαδικασία, έχοντας ένα ορισμένο χώρο που θέλουμε να προσεγγίσουμε, όπως υπολογίσαμε στο προηγούμενο βήμα και πραγματοποιώντας έπειτα βελτιστοποίηση [25][26]. Στην χαμηλή διάσταση χρησιμοποιείται μία οικογένεια καμπυλών της μορφής $q_{i|j} = \left(1 + a(y_i - y_j)^{2b}\right)^{-1}$, για να ποσοτικοποιήσει τις πιθανότητες των αποστάσεων των σημείων [28] και απαιτείται η ελαχιστοποίση της συνάρτησης κόστους της εγκάρσιας εντροπίας, που καλείται να συγκρίνει την απόκλιση των δύο ακανόνιστων τοπολογικών χώρων, η οποία έπειτα από εφαρμογή στοχαστικής γραμμικής παλινδρόμησης λαμβάνει την εξής μορφή, $\mathrm{C}_l(\mathrm{X}, \mathrm{Y}) = \sum_{i=1}^{l} \lambda_i C(X_i, Y_i)$ [28]

Ο *UMAP* περιέχει μία σειρά από υπερπαραμέτρους που καθορίζουν το τελικό του αποτέλεσμα και την τελική διασπορά των σημείων. Σ' αυτές τις υπερπαραμέτρους παρουσιάζονται ο αριθμός των γειτόνων (*n_neighbors*), η ελάχιστη απόσταση (*min_dist*), οι διαστάσεις προβολής (*n_components*) και ο αριθμός εποχών βελτιστοποίησης [25]. Η πιο σημαντική παράμετρος αντιστοιχεί στον αριθμό γειτόνων, η οποία καθορίζει το επίπεδο ισορροπίας μεταξύ της τοπικής και ολικής αποτίμησης του μειωμένου χώρου προβολής [25][26]. Μικρές τιμές συνήθως δίνουν προβάδισμα στην τοπική συνεκτικότητα των σημείων, ενώ υψηλές τιμές απεικονίζουν σε μεγαλύτερο βαθμό την συνολική δόμηση των δεδομένων, αγνοώντας παρ' όλα αυτά τις λεπτομερείς συσχετίσεις των σημείων [25][26]. Τέλος, η παράμετρος της ελάχιστης απόστασης καθορίζει το πόσο κοντά «πακετάρονται» τα σημεία μεταξύ τους και αποκτά ιδιαίτερο ενδιαφέρον, αν το ενδιαφέρον της μελέτης αφορά την ικανότητα ομαδοποίησης των σημείων μέσω του αλγορίθμου [26].

3.4 Δύο σύγχρονες προσεγγίσεις μείωσης διάστασης

Η ανάλυση κύριων συνιστωσών (*PCA*) , ο *t-SNE* και ο *UMAP*, ταυτοποιήθηκαν στα παραπάνω κεφάλαια ως τρεις από τους πιο ευρέως διαδεδομένους αλγορίθμους στην μείωση διάστασης δεδομένων μεγάλου όγκου, με ιδιαίτερο βαθμό χρησιμότητας στα προβλήματα γονιδιακής έκφρασης [11]. Η δημοτικότητα των παραπάνω τεχνικών και η ισχυρή μαθηματική τους «ταυτότητα», αν και βασίζεται σε διαφορετικά πρότυπα, τις καθιστά ως βασικά μέσα σύγκρισης απόδοσης με νέους αλγορίθμους και μεθόδους ή και ακόμα ως βασικό εργαλείο διεκπεραίωσης συνδυαστικών αλγοριθμικών μεθοδολογιών, όπως θα μελετήσουμε στο παρόν κεφάλαιο.

Αναζητώντας σε σύγχρονες βιβλιογραφικές αναφορές επιλέγουμε να μελετήσουμε δύο νεοσύστατες αλγοριθμικές μεθόδους, τον *Random Projection and Geodesic Distances t-Stochastic Neighbor Embedding* [6] και τον *Random Projection Ensemble Visualization* [29], καθώς η εφαρμογή τους σε δεδομένα *single cell RNA-sequencing* μας ενθαρρύνουν να τους δοκιμάσουμε και να τους αξιολογήσουμε. Οι δύο αλγόριθμοι συνιστούν, όπως αναφέραμε, ένα είδος συνδυαστικής εκτέλεσης γνωστών αλγοριθμικών μεθόδων μείωσης διάστασης με ενδιάμεσα επεξεργαστικά βήματα [6][29]. Στόχος βέβαια, εκτός από την μείωση διάστασης των δεδομένων μεγάλης κλίμακας είναι και η οπτικοποίηση στο δισδιάστατο χώρο των πολυδιάστατων στοιχείων, επιτυγχάνοντας να αποκαλύψουν διακριτά μοτίβα, αν αυτό είναι εφικτό [6][29]. Τα βήματα συνοψίζονται στους παρακάτω πίνακες και καθώς γίνεται αντιληπτό ότι τα περισσότερα από αυτά είναι κοινά και στις δύο προσεγγίσεις θεωρείται σκόπιμο να αναφερθούμε προκαταβολικά στο θεωρητικό υπόβαθρο των τυχαίων προβολών μίας ακόμη τεχνικής μείωσης διάστασης, προτού αναλύσουμε κάθε αλγόριθμο ξεχωριστά.

3.4.1 Μέθοδος τυχαίων προβολών

Η τεχνική των τυχαίων προβολών αποτελεί μία μαθηματική μέθοδος, έχοντας ως βασικό θεμέλιο στην προσέγγιση της το λήμμα των ***Johnson–Lindenstrauss*** [30]**.** Η συγκεκριμένη θεμελιώδης έννοια αποδεικνύει ότι ένα σύνολο σημείων σε ένα υψηλά Ευκλείδιο διανυσματικό χώρο μπορούν να αποτυπωθούν σε ένα μικρότερο χώρο, χωρίς να προκληθεί αλλοίωση των αποστάσεων μεταξύ οποιωνδήποτε ζευγών σημείων

[30][31]. Η διατήρηση των αρχικών αποστάσεων των σημείων εξαρτάται από ένα παράγοντα παραμόρφωσης ε της αρχικής δομής που κυμαίνεται μεταξύ του 0 και του 1 ($0 < \varepsilon < 1$) [30][31].

$$(1-\varepsilon) \parallel u - v \parallel^2 < \parallel p(u) - p(v) \parallel^2 < (1+\varepsilon) \parallel u - v \parallel^2$$

Μέσω του θεωρήματος καθίσταται δυνατή η εύρεση ενός p τυχαίου υποχώρου μικρότερης διάστασης από τον αρχικό, ο οποίος δεν προκαλεί αλλοίωση των αποστάσεων σε βαθμό $0 < \varepsilon < 1$. [30][31][32]. Με αυτή την μέθοδο μπορούμε να κατασκευάσουμε μία τυχαία μήτρα $R_{d \times p}$, μικρότερης διάστασης από τα αρχικά δεδομένα $X_{n \times d}$ και στην συνέχεια να προβάλλουμε τον αρχικό χώρο d στο τυχαίο υποχώρο p [30][31][32].

$$A_{n \times p} = X_{n \times d} \times R_{d \times p}$$

Σημείο κλειδί της αποτελεσματικότερης εφαρμογής της μεθόδου αποτελεί η επιλογή της τυχαίας μήτρας και ο τρόπος με τον οποίο κατανέμονται οι τιμές της. Συνήθης τακτική αποτελεί η δημιουργία ενός διανυσματικού χώρου του οποίου τα στοιχεία ακολουθούν μία Γκαουσσιανή κατανομή $N(0, \frac{1}{\Pi\lambda\acute{\eta}\theta o\varsigma\ \Sigma υνιστωσών})$ [30][31][32], ή μία ομοιόμορφη κατανομή *N(0,1)* [6]. Επιπλέον ιδιαίτερα αποδοτική έχει αποδειχτεί η χρήση τυχαίας αραιής μήτρας [33], που λειτουργεί βάσει του βαθμού πυκνότητας που ορίζουμε ώστε να επιλέξει την καλύτερη δυνατή προβολή στοιχείων, τόσο στην ποιότητα των αποτελεσμάτων, όσο στην δέσμευση μνήμης και χρονικής υπολογιστικής πολυπλοκότητας [33].

Το θεώρημα διευκρινίζει τον ελάχιστο αριθμό των δυνατών p διαστάσεων που επαρκούν για να αποτυπώσουν την μειωμένη αναπαράσταση με όσο το δυνατόν καλύτερη διατήρηση της αρχικής δομής, δεδομένου του παράγοντα αλλοίωσης ε [30][31].

$$\pi\lambda\acute{\eta}\theta o\varsigma\ \nu\acute{\epsilon}\omega\nu\ \sigma υνιστωσών \geq \frac{4\log(\pi\lambda\acute{\eta}\theta o\varsigma\ \delta ειγμάτων)}{\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)}$$ [30]

Εν κατακλείδι, σημαντικό είναι ότι πολλές φορές δεν επαρκεί μία μόνο εφαρμογή τυχαίας προβολής, αλλά είναι απαραίτητο να χρησιμοποιηθεί μία αλγοριθμική προσέγγιση πολλαπλών τυχαίων προβολών [6][34][29]. Η πολλαπλή

εφαρμογή τυχαίας προβολής στα αρχικά δεδομένα μπορεί να αποτελέσει ιδιαίτερα χρήσιμο εργαλείο στην προσέγγιση της καλύτερης δυνατής επιλογής *p* υποχώρου, αλλά και στην προσαρμογή τους σε μία γραμμική μείωση διάστασης [34] και εν συνεχεία στην διοχέτευση τους σε μία πιο σύνθετη αλγοριθμική προσέγγιση μείωσης διάστασης (*t-sne, Isomap, Multidimensional Scaling*) [6][29][34], εξασφαλίζοντας καλύτερη απόδοση και αποτίμηση του πολυεπίπεδου χάρτη δεδομένων [34].

3.4.2 Μέθοδος Πολλαπλών Τυχαίων Προβολών και Συντομότερων Μονοπατιών σε t(κατανομή)-Στοχαστική Αφομοίωση Γειτονικών Σημείων

Το όνομα του συγκεκριμένου αλγορίθμου προέρχεται από την εφαρμογή τριών βασικών τεχνικών, την μέθοδο τυχαίων προβολών (*Random Projections*), την αναζήτηση συντομότερων μονοπατιών μεταξύ δύο κορυφών (*Geodesic Distances*) και του *t-Stochastic Neighbor Embedding* [6].

Η κεντρική ιδέα του *RGt-SNE* είναι η αναδιάταξη των κατά ζευγών αποστάσεων μεταξύ των κυττάρων-δειγμάτων (εφόσον εξετάζουμε γονιδιακά δεδομένα) σε μία νέα μήτρα αποστάσεων που θα περιγράφει σε μία όσο το δυνατόν πιο ρεαλιστική «απεικόνιση» τις διαφορές μεταξύ των κυττάρων [6]. Στην συνέχεια, η κατασκευασμένη μήτρα αποστάσεων διοχετεύεται ως είσοδος στον *t-SNE* έναντι των δεδομένων σε μη επεξεργασμένη μορφή [6]. Ο *RGt-SNE* εκμεταλλεύεται την αλγοριθμική απλότητα της μεθόδου τυχαίων προβολών για την προβολή των αρχικών δεδομένων σε ένα μικρότερης διάστασης χώρο σε ελάχιστη χρονική διάρκεια, διατηρώντας ατόφιο όσον το δυνατόν μεγαλύτερο όγκο χρήσιμης πληροφορίας [6]. Ο νεοσύστατος χώρος, αφού έχει «απλοποιηθεί», παρέχει πλέον την δυνατότητα στην αναζήτηση των *k* κοντινότερων γειτόνων, ο οποίος, αν και συνηθώς εμφανίζεται σε εφαρμογές επιβλεπόμενης μάθησης και κατηγοριοποίησης, παρουσιάζει και ενδιαφέρον στην μη επιβλεπόμενη μάθηση, καθώς δημιουργούνται γράφοι που προσπαθούν να αναδείξουν την τοπολογία των κυττάρων, χωρίς να χρειαστούν να αναζητήσουν μοτίβα γύρω απ' τις γεωμετρικές ιδιότητες των διανυσματικών χώρων (όπως προσπαθεί να προσεγγίσει ο *UMAP*) [6]. Έτσι, η κατασκευή γράφων, προσπαθεί

να χαρτογραφήσει τα προφίλ γονιδιακής έκφρασης μεταξύ των κυττάρων, αφού ως κόμβοι θεωρούνται τα κύτταρα-δείγματα, ενώ οι ακμές αποτυπώνουν ουσιαστικά τον βαθμό ομοιότητας μεταξύ των διαφορετικών κυττάρων [6]. Το αποτέλεσμα του παραπάνω συνδυασμού καθορίζει, ύστερα από περεταίρω επεξεργασία, την επιθυμητή μήτρα αποστάσεων, με την οποία επανακαθορίζουμε τις πραγματικές αποκλίσεις μεταξύ των κυτταρικών στοιχείων, προσπαθώντας να εισάγουμε τα δεδομένα στον *t-SNE* σε μία πρώιμη βελτιστοποιημένη μορφή, ώστε να αποτυπωθεί στον δισιδιάστατο χώρο [6].

Τα αλγοριθμικά βήματα του *RGt-SNE* δύνανται να συνοψιστούν σε τρία κύρια σκέλη [6]:

I. Δεδομένης μίας αρχικής πολυδιάστατης μήτρας $A \in R^{Sxd}$, όπου s το πλήθος δειγμάτων και d το αρχικό πλήθος διαστάσεων (μεγάλος αριθμός), μετασχηματίζουμε τα δεδομένα με τέτοιο τρόπο ώστε να κατασκευαστεί μία μήτρα ομοιότητας μεταξύ των δειγμάτων. Για να επιτευχθεί αυτό, εφαρμόζουμε τεχνικές μείωσης της αρχικής διάστασης d, σε μία πολύ μικρότερη διάσταση r ($r << d$) μέσω της τεχνικής τυχαίων προβολών. Έτσι δημιουργούμε μία τυχαία μήτρα R_{sxr} και εν συνεχεία προβάλλουμε την αρχική μήτρα με την τυχαία μέσω της πράξης $B^{RP}_{s\times r} = A_{s\times d} \times R_{sxr}$. Για να επιτευχθεί η όσο το δυνατόν ακριβέστερη προσέγγιση της μήτρας ομοιότητας, ο αλγόριθμος δεν περιορίζεται σε μία απλή προβολή σε μικρότερη διάσταση, αλλά ακολουθεί μία επαναληπτική διαδικασία κατασκευάζοντας L νέους υποχώρους διάστασης r. Σε κάθε νέο r προβαλλόμενο χώρο αναζητούμε τους k κοντινότερους γείτονες για κάθε δείγμα, συγκρατώντας τους σε μία νέα μήτρα της μορφής $K^i_{s\times k}$. Σε κάθε βήμα επανάληψης i, λοιπόν αντιστοιχούν k κοντινότεροι γείτονες πάνω στον Ευκλείδιο χώρο αποστάσεων για κάθε δείγμα, στο νέο προβαλλόμενο χώρο. Τα αποτελέσματα, έπειτα από L αριθμό επαναλήψεων, συγκεντρώνονται σε νέο μητρώο $S_{L\times s\times k}$ που περιέχει L πίνακες της μορφής $K^i_{s\times k}$, όπως εξηγήσαμε παραπάνω. Στην συνέχεια, ορίζουμε ως βαθμό ομοιότητας μεταξύ των δύο δειγμάτων (i, j) το πλήθος των περιπτώσεων των οποίων κάθε δείγμα j εμφανίστηκε ως ένας εκ των κοντινότερων

γειτόνων του δείγματος i. Έτσι, κατασκευάζεται ο τετραγωνικός πίνακας ομοιότητας $PD_{S\times S}$, ο οποίος περιγράφει το κατά πόσο όμοιο είναι κάθε δείγμα σε σχέση με όλα τα άλλα, σε όλο το σύνολο δειγμάτων. Είναι ξεκάθαρο πως ο πίνακας θα έχει ιδιαίτερα υψηλό ποσοστό αραιότητας, καθώς θα περιέχει πολλά μηδενικά στοιχεία, ενώ η μέγιστη τιμή ενός στοιχείου μπορεί να ίση με L (δηλαδή το δείγμα j εμφανίστηκε L φορές ως ένας από τους κοντινότερους γείτονες του δείγματος i) , δηλαδή το πλήθος των επαναλήψεων που ορίζεται [6].

II. Από την στιγμή που ορίσαμε τον τρόπο κατασκευής της μήτρας ομοιότητας, σημαντικό είναι να ξεκαθαριστεί το πως θα γίνει η μετάβαση της σε πίνακα αποστάσεων. Στην πραγματικότητα, δεν υπάρχει κάποιος συγκεκριμένος τρόπος μετάβασης και η εναλλαγή γενικά αποτελεί ένα ενδιαφέρον αντικείμενο συζήτησης. Στην προκειμένη περίπτωση επιλέγεται να γίνει αντιστροφή των στοιχείων, δημιουργώντας ένα πίνακα $W_{S\times S}$. Προφανώς, τα μηδενικά απαγορεύεται να αντιστραφούν και έτσι παραμένουν άθικτα στον πίνακα αποστάσεων. Παρ' όλα αυτά είναι σημαντικό να αναφερθεί ότι η σχέση μεταξύ πίνακα ομοιότητας και πίνακα αποστάσεων είναι αντιστρόφως ανάλογη, δηλαδή δύο πολύ όμοια στοιχεία μοιράζονται ένα υψηλό βαθμό ομοιότητας, ο οποίος με την αντιστροφή μετατρέπεται σε πολύ μικρή τιμή που συγκλίνει στο μηδέν. Το πρόβλημα, όμως, που προκύπτει αφορά την ύπαρξη απείρακτων λόγω αντιστροφής μηδενικών τιμών, τα οποία στην πραγματικότητα, ενώ αφορούν ιδιαίτερα ανόμοια στοιχεία, μπορούν να ληφθούν ως όμοια, παρότι δεν είναι. Το ζήτημα αυτό επιλύεται μέσω της αναζήτησης το κοντινότερου μονοπατιού (*geodesic distances*), με το οποίο θεωρούμε έναν γράφο με κόμβους κάθε δείγμα και ως βάρος ακμής μεταξύ δύο δειγμάτων i, j, το στοιχείο $\mathrm{W}(i, j)$ του πίνακα απόστασης. Η εύρεση του κοντινότερου μονοπατιού κάθε ζεύγους δειγμάτων επιλύει το παραπάνω πρόβλημα, καθώς θεωρεί τα μηδενικά ως μη συνδεδεμένα μονοπάτια στον γράφο αναζήτησης και έτσι αποθηκεύει τον ελάχιστο αριθμό ακμών που χρειάζεται ένα δείγμα i για να φτάσει στο δείγμα j. Υπάρχουν διάφοροι αλγόριθμοι

αναζήτησης συντομότερου μονοπατιού, με τον πιο γνωστό να αποτελεί η προσέγγιση του *Djikstra* [6].

III. Το τελικό σκέλος αφορά την εισαγωγή της μήτρας $D_{S\times S}$ που υπολογίστηκε από τον αλγόριθμο του *Djikstra* στον τελικό αλγόριθμο μείωσης διάστασης *t-SNE*, τον οποίο αναλύσαμε στην υποενότητα 3.2, οπτικοποιώντας τα αποτελέσματα στο δισδιάστατο χώρο [6].

Αλγόριθμος 3 *Random Projection and Geodesic Distances t-SNE Visualization*[6]

1: **Διαδικασία** *RGt-SNE* ($As \times d$, L, r) ► *A: δεδομένα εισόδου,* L: *πλήθος τυχαίων υποχώρων,* r: *διάσταση προβολής*

2: ***Για*** $i = 1 : L$ ***επανέλαβε***

3: $R_{d\times r}$ = *τυχαία μήτρα ()*

4: $B^i_{s\times r} = A_{S\times d} \times R_{d\times r}$

5: $K^i_{s\times k}$ = $knn(B^i_{s\times r})$ ► *Υπολόγισε κ κοντινότερους γείτονες*

6: $S_{L\times s\times k} = VOTE^L_{i=1} K^i_{s\times k}$ ► *Συγκέντρωσε τα αποτελέσματα του kNN*

7: $PD_{S\times S} = COUNT^L_{i=1} S^i_{s\times k}$ ► *Υπολόγισε βαθμό ομοιότητας*

8: $W_{S\times S} = 1./PD_{S\times S}$ ► *Δημιουργία πίνακα αποστάσεων απ' τις ομοιότητες μέσω αντιστροφής κάθε στοιχείου*

9: $D_{S\times S} = Djikstra\ (W_{S\times S})$

10: $y_{S\times 2} = t-SNE\ (D_{S\times S})$

11: ***Επέστρεψε*** $y_{S\times 2}$ ►*Δισδιάστατα δεδομένα για οπτικοποίηση*

3.4.*3 Οπτικοποίηση Συνόλου Τυχαίων Προβολών*

Ο δεύτερος αλγόριθμος που ανακτήθηκε έπειτα από βιβλιογραφική έρευνα, έχοντας ως κεντρικό άξονα αναζήτησης την εφαρμογή σε δεδομένα γονιδιακής έκφρασης, είναι ο αλγόριθμος πολλαπλών τυχαίων προβολών για οπτικοποίηση δεδομένων μεγάλου όγκου (*Random Projection Enesemble Visualisation)* [29]. Όπως και στον *RGt-SNE* [6], έτσι και στον *RPEV*, το βασικό προεπεξεργαστικό στάδιο αφορά

τον καθορισμό μίας νέας μήτρας αποστάσεων, η οποία περιγράφει το επίπεδο ανομοιότητας μεταξύ των διαφορετικών κυτταρικών πληθυσμών σε χαμηλότερη διάσταση, υπερβαίνοντας τις δυσκολίες που προσφέρουν τα πολυδιάστατα δεδομένα, στα οποία οι αποστάσεις μεταξύ των δειγμάτων τείνουν να ομογενοποιούνται και να συγκλίνουν μεταξύ τους [29].

Τα πρώτα βήματα του αλγορίθμου είναι παρόμοια με αυτά που προηγήθηκαν στην υποενότητα *3.4.2* και έτσι η αναδιάταξη των ομοιοτήτων των δειγμάτων *scRNA-sequencing* μεταξύ τους επιτυγχάνεται μέσα από την απλή τεχνική των πολλαπλών τυχαίων προβολών στην αρχική μήτρα δεδομένων, επιδιώκοντας να παράξει συμπυκνωμένους *r-υποχώρους* για κάθε έναν από τους οποίους αναζητούνται οι k κοντινότεροι γείτονες [29]. Η συγκέντρωση των πολλαπλών τυχαίων προβολών σε συνδυασμό με τον αλγόριθμο *kNN* ενδέχεται να αποκαλύψει σημαντικές ιδιότητες μεταξύ των κυτταρικών δειγμάτων.

Τα αλγοριθμικά βήματα του *RPEV* για ένα αρχικό σύνολο δεδομένων $D_{n\times a}$ (όπου n το πλήθος δειγμάτων και a ο υψηλός χώρος διαστάσεων) συνοψίζονται [29]:

a) Ορίζοντας έναν αριθμό επαναληπτικών διαδικασιών L, μαζί με όσο τον δυνατόν ελάχιστο r-διάστατο χώρο προβολής (όπου $r << a$) και διενεργώντας βρόγχο επανάληψης με τον οποίο η αρχική μήτρα προβάλλεται σε μικρότερη διάσταση μέσω της διαδικασίας υπολογισμού $D_{n\times r}^{RP\,[i]} = D_{n\times a} \times R_{a\times r}$, όπου $R_{a\times r}$ η τυχαία μήτρα προβολής. Κάθε βρόγχος επανάληψης διαθέτει ως τελική διεργασία την εφαρμογή του αλγορίθμου *kNN* αναζητώντας για κάθε δείγμα τα k δείγματα που εμφανίζονται ως κοντινότεροι γείτονες [29].

b) Η τελική συγκεντρωτική μορφή αποτυπώνει ουσιαστικά L υποπίνακες οι οποίοι για κάθε δείγμα αποθηκεύουν τα k δείγματα που βρίσκονται πιο κοντά του. Πάνω σε αυτούς τους L υποπίνακες υπολογίζουμε τον βαθμό ομοιότητας μεταξύ δύο δειγμάτων i, j μετρώντας το πόσες φορές ένα στοιχείο j εμφανίστηκε ως κοντινότερος γείτονας του i. Η μέγιστη τιμή κάθε στοιχείου στον κατασκευασμένο πίνακα ομοιότητας μπορεί να ισοδυναμεί με τον αριθμό επαναλήψεων L [29].

c) Ο πίνακας ομοιότητας $S_{s\times s}$ μετασχηματίζεται σε άνω συμμετρικό πίνακα αποστάσεων εναλλάσσοντας τις μηδενικές τιμές με 0.01 και διοχετεύοντας τον, στον τελικό αλγόριθμο οπτικοποίησης σε δισδιάτατη

απεικόνιση *Multidimensional Scaling*. Ο *Multidimensional Scaling* συνιστά ένα αλγόριθμο μείωσης διαστάσεων, ο οποίος, δεδομένης της απόστασης δύο σημείων σε πολυδιάστατο επίπεδο, επιδιώκει να οπτικοποιήσει τα σημεία βασιζόμενος στην γεωμετρική τους κατανομή στον χώρο [29] [35].

***Αλγόριθμος 4 Random Projection Ensemble Visualization* [29]**

1: ***Διαδικασία*** $RPEV(D_{n\times a})$

2: *Καθόρισε L και r.*

3: ***Όσο*** $i \leq L$ ***επανέλαβε:***

4: *Δημιούργησε τυχαία μήτρα* $R_{n\times r}$

5: $D_{n\times r}^{RP[i]} = D_{n\times a} \times R_{a\times r}$

6: $K_{n\times k}^{i} = knn(D_{n\times r}^{RP[i]})$ ► *Εφάρμοσε αναζήτηση K κοντινότερων γειτόνων*

7: $S = VOTE_{i=1}^{L} K_{s\times k}^{i}$ ► *Συνδύασε τα αποτελέσματα του βήματος 6*

8: *Μετέτρεψε τον πίνακα ομοιότητας σε άνω συμμετρικό πίνακα αποστάσεων.*

9: *M = MDS(S)* ► *Υπολόγισε την δισδιάστατη απεικόνιση*

10: ***Επέστρεψε*** *M*

Κεφάλαιο 4ο

Single cell RNA-seq ανάλυση για την έκφραση γονιδίων σε δύο διαφορετικά σύνολα δεδομένων

Αφού καλύψαμε το θεωρητικό υπόβαθρο των αλγορίθμων μείωσης διάστασης δεδομένων μεγάλης κλίμακας, στόχος μας στο παρόν κεφάλαιο είναι η πρακτική εφαρμογή τους και η ερμηνεία των αποτελεσμάτων τους. Η ανάλυση γίνεται πάνω σε δύο πειραματικά δεδομένα *single cell RNA sequencing*, όπου εφαρμόζονται τεχνικές μείωσης διάστασης και κατόπιν διενεργείται η οπτικοποίηση τους σε δισδιάστατη αναπαράσταση. Οι τρεις άξονες πάνω στους οποίους στηρίζουμε την ανάλυση μας αφορούν την περιγραφή και ταυτοποίηση του εκάστοτε συνόλου δεδομένων, την μείωση διάστασης και οπτικοποίηση τους σε κάθε μία από τις προαναφερόμενες αλγοριθμικές τεχνικές και εν τέλει την αξιολόγηση τους.

4.1 Περιγραφή χαρακτηριστικών συνόλου δεδομένων γονιδιακής έκφρασης καρκινοπαθών ασθενών

Η αποτελεσματική τεχνική *single cell RNA sequencing* για την ποσοτικοποίηση της γενετικής έκφρασης έχει ωθήσει διάφορους ερευνητές και πανεπιστημιακές μονάδες στην μελέτη διάφορων καρκινικών ασθενειών, με κλασσικό παράδειγμα να αποτελεί η ***The Cancer Genome Atlas (TCGA)*** , ένας κατάλογος που συγκεντρώνει δεδομένα για γενετικές μεταλλάξεις που προκαλούν καρκίνο χρησιμοποιώντας τεχνικές *sequencing* και βιοπληροφορική. Στην περίπτωση μας ανακτούμε δεδομένα(*dataset*) που αφορούν την γονιδιακή έκφραση διαφόρων ειδών καρκίνων μέσω της *Kaggle,* συσχετιζόμενο με *το The cancer genome atlas pan-cancer analysis project*. Το συγκεκριμένο *dataset* αφορά την ανάλυση έκφρασης γονιδίων για διαφόρων ειδών καρκίνων. Περιλαμβάνει δύο αρχεία σε μορφή *.csv* που μας διαθέτουν την *RNA sequencing* ανάλυση για διάφορους ασθενείς(*data.csv*) και την διάγνωση για την μορφή καρκίνου που πάσχουν(*label.csv*). Μέσω του πακέτου *pandas* διαβάζουμε

τα δύο αρχεία, τα εμφανίζουμε στην μορφή *dataframe* και κάνουμε μία αρχική διερεύνηση στο περιεχόμενο των δύο αρχείων μας.

Ξεκινώντας με το *dataframe* που περιέχει την αλληλούχιση *RNA*, παρατηρούμε ότι περιέχονται αριθμητικά δεδομένα τα οποία αναπαριστούν την τιμή έκφρασης πολλών διαφορετικών γονιδίων σε ένα υψηλό δειγματικό χώρο. Τα δεδομένα μας είναι ετερογενή, δηλαδή κάθε δείγμα αφορά διαφορετικό ασθενή, επομένως αναμένονται διάφορες κλιμακώσεις στις τιμές κάθε γονιδίου και μεταξύ των δειγμάτων-ασθενών. Ο δημιουργός της συγκεκριμένης βάσης δεδομένων θεώρησε περιττό να αναφέρει τα πραγματικά ονόματα των γονιδίων και προτίμησε να χρησιμοποιήσει ετικέτες της γενικευμένης μορφής γονίδιο_1,γονίδιο_2 κ.ο.κ. Τέλος, ο πίνακας εγγραφών με τις διαγνώσεις περιέχει απλά το πόρισμα για την μορφή καρκίνου κάθε ασθενούς.

Τα βασικά χαρακτηριστικά που μπορούμε να εξάγουμε σε πρώτο βαθμό για τα δεδομένα μας συνιστούν:

- *Αριθμός στοιχείων/Μέγεθος δεδομένων:* Το *dataset* αποτελείται από *20531* γονίδια (στήλες) που αφορούν *801* δείγματα (διαφορετικούς) ασθενείς (γραμμές). Με όρους γραμμικής άλγεβρας αναφερόμαστε σε πίνακα 801×20531, ενώ ο όγκος των δεδομένων ανέρχεται στα *16445331* στοιχεία(στην περίπτωση μας μετρήσεις έκφρασης γονιδίου). Σημαντική θεωρείται η ανίχνευση μη γνωστών (*NULL,NAN*) τιμών, οι οποίες αλλοιώνουν βασικές μετρήσεις και προκαλούν προβλήματα στην χρήση πολλών αλγορίθμων. Οι τεχνικές που πραγματεύονται τον τρόπο διαχείρισης των άγνωστων αυτών τιμών αποτελούν ενδιαφέρουσα μελέτη, ιδιαίτερα όταν αφορούν μεγάλο ποσοστό των δεδομένων(ακόμα και πάνω από 15%), καθώς υπάρχει κίνδυνος απώλειας σημαντικής πληροφορίας. Στην περίπτωση μας δεν υπάρχουν άγνωστες τιμές, με αποτέλεσμα να μην χρειάζεται κάποια δραστική παρέμβαση.

- *Απομάκρυνση περιττών τιμών:* Παρακολουθώντας τα δεδομένα με μία πρώτη εποπτεία στο εύρος των τιμών ανά γονίδιο, διακρίνεται ότι υπάρχουν γονίδια με υψηλή συχνότητα εμφάνισης μηδενικών τιμών. Ύστερα από επεξεργασία των δεδομένων εντοπίζονται γονίδια που περιέχουν μόνο μηδενικές τιμές. Κρίνεται, έτσι σημαντικό να απορριφθούν τα γονίδια με αυτά τα χαρακτηριστικά. Καθώς απορρίπτουμε αυτά τα γονίδια, μεταφερόμαστε από τις

20531 διαστάσεις στις *20264*, όχι αρκετά μεγάλη μείωση άλλα ίσως χρήσιμη για την απώλεια θορύβου και μείωση της αραιότητας (*sparsity*) του πίνακα.

- *Κατηγοριοποίηση δεδομένων:* Η ύπαρξη του δεύτερου *dataframe* μας γνωστοποιεί το είδος καρκίνου που πάσχει κάθε ασθενής μέσω της στήλης *Class*. Όπως αναφέρθηκε προηγουμένως, σημαντική είναι η διαδικασία ανίχνευσης μη υπαρκτών τιμών (*missing values*), πάρα όλα αυτά παρατηρείται πλήρης κατηγοριοποίηση των δειγμάτων. Ύστερα από αναζήτηση προκύπτει ότι διαθέτουμε 5 κλάσεις, οι οποίες αφορούν πέντε διαφορετικούς τύπους καρκίνου με το πλήθος ανά κατηγορία να αντιστοιχεί σε:

    1. *Breast Carcinoma (BRCA) για 300 ασθενείς.*
    2. *Kidney Renal Clear Cell (KIRC) για 146 ασθενείς*
    3. *Lungs Adenocrinoma (LUAD) για 141 ασθενείς.*
    4. *Prostate Adenocrinoma (PRAD) για 136 ασθενείς.*
    5. *Colon Adenocrinoma (COAD) για 78 ασθενείς.*

Παρατηρούμε ότι οι περιπτώσεις καρκίνου του μαστού ξεπερνούν κατά πολύ τις άλλες κλάσεις στην συχνότητα εμφάνισης. Στα ίδια επίπεδα κυμαίνονται οι περιπτώσεις *KIRC, PRAD, LUAD*, ενώ η κατηγορία *COAD* εμφανίζει τα λιγότερα.

Συνοπτικά, με την μελέτη των βασικών χαρακτηριστικών αντιλαμβανόμαστε ότι με την ύπαρξη ετικετών για κάθε δείγμα μπορούμε να εφαρμόσουμε ποικίλες μορφές επιβλεπόμενης μάθησης, όπως εκπαίδευση ενός μοντέλου πρόβλεψης μορφής καρκίνου. Παράλληλα, όμως, η παρουσία ενός μεγάλου όγκου μεταβλητών(τα *20531* γονίδια) καθιστούν δύσκολη την προσπάθεια ανάλυσης και επίβλεψης των δεδομένων. Επιπρόσθετα, δεν είναι δυνατή η οπτικοποίηση του πολυδιάστατου χώρου που καταγράφει την συμπεριφορά και σύγκλιση των σημείων στον χώρο. Η ανάγκη μείωση της διαστατικότητας γίνεται, επομένως, εμφανής μαζί με την εφαρμογή τεχνικών οπτικοποίησης που θα συντηρούν και δεν θα αλλοιώνουν την αρχική δομή των δεδομένων.

Για το στάδιο της προ-επεξεργασίας (*pre-proscessing step*) θα εφαρμόσουμε μία απλή κλιμάκωση, χρησιμοποιώντας την πιο απλή εκδοχή του *Standard Scaler* από το πακέτο του *sklearn* της *Python*.

4.2 Περιγραφή χαρακτηριστικών συνόλου δεδομένων γονιδιακής έκφρασης καλλιεργειών με στελέχη κορωναϊού

Η πανδημία του *SARSCov2*, που προκαλεί την ασθένεια του *Covid-19,* ώθησε την επιστημονική κοινότητα στην άμεση έρευνα και χαρτογράφηση του γονιδιώματος του ιού. Ήδη από τις αρχές εμφάνισης του ιού έχει γίνει αποτελεσματική χαρτογράφηση και μαζί με τις τεχνικές αλληλούχισης νέας γενιάς, οι οποίες εγγυούνται υψηλή απόδοση, αναζητούνται διάφορες ιδιότητες ανάμεσα στα γονίδια και τις αλληλεπιδράσεις με τα στελέχη του ιού σε κυτταρικό επίπεδο. Διάφορες πλατφόρμες όπως η *GEO,* δημοσιεύουν και ανανεώνουν το περιεχόμενο τους με διάφορες αναλύσεις σχετικά με τον *Covid-19*. Παραμένοντας πιστοί στην τεχνική *scRNA sequencing*, η οποία ποσοτικοποιεί την έκφραση γονιδίων σε διάφορα κύτταρα, ανακτούμε δεδομένα σχετικά με τον πανδημικό «φαινόμενο» από την *GEO(GSE148729)* [36].

Το συγκεκριμένο *dataset* συνιστά ένα τεράστιο σύνολο δεδομένων, το οποίο μελετά τις γονιδιακές επιδράσεις και τις κυτταρικές οδούς, που εμφανίζουν περιπτώσεις μολύνσεων δύο διαφορετικών στελεχών κορωνοϊού, του *SARSCoV1* και του *SARSCoV2*, που προκαλούν σοβαρά αναπνευστικά προβλήματα. Η κλινική μελέτη αφορούσε την μόλυνση ποικίλων κυτταρικών τύπων με στελέχη των δύο ανωτέρω ιών σε διάφορες χρονικές φάσεις, και η αλληλούχιση με διάφορες μεθόδους [36]. Η ανάκτηση αφορούσε δύο αρχεία σε μορφή *text*, τα οποία περιλαμβάνουν τις μετρήσεις μέσω *RNA sequencing* (*GSE148729_Calu3_scRNAseq_morethan1000genes_Raw counts_tr.txt*) και τα χαρακτηριστικά γνωρίσματα κάθε κυτταρικού τύπου που αναλύουμε (*GSE148729_Calu3_scRNAseq_morethan1000genes_metadata.txt*) [36]. Μέσω του πακέτου *pandas* διαβάζουμε τα δύο αρχεία, τα εμφανίζουμε στην μορφή *dataframe* και κάνουμε μία αρχική διερεύνηση στο περιεχόμενο των δύο αρχείων μας.

Ξεκινώντας με το *dataframe* που περιέχει την *scRNA seq* ανάλυση, παρατηρούμε ότι περιέχονται αριθμητικά δεδομένα, τα οποία αναπαριστούν την τιμή έκφρασης πολλών διαφορετικών γονιδίων σε ένα υψηλό δειγματικό χώρο. Οι δημιουργοί της συγκεκριμένης βάσης δεδομένων αναφέρουν τα πραγματικά ονόματα των γονιδίων και αναθέτουν ως τιμή δείκτη το χαρακτηριστικό κωδικό ταυτοποίησης κάθε κυττάρου. Έτσι κάθε πλειάδα διαθέτει το δικό της ξεχωριστό αναγνωριστικό, ενώ διατίθεται ένα σύνολο κοινών γονιδίων ως στήλες. Τέλος, το *dataframe* με τα

γνωρίσματα κάθε κυττάρου περιλαμβάνει δώδεκα στήλες (αφού ορίσουμε και εδώ ως δείκτη κάθε πλειάδας κοινά αναγνωριστικά με το προηγούμενο *dataframe*), από τις οποίες θα μας απασχολήσουν συγκριμένα οι στήλες *infect, strain* και *orig.ident.*

Προτού προβούμε σε οποιαδήποτε ανάλυση και μείωση της διαστατικότητας, συνηθίζεται να περιγράφονται βασικά χαρακτηριστικά των δεδομένων που διαθέτουμε (πλήθος στοιχείων, ανίχνευση μη υπαρκτών τιμών, ιστογράμματα κ.α) ,ενώ απαραίτητο θεωρείται πολλές φορές και το στάδιο της προ-επεξεργασίας (κανονικοποίηση τιμών ή και λογαριθμήσεις για υψηλά και χαμηλά εκφραζόμενα γονίδια). Για το στάδιο της προ-επεξεργασίας (*pre-proscessing step*) αυτήν την φορά θα εφαρμόσουμε μία απλή κανονικοποίηση, χρησιμοποιώντας την πιο απλή εκδοχή του *Normalizer* από το πακέτο του *sklearn* της *Python*.

Τα βασικά χαρακτηριστικά που μπορούμε να εξάγουμε σε πρώτο βαθμό για τα δεδομένα μας συνιστούν:

- <u>*Αριθμός στοιχείων/Μέγεθος δεδομένων:*</u> Το *dataset* αποτελείται από *27072* ονομαστικά γονίδια (στήλες), που αφορούν *48890* δείγματα (πλειάδες) επιθηλιακών κυττάρων, τα οποία μολύνονται με στελέχη των ιών *SARSCoV1* και *SARSCoV2*, και συλλέγονται σε διάφορες φάσεις ανάπτυξης. Με όρους γραμμικής άλγεβρας αναφερόμαστε σε πίνακα 48890×27072, ενώ ο όγκος των δεδομένων ανέρχεται στα 1.323.550.080 στοιχεία (στην περίπτωση μας μετρήσεις έκφρασης γονιδίου). Σημαντική θεωρείται η ανίχνευση μη γνωστών (*NULL,NAN*) τιμών, οι οποίες αλλοιώνουν βασικές μετρήσεις και προκαλούν προβλήματα στην χρήση πολλών αλγορίθμων. Οι τεχνικές που πραγματεύονται τον τρόπο διαχείρισης των άγνωστων αυτών τιμών αποτελούν ενδιαφέρουσα μελέτη, ιδιαίτερα όταν αφορούν μεγάλο ποσοστό των δεδομένων(ακόμα και πάνω από 15%) καθώς υπάρχει κίνδυνος απώλειας σημαντικής πληροφορίας. Στην περίπτωση μας δεν υπάρχουν άγνωστες τιμές, με αποτέλεσμα να μην χρειάζεται κάποια δραστική παρέμβαση.

- <u>*Κατηγοριοποίηση δεδομένων:*</u> Η ύπαρξη του δεύτερου *dataframe* μας γνωστοποιεί διάφορα γνωρίσματα για κάθε δείγμα που διαθέτουμε. Εν αντιθέσει με το προηγούμενο σύνολο δεδομένων, στην προκειμένη περίπτωση αναγνωρίζουμε την παρουσία τριών διαφορετικών ειδών ετικετών, οι οποίες χαρακτηρίζουν τα δεδομένα. Διερευνώντας κάθε είδος γνωρίσματος αποκαλύπτεται ότι:

- Η πρώτη στήλη ***infect*** περιγράφει αν κάποιο κύτταρο μολύνθηκε από κάποιο στέλεχος των δύο διαφορετικών ιών ή όχι. Στο σύνολο διαθέτουμε *36262* μολυσμένα κύτταρα και *12628* μη μολυσμένα.

- Η δεύτερη στήλη ***strain*** αναφέρεται στο είδος του στελέχους του ιού, δηλαδή αν ανήκει στο *SARSCoV1* ή *SARSCoV2*. Από την αναζήτηση προκύπτει ότι έχουμε *16243* (*33*%) εγγραφές για το *SARSCoV1, 16135* (*33*%) για το *SARSCoV2* και *16512* άγνωστες περιπτώσεις (*34*%).

- Η τρίτη στήλη περιγράφει το είδος κυττάρου και την φάση ανάπτυξης καλλιέργειας σε ώρες.

Για να γίνουν αντιληπτά καλύτερα τα δεδομένα, τα ομαδοποιούμε με τέτοιο τρόπο ώστε να διακρίνεται βάση περίπτωση μόλυνσης, τι στέλεχος έχει εισχωρήσει. Έτσι διακρίνονται:

1. *16243* μολυσμένα και με στέλεχος *SARSCoV2* κύτταρα
2. *16135* μολυσμένα και με στέλεχος *SARSCoV1* κύτταρα
3. *3884* μολυσμένα και με άγνωστο στέλεχος ιού κύτταρα
4. *12628* μη μολυσμένα κύτταρα, συνεπάγεται ότι δεν υπάρχει στέλεχος ενεργό

Ένα ζήτημα είναι ότι τα συγκεκριμένα δεδομένα είναι ιδιαίτερα υπέρογκα , γεγονός που υπερβαίνει τις επεξεργαστικές δυνατότητες και απαιτήσεις μνήμης του μηχανήματος εργασίας που διαθέτουμε. Γι' αυτό το λόγο προβαίνουμε σε μία δειγματοληψία δεδομένων, με στρωματοποιημένη χρήση ετικετών σε όλες τις επιλεγμένες εγγραφές και με σεβασμό στην ποσοστιαία κατανομή των κατηγοριών. Αυτό σημαίνει ότι δειγματοληπτούμε τυχαία 2000 εγγραφές με στόχο να πετύχουμε την ομαδοποίηση που αναφέραμε παραπάνω (33% *SarsCov2*, 33% *SarsCov1*, 8% *Unknown Infection*, 26% *Uninfected*). Επομένως, το μετασχηματισμένο *dataset* που διαθέτουμε μεταβαίνει σε 2000 εγγραφές και *27072* στήλες, περιέχοντας:

I. 660 *Infected* και *SARSCoV1*

II. 664 *Infected* και *SARSCoV2*

III. 159 *Infected* και U*nknown Infection*

IV. 517 *Uninfected*

- *Απομάκρυνση περιττών τιμών:* Παρακολουθώντας τα δεδομένα με μία πρώτη εποπτεία στο εύρος των τιμών ανά γονίδιο, διακρίνεται ότι υπάρχουν γονίδια με υψηλή συχνότητα εμφάνισης μηδενικών τιμών. Ύστερα από επεξεργασία των δεδομένων εντοπίζονται γονίδια που περιέχουν μόνο μηδενικές τιμές. Κρίνεται έτσι σημαντικό να απορριφθούν τα γονίδια με αυτά τα χαρακτηριστικά, διότι δεν έχουν να προσφέρουν κάτι σημαντικό, ενώ ενδέχεται να επιβραδύνουν άσκοπα την διαδικασία ανάλυσης, έστω και κάποια δευτερόλεπτα παραπάνω. Καθώς απορρίπτουμε αυτά τα γονίδια, μεταφερόμαστε από τις *27072* διαστάσεις στις *17934*, σχετικά μέτρια μείωση άλλα ίσως χρήσιμη για την απώλεια θορύβου και μείωση της αραιότητας (*sparsity*) του πίνακα.

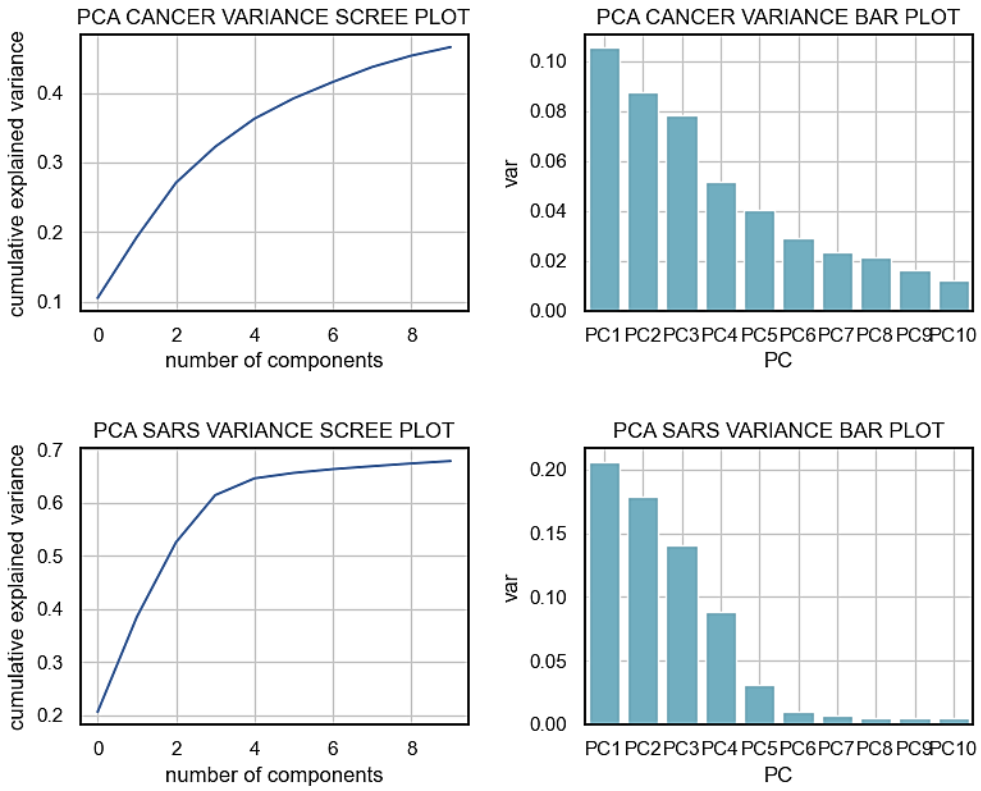
4.3 Μείωση διάστασης και οπτικοποίηση

Οι αλγοριθμικές τεχνικές που εφαρμόζονται αφορούν τον ευρέως διαδεδομένο αλγόριθμο της ανάλυσης κύριων συνιστωσών (*PCA*), μαζί με το *t-SNE* και τον *UMAP*. Επιπρόσθετα, αναλύεται και αξιολογείται η επίδοση δύο προτεινόμενων αλγορίθμων του *PREV* και του *RGt-SNE*.

4.3.1 Εφαρμογή Ανάλυσης Κύριων Συνιστωσών και αποτελέσματα

Η ανάλυση κύριων συνιστωσών αποτελεί ίσως τον πιο ευρέως διαδεδομένο αλγόριθμο για την μείωση διάστασης. Στόχος της είναι να αποτυπώσει βέλτιστα την διακύμανση των δεδομένων σε ένα μικρότερο *r*-διάστατο χώρο, ελαχιστοποιώντας το μέσο τετραγωνικό σφάλμα [20]. Για ένα *N*-διάστατο χώρο μπορεί να προκύψουν *N* διαφορετικές προβαλλόμενες ευθείες *PCs*, αν το πλήθος τους είναι μικρότερο απ' το αντίστοιχο πλήθος δειγμάτων [20]. Στην αντίθετη περίπτωση το πλήθος των κύριων συνιστωσών ισούται με το πλήθος εγγραφών των δεδομένων [20]. Κάθε μία από τις συνιστώσες αποτυπώνουν ένα ποσοστό της συνολικής διακύμανσης. Γίνεται κατανοητό, λοιπόν, πως επιλέγουμε τον *r*-διάστατο χώρο προβολής παρακολουθώντας πόσες *PCs* μας προσδίδουν υψηλότερη αναπαράσταση των δεδομένων. Συνήθως, η *PCA* χρησιμοποιείται σαν προεπεξεργαστικό εργαλείο, με σκοπό την διοχέτευση των νέων δεδομένων σε πιο σύνθετες μεθόδους(π.χ. *t-SNE*) [11]. Στην προκειμένη περίπτωση εξετάζουμε τον αντίκτυπο της *PCA* στα δεδομένα που διαθέτουμε και προσπαθούμε να ερμηνεύσουμε τα αποτελέσματα, παρακολουθώντας τις κύριες συνιστώσες και την απόδοση της οπτικοποίησης των σημείων.

Η αρχική προσέγγιση φανερώνει το ποσοστό της διακύμανσης που περιγράφουν οι κύριες συνιστώσες για να γίνει διακριτό το πλήθος το οποίο χρειαζόμαστε για να περιγράψουμε τα δεδομένα, με όσο το δυνατό ελάχιστη απώλεια πληροφορίας. Συνήθως, θεωρείται ικανοποιητικό να επιλέγουμε κύριες συνιστώσες, οι οποίες περιγράφουν τουλάχιστον το 85 με 90 τοις εκατό της συνολικής διακύμανσης. Παρακάτω φανερώνεται μία αρχική μείωση διάστασης για τις 10 πρώτες κύριες συνιστώσες στα δύο πακέτα δεδομένων που διαθέτουμε. Τα διαγράμματα προσπαθούν να φανερώσουν το εύρος διακύμανσης στις 10 πρώτες κύριες συνιστώσες μέσω γραφημάτων αθροιστικής γραμμής και ράβδων.



Πίνακας 2 Αποτίμηση συνολικής διακύμανσης απ' τις 10 πρώτες κύριες συνιστώσες στο καρκινικό dataset (πάνω) και στο dataset με καλλιέργειες κορωνοϊού(κάτω).

Η πρώτη κύρια συνιστώσα αποτυπώνει το μεγαλύτερο ποσοστό τις συνολικής διακύμανσης σε σχέση με τις άλλες συνιστώσες. Τα παραπάνω διαγράμματα παρουσιάζουν ότι για το *SARSCovid Dataset* η πρώτη κύρια συνιστώσα αντιπροσωπεύει το 20 τοις εκατό της συνολικής διακύμανσης, ποσοστό ιδιαίτερα μέτριο. Μία προβολή στο δισδιάστατο ή τρισδιάστατο χώρο θα μας έδινε μόνο το 38 και 53 τοις εκατό της συνολικής διακύμανσης (στην βέλτιστη περίπτωση). Η επιλογή σε 10 διαστάσεις συντηρεί περίπου το 70 τοις εκατό, το οποίο είναι σχετικά κοντά στο κατώφλι που ορίσαμε, αλλά ταυτόχρονα θα θεωρηθεί ανεπαρκής.

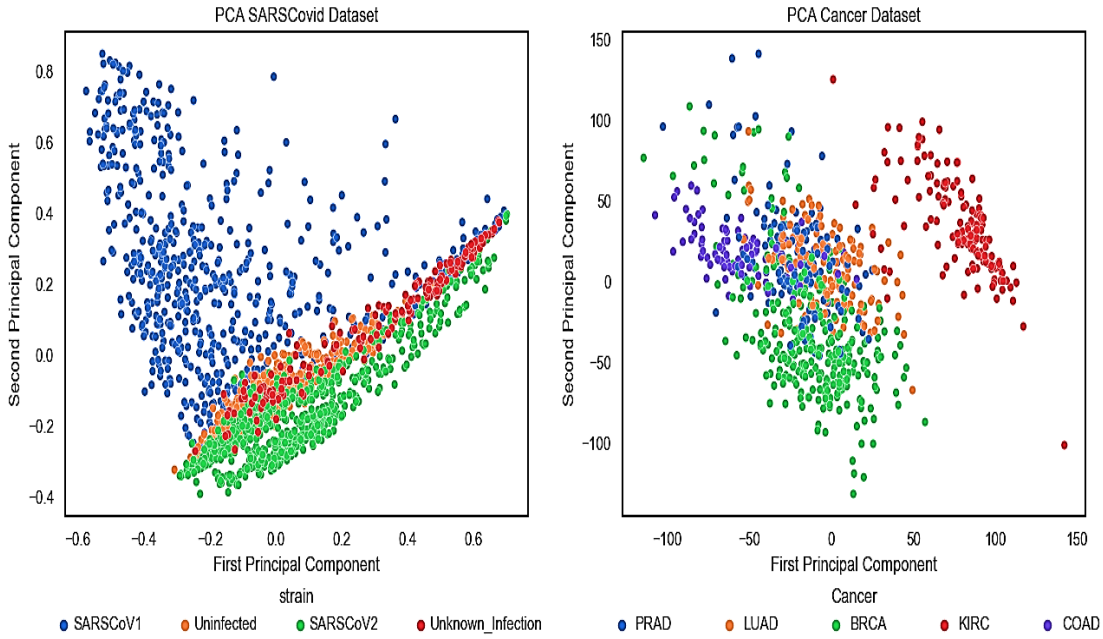
Αντίστοιχα, για το σετ δεδομένων *Cancer Dataset* τα παραπάνω διαγράμματα παρουσιάζουν ότι η πρώτη κύρια συνιστώσα αντιπροσωπεύει το *10.5* τοις εκατό της συνολικής διακύμανσης, ποσοστό ιδιαίτερα χαμηλό. Μία προβολή στο δισδιάστατο ή τρισδιάστατο χώρο θα μας έδινε μόνο το *19.3* και *26.1* τοις εκατό της συνολικής

διακύμανσης (στην βέλτιστη περίπτωση). Ακόμα και η επιλογή σε 10 διαστάσεις θεωρείται ανεπαρκής καθώς δεν συντηρεί ούτε το 50 τοις εκατό.

Κατά πάσα πιθανότητα, η οπτικοποίηση και των δύο συνόλων δεδομένων στον δισδιάστατο χώρο δεν θα μπορεί να εξάγει μοτίβα στα δεδομένα, ούτε θα αναπαριστά την πραγματική διακύμανση των σημείων, καθώς υπάρχει μεγάλη απώλεια πληροφορίας.

Στην παρακάτω εικόνα *4.3.1* η αναπαράσταση αφορά την προβολή των σημείων στο δισδιάστατο χώρο χρησιμοποιώντας τις πρώτες δύο κύριες συνιστώσες. Σε κάθε διάγραμμα διασποράς χρωματίζουμε τα σημεία καθώς προβάλλονται στον χώρο βάσει της αληθινής τους κατηγορίας, όπως αυτή περιγράφεται στα δύο σύνολα δεδομένων που κατέχουμε.

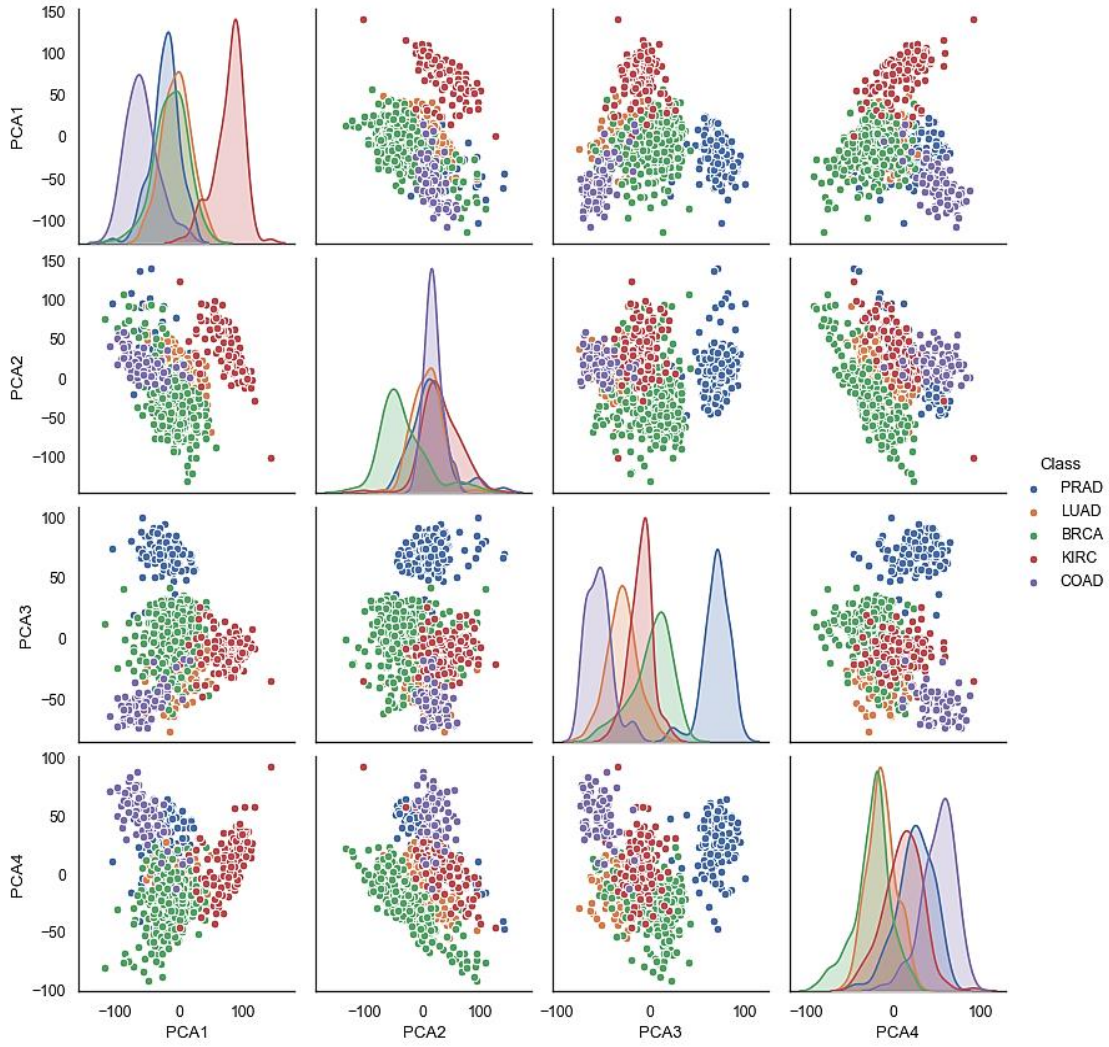
Όπως προαναφέρθηκε, δεν αναμενόταν κάποιο ικανοποιητικό αποτέλεσμα, καθώς το ποσοστό της συνολικής διακύμανσης των σημείων που αναπαράγουμε στο χώρο δεν επαρκεί να φανερώσει την πραγματική ροή των δεδομένων και να συντηρήσει την αρχική τους δομή. Στην παρακάτω εικόνα, καθώς χρωματίζουμε τα σημεία με βάση τις κλάσεις που ανήκουν, φανερώνεται ότι τα σημεία των κατηγοριών *SARSCoV2, Uninfected* και *Unknown_Infection*, επικαλύπτονται μεταξύ τους, ενώ δεν δημιουργούν καθαρές «ομάδες». Επιπλέον, τα σημεία της κατηγορίας *SARSCoV1* διασπείρονται προς πάσα κατεύθυνση, χωρίς παρ' όλα αυτά να μπορέσουμε να διακρίνουμε αν όντως περιέχουν διαφοροποιήσεις σε επίπεδο γονιδιακής έκφρασης με τις άλλες τρεις κατηγορίες. Βάσει του παρακάτω διαγράμματος διασποράς δύσκολα θα μπορούσαμε να πούμε ότι τα σημεία της προαναφερθείσας κλάσης εμφανίζουν ομοιότητες, αφού δεν δημιουργείται κάποια συμπαγής δομή.



Εικόνα 4.3.1 Διάγραμμα διασποράς δύο πρώτων κύριων συνιστωσών σε κανονικοποιημένα δεδομένα για το SARSCovid Dataset(GSE148729) (αριστερά) και Cancer Dataset (δεξιά)

Για το *Cancer Dataset*, καθώς χρωματίζουμε τα σημεία βάση τις κλάσεις που ανήκουν, φανερώνεται ότι τα σημεία των κατηγοριών *BRCA, PRAD,LUAD,COAD* επικαλύπτονται μεταξύ τους. Το γεγονός ότι η κατηγορία *KIRC* φαίνεται να ξεχωρίζει και να ομαδοποιείται απόμακρα από τις άλλες κλάσεις ίσως να είναι τυχαίο και έτσι δεν μπορεί να ληφθεί ασφαλές συμπέρασμα. Σε περίπτωση που προβάλλαμε άλλες συνιστώσες μεταξύ τους πιθανότατα κάποιες άλλες ομάδες να φαίνονταν να ξεχωρίζουν, όπως η κλάση *KIRC* σε αυτό το παράδειγμα. Παρ' όλα αυτά το αποτέλεσμα δεν επαρκεί να αποτυπώσει την αληθινή διακύμανση των δεδομένων και έτσι κανένα συμπέρασμα δεν είναι ασφαλές. Στην εικόνα *4.3.2* παρακολουθούμε τα αποτελέσματα διαφόρων συνδυασμών των τεσσάρων πρώτων κύριων συνιστωσών, επιβεβαιώνοντας την διαφορετική προβολή των σημείων στον χώρο.

Καθώς συνδυάζουμε διαφορετικές συνιστώσες δεν παρατηρείται κάποια αλλαγή, παρά μόνο ότι υπάρχει άλλη ομάδα να ξεχωρίζει διακριτά από τις άλλες σε σχέση με την αρχική οπτικοποίηση. Αυτή η ομάδα είναι η *PRAD*, αλλά και πάλι τονίζεται ότι δεν μπορεί να βγει κάποιο ασφαλές συμπέρασμα, καθώς μειώνουμε ακόμα περισσότερο την αποτίμηση στην συνολική διακύμανση στις κύριες συνιστώσες.

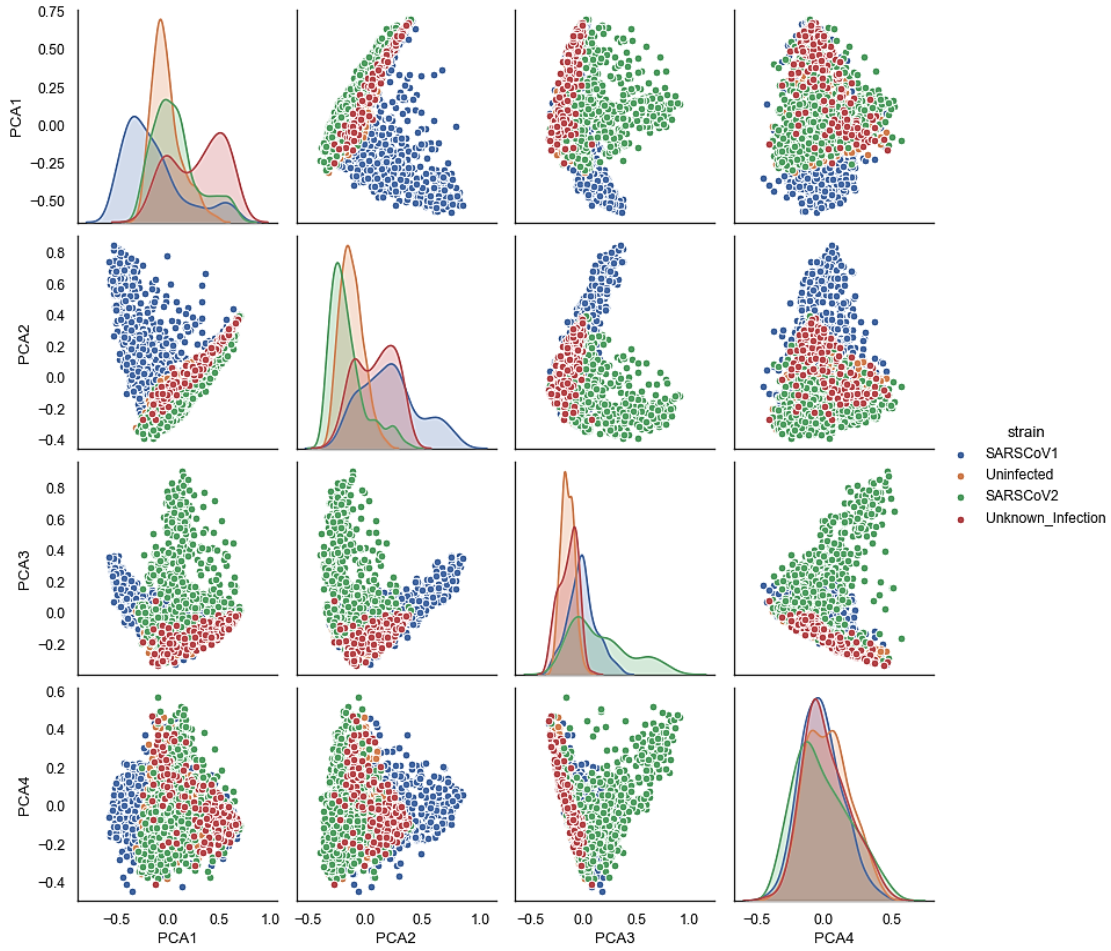


Εικόνα 4.3.2 Διαγράμματα διασποράς των τεσσάρων πρώτων συνδυαστικών προβολών για το Cancer Dataset (30% διακύμανσης)

Για να περιγράψουμε ένα επαρκές ποσοστό της συνολικής διακύμανσης και να έχουμε όσο το δυνατόν καλύτερη διατήρηση της αρχικής δομής χρειαζόμαστε τουλάχιστον 183 κύριες συνιστώσες για το 80 τοις εκατό και 800 για την βέλτιστη αναπαράσταση.

Επαναλαμβάνουμε την ίδια διαδικασία συνδυάζοντας διαφορετικές συνιστώσες και για το πακέτο δεδομένων με τις περιπτώσεις κορωνοϊού. Δεν παρατηρείται κάποια αλλαγή, τα σημεία συνεχίζουν να συγκεντρώνονται μεταξύ τους χωρίς να αποκαλύπτουν κάποιο ενδιαφέρον μοτίβο. Ταυτόχρονα, εμφανίζουμε την κατανομή πυκνότητας των σημείων ανά κατηγορία στις τέσσερις διαφορετικές συνιστώσες. Οι κατανομές φαίνεται να συγκλίνουν στην τέταρτη συνιστώσα, με αποτέλεσμα η προβολή με την συγκεκριμένη *PC* να εμφανίζει ακόμα πιο συμπυκνωμένα σημεία και ακόμα πιο δυσδιάκριτες διαφοροποιήσεις ανά κλάση. Για

να περιγράψουμε ένα επαρκές ποσοστό της συνολικής διακύμανσης και να έχουμε όσο το δυνατόν καλύτερη διατήρηση της αρχικής δομής χρειαζόμαστε τουλάχιστον 151 κύριες συνιστώσες για το 80 τοις εκατό και 1480 για την βέλτιστη αναπαράσταση. Γίνεται κατανοητό, λοιπόν, ότι η οπτικοποίηση των δεδομένων μέσω της *PCA* δεν μπορεί να μας δώσει ασφαλή συμπεράσματα.



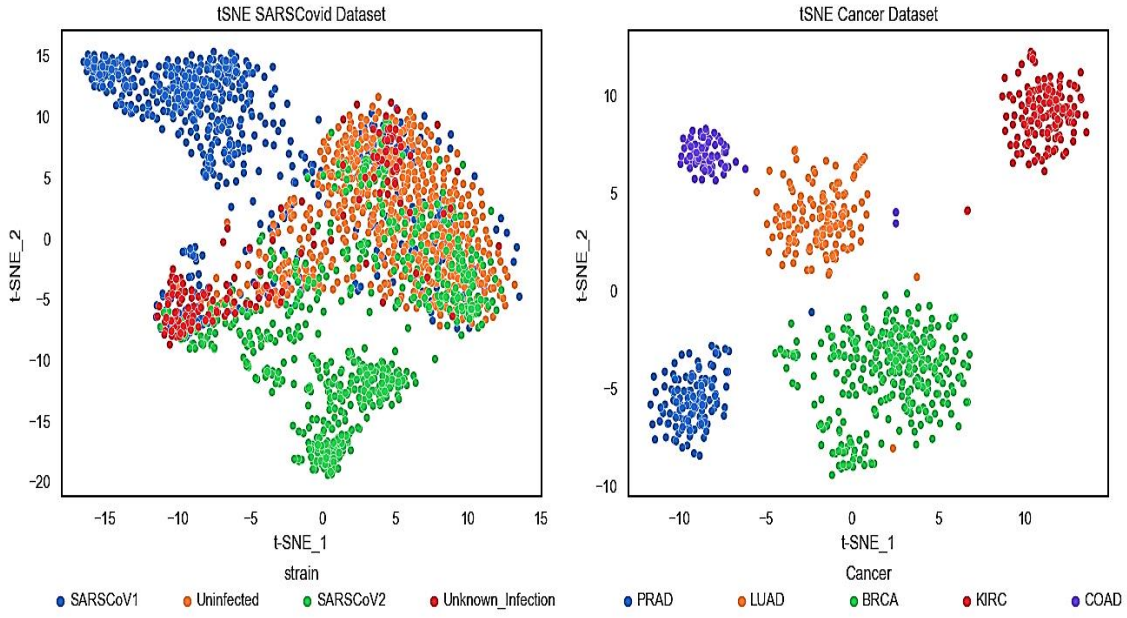
Εικόνα 4.3.3 Διαγράμματα διασποράς των τεσσάρων πρώτων συνδυαστικών προβολών (65% διακύμανσης)

4.3.2 Εφαρμογή t (κατανομή) - Στοχαστική Αφομοίωση Γειτονικών σημείων και αποτελέσματα

Ο *t-SNE* αποτελεί μία σύγχρονη αλγοριθμική τεχνική για την μείωση διάστασης σε δεδομένα με υψηλό όγκο πληροφορίας. Συνιστά σημαντικό εργαλείο στην

οπτικοποίηση των σημείων από υψηλότερης σε χαμηλότερης διάστασης χώρο (συνήθως δισδιάστατο και τρισδιάστατο), ενώ σε αντίθεση με την *PCA* καταφέρνει να επεξεργάζεται δεδομένα με πιο πολύπλοκη ενσωμάτωση στον χώρο [21]. Ο *t-SNE* αποτελεί πλέον από τα πιο ευρέως διαδεδομένα εργαλεία στην βιοπληροφορική και ιδιαίτερα στις αναλύσεις *RNA-seq* για την επίτευξη της μείωσης διαστατικότητας και της παρακολούθησης της διακύμανσης των σημείων στον ύστερα από την προβολή δισδιάστατο ή τρισδιάστατο άξονα [11]

Η λειτουργία του *t-SNE* εξαρτάται σε μεγάλο βαθμό από τον τρόπο με τον οποίο χειριζόμαστε τις παραμέτρους που τον απαρτίζουν. Η παραμετροποίηση του *t-SNE* δεν εξετάζεται ως βάθος στην παρούσα εργασία, με αποτέλεσμα να χρησιμοποιούνται οι προκαθορισμένες ρυθμίσεις από την υλοποίηση που υπάρχει στο *sklearn* (*learning_rate: 200.0* , *early_exaggeration: 12.0 κ.ο.κ*), ενώ ακόμα για το *perplexity* επιλέγουμε την προκαθορισμένη τιμή της υλοποίησης του *sklearn* που ισούται με 100.



Εικόνα 4.3.4 Διάγραμμα διασποράς t-SNE σε δύο διαστάσεις (perplexity = 100) για το SARSCovid Dataset(GSE148729) (αριστερά) και Cancer Dataset (δεξιά).

Τα παραπάνω διαγράμματα διασποράς(*4.3.4*) αφορά την απεικόνιση των σημείων στον χώρο, με ταυτόχρονο χρωματισμό βάσει κλάσης. Ο αλγόριθμος επιτυγχάνει να οπτικοποιήσει την προβολή των σημείων σε δύο διαστάσεις παρουσιάζοντας ένα πιο ευδιάκριτο διαχωρισμό κλάσεων. Η αναπαραγωγή αλγορίθμου για διάφορες παραμέτρους αποτελεί γενική συμβουλή για την καλύτερη θεώρηση του αλγορίθμου, ενώ σημαντικό στοιχείο είναι ότι εξ ορισμού ο *t-SNE*

υπολογίζει και καθορίζει την ομοιότητα των κοντινότερων σημείων μεταξύ τους, προκαλώντας απώλεια σε πολλές περιπτώσεις της αρχικής και πραγματικής δομής των σημείων [21][22][24].

Για το *Cancer Dataset* παρατηρούμε ότι τα σημεία ομαδοποιούνται σε ικανοποιητικό βαθμό ανά κατηγορία, χωρίς να υπάρχει επικάλυψη μεταξύ διαφορετικών κατηγοριών. Στην κατηγορία *BRCA* διακρίνεται ότι κάποια σημεία σχηματίζουν μία «υποσυστάδα» υποδηλώντας πιθανότατα ότι δεν ευθύνονται ίσως σε όλες τις περιπτώσεις καρκίνου του μαστού τα ίδια γονίδια. Ταυτόχρονα, δύο σημεία της κατηγορίας *LUAD* φαίνεται ότι διαθέτουν περισσότερες ομοιότητες με τα σημεία της κατηγορίας *BRCA* καθώς ομαδοποιούνται μαζί τους (π.χ. δείγμα 129). Ταυτόχρονα, δύο σημεία για την κατηγορία *COAD* και ένα δείγμα για τις *PRAD, KIRC* διαχέονται «άτακτα» στον χώρο, παρουσιάζοντας ίσως ανεξάρτητη φύση, σε σχέση με τα στοιχεία όμοιας κλάσης. Ο αλγόριθμος του *t-SNE* εξαιτίας της μαθηματικής του θεώρησης καταφέρνει να αποτυπώσει με μεγαλύτερη ακρίβεια τις συσχετίσεις των σημείων μέσα σε μία συστάδα(τοπικά), χωρίς να μπορεί να εξηγήσει σε πολλές περιπτώσεις τις «διασυσταδικές» σχέσεις. Αν και σε όλες τις περιπτώσεις διαχωρίζονται τα σημεία ανά κατηγορία καρκίνου σε ικανοποιητικό βαθμό, φαίνεται οι κλάσεις *BRCA,LUAD* και *PRAD* να έρχονται κοντά. Οι αποστάσεις στην πραγματικότητα ίσως να μην σημαίνουν τίποτα και έτσι δεν μπορούμε να αναγνωρίσουμε εύκολα, αν παραδείγματος χάριν η κατηγορία *BRCA* έχει περισσότερες ομοιότητες με την κατηγορία *LUAD* και όχι με την κατηγορία *KIRC*. Σε γενικές γραμμές ο αλγόριθμος καταφέρνει να ομαδοποιήσει τα σημεία ανά κατηγορία σε ικανοποιητικό βαθμό και πιθανότατα τα γονίδια που εκφράζουν κάθε μορφή καρκίνου να ξεχωρίζουν ανά κατηγορία, διαθέτοντας τα χαρακτηριστικά που επισημάναμε παραπάνω.

Σίγουρα σε σχέση με το προηγούμενο *dataset*, για το *SARSCovid* οι «συστάδες» που προκύπτουν δεν ξεχωρίζουν τόσο πολύ η μία από την άλλη. Σαφώς, παρακολουθώντας το παραπάνω διάγραμμα διασποράς μπορούμε να διακρίνουμε την κατηγορία των σημείων *SARSCoV1* να ομαδοποιείται ξεχωριστά κατά κύριο λόγο από τις υπόλοιπες κατηγορίες. Το ομόλογο στέλεχος κορωνοϊού παρουσιάζει σημεία που ξεχωρίζουν από τις υπόλοιπες κατηγορίες, αλλά εν αντιθέσει με την *SARSCoV1*, η «συστάδα» που δημιουργείται δεν είναι ιδιαίτερα συμπαγής. Τουναντίον, διακρίνεται η διάσπαση σε δύο ξεχωριστές ομάδες. Τα μη μολυσμένα κύτταρα (πορτοκαλί χρώμα) δημιουργούν μία μεγάλη ομάδα, πάνω στην οποία όμως έλκονται πολλά δείγματα που ανήκουν στις άλλες κλάσεις, αλλά ως επί το πλείστον στην κατηγορία *SARSCoV2*. Σε

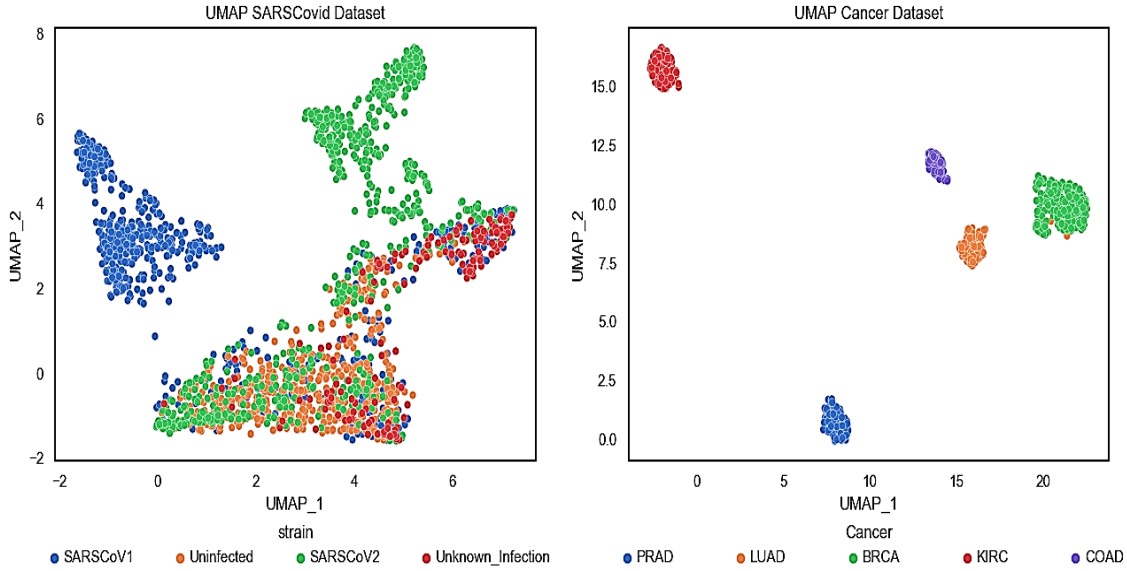
γενικές γραμμές υπάρχουν αρκετά σημεία των δύο στελεχών του ιού, τα οποία διαχέονται σε όλο το εύρος του χώρου και κυρίως του *SARSCoV2*. Το πιο σύνθετο μέρος ίσως να αποτελεί το σημείο τομής πολλών κατηγοριών, κυρίως του μη γνωστού είδους μόλυνσης (κόκκινο χρώμα) και *SARSCoV1, SARSCoV2*. Το παραπάνω διάγραμμα δύσκολα μας παρέχει πληροφορίες για το αν υπάρχουν διακριτές ομάδες, κάποιος θα μπορούσε να εντοπίσει 4 «συστάδες» ομοιότητας, κάποιος άλλος ακόμα και 6 θεωρώντας ότι υπάρχουν πιο ακανόνιστες διαβαθμίσεις.

Ο *t-SNE* λειτουργεί εξ ορισμού δίνοντας βαρύτητα στις τοπικές συσχετίσεις των σημείων [21][28]. Το γεγονός αυτό μας ενθαρρύνει να διατυπώσουμε την άποψη ότι τα σημεία μέσα σε μία «συστάδα» είναι σίγουρα αρκετά όμοια μεταξύ τους, αλλά δύσκολα μπορούμε να κρίνουμε τις «διασυσταδικές» σχέσεις [24][28]. Η πλειοψηφία των δειγμάτων *SARSCoV1* παρουσιάζει ξεκάθαρη ομοιότητα μεταξύ τους, ενώ το ίδιο ισχύει για τον *SARSCoV2* και τα μη μολυσμένα κύτταρα. Αυτό πιθανότατα να μετουσιώνεται σε ιδιαίτερες διαφορές σε επίπεδο γονιδιακής έκφρασης. Παρ' όλα αυτά δεν μπορούμε να εκφέρουμε άποψη αν η «συστάδα» με την κατηγορία *SARSCoV1* είναι πιο κοντά σε ομοιότητα εξαιτίας της θέσης της με αυτή της *Uninfected*. Αυτή η αδιευκρίνιστη συνθήκη δυσχεραίνει το έργο μας, ώστε να μπορέσουμε να ερμηνεύσουμε το τι συμβαίνει στην ιδιαίτερη σύγκλιση των σημείων σε εύρος *(-10,5)* στον άξονα *t-SNE_1* και *(-10, 0)* στον άξονα *t-SNE_2*.

4.3.3 Εφαρμογή Ομοιόμορφης Προσέγγισης και Προβολής Πολύπτυχου Μορφώματος και αποτελέσματα

Ο *UMAP* τα τελευταία χρόνια αποτελεί ένα πολύ ευέλικτο εργαλείο στην μείωση διάστασης και την οπτικοποίηση της διακύμανσης των δεδομένων στον χώρο και χρησιμοποιείται ιδιαίτερα για την ανάλυση *scRNA Seq* [11]. Ο *UMAP,* εν αντιθέσει με τον *t-SNE,* βοηθά στην συντήρηση της ολικής δομής των δεδομένων, καθώς κάνει την προβολή σε πολλή χαμηλότερης διάστασης χώρο σε σχέση με την αρχική διαστρωμάτωση [25][28]. Όπως αναφέρθηκε προηγουμένως, τόσο ο *UMAP*, όσο και *t-SNE* εξαρτώνται από τον τρόπο με τον οποίο χειρίζονται τις παραμέτρους [21][25]. Στην περίπτωση μας δεν γίνεται εκτεταμένη ανάλυση των παραμέτρων και έτσι επιλέγεται για την πιο σημαντική υπερπαράμετρο των κοντινότερων γειτόνων η τιμή 50. Σημειώνεται ότι η τιμή που χρησιμοποιείται για την παράμετρο *min_distance* είναι η προκαθορισμένη(*default*) και εκτιμάται σε *0.1*, έχοντας ως αποτέλεσμα να

δημιουργεί όσο το δυνατόν πιο «σφιχτές συστάδες» [26]. Τονίζεται ότι για την εφαρμογή του αλγορίθμου χρησιμοποιούνται εντολές από το *API* του δημιουργού σε γλώσσα *Python* [26]. Έτσι, λοιπόν, προβάλλουμε τα σημεία σε δύο διαστάσεις και τα χρωματίζουμε βάση των διαφορετικών κλάσεων.



Εικόνα 4.3.5 Διάγραμμα διασποράς μετά από μείωση διάστασης UMAP για το SARSCovid Dataset(GSE148729) (αριστερά) και Cancer Dataset (δεξιά).

Η διαφορετική θεωρητική προσέγγιση του *UMAP* προσφέρει διαφορετικά πλεονεκτήματα έναντι του *t-SNE*. Συνήθως, ο αλγόριθμος αυτός προσπαθεί να ισορροπήσει την συσχέτιση των σημείων σε τοπικό επίπεδο και στο συνολικό εύρος της αρχικής δομής [25][28]. Το πιο σύνθετο μαθηματικό υπόβαθρο του *UMAP* τον καθιστά πιο έγκυρο όσον αφορά την διατήρηση πληροφορίας των αρχικών αποστάσεων μεταξύ των σημείων. Συχνά, όσο αυξάνουμε την παράμετρο των κοντινότερων γειτόνων, τόσο πιο ρεαλιστική αναπαράσταση επιτυγχάνεται [26].

Παρατηρώντας το παραπάνω πλέγμα (*4.3.5*) διακρίνουμε για το δεύτερο σύνολο δεδομένων (δεξιά) την δημιουργία πέντε «συστάδων», όσες δηλαδή και οι διαφορετικές κλάσεις που διαθέτουμε. Η κατηγορία *KIRC* απομονώνεται σε κάθε προσέγγιση και ενδέχεται να παρουσιάζει αρκετές διαφορές στα γονίδια που ευθύνονται γι' αυτήν. Επιπροσθέτως, οι κατηγορίες *BRCA*,*LUAD* και *COAD*, (με κάθε επιφύλαξη ως προς τον τρόπο με τον οποίο ο *UMAP* ερμηνεύει τις «διασυσταδικές» αποστάσεις) ενδέχεται να συγκλίνουν η μία με την άλλη, ενώ ταυτόχρονα δύο σημεία που ανήκουν στην κατηγορία *LUAD* εμφανίζουν περισσότερες ομοιότητες με την κατηγορία *BRCA*. Τέλος, η κατηγορία *PRAD* καταφέρνει, επίσης να παραμείνει διαχωρισμένη. Ενδέχεται, λοιπόν, τα γονίδια που εκφράζουν τα είδη καρκίνων για το

PRAD και *KIRC* να εμφανίζουν μεγαλύτερες διαφοροποιήσεις σε σχέση με τα άλλα τρία είδη, ενώ είναι άκρως θετικό το γεγονός ότι διαφαίνονται συμπαγείς δομές σε κάθε είδους κλάση που διακρίνεται.

Σε αντίθεση με το προαναφερθέντα σύνολο δεδομένων, η οπτικοποίηση στο ομόλογο *dataset* δεν είναι αντίστοιχα καλή. Μελετώντας το παραπάνω διάγραμμα διασποράς δεν μπορούμε να πούμε ότι διαφέρει απόλυτα, με το αντίστοιχο του στον *t-SNE*. Ο *UMAP* κατορθώνει να διαχωρίσει τα σημεία που ανήκουν στις κατηγορίες *SARSCoV1* και *SARSCoV2*. Εδώ, ο *SARSCoV1* (μπλε χρώμα) φαίνεται να παρουσιάζει δύο υποομάδες, φανερώνοντας ότι ναι μεν πιθανότατα διαφοροποιούνται σε επίπεδο γονιδιακής έκφρασης από τις υπόλοιπες κατηγορίες, αλλά και ότι ίσως υπάρχουν και εσωτερικές διαβαθμίσεις μεταξύ των γονιδιών. Αντίστοιχα, κάποια σημεία του *SARSCoV2* διαχωρίζονται, αλλά η σύσταση τους είναι ακανόνιστη. Ταυτόχρονα, πολλά σημεία κατηγορίας *SARSCoV2-Uninfected* και *Unknown_Infection-SARSCoV1-SARSCoV2* «ομαδοποιούνται» μαζί . Πιθανότατα, λοιπόν, τα γονιδιακά μοτίβα αρκετών σημείων που έχουν μολυνθεί με τα δύο στελέχη του ιού και αυτά των μη μολυσμένων κλάσεων να μην παρουσιάζουν εκτεταμένες διαφορές. Σ*'* αυτή την περίπτωση η ασάφεια στο πως αντιλαμβανόμαστε τις ομοιότητες των «συστάδων» ίσως να προκαλέσει σύγχυση σε εφαρμογές κατηγοριοποίησης. Δηλαδή, μη μολυσμένα κύτταρα εύκολα να προβλέπονται ως μολυσμένα και κυρίως με στέλεχος του ιού *SARSCoV2,* ενώ στην περίπτωση της άγνωστης επιμόλυνσης κυττάρων παρατηρείται ότι εύκολα μπορεί να αναληφθούν ως περιπτώσεις *SARSCoV2* (κυρίως) και *SARSCoV1*. Τονίζεται, βέβαια, ότι στις περιπτώσεις άγνωστης μόλυνσης δεν γνωρίζουμε στην πραγματικότητα αν αποτελούν και περιπτώσεις *SARSCoV1* ή *SARSCoV2*. Αποτελούν κύτταρα που γνωρίζουμε ότι μολύνθηκαν, αλλά τα αφήσαν χωρίς διάγνωση [36]. Μία καλύτερη προσέγγιση ίσως θα μπορούσε να μας δώσει μία καλύτερη βάση , ώστε να μπορούσαμε να υποθέσουμε ποιο στέλεχος έχει μεγαλύτερη πιθανότητα να εισχώρησε στα κύτταρα.

4.3.4 Εφαρμογή προτεινόμενων αλγορίθμων σε πειραματικά δεδομένα

Στη παρούσα εργασία έγινε γνωστό, αρχικά, ότι θα αναλυθεί η λειτουργία και θα αξιολογηθεί η πορεία δύο προτεινόμενων μεθόδων, του *RGt-SNE* και του *RPEV*. Οι δύο προτεινόμενες αλγοριθμικές μέθοδοι στα αρχικά τους βήματα παρουσιάζουν ομοιότητες με αποτέλεσμα η περιγραφή τους να γίνει παράλληλα. Για την υλοποίηση τους χρησιμοποιούμε πακέτο εντολών που διατίθενται στην βιβλιοθήκη *sklearn*, τόσο για τον υπολογισμό των κοντινότερων γειτόνων όσο και για τις πολλαπλές τυχαίες προβολές, την μέθοδο *Djikstra* και τις τελικές εκτελέσεις αλγορίθμων μείωσης διάστασης που αφορούν τον *t-SNE* (για το *RG-tSNE*) και τον *Multidiamensional Scaling* (για τον *RPEV*).

Αρχικά, τονίζεται ότι εδώ δεν θα εφαρμοστεί κανενός είδος κανονικοποίησης της αρχικής μήτρας δεδομένων με σκοπό την μελέτη των αλγορίθμων στην αρχική κατανομή των σημείων. Οι δύο αλγοριθμικές μέθοδοι βασίζονται στην συγκεντρωτική μέθοδο πολλαπλών τυχαίων προβολών για την μείωση διάστασης σε πρώτη φάση και την εύρεση των κοντινότερων γειτόνων για κάθε δείγμα [6][29]. Σκοπός είναι να δημιουργηθεί μία τοπολογία που θα περιγράφει την ομοιότητα των δειγμάτων μεταξύ τους, βασισμένη στον συνδυασμό των αποτελεσμάτων της *kNN* αναζήτησης [6][29].

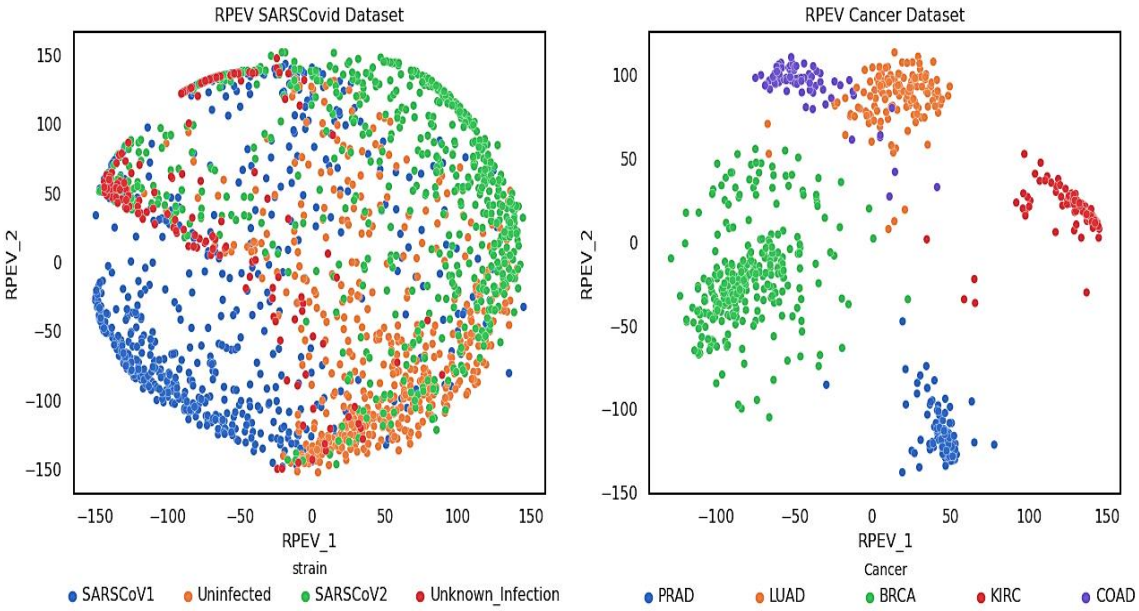
Οι δύο αλγοριθμικές προσεγγίσεις περιέχουν μία σειρά παραμέτρων, όπως το πλήθος επαναλήψεων τυχαίων προβολών, ο αριθμός κοντινότερων γειτόνων και το νέο πλήθος διαστάσεων. Στην παρούσα εργασία δεν γίνεται μελέτη της παραμετροποίησης των μεθόδων εις βάθος, παρ' όλα αυτά η εξαγωγή των αποτελεσμάτων έγινε ύστερα από εξαντλητική δοκιμή παραμέτρων. Τονίζεται ότι η τυχαία μήτρα επιλέχθηκε και στις δύο περιπτώσεις να ακολουθεί Γκαουσιανή κατανομή, καθώς βρέθηκε ότι διαχειρίζεται καλύτερα τα επόμενα αλγοριθμικά βήματα.

4.3.4.1 Οπτικοποίηση Συνόλου Τυχαίων Προβολών (Αλγόριθμος 4) και αποτελέσματα

Για τον SARSCovid Dataset: Βασικό είναι να καθορίσουμε τις βασικές παραμέτρους που ορίζονται ως πρώτο βήμα στον Αλγόριθμο 4. Αναθέτουμε σε 7 τον αριθμό νέων προβολών, σε 80 τον νέο υποχώρο, ενώ για να μπορέσουμε να εξάγουμε ένα αρκετά καλό αποτέλεσμα σε 117 τον αριθμό κοντινότερων γειτόνων σε ένα μετρικό σύστημα, που αφορά την συσχέτιση (*correlation*) των σημείων. Κατασκευάζονται λοιπόν 7 τυχαίες μήτρες Γκαουσιανής κατανομής, οι οποίες προβάλλονται με την αρχική μήτρα δεδομένων. Μεταβαίνουμε λοιπόν από τον χώρο των 17.934 διαστάσεων σε 80, εκμεταλλευόμενοι τα πλεονεκτήματα που μας δίνει το λήμμα των *Johnson-Lindestrauss* [31]. Ύστερα, διενεργείται η αναζήτηση 117 κοντινότερων γειτόνων στον προβαλλόμενο πίνακα εξάγοντας τα αποτελέσματα σε μορφή γράφου, που περιέχει μηδενικές τιμές για τα ασύνδετα δείγματα και ένα για τους κοντινότερους γείτονες. Η επαναληπτική άθροιση της μήτρας κατασκευάζει τον επιθυμητό πίνακα 2000×2000, ομοιότητας με μεγαλύτερη ευκολία και επιτάχυνση. Το μέγιστο εύρος βαθμού ομοιότητας ανέρχεται στον αριθμό επαναλήψεων, δηλαδή ίσο με 7. Αφού αντιστρέψουμε την μήτρα ομοιότητας, κατασκευάζουμε τον άνω συμμετρικό πίνακα αποστάσεων, ο οποίος διοχετεύεται στον *MDS*, διατηρώντας την μετρική στο *precomputed* και το *max_iteration* στο 1000.

Για τον Cancer Dataset: Αναθέτουμε σε 10 τον αριθμό νέων προβολών, σε 50 τον νέο υποχώρο, ενώ για να μπορέσουμε να εξάγουμε ένα αρκετά καλό αποτέλεσμα σε 53 το αριθμό κοντινότερων γειτόνων σε ένα μετρικό σύστημα, που αφορά τον Ευκλείδιο χώρο. Κατασκευάζονται λοιπόν 10 τυχαίες μήτρες Γκαουσιανής κατανομής, οι οποίες προβάλλονται με την αρχική μήτρα δεδομένων. Μεταβαίνουμε, λοιπόν, από τον χώρο των 20.264 διαστάσεων σε 50, εκμεταλλευόμενοι τα πλεονεκτήματα που μας δίνει το λήμμα των *Johnson-Lindestrauss* [31]. Ύστερα, διενεργείται η αναζήτηση 53 κοντινότερων γειτόνων στον προβαλλόμενο πίνακα εξάγοντας τα αποτελέσματα σε μορφή γράφου, που περιέχει μηδενικές τιμές για τα ασύνδετα δείγματα και ένα για τους κοντινότερους γείτονες. Η επαναληπτική άθροιση της μήτρας κατασκευάζει τον επιθυμητό πίνακα 801×801, ομοιότητας με μεγαλύτερη ευκολία και επιτάχυνση. Το μέγιστο εύρος βαθμού ομοιότητας ανέρχεται στον αριθμό επαναλήψεων, δηλαδή ίσο με 10. Αφού αντιστρέψουμε την μήτρα ομοιότητας, κατασκευάζουμε τον άνω

συμμετρικό πίνακα αποστάσεων, ο οποίος διοχετεύεται στον *MDS*, διατηρώντας την μετρική στο *precomputed* και το *max_iteration* στο 1000.



Εικόνα 4.3.5 Δισδιάστατο διάγραμμα διασποράς του RPEV χρωματισμένο βάσει ετικετών για το SARSCovid Dataset(GSE148729) (αριστερά) και Cancer Dataset (δεξιά).

Παρατηρούμε ότι στον συγκεκριμένο αλγόριθμο φανερώνοται δύο πιο ξεκάθαρα σύνολα σημείων για τις κατηγορίες *SARSCoV1* και *SARSCoV2*. Υπάρχει υψηλή συγκέντρωση διάσπαρτων σημείων από κατηγορίες *Uninfected* και *SARSCoV2* στις περιοχές *RPEV_1(0,100)* και *RPEV_2(-150,-50)*. Η κατηγορία *Uninfected* κατορθώνει να σχηματίσει ένα σύμπλοκο, το οποίο έλκει, όμως, αρκετά σημεία από τις κατηγορίες *SARSCoV2* και *Unkonown_Infection*. Τέλος, σχηματίζεται και μία διακριτή ομοιότητα μεταξύ δειγμάτων της κλάσης *Unknown_Infection* στην κορυφή του διαγράμματος.

Όσον αφορά το δεύτερο σύνολο δεδομένων, τα αποτελέσματα είναι πιο ενθαρρυντικά, καθότι υπάρχει ευδιάκριτη κατανομή των σημείων ανά κατηγορία στον χώρο. Οι κατηγορίες *BRCA, KIRC* και *PRAD* παρουσιάζουν έναν αρκετά σαφή διαχωρισμό, αν και πολλά από τα σημεία τους διαχέονται στον χώρο, κυρίως της *BRCA*. Από την άλλη μεριά, οι κλάσεις *LUAD* και *COAD*, αν και εμφανίζουν σαφείς «συστάδες», συγκλίνουν μεταξύ τους με ελάχιστα σημεία της κατηγορίας *COAD* να στοιχίζονται με αυτά της *LUAD*.

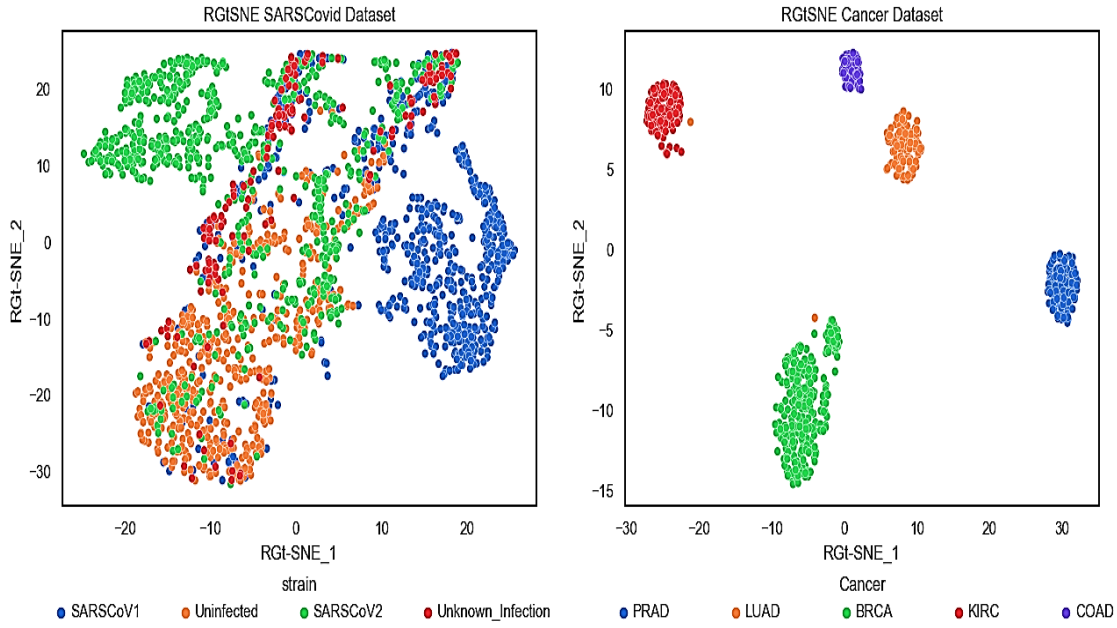
4.3.4.2 Μέθοδος Πολλαπλών Τυχαίων Προβολών και Κοντινότερων Αποστάσεων σε t κατανομή-Στοχαστική Αφομοίωση Γειτονικών Σημείων και αποτελέσματα

Για τον SARSCovid: Όπως αναφέρθηκε προηγουμένως, βασικό είναι να καθορίσουμε τις βασικές παραμέτρους που ορίζονται ως πρώτο βήμα στον Αλγόριθμο 3. Έτσι, λοιπόν, θέτουμε ως αριθμό επαναλήψεων ίσο με 7, ως διάσταση μείωσης μέσω τυχαίων προβολών ίση με 500 και ως αριθμό κοντινότερων γειτόνων ίσο με 5. Κατασκευάζονται, λοιπόν, 7 τυχαίες μήτρες Γκαουσιανής κατανομής, οι οποίες προβάλλονται με την αρχική μήτρα δεδομένων. Μεταβαίνουμε, λοιπόν, από τον χώρο των 17.934 διαστάσεων σε 500, εκμεταλλευόμενοι τα πλεονεκτήματα που μας δίνει το λήμμα των *Johnson-Lindestrauss* [31]. Ύστερα, διενεργείται η αναζήτηση 5 κοντινότερων γειτόνων στον προβαλλόμενο πίνακα εξάγοντας τα αποτελέσματα σε μορφή γράφου, που περιέχει μηδενικές τιμές για τα ασύνδετα δείγματα και ένα για τους κοντινότερους γείτονες. Η επαναληπτική άθροιση της μήτρας κατασκευάζει τον επιθυμητό πίνακα ομοιότητας με μεγαλύτερη ευκολία και επιτάχυνση. Το μέγιστο εύρος βαθμού ομοιότητας ανέρχεται στον αριθμό επαναλήψεων, δηλαδή ίσο με 7. Αφού αντιστρέψουμε την μήτρα ομοιότητας, διενεργούμε αναζήτηση συντομότερου μονοπατιού μέσω αλγορίθμου *Djikstra*, ο οποίος βάσει της συνάρτησης του *sklearn* αγνοεί τα μονοπάτια με μηδενικά βάρη ως ασύνδετα. Η τελική μήτρα αποστάσεων διοχετεύεται στον *t-SNE*, διατηρώντας την μετρική στο *precomputed* και το *perplexity* στο 100.

Για το Cancer Dataset: Θέτουμε ως αριθμό επαναλήψεων ίσο με 10, ως διάσταση μείωσης μέσω τυχαίων προβολών ίση με 50 και ως αριθμό κοντινότερων γειτόνων ίσο με 5. Κατασκευάζονται, λοιπόν, 10 τυχαίες μήτρες Γκαουσιανής κατανομής, οι οποίες προβάλλονται με την αρχική μήτρα δεδομένων. Μεταβαίνουμε από τον χώρο των 20.264 διαστάσεων σε 50, εκμεταλλευόμενοι τα πλεονεκτήματα που μας δίνει το λήμμα των *Johnson-Lindestrauss* [31]. Ύστερα διενεργείται η αναζήτηση 5 κοντινότερων γειτόνων στον προβαλλόμενο πίνακα εξάγοντας τα αποτελέσματα σε μορφή γράφου, που περιέχει μηδενικές τιμές για τα ασύνδετα δείγματα και ένα για τους κοντινότερους γείτονες. Η επαναληπτική άθροιση της μήτρας κατασκευάζει τον επιθυμητό πίνακα ομοιότητας με μεγαλύτερη ευκολία και επιτάχυνση. Το μέγιστο εύρος βαθμού ομοιότητας ανέρχεται στον αριθμό επαναλήψεων, δηλαδή ίσο με 10. Αφού αντιστρέψουμε την μήτρα ομοιότητας, διενεργούμε αναζήτηση συντομότερου

μονοπατιού μέσω αλγορίθμου *Djikstra*, ο οποίος βάσει της συνάρτησης του *sklearn* αγνοεί τα μονοπάτια με μηδενικά βάρη ως ασύνδετα. Η τελική μήτρα αποστάσεων διοχετεύεται στον *t-SNE*, διατηρώντας την μετρική στο *precomputed* και το *perplexity* στο 100.

Η τελική αναπαράσταση σε δύο διαστάσεις λαμβάνει την εξής μορφή:



Εικόνα 4.3.6 Δισδιάστατο διάγραμμα διασποράς του RG-tSNE χρωματισμένο βάσει ετικετών για το SARSCovid Dataset(GSE148729) (αριστερά) και Cancer Dataset (δεξιά).

Στο παραπάνω διάγραμμα διασποράς παρουσιάζεται η επιρροή της διαδικασίας επαναπροσδιορισμού των αποστάσεων των κυτταρικών δειγμάτων και η μετέπειτα εκτέλεση του *t-SNE*. Σε αυτή την προσέγγιση παρατηρούμε ότι ξανά αρκετά σημεία των κατηγοριών *SARSCoV1(μπλε)* και *SARSCoV2(πράσινο)* ξεχωρίζουν καθιστώντας σοβαρή την πιθανότητα η γονιδιακή έκφραση να διαφέρει στις δύο αυτές περιπτώσεις μόλυνσης. Από την άλλη, η κατηγορία των μη μολυσμένων κυττάρων(πορτοκαλί) παρουσιάζει μία πιο διάσπαρτη κατανομή σημείων στον χώρο. Σε σχέση με τις προηγούμενες προσεγγίσεις, τα περισσότερα μολυσμένα σημεία ανεξαρτήτου ιϊκού στελέχους πακετάρονταν με αυτά της κατηγορίας *Uninfected*. Αντίθετα, εδώ τα περισσότερα από αυτά τα σημεία αρχίζουν να διαχέονται στον χώρο, με την πλειοψηφία να σχηματίζει δύο διαφορετικές «ομαδοποιήσεις» στις περιοχές *[-10,10]* στον *t-SNE_1* και *[-15,10]* στον *t-SNE_2*. Στις περιοχές αυτές μαζεύονται δείγματα των τριών κατηγοριών μολυσμένων κυττάρων. Παράλληλα, είναι ξεκάθαρο πως υπάρχει μεγάλος αριθμός διάσπαρτων σημείων από τις ίδιες κατηγορίες, αλλά κυρίως από τις

Uninfected και *SARSSCoV2*. Τέλος, η εντύπωση που δίνεται είναι ότι παρά τις δύο διακριτές διαφοροποιήσεις σημείων ανά είδος ιού, εδώ μπορούμε να υποθέσουμε ότι σχηματίζονται υποομάδες στις ξεκάθαρες αυτές «συστάδες», υποδηλώνοντας ότι ενδεχομένως να υπάρχουν διαφορές στο πλαίσιο έκφρασης γονιδίων, έστω και σε ελάχιστο βαθμό ακόμα και σε ίδιες περιπτώσεις μόλυνσης.

Αντίθετα, οι ομαδοποιήσεις των καρκινικών κλάσεων γίνονται με ιδιαίτερη επιτυχία, καθώς προκύπτουν πέντε ευδιάκριτες ομάδες. Στην κατηγορία *BRCA* ίσως μπορούμε να διακρίνουμε δύο υποομάδες, ενώ ένα και μόνο σημείο της *LUAD* συγκλίνει κοντά τους. Γενικά, παρατηρούνται μόνο δύο σημεία της κλάσης *LUAD*, ένα κοντά στην *BRCA* και ένα κοντά στην *KIRC*.

4.4 Αξιολόγηση αλγοριθμικών τεχνικών

Αφού ερμηνεύσαμε και διερευνήσαμε τα διαγράμματα διασποράς, ύστερα από την μείωση διάστασης μέσω μίας σειράς αλγορίθμων ολοκληρώνουμε την ανάλυση τους, αξιολογώντας την απόδοση τους. Θέτοντας τα κατάλληλα κριτήρια αξιολόγησης μπορούμε να ορίσουμε το πλαίσιο με το οποίο θα αναγνωρίσουμε την πιο αποδοτική μέθοδο. Η έννοια της οπτικοποίησης, μαζί με τις υποθέσεις που προκύπτουν για την υφή και «φύση» των δειγμάτων, παρουσιάστηκε στις παραπάνω ενότητες. Επιλέγουμε, λοιπόν, να αξιολογήσουμε τα αποτελέσματα της μείωσης διάστασης με κατάλληλες μετρικές ομαδοποίησης που αφορούν την διαχωρισιμότητα των «συστάδων» και το πόσο συμπαγείς είναι [6][29]. Με αυτή την προσέγγιση, οι μετρικές που κατορθώνουν να πσοσοτικοποιήσουν τα δύο αυτά επιθυμητά χαρακτηριστικά είναι τα εσωτερικά μέτρα αξιολόγησης ομαδοποίησης. Τα εσωτερικά μέτρα δεν απαιτούν εξωτερική πληροφορία για τα δεδομένα, αλλά μελετούν τις «διασυσταδικές» και «ενδοσυσταδικές» σχέσεις [37]. Για να γίνει πιο κατανοητό, ως συνεκτικότητα στις ομάδες χαρακτηρίζουμε το πόσο κοντά στοιχίζονται τα «ενδοσυσταδικά» σημεία μεταξύ τους, ενώ ως διαχωρισιμότητα τον βαθμό με τον οποίο διακρίνεται ξεκάθαρα η μία «συστάδα» από την άλλη, δίνοντας την «διασυσταδική» συσχέτιση [37]. Οι υπολογισμοί βασίζονται στην μήτρα απόστασης ή αλλιώς ανομοιότητας των δειγμάτων [37]. Υπενθυμίζουμε ότι δεν πρόκειται για μελέτη αλγορίθμων ομαδοποίησης, άλλα

κάθε προσέγγιση και ιδίως οι αλγόριθμοι *RGt-SNE* και *RPEV* επιδιώκουν να κατασκευάσουν την πιο πιστή εκδοχή της πολυδιάστατης αναπαράστασης σε όσο το δυνατόν καλύτερη απεικόνιση, παρακολουθώντας τις τοπικές και ολικές βαθμίδες ομοιότητας των σημείων [6][29] .

Οι μετρικές που θα εξυπηρετήσουν τις παραπάνω συνθήκες συνιστούν ο *silhouette coefficient*, ίσως ο πιο κοινός τρόπος αξιολόγησης, ο δείκτης των *Davies-Bouldin* και ο δείκτης *Calinski-Harabasz*. Με τους δείκτες αυτούς προσπαθούμε να ποσοτικοποιήσουμε τις παραπάνω συνθήκες και να συνοψίσουμε την αποτελεσματικότητα των μεθόδων. Συγκεκριμένα, ο *silhouette coefficient* επιδιώκει να υπολογίσει την μέση απόσταση μεταξύ των συστάδων, ο δείκτης *Calinski-Harabasz* γνωστός και ως κριτήριο ποσοστιαίας διασποράς, λαμβάνει υπόψιν τον διασκορπισμό των σημείων μεταξύ δύο συστάδων και των «ενδοσυσταδικών» διακυμάνσεων αντίστοιχα [29] και ο *Davies-Bouldin* μελετά τον μέσο βαθμό ομοιότητας μεταξύ μίας συστάδας και της κοντινότερης της. Οι υπολογισμένες τιμές για κάθε μετρική ερμηνεύεται διαφορετικά. Για τον *silhouette coefficient* καλές θεωρούνται τιμές μεγαλύτερες του μηδενός και όσο πιο κοντά στο ένα, για το *Calinski-Harabasz* όσο μεγαλώνει το τελικό αποτέλεσμα τόσο καλύτερη απόδοση, ενώ αντίθετα για τον δείκτη *Davies-Bouldin* αναζητούνται τιμές που συγκλίνουν στο μηδέν [37]. Βέβαια, είναι σχεδόν ακατόρθωτο κάποιος αλγόριθμος να υπερτερεί και στους τρεις δείκτες [37]. Οι παρακάτω πίνακες φανερώνουν την απόδοση κάθε αλγορίθμου για κάθε μετρική, ενώ λόγω της χρονικής απόκλισης των τεχνικών αναγράφεται και η διάρκεια υλοποίησης. Αφού αναφερθήκαμε στην αρχή της παρούσας υποενότητας για την διαισθητική εντύπωση που μας προκαλούν τα διαγράμματα, θα μπορούσαμε να υποθέσουμε ότι δεν περιμένουμε ιδιαίτερα ενθαρρυντικά αποτελέσματα όσον αφορά τις αξιολογήσεις στο *SARSCovid Dataset*, καθώς στην πλειοψηφία των αναλύσεων αναφερθήκαμε για ακανόνιστες συσταδικές δομήσεις και για διφορούμενο πλήθος διακριτών ομαδοποιήσεων, ιδιαίτερα βάσει της κατηγορίας του εκάστοτε σημείου. Αντίθετα, στο *Cancer Dataset* οι ευδιάκριτες συστάδες θα μας προδιάθεταν για πιο ενθαρρυντικά αποτελέσματα.

4.4.1 Αξιολόγηση αλγοριθμικών τεχνικών για τις καλλιέργειες κορωνοϊού

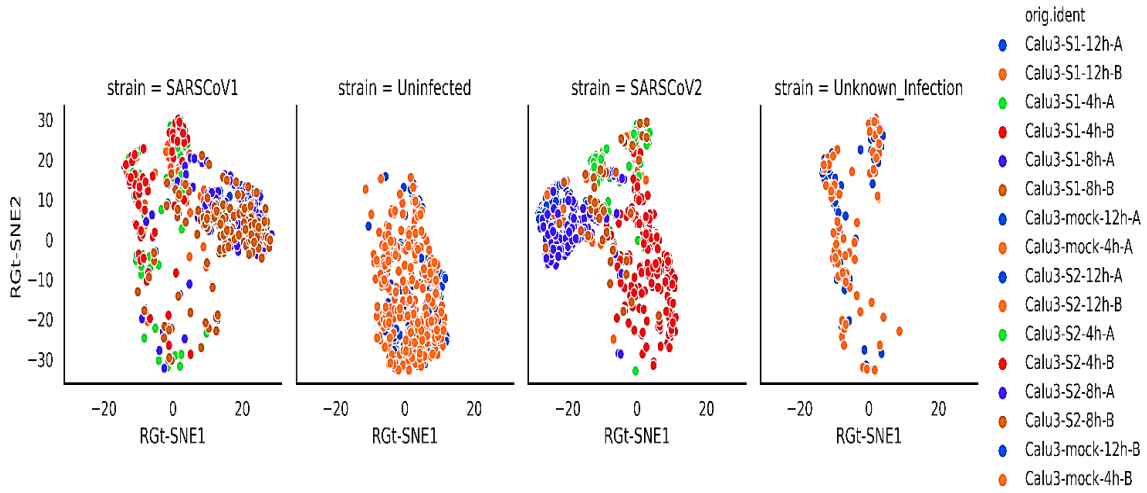
Αλγόριθμοι και μετρικές			
Αλγόριθμοι (χρόνος)	*Silhouette*	*Davies-Bouldin*	*Calinski-Harabasz*
PCA (6 *seconds*)	-0.020	2.65	315.54
t-SNE (4 *minutes*)	0.128	**1.773**	**577.34**
UMAP (45 *seconds)*	0.073	2.061	275.41
RGt-SNE (1.5 *min*)	**0.140**	3.586	440.14
RPEV (2 *minutes*)	0.100	2.580	370.45

Πίνακας 3 Αξιολόγηση βάσει μετρικών των αλγορίθμων μείωσης διάστασης για SARSCovid Dataset

Σύμφωνα με τον παραπάνω πίνακα, ο *RGt-SNE* αποδίδει καλύτερα στον δείκτη *Silhouette Coefficien*t, ενώ ο *t-SNE* τον ξεπερνάει στους άλλους δύο δείκτες. Δύσκολα μπορούμε να καθορίσουμε ποιος από τους δύο αλγορίθμους ταιριάζει στην προσέγγιση της καλύτερης μεθόδου μείωσης διάστασης. Βέβαια, υπόψιν πρέπει να ληφθεί ότι ο *RGt-SNE* επιταχύνει κατά 60% τον *t-SNE*, κάνοντας την εκτέλεση πολύ πιο γρήγορη. Δεδομένων των δεικτών, ο *RGt-SNE* φαίνεται ότι αντιλαμβάνεται καλύτερα τις ξεχωριστές «συστάδες», θεωρώντας πιο καθαρές τις μέσες αποστάσεις μεταξύ τους. Αντίθετα, ο *t-SNE* προσεγγίζει καλύτερα τον τρόπο με τον οποίο τα σημεία παρουσιάζουν διακύμανση μέσα στις ομάδες. Ο διασκορπισμός των σημείων είναι, δηλαδή, πιο ομαλός. Ο *PCA* παρουσιάζει την χειρότερη απόδοση, καθώς δυσκολεύεται να παραδώσει μία ικανή αποτίμηση της διασποράς σε μικρότερη διάσταση. Ίσως ως προεπεξεργαστικό εργαλείο να είναι πιο χρήσιμος και ιδιαίτερα σε κάποιας μορφής συνδυαστική ανάλυση, όπως στον *RPEV* και *RGt-SNE* [6][29]. Τέλος, σημαντικό είναι ότι ο *RPEV* φαίνεται να είναι ο τρίτος κατά σειρά αλγόριθμος με καλύτερη απόδοση, ξεπερνώντας ακόμα και τον *UMAP*.

Επιλέγοντας τον *RGt-SNE* δίνουμε μία νέα οπτική στον τρόπο που οπτικοποιούμε τα δεδομένα και παρατηρούμε ανά περίπτωση μόλυνσης την χρονική διάρκεια επώασης. Σύμφωνα, λοιπόν, με τις παρακάτω γραφικές παραστάσεις και σε συνδυασμό με το συνολικό διάγραμμα της υποενότητας *4.3.4.2* παρατηρούμε ότι κατά την πάροδο του χρόνου τα στελέχη του ιού ενδεχομένως να παρουσιάζουν όλο και περισσότερες διαφορές. Αγνοώντας την κατηγορία *Uninfected* που παρουσιάζει

σταθερή ροή, οι άλλες περιπτώσεις εμφανίζουν πιο ξεκάθαρες διαφορές ύστερα από 8 και 12 ώρες επώασης. Σε πιο πρώιμες χρονικές στιγμές τα σημεία των δύο στελεχών του ιού παρουσιάζουν σχετικά πιο συγκλίνουσες τοποθετήσεις στον χώρο και κυρίως με την περίπτωση του *Unknown_Infection*.



Εικόνα 4.4.1 Παρακολούθηση διασποράς για διάφορες χρονικές στιγμές επώασης σε κάθε κλάση για τον RGt-SNE

Συμπερασματικά, η συνολική διερεύνηση των γραφημάτων οπτικοποίησης σε συνδυασμό με τις μετρικές αξιολόγησης, μας δίνουν την δυνατότητα να υποθέσουμε ότι ενδεχομένως η πλειοψηφία των σημείων για *SARSCoV1*, *SARSCoV2* και *Uninfected* παρουσιάζουν εμφανείς διαφοροποιήσεις. Παρ' όλα αυτά, διακρίνεται και υψηλός βαθμός επικάλυψης μεταξύ κλάσεων *SARSCoV1*, *SARSCoV2* και *Unknown_Infection* καθιστώντας δύσκολη την διάκριση τους. Ως επί το πλείστον τα μολυσμένα κύτταρα πιθανότατα να εκφράζονται από διαφορετικά γονίδια, ανάλογα τον ιό από τον οποίο προσβάλλονται. Λαμβάνοντας υπόψιν και την παραπάνω συνθήκη που αναλύθηκε για τον χρόνο επώασης, τα κύτταρα που συλλέγονται μετά από 4 ώρες από την εισβολή του ιού, ίσως να διακρίνονται πολύ πιο δύσκολα. Σε αυτό το διάστημα φαίνεται παρακολουθώντας τις εικόνες *4.3.6* και *4.4.1* οι ισορροπίες διάκρισης μεταξύ των μολυσμένων κυττάρων να είναι πιο εύθραυστες. Πιθανόν, σε περιπτώσεις κατηγοριοποίησης και ακόμα και σε αλγόριθμους που βασίζονται στον υπολογισμό αποστάσεων, τα αποτελέσματα να είναι αρκετά καλά λαμβάνοντας υπόψιν ότι και ο *RGt-SNE* προσπαθεί να χρησιμοποιεί μία όσο το δυνατόν ρεαλιστική απεικόνιση του βαθμού ανομοιότητας των σημείων. Συνοψίζοντας, σύμφωνα με τα παραπάνω, ενδεχομένως να περιμέναμε σε μία εφαρμογή κατηγοριοποίησης υψηλό βαθμό επιτυχίας στην ανάθεση κλάσεων για τις ετικέτες *SARSCoV1*, *SARSCoV2* και

Uninfected με ένα μέτριο βαθμό λάθους αναγνώρισης μεταξύ τους, ενώ πιθανόν θα περιμέναμε χαμηλό βαθμό ακρίβειας στις δύο περιοχές που η κατηγορία *Unknown_Infection* δημιουργεί πυκνώσεις με τους *SARSCoV1* και *SARSCoV2* και ίσως για χαμηλές χρονικές στιγμές επώασης. Επιπρόσθετα, αλγόριθμοι που δεν αποδίδουν συνήθως καλά σε δεδομένα με υψηλή επικάλυψη σημείων ποικίλων κατηγοριών, ίσως το αποτέλεσμα να είναι ακόμα χειρότερο.

4.4.2 Αξιολόγηση αλγοριθμικών τεχνικών για τις περιπτώσεις καρκίνου

Αλγόριθμοι και μετρικές			
Αλγόριθμοι (χρόνος)	*Silhouette*	*Davies-Bouldin*	*Calinski-Harabasz*
PCA (1,7 *seconds*)	0.1471	2.0650	268.68
t-SNE (48 *seconds*)	0.6547	0.4600	2306.47
UMAP (24 *seconds)*	0.8049	**0.2323**	**18723.26**
RGt-SNE (16 *sec*)	**0.8224**	0.2440	9502.47
RPEV (1,5 *minutes*)	0.6718	0.3592	1928.68

Πίνακας 4 Αξιολόγηση βάσει μετρικών των αλγορίθμων μείωσης διάστασης για το Cancer Dataset

Οι παραπάνω μετρικές καθιστούν σαφές ότι ο αλγόριθμος *UMAP* αποδίδει καλύτερα από όλες τις άλλες τεχνικές, παρουσιάζοντας τις μεγαλύτερες τιμές για τα κριτήρια του *Davies-Bouldin* και *Calinski-Harabasz Index*, επιβεβαιώνοντας την φήμη του. Οι τιμές αυτές φανερώνουν ότι με την μείωση διάστασης μέσω του *UMAP*, δημιουργούνται πιο διακριτές ομάδες με χαμηλή «ενδοσυσταδική» διακύμανση και χαμηλή επικάλυψη μεταξύ σημείων ποικίλων ομάδων. Επιβεβαιώνεται ότι η χρήση της *PCA* (μοναδικού αλγορίθμου γραμμικής μετατροπής) δεν επαρκεί ώστε να δημιουργήσει διακριτά μοτίβα, παρουσιάζοντας την χαμηλότερη απόδοση στο σύνολο. Η λειτουργία της *t-SNE* είναι αρκετά ικανοποιητική, καταφέρνει να ομαδοποιήσει τα σημεία σε υψηλό βαθμό, αλλά, όπως αναλύσαμε προηγουμένως, εμφανίζει περισσότερα σημεία σε άλλες ομάδες, ασχέτως αν δεν υπάρχει επικάλυψη με υψηλή «ενδοσυσταδική» διακύμανση, ενισχύοντας την παρατήρηση για την μικρή υπομονάδα που διακρίνεται στην κατηγορία *BRCA*. Όσον αφορά τις προτεινόμενες μεθόδους(*RGt-SNE, RPEV*), φαίνεται ότι ο *RPEV* παρουσιάζει καλύτερα αποτελέσματα σε σχέση με τον *t-SNE*.

Ο *RGt-SNE* είναι ο δεύτερος σε βαθμολογία αλγόριθμος (αν και ο *Silhouette Score* αποδίδει καλύτερα), παρουσιάζοντας καλό διαχωρισμό ομάδων και ελάχιστα σημεία που ομαδοποιούνται με διαφορετική κλάση. Συγκρίνοντας την απόδοση του με τον *t-SNE,* οι διαφορές φαίνεται ότι εμπίπτουν στο γεγονός πως ο ο *RGt-SNE* εμφανίζει πιο σφιχτές ομάδες με ιδιαίτερα χαμηλή διακύμανση.

Συμπερασματικά, αναλογιζόμενοι τις παραπάνω αναλύσεις και αξιολογήσεις διακρίνεται ότι η φύση των δεδομένων προσφέρει σε μεγάλο βαθμό ικανοποιητική οπτικοποίηση και μείωση της αρχικής τεράστιας διάστασης στους περισσότερους αλγόριθμους. Η ανάλυση κύριων συνιστωσών ίσως να είναι περισσότερο χρήσιμη ως ένα προεπεξεργαστικό εργαλείο για μείωση διάστασης σε βέλτιστο χώρο και ύστερα διοχέτευσης στους πιο αποδοτικούς αλγορίθμους.

Συνολικά, οι κατηγορίες *KIRC* και *PRAD* τείνουν να ξεχωρίζουν από τις άλλες ομάδες, δηλαδή εμφανίζουν πιθανότατα διακριτή διαφοροποίηση τόσο μεταξύ τους όσο και με τις άλλες ομάδες. Όλες οι προσεγγίσεις εμφανίζουν μία εμφανή τάση των ομάδων *BRCA-LUAD* και *LUAD-COAD* να συγκλίνουν μεταξύ τους, αν και στην πλειονότητα των περιπτώσεων δεν υπάρχει επικάλυψη, φανερώνοντας πιθανότατα ότι κάποια δείγματα δεν μπορούν να κατηγοριοποιηθούν εύκολα, καθώς εμφανίζουν περισσότερες ομοιότητες με γειτονικές ομάδες. Σύμφωνα με τα παραπάνω, τα γονίδια που εκφράζουν τις περιπτώσεις καρκίνου *KIRC* και *PRAD* πιθανόν να είναι διαφορετικά τόσο μεταξύ τους όσο και με τις άλλες τρεις ομάδες και ίσως η κατηγοριοποίηση να είναι ιδιαίτερα αποτελεσματική γι’ αυτές τις δύο κλάσεις. Ενδεχομένως και η πλειοψηφία των εκφραζόμενων γονιδίων των τριών άλλων κλάσεων να εμφανίζει έντονες διαφοροποιήσεις, αλλά και αρκετά όμοιες περιπτώσεις, δυσχεραίνοντας σε κάποιες περιπτώσεις τόσο την ομαδοποίηση όσο και την κατηγοριοποίηση των δεδομένων.

Κεφάλαιο 5ο

Διερευνητική κατηγοριοποίηση βάσει αποτελεσμάτων μείωσης διάστασης

Η παρούσα πτυχιακή εργασία έχει ως σκοπό την παρουσίαση και εφαρμογή αλγοριθμικών τεχνικών μείωσης διάστασης. Αφού παρουσιάσαμε, αξιολογήσαμε και αναλύσαμε τις προαναφερόμενες μεθόδους, διατυπώσαμε και ένα σύνολο «υποθέσεων» που προέκυψαν και μέσω της οπτικοποίησης των δεδομένων σε δισδιάστατη αναπαράσταση. Στο παρόν τελευταίο κεφάλαιο, λοιπόν, εφαρμόζουμε απλές τεχνικές κατηγοριοποίησης των δεδομένων, χωρίς να εμβαθύνουμε στο θεωρητικό πλαίσιο, προσπαθώντας να αναγνωρίσουμε αν τα αποτελέσματα της μείωσης διάστασης κατάφεραν να μας προϊδεάσουν για το πως μπορούν να εξελιχθούν οι εφαρμογές επιβλεπόμενης μάθησης.

5.1 Εφαρμογή αλγορίθμων κατηγοριοποίησης για τις καλλιέργειες κορωναϊού

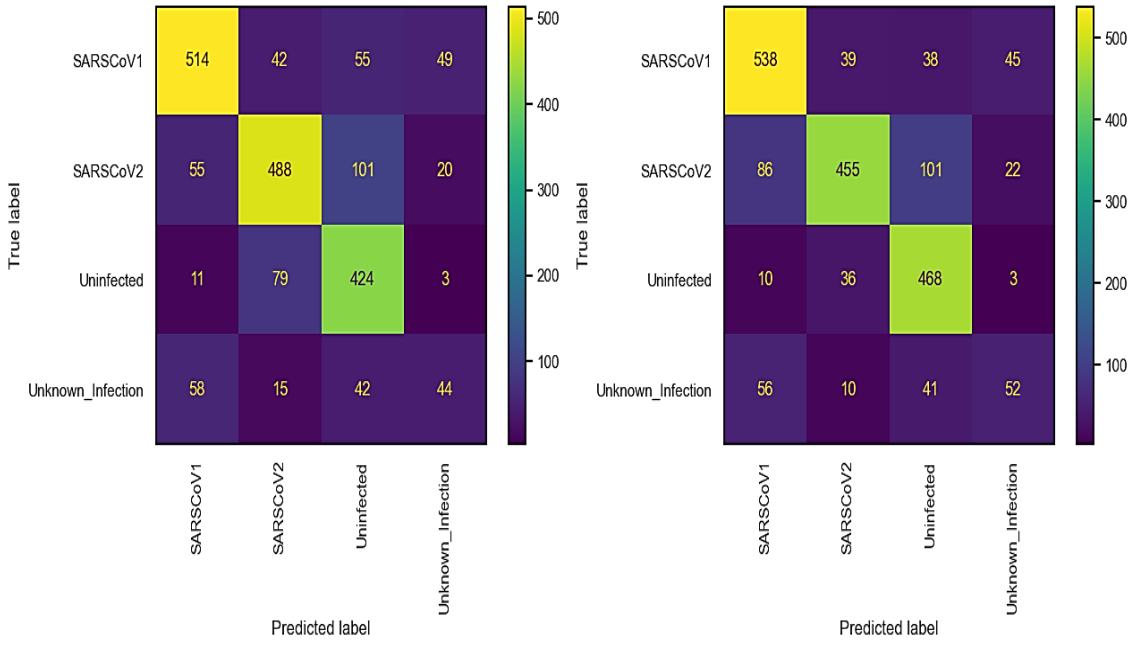
Σύμφωνα με την ενδελεχή έρευνα [15], η μείωση διάστασης διευκολύνει εφαρμογές κατηγοριοποίησης τόσο βάσει της συμπίεσης δεδομένων, όσο και με την ερμηνεία των αποτελεσμάτων οπτικοποίησης, που περιγράφουν την διακύμανση των σημείων στον χώρο και αποκαλύπτουν συχνά ενδιαφέροντα μοτίβα [6][15][25] [29]. Η ανάλυση πρωτευόντων συνιστωσών, όπως και σε πολλές περιπτώσεις και ο *UMAP* λειτουργούν ως ιδιαίτερα αποτελεσματικά εργαλεία στην βελτίωση της απόδοσης των αλγορίθμων κατηγοριοποίησης [26]. Το θεωρητικό τους υπόβαθρο είναι αναπτυγμένο, ώστε να αναζητούν μία ισορροπημένη μετάβαση από την συνολική αρχική δομή σε όσο το δυνατόν μικρότερη αναπαράσταση [6]. Από την άλλη μεριά, οι αλγόριθμοι *t-SNE, RGt-SNE* και *RPEV* περιορίζονται με τα μέχρι τώρα υπάρχοντα δεδομένα σε ιδιαίτερα αποτελεσματικές απεικονίσεις της διασποράς των σημείων και στην αποκάλυψη τοπικών ομοιοτήτων [6].

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, σε κάθε περίπτωση μείωσης διάστασης εντοπίσαμε περιοχές όπου υπήρχε διακριτός διαχωρισμός σημείων ανά κατηγορία (*SARSCoV1, SARSCoV2, Uninfected*), αρκετά διάσπαρτα σημεία στον χώρο ανεξαρτήτου κλάσης και δύο ιδιαίτερες περιοχές όπου σημεία κατηγοριών στοιχίζονται μεταξύ τους (*SARSCoV1, SARSCoV2, Unknown_Infection)*. Αυτή, λοιπόν, είναι η αίσθηση που μας παρείχε η διαδικασία οπτικοποιήσης. Δηλαδή, πιθανόν περίπου 5 «ομάδες» σημείων με υψηλή ομοιότητα σε εσωτερικό επίπεδο (δεν γνωρίζουμε τις «διασυσταδικές» συσχετίσεις [28]) και ένα σεβαστό πλήθος σημείων που διασπείρονται ακανόνιστα στον χώρο. Στόχος μας, λοιπόν, στην παρούσα ενότητα είναι να εφαρμόσουμε κατηγοριοποίηση των δεδομένων, προσπαθώντας να ανακαλύψουμε εάν η διαισθητική εντύπωση που μας έδωσε η αναπαράσταση των δεδομένων μας σε δισδιάστατη προβολή είναι όντως ρεαλιστική και κατά πόσο οι αλγόριθμοι ταξινόμησης συγκλίνουν στον τρόπο με τον οποίο ερμηνεύσαμε τα αποτελέσματα. Δεν γίνεται κάποια εξειδικευμένη ανάλυση ως προς την κατηγοριοποίηση των δεδομένων, απλά εφαρμόζονται γνωστές μέθοδοι με διαφορετικό θεωρητικό υπόβαθρο, πλεονεκτήματα και περιορισμούς. Έτσι, μέσω της διαδικασίας διασταυρωμένης επικύρωσης (*cross-validation*) διαιρούμε σε 10 μέρη το σύνολο δεδομένων, και εκτελούμε τους αλγορίθμους επιδιώκοντας να αποκτήσουμε όσο το δυνατόν πιο αντικειμενικά αποτελέσματα. Ο τελικός έλεγχος γίνεται μέσω του πίνακα σύγχυσης 4×4, στον οποίο συγκεντρώνονται το σύνολο των αληθινών και προβλεπόμενων παρατηρήσεων ανά κλάση, δίνοντας πληροφορίες για την τάση ορισμένων κατηγοριών να συγχέονται μεταξύ τους, δηλαδή την απάντηση στο πόσο ρεαλιστικά είναι τα συμπεράσματα μας και οι υποθέσεις μας. Χωρίζουμε την ανάλυση μας σε δύο σκέλη. Στο πρώτο σκέλος θεωρούμε δύο αλγορίθμους ταξινόμησης και εξετάζουμε τις αποδόσεις βάσει των συμπερασμάτων και των ενδεχομένων που παραθέσαμε. Στο δεύτερο σκέλος εξετάζουμε το πως επηρεάζεται η οπτικοποίηση εκμεταλλευόμενοι τις προβλέψεις που διαθέτουμε. Υπενθυμίζεται ότι το κεντρικό νόημα του παρόντος κεφαλαίου δεν αποτελεί κάποια εξειδικευμένη ανάλυση της ταξινόμησης, αλλά μία διερευνητική προσέγγιση και εκτίμηση της επιρροής της μείωσης διάστασης [15].

Στο πρώτο σκέλος λαμβάνοντας υπόψιν την οπτικοποίηση και αξιολόγηση όλων των τεχνικών και κυρίως των τριών πιο ισχυρών *(RG-tNE,t-SNE και RPEV)*, αναφέραμε μία σειρά συμπερασμάτων και πιθανών σεναρίων. Οι δύο αλγόριθμοι ταξινόμησης αφορούν απλοποιημένες προσεγγίσεις των:

- ***K Nearest Neighbors:*** Ένας από τους πιο απλούς αλγορίθμους που προσπαθεί να κατηγοριοποιήσει σημεία αναζητώντας του *k* κοντινότερους γείτονες βάσει ενός μετρικού συστήματος αποστάσεων (π.χ. Ευκλείδιο, *Manhattan* κ.ο.κ). Δεν κάνει υποθέσεις πάνω στα δεδομένα (μη παραμετρικός) και με ιδιαίτερα καλή απόδοση σε περιπτώσεις πολλών ετικετών. Παρ' όλα αυτά, την απόδοση του δυσχεραίνει ο υψηλός αριθμός διαστάσεων, η ανισορροπία μεταξύ κλάσεων και η ύπαρξη ακραίων σημείων (*outliers*) [38][39].
- ***Support Vector Machine:*** Ο μηχανισμός του βασίζεται στην εύρεση μίας κατάλληλης προβολής των σημείων (γραμμικής ή μη), όπου τα πιο κοντινά σημεία κάθε κλάσης χωρίζονται με όσο το δυνατόν μέγιστο διάστημα. Συνήθως αποδίδει καλά σε μεγάλα *dataset*, όταν οι κλάσεις έχουν υψηλότερο βαθμό διαχωρισμού και τα ακραία σημεία(*outliers*) δεν επηρεάζουν δραστικά. Τα μειονεκτήματα του είναι ότι είναι σχετικά αργός και αλλοιώνεται όταν υπάρχει μεγάλος βαθμός επικάλυψης μεταξύ κατηγοριών [39][40].

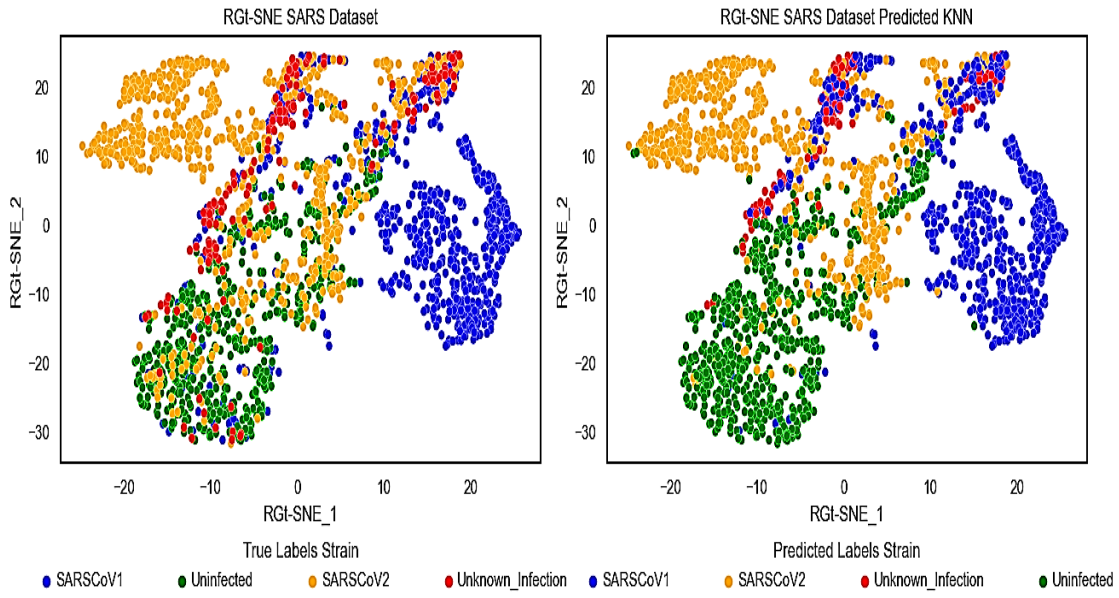
Αφού περιγράψαμε τους αλγορίθμους ταξινόμησης και τα θετικά και τα αρνητικά της κάθε περίπτωσης κατηγοριοποίησης εξάγουμε σχήμα πινάκων σύγχυσης που φανερώνει την σχέση των προβλέψεων έναντι των αληθινών ετικετών τις οποίες διαθέτουμε.



Πίνακας 5 Πίνακας σύγχυσης kNN k=7 (αριστερά) και SVM (δεξιά) για το SARSCovid Dataset

Οι παραπάνω πίνακες σύγχυσης παρουσιάζουν το πως καταφέρνει κάθε αλγόριθμος να εκπαιδευτεί μέσω της διαδικασίας *cross-validation*. Όπως προβλέψαμε, ένας αλγόριθμος, όπως ο *kNN*, κατορθώνει να ταξινομήσει τα περισσότερα σημεία των κατηγοριών *SARSCoV1* και *SARSCoV2 και Uninfected* το ίδιο και ο *SVM*. Οι κατηγορίες *SARSCoV2* και *Uninfected*, σύμφωνα με τις απεικονίσεις έπειτα από την μείωση διάστασης, παρουσίασαν αρκετά σημεία διάσπαρτα στον χώρο έναντι του *SARSCoV1* και έτσι ο *kNN* ενδέχεται να μπερδεύει τις δύο αυτές κατηγορίες, καθώς αποδεικνύεται ότι μεταξύ τους φανερώνεται μεγαλύτερη απώλεια ακρίβειας (101 και 79). Όσο για την κατηγορία *Unknown_Infection*, και στις δύο περιπτώσεις τα αποτελέσματα δεν είναι καλά. Ο *SVM* εντοπίζει ελάχιστα σημεία παραπάνω από τον *kNN* για το στέλεχος *SARCoV1*, πιθανόν αυτά που εμφανίζονται σε πιο ακραία σημεία, καθώς σαν αλγόριθμος επηρεάζεται εξ ορισμού λιγότερο από αυτά [40].

Χρησιμοποιώντας την μείωση διάστασης μέσω του *RGt-SNE* δοκιμάζουμε να χρωματίσουμε στην γραφική παράσταση διασποράς τα σημεία βάσει των προβλέψεων που έκαναν ο *kNN* και ο *SVM*. Στην παρακάτω εικόνα παρουσιάζουμε την οπτικοποίηση του *RGt-SNE* με ανάθεση ετικετών βασισμένη στην πραγματικότητα και στις προβλέψεις του *kNN*.



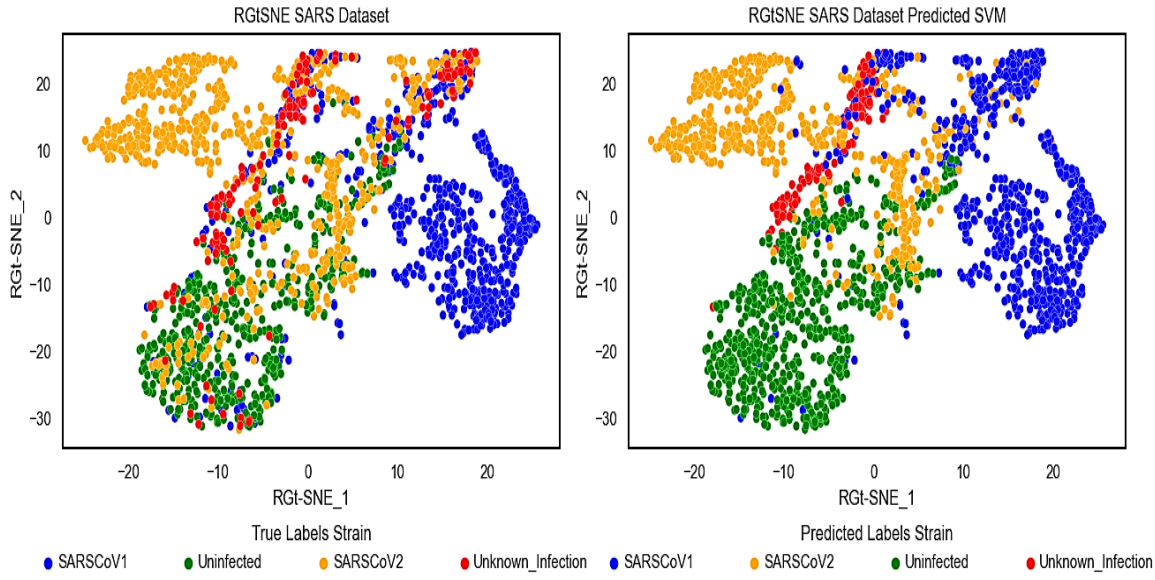
Εικόνα 5.1 Διάγραμμα RGt-SNE με πραγματικές ετικέτες (αριστερά) και μέσω κατηγοριοποίησης kNN (δεξιά) για το SARSCovid Dataset

Η συγκεκριμένη απεικόνιση έχει ως στόχο να προσπαθήσει να μας δώσει τον τρόπο με τον οποίο γίνεται αντιληπτή η διάκριση των σημείων στον χώρο βάσει των προβλέψεων του *kNN*, μέσω της προβολής σε δύο διαστάσεις. Όπως αναμέναμε, τα πιο

ακραία σημεία κάθε κλάσης δεν βρέθηκαν με επιτυχία, ιδιαίτερα στην περίπτωση μεταξύ του *SARSCoV2* και *Uninfected*. Οι δύο προσεγγίσεις φανερώνουν την αδυναμία του *kNN* στην ταξινόμηση ακραίων σημείων οποιασδήποτε ετικέτας, ενώ στις επικαλυπτόμενες περιοχές που αναφέραμε δίνει προβάδισμα στον *SARSCoV1* και ταυτόχρονα εκεί εντοπίζει σχεδόν όλες τις περιπτώσεις *Unknown_Infection*, ιδιαίτερα αυτές που προβάλλονται σε εύρος (0,20) στον *RGt-SNE_2* άξονα.

Απ' την άλλη μεριά, βασιζόμενοι στην ίδια κεντρική ιδέα αξιολογούμε και την διασπορά σημείων μέσω χρωματισμού των προβλέψεων του *SVM*. Στις περιοχές με μεγάλη επικάλυψη κοινών κλάσεων που αναφέραμε στην υποενότητα *4.3.4.2* η κατηγοριοποίηση δεν κατορθώνει να εμφανίσει καλή απόδοση. Γενικά, ακόμα και για μηχανισμό που έχει μεγαλύτερη ευχέρεια σε ακραία σημεία, σε αυτό το σύνολο δεδομένων η ταξινόμηση τους δεν φαίνεται ξεκάθαρη. Ο *SVM* δίνει περισσότερες σωστές κατηγοριοποιήσεις *SARSCoV1* και *Uninfected*, στερώντας πιο πολλά σημεία στον *SARSCoV2*.

Συνολικά, αν και ο *RGt-SNE* μας παρέθεσε κάποιες αρκετά διακριτές περιοχές που στοιχίζονται οι κλάσεις *SARSCoV1* και *SARSCoV2*, υπάρχουν περιπτώσεις και στους δύο αλγόριθμους (ελάχιστες στον *kNN* και περισσότερες στον *SVM)* που αυτή η υποτιθέμενη διάκριση δεν ικανοποιήθηκε.



Εικόνα 5.2 Διάγραμμα RGt-SNE με πραγματικές ετικέτες (αριστερά) και μέσω κατηγοριοποίησης SVM (δεξιά) για το SARSCovid Dataset

Οι μετρικές αξιολόγησης κατηγοριοποίησης για του δύο αλγορίθμους μας παρουσιάζουν και την συνολική εικόνα σε επίπεδο ακρίβειας και ανάκλησης.

Labels (Support)	*Precision*	*Recall*	*F1_Score*
Αλγόριθμος	***K***	***Nearest***	***Neighbors***
SARSCoV1 (660)	*0.81*	*0.78*	*0.79*
SARSCoV2 (664)	*0.78*	*0.73*	*0.76*
Uninfected (517)	*0.68*	*0.82*	*0.74*
Unknown (159)	*0.38*	*0.28*	*0.32*
accuracy			*0.73*
Macro_Average	*0.66*	*0.65*	*0.65*
Weight_Average	*0.73*	*0.73*	*0.73*
Αλγόριθμος	***Support***	***Vector***	***Machines***
SARSCoV1 (660)	*0.78*	*0.82*	*0.80*
SARSCoV2 (664)	*0.84*	*0.69*	*0.76*
Uninfected (517)	*0.72*	*0.91*	*0.80*
Unknown (159)	*0.43*	*0.33*	*0.37*
accuracy			*0.76*
Macro_Average	*0.69*	*0.68*	*0.68*
Weight_Average	*0.76*	*0.76*	*0.75*

Πίνακας 7 Μετρικές αξιολόγησης κατηγοριοποίησης των δύο αλγορίθμων ταξινόμησης KNN και SVM. Μετρικές ακρίβειας, ανάκλησης, F1 για κάθε κατηγορία και συνολικά .

Οι τρεις από τις τέσσερεις κατηγορίες εμφανίζουν παραπλήσιο πλήθος ετικετών ενώ η κατηγορία *Unknown Infection* πολύ μικρότερο. Η ανάκληση(*recall*) μας παρουσιάζει ένα μέτρο κατά το οποίο μπορούμε να εκτιμήσουμε την πιθανότητα κάθε στοιχείο που εξετάζουμε να προβλεφθεί σωστά σε σύγκριση με ολόκληρο το σύνολο δεδομένων. Η ακρίβεια(*precision*), από την άλλη μεριά, επιδιώκει να προσδιορίσει κατά πόσο μπορούμε να εμπιστευτούμε ένα μοντέλο να προβλέψει ένα δείγμα ατομικώς ορθά. Παράλληλα, παρουσιάζεται και η αμοιβαία συνεισφορά των δύο αυτών κριτηρίων μέσω της αμοιβαίας συνδρομής που περιγράφεται απ' τον δείκτη *F1*.

Ο παραπάνω πίνακας μας παρουσιάζει μία ξεκάθαρη οπτική για τον τρόπο με τον οποίο οι δύο αλγόριθμοι αποδίδουν. Αν και οι τιμές απόδοσης (*accuracy*)

εμφανίζουν ελάχιστες αποκλίσεις, παρατηρούμε ότι ο *SVM* εμφανίζει διακριτή βελτίωση των δεικτών της κατηγορίας *Unknown_Infection* και *Uninfected*. Από την στιγμή που αναφερόμαστε σε δύο κατηγορίες με υψηλή διάχυση στον χώρο, αποτελεί σημαντικό κριτήριο ότι ο *SVM* αντιλαμβάνεται καλύτερα τις δύο αυτές κατηγορίες. Σε γενικές γραμμές θα αποτελούσε σημαντικό κριτήριο επιλογής μεθόδου η καλύτερη προσέγγιση στην κατηγορία *Unknown_Infection,* καθώς φανερώνεται μία χαμηλή συγκέντρωση τιμών με αποτέλεσμα ο εντοπισμός τους να υποσκιάζονται από αυτές μεγαλύτερης κλάσης.

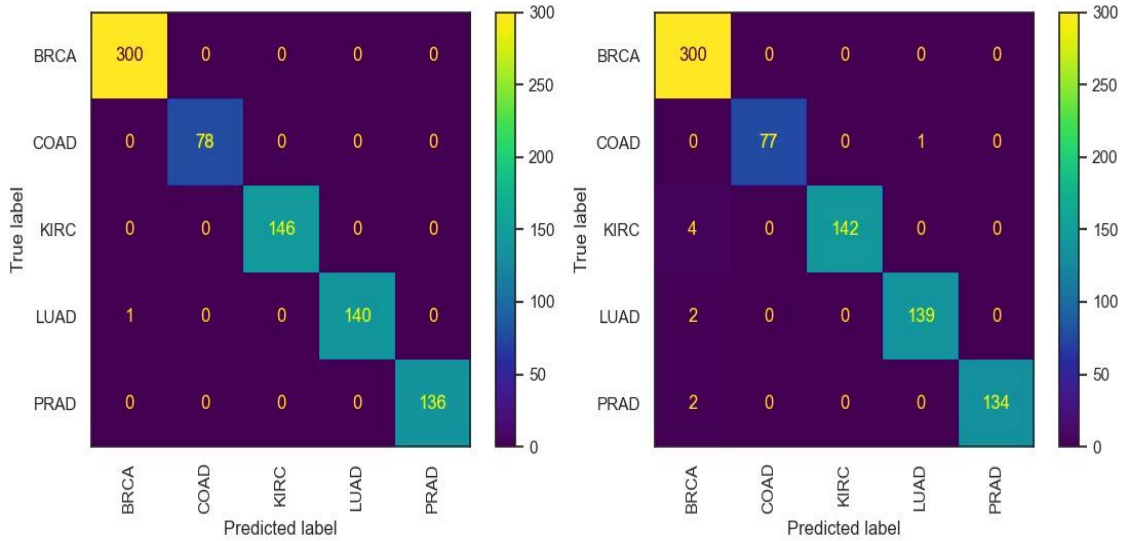
5.2 Εφαρμογή αλγορίθμων κατηγοριοποίησης για τα καρκινικά δείγματα

Ακολουθώντας την ίδια τακτική προσέγγιση με την προηγούμενη ενότητα χρησιμοποιούμε τους δύο αλγορίθμους κατηγοριοποίησης των *k* κοντινότερων γειτόνων και *Support Vector Machines*. Όπως αναλύσαμε στο κεφάλαιο 4, όλοι οι αλγόριθμοι μείωσης διάστασης που διενεργήθηκαν παρουσίασαν ιδιαίτερα υψηλή απόδοση και σχεδόν πέτυχαν να διαχωρίσουν τα σημεία ανά κλάση όσο το δυνατόν καλύτερα. Οι δύο καλύτερες βάσει εσωτερικών μέτρων αξιολόγησης μέθοδοι αφορούν αυτούς του *RGt-SNE* και *UMAP*. Σε αυτές τις τεχνικές παρουσιάζονται ιδιαίτερα διακριτές «ομάδες», σφιχτά δεμένες με σχεδόν απόλυτη σύγκλιση των σημείων κοινών κατηγοριών. Τα αποτελέσματα της μείωσης διάστασης και της οπτικοποίησης του συγκεκριμένου συνόλου δεδομένων για περιπτώσεις καρκίνου παρουσίασαν 5 καθαρές «συστάδες», χωρίς να υπάρχει κάποια επικάλυψη μεταξύ τους, παρουσιάζοντας δύο μόνο ακραία σημεία, τα οποία ανήκουν στην κατηγορία *LUAD*. Η διαδικασία ανάλυσης των αποτελεσμάτων στο προηγούμενο κεφάλαιο μας οδήγησε στην ανάπτυξη υποθέσεων σχετικά με την πιθανή διάκριση διαφορετικών λειτουργικών γονιδίων ανά κλάση και με την ενδεχόμενη υψηλής ποιότητας κατηγοριοποίηση. Προφανώς, δεν θεωρούμε απόλυτη την προαναφερθείσα θεώρηση και επιφυλασσόμαστε για την πορεία και την απόδοση των τεχνικών ταξινόμησης. Όπως εξηγήθηκε και στην προηγούμενη ενότητα, αναζητούμε ενδείξεις για το πόσο κοντά οι παραπάνω αλγόριθμοι προσεγγίζουν την «αλήθεια» και δύναται να μας δώσουν μία διαισθητική προσέγγιση για την διασπορά των σημείων στο χώρο.

Χρησιμοποιώντας, λοιπόν, τις υποθέσεις που διατυπώσαμε παραπάνω αναμένουμε μία υψηλή απόδοση των αλγορίθμων κατηγοριοποίησης που διαθέτουμε.

Αν τα μοντέλα ακολουθούσαν πιστά την διασπορά των παραπάνω διαγραμμάτων , πιθανότατα η πλειοψηφία των σημείων να ταξινομείται, με πιο ευδιάκριτες περιπτώσεις λάθους τα δύο ακραία σημεία της κλάσης *LUAD*. Βέβαια, οι αποστάσεις μέσω των ευδιάκριτων ομάδων συνήθως μπορεί να μην σημαίνουν και τίποτα, οπότε δύσκολα να μπορούμε να κάνουμε μία πρόβλεψη για την σύγχυση που μπορεί να υπάρχει μεταξύ κατηγοριών. Προσεγγίζουμε πιο εύκολα τα δύο σημεία που αναφέρθηκε ότι πιθανόν να συγχύζονται με την κατηγορία *BRCA*.

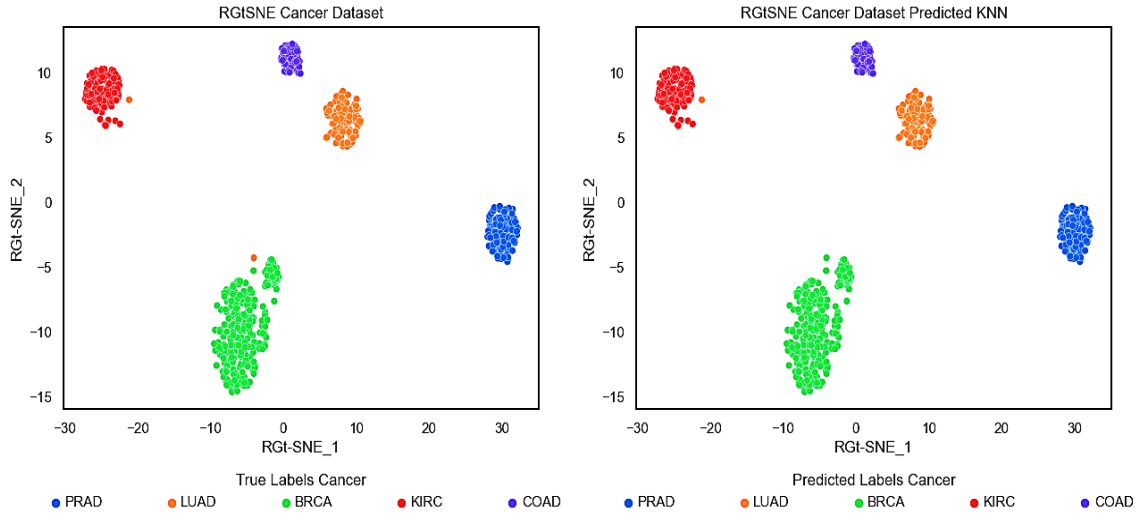
Ύστερα, λοιπόν, από την διαδικασία διασταυρωμένης επικύρωσης (*cross-validation)* για τους αλγορίθμους *kNN* και *SVM*, κατασκευάζουμε τους πίνακες σύγχυσης 5×5 για τις αληθινές και προβλεπόμενες ετικέτες.



Πίνακας 6 Πίνακας σύγχυσης kNN k=7 (αριστερά) και SVM (δεξιά) για το Cancer Dataset

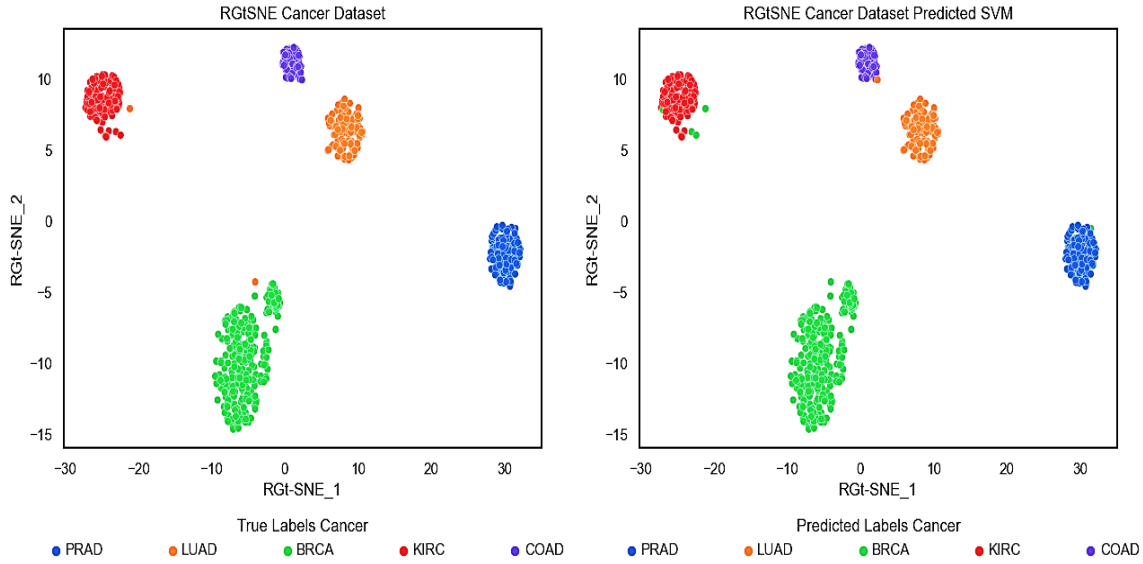
Οι παραπάνω πίνακες παρουσιάζουν το κατά πόσο οι τεχνικές ταξινόμησης καταφέρνουν να εκπαιδευτούν αποτελεσματικά σε όλο το φάσμα των συνόλων δεδομένων. Παρατηρούμε ότι οι αρχικές μας εκτιμήσεις επαληθεύονται, καθότι σχεδόν όλα τα σημεία ταξινομούνται σωστά πλην ενός στην περίπτωση του *kNN* και *9* σ' αυτή του *SVM*. Για το *kNN* γίνεται αντιληπτό στο παρακάτω σχήμα το πως ένα από τα ακραία σημεία της κατηγορίας *LUAD* συγχέεται με αυτό της *BRCA*. Παρ' όλα αυτά η εύρεση όλων των άλλων κατηγοριών είναι τέλεια, χωρίς να υπάρχει κάποια απώλεια ή

σύγχυση.



Εικόνα 5.3 Διάγραμμα RGt-SNE με πραγματικές ετικέτες (αριστερά) και μέσω κατηγοριοποίησης kNN (δεξιά) για το Cancer Dataset

Τα παραπάνω διαγράμματα αντικατοπτρίζουν τις οπτικοποιήσεις των σημείων σε δισδιάστατη αναπαράσταση μέσω του *RGt-SNE* (κάλλιστα θα μπορούσαμε να βάλουμε και τον *UMAP*), υποδεικνύοντας την αποτελεσματικότητα του *kNN* και τα αδύναμα σημεία. Ως ένας αλγόριθμος που βασίζεται στην εύρεση των αποστάσεων μεταξύ των σημείων, φανερώνονται αδυναμίες στην ακριβή ταξινόμηση ακραίων σημείων(*outliers*), οι οποίες επαληθεύονται βάσει των διαγραμμάτων διασποράς, αναδεικνύοντας την ικανότητα των μεθόδων μείωσης διάστασης να προσεγγίσουν τις τοπικές συσχετίσεις με ιδιαίτερη ευχέρεια.



Εικόνα 5.4 Διάγραμμα RGt-SNE με πραγματικές ετικέτες (αριστερά) και μέσω κατηγοριοποίησης SVM (δεξιά) για το Cancer Dataset

Από την άλλη μεριά, ο *SVM* παρουσιάζει περισσότερες περιπτώσεις λανθασμένης ταξινόμησης. Η μόνη κατηγορία που εμφανίζει σωστά όλα τα δείγματα είναι η *BRCA*, αν και παρουσιάζεται υψηλός αριθμός ψευδώς αρνητικών προβλέψεων. Από την κατηγορία *KIRC* 4 σημεία συγχέονται με αυτά της *BRCA* και το ίδιο συμβαίνει με δύο σημεία της *PRAD*. Επιπλέον, και τα δύο ακραία σημεία που εντοπίζονται για την *LUAD* κατηγοριοποιούνται ως *BRCA*, ενώ ένα σημείο της *COAD* συγχέεται με την κλάση *LUAD*, κάτι που φανέρωσε μόνο ο αλγόριθμος του *RPEV*.

Labels (Support)	*Precision*	*Recall*	*F1_Score*
Αλγόριθμος	***K***	***Nearest***	***Neighbors***
BRCA (300)	*1.00*	*1.00*	*1.00*
COAD (78)	*1.00*	*1.00*	*1.00*
KIRC (146)	*1.00*	*1.00*	*1.00*
LUAD (141)	*1.00*	*0.99*	*1.00*
PRAD (136)	*1.00*	*1.00*	*1.00*
accuracy			*1.00*
Macro_Average	*1.00*	*1.00*	*1.00*
Weight_Average	*1.00*	*1.00*	*1.00*
Αλγόριθμος	***Support***	***Vector***	***Machines***
BRCA (300)	*0.97*	*1.00*	*0.99*
COAD (78)	*1.00*	*0.99*	*0.99*
KIRC (146)	*1.00*	*0.97*	*0.99*
LUAD (141)	*0.99*	*0.99*	*0.99*
PRAD (136)	*1.00*	*0.99*	*0.99*
accuracy			*0.99*
Macro_Average	*0.99*	*0.99*	*0.99*
Weight_Average	*0.99*	*0.99*	*0.99*

Πίνακας 8 Μετρικές αξιολόγησης κατηγοριοποίησης των δύο αλγορίθμων ταξινόμησης KNN και SVM. Μετρικές ακρίβειας, ανάκλησης, F1 για κάθε κατηγορία και συνολικά .

Σ' αυτή την περίπτωση δεν υπάρχουν πολλά πράγματα να αναλυθούν καθώς οι αποδόσεις ταξινόμησης αγγίζουν το τέλειο στο *KNN* και στον *SVM*. Το μόνο που ίσως

αξίζει σχολιασμό είναι ότι υπάρχει μεγάλη πιθανότητα σε νέες καταχωρήσεις οι οποίες να εμφανίζουν πιο ακραία συμπεριφορά η προσέγγιση κατηγοριοποίησης να μην εμφανίσει ανάλογη απόδοση.

5.3 Συμπεράσματα πάνω στην μείωση διάστασης και την κατηγοριοποίηση

Το κεντρικό θέμα που κληθήκαμε να αναλύσουμε αποτέλεσε η μείωση διάστασης. Η επίλυση προβλημάτων υψηλής διαστατικότητας συνθέτει ένα πολύπλοκο μοτίβο, το οποίο προκαλεί δυσκολίες στον τρόπο που αντιλαμβανόμαστε την διακύμανση των σημείων και τον τρόπο που κατανοούμε τις βασικές ιδιότητες τους. Η εφαρμογή διάφορων μεθόδων μείωσης διάστασης, τόσο πάνω στο μαθηματικό τους υπόβαθρο, όσο και στις επιμέρους λειτουργίες τους μας χρησιμεύει, ώστε να αποκαλυφθούν σημαντικά πρότυπα πάνω στα δεδομένα. Ακόμα και αν δεν είμαστε σίγουροι για το πόσο ρεαλιστικά είναι τα αποτελέσματα οπτικοποίησης και μετασχηματισμού δεδομένων, αντιληφθήκαμε ότι μας παρέδωσαν ένα σημαντικό πλαίσιο ανάγνωσης, που σε πρώτο βαθμό δύσκολα θα μπορούσαμε να αναλύσουμε, εξαιτίας του υψηλού χώρου δειγμάτων και χαρακτηριστικών.

Η οπτική γωνία με την οποία αναλύσαμε την κατηγοριοποίηση των δεδομένων έγινε με σκοπό την ενίσχυση των ισχυρισμών μας πάνω στα αποτελέσματα και την ανάδειξη της σημασίας της μείωσης διάστασης. Ανακαλώντας το θεωρητικό υπόβαθρο των μεθόδων κατηγοριοποίησης ορίσαμε ένα πλαίσιο πάνω στο οποίο περιμέναμε να αποδώσουν, χωρίς βέβαια να προεξοφλούμε άκριτα το αποτέλεσμα. Τα ίδια τα αποτελέσματα φανέρωσαν ότι τα πλεονεκτήματα και τα μειονεκτήματα του εκάστοτε αλγορίθμου συμβαδίζουν σε ένα τεράστιο βαθμό(όχι απόλυτα) με τις υποθέσεις και το πλαίσιο που αναλύσαμε την οπτικοποίηση των αποτελεσμάτων, δεδομένου των αληθινών ετικετών.

Εν κατακλείδι, όλες οι εφαρμογές και αναλύσεις στις οποίες προβήκαμε πάνω στο σύνολο δεδομένων, αν και η πλειοψηφία σημείων για τις κατηγορίες *SARSCoV1*, *SARSCoV2* και *Uninfected* διακρίνονται καθαρά, ενδεχομένως να παρουσιάζουν περισσότερες ομοιότητες στην γονιδιακή έκφραση από όσο αρχικά προσεγγίσαμε, κυρίως μεταξύ των ζευγαριών *SARSCoV1-SARSCoV2* και *SARSCoV2-Uninfeted.*

Αποτέλεσμα είναι ορισμένοι αλγόριθμοι(όπως ο *SVM* και πολύ λιγότερο ο *kNN)* να χάνουν αρκετά τέτοια δείγματα. Τέλος, αποκαλύφθηκε πως ιδιαίτερη μεταχείριση θα χρειαστούν δείγματα που ανήκουν στην κατηγορία *Unknown_Infection,* καθώς η συμπεριφορά και των δύο αλγορίθμων στην εξεύρεση αυτών των σημείων θεωρείται ανεπαρκής. Η πειραματική ανάλυση την οποία εκτελέσαμε παραπάνω, πιθανόν σε μεγάλο βαθμό ενισχύει το ενδεχόμενο που εξετάσαμε στην υποενότητα *4.4* για τον διαφορετικό τρόπο με τον οποίο αποκωδικοποιούνται τα δεδομένα εξαιτίας του χρόνου επώασης. Πιθανότατα, να επιβεβαιώνεται ο ισχυρισμός ότι τα δείγματα με μικρό χρόνο επώασης του κάθε ιού να παρουσιάζουν πιο δυσδιάκριτες διαφορές μεταξύ τους και ενδεχομένως, τα γονιδιακά μοτίβα να είναι αρκετά κοινά στην αρχή και να διαφοροποιούνται σε ένα πιο ευρύτερο πέρασμα του χρόνου (8 με 12 ώρες).

Στο ίδιο πλαίσιο ανάγνωσης για το δεύτερο σύνολο δεδομένων που αφορά τα καρκινικά δείγματα, η μείωση διάστασης και κατ' επέκταση η οπτικοποίηση σε δισδιάστατη αναπαράσταση μας έδωσε πολύ πιο ενθαρρυντικά αποτελέσματα. Τα σημεία κατά κύριο λόγο συσπειρώνονται με εκείνα που ανήκουν σε κοινή κατηγορία, εμφανίζοντας διακριτές δομές. Αν και δεν πρόκειται για ανάλυση ομαδοποίησης, η αξιολόγηση μέσω εσωτερικών μετρικών επικύρωσης παρουσίασαν ιδιαίτερα υψηλές αποδόσεις. Αυτές οι διακριτές δομές μαζί με τα οποιαδήποτε ακραία σημεία που εντοπίσαμε σε κάθε διάγραμμα διασποράς επαληθεύτηκαν μέσω των μεθόδων κατηγοριοποίησης. Παρατηρήσαμε την υψηλή απόδοση ενός αλγορίθμου που βασίζεται στην εύρεση μετρικών αποστάσεων μεταξύ των σημείων, όπως *kNN* , αλλά ενδέχεται ερευνώντας και τα αποτελέσματα του *SVM,* μπορούμε να δούμε ότι υπάρχουν δείγματα διαφορετικών κατηγοριών, όπως η *KIRC-BRCA* και η *PRAD-BRCA*, με περισσότερα κοινά από όσα αναμέναμε. Ακόμα και οι κατηγορίες *COAD-LUAD* παρουσίασαν ένα σημείο σύγχυσης κάτι που μας έκανε διακριτό μόνο ο αλγόριθμος *RPEV*. Μία ενδεχόμενη προσέγγιση πάνω στα δεδομένα αφορά ότι τα γονίδια εμφανίζουν μία αρκετά διακριτή διαφοροποίηση ανά κατηγορία, κάτι που θα εξηγούσε την υψηλή απόδοση κατά την διαδικασία κατηγοριοποίησης. Η επικρατούσα κατηγορία η *BRCA* ίσως παρουσιάζει αρκετά δείγματα πιο διάσπαρτα στον χώρο, όπως παρουσιάζεται στον *RPEV* δημιουργώντας σύγχυση με τις κατηγορίες *KIRC, PRAD* και *LUAD* και επομένως πιθανότατα να υπάρχει σύγκλιση σε επίπεδο γονιδιακής έκφρασης σε αυτές τις κατηγορίες περισσότερο από όσο αναμέναμε.

Βιβλιογραφικές Αναφορές

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. USA: Prentice Hall Press, 2009.

[2] "Big Data trend." http://www.enterpriseappstoday.com/business-intelligence/big-data-and-beyond-10-bi-trends-for-2013.html (accessed Dec. 31, 2020).

[3] Chethan Kumar GN, "Artificial Intelligence: Definition, Types, Examples, Technologies | by Chethan Kumar GN | Medium," *Aug 31, 2018*. https://medium.com/@chethankumargn/artificial-intelligence-definition-types-examples-technologies-962ea75c7b9b (accessed Dec. 31, 2020).

[4] "Τεχνητή Νοημοσύνη (Artificial Intelligence) – Τι είναι και γιατί είναι σημαντική | SAS." https://www.sas.com/el_gr/insights/analytics/what-is-artificial-intelligence.html (accessed Dec. 31, 2020).

[5] "Top 10 Real World Artificial Intelligence Applications | AI Applications | Edureka." https://www.edureka.co/blog/artificial-intelligence-applications/#AI in HealthCare (accessed Dec. 31, 2020).

[6] A. G. Vrahatis, S. K. Tasoulis, G. N. Dimitrakopoulos, and V. P. Plagianakos, "Visualizing High-Dimensional Single-Cell RNA-seq Data via Random Projections and Geodesic Distances," *2019 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2019*, 2019, doi: 10.1109/CIBCB.2019.8791482.

[7] T. Finnegan, A. Hall, and J. M. Skopek, *Identification and genomic data Identification and genomic data Acknowledgements Identification and genomic data*. 2017.

[8] C. Huttenhower and O. Hofmann, "A Quick Guide to Large-Scale Genomic Data Mining," *PLoS Comput. Biol.*, vol. 6, no. 5, p. e1000779, May 2010, doi: 10.1371/journal.pcbi.1000779.

[9] S. C. G. B. G. © 2008 N. E. Warren C. Lathe III (OpenHelix), Jennifer M. Williams (OpenHelix), Mary E. Mangan (OpenHelix) & Donna Karolchik (University of California and D. (2008) G. D. R. C. and P. N. E. 1(3):2

Citation: Lathe, W., Williams, J., Mangan, M. & Karolchik, “Genomic Data Resources: Challenges and Promises,” *https://www.nature.com/scitable/*, [Online]. Available: https://www.nature.com/scitable/topicpage/genomic-data-resources-challenges-and-promises-743721/#TB_inline?height=300&width=400&inlineId=trOutLine.

[10] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications,” *Genome Medicine*, vol. 9, no. 1. BioMed Central Ltd., pp. 1–12, Aug. 18, 2017, doi: 10.1186/s13073-017-0467-4.

[11] M. D. Luecken and F. J. Theis, “Current best practices in single-cell RNA-seq analysis: a tutorial,” *Mol. Syst. Biol.*, vol. 15, no. 6, Jun. 2019, doi: 10.15252/msb.20188746.

[12] S. Glen, “Dimensionality & High Dimensional Data: Definition, Examples, Curse of,” *StatisticsHowTo.com*. https://www.statisticshowto.com/dimensionality/.

[13] Great Learning Team, “What is Curse of Dimensionality in Machine Learning?” https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/ (accessed Jan. 01, 2021).

[14] Tony Yiu, “The Curse of Dimensionality. Why High Dimensional Data Can Be So… | by Tony Yiu | Towards Data Science,” *Jul 20, 2019*. https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e (accessed Jan. 01, 2021).

[15] P. O. Box, L. Van Der Maaten, E. Postma, and J. Van Den Herik, “Tilburg centre for Creative Computing Dimensionality Reduction: A Comparative Review Dimensionality Reduction: A Comparative Review,” 2009. [Online]. Available: http://www.uvt.nl/ticc.

[16] “Generalized eigenvector - Wikipedia.” https://en.wikipedia.org/wiki/Generalized_eigenvector (accessed Jan. 01, 2021).

[17] J. Shlens, “A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS Derivation, Discussion and Singular Value Decomposition,” 2003.

[18] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065. Royal

Society of London, Apr. 13, 2016, doi: 10.1098/rsta.2015.0202.

[19] J. Lever, M. Krzywinski, and N. Altman, “Principal component analysis,” *Nat. Methods*, vol. 14, no. 7, pp. 641–642, 2017, doi: 10.1038/nmeth.4346.

[20] M. J. Zaki and W. M. Jr, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. USA: Cambridge University Press, 2014.

[21] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” 2008.

[22] “How t-SNE works — openTSNE 0.3.13 documentation.” https://opentsne.readthedocs.io/en/latest/tsne_algorithm.html#t-sne (accessed Jan. 01, 2021).

[23] “T-Distribution / Student’s T: Definition, Step by Step Articles, Video.” https://www.statisticshowto.com/probability-and-statistics/t-distribution/ (accessed Jan. 01, 2021).

[24] “t-SNE – Laurens van der Maaten.” https://lvdmaaten.github.io/tsne/ (accessed Jan. 01, 2021).

[25] L. Mcinnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” 2020.

[26] “How UMAP Works — umap 0.5 documentation.” https://umap-learn.readthedocs.io/en/latest/how_umap_works.html (accessed Jan. 02, 2021).

[27] “Understanding UMAP.” https://pair-code.github.io/understanding-umap/supplement.html (accessed Jan. 02, 2021).

[28] “How Exactly UMAP Works. And why exactly it is better than tSNE | by Nikolay Oskolkov | Towards Data Science.” https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668 (accessed Jan. 02, 2021).

[29] S. K. Tasoulis, A. G. Vrahatis, S. V. Georgakopoulos, and V. P. Plagianakos, “Visualizing High-dimensional single-cell RNA-sequencing data through multiple Random Projections,” *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 5448–5450, 2019, doi: 10.1109/BigData.2018.8622170.

[30] “The Johnson-Lindenstrauss bound for embedding with random projections — scikit-learn 0.24.0 documentation.” https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_johnson_lindenstrauss_bound.html#sphx-glr-auto-examples-miscellaneous-plot-johnson-lindenstrauss-bound-py (accessed Jan. 01, 2021).

[31] J.-L. Lemma, S. Dasgupta, and A. Gupta, “An elementary proof of the.”

[32] S. Dasgupta, “Experiments with Random Projection.”

[33] “sklearn.random_projection.SparseRandomProjection — scikit-learn 0.24.0 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.random_projection.SparseRandomProjection.html (accessed Jan. 01, 2021).

[34] C. Hegde, M. B. Wakin, and R. G. Baraniuk, “Random Projections for Manifold Learning.”

[35] “2.2. Manifold learning — scikit-learn 0.16.1 documentation.” https://scikit-learn.org/0.16/modules/manifold.html#manifold (accessed Jan. 02, 2021).

[36] “GEO Accession viewer.” https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148729 (accessed Jan. 02, 2021).

[37] J.-O. Palacio-Niño and F. Berzal, “Evaluation Metrics for Unsupervised Learning Algorithms.”

[38] D. Luqman Abd AL-Nabi and S. Shukri Ahmed, “Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation),” 2013. [Online]. Available: www.iiste.org.

[39] S. Gupta, “Pros and cons of various Machine Learning algorithms | by Shailaja Gupta | Towards Data Science.” https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6 (accessed Jan. 10, 2021).

[40] R. Kumar and R. Verma, “Classification Algorithms for Data Mining: A Survey.”