Σελίδα

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΙΩΑΝΝΙΝΩΝ

ΣΧΟΛΗ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΜΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΔΙΚΤΥΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗ ΔΙΑΝΥΣΜΑΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΓΡΑΦΩΝ ΒΙΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Οδυσσέας Καραδήμας

Επιβλέπων: Πέτρος Καρβέλης

Επίκουρος Καθηγητής

Άρτα, Οκτώβριος, 2024

MACHINE LEARNING TECHNIQUES FOR VECTORIZED REPRESENTATION OF BIOLOGICAL DATA GRAPHS

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, 04/10/2024

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

   Πέτρος Καρβέλης,

   Επίκουρος Καθηγητής

2. Μέλος επιτροπής

   Χρυσόστομος Στύλιος,

   Καθηγητής

3. Μέλος επιτροπής

   Ανδρέας Σκορίλας,

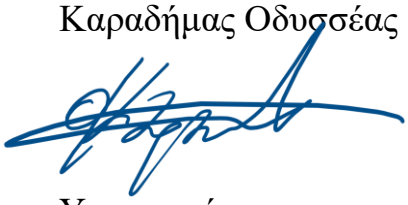
   Καθηγητής

© Καραδήμας Οδυσσέας, 2024.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εκ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Καραδήμας Οδυσσέας



Υπογραφή

ΠΕΡΙΛΗΨΗ

Ο αυξανόμενος όγκος βιολογικών δεδομένων που παράγονται από πειράματα παρουσιάζει σημαντικές προκλήσεις για την ανάλυση και την ερμηνεία, ενώ συχνά αφήνει ανεξερεύνητες πολύτιμες γνώσεις λόγω υπολογιστικών περιορισμών. Η αναπαράσταση μεγάλων βιολογικών συνόλων δεδομένων σε διανυσματική μορφή αποτελεί μια κρίσιμη προσέγγιση για την αντιμετώπιση αυτών των προκλήσεων, προσφέροντας ένα κλιμακούμενο και αποτελεσματικό μέσο για την επεξεργασία και την ανάλυση πολύπλοκων δεδομένων. Η παρούσα διατριβή επικεντρώνεται στην αναπαράσταση δεδομένων πρωτεϊνικής αλληλεπίδρασης (Protein – Protein Interactions, PPI), τα οποία δομούνται σε μορφή γράφου, γεγονός που δημιουργεί σημαντικές δυσκολίες για τις καθιερωμένες μεθοδολογίες ανάλυσης της θεωρίας γράφων, όταν εφαρμόζονται σε σύνολα δεδομένων μεγάλης κλίμακας.

Η έρευνα αυτή αναδεικνύει τη σημασία της διανυσματικής αναπαράστασης στην υπέρβαση αυτών των περιορισμών, παρέχοντας μια νέα προσέγγιση που αξιοποιεί προηγμένες τεχνικές μείωσης των διαστάσεων για την οπτικοποίηση δεδομένων υψηλής διάστασης σε διαχειρίσιμες μορφές. Με τη μετατροπή των δεδομένων PPI σε διανυσματική μορφή, επιτρέπουμε την αποτελεσματικότερη ομαδοποίηση και τον εντοπισμό λειτουργικών σχέσεων μεταξύ πρωτεϊνών, διευκολύνοντας την ανακάλυψη κρυμμένων μοτίβων και γνώσεων. Η εργασία αυτή αποτελεί σημαντική πρόοδο στον τομέα της βιοπληροφορικής, προσφέροντας ισχυρά εργαλεία για την ανάλυση πολύπλοκων βιολογικών συστημάτων και ανοίγοντας το δρόμο για μελλοντικές έρευνες και ανακαλύψεις.

Λέξεις-κλειδιά: βιολογικά δεδομένα, διανυσματοποίηση, πρωτεϊνικές αλληλεπιδράσεις, γράφοι, μηχανική μάθηση, βιοπληροφορική

ABSTRACT

The increasing volume of biological data generated from experiments presents significant challenges for analysis and interpretation, often leaving valuable insights undiscovered due to computational limitations. Representing large biological datasets in vector form is a crucial approach to address these challenges, offering a scalable and efficient means to process and analyze complex data. This thesis focuses on the representation of Protein-Protein Interaction (PPI) data, traditionally structured in graph format, which poses significant difficulties for traditional graph theory approaches, when applied to large-scale datasets.

Our research highlights the importance of vector representation in overcoming these limitations, providing a novel approach that leverages advanced dimensionality reduction techniques to visualize high-dimensional data in manageable forms. By converting PPI data into vector form, we enable more effective clustering and identification of functional relationships between proteins, facilitating the discovery of hidden patterns and insights. This work represents a significant advancement in the field of bioinformatics, offering powerful tools for the analysis of complex biological systems and paving the way for future research and breakthroughs.

Keywords: biological data, vectorization, protein interactions, graphs, machine learning, bioinformatics.

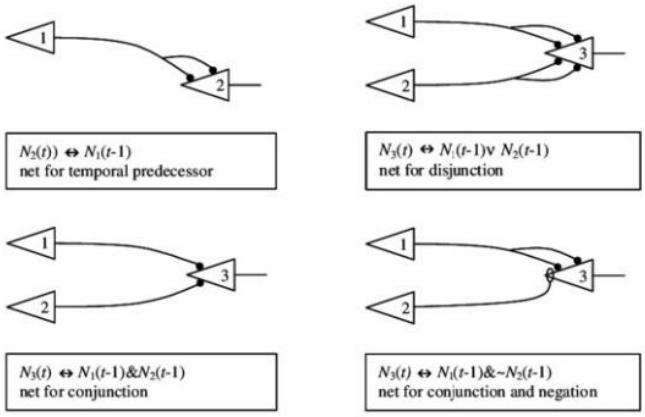
ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ

Εισαγωγή

1.1 Τεχνητή Νοημοσύνη – Μηχανική Μάθηση

Η σύλληψη της ιδέας της υπολογιστικής νοημοσύνης αποδίδεται στον Alan Turing, ο οποίος πρώτος διατύπωσε και προσπάθησε να απαντήσει το ερώτημα αν οι μηχανές (υπολογιστές) μπορούν να σκεφτούν (Turing, 1950), και διατύπωσε την επιδίωξη της «αυτόνομης σκέψης» των υπολογιστών. Λίγα χρόνια νωρίτερα οι Walter Pitts και Warren McCulloch δημοσίευσαν το πρώτο μαθηματικό μοντέλο ενός νευρωνικού δικτύου, όπως παρουσιάζεται στην Εικόνα 1, που δημιουργεί αλγορίθμους που μιμούνται τη διαδικασία της ανθρώπινης σκέψης (Mcculloch & Pitts, 1943; Piccinini, 2004).



Εικόνα 1. Το μοντέλο νευρωνικού δικτύου των Pitts και McCulloh. Για να στείλει έναν παλμό εξόδου, κάθε νευρώνας πρέπει να λάβει δύο διεγερτικές εισόδους και καμία ανασταλτική είσοδο. Οι γραμμές που τελειώνουν σε τελεία αντιπροσωπεύουν διεγερτικές συνδέσεις- οι γραμμές που τελειώνουν σε στεφάνι αντιπροσωπεύουν ανασταλτικές συνδέσεις. Από: (Mcculloch & Pitts, 1943).

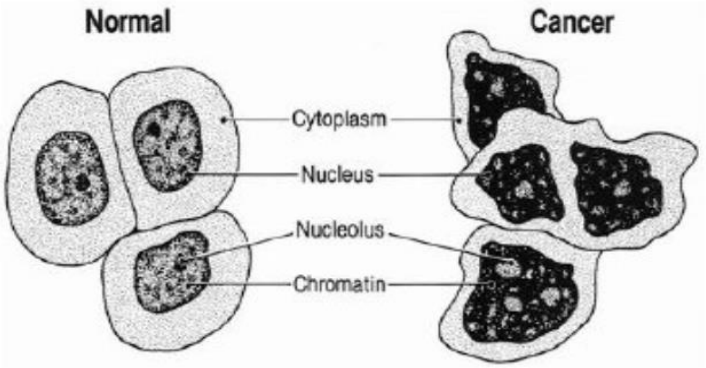
Η μοντελοποίηση της ανθρώπινης σκέψης με απλές λογικές μαθηματικές σχέσεις, σε συνδυασμό με την ανάπτυξη υπολογιστών με την ικανότητα εκτέλεσης μαθηματικών πράξεων, οδήγησε στη σύλληψη του όρου της «τεχνητής νοημοσύνης» (ΤΝ), που χαρακτηρίζει συστήματα που επιδεικνύουν ευφυή συμπεριφορά, αναλύοντας το περιβάλλον τους και αναλαμβάνοντας δράσεις - με κάποιο βαθμό αυτονομίας - για την επίτευξη συγκεκριμένων στόχων (High-Level Expert Group on Artificial Intelligence, 2019). Η τεχνητή νοημοσύνη ορίζεται ως η μίμηση, από τους υπολογιστές, της έμφυτης ανθρώπινης νοημοσύνης (Sheikh et al., 2023).

Η μηχανική μάθηση είναι ένα υποσύνολο της τεχνητής νοημοσύνης που δίνει τη δυνατότητα στους υπολογιστές να μαθαίνουν και να βελτιώνονται από την εμπειρία τους χωρίς να προγραμματίζονται επί τούτου. Περιλαμβάνει αλγορίθμους που επιτρέπουν στα συστήματα

να αναλύουν δεδομένα, να εντοπίζουν μοτίβα και να λαμβάνουν αποφάσεις ή προβλέψεις. Η επιτυχία της βασίζεται στην εκπαίδευση μοντέλων με ποικίλες και αντιπροσωπευτικές πληροφορίες για την ενίσχυση της. Η μηχανική μάθηση αναμένεται να μεταμορφώσει τον εικοστό πρώτο αιώνα (Alam, 2023). Η ραγδαία, πρόσφατη πρόοδος στην υποκείμενη αρχιτεκτονική και τους αλγορίθμους της καθώς και η αύξηση στη διαθέσιμη επεξεργαστική ισχύ και το μέγεθος των συνόλων δεδομένων, έχουν οδηγήσει σε αύξηση της ικανότητας των υπολογιστών σε διάφορους τομείς. Αυτοί περιλαμβάνουν την οδήγηση ενός οχήματος, τη μετάφραση γλωσσών, τα γλωσσικά μοντέλα (Nichols et al., 2019), τη διάγνωση ασθενειών ή την άμυνα κατά κυβερνοεπιθέσεων (D. S. Watson, 2023).

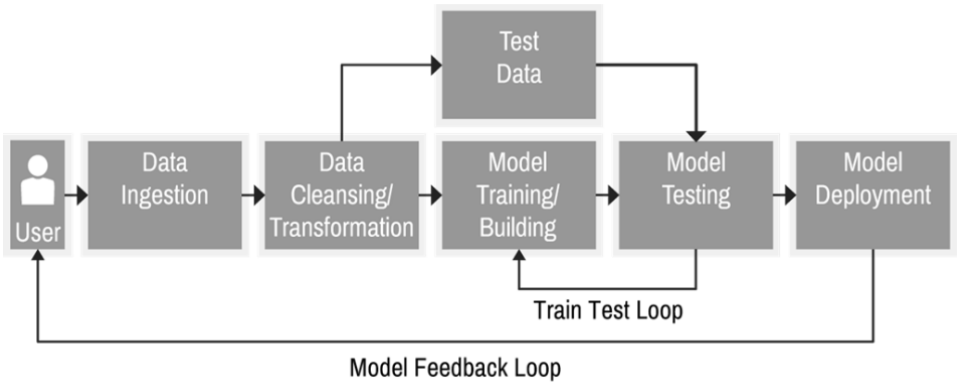
Η μηχανική μάθηση θεωρείται γενικά ότι περιλαμβάνει αυτόματες υπολογιστικές διαδικασίες, βασισμένες σε λογικές ή δυαδικές πράξεις, που μαθαίνουν μια εργασία από μια σειρά παραδειγμάτων (Fulkerson, 1995; Oladipupo, 2010). Πρόκειται για έναν γενικό όρο που αναφέρεται σε ένα ευρύ φάσμα αλγορίθμων που πραγματοποιούν έξυπνες προβλέψεις με βάση ένα σύνολο δεδομένων. Αυτά τα σύνολα δεδομένων είναι συχνά μεγάλα, αποτελούμενα ίσως από εκατομμύρια παρατηρήσεις. Η πρόσφατη πρόοδος στη μηχανική μάθηση έχει επιτύχει αυτό που φαίνεται να είναι ένα ανθρώπινο επίπεδο σημασιολογικής κατανόησης και εξαγωγής πληροφοριών, και μερικές φορές την ικανότητα εντοπισμού μοτίβων με μεγαλύτερη ακρίβεια από τους ανθρώπους εξειδικευμένους πάνω στο εκάστοτε αντικείμενο (Nichols et al., 2019). Σήμερα, η μηχανική μάθηση εστιάζει στην επίλυση των εξής προβλημάτων (Sodhi et al., 2019):

- **Ταξινόμηση:** χρησιμοποιείται για τον προσδιορισμό της κατηγορίας στην οποία ανήκει ένα αντικείμενο. Για παράδειγμα, αν ένα μήνυμα ηλεκτρονικής αλληλογραφίας είναι ανεπιθύμητο ή αν ένα κύτταρο είναι καρκινικό (Εικόνα 2).
- **Παλινδρόμηση:** χρησιμοποιείται για την πρόβλεψη μίας ή περισσοτέρων αριθμητικών τιμών που σχετίζεται με ένα αντικείμενο. Για παράδειγμα, η πρόβλεψη της τιμής της μετοχής.



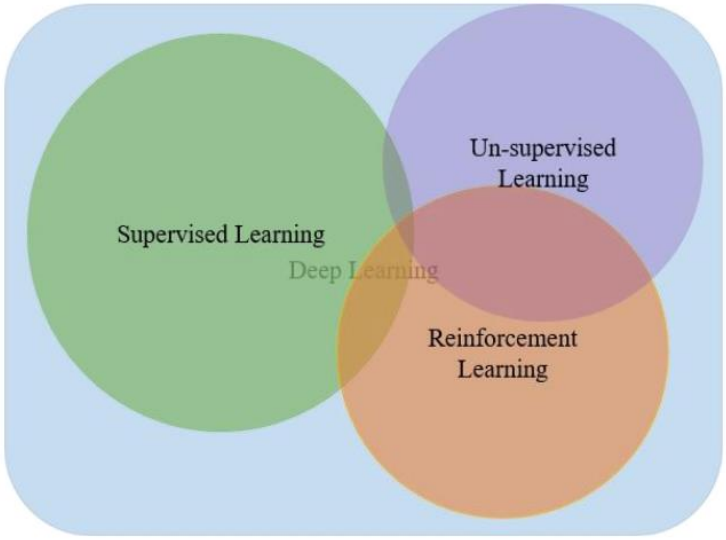
Εικόνα 2. Σύγκριση μεταξύ ενός υγιούς (αριστερά) και ενός καρκινικού (δεξιά) κυττάρου του ανθρώπου. Από:(Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egyp et al., 2014).

- **Ομοιότητα:** χρησιμοποιείται για την ανάκτηση παρόμοιων αντικειμένων ή για την εύρεση ανωμαλιών στη συμπεριφορά. Για παράδειγμα, αναζήτηση παρόμοιων εικόνων ή ανίχνευση εξαπάτησης στη συμπεριφορά του χρήστη.
- **Κατάταξη:** χρησιμοποιείται για την βαθμολόγηση σχετικών δεδομένων σύμφωνα με μια συγκεκριμένη είσοδο. Για παράδειγμα, ο αλγόριθμος PageRank (Brin & Page, 1998) της Google που καθορίζει τη σειρά προβολής των αποτελεσμάτων στη μηχανή αναζήτησής της.



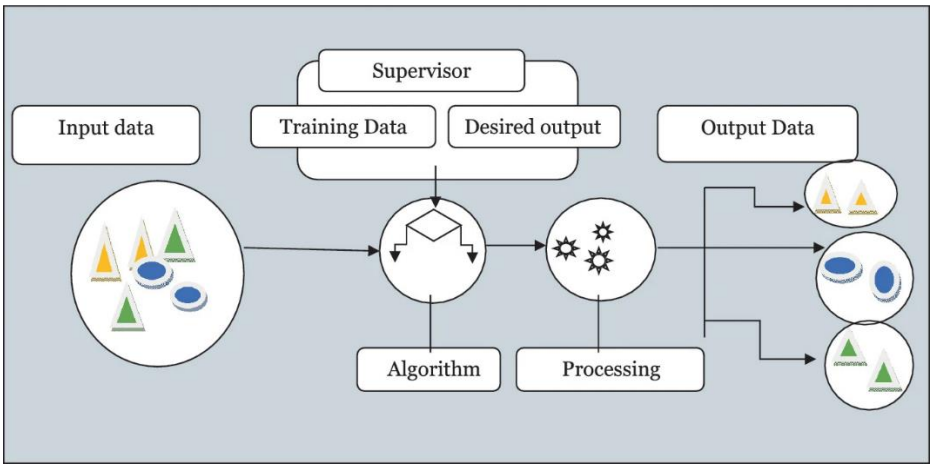
Εικόνα 3. Διάγραμμα ροής ενός μοντέλου μηχανικής μάθησης. Η ροή εργασιών αρχίζει με την είσοδο δεδομένων από το χρήση. Στη συνέχεια πραγματοποιείται καθαρισμός και μετασχηματισμός των δεδομένων σε κατάλληλη μορφή. Τα δεδομένα ελέγχονται και, αν δεν επιφέρουν τα επιθυμητά αποτελέσματα από το μοντέλο μηχανικής μάθησης, τροποποιούνται εκ νέου μέχρι να επιτευχθεί ο στόχος τους. Έχοντας λάβει την κατάλληλη μορφή, τα δεδομένα εισάγονται στο μοντέλο για να εκπαιδευτεί σε αυτά. Ακολουθεί η αξιολόγηση του μοντέλου, με τροποποίηση και βελτιστοποίηση των παραμέτρων εκπαίδευσης ώστε να επιτευχθεί η βέλτιστη απόδοση. Μετά την ολοκλήρωση της βελτιστοποίησης, το μοντέλο καθίσταται διαθέσιμο για χρήση. Από: (Sodhi et al., 2019).

Η γενική αρχή λειτουργίας ενός μοντέλου μηχανικής μάθησης παρουσιάζεται στην Εικόνα 3. Οι αλγόριθμοι μηχανικής μάθησης ταξινομούνται, με βάση τον τρόπο λειτουργίας του αλγορίθμου, στις παρακάτω γενικές κατηγορίες (Εικόνα 4):



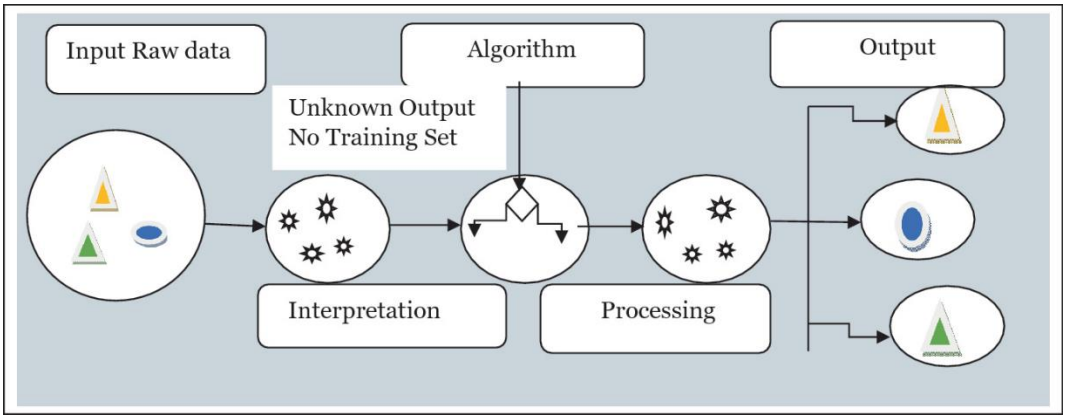
Εικόνα 4. Κατηγορίες της βαθιάς μάθησης. Από: (Alom et al., 2018).

- **Μάθηση με Επίβλεψη** (Εικόνα 5)**:** Ο αλγόριθμος παράγει μια συνάρτηση που αντιστοιχίζει τις εισόδους στις επιθυμητές εξόδους (Q. Liu & Wu, 2012). Ένα τυπικό παράδειγμα προβλήματος που αντιμετωπίζεται με Μάθηση με Επίβλεψη είναι αυτό της ταξινόμησης: το μοντέλο καλείται να μάθει (δηλαδή να εντοπίσει τις καλύτερες παραμέτρους, με βάση κάποιο κριτήριο) μιας συνάρτησης η οποία κατατάσσει ένα διάνυσμα σε μια από πολλές κατηγορίες, εξετάζοντας πολλά παραδείγματα εισόδου-εξόδου της συνάρτησης. Οι αλγόριθμοι όπως: Εγγύτερος Γείτονας, Ταξινομητής Bayes, Δέντρα Αποφάσεων, Γραμμική Παλινδρόμηση, Μηχανές Διανυσμάτων Υποστήριξης και Νευρωνικά Δίκτυα (Haykin & Haykin, 2009) είναι τα πιο χαρακτηριστικά παραδείγματα όπου εφαρμόζεται Μάθηση με Επίβλεψη (Hoda, 2016).



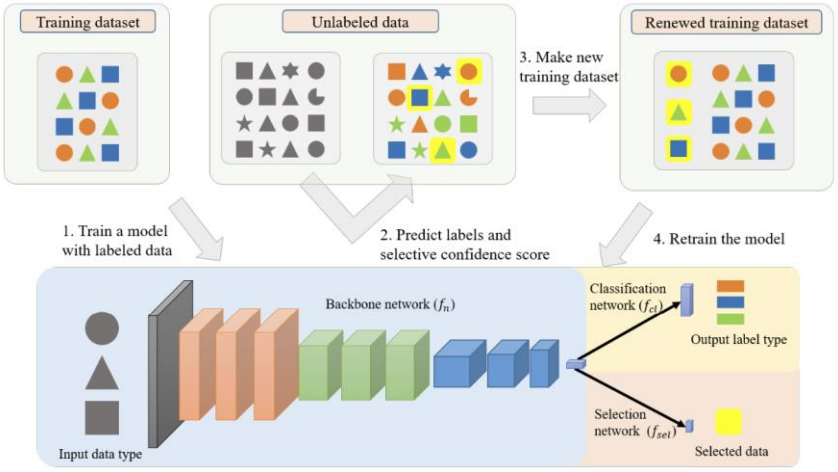
Εικόνα 5. Διαγραμματική απεικόνιση της αρχιτεκτονικής λειτουργίας ενός αλγορίθμου μηχανικής μάθησης με επίβλεψη. Για την εκπαίδευση του μοντέλου χρησιμοποιούνται επισημασμένα σύνολα δεδομένων με το επιθυμητό αποτέλεσμα. Το μοντέλο εκπαιδεύεται να διαχειρίζεται ανάλογης μορφής δεδομένα εισόδου, και να εξάγει κάποιο από τα επιθυμητά αποτελέσματα πάνω στα οποία έχει εκπαιδευτεί. Οποιασδήποτε μορφής δεδομένα και αν εισαχθούν σε ένα μοντέλο μηχανικής μάθησης με επίβλεψη, η έξοδός του θα είναι κάποια από αυτές για τις οποίες έχει εκπαιδευθεί. Από: (Karuppusamy et al., 2022).

- **Μάθηση χωρίς Επίβλεψη** (Εικόνα 6)**:** Στον αλγόριθμο παρέχονται μη επισημασμένα σύνολα δεδομένων εισόδου και ο αλγόριθμος μάθησης, χωρίς επίβλεψη, παράγει μια συνάρτηση για τον εντοπισμό κρυφών δομών στο σύνολο δεδομένων σύμφωνα με πρότυπα, ομοιότητες και διαφορές που υπάρχουν μεταξύ των δεδομένων, χωρίς προηγούμενη εκπαίδευση (D. S. Watson, 2023). Το Unsupervised Learning είναι εξαιρετικά χρήσιμο, διότι υπάρχουν πάρα πολλά, και συνεχώς παράγονται περισσότερα, μη επισημασμένα δεδομένα, η επισήμανση των οποίων είναι χρονοβόρα και κοστοβόρα. Επίσης, μπορεί να χρησιμοποιηθεί για τη διερεύνηση άγνωστων ή μη επεξεργασμένων δεδομένων (S. Naeem et al., 2023).



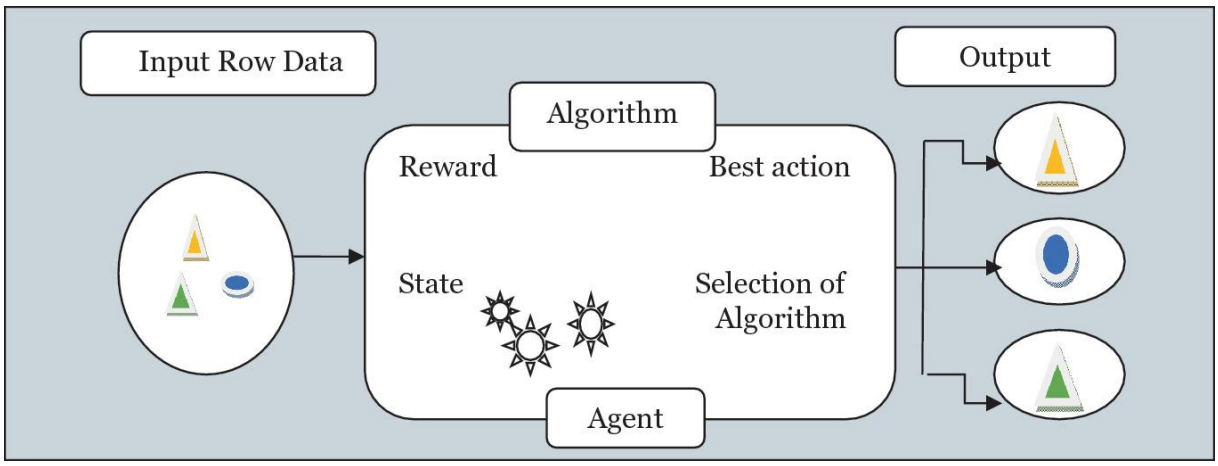
Εικόνα 6. Διαγραμματική απεικόνιση της αρχιτεκτονικής λειτουργίας ενός αλγορίθμου μηχανικής μάθησης χωρίς επίβλεψη. Σε αυτή την περίπτωση δεν υπάρχει επισημασμένο σύνολο δεδομένων εκπαίδευσης, και είναι αρμοδιότητα του μοντέλου η «ερμηνεία» των εισαγόμενων δεδομένων και η εξαγωγή αποτελεσμάτων. Οι πιθανές έξοδοι του μοντέλου δεν είναι γνωστές από πριν. (Karuppusamy et al., 2022).

- **Ημι-επιβλεπόμενη Μάθηση** (Εικόνα 7)**:** Συνδυάζει τόσο επισημασμένα όσο και μη επισημασμένα δεδομένα (C A Padmanabha Reddy et al., 2018).



Εικόνα 7. Διαγραμματική απεικόνιση της αρχιτεκτονικής λειτουργίας ενός αλγορίθμου ημι-επιβλεπόμενης μάθησης. Από: Semi-Supervised Learning. (Jeong et al., 2020).

- **Ενισχυτική Μάθηση (Reinforcement Learning)** (Εικόνα 8)**:** Ο αλγόριθμος μαθαίνει μια πολιτική για το πώς να ενεργεί, δεδομένης μιας εξωτερικής παρατήρησης. Κάθε ενέργεια έχει κάποιο αντίκτυπο στο περιβάλλον και το περιβάλλον παρέχει ανατροφοδότηση που καθοδηγεί τον αλγόριθμο μάθησης (M. Naeem et al., 2020).

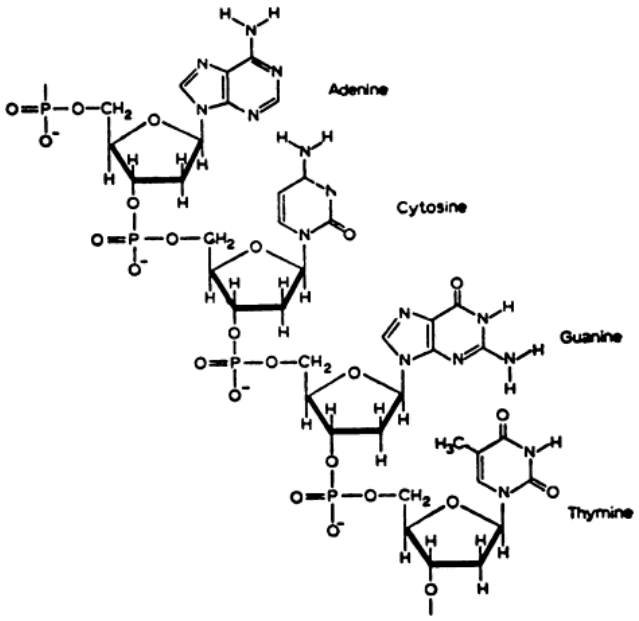


Εικόνα 8. Διαγραμματική απεικόνιση της ενισχυτικής μάθησης. Ο αλγόριθμος λαμβάνει επιβράβευση όταν παράγει το επιθυμητό αποτέλεσμα, η οποία δεν λαμβάνεται όταν παράγει άλλα αποτελέσματα, πέραν του επιθυμητού. Με τον τρόπο αυτό το μοντέλο εκπαιδεύεται για την παραγωγή των βέλτιστων αποτελεσμάτων μέσα από τη διαδικασία της επιβράβευσης. Από: (Karuppusamy et al., 2022).

1.2 Βιολογικά δεδομένα

Η επιστήμη της Βιολογίας ορίζεται ως η επιστήμη που μελετά τη ζωή, ή αλλιώς τα ζωντανά συστήματα. Ένα ζωντανό σύστημα είναι ένα μοναδικό σύνολο χημικών ουσιών που είναι ικανό να χρησιμοποιεί την ενέργεια για να οργανώνει την ύλη σύμφωνα με το πληροφοριακό της περιεχόμενο με τρόπο τέτοιο που οδηγεί στην αυτοσυντήρηση (Brown & Brown, 1984). Ο όρος απαντάται για πρώτη φορά σε επιστημονικά κείμενα από τον J.B. Lamarck το 1802 στο βιβλίο του «Hydrogeology», (Lamarck, 1802). Οι ρίζες, ωστόσο, της επιστήμης, πολύ πριν την επινόηση του όρου, βρίσκονται στην Αρχαία Ελλάδα και τον Αριστοτέλη, ο οποίος έχει αναγνωριστεί και είναι κοινός αποδεκτός ως ο «πατέρας της Βιολογίας και Ζωολογίας» (McRae, 1890).

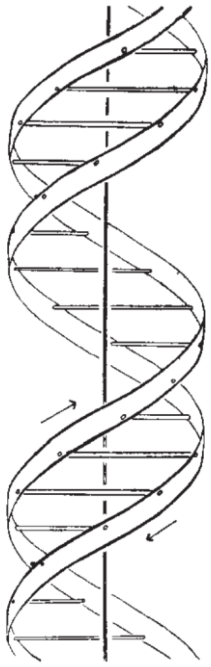
Η επιστήμη της βιολογίας, με τις ρίζες της στην αρχαία Ελλάδα και τις αρχικές παρατηρήσεις του Αριστοτέλη, εξελίχθηκε σταδιακά, αγγίζοντας θεμελιώδη ερωτήματα για τη φύση της ζωής. Η τυποποίηση του όρου "βιολογία" από τον J.B. Lamarck το 1802 σηματοδότησε την έναρξης μιας νέας εποχής, θέτοντας τις βάσεις για μια πιο συστηματική και επιστημονική μελέτη των ζωντανών οργανισμών.



Εικόνα 9. Η χημική δομή του νουκλεϊκού οξέος, λίγο μετά την ανακάλυψη της τριτοταγούς δομής του. Από: (Todd, 1954).

Ωστόσο, η πιο κομβική στιγμή στην ιστορία της βιολογίας ήταν το 1953 με την αποκάλυψη της τριτοταγούς δομής του DNA (J. Watson & Crick, 1953), όπως φαίνεται στην Εικόνα 10. Αυτή η επαναστατική ανακάλυψη αποκάλυψε το μυστήριο της κληρονομικότητας, αποδεικνύοντας ότι το DNA, με τη διπλή έλικα και τις βάσεις του (Εικόνα 9), αποτελεί τον

γενετικό κώδικα που καθορίζει τα χαρακτηριστικά κάθε οργανισμού. Η αποκρυπτογράφηση του DNA άνοιξε τον δρόμο για μια άνευ προηγουμένου έκρηξη γνώσης, θέτοντας τα θεμέλια για την ανάπτυξη της σύγχρονης βιολογίας.



Εικόνα 10. Το μοντέλο της διπλής έλικας που περιγράφει την τριτοταγή δομή του DNA, όπως δημοσιεύθηκε στο πρωτότυπο άρθρο των Watson και Crick το 1953.

Σήμερα, η βιολογία έχει εξελιχθεί σε μια πολυδιάστατη επιστήμη, με κλάδους όπως η γενετική, η βιοχημεία, η μοριακή βιολογία, η βιοτεχνολογία, η βιοπληροφορική και η νευροβιολογία να ακμάζουν. Η επανάσταση του DNA έθεσε τα θεμέλια για μια εποχή ραγδαίας προόδου στη βιολογία, με απίστευτες δυνατότητες για την κατανόηση της ζωής, την αντιμετώπιση ασθενειών και τη βελτίωση της ανθρώπινης υγείας. Πλέον, η βιολογία δεν περιορίζεται μόνο στα εργαστήρια, αλλά επηρεάζει πολλές πτυχές της ζωής των ανθρώπων, από την ιατρική και τη φαρμακευτική έως τη γεωργία και την προστασία του περιβάλλοντος.

Η επανάσταση του DNA έθεσε τα θεμέλια για μια νέα εποχή στη βιολογία, όχι μόνο αποκαλύπτοντας τα μυστήρια της κληρονομικότητας, αλλά και ανοίγοντας τον δρόμο για την παραγωγή τεράστιας ποσότητας πληροφορίας. Η ανάπτυξη τεχνολογιών αλληλούχισης γονιδιώματος, όπως η Next-Generation Sequencing (NGS), επέτρεψε στους επιστήμονες να αποκρυπτογραφήσουν το DNA οργανισμών με τεράστια ταχύτητα και ακρίβεια.

Σημαντικότατο ρόλο στην εδραίωση των σύγχρονων μεθοδολογιών και την ταχεία πρόοδο της επιστήμης της βιολογίας έπαιξε η υλοποίηση του Human Genome Project (International Human Genome Sequencing Consortium et al., 2001). Το Human Genome Project ήταν μια διεθνής ερευνητική πρωτοβουλία που ξεκίνησε το 1990 και ολοκληρώθηκε το 2003. Ο κύριος στόχος του ήταν η χαρτογράφηση και η αποκωδικοποίηση ολόκληρου του ανθρώπινου γονιδιώματος, δηλαδή η αλληλουχία περίπου 3 δισεκατομμυρίων βάσεων του DNA που συνθέτουν τον ανθρώπινο οργανισμό (Hood & Rowen, 2013). Το HGP παρείχε τη βάση για την κατανόηση της γενετικής σύνθεσης των ανθρώπων και έφερε επανάσταση στον τομέα της βιολογίας και της βιοπληροφορικής.

Το HGP δημιούργησε τεράστιες ποσότητες γενετικών δεδομένων, τα οποία αποτελούν μια από τις βασικές πηγές βιολογικών δεδομένων που χρησιμοποιούνται σήμερα. Οι βάσεις δεδομένων που δημιουργήθηκαν από το HGP επιτρέπουν στους επιστήμονες να ερευνούν γονίδια, να αναλύουν γενετικές μεταλλάξεις και να προβλέπουν την πιθανότητα εμφάνισης διαφόρων ασθενειών. Πηγές δεδομένων που προέκυψαν από το HGP, όπως το GenBank και το NCBI, έχουν τεράστια σημασία στη βιοϊατρική έρευνα και την ανάπτυξη νέων φαρμάκων (Nurk et al., 2022).

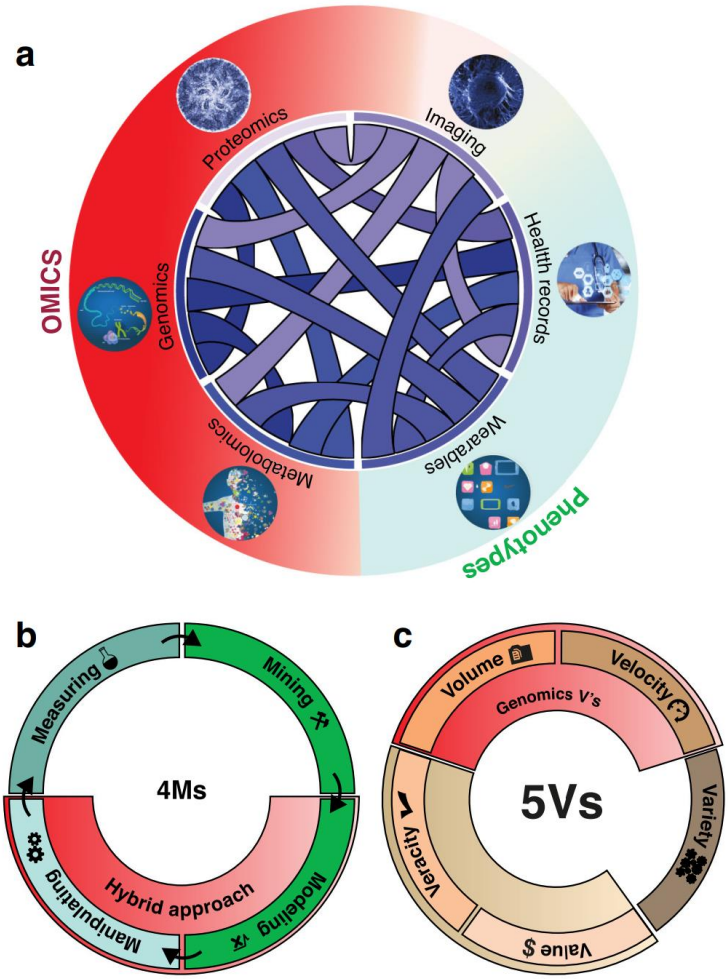
Το έργο αυτό αποτέλεσε τη βάση για την ανάπτυξη και άλλων γονιδιωματικών πρωτοβουλιών, όπως η Next-Generation Sequencing (NGS), η οποία επιτρέπει την αλληλούχιση γονιδιωμάτων με μεγαλύτερη ταχύτητα και ακρίβεια. Σήμερα, οι τεχνικές που βασίζονται στο HGP χρησιμοποιούνται ευρέως στην εξατομικευμένη ιατρική, όπου οι θεραπείες σχεδιάζονται με βάση το γενετικό προφίλ κάθε ασθενούς.

Αυτή η πληθώρα βιολογικών δεδομένων, που ονομάζονται και "ομικά δεδομένα" δημιούργησε νέες προκλήσεις και ευκαιρίες για την βιολογική έρευνα. Η ανάλυση αυτών των δεδομένων, απαιτεί προηγμένες μεθόδους βιοπληροφορικής και τεχνητής νοημοσύνης. Πλέον, η βιολογία βασίζεται σε μεγάλο βαθμό στα δεδομένα, με την εξαγωγή χρήσιμων πληροφοριών από τεράστιες ποσότητες ομικών δεδομένων να αποτελεί το κλειδί για την κατανόηση πολύπλοκων βιολογικών φαινομένων. Αυτή η στροφή έχει φέρει επανάσταση σε διάφορους τομείς της βιολογίας, από την ανάπτυξη νέων φαρμάκων και θεραπειών έως την πρόβλεψη ασθενειών και την εξατομικευμένη ιατρική. Η βιολογία των δεδομένων έχει ανοίξει νέους ορίζοντες για την κατανόηση της ζωής και την αντιμετώπιση των σημαντικότερων προκλήσεων που αντιμετωπίζει η ανθρωπότητα.

Για την πληρέστερη κατανόηση πολύπλοκων βιολογικών φαινομένων, όπως πολλές ανθρώπινες ασθένειες ή ποσοτικά χαρακτηριστικά σε ζώα/φυτά, παράγονται τεράστιες ποσότητες και πολλαπλοί τύποι «μεγάλων» δεδομένων από πολύπλοκες μελέτες. Μέχρι πρόσφατα, λόγω του υψηλού κόστους και των δυσκολιών υλοποίησης των πειραματικών διαδικασιών αλληλούχισης γονιδιώματος, η παραγωγή βιολογικών δεδομένων ήταν το σημείο συμφόρησης της έρευνας, ενώ τώρα είναι η εξόρυξη δεδομένων ή η εξαγωγή χρήσιμων βιολογικών πληροφοριών από μεγάλα, περίπλοκα σύνολα δεδομένων (Xu & Jackson, 2019).

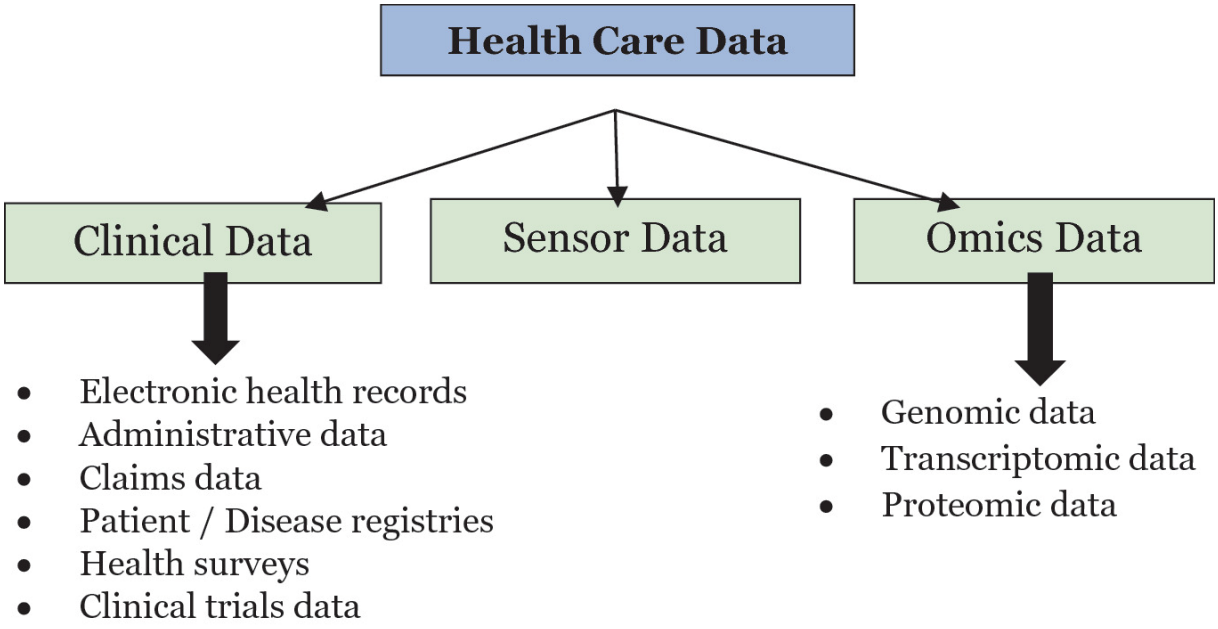
Ο μαζικός όγκος δεδομένων αλληλούχισης DNA και RNA, το αποτέλεσμα των τεχνολογιών υψηλής απόδοσης στη μοριακή και κυτταρική βιολογία που παράγουν τεράστιο όγκο - ομικών δεδομένων (omics) που σχετίζονται με γονιδιώματα, μεταγραφώματα, πρωτεϊνώματα, μεταβολισμούς, αλληλεπιδράσεις κ.λπ. έχουν φέρει μεγάλες προκλήσεις «μεγάλων δεδομένων» και στη βιολογία. Τα τρέχοντα έργα αλληλούχισης γονιδιωμάτων (Whole Genome Sequencing – WGS) οδηγούν σε μεγάλη αύξηση δεδομένων αλληλούχισης (Pal et al., 2020). Εκτός από τα ομικά δεδομένα, βιομετρικά, απεικονιστικά και κλινικά δεδομένα παράγονται επίσης με ταχύτατο ρυθμό σήμερα. Οι πηγές και τα χαρακτηριστικά των μεγάλων βιολογικών δεδομένων συνοψίζονται στην Εικόνα 11.

Η υπολογιστική βιολογία, που είναι ο ερευνητικός τομέας όπου συναντώνται η μοριακή βιολογία και η επιστήμη των υπολογιστών, λόγω της διεπιστημονικής της φύσης έδωσε μια νέα οπτική στα προβλήματα και των δύο πλευρών. Η βιολογική πλευρά εφοδιάστηκε με μοντέλα και μεθόδους από την επιστήμη των υπολογιστών, που επιτρέπουν την αποτελεσματική επίλυση βιολογικών προβλημάτων. Η πλευρά της επιστήμης των υπολογιστών κέρδισε επίσης πολλά από αυτή τη διασταύρωση, μεταξύ άλλων ορίστηκαν και αναλύθηκαν νέες κατηγορίες γραφημάτων χρήσιμων για τη μοντελοποίηση προβλημάτων (Blazewicz & Kasprzak, 2012).



Εικόνα 11. Σύνοψη της επιστήμης των βιοϊατρικών δεδομένων. a) Η επιστήμη των βιοϊατρικών δεδομένων προέκυψε ως αποτέλεσμα της δημιουργίας συνόλων δεδομένων μεγάλης κλίμακας που συνδέουν τη γονιδιωματική, τη μεταβολική, τις φορητές συσκευές, την πρωτεομική, τα αρχεία υγείας και την απεικόνιση με τη στατιστική και την επιστήμη των υπολογιστών. b) Το πλαίσιο των 4 M διαδικασιών για την αξιοποίηση των βιοϊατρικών δεδομένων. c) Τα 5 V των μεγάλων δεδομένων. Από (Navarro et al., 2019)

1.3 Πηγές Βιολογικών Δεδομένων



Εικόνα 12. Πηγές προέλευσης βιοϊατρικών δεδομένων. Από: (Karuppusamy et al., 2022).

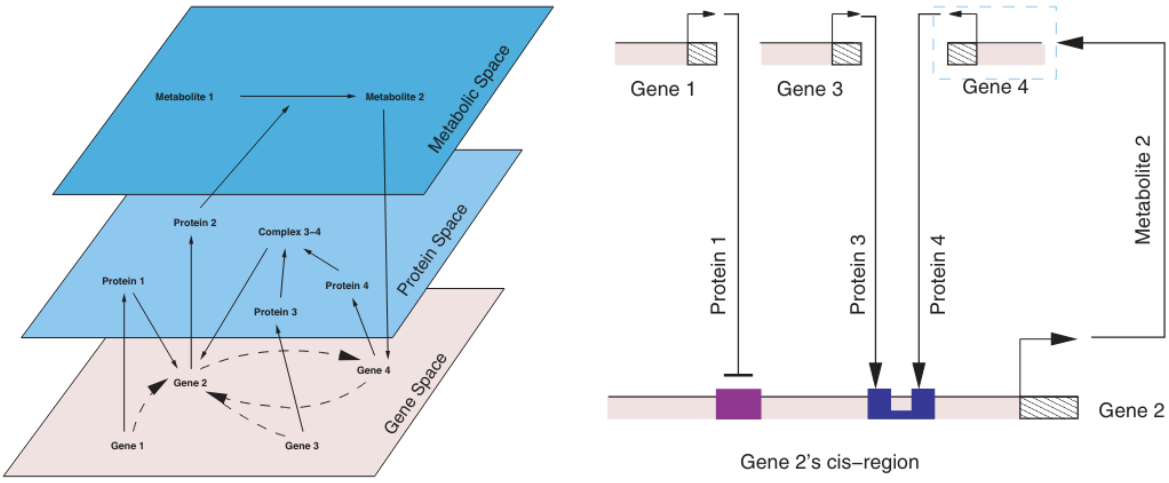
Η επιστήμη της βιολογίας αποτελεί μια πλούσια πηγή δεδομένων που μπορούν να χρησιμοποιηθούν για την ανάλυση και την κατανόηση σύνθετων βιολογικών συστημάτων. Ορισμένες κατηγορίες βιολογικών δεδομένων είναι:

i. **Αλληλεπιδράσεις πρωτεϊνών:** Μπορούν να αναπαρασταθούν ως δίκτυα πρωτεϊνών που αποτυπώνουν τις δομικές και λειτουργικές αλληλεπιδράσεις μεταξύ πρωτεϊνών σε ένα κύτταρο ή οργανισμό (Safari-Alighiarloo et al., 2014). Τα δυναμικά δίκτυα πρωτεϊνών εξετάζουν τις αλλαγές στις αλληλεπιδράσεις μεταξύ των πρωτεϊνών σε διαφορετικές συνθήκες ή χρονικές περιόδους (πχ λήψη φαρμάκου, μεταβολές περιβαλλοντικών παραγόντων, στρες, ανάπτυξη οργανισμού, εξέλιξη πορείας νόσου κλπ) (Noori et al., 2022). Υπάρχουν διάφορες βάσεις δεδομένων στο διαδίκτυο, από τις οποίες μπορούν να αντληθούν πληροφορίες για δίκτυα πρωτεϊνικών αλληλεπιδράσεων (Farooq et al., 2021), όπως BioGRID (Oughtred et al., 2019), STRING (Szklarczyk et al., 2023), HPIDB 2.0 (Ammari et al., 2016), DIP (Salwinski, 2004), IntAct (Orchard et al., 2014), HPRD (Keshava Prasad et al., 2009).

ii. **Δίκτυα γονιδιακής έκφρασης:** Τα ρυθμιστικά δίκτυα γονιδιακής έκφρασης (GRNs, Gene Regulatory Networks), περιλαμβάνουν αλληλεπιδράσεις μεταξύ μεγάλου αριθμού γονιδίων και των ρυθμιστών τους (μεταγραφικοί παράγοντες, οπερόνια, παράγοντες συνέκφρασης κλπ) και συνήθως αποτυπώνονται σε διαγράμματα που χρησιμοποιούνται για την οπτικοποίηση των ρυθμιστικών σχέσεων (MacNeil &

Walhout, 2011). Ένα δίκτυο γονιδιακής ρύθμισης αποτελείται από γονίδια, cis-στοιχεία και ρυθμιστές. Οι ρυθμιστές είναι τις περισσότερες φορές πρωτεΐνες, που ονομάζονται παράγοντες μεταγραφής, αλλά μικρά μόρια, όπως RNA και μεταβολίτες, συμμετέχουν επίσης μερικές φορές στη συνολική ρύθμιση. Οι αλληλεπιδράσεις και η πρόσδεση των ρυθμιστών στα cis-στοιχεία στην cis-περιοχή των γονιδίων ελέγχουν το επίπεδο της γονιδιακής έκφρασης κατά τη διάρκεια της μεταγραφής. Οι cis-περιοχές χρησιμεύουν για τη συγκέντρωση των σημάτων εισόδου, με τη μεσολάβηση των ρυθμιστών, και έτσι επιφέρουν ένα πολύ συγκεκριμένο σήμα γονιδιακής έκφρασης. Τα γονίδια, οι ρυθμιστές και οι ρυθμιστικές συνδέσεις μεταξύ τους, μαζί με ένα σχήμα ερμηνείας σχηματίζουν γονιδιακά δίκτυα (Εικόνα 13)(Filkov, 2005).

Οι πληροφορίες για την έκφραση των γονιδίων στους επιμέρους ιστούς μπορούν να εξαχθούν είτε με μικροσυστοιχίες (microarrays), είτε με αλληλούχιση RNA (RNA Sequencing). Το πλεονέκτημα της χρήσης των μικροσυστοιχιών είναι ο ταυτόχρονος υπολογισμός του επιπέδου έκφρασης για χιλιάδες γονίδια, ενώ η αλληλούχιση RNA παρέχει μια πιο λεπτομερή ανάλυση της έκφρασης συγκεκριμένων γονιδίων, ισομόρφων mRNA και μη κωδικού RNA (ncRNA). Ανάκτηση δεδομένων γονιδιακής έκφρασης μπορεί να πραγματοποιηθεί από πηγές όπως NCBI (Cantelli et al., 2022), EBI (Sayers et al., 2022), Bgee (Bastian et al., 2021).



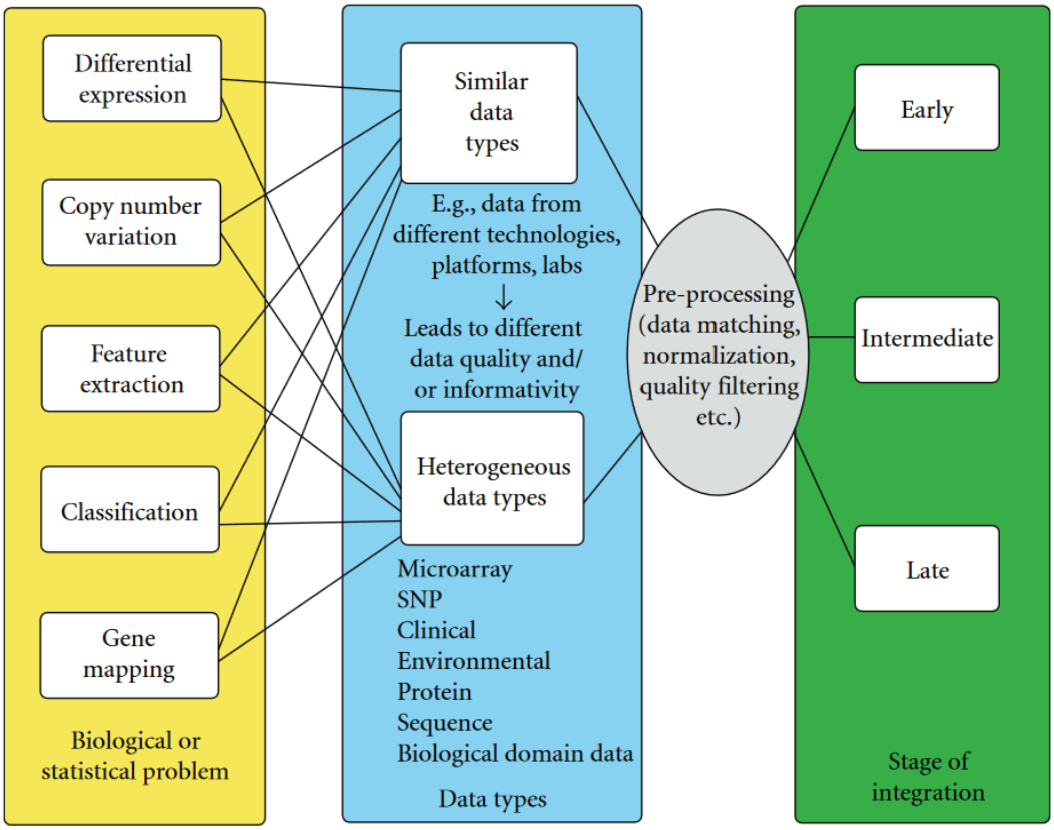
Εικόνα 13. Παράδειγμα ενός υποθετικού γονιδιακού δικτύου. Αριστερά παρουσιάζονται τα διάφορα επίπεδα στα οποία ρυθμίζεται η έκφραση και η λειτουργία ενός γονιδίου από άλλα γονίδια, πρωτεΐνες και μεταβολίτες. Δεξιά παρουσιάζεται ένα διάγραμμα σε μορφή δικτύου που παρουσιάζει τις αλληλεπιδράσεις μεταξύ των γονιδίων 1, 3 και 4, καθώς και των προϊόντων τους (πρωτεϊνών) στην έκφραση του γονιδίου 1. Από: (Filkov, 2005).

iii. **Μεταβολικά Δίκτυα:** Η μεταβολική διαδικασία είναι απαραίτητη για τη διατήρηση της ζωής. Στο μεταβολισμό ορισμένα υλικά διασπώνται για να αποδώσουν ενέργεια για τις ζωτικές διεργασίες, ενώ άλλες ουσίες, απαραίτητες για τη ζωή, συντίθενται. Το μεταβολικό δίκτυο, ένα χαρακτηριστικό πολύπλοκο δίκτυο που περιλαμβάνει όλους τους μεταβολίτες και τις ενζυμικά καταλυόμενες αντιδράσεις που συμβαίνουν σε ένα ζωντανό κύτταρο, καθώς και τις αλληλεπιδράσεις μεταξύ των αντιδρώντων και των ενζύμων, είναι μια αφηρημένη αναπαράσταση του κυτταρικού μεταβολισμού. Η ανάλυση των μεταβολικών δικτύων μπορεί να συμβάλει στην κατανόηση και αξιοποίηση της κυτταρικής μεταβολικής διαδικασίας, προκειμένου να προωθηθεί η ανάπτυξη της τεχνολογίας των ζυμώσεων και της ιατρικής βιομηχανίας. Από την άλλη πλευρά, η τοπολογία των μεταβολικών δικτύων αντανακλά τη δυναμική του σχηματισμού και της εξέλιξής τους. Η μελέτη αυτού του πεδίου μπορεί να βοηθήσει στην κατανόηση της εξελικτικής ιστορίας της ζωής (Zhao et al., 2006).

   Τα μεταβολικά δίκτυα, επομένως, αναπαριστούν τις χημικές μετατροπές που λαμβάνουν χώρα σε ένα κύτταρο ή οργανισμό (Ramon & Stelling, 2023). Δεδομένα μεταβολικών δικτύων μπορούν να αντληθούν από βάσεις δεδομένων όπως reg-genome, genexplain, KEGG (Kanehisa & Goto, 1999), BioCyc (Karp et al., 2019).

iv. **Βιοϊατρικά δεδομένα:** Μπορούν να είναι είτε κλινικά δεδομένα ασθενών, όπως κλινική εικόνα, συμπτωματολογία, αποτελέσματα εξετάσεων, ιατρικό ιστορικό, διαγνώσεις, θεραπευτικές αγωγές και αποτελέσματα, είτε και δεδομένα εικόνων ιατρικής απεικόνισης, όπως ακτινογραφίες, αξονικές/μαγνητικές τομογραφίες, υπέρηχοι και άλλες ιατρικές εικόνες (Shortliffe & Barnett, 2006).

Η ύπαρξη πλούσιων πηγών βιολογικών δεδομένων, όπως παρουσιάζεται στην Εικόνα 14, ανοίγει νέες δυνατότητες για την ανάλυση και την κατανόηση σύνθετων βιολογικών συστημάτων (Hamid et al., 2008). Η επιλογή της κατάλληλης πηγής δεδομένων εξαρτάται από το είδος της μελέτης και το ερώτημα που διερευνάται.



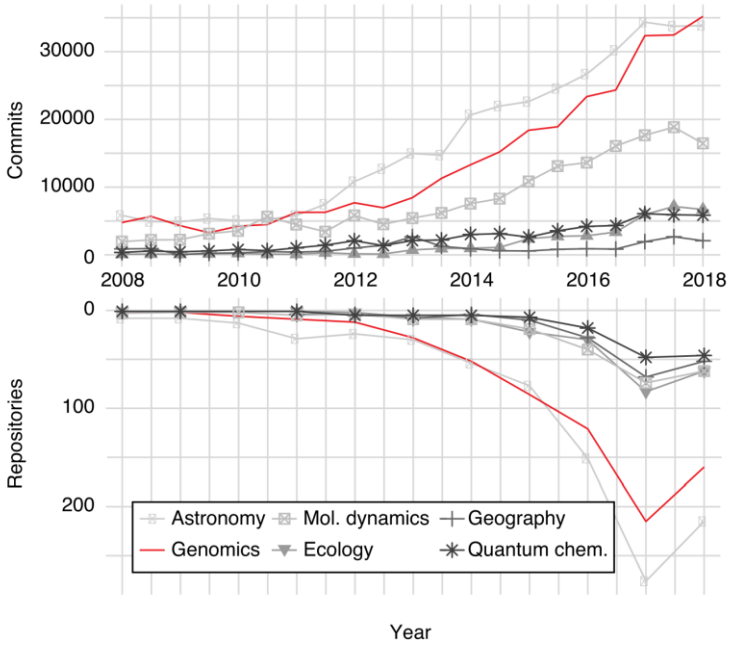
Εικόνα 14. Εννοιολογικό πλαίσιο για την ενσωμάτωση δεδομένων στη γενετική και τη γονιδιωματική. Από: (Hamid et al., 2008).

1.4 Προκλήσεις στην επεξεργασία και ανάλυση βιολογικών δεδομένων

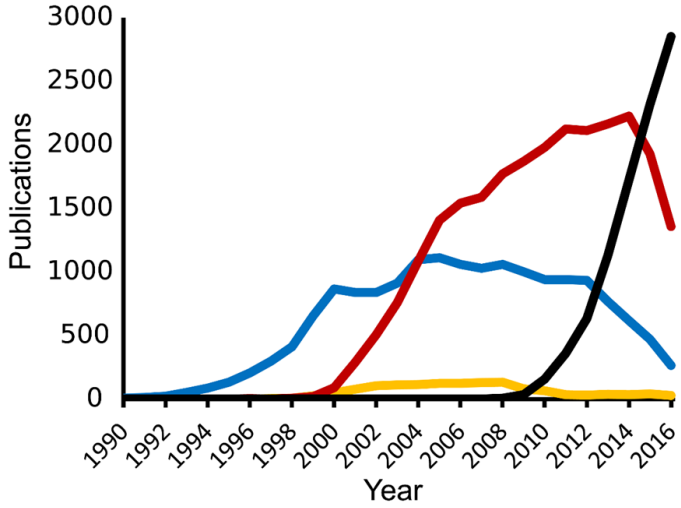
Η επεξεργασία και η ανάλυση βιολογικών δεδομένων αποτελεί ένα συναρπαστικό αλλά και απαιτητικό πεδίο. Η ύπαρξη πλούσιων και ετερογενών δεδομένων φέρνει στο προσκήνιο ορισμένες προκλήσεις που χρήζουν ιδιαίτερης προσοχής.

i. Όγκος και ετερογένεια δεδομένων:

Η ταχύτητα με την οποία παράγονται βιολογικά δεδομένα είναι ιλιγγιώδης (Εικόνα 15, Εικόνα 16 και Εικόνα 18), θέτοντας ζητήματα αποθήκευσης, διαχείρισης και, κυρίως, επεξεργασίας. Την τελευταία δεκαετία ιδίως παρατηρήθηκε μια έκρηξη στον όγκο των διαθέσιμων δεδομένων βιολογικών αλληλουχιών, λόγω της ταχείας προόδου των συσκευών αλληλούχισης υψηλής απόδοσης (Lowe et al., 2017). Ωστόσο, ο όγκος των βιολογικών δεδομένων γίνεται τόσο μεγάλο που οι παραδοσιακές πλατφόρμες και μέθοδοι ανάλυσης δεδομένων δεν μπορούν πλέον να ανταποκριθούν στην ανάγκη ταχείας εκτέλεσης εργασιών ανάλυσης δεδομένων στις βιοεπιστήμες (Yin et al., 2017).



Εικόνα 15. Υιοθέτηση του ανοικτού κώδικα στη γονιδιωματική και σε άλλες υποκατηγορίες της επιστήμης των δεδομένων. Ο αριθμός των κοινοποιήσεων στο GitHub (επάνω) και των νέων αποθετηρίων στο GitHub (κάτω) ανά έτος για διάφορα υποπεδία. Τα αποθετήρια των υποπεδίων επιλέχθηκαν με βάση θεματικές κατηγορίες του GitHub, όπως η γονιδιωματική, η αστρονομία, η γεωγραφία, η μοριακή δυναμική (Mol. Dynamics), η κβαντική χημεία (Quantum Chem.) και η οικολογία. Από: (Navarro et al., 2019).



Εικόνα 16. Δημοσιευμένες εργασίες από το 1990, που αναφέρονται σε αλληλούχιση RNA (μαύρο), σε μικροσυστοιχίες RNA (κόκκινο), σε αλληλουχίες ετικετών (expressed sequence tag) (μπλε) και σε σειριακή ανάλυση γονιδιακής έκφρασης (κίτρινο). Από: (Lowe et al., 2017).

Επειδή τα βιολογικά δεδομένα προέρχονται από ποικίλες πηγές, με διαφορετικές μορφές και δομές, από διαφορετικές πειραματικές διαδικασίες, με τη χρήση διαφορετικού εξοπλισμού ή/και μεταβαλλόμενες περιβαλλοντικές συνθήκες και άλλους αστάθμητους παράγοντες, η συγκέντρωση, ενοποίηση και ανάλυσή τους καθίσταται ιδιαίτερα δύσκολη και απαιτητική.

ii. Θόρυβος και σφάλματα:

Τα βιολογικά δεδομένα συχνά επηρεάζονται από θόρυβο και σφάλματα μέτρησης, που οφείλονται σε τεχνικούς περιορισμούς ή βιολογική μεταβλητότητα. Έλλειψη συντήρησης των ιατρικών μηχανημάτων, ελλιπής βαθμονόμηση, αποκλίσεις κατά την εφαρμογή των πειραματικών πρωτοκόλλων και ατελής σχεδιασμός ενός πειράματος μπορεί να εισάγουν σημαντικά λάθη στα παραγόμενα δεδομένα (Ishak & Salim, 2006). Η ύπαρξη σφαλμάτων στα δεδομένα μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα και να επηρεάσει την αξιοπιστία των αναλύσεων.

Ο θόρυβος αποτελεί, επίσης, σημαντικό ζήτημα, καθώς είναι πολύ δύσκολο να αντιμετωπιστεί εκ των υστέρων, ενώ η εισαγωγή του στα δεδομένα μπορεί να οφείλεται σε αστάθμητους περιβαλλοντικούς παράγοντες. Συνολικά, γίνονται προσπάθειες αποσφαλμάτωσης, αποθορυβοποίησης και ομογενοποίησης βιολογικών δεδομένων διαφορετικών πηγών και προελεύσεων, καθώς αποτελεί απαραίτητο βήμα για τη μετάβαση σε μετααναλύσεις και συγκριτικές μελέτες μεγάλης κλίμακας (Khanal et al., 2021).

iii. Ερμηνεία αποτελεσμάτων:

Η ταχύτατη εξέλιξη της βιολογικής γνώσης απαιτεί συνεχή ενημέρωση και προσαρμογή των μεθόδων ανάλυσης και ερμηνείας. Επίσης, η βιολογία χαρακτηρίζεται από σύνθετες και εξαιρετικά πολύπλοκες αλληλεπιδράσεις σε διάφορα επίπεδα, πολλές από τις οποίες δεν έχουν ανακαλυφθεί ακόμη, καθιστώντας δύσκολη την αποκάλυψη των υποκείμενων μηχανισμών.

Λόγω των ταχέων τεχνολογικών εξελίξεων, έχουν καταστεί διαθέσιμοι διάφοροι τύποι γονιδιωματικών και πρωτεομικών δεδομένων με διαφορετικά μεγέθη, μορφές και δομές. Μεταξύ αυτών είναι η γονιδιακή έκφραση, ο πολυμορφισμός ενός νουκλεοτιδίου, η διακύμανση του αριθμού αντιγράφων και οι αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης/γονιδίου-γονιδίου. Κάθε ένας από αυτούς τους διαφορετικούς τύπους δεδομένων παρέχει μια διαφορετική, εν μέρει ανεξάρτητη και συμπληρωματική, άποψη ολόκληρου του γονιδιώματος. Ωστόσο, η κατανόηση των λειτουργιών των γονιδίων, των πρωτεϊνών και άλλων πτυχών του γονιδιώματος απαιτεί περισσότερες πληροφορίες από αυτές που παρέχει κάθε ένα από τα σύνολα δεδομένων (Hamid et al., 2008).

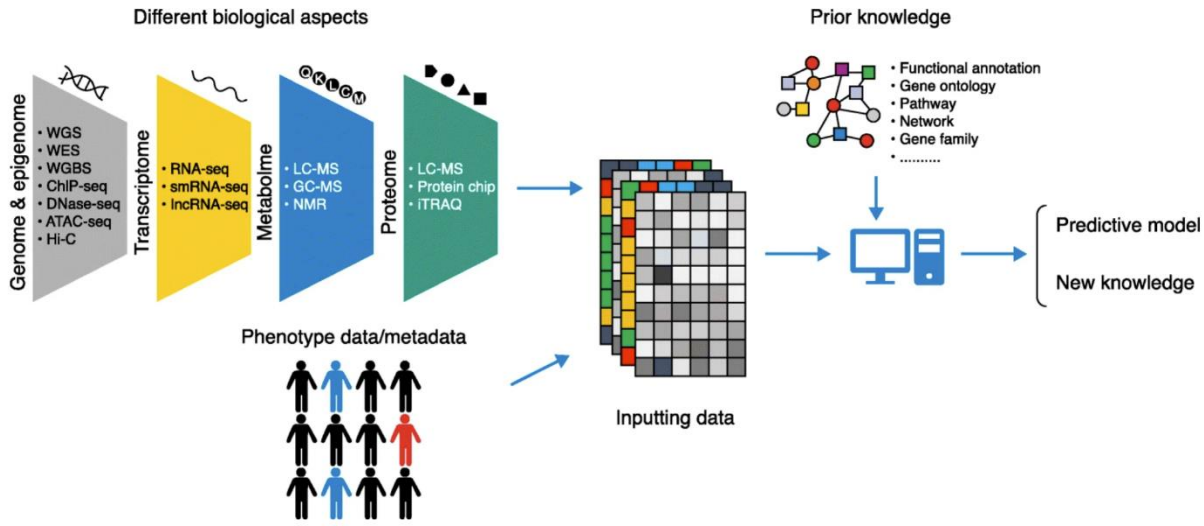
iv. Απαιτήσεις σε υπολογιστικούς πόρους:

Με την ταχεία πρόοδο της τεχνολογίας αλληλούχισης υψηλής απόδοσης, τα βιολογικά δεδομένα αυξάνονται εκθετικά, δημιουργώντας την ανάγκη για αποτελεσματικούς και κλιμακούμενους αλγορίθμους και υπολογιστικές πλατφόρμες υψηλής απόδοσης για ανάλυση μεγάλων βιολογικών δεδομένων. Ωστόσο, οι παραδοσιακές υπολογιστικές πλατφόρμες υψηλής απόδοσης είναι ανεπαρκείς για την κάλυψη της ζήτησης για ταχεία ανάλυση δεδομένων στη βιοπληροφορική έρευνα. Η ανάλυση βιολογικών δεδομένων αλληλουχιών αποτελεί πρόκληση λόγω της πολυπλοκότητάς τους και της μεγάλης απαίτησης σε υπολογιστικό χρόνο (Yeh et al., 2023).

Σήμερα, πολλές φορές οι ερευνητές αναγκάζονται να υποβαθμίσουν την ακρίβεια και την αξιοπιστία των αποτελεσμάτων τους, ώστε οι αναλύσεις τους να είναι εφικτές σε πραγματικό χρόνο (Cirillo et al., 2018).

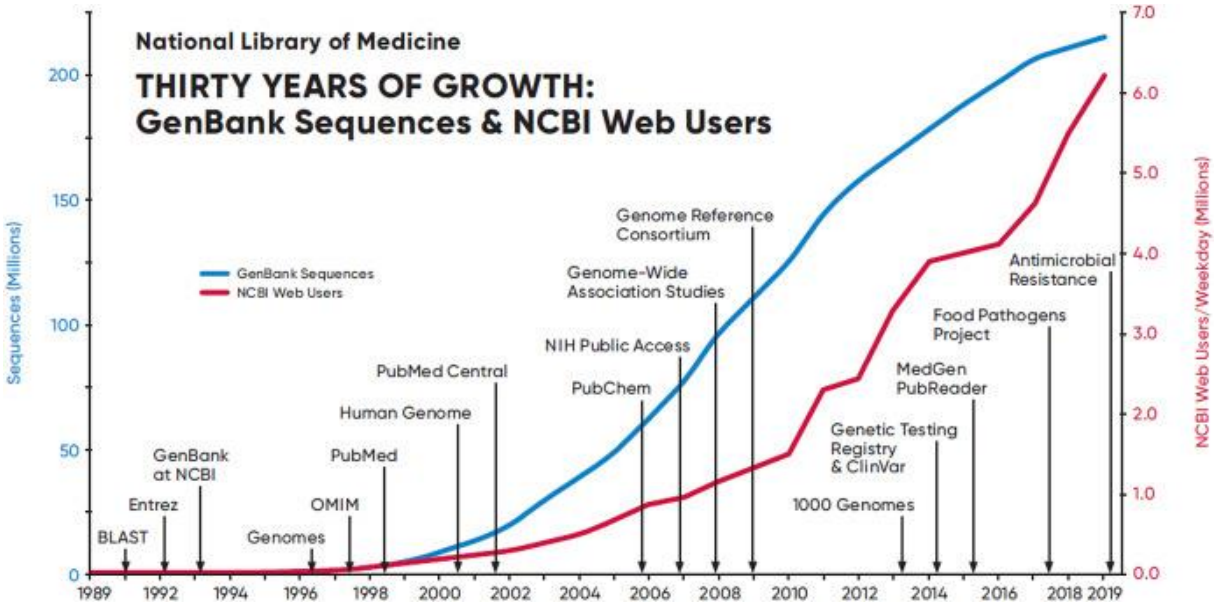
Ωστόσο, παρά τις προκλήσεις, η συνεχής ανάπτυξη τεχνολογιών και μεθόδων ανοίγει νέες δυνατότητες για την αποτελεσματικότερη επεξεργασία και ανάλυση βιολογικών δεδομένων. Η συνεργασία μεταξύ επιστημόνων διαφορετικών πεδίων, όπως βιολογία, πληροφορική και στατιστική, αποτελεί βασικό παράγοντα για την αντιμετώπιση των προκλήσεων και την πρόοδο στην έρευνα (Heidorn et al., 2007; Myneni & Patel, 2010).

1.5 Μηχανική Μάθηση στη Βιολογία



Εικόνα 17. Η μηχανική μάθηση αξιοποιεί διαθέσιμα βιολογικά δεδομένα από πολυάριθμες πηγές: Γονιδίωμα και επιγενετικές τροποποιήσεις, μεταγράφωμική, μεταβολωμική και πρωτεομική τεχνολογία παράλληλα με πληθυσμιακά δεδομένα και μεταδεδομένα φαινοτύπου. Όλη αυτή η πληροφορία κωδικοποιείται κατάλληλα και συνδυάζεται με υπάρχουσα γνώση από πειράματα, functional annotation, την οντολογία γονιδίων, σηματοδοτικά μονοπάτια, δίκτυα αλληλεπιδράσεων ή/και συνέκφρασης, καθώς και πληροφορίες για οικογένειες γονιδίων, ώστε να δημιουργήσει ένα μοντέλο πρόβλεψης, ή/και να παράξει νέα γνώση. Από: (Xu & Jackson, 2019).

Η διαθεσιμότητα μεγάλων δεδομένων όχι μόνο δημιουργεί άνευ προηγουμένου ευκαιρίες εμβάθυνσης στην ανάλυση βιολογικών συστημάτων, αλλά θέτει επίσης νέες προκλήσεις για την εξόρυξη και την ανάλυση δεδομένων (Pal et al., 2020). Πρόσφατα, η μηχανική μάθηση και η βαθιά μάθηση έχουν γίνει οι πιο σύγχρονες τεχνικές για την ανάλυση βιολογικών μεγάλων δεδομένων (Angermueller et al., 2016), (Webb, 2018; Xu & Jackson, 2019). Επί του παρόντος αποτελούν την τελευταία τεχνολογική εξέλιξη στην επίλυση εργασιών που σχετίζονται με την πρόβλεψη. Τα βιολογικά μεγάλα δεδομένα περιλαμβάνουν δεδομένα αλληλούχισης (sequencing data), δεδομένα εικόνας, γενικά δεδομένα πίνακα κ.λπ (Ching et al., 2018; Mirza et al., 2019). Η στατιστική και τα υπολογιστικά μαθηματικά που σχετίζονται με τη μηχανική και τη βαθιά μάθηση βοηθούν σημαντικά στην ανάλυση τέτοιων δεδομένων (Perakakis et al., 2018). Όμως, η μέγιστη εξαγωγή πληροφοριών από τέτοια βιολογικά μεγάλα δεδομένα μπορεί να γίνει αποτελεσματικά όταν οι τεχνικές αυτές εκτελούνται σε μια πλατφόρμα μεγάλων δεδομένων.

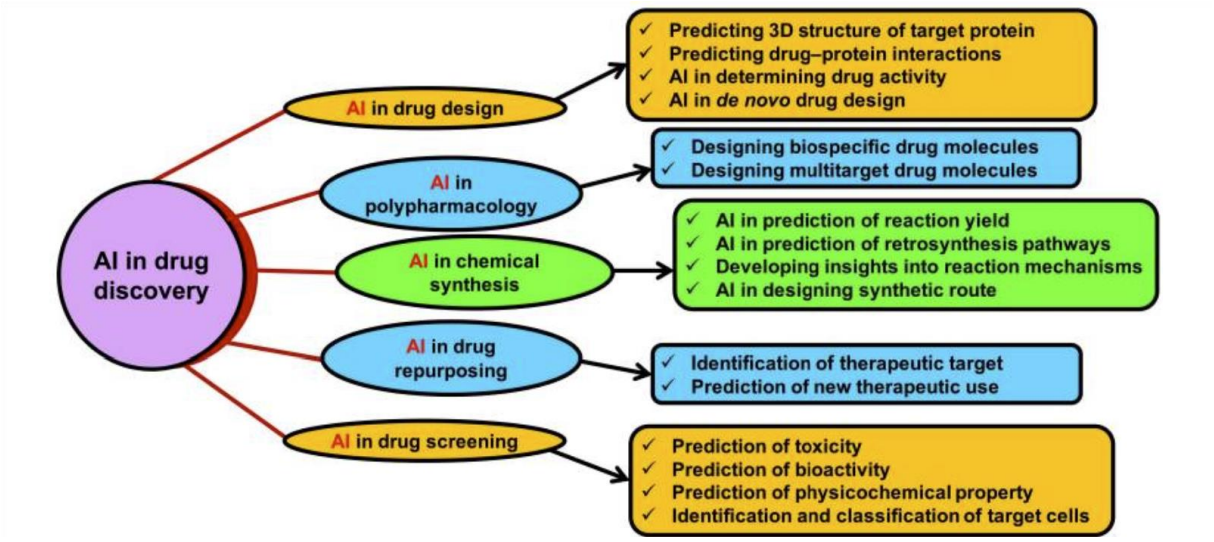


Εικόνα 18. Πλήθος αλληλουχιών που έχουν αναρτηθεί στη GenBank (μπλε) και πλήθος διαδικτυακών χρηστών του NCBI (κόκκινο). Και στις δύο μεταβλητές παρατηρείται ραγδαία αύξηση. (Gaffney et al., 2020).

Ως αποτέλεσμα αυτών των εξελίξεων, οι τεχνικές ML χρησιμοποιούνται πλέον σε ένα ευρύ φάσμα εφαρμογών στη βιολογία και την ιατρική (Kushwaha et al., 2024), όπως:

Ανακάλυψη φαρμάκων (Εικόνα 19)**:** Η επιστήμη της χημείας έχει ταυτοποιήσει περισσότερα από 10^{60} χημικά μόρια, επομένως ο δειγματικός χώρος αναζήτησης και ανάπτυξης νέων φαρμάκων είναι τεράστιος και αχανής (Mak & Pichika, 2019). Ωστόσο, η έλλειψη προηγμένων τεχνολογιών, μέχρι πρόσφατα, περιόριζε τη διαδικασία ανάπτυξης

φαρμάκων, καθιστώντας τη πολύ χρονοβόρα και δαπανηρή, πρόβλημα το οποίο ήρθε να αντιμετωπίσει η τεχνητή νοημοσύνη (Vyas et al., 2018). Η τεχνητή νοημοσύνη μπορεί να αναγνωρίσει τις υποψήφιες φαρμακευτικές ενώσεις και τις κύριες δραστικές ουσίες τους, και να παρέχει ταχύτερη επικύρωση του φαρμακευτικού στόχου και βελτιστοποίηση του σχεδιασμού της δομής του φαρμάκου (Sellwood et al., 2018).

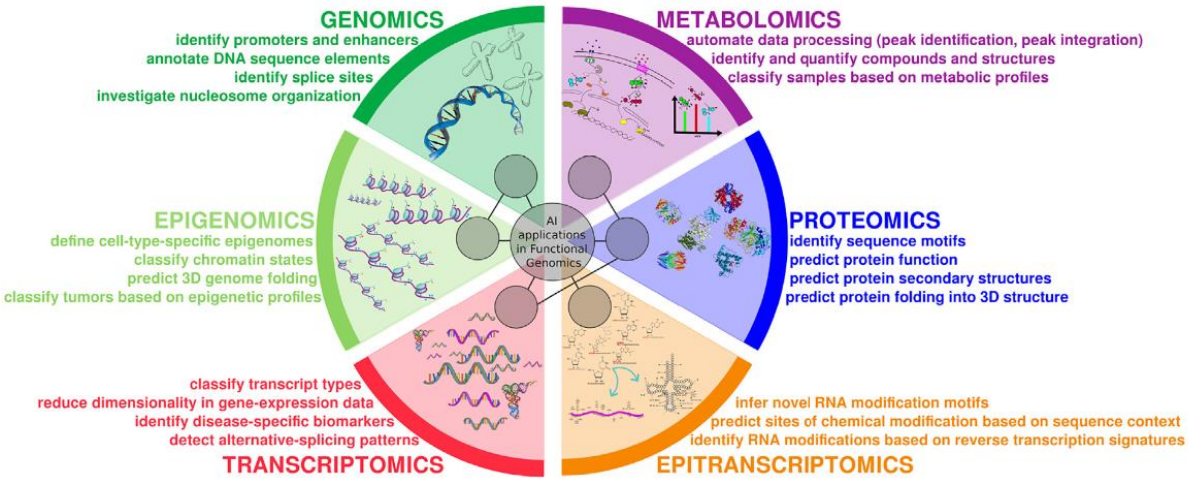


Εικόνα 19. Ο ρόλος της τεχνητής νοημοσύνης στα διάφορα στάδια και τα επιμέρους πεδία της ανακάλυψης νέων φαρμάκων. Από: (Paul et al., 2021).

Διαγνωστική: Παρά την αλματώδη πρόοδο της ιατρικής, η αποτελεσματική διάγνωση των ασθενειών εξακολουθεί να αποτελεί πρόκληση σε παγκόσμια κλίμακα. Η ανάπτυξη πρώιμων διαγνωστικών εργαλείων αποτελεί συνεχή πρόκληση λόγω της πολυπλοκότητας των διαφόρων μηχανισμών της νόσου και των υποκείμενων συμπτωμάτων. Η τεχνητή νοημοσύνη μπορεί να φέρει επανάσταση σε διάφορες πτυχές της υγειονομικής περίθαλψης, συμπεριλαμβανομένης της διάγνωσης (Alowais et al., 2023). Το ML, εν ολίγοις, μπορεί να βοηθήσει στη λήψη αποφάσεων, στη διαχείριση της ροής εργασίας και στην αυτοματοποίηση εργασιών με έγκαιρο και οικονομικά αποδοτικό τρόπο. Συνελικτικά νευρωνικά δίκτυα (CNN) και τεχνικές εξόρυξης δεδομένων που βοηθούν στον εντοπισμό μοτίβων δεδομένων έχουν ευρεία εφαρμογή στον εντοπισμό βασικών μοτίβων ανίχνευσης ασθενειών μεταξύ μεγάλων συνόλων δεδομένων. Αυτά τα εργαλεία είναι ιδιαίτερα χρήσιμα στα συστήματα υγειονομικής περίθαλψης για τη διάγνωση (Miltiadous et al., 2021), την πρόβλεψη ή την ταξινόμηση ασθενειών (Ahsan et al., 2022).

Λειτουργική Γονιδιωματική (Εικόνα 20)**:** Το ML χρησιμοποιείται για την ανάλυση γονιδιακών δεδομένων για να κατανοήσουμε την κληρονομικότητα των ασθενειών, να

αναπτύξουμε προγνωστικά μοντέλα και να εξατομικεύσουμε τη θεραπεία. Τα ομικά δεδομένα μπορούν εύκολα να γίνουν πολύ ογκώδη και πολύπλοκα για να διερευνηθούν μέσω οπτικής ανάλυσης ή στατιστικών συσχετίσεων. Το γεγονός αυτό έχει ενθαρρύνει τη χρήση της τεχνητής νοημοσύνης, η οποία είναι σε θέση όχι μόνο να διαχειρίζεται όγκους δεδομένων που είναι δύσβατοι για τα ανθρώπινα μυαλά, αλλά και να εξάγει πληροφορίες που υπερβαίνουν την τρέχουσα κατανόηση του υπό διερεύνηση συστήματος και, κυρίως, να βελτιώνεται αυτόματα μέσω της εμπειρίας που αποκτάται σε δεδομένα εκπαίδευσης (Caudai et al., 2021).



Εικόνα 20.Εφαρμογές της τεχνητής νοημοσύνης στη λειτουργική γονιδιωματική. Από: (Caudai et al., 2021).

Ορισμένα παραδείγματα της χρήσης τεχνικών ML στη βιολογία και την ιατρική παρουσιάζονται παρακάτω:

DeepMind AlphaFold: Το AlphaFold είναι ένα σύστημα βαθιάς μάθησης που μπορεί να προβλέψει με ακρίβεια την τρισδιάστατη δομή των πρωτεϊνών (Jumper et al., 2021). Αυτή η τεχνολογία έχει τη δυνατότητα να επαναστατήσει την ανάπτυξη φαρμάκων και άλλων βιοϊατρικών θεραπειών.

Earth Biological Genome Project: Το Earth BioGenome Project (EBP) έχει ως στόχο την αλληλούχιση, την καταγραφή και τον χαρακτηρισμό των γονιδιωμάτων όλης της ευκαρυωτικής βιοποικιλότητας της γης, εντός 10 ετών. Η αξιοποίηση της τεχνητής νοημοσύνης αποτελεί απαραίτητο εργαλείο για την υλοποίηση ενός τόσο μεγαλεπήβολου εγχειρήματος (Lewin et al., 2018).

Global Microbial Identifier: Το Global Microbial Identifier (GMI) είναι ένα ερευνητικό πρόγραμμα με στόχο την ανάπτυξης μιας ταχείας και ακριβούς μεθόδου ταυτοποίησης μικροοργανισμών, συμπεριλαμβανομένων και δυνητικά παθογόνων. Αξιοποιεί τεχνολογίες τεχνητής νοημοσύνης για την αναγνώριση για το χαρακτηρισμό ενός μικροοργανισμού ως παθογόνου, βάσει συγκεκριμένων αλληλουχιών και μοτίβων στο γονιδίωμά του, που σχετίζονται με ήδη ταυτοποιημένα παθογόνα (Deng et al., 2017).

ANTIBIOGO: Το ANTIBIOGO είναι μια δωρεάν εφαρμογή για κινητά ικανή να αναλύει εικόνες τριβλύων δοκιμών αντιμικροβιακής ευαισθησίας (Antimicrobial Susceptibility Testing, AST) συνδυάζοντας αλγορίθμους μηχανικής μάθησης και επεξεργασίας εικόνας, ώστε να προτείνει τη χορήγηση του κατάλληλου αντιβιοτικού για κάθε παθογόνο. (Pascucci et al., 2020, 2021)

Galaxy: Το Galaxy είναι μια διαδικτυακή πλατφόρμα αυτοματοποίησης και ανταλλαγής ροών εργασίας στη βιοπληροφορική. Χρησιμοποιεί τεχνητή νοημοσύνη για την αυτοματοποίηση χρονοβόρων και επαναλαμβανόμενων διαδικασιών όπως η προεπεξεργασία και η τυποποιημένη ανάλυση δεδομένων, μειώνοντας σημαντικά το χρόνο που απαιτείται για τις βιοπληροφορικές αναλύσεις (Larivière et al., 2023).

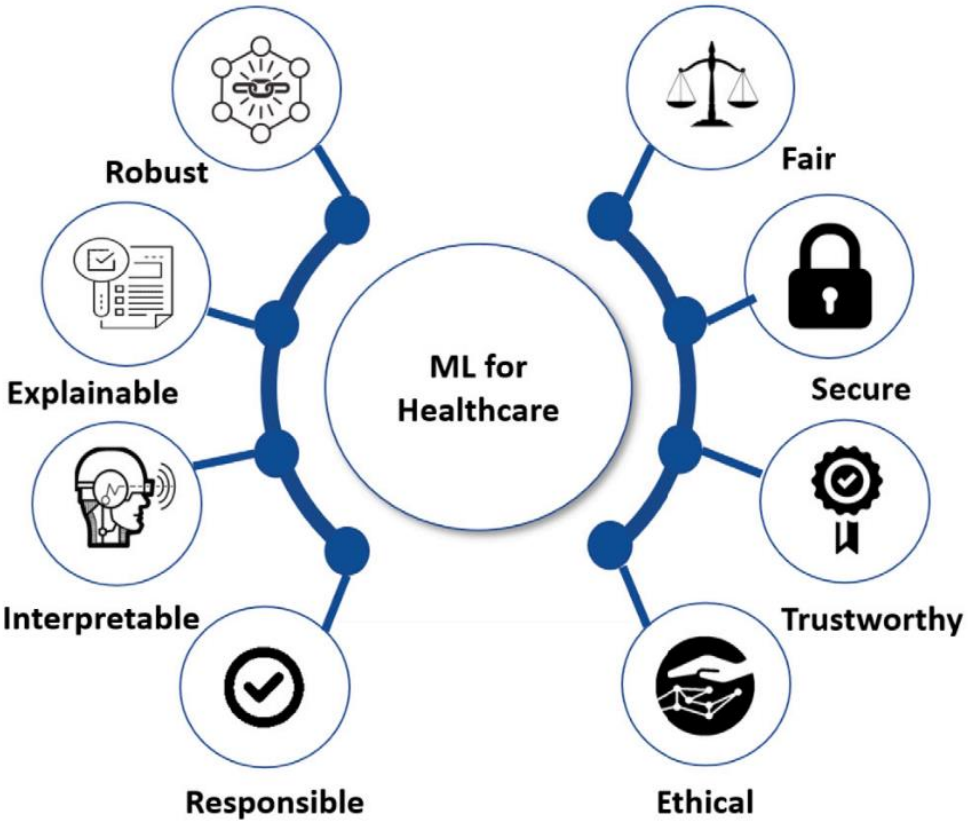
Watson for Oncology: Το Watson for Oncology είναι ένα σύστημα ML που χρησιμοποιείται για να βοηθήσει τους ογκολόγους να λάβουν εξατομικευμένες αποφάσεις θεραπείας για τους ασθενείς τους. Το σύστημα λαμβάνει υπόψη μια ποικιλία δεδομένων, συμπεριλαμβανομένων των γονιδιακών δεδομένων, των δεδομένων απεικόνισης και του ιατρικού ιστορικού του ασθενούς (C. Liu et al., 2018).

Εξατομικευμένη ιατρική: Η ML χρησιμοποιείται για την ανάπτυξη στοχευμένων θεραπευτικών σχημάτων και προσεγγίσεων, εξατομικευμένων για κάθε ασθενή, λαμβάνοντας υπόψιν το πλήρες ιατρικό ιστορικό, την κληρονομικότητα, το μεταβολικό προφίλ, τις συνήθειες της καθημερινότητάς του κ.ο.κ. στοχεύοντας σε μεγαλύτερη ακρίβεια και καλύτερα θεραπευτικά αποτελέσματα και βελτιώνοντας την πρόγνωση (Schork, 2019).

Παρά τις πολλές υποσχόμενες εφαρμογές της, η χρήση τεχνικών ML στη βιολογία και την ιατρική αντιμετωπίζει επίσης ορισμένες προκλήσεις, η σημαντικότερη εκ των οποίων αφορά την ποιότητα και την ετερογένεια των δεδομένων (Ishak & Salim, 2006). Η τεχνητή νοημοσύνη απαιτεί μεγάλη ποσότητα δεδομένων προς εκπαίδευση, ώστε τα μοντέλα να διδαχθούν επαρκώς και να μπορούν να είναι αξιόπιστα. Η ομοιογένεια και η ποιότητα των

δεδομένων αποτελούν απαραίτητη προϋπόθεση για τη σωστή εφαρμογή της μηχανικής μάθησης και της τεχνητής νοημοσύνης στη βιολογία (Ching et al., 2018; Mirza et al., 2019).

Μια ακόμη πρόκληση που καλείται να αντιμετωπίσει η εφαρμογή της τεχνητής νοημοσύνης στην βιοϊατρική έχει να κάνει με την ερμηνεία και επεξήγηση των αποτελεσμάτων που παράγει. Η τεχνητή νοημοσύνη με μοντέλα βαθιάς μάθησης έχει εφαρμοστεί ευρέως στους τομείς της ιατρικής απεικόνισης και υγειονομικής περίθαλψης. Στον ιατρικό τομέα, κάθε κρίση ή απόφαση είναι γεμάτη κινδύνους. Ένας γιατρός θα κρίνει προσεκτικά αν ένας ασθενής είναι άρρωστος, προτού σχηματίσει μια λογική εξήγηση με βάση τα συμπτώματα του ασθενούς ή/και μια εξέταση. Ως εκ τούτου, για να είναι ένα βιώσιμο και αποδεκτό εργαλείο, η τεχνητή νοημοσύνη πρέπει να μιμείται τις ανθρώπινες ικανότητες κρίσης και ερμηνείας. Συγκεκριμένα, η εξηγήσιμη τεχνητή νοημοσύνη (XAI) στοχεύει στην εξήγηση των πληροφοριών πίσω από το μοντέλο μαύρου κουτιού της βαθιάς μάθησης που αποκαλύπτει πώς λαμβάνονται οι αποφάσεις (Chaddad et al., 2023).



Εικόνα 21.Τα βασικά χαρακτηριστικά που πρέπει να διαθέτει η μηχανική μάθηση ώστε να εφαρμοστεί στην υγεία. Από: (Rasheed et al., 2022).

Επιπλέον, η χρήση της τεχνητής νοημοσύνης στη βιολογία, και τον τομέα της υγείας γενικότερα, εγείρει ορισμένα ηθικά ζητήματα που αφορούν την αξιοπιστία, την ασφάλεια

και την ορθή χρήση των δεδομένων (Gundersen & Bærøe, 2022; Rasheed et al., 2022). Παρά τις εντυπωσιακές επιδόσεις των αλγορίθμων μηχανικής μάθησης, πολλές πρόσφατες μελέτες έχουν εγείρει ανησυχίες σχετικά με την ασφάλεια και την ανθεκτικότητα των ML μοντέλων (Qayyum et al., 2021). Για παράδειγμα τα μοντέλα βαθιάς μάθησης είναι αυστηρά ευάλωτα σε προσεκτικά σχεδιασμένα αντιπαραδείγματα (Szegedy et al., 2014). Η εφαρμογή τεχνητής νοημοσύνης στην υγειονομική περίθαλψη βασίζεται στη συλλογή και αποθήκευση ευαίσθητων προσωπικών δεδομένων υγείας, συμπεριλαμβανομένων ιατρικών φακέλων, γενετικών πληροφοριών και δεδομένων παρακολούθησης σε πραγματικό χρόνο. Η εκτεταμένη χρήση αυτών των δεδομένων για την ανάλυση εισάγει εγγενείς κινδύνους για την ιδιωτική ζωή των ατόμων. Η μη εξουσιοδοτημένη πρόσβαση, οι παραβιάσεις δεδομένων και ο ακατάλληλος χειρισμός των προσωπικών πληροφοριών υγείας μπορεί να οδηγήσει σε σοβαρές συνέπειες, όπως κλοπή ταυτότητας, διακρίσεις ή παραβίαση του απορρήτου των ασθενών (Frank & Olaoye, 2024). Συνοπτικά, τα χαρακτηριστικά που πρέπει να διαθέτει η μηχανική μάθηση για να χρησιμοποιείται ορθά στην ιατρική περίθαλψη παρουσιάζονται στην Εικόνα 21.

Παρά αυτές τις προκλήσεις, η μελλοντική χρήση τεχνικών ML στη βιολογία και την ιατρική αναμένεται να είναι ραγδαία. Νέα αλγόριθμοι ML, μεγαλύτερες ποσότητες δεδομένων και αυξημένη υπολογιστική ισχύς θα οδηγήσουν σε ακόμα πιο καινοτόμες εφαρμογές που θα βελτιώσουν την υγεία και την ευημερία των ανθρώπων (Ching et al., 2018).

Η χρήση τεχνικών μηχανικής μάθησης στη βιολογία και την ανάλυση ιατρικών δεδομένων έχει επαναστατήσει τον τρόπο με τον οποίο διεξάγεται η βιοϊατρική έρευνα και η ιατρική περίθαλψη (Caudai et al., 2021; Hassoun et al., 2022). Η τεχνολογία έχει τη δυνατότητα να βελτιώσει σημαντικά την ακρίβεια της διάγνωσης, την αποτελεσματικότητα της θεραπείας και την ποιότητα ζωής των ασθενών. Καθώς η τεχνολογία ML συνεχίζει να εξελίσσεται, μπορούμε να περιμένουμε ακόμα πιο καινοτόμες εφαρμογές που θα αλλάξουν το μέλλον της ιατρικής.

1.6 Το πρόβλημα και οι στόχοι της εργασίας

Η αναπαράσταση βιολογικών δεδομένων σε μορφή γραφήματος αποτελεί μια σημαντική πρόκληση στον τομέα της βιολογικής έρευνας. Τα βιολογικά δεδομένα, όπως οι αλληλεπιδράσεις πρωτεϊνών ή τα γενετικά δίκτυα, συχνά παρουσιάζουν περίπλοκες σχέσεις και αλληλεξαρτήσεις που δύσκολα μπορούν να αποτυπωθούν με παραδοσιακές μεθόδους. Η ανάπτυξη αποτελεσματικών τεχνικών αναπαράστασης γραφημάτων για βιολογικά

δεδομένα είναι απαραίτητη για την καλύτερη κατανόηση των βιολογικών διεργασιών, την ανάπτυξη νέων θεραπειών και φαρμάκων, καθώς και την ανακάλυψη νέων σηματοδοτικών μονοπατιών ή τις αλληλεπιδράσεις μεταξύ των ήδη γνωστών.

Η παρούσα διπλωματική εργασία έχει ως στόχο:

- Να ερευνήσει και να αξιολογήσει διάφορες τεχνικές μηχανικής μάθησης για την αναπαράσταση βιολογικών δεδομένων σε διανυσματική μορφή.
- Να αναπτύξει βελτιωμένες τεχνικές αναπαράστασης γραφημάτων που λαμβάνουν υπόψη τα μοναδικά χαρακτηριστικά των βιολογικών δεδομένων.
- Να εφαρμόσει τις προτεινόμενες τεχνικές σε πραγματικά βιολογικά δεδομένα και να αξιολογήσει την αποτελεσματικότητά τους.
- Να συμβάλει στην ανάπτυξη νέων εργαλείων και μεθόδων για την ανάλυση βιολογικών δεδομένων με βάση γραφήματα.

Επιτυγχάνοντας τους παραπάνω στόχους, η παρούσα διπλωματική εργασία αναμένεται να συμβάλει σημαντικά στην πρόοδο του πεδίου της αναπαράστασης βιολογικών δεδομένων με βάση τους γράφους. Η θεωρία γραφημάτων βρίσκει ευρεία εφαρμογή στη μοντελοποίηση βιολογικών προβλημάτων, χάρη στη συνδυαστική φύση των νουκλεϊκών οξέων (Blazewicz & Kasprzak, 2012). Τα αποτελέσματα της έρευνας θα μπορούσαν να αξιοποιηθούν από βιολόγους, πληροφορικούς και άλλους επιστήμονες για:

- Βελτίωση της κατανόησης των βιολογικών διεργασιών και των σχέσεων μεταξύ βιολογικών οντοτήτων.
- Ανάπτυξη νέων θεραπειών και φαρμάκων.
- Ανακάλυψη νέων βιολογικών μονοπατιών και στόχων φαρμάκων.
- Δημιουργία νέων εργαλείων και μεθόδων για την ανάλυση βιολογικών δεδομένων.

Συνοψίζοντας, η συγκεκριμένη εργασία εστιάζει σε ένα σημαντικό πρόβλημα στον τομέα της βιολογικής έρευνας και έχει ως στόχο την ανάπτυξη νέων τεχνικών αναπαράστασης γραφημάτων για βιολογικά δεδομένα. Τα αποτελέσματα της έρευνας θα μπορούσαν να έχουν σημαντικό αντίκτυπο στην κατανόηση των βιολογικών διεργασιών, την επιτάχυνση πολύπλοκων και πολυπαραγοντικών αναλύσεων δεδομένων γονιδιακής έκφρασης, την καθιέρωση πιο αποδοτικών μεθοδολογιών και στην ανάπτυξη νέων μεταβολικών

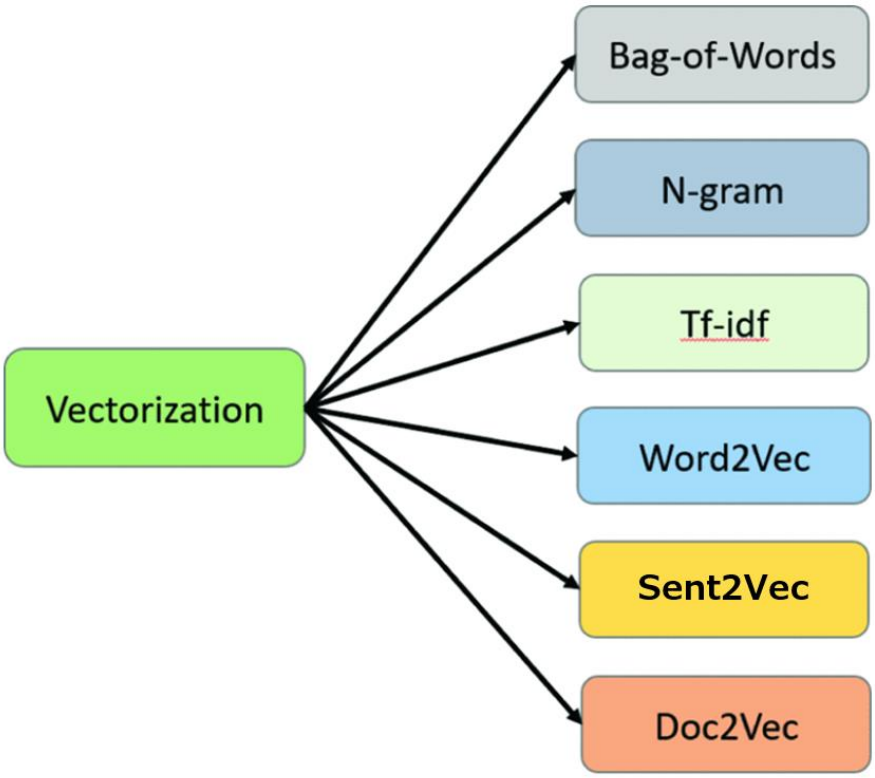
μονοπατιών που σχετίζονται με ασθένειες και κατ’ επέκταση νέων θεραπειών και φαρμάκων.

2 Διανυσματική Αναπαράσταση Δεδομένων

2.1 Διανυσματοποίηση

Με την απίστευτη αύξηση του όγκου των δεδομένων κειμένου, αυξάνεται η ανάγκη για την ανάπτυξη αποτελεσματικών μεθόδων επεξεργασίας και ανάλυσής τους. Η μετατροπή πληροφορίας σε διανύσματα αποτελεί μια αποτελεσματική μέθοδο για την εξαγωγή σημαντικών χαρακτηριστικών και τη συμπύκνωση δεδομένων, ελαχιστοποιώντας την απώλεια χρήσιμης πληροφορίας. Μια πηγή δεδομένων με τεράστιο όγκο και ποικιλομορφία, που μπορεί να περιέχει σημαντική πληροφορία είναι τα κείμενα. Επομένως, η εξαγωγή χαρακτηριστικών από κείμενο αποτελεί επιτακτική ανάγκη για την επίλυση πολλών προβλημάτων εξόρυξης δεδομένων κειμένου και ανάκτησης πληροφοριών (Mansour et al., 2022).

Οι παραδοσιακές μέθοδοι επεξεργασίας και ανάλυσης δεδομένων δυσκολεύονται να αντιμετωπίσουν το πρόβλημα της εξαγωγής πληροφοριών από μεγάλα δεδομένα κειμένου. Επομένως, είναι απαραίτητο να χρησιμοποιηθεί η ευφυής επεξεργασία με τη βοήθεια της μηχανικής μάθησης για την εξαγωγή πληροφοριών από τέτοιου είδους δεδομένα. Επειδή τα δεδομένα κειμένου δεν μπορούν να χρησιμοποιηθούν άμεσα για την εκπαίδευση των παραμέτρων ενός μοντέλου μηχανικής μάθησης, είναι απαραίτητο να διανυσματοποιηθούν τα αρχικά δεδομένα κειμένου και να γίνουν αριθμητικά, και στη συνέχεια μπορεί να πραγματοποιηθεί η λειτουργία εξαγωγής χαρακτηριστικών (Yang et al., 2022).



Εικόνα 22. Διάφορες τεχνικές για τη διανυσματοποίηση κειμένου. Τροποποίηση από: (Rani et al., 2022).

2.2 Διανυσματοποίηση Κειμένου

Έχουν αναπτυχθεί διάφορες προσεγγίσεις για τη διανυσματική αναπαράσταση κειμένων (Εικόνα 22):

Bag of Words (BoW): Είναι η πιο απλή τεχνική από όλες. Δημιουργεί ένα λεξιλόγιο για όλες τις μοναδικές λέξεις από το κείμενο και στη συνέχεια δημιουργεί ένα διάνυσμα που περιέχει τη συχνότητα εμφάνισης των μοναδικών λέξεων. Το μέγεθος του διανύσματος θα είναι ίσο με τον αριθμό των μοναδικών λέξεων, δηλαδή τον αριθμό του λεξιλογίου (Qader et al., 2019).

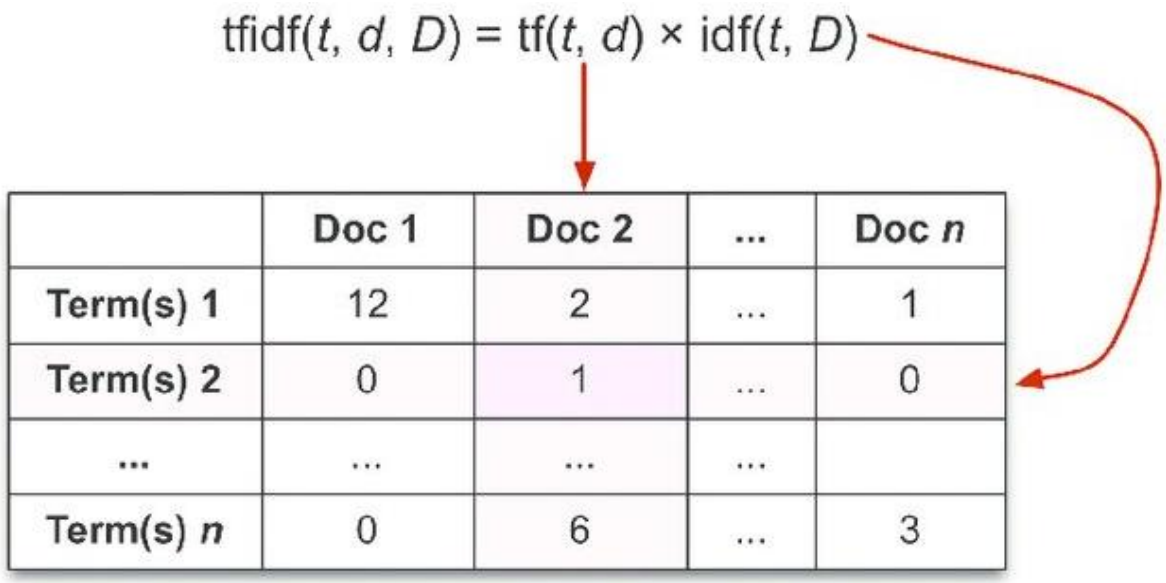
N-Gram: Αποτελεί επέκταση του BoW. Φτιάχνει μια στήλη γειτονικών λέξεων μήκους n για να σχηματίσει σημασιολογικό πλαίσιο. Από τη μικρή τιμή του n δεν μπορούν να εξαχθούν επαρκείς πληροφορίες. Όσο αυξάνεται η τιμή του n, το σύστημα μαθαίνει με μεγαλύτερη ακρίβεια. Με τη μεγάλη τιμή του n, το μέγεθος του πίνακα θα είναι επίσης μεγάλο. Έτσι, η επιλογή της σωστής τιμής του n είναι επίσης σημαντική (Cavnar & Trenkle, 2012).

TF-IDF (Εικόνα 23)**:** Το TF-IDF (term frequency-inverse document frequency) είναι μια μέθοδος στάθμισης όρων που χρησιμοποιείται συνήθως για την αναπαράσταση εγγράφων κειμένου ως διανύσματα. Έστω $T = \{t_1, t_2...,t_n\}$ το σύνολο όλων των όρων που εμφανίζονται στο εξεταζόμενο σύνολο εγγράφων. Τότε ένα έγγραφο d_i αναπαρίσταται από ένα n-διάστατο διάνυσμα πραγματικών τιμών $x_i = (x_{i1}, x_{i2},..., x_{in})$ με μία συνιστώσα για κάθε πιθανό όρο από το T.

Το βάρος x_{ij} που αντιστοιχεί στον όρο t_j στο έγγραφο d_i είναι συνήθως ένα γινόμενο τριών μερών: ένα που εξαρτάται από την παρουσία ή τη συχνότητα του t_j στο d_i, ένα που εξαρτάται από την παρουσία του t_j στο σύνολο του σώματος κειμένων και ένα μέρος κανονικοποίησης που εξαρτάται από το d_j . Συνήθως ο συντελεστής στάθμισης TF-IDF ορίζεται ως εξής:

$$x_{ij} = TF_i \cdot IDF_i \cdot \left(\sum_j \left(TF_{ij} \cdot IDF_j \right)^2 \right)^{-1/2}$$

Όπου TF_{ij} είναι η συχνότητα των όρων (δηλαδή ο αριθμός των εμφανίσεων) του t_j στο d_i και IDF_j είναι η *IDF* του t_j, που ορίζεται ως $log(N/DF_j)$, όπου N είναι ο αριθμός των εγγράφων στο σώμα εγγράφων και DF_j είναι η συχνότητα των εγγράφων του t_j (δηλαδή ο αριθμός των εγγράφων στα οποία εμφανίζεται το t_j). Το τμήμα κανονικοποίησης της εξίσωσης εξασφαλίζει ότι το διάνυσμα έχει ευκλείδειο μήκος ίσο με 1 (Sammut & Webb, 2010).



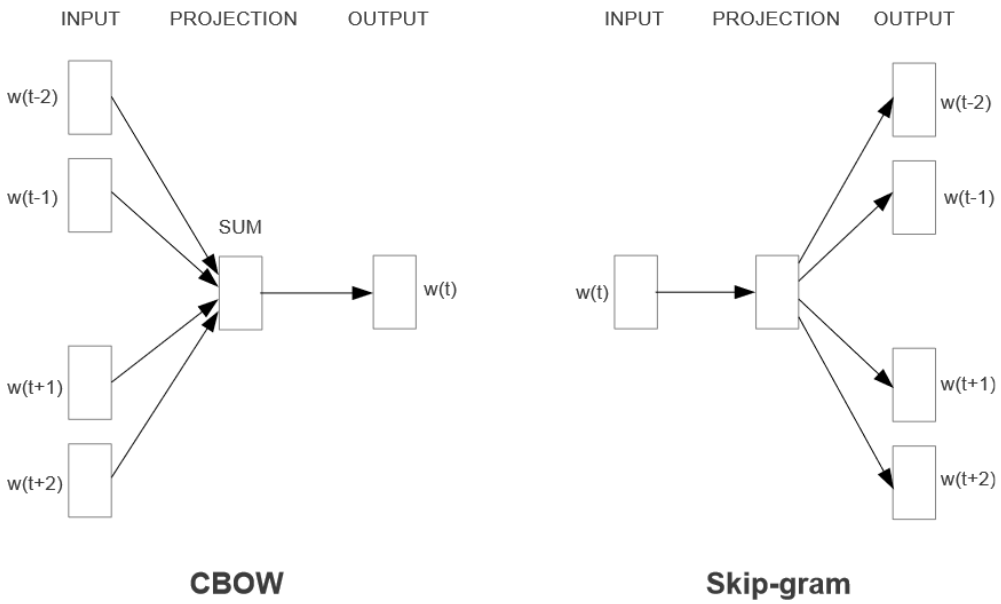
Εικόνα 23. Παράδειγμα μεθόδου διανυσματοποίησης κειμένου TF - IDF. Από: (al-Khateeb & Epiphaniou, 2016).

Παρόλο που τα μοντέλα Bag of Words (BoW), N-gram και TF-IDF αποτέλεσαν βασικά εργαλεία στην επεξεργασία φυσικής γλώσσας (NLP) για δεκαετίες, παρουσιάζουν σημαντικούς περιορισμούς που θέτουν όρια στην αποτελεσματικότητά τους. Το σημαντικότερο μειονέκτημά τους είναι η έλλειψη σημασιολογικής κατανόησης του κειμένου. Τα μοντέλα BoW και N-gram αντιμετωπίζουν τις λέξεις ως απομονωμένες μονάδες, αγνοώντας τη σημασιολογική τους σχέση και το πλαίσιο στο οποίο χρησιμοποιούνται. Αυτό μπορεί να οδηγήσει σε λανθασμένες ερμηνείες και αδυναμία κατανόησης της ουσίας του κειμένου. Τα μοντέλα tf-idf, αν και λαμβάνουν υπόψη τη συχνότητα εμφάνισης των λέξεων, δεν μπορούν να εξάγουν σημασιολογικές σχέσεις μεταξύ τους. Αυτό περιορίζει την ικανότητά τους να κατανοήσουν το ύφος και την ουσία του κειμένου και να εντοπίσουν κρυφές έννοιες και λανθάνοντα νοήματα.

2.2.1 Word2vec

Το Word2vec αποτελεί μια δημοφιλή τεχνική μηχανικής μάθησης για την αναπαράσταση λέξεων, η οποία μπορεί να εφαρμοστεί και σε γραφήματα για την αναπαράσταση κόμβων. Η βασική ιδέα του Word2vec είναι να αναπαραστήσει κάθε λέξη ως ένα διάνυσμα στον πραγματικό χώρο, όπου η θέση του διανύσματος στο χώρο αντικατοπτρίζει τις σημασιολογικές σχέσεις της λέξης με άλλες λέξεις (Mikolov et al., 2013). Υπάρχουν δύο βασικά μοντέλα Word2vec (Εικόνα 24):

- **CBOW (Continuous Bag-of-Words):** Προσπαθεί να προβλέψει μια τρέχουσα λέξη (στόχος) δεδομένου του περιβάλλοντός της (γύρω λέξεις).
- **Skip-gram:** Προσπαθεί να προβλέψει το περιβάλλον μιας λέξης (στόχος) δεδομένης της λέξης.



Εικόνα 24. Παρουσίαση των δύο βασικών μοντέλων λειτουργίας του Word2Vec αλγορίθμου. Το CBOW προβλέπει μια λέξη από τις γειτονικές της, ενώ το Skip-gram προβλέπει τις γειτονικές λέξεις μιας δεδομένης. Από: (Mikolov et al., 2013).

Και τα δύο μοντέλα μαθαίνουν διανυσματικές αναπαραστάσεις λέξεων με βάση στατιστικές πληροφορίες από ένα μεγάλο σώμα κειμένου.

Για να εφαρμοστεί το Word2vec σε γραφήματα, κάθε κόμβος αντιμετωπίζεται ως μια λέξη και οι συνδέσεις μεταξύ κόμβων αντιμετωπίζονται ως λέξεις στο κείμενο. Στη συνέχεια, ένα από τα μοντέλα Word2vec εκπαιδεύεται στο γράφημα, μαθαίνοντας διανυσματικές αναπαραστάσεις για κάθε κόμβο (Li et al., 2023).

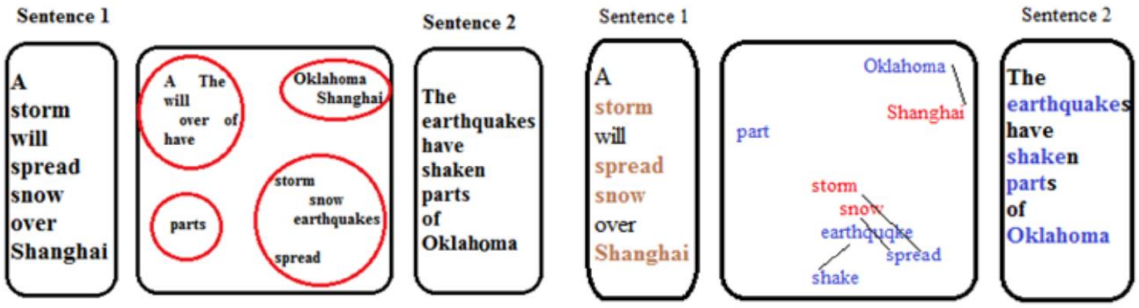
Η εφαρμογή της μεθόδου Word2vec παρουσιάζει ορισμένα σημαντικά πλεονεκτήματα, όπως η απλότητα στην κατανόηση αλλά και την υλοποίηση, η αποτελεσματικότητα, η απόδοση, σε σύγκριση πάντα με τις κλασσικές τεχνικές ανάλυσης γράφων, και η ευελιξία, καθώς η συγκεκριμένη τεχνική μπορεί να εφαρμοστεί σε διάφορους τύπους γραφημάτων (διμερή, υπερ-γραφήματα κλπ) (Mikolov et al., 2013).

Ωστόσο, η τεχνική αυτή παρουσιάζει και ορισμένα σημαντικά μειονεκτήματα. Το σημαντικότερο αυτών είναι ότι μετατρέπει κάθε κόμβο σε ένα σύνολο από λέξεις που αντιστοιχούν στους γείτονες με τους οποίους συνδέεται, αγνοώντας τη δομή του γραφήματος στο επίπεδο, και τις έμμεσες συσχετίσεις. Επιπλέον, τα διανυσματικά ενσωματώματα (embeddings) που παράγονται από το Word2vec μπορεί να είναι δύσκολο

να ερμηνευθούν άμεσα, επομένως έχει νόημα μόνο σε περιπτώσεις συγκρίσεων μεταξύ των κόμβων ενός γραφήματος.

Συνοψίζοντας, το Word2vec αποτελεί μια χρήσιμη τεχνική μηχανικής μάθησης για την αναπαράσταση γραφημάτων. Η απλότητα, η αποτελεσματικότητα και η ευελιξία του το καθιστούν μια δημοφιλή επιλογή για διάφορες εφαρμογές. Ωστόσο, είναι σημαντικό να ληφθούν υπόψη οι περιορισμοί του Word2vec, πριν επιλεγεί η χρήση του.

2.2.2 Sent2vec

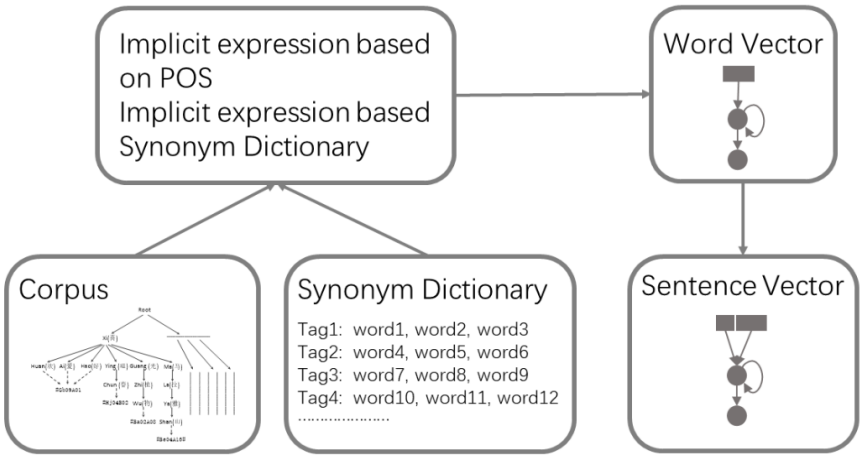


Εικόνα 25. Διανυσματική αναπαράσταση των επιμέρους λέξεων δύο προτάσεων με τη μέθοδο word2vec (αριστερά) και λαμβάνοντας υπόψιν το πλαίσιο της πρότασης στην οποία ανήκουν με τη μέθοδο sent2vec (δεξιά). Από: (Abdolahi & Zahedi, 2019).

Το Sent2Vec είναι μια τεχνική μηχανικής μάθησης που παράγει διανυσματικές αναπαραστάσεις προτάσεων, επιτρέποντας στους υπολογιστές να κατανοούν και να επεξεργάζονται τη φυσική γλώσσα πιο αποτελεσματικά σε σχέση με το Word2Vec. Στην Εικόνα 25 συγκρίνεται η αποτελεσματικότητα των δύο μεθόδων για την ίδια πρόταση κειμένου. Πρόκειται για μια μέθοδο μάθησης χωρίς επίβλεψη που μαθαίνει να αναπαριστά τις προτάσεις ως διανύσματα σε έναν χώρο υψηλών διαστάσεων, όπου παρόμοιες προτάσεις αντιστοιχίζονται σε πιο κοντινά σημεία του διανυσματικού χώρου (Abdolahi & Zahedi, 2019). Το Sent2Vec βασίζεται στην ιδέα ότι το νόημα μιας πρότασης μπορεί να αποτυπωθεί από τον μέσο όρο των σημασιών των λέξεων που την αποτελούν. Υποθέτει ότι λέξεις με παρόμοιες σημασίες θα έχουν παρόμοιες διανυσματικές αναπαραστάσεις και ότι οι προτάσεις που περιέχουν αυτές τις λέξεις θα έχουν επίσης παρόμοιες διανυσματικές αναπαραστάσεις (Pagliardini et al., 2018).

Το Sent2Vec απαιτεί προ-εκπαιδευμένες ενσωματώσεις λέξεων (word embeddings), οι οποίες είναι διανυσματικές αναπαραστάσεις μεμονωμένων λέξεων. Αυτές οι ενσωματώσεις συνήθως λαμβάνονται από Word2Vec μοντέλα. Το Sent2Vec υπολογίζει την ενσωμάτωση πρότασης με τον μέσο όρο των ενσωματώσεων λέξεων των λέξεων της πρότασης. Αυτό μπορεί να γίνει με τη χρήση απλού μέσου όρου ή σταθμισμένου μέσου όρου, όπου τα βάρη

αποδίδονται με βάση τη σημασία της λέξης ή τη θέση της στην πρόταση (X. Wang et al., 2018). Η Εικόνα 26 παρουσιάζει το διάγραμμα ροής εργασιών ενός Sent2Vec μοντέλου. Η ενσωμάτωση πρότασης μπορεί να εξηγηθεί ως η αναπαράσταση της σημασιολογίας μιας πρότασης σε μορφή ενός ενιαίου αριθμητικού διανύσματος (Moghadasi & Zhuang, 2020).



Εικόνα 26. Διάγραμμα ροής εργασιών της μεθόδου Sent2Vec. Από: (X. Wang et al., 2018).

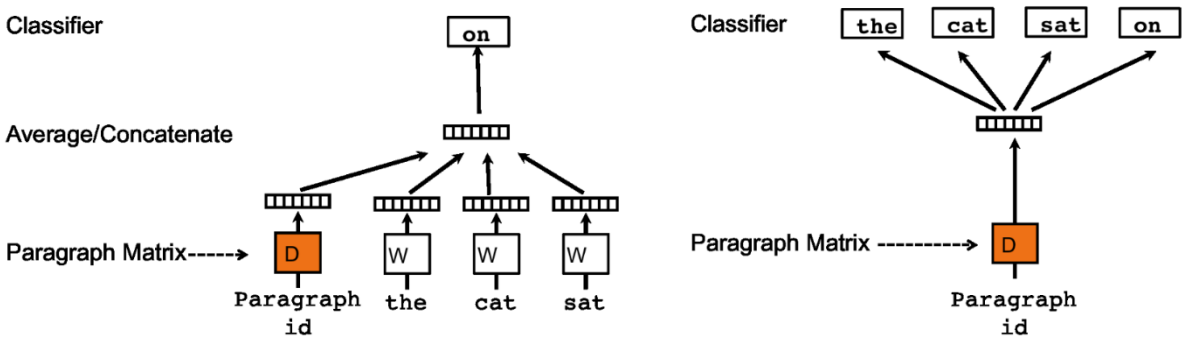
Η μέθοδος Sent2Vec βρίσκει εφαρμογές στον έλεγχο ομοιότητας και ομαδοποίηση προτάσεων, καθώς και την παραγωγή προτάσεων. Η ομοιότητα μεταξύ δύο προτάσεων μπορεί να υπολογιστεί με τη μέτρηση της απόστασης μεταξύ των αντίστοιχων διανυσματικών αναπαραστάσεών τους στον πολυδιάστατο χώρο, με μετρικές όπως η ομοιότητα συνημίτονου ή η ευκλείδεια απόσταση.

Το Sent2Vec αποτελεί μια απλή και αποτελεσματική μέθοδο στη σύλληψη και διανυσματική αποτύπωση του νοήματος μιας ολόκληρης πρότασης. Στηρίζεται σε καθιερωμένες μεθόδους ενσωμάτωσης λέξεων, και επομένως είναι εύκολα υλοποιήσιμη, ωστόσο, περιορίζεται από την ποιότητα των προεκπαιδευμένων λέξεων, από τις οποίες εξαρτάται.

2.2.3 Doc2Vec

Το Doc2Vec είναι μια επέκταση του Sent2Vec, καθώς οι προτάσεις αποτελούν μέρος των εγγράφων, και η διαδικασία απόκτησης των Doc2Vec ενσωματώσεων είναι παρόμοια με εκείνη του Sent2Vec. Ενώ οι ενσωματώσεις λέξεων ενός επιπέδου που παράγονται από Word2Vec μοντέλα είναι αναμφίβολα οι πιο συχνά χρησιμοποιούμενες από τις ενσωματώσεις λέξεων, εξακολουθούν να περιορίζονται στην καταγραφή μόνο της συντακτικής και σημασιολογικής πληροφορίας των λέξεων από μια μεγάλη συλλογή μη

επισημασμένων κειμένων (Le & Mikolov, 2014). Ενώ αυτές οι μέθοδοι αρκούν για την ομαδοποίηση εγγράφων, δεν μπορούν να χειριστούν αποτελεσματικά πιο σύνθετες εργασίες επεξεργασίας φυσικής γλώσσας, όπως ερωταπαντήσεις, επεξήγηση κειμένου, εντοπισμός ονοματικών οντοτήτων, ανάλυση συναισθήματος κλπ, καθώς παράγουν ανεξάρτητες από το περιβάλλον ενσωματώσεις με περιορισμένη ικανότητα για την αποσαφήνιση του νοήματος του λόγου (Word Sense Disambiguation, WSD).



Εικόνα 27. Αριστερά: Το Distributed Memory Paragraph Vectors (dmpv) μοντέλο για την εκμάθηση διανύσματος παραγράφων. Λειτουργεί με παρόμοιο τρόπο με το Word2Vec. Η μόνη αλλαγή είναι το πρόσθετο token παραγράφου που αντιστοιχίζεται σε διάνυσμα μέσω του πίνακα D. Σε αυτό το μοντέλο, η συνένωση ή ο μέσος όρος αυτού του διανύσματος παραγράφου με τα διανύσματα των τριών λέξεων χρησιμοποιείται για την πρόβλεψη της τέταρτης λέξης. Το διάνυσμα της παραγράφου αντιπροσωπεύει τις πληροφορίες που λείπουν από το τρέχον πλαίσιο και μπορεί να λειτουργήσει ως αποτύπωση της θεματολογίας της παραγράφου.

Δεξιά: Το Distributed Bag of Words (dbow) μοντέλο για την εκμάθηση διανύσματος παραγράφων. Σε αυτό αγνοείται πλήρως το νόημα των επιμέρους λέξεων στην είσοδο, και το μοντέλο προβλέπει λέξεις που επιλέγονται τυχαία από την παράγραφο στην έξοδο. Το διάνυσμα παραγράφων εκπαιδεύεται για να προβλέπει τις λέξεις σε ένα μικρό παράθυρο.

Από: (Le & Mikolov, 2014).

Η ανάλυση μεγαλύτερων προτάσεων, παραγράφων ή εγγράφων διαφορετικού μήκους απαιτεί τεχνικές ενσωμάτωσης σε μακρο-επίπεδο και το Doc2Vec έχει σχεδιαστεί για τέτοια σενάρια. Το Doc2Vec είναι μια επέκταση του αλγορίθμου Word2Vec για την εκμάθηση συνεχών αναπαραστάσεων μεγαλύτερων τμημάτων κειμένου, όπως προτάσεις, παράγραφοι ή ολόκληρο το έγγραφο, από την άποψη της ενσωμάτωσης των συστατικών λέξεων. Ένα πρόσθετο token πρότασης/παραγράφου παράγεται και χρησιμοποιείται για να προκύψουν τα διανύσματα ολόκληρου του εγγράφου (Singh & Shashi, 2019). Δυο διαφορετικές υλοποιήσεις Doc2Vec μοντέλων παρουσιάζονται στην Εικόνα 27.

2.3 Διανυσματική αναπαράσταση γράφων

Οι τεχνικές διανυσματικής αναπαράστασης κειμένου Word2Vec, Sent2Vec και Doc2Vec αποσκοπούν στην ανάλυση λέξεων, προτάσεων και εγγράφων με τη μετατροπή τους σε

διανύσματα που διατηρούν τις σημασιολογικές τους σχέσεις. Αυτές οι τεχνικές έχουν αποδειχθεί εξαιρετικά χρήσιμες για την ανάλυση και την κατανόηση του κειμένου, παρέχοντας ισχυρά εργαλεία για την επεξεργασία γλωσσικών δεδομένων. Ωστόσο, τα βιολογικά δεδομένα και άλλες μορφές σύνθετων πληροφοριών συχνά απαιτούν αναπαράσταση που να λαμβάνει υπόψη τη δομή και τις σχέσεις των στοιχείων τους, όπως συμβαίνει στους γράφους. Στη συνέχεια εξετάζονται τεχνικές διανυσματικής αναπαράστασης γράφων, οι οποίες την αποτύπωση της πολυπλοκότητας και των αλληλεπιδράσεων των δεδομένων σε μορφή γράφου, προσφέροντας νέες δυνατότητες για την ανάλυση και την ερμηνεία σύνθετων δομών.

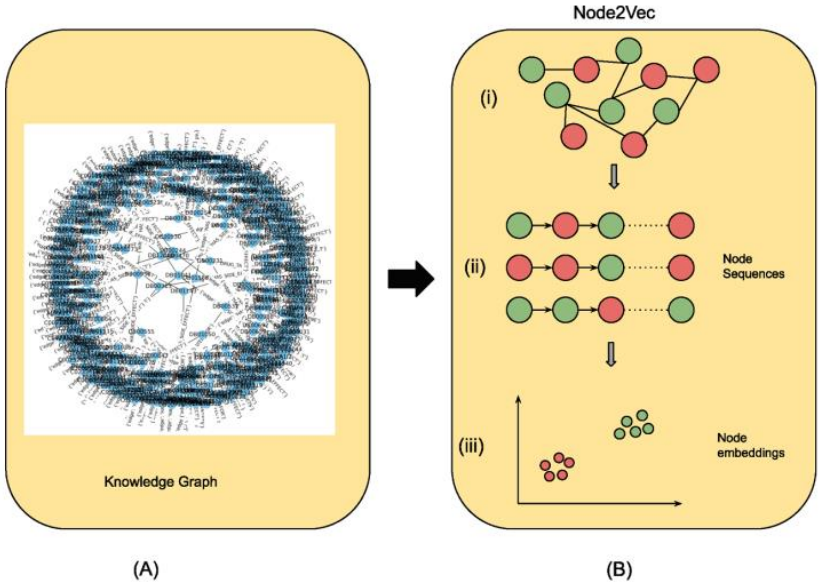
2.3.1 Node2vec

Το Node2vec αποτελεί μια δημοφιλή τεχνική μηχανικής μάθησης για την αναπαράσταση γραφημάτων, η οποία βασίζεται στην ιδέα της "τυχαίας περιήγησης" για την εξερεύνηση του γραφήματος και την δημιουργία διαφορετικών γειτονιών για κάθε κόμβο.

Σε αντίθεση με το παραδοσιακό Word2vec, το Node2vec δεν χρησιμοποιεί σταθερές γειτονίες για κάθε κόμβο. Αντίθετα, χρησιμοποιεί τυχαία βήματα για να εξερευνήσει το γράφημα και να δημιουργήσει διαφορετικές γειτονίες για κάθε κόμβο (Εικόνα 28, Εικόνα 29 και Εικόνα 30).

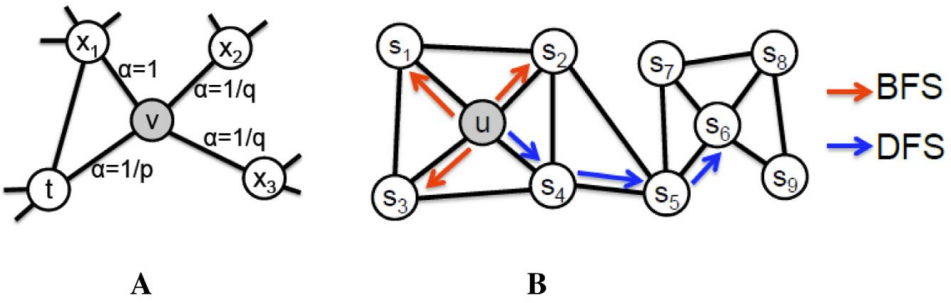
Η τυχαία περιήγηση στο Node2vec υλοποιείται με δύο βασικούς παραμέτρους:

- **p: Ο παράγοντας επιστροφής.** Ορίζει την τάση επιστροφής σε έναν προηγούμενο κόμβο κατά τη διάρκεια της περιήγησης.
- **q: Ο παράγοντας προώθησης.** Ορίζει την τάση εξερεύνησης νέων κόμβων κατά τη διάρκεια της περιήγησης.



Εικόνα 28. Αναπαράσταση της λειτουργίας του Node2Vec αλγορίθμου. Εφαρμοζόμενος σε ένα γράφο (A), πραγματοποιείται η προσπέλαση με τη μέθοδο της τυχαίας περιήγησης (i) και προκύπτουν ακολουθίες κόμβων από τις επιμέρους διαδρομές που ακολουθήθηκαν (ii). Οι διαδρομές στη συνέχεια μετατρέπονται σε διανύσματα και παρουσιάζονται στο διανυσματικό χώρο σε δύο διαστάσεις (iii). Από: (Joshi et al., 2022).

Διαφορετικές τιμές των p και q οδηγούν σε διαφορετικές στρατηγικές εξερεύνησης, οι οποίες μπορούν να αποτυπώσουν διαφορετικές όψεις της δομής του γραφήματος. Στη συνέχεια, το Node2vec εκπαιδεύεται για να μάθει διανυσματικές αναπαραστάσεις για κάθε κόμβο, λαμβάνοντας υπόψη τις διαφορετικές γειτονίες που δημιουργήθηκαν κατά την τυχαία περιήγηση (Grover & Leskovec, 2016).



Εικόνα 29. Ένας τυχαίος περίπατος με μεροληψία, όπως περιεγράφηκε από τους (Grover & Leskovec, 2016), δημιουργούς του Node2Vec.

Το Node2vec χρησιμοποιεί τυχαίους περιπάτους με μεροληψία για τη δημιουργία των πιθανοτήτων μετάβασης με σταθερό μήκος l, όπου l είναι το μήκος των ακολουθιών κόμβων που πρέπει να διανυθούν. Έστω ότι n_i συμβολίζεται ο *i*-οστός κόμβος στον περίπατο που ξεκινά με $n_0 = u$. Ο κόμβος n_i λαμβάνεται με την ακόλουθη πιθανότητα:

$$\mathrm{P}(n_i = x | n_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{z}, \alpha v\ (v, x) \in E \\ 0, \alpha v\ (v, x) \notin E \end{cases} \quad \text{Εξίσωση (1)}$$

όπου π_{vx} είναι η μη κανονικοποιημένη πιθανότητα μετάβασης μεταξύ των κόμβων ν και χ και z είναι η σταθερά κανονικοποίησης. Το node2vec εισάγει το βάρος άλφα (α) έτσι ώστε:

$$\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx} \quad \text{Εξίσωση (2)}$$

όπου w_{vx} είναι το βάρος ακμής (ίσο με 1 για μη σταθμισμένους γράφους) και:

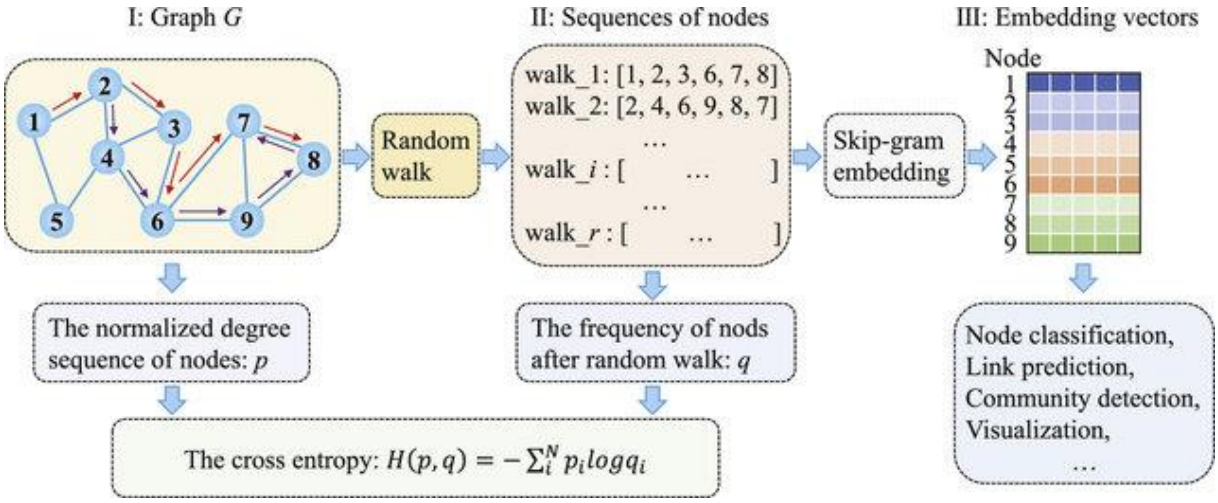
$$\alpha_{pq} = \begin{cases} \frac{1}{p}, \alpha v\ d_{tx} = 0 \\ 1, \alpha v\ d_{tx} = 1 \\ \frac{1}{q}, \alpha v\ d_{tx} = 2 \end{cases} \quad \text{Εξίσωση (3)}$$

όπου το d_{tx} δηλώνει την απόσταση συντομότερης διαδρομής μεταξύ των κόμβων t και x. Η Εξίσωση (1) δίνει την πιθανότητα συμπερίληψης του κόμβου στον τυχαίο περίπατο ενώ η Εξίσωση (2) δίνει την πιθανότητα μετάβασης που εισάγεται στην Εξίσωση (1). Η πιθανότητα μετάβασης προκύπτει από μια παράμετρο α που εισάγεται από το Node2vec. Η τιμή του α λαμβάνεται χρησιμοποιώντας δύο υπερπαραμέτρους που ονομάζονται p και q, οι οποίες δίνονται στην Εξίσωση (3). Οι δύο παράμετροι p και q ελέγχουν τη στρατηγική αναζήτησης μεταξύ Depth First Search (DFS) και Breadth First Search (BFS), φέρνοντας έτσι πιο αποτελεσματικές ακολουθίες κόμβων. Η αύξηση της τιμής p βοηθά την αναζήτηση με τυχαίο περίπατο βαθιά μέσα στο δίκτυο (DFS), ενώ η αύξηση της παραμέτρου q βοηθά στην ευρύτερη αναζήτηση, μέσα στο τοπικό δίκτυο του κόμβου προέλευσης (BFS). Οι ακολουθίες κόμβων τροφοδοτούνται στο μοντέλο Word2vec για τη δημιουργία word embeddings. Το Node2vec χρησιμοποιεί είτε το μοντέλο Skip-gram, όπου δεδομένης μιας λέξης, προβλέπονται οι γειτονικές της, είτε το CBOW, όπου προβλέπει την κεντρική λέξη όταν δίνονται οι γειτονικές της (Joshi et al., 2022).

Τα σημαντικότερα πλεονεκτήματα της Node2vec τεχνικής είναι η ευελιξία, καθώς μπορεί να ελέγξει διάφορες όψεις της δομής του γραφήματος μέσω των παραμέτρων p και q, και η αποτελεσματικότητα, διότι έχει αποδειχθεί ότι παράγει ακριβείς και χρήσιμες διανυσματικές αναπαραστάσεις για γραφήματα.

Το Node2vec βασίζεται στην ίδια βασική ιδέα με το Word2vec, δηλαδή της περιγραφής ενός κόμβου βάσει της γειτονιάς του, καθιστώντας το εύκολα κατανοητό και υλοποιήσιμο.

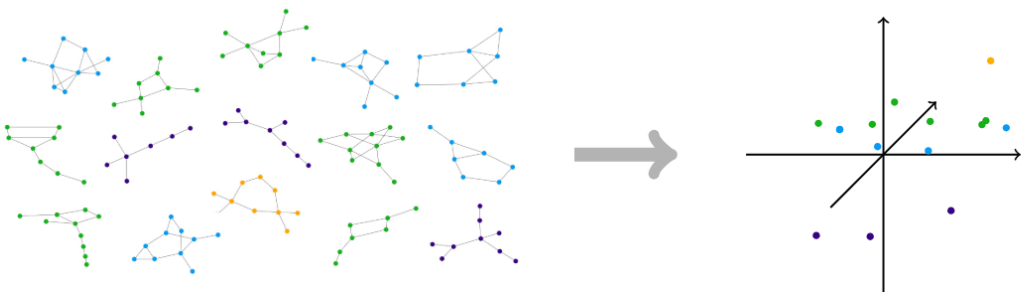
Η εκπαίδευση του Node2vec σε μεγάλα γραφήματα μπορεί να είναι χρονοβόρα και να απαιτεί σημαντική υπολογιστική ισχύ, ενώ η επιλογή των κατάλληλων τιμών για τους παραμέτρους p και q μπορεί να είναι δύσκολη και να επηρεάσει σημαντικά την ποιότητα των αποτελεσμάτων.



Εικόνα 30. Η αρχιτεκτονική του node2vec. Υπάρχουν τρεις διαφορετικοί τρόποι αναπαράστασης ενός δικτύου: Ι ως γράφος, ΙΙ ως ακολουθίες κόμβων και ΙΙΙ ως διανύσματα ενσωμάτωσης (embeddings). Από: (Li & Yang, 2024).

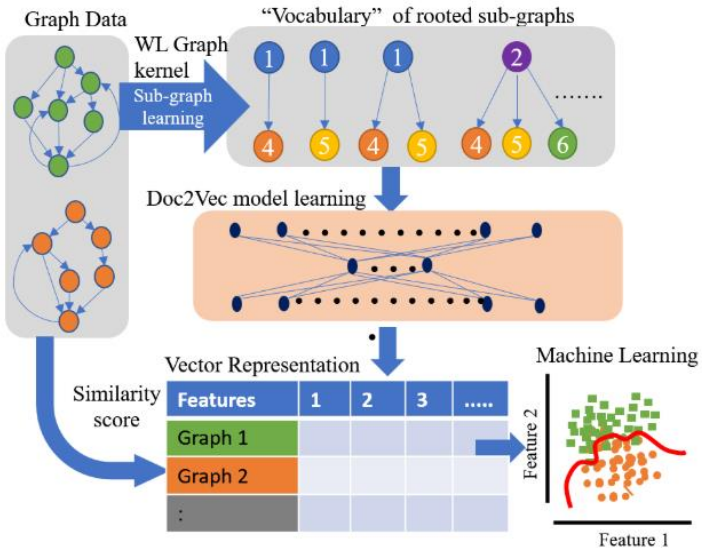
Επομένως, το Node2vec αποτελεί μια ισχυρή και ευέλικτη τεχνική μηχανικής μάθησης για την αναπαράσταση γραφημάτων. Η ικανότητά του να ελέγχει διάφορες όψεις της δομής του γραφήματος το καθιστά μια χρήσιμη επιλογή για διάφορες εφαρμογές. Ωστόσο, είναι σημαντικό να ληφθούν υπόψιν οι περιορισμοί του Node2vec, όπως η υπολογιστική δαπανηρότητα και η επιλογή παραμέτρων.

2.3.2 Graph2vec



Εικόνα 31. Αναπαράσταση γράφων ως σημεία σε ένα διανυσματικό χώρο δύο διαστάσεων – Η ιδέα του Graph2Vec. Από: (Grohe, 2020).

Το Graph2vec, όπως και το Word2vec και το Node2vec, αποτελεί μια δημοφιλή τεχνική μηχανικής μάθησης για την αναπαράσταση γραφημάτων. Σε αντίθεση με το Word2vec, το οποίο εστιάζει στην αναπαράσταση επιμέρους λέξεων (κόμβων), το Graph2vec παράγει μια συνολική αναπαράσταση για ολόκληρο το γράφημα (Εικόνα 31).



Εικόνα 32. Σχεδιάγραμμα εργασιών του Graph2Vec αλγορίθμου. Ως είσοδο στο μοντέλο εισάγεται ένας γράφος (πράσινο), για τον οποίο δημιουργείται ένα λεξικό από rooted subgraphs, το οποίο στη συνέχεια χρησιμοποιείται για την εκπαίδευση ενός Doc2Vec μοντέλου. Κάθε γράφος μετατρέπεται σε μορφή διανύσματος ορισμένων διαστάσεων (features). Από: (Liyanage et al., 2023).

Το Graph2vec βασίζεται στην ιδέα της "γειτονιάς", η οποία ορίζει το σύνολο των κόμβων που είναι άμεσα συνδεδεμένοι με έναν δεδομένο κόμβο. Στη συνέχεια, το Graph2vec εκπαιδεύεται για να μάθει μια συνολική διανυσματική αναπαράσταση για το γράφημα, λαμβάνοντας υπόψη τόσο τις γειτονιές όσο και τις σχέσεις μεταξύ των κόμβων (Εικόνα 32 και Εικόνα 33).

Σε αντίθεση με το Word2vec που εστιάζει στην αναπαράσταση επιμέρους λέξεων (κόμβων), το Graph2vec παράγει μια συνολική διανυσματική αναπαράσταση για ολόκληρο το γράφημα. Για τη δημιουργία ενός τελικού συγκεντρωτικού διανύσματος μπορούν να ακολουθηθούν διαφορετικές προσεγγίσεις:

- **Βασισμένες σε Word2vec:** Θεωρεί ολόκληρο τον γράφο ως ένα έγγραφο και τους rooted subraphs γύρω από κάθε κόμβο του γράφου ως λέξεις που συνθέτουν το έγγραφο (Narayanan et al., 2017). Χρησιμοποιεί το Word2vec σε κάθε κόμβο και

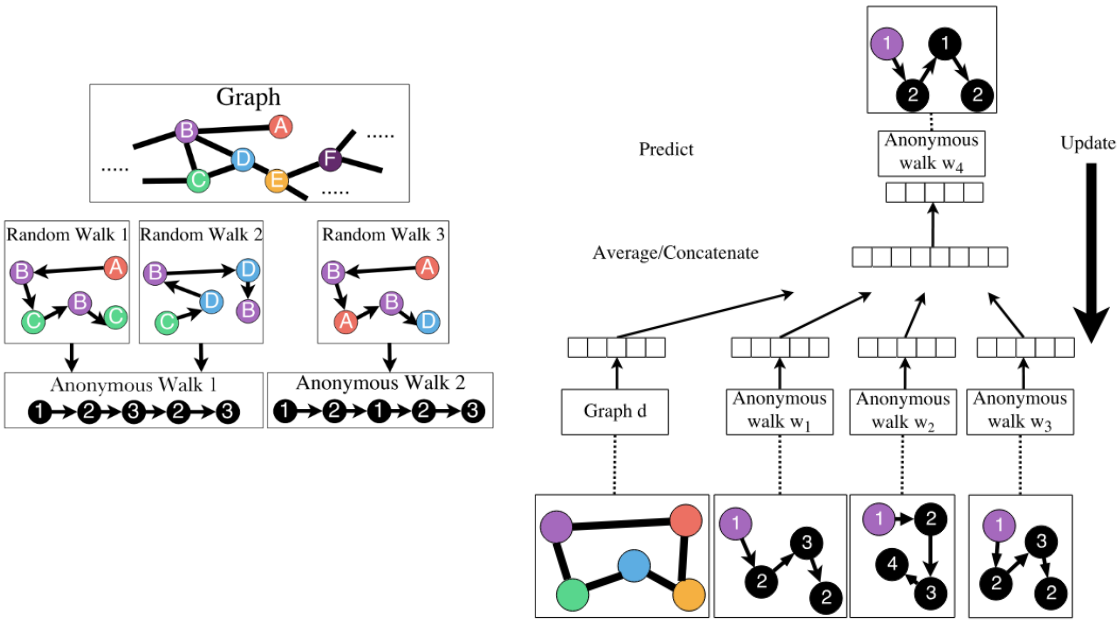
στη συνέχεια υπολογίζει το μέσο όρο αυτών, ή τα συγκεντρώνει σε ένα συνεχόμενο διάνυσμα (concatenation).

- **Βασισμένες σε Node2vec:** Χρησιμοποιεί το Node2vec σε κάθε κόμβο και στη συνέχεια, όπως και παραπάνω, είτε εξάγει ένα μέσο όρο όλων των διανυσμάτων, είτε τα συνενώνει σε ένα διάνυσμα.
- **Νευρωνικά δίκτυα:** Χρησιμοποιεί νευρωνικά δίκτυα για να μάθει απευθείας μια διανυσματική αναπαράσταση για ολόκληρο το γράφημα από τα δεδομένα του γραφήματος Χαρακτηριστικό παράδειγμα ενός τέτοιου μοντέλου είναι το GraphSAGE (Hamilton et al., 2018).

Η επιλογή της μεθόδου εξαγωγής ενός συνολικού διανύσματος για όλο το γράφημα, από τα επιμέρους διανύσματα ανά κόμβο, έχει μεγάλη σημασία, και εξαρτάται από το σκοπό της ανάλυσης που πραγματοποιείται κάθε φορά. Η εξαγωγή μέσου όρου δημιουργεί ένα μικρό διάνυσμα για όλο το γράφημα, ίσο με το μέγεθος του διανύσματος που αναπαριστά κάθε κόμβο. Επομένως, περαιτέρω αναλύσει είναι υπολογιστικά πιο αποδοτικές σε σύγκριση με τη μέθοδο της συνένωσης διανυσμάτων, όπου παράγεται ένα πολύ μεγάλης διάστασης διάνυσμα. Ωστόσο, το σύνολο της πληροφορίας από κάθε κόμβο διατηρείται αναλλοίωτο σε αυτή την περίπτωση.

Η αναπαράσταση ενός γραφήματος ως ένα μόνο διάνυσμα με τη μέθοδο Graph2vec μπορεί να είναι χρήσιμη για διάφορους λόγους, όπως η απλοποίηση, η συγκριτική ανάλυση, η εξόρυξη δεδομένων και η μηχανική μάθηση. Χρησιμοποιείται για την απλοποίηση της δομής του γραφήματος σε μια απλή μορφή, η οποία μπορεί να είναι πιο εύκολη στην επεξεργασία και ανάλυση. Στην περίπτωση ανάγκης σύγκρισης και ταξινόμησης γραφημάτων, είναι πολύ αποδοτικό να υπάρχει ένα μόνο διάνυσμα που να περικλείει όλη την πληροφορία κάθε γραφήματος. Το Graph2vec χρησιμοποιείται επίσης για την εξόρυξη δεδομένων και εξαγωγή πληροφοριών από γραφήματα, όπως ανίχνευση κοινοτήτων ή αναγνώριση μοτίβων. Τέλος, ένα γράφημα που έχει αποτυπωθεί ως ένα διάνυσμα μπορεί να

χρησιμοποιηθεί ως είσοδος σε αλγόριθμους μηχανικής μάθησης για εργασίες όπως η πρόβλεψη ή η ταξινόμηση γραφημάτων (Grohe, 2020).



Εικόνα 33. Αριστερά: Παράδειγμα της έννοιας του ανώνυμου περιπάτου. Δύο διαφορετικοί τυχαίοι περίπατοι 1 και 2 του γραφήματος αντιστοιχούν στον ίδιο ανώνυμο περίπατο 1. Ένας τυχαίος περίπατος 3 αντιστοιχεί σε έναν άλλο ανώνυμο περίπατο 2.

Δεξιά: Ένα πλαίσιο για την εκμάθηση ενσωματωμάτων ανώνυμων περιπάτων με βάση τα δεδομένα. Ο γράφος αναπαρίσταται από ένα διάνυσμα d και οι ανώνυμοι περίπατοι αναπαρίστανται από τις γραμμές του πίνακα W. Όλοι οι συνυπάρχοντες ανώνυμοι περίπατοι ξεκινούν από τον ίδιο κόμβο σε έναν γράφο. Στόχος είναι η πρόβλεψη ενός περιπάτου-στόχου w_4 από τους περιβάλλοντες περιπάτους (w_1, w_2, w_3) και ένα διάνυσμα γράφου d. Αρχικά υπολογίζεται ο μέσος όρος των ενσωματωμάτων των περιβαλλόντων περιπάτων και στη συνέχεια συνδυάζεται με ένα διάνυσμα γράφου για την πρόβλεψη του διανύσματος-στόχου.

Από: (Ivanov & Burnaev, 2018).

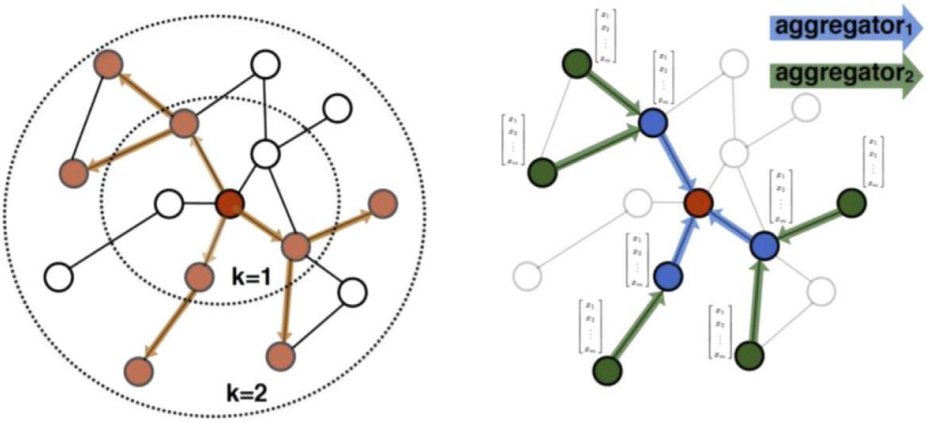
Συνολικά, το Graph2vec αποτελεί μια ισχυρή τεχνική για την αναπαράσταση γραφημάτων, προσφέροντας μια συνολική άποψη για τη δομή και τις σχέσεις του γραφήματος. Η ευελιξία και η ικανότητά του να λαμβάνει υπόψη σχέσεις το καθιστούν μια χρήσιμη επιλογή για διάφορες εφαρμογές. Ωστόσο, είναι σημαντικό να ληφθούν υπόψη οι περιορισμοί του Graph2vec, όπως η υπολογιστική δαπανηρότητα και η εξάρτηση από το μοντέλο εκπαίδευσης και τη μεθοδολογία εφαρμογής.

2.3.3 GraphSAGE

Το GraphSAGE αποτελεί μια δημοφιλή τεχνική μηχανικής μάθησης για την αναπαράσταση γραφημάτων, η οποία βασίζεται σε νευρωνικά δίκτυα για να μάθει διανυσματικές αναπαραστάσεις για κάθε κόμβο στο γράφημα.

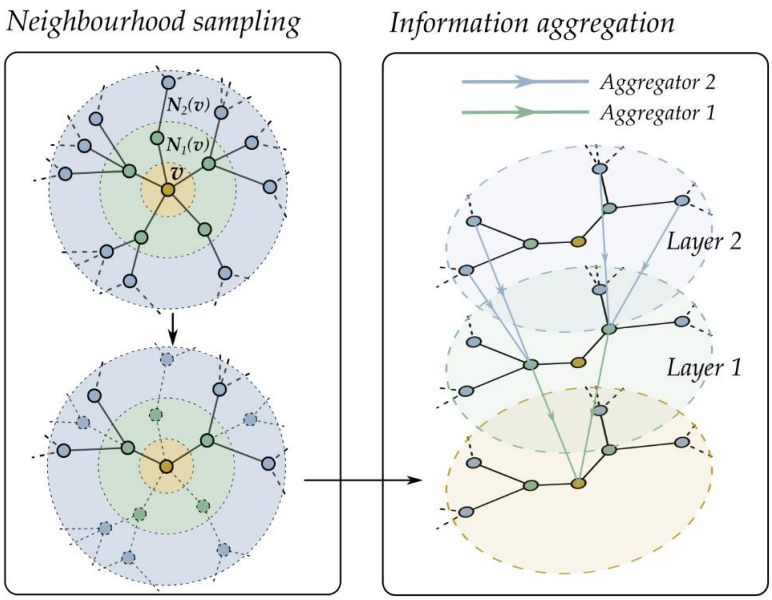
Σε αντίθεση με το Word2vec και το Node2vec, το GraphSAGE εστιάζει στην εκπαίδευση ενός μοντέλου νευρωνικού δικτύου που μαθαίνει να προβλέπει γειτονίες κάθε κόμβου. Η εκπαίδευση του GraphSAGE υλοποιείται σε 2 βήματα (Hamilton et al., 2018) και παρουσιάζεται διαγραμματικά στις Εικόνες 34 και 35:

- **Aggregation:** Συλλέγει πληροφορίες από γειτονικούς κόμβους κάθε κόμβου, λαμβάνοντας υπόψη τις βαρύτητες των συνδέσεων.
- **Transformation:** Επεξεργάζεται τις συγκεντρωμένες πληροφορίες με ένα νευρωνικό δίκτυο για να παράγει μια διανυσματική αναπαράσταση για τον κόμβο.



Εικόνα 34. Εξερεύνηση γειτονιών και συνένωση πληροφοριών στο μοντέλο GraphSAGE. Από: (Hamilton et al., 2018).

Το σημαντικότερο πλεονέκτημα του GraphSAGE είναι η αποδοτικότητά του, καθώς μπορεί να εκπαιδευτεί με πολύ αποτελεσματικό τρόπο σε μεγάλα γραφήματα. Μπορεί, επίσης, να εφαρμοστεί σε διάφορους τύπους γραφημάτων, λαμβάνοντας υπόψη διαφορετικούς τύπους γειτονιών. Παράλληλα, ενσωματώνει τα πλεονεκτήματα των νευρωνικών δικτύων, όπως την ικανότητα μάθησης μη γραμμικών σχέσεων.



Εικόνα 35. Οπτική αναπαράσταση των διαδικασιών δειγματοληψίας και συνένωνησς (sampling and aggregation) του GraphSAGE. Εν προκειμένω ο κόμβος-στόχος είναι ο v, και $N_1(v)$ και $N_2(v)$ οι γειτονιές ενός και δύο βημάτων απόστασης από τον κόμβο v. Από: (Li et al., 2023).

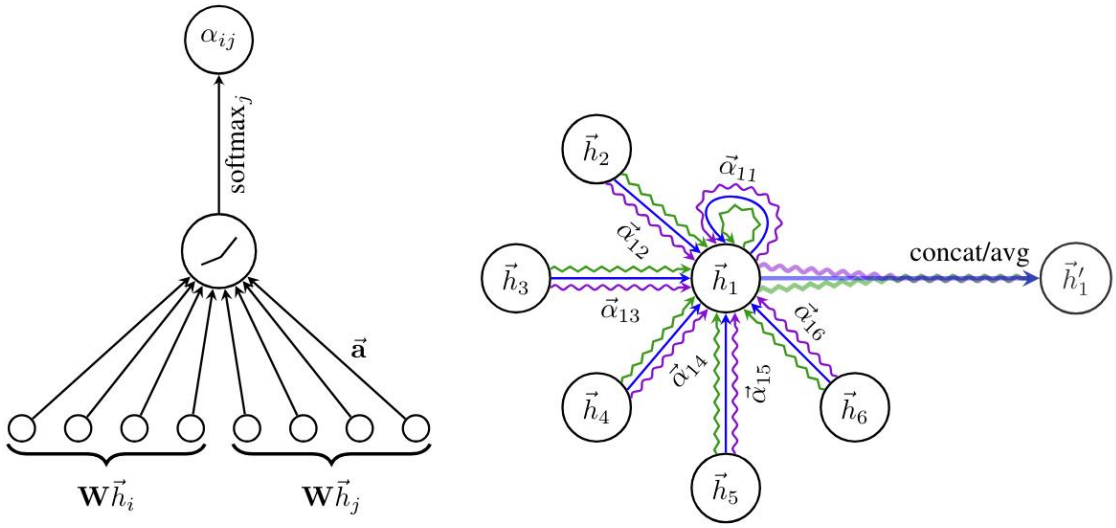
Το GraphSAGE αποτελεί μια ισχυρή και ευέλικτη τεχνική μηχανικής μάθησης για την αναπαράσταση γραφημάτων, η οποία εστιάζει στην εκμάθηση διανυσματικών αναπαραστάσεων για κάθε κόμβο με βάση τις γειτονιές του. Η ευελιξία, η αποδοτικότητα και η σχέση του με νευρωνικά δίκτυα το καθιστούν μια δημοφιλή επιλογή για διάφορες εφαρμογές. Ωστόσο, είναι σημαντικό να ληφθούν υπόψη οι περιορισμοί του GraphSAGE, όπως η εξάρτηση από τον τύπο γειτονιάς και η υπολογιστική δαπανηρότητα.

2.3.4 GAT (Graph Attention Networks)

Ο μηχανισμός προσοχής (attention) (Vaswani et al., 2023) έχει γίνει πλέον το πιο συχνά χρησιμοποιούμενο πρότυπο σε εργασίες που βασίζονται σε ακολουθίες (Bahdanau et al., 2016; Gehring et al., 2017) . Ένα από τα πλεονεκτήματα του μηχανισμού προσοχής είναι ότι επιτρέπει τη διαχείριση εισόδων μεταβλητού μεγέθους, εστιάζοντας στα πιο σημαντικά μέρη της εισόδου, κάθε φορά, για να λάβει αποφάσεις. Ο μηχανισμός προσοχής χρησιμοποιείται για τον υπολογισμό μιας αναπαράστασης μιας ενιαίας ακολουθίας αναφέρεται συνήθως ως αυτοπροσοχή (self-attention).

Τα Graph Attention Networks (GATs) είναι νευρωνικά δίκτυα που λειτουργούν σε δεδομένα δομημένα σε μορφή γράφου, χρησιμοποιώντας κρυμμένα επίπεδα αυτοπροσοχής. Το επίπεδο προσοχής (attention layer) του γράφου που χρησιμοποιείται σε όλα αυτά τα δίκτυα είναι υπολογιστικά αποδοτικό (δεν απαιτεί δαπανηρές πράξεις πινάκων και είναι

παραλληλίσιμο σε όλους τους κόμβους του γράφου), επιτρέπει την (έμμεση) ανάθεση διαφορετικής σημασίας σε διαφορετικούς κόμβους μέσα σε μια γειτονιά, ενώ μπορεί να χειριστεί γειτονιές διαφορετικού μεγέθους, και δεν εξαρτάται από τη γνώση ολόκληρης της δομής του γράφου εκ των προτέρων (Veličković et al., 2018). Συνοπτικά η λειτουργία του μηχανισμού προσοχής του GAN παρουσιάζεται στην Εικόνα 36.

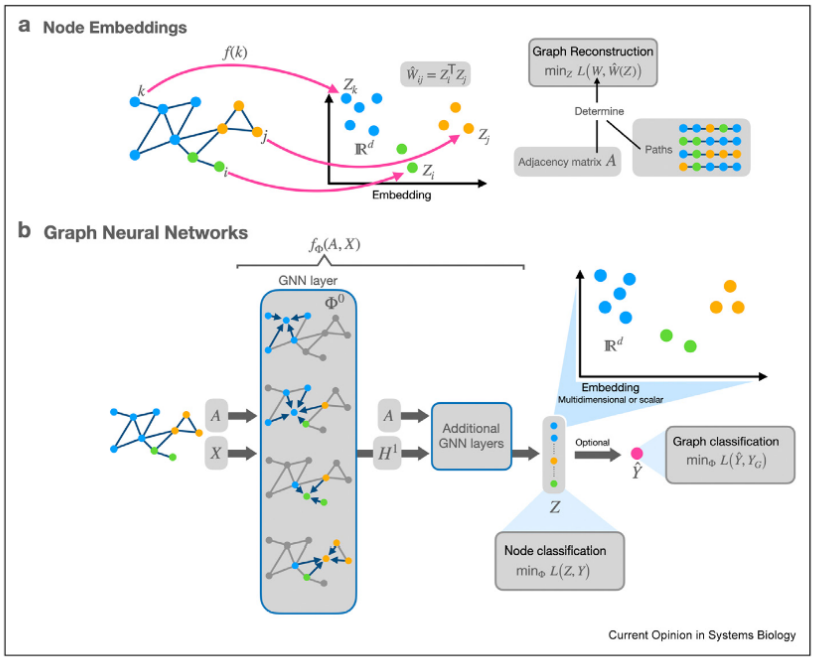


Εικόνα 36. Η λειτουργία του μηχανισμού προσοχής του GAN.

Αριστερά: Μηχανισμός Προσοχής. Κάθε κόμβος στο γράφημα έχει ένα σχετικό διάνυσμα χαρακτηριστικών (που αναπαρίσταται από τους κύκλους στο κάτω μέρος). Αυτά τα χαρακτηριστικά μετασχηματίζονται με τη χρήση πινάκων βαρύτητας ($W\vec{h}_i, W\vec{h}_j$), ώστε να προκύψει ένα νέο σύνολο χαρακτηριστικών σχετικών με τον υπολογισμό της προσοχής. Τα μετασχηματισμένα χαρακτηριστικά χρησιμοποιούνται για τον υπολογισμό των συντελεστών προσοχής (α_{ij}), που αντιπροσωπεύουν τη σημασία των χαρακτηριστικών του κόμβου j για τον κόμβο i. Αυτό γίνεται συχνά με τη χρήση ενός μηχανισμού όπως ο υπολογισμός του εσωτερικού γινομένου, ακολουθούμενος από μια συνάρτηση softmax για την κανονικοποίηση των συντελεστών. Η τελική αναπαράσταση του κόμβου i (h_i) υπολογίζεται ως σταθμισμένο άθροισμα των χαρακτηριστικών των γειτονικών του κόμβων, όπου τα βάρη είναι οι συντελεστές προσοχής (α_{ij}).

Δεξιά: Προσοχή πολλαπλών κεφαλών. Η αρχιτεκτονική GAT χρησιμοποιεί παράλληλα πολλαπλούς μηχανισμούς προσοχής (κεφαλές), καθένας από τους οποίους εστιάζει σε διαφορετικές πτυχές των χαρακτηριστικών των κόμβων. Αυτό επιτρέπει στο μοντέλο να αποτυπώσει μια πληρέστερη αναπαράσταση της δομής του γράφου. Οι έξοδοι από κάθε κεφαλή προσοχής είτε συνενώνονται είτε υπολογίζονται κατά μέσο όρο για να σχηματιστεί η τελική αναπαράσταση κόμβου (h'_i) για τον κόμβο i.

Από: (Veličković et al., 2018).



The two presented methods for graph representation learning: (*a*) Node embeddings and (*b*) Graph Neural Networks. (***a***) Nodes are mapped to a low dimensional space in which their representation should resemble a graph property *W* which can be computed from the adjacency matrix or paths on the graphs. The prediction of this similarity matrix $\hat{W}_{ij}$ is the inner product between nodes v_i and v_j. (***b***) Illustration of a GNN which takes the adjacency matrix *A* as well as node features *X* as inputs and then aggregates information over neighbourhoods to update the representation of nodes H^i. The final representation *Z* of nodes (coloured dots in grey box) can be multidimensional or scalar for node classification. For graph classification tasks with labels Y_G, this embedding can be projected further to just one prediction for the whole graph ($\hat{Y}$).

Εικόνα 37. Σύγκριση μεθόδων εκμάθησης αναπαράστασης γράφων: (α) Ενσωμάτωση κόμβων και (β) Νευρωνικά δίκτυα γράφων. (α) Οι κόμβοι απεικονίζονται σε ένα χώρο χαμηλών διαστάσεων στον οποίο η αναπαράστασή τους θα πρέπει να μοιάζει με μια ιδιότητα W του γράφου, η οποία μπορεί να υπολογιστεί από τον πίνακα γειτνίασης ή τα μονοπάτια στους γράφους. Η πρόβλεψη αυτού του πίνακα ομοιότητας $\widehat{W_{IJ}}$ είναι το εσωτερικό γινόμενο μεταξύ των κόμβων v_i και v_j . (β) Απεικόνιση ενός GNN το οποίο λαμβάνει τον πίνακα γειτνίασης A καθώς και τα χαρακτηριστικά των κόμβων X ως εισόδους και στη συνέχεια συγκεντρώνει πληροφορίες σχετικά με τις γειτονιές για να ενημερώσει την αναπαράσταση των κόμβων H^i . Η τελική αναπαράσταση Z των κόμβων (χρωματιστές κουκκίδες στο γκρι πλαίσιο) μπορεί να είναι πολυδιάστατη ή κλιμακωτή για την ταξινόμηση των κόμβων. Για εργασίες ταξινόμησης γραφημάτων με ετικέτες Y_G, αυτή η ενσωμάτωση μπορεί να προβάλλεται περαιτέρω σε μία μόνο πρόβλεψη για ολόκληρο το γράφημα ($\hat{Y}$). Από: (Hetzel et al., 2021).

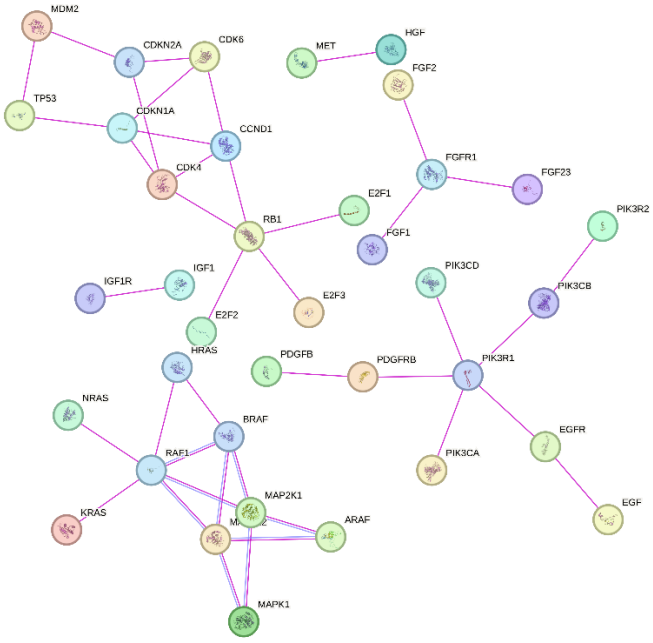
3 Συλλογή Δεδομένων

Στο κεφάλαιο αυτό θα περιγραφεί η διαδικασία συλλογής και προετοιμασίας των δεδομένων που χρησιμοποιήθηκαν για τις αναλύσεις που πραγματοποιήθηκαν.

3.1 Δεδομένα γράφων

Τα δεδομένα γράφων που αντιστοιχούν σε δίκτυα πρωτεϊνικών αλληλεπιδράσεων από ασθενείς με διάφορους τύπους καρκίνου προέρχονται από τη βάση δεδομένων String (Szklarczyk et al., 2023). Η String αποτελεί μια ολοκληρωμένη βάση δεδομένων πρωτεϊνικών αλληλεπιδράσεων, η οποία συγκεντρώνει πληροφορίες από διάφορες πηγές, όπως πειράματα, βιβλιογραφία και υπολογιστικές. Η βάση περιλαμβάνει δεδομένα για διάφορους οργανισμούς, συμπεριλαμβανομένου του ανθρώπου. Τα δεδομένα πρωτεϊνικών αλληλεπιδράσεων αναπαρίστανται ως γράφοι, με τους κόμβους να αντιστοιχούν στις πρωτεΐνες και τις ακμές στις αλληλεπιδράσεις μεταξύ τους, όπως φαίνεται στο παράδειγμα της Εικόνας 38.

Αρχικά επιλέχθηκαν 16 σύνολα δεδομένων από τη String που εστιάζουν σε ανθρώπινα δεδομένα πρωτεϊνικών αλληλεπιδράσεων για 11 κοινούς τύπους καρκίνου: μαστού, παχέος εντέρου, γλαυκώματος, νεφρών, λευχαιμίας, ήπατος, πνευμόνων, λεμφαδένων, μελανώματος, προστάτη και θυρεοειδούς.



Εικόνα 38. Παράδειγμα γράφου πρωτεϊνικών αλληλεπιδράσεων από πειραματικά δεδομένα, που παρήχθη από τη βάση δεδομένων String.

Στη συνέχεια επιλέχθηκαν να διατηρηθούν μόνο οι πειραματικά παρατηρηθείσες αλληλεπιδράσεις πρωτεϊνών, ώστε να διασφαλιστεί η αξιοπιστία των δεδομένων και η εστίαση σε αλληλεπιδράσεις με βιολογική σημασία.

Τα σύνολα δεδομένων που ανακτήθηκαν από τη String έχουν τη μορφή λίστας ακμών και περιλαμβάνουν πολλά πεδία με πληροφορίες. Για το σκοπό των αναλύσεων που πραγματοποιήθηκαν στα πλαίσια της συγκεκριμένης εργασίας, μόνο οι δύο πρώτες στήλες διατηρήθηκαν και χρησιμοποιήθηκαν. Τα πρώτα δύο πεδία (node1 και node2) αντιστοιχούν στις συντομογραφικές ονομασίες των πρωτεϊνών (*The Role of Protein and Amino Acids in Sustaining and Enhancing Performance*, 1999), μεταξύ των οποίων παρατηρείται η αλληλεπίδραση. Επομένως, κάθε καταχώρηση αντιστοιχεί σε μία ακμή στο γράφο πρωτεϊνικής αλληλεπίδρασης. Τα υπόλοιπα πεδία των συνόλων δεδομένων παρέχουν επιπλέον πληροφορίες για την γονιδιακή ομολογία, τη συνέκφραση, τη χρωμοσωμική απόσταση κ.ο.κ.. Για τους σκοπούς της συγκεκριμένης ερευνητικής εργασίας, μόνο τα δύο πρώτα πεδία των κόμβων χρησιμοποιήθηκαν.

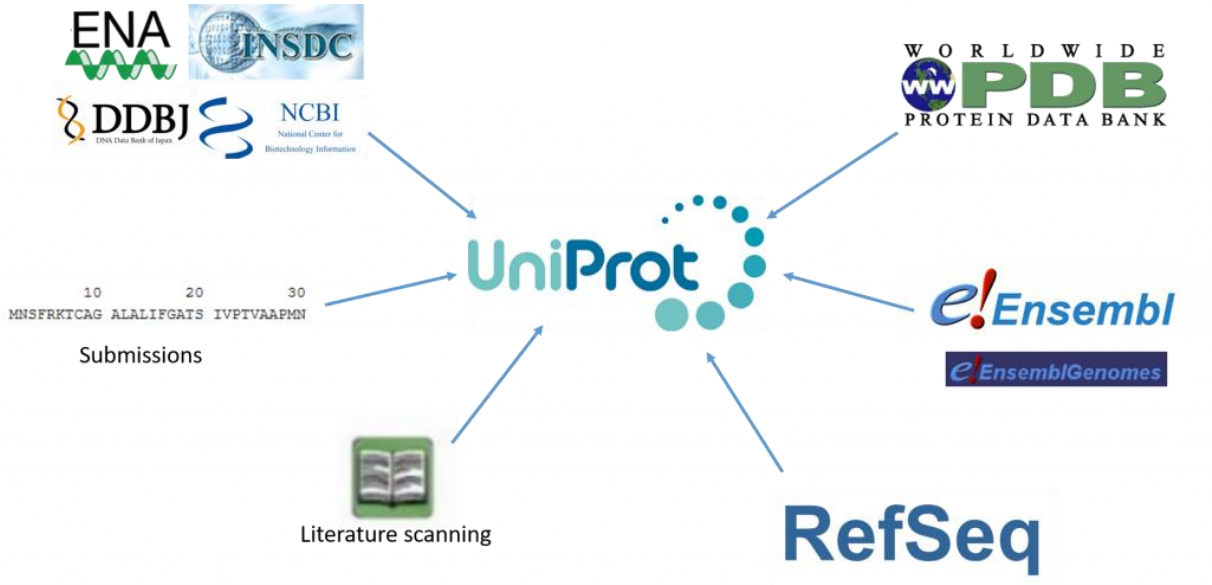
3.2 Δεδομένα κειμένου

Οι πληροφορίες κειμένου που περιγράφει τη λειτουργικότητα κάθε πρωτεΐνης αντλήθηκαν από τη βάση δεδομένων UniProt (The UniProt Consortium, 2007), που αποτελεί μια ολοκληρωμένη βάση δεδομένων πρωτεϊνών, η οποία παρέχει πληροφορίες για την αλληλεπίδραση, τη δομή, τη λειτουργία και άλλα χαρακτηριστικά πρωτεϊνών.

Η UniProt είναι μια βάση δεδομένων με ελεύθερα προσβάσιμες πληροφορίες για την αλληλουχία και τη λειτουργία των πρωτεϊνών. Περιέχει μεγάλο όγκο πληροφοριών σχετικά με τη βιολογική λειτουργία των πρωτεϊνών που προέρχονται από την ερευνητική βιβλιογραφία. Διατηρείται από την κοινοπραξία UniProt, η οποία αποτελείται από διάφορους ευρωπαϊκούς οργανισμούς βιοπληροφορικής και ένα ίδρυμα από την Ουάσινγκτον των ΗΠΑ.

Το UniProtKB/Swiss-Prot είναι μια μη πλεονάζουσα βάση δεδομένων πρωτεϊνικών αλληλουχιών με χειροκίνητο σχολιασμό. Συνδυάζει πληροφορίες που προέρχονται από την επιστημονική βιβλιογραφία και από υπολογιστικές αναλύσεις που έχουν αξιολογηθεί από βιοκριτές. Στόχος της UniProtKB/Swiss-Prot είναι να παρέχει όλες τις γνωστές σχετικές πληροφορίες για μια συγκεκριμένη πρωτεΐνη (Εικόνα 39). Ο σχολιασμός αναθεωρείται τακτικά για να συμβαδίζει με τα τρέχοντα επιστημονικά ευρήματα. Ο χειροκίνητος

σχολιασμός μιας καταχώρησης περιλαμβάνει λεπτομερή ανάλυση της πρωτεϊνικής αλληλουχίας και της επιστημονικής βιβλιογραφίας.



Εικόνα 39. Πηγές δεδομένων από τις οποίες αντλεί πληροφορίες η UniProt. Από: https://www.uniprot.org/.

Η UniProt επιλέχθηκε για την άντληση δεδομένων για τις αναλύσεις της συγκεκριμένης εργασίας, λόγω της αξιοπιστίας και της πληρότητάς της, που την καθιστούν ιδανική για τη βιοπληροφορική ανάλυση. Η UniProt είναι αναγνωρισμένη διεθνώς και χρησιμοποιείται ευρέως από την επιστημονική κοινότητα, διασφαλίζοντας έτσι ότι τα δεδομένα που αντλήθηκαν είναι επικυρωμένα και αποδεκτά.

Για την ανάκτηση των κειμένων που περιγράφουν τη λειτουργικότητα κάθε πρωτεΐνης αξιοποιήθηκε η δωρεάν πρόσβαση στο API που παρέχει η UniProt (Patient et al., 2008; The UniProt Consortium et al., 2021). Αναλυτικά τα βήματα που ακολουθήθηκαν παρουσιάζονται παρακάτω:

i. **Φόρτωση Δικτύου Πρωτεϊνικής Αλληλεπίδρασης:**

   Η διαδικασία αρχίζει με τη φόρτωση ενός δικτύου πρωτεϊνικής αλληλεπίδρασης από τα διαθέσιμα σύνολα δεδομένων. Η ανάκτηση των πληροφοριών κειμένου πραγματοποιούνταν σε πραγματικό χρόνο κάθε φορά, επομένως το δίκτυο πρέπει να έχει ήδη φορτωθεί στη μνήμη ώστε να γίνει η προσπέλαση των κόμβων.

ii. **Ανάκτηση Πληροφοριών Κειμένου για Κάθε Πρωτεΐνη:**

Για κάθε κόμβο στο δίκτυο (πρωτεΐνη) δημιουργείται προγραμματιστικά ένα ερώτημα (query) προς το API της UniProt (Patient et al., 2008), προσδιορίζοντας κάθε φορά διαφορετικό όνομα πρωτεΐνης. Στο ερώτημα επίσης ορίζεται και το όνομα του οργανισμού (Nightingale et al., 2017), καθώς η UniProt παρέχει πληροφορίας για μεγάλη ποικιλία οργανισμών, και όχι μόνο για τον άνθρωπο, που είναι ο στόχος της συγκεκριμένης ερευνητικής εργασίας.

Εάν η αναζήτηση είναι επιτυχής και το ερώτημα λάβει απάντηση, μια αναλυτική περιγραφή της πρωτεΐνης επιστρέφεται. Η απάντηση αυτή περιλαμβάνει πολλή μεγάλη ποσότητα πληροφορίας, ένα μικρό μόνο μέρος της οποίας αντιστοιχεί στην περιγραφή της λειτουργικότητας της πρωτεΐνης, που είναι το ζητούμενο.

iii. Επεξεργασία Κειμένου Περιγραφής:

Το κείμενο περιγραφής που λαμβάνεται από την UniProt μπορεί να περιέχει ετικέτες HTML, κενές γραμμές και άλλα περιττά στοιχεία, τα οποία αφαιρούνται προγραμματιστικά, ώστε στο τέλος να διατηρηθεί μόνο το απλό κείμενο. Αυτό στη συνέχεια διασπάται σε λέξεις (tokens) και απορρίπτονται άσχετες ή συχνές λέξεις (stop words) (Sahani, 2023).

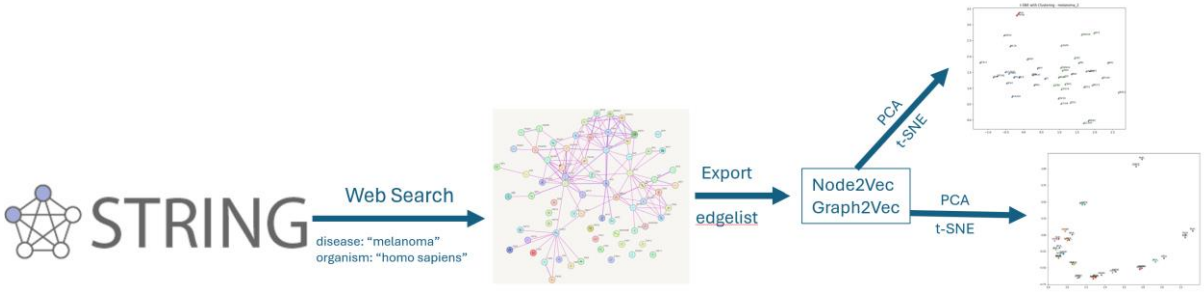
iv. Δημιουργία Λεξικού με Πληροφορίες Κειμένου:

Έπειτα δημιουργείται ένα λεξικό όπου κάθε κλειδί αντιστοιχεί σε μια πρωτεΐνη (όνομα γονιδίου) και η αντίστοιχη τιμή είναι το επεξεργασμένο κείμενο περιγραφής της πρωτεΐνης. Αυτό το λεξικό περιέχει πληροφορίες κειμένου για κάθε πρωτεΐνη στο δίκτυο, και είναι αυτό που θα αξιοποιηθεί στη συνέχεια για την εξαγωγή ενσωματωμάτων με τις διαφορετικές μεθόδους που θα περιγραφούν στο επόμενο κεφάλαιο.

4 Μεθοδολογία

Σε αυτό το κεφάλαιο περιγράφεται λεπτομερώς η μεθοδολογία που ακολουθήθηκε στην παρούσα εργασία για την εξαγωγή διανυσματικών αναπαραστάσεων από δίκτυα πρωτεϊνικής αλληλεπίδρασης σε διάφορους τύπους καρκίνου, με τη χρήση διαφορετικών μεθόδων και την αξιοποίηση διαφορετικών πηγών δεδομένων. Ο στόχος της μελέτης είναι η εφαρμογή διαφορετικών μεθόδων διανυσματικής αναπαράστασης γράφων και η αξιολόγηση της απόδοσής τους.

4.1 Διανυσματοποίηση δικτύων πρωτεϊνικών αλληλεπιδράσεων



Εικόνα 40. Διάγραμμα ροής της μεθοδολογίας διανυσματοποίησης των γράφων πρωτεϊνικών αλληλεπιδράσεων διαφόρων τύπων καρκίνου, που αντλήθηκαν από τη UniProt.

4.1.1 Graph2Vec

Αρχικά, εφαρμόστηκε η τεχνική Graph2Vec για τη διανυσματοποίηση όλων των γράφων πρωτεϊνικών αλληλεπιδράσεων που συγκεντρώθηκαν από τη String. Το Graph2Vec είναι μια μέθοδος που μετατρέπει ολόκληρους γράφους σε μοναδικά ενσωματώματα. Βασίζεται στην ιδέα της εκμάθησης αναπαραστάσεων για κόμβους και τις συνδυάζει για να παραχθούν συνολικές αναπαραστάσεις γράφων.

Κάθε γράφος, που αντιπροσωπεύει ένα σύνολο πρωτεϊνικών αλληλεπιδράσεων, μετατρέπεται σε ένα μοναδικό ενσωμάτωμα. Αυτά τα ενσωματώματα αποτυπώνουν τις δομικές και λειτουργικές ιδιότητες των γράφων. Αρχικά χρησιμοποιήθηκε ένα νευρωνικό δίκτυο για την εκπαίδευση του Graph2Vec στα διαθέσιμα δεδομένα. Το δίκτυο μαθαίνει να δημιουργεί ενσωματώματα γράφων που διατηρούν τις πληροφορίες για τις δομές και τις αλληλεπιδράσεις των πρωτεϊνών.

Μετά την ολοκλήρωση της εκπαίδευσης, το Graph2Vec παράγει τα τελικά ενσωματώματα για κάθε γράφο. Αυτά τα ενσωματώματα μπορούν να χρησιμοποιηθούν σε διάφορες

αναλυτικές εφαρμογές, όπως η κατηγοριοποίηση και η ομαδοποίηση πρωτεϊνικών αλληλεπιδράσεων.

Τα ενσωματώματα παρέχουν μια συνολική και περιεκτική αναπαράσταση των πρωτεϊνικών αλληλεπιδράσεων που παρατηρούνται σε κάθε δίκτυο, και προσφέρουν μια εποπτική εικόνα στις ομοιότητες των δικτύων πρωτεϊνικής αλληλεπίδρασης μεταξύ διαφόρων τύπων καρκίνου

4.1.2 Node2Vec

Στη συνέχεια, πραγματοποιήθηκε μετάβαση στο αμέσως επόμενο επίπεδο και διανυσματοποιήθηκε κάθε επιμέρους κόμβος του κάθε δικτύου. Η διαδικασία που ακολουθήθηκε για την εξαγωγή των ενσωματωμάτων για κάθε πρωτεΐνη με τη μέθοδο node2vec παρουσιάζονται στη συνέχεια:

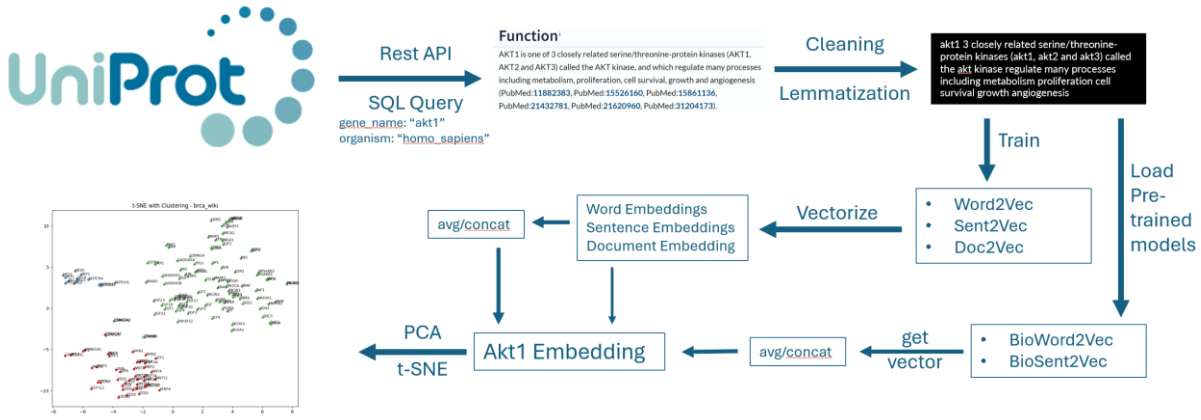
i. **Προεπεξεργασία και φόρτωση Δεδομένων:**

Αρχικά, κάθε δίκτυο πρωτεϊνικής αλληλεπίδρασης φορτώθηκε σε μορφή λίστας ακμών, όπου κάθε γραμμή αντιπροσώπευε μια αλληλεπίδραση μεταξύ δύο πρωτεϊνών. Στη συνέχεια το κάθε δίκτυο τροποποιήθηκε κατάλληλα, και αφαιρέθηκε η περιττή πληροφορία και τα δεδομένα μετατράπηκαν στη συνέχεια σε αντικείμενα της κατάλληλης βιβλιοθήκης networkx, ώστε να είναι αναγνωρίσιμα και συμβατά με τη βιβλιοθήκη Node2vec.

ii. **Εκπαίδευση Node2vec και λήψη ενσωματωμάτων:**

Αρχικά, πριν την έναρξη της διαδικασίας εκπαίδευσης, ορίστηκαν παράμετροι όπως το μήκος της διαδρομής, ο αριθμός των διαδρομών ανά κόμβο και το πλήθος των διαστάσεων του embedding που θα παραχθεί για κάθε κόμβο. Έπειτα, το μοντέλο Node2vec εκπαιδεύτηκε στο προεπεξεργασμένο δίκτυο πρωτεϊνικής αλληλεπίδρασης. Για κάθε κόμβο, που αντιστοιχεί σε μια πρωτεΐνη, λήφθηκε το αντίστοιχο ενσωμάτωμά της από το εκπαιδευμένο μοντέλο Node2vec. Τα ενσωματώματα αποθηκεύτηκαν σε μορφή αρχείου για περαιτέρω χρήση.

4.2 Διανυσματοποίηση κειμένου λειτουργικότητας πρωτεϊνών



Εικόνα 41. Διάγραμμα ροής της μεθοδολογίας που εφαρμόστηκε για τη διανυσματοποίηση του κειμένου περιγραφής των πρωτεϊνικών αλληλεπιδράσεων που αντλήθηκε από τη UniProt.

4.2.1 Word2Vec

Τα κείμενα περιγραφής της λειτουργικότητας κάθε πρωτεΐνης κάθε συνόλου δεδομένων, όπως ανακτήθηκαν από τη βάση δεδομένων UniProt, χρησιμοποιήθηκαν στη συνέχεια ως είσοδοι σε ένα word2vec μοντέλο, για την παραγωγή embeddings για κάθε πρωτεΐνη-κόμβο των επιμέρους συνόλων δεδομένων. Τα βήματα που ακολουθήθηκαν περιγράφονται παρακάτω:

i. Προεπεξεργασία Δεδομένων:

Αρχικά πραγματοποιήθηκε καθαρισμός των αντληθέντων κειμένων, με αφαίρεση σημείων στίξης, κεφαλαίων γραμμάτων και ειδικών χαρακτήρων, ώστε το κείμενο να μετατραπεί σε μια πιο απλή μορφή που να μπορεί να τη διαχειριστεί το word2vec μοντέλο. Στη συνέχεια κάθε περιγραφή κατακερματίστηκε σε tokens και αφαιρέθηκαν συχνά χρησιμοποιούμενες λέξεις χωρίς νόημα (πχ άρθρα, αντωνυμίες, σύνδεσμοι, προθέσεις κ.λπ.) (Tohti et al., 2017). Τέλος εφαρμόστηκαν τεχνικές stemming/lemmatization με σκοπό την αναγωγή των λέξεων που απομονώθηκαν, στις ριζικές τους μορφές, δηλαδή χωρίς καταλήξεις, αυξήσεις κλπ. Με τον τρόπο αυτό διατηρείται η σημασιολογία κάθε λέξης, κάτι που μπορεί να εκμεταλλευτεί το μοντέλο word2vec για την πιο αποδοτική εκπαίδευσή του (Senders, 2021).

ii. Εκπαίδευση Word2vec και λήψη embeddings:

Για την υλοποίηση της μεθόδου Word2Vec χρησιμοποιήθηκε η βιβλιοθήκη GenSim (L. Wang et al., 2024) της python. Το Word2Vec μοντέλο εκπαιδεύτηκε με τον CBOW

αλγόριθμο, και το μέγεθος των παραγόμενων διανυσμάτων ορίστηκε στις 100 διαστάσεις. Μετά την ολοκλήρωση της εκπαίδευσης του μοντέλου, πραγματοποιήθηκε ανάκτηση των embeddings για κάθε πρωτεΐνη-κόμβο. Τα embeddings στη συνέχεια αποθηκεύτηκαν σε ξεχωριστό αρχείο με μορφή αρχείου τιμών χωρισμένων με κόμμα (csv) για περαιτέρω χρήση και σύγκριση.

Η παραπάνω διαδικασία εκπαίδευσης εφαρμόστηκε για όλα τα σύνολα δεδομένων. Για κάθε σύνολο δεδομένων, που αντιστοιχεί σε ένα δίκτυο, δημιουργήθηκε ένα αρχείο που περιέχει το όνομα κάθε κόμβου – πρωτεΐνης και τις συντεταγμένες του διανύσματος σε κάθε διάσταση.

4.2.2 BioWord2Vec

Για την εφαρμογή της τεχνικής BioWord2Vec για την εξαγωγή διανυσμάτων για κάθε κόμβου των δικτύων πρωτεϊνικών αλληλεπιδράσεων χρησιμοποιήθηκε ένα προεκπαιδευμένο μοντέλο με διανύσματα (keyed vector pretrained model) 200 διαστάσεων, το οποίο έχει εκπαιδευτεί σε βιολογικά και ιατρικά κείμενα από την PubMed και στις σχέσεις μεταξύ των επιμέρους όρων από το γράφο γνώσης του MeSH. Το προεκπαιδευμένο αυτό μοντέλο αποτελεί ερευνητικό έργο των (Zhang et al., 2019) και διατίθεται ελεύθερα.

Η διαδικασία προεπεξεργασίας των δεδομένων κειμένου που εφαρμόστηκε δεν παρουσιάζει καμία διαφορά σε σχέση με την μέθοδο Word2Vec. Ωστόσο, στην BioWord2Vec δεν πραγματοποιήθηκε εκπαίδευση του μοντέλου, καθώς χρησιμοποιήθηκε το ήδη προ-εκπαιδευμένο μοντέλο των (Zhang et al., 2019). Στο μοντέλο αυτό εισήχθη το όνομα κάθε πρωτεΐνης-κόμβου των δικτύων πρωτεϊνικών αλληλεπιδράσεων, και από το όνομα και μόνο παράχθηκαν τα αντίστοιχα embeddings. Η επιλογή αυτή βασίστηκε στο γεγονός ότι το μοντέλο έχει εκπαιδευτεί ήδη σε βιολογικά και ιατρικά δεδομένα, και στην υπόθεση ότι έχει συμπεριλάβει το νόημα των ονομάτων των πρωτεϊνών στα προεκπαιδευμένα διανύσματά του.

Για ορισμένες πρωτεΐνες δεν υπήρχε καταχώρηση στο λεξικό του προεκπαιδευμένου μοντέλου, γεγονός που οδήγησε σε ορισμένα μηδενικά διανύσματα για κάποιες πρωτεΐνες, περιορίζοντας την αποτελεσματικότητα της συγκεκριμένης μεθόδου. Τέλος πραγματοποιήθηκε ανάκτηση των embeddings, τα οποία και αποθηκεύτηκαν σε αρχείο csv για μελλοντική χρήση.

4.2.3 BioSent2Vec

Η εφαρμογή της μεθόδου BioSent2Vec στηρίζεται επίσης σε ένα προεκπαιδευμένο μοντέλο διανυσμάτων, το οποίο έχει εκπαιδευτεί σε βιολογικά και ιατρικά κείμενα από την PubMed και στις σχέσεις μεταξύ των επιμέρους όρων από το γράφο γνώσης του MeSH. Πρόκειται για ένα μοντέλο με διανύσματα 700 διαστάσεων που αναπτύχθηκε και εκπαιδεύτηκε από τους (Chen et al., 2019). Ωστόσο, το BioSentVec μοντέλο διαφέρει από το BioWordVec μοντέλο καθώς δεν εκπαιδεύθηκε στην κατανόηση του νοήματος μεμονωμένων λέξεων, αλλά φράσεων/προτάσεων. Γι' αυτό και τα διανύσματα που παράγει είναι πολύ περισσότερων διαστάσεων (700) αντί για 200 στο αντίστοιχο BioWordVec μοντέλο, ώστε να συλλάβουν το πολύ πιο πολύπλοκο και πολυδιάστατο νόημα που μπορεί να εμπεριέχει μια πρόταση, σε σχέση με μια μεμονωμένη λέξη.

Η προεπεξεργασία και ο καθαρισμός των προτάσεων δεν περιλαμβάνει την αφαίρεση λέξεων ή την αναγωγή στη ρίζα, όπως στο Word2Vec. Πραγματοποιείται όμως μετατροπή όλων των χαρακτήρων σε πεζούς, και στη συνέχεια αναγνώριση λεξικών μονάδων (tokenization) για τη δημιουργία των προτάσεων που θα εισαχθούν στο BioSentVec μοντέλο. Οι προτάσεις προέρχονται από τα κείμενα που περιγράφουν τη λειτουργικότητα κάθε πρωτεΐνης-κόμβου από τη UniProt. Εφόσον το κείμενο της περιγραφής κάθε πρωτεΐνης περιλαμβάνει περισσότερες από μία προτάσεις, θα προκύψουν και περισσότερα από ένα embeddings για τη συγκεκριμένη πρωτεΐνη. Ο μέσος όρος των επιμέρους embeddings των προτάσεων αποτελεί το τελικό embedding κάθε πρωτεΐνης. Για κάθε πρωτεΐνη κάθε δικτύου πρωτεϊνικών αλληλεπιδράσεων παρήχθη ένα embedding, τα οποία αποθηκεύτηκαν στη συνέχεια σε ένα αρχείο csv για κάθε σύνολο δεδομένων, για μελλοντική χρήση και περαιτέρω αναλύσεις.

4.2.4 BioDoc2Vec

Η τεχνική BioDoc2Vec αξιοποιεί τα ενσωματώματα του μοντέλου BioDocVec προεκπαιδευμένων διανυσμάτων για τη δημιουργία πληρέστερων και ακριβέστερων διανυσματικών αναπαραστάσεων των πρωτεϊνών. Για την υλοποίησή της χρησιμοποιήθηκε πάλι το BioWordVec μοντέλο των (Zhang et al., 2019), αλλά αυτή τη φορά εισήχθησαν όλες οι λέξεις της περιγραφής κάθε πρωτεΐνης. Σε αντίθεση με την BioWord2Vec τεχνική, που περιεγράφηκε παραπάνω, και παράγει ένα μοναδικό διάνυσμα για το όνομα κάθε πρωτεΐνης, η BioDoc2Vec προσπαθεί να συμπεριλάβει όλη την πληροφορία των περιγραφών των πρωτεϊνών.

Για την υλοποίηση της μεθόδου αυτής χρησιμοποιήθηκαν οι περιγραφές των πρωτεϊνών που ανακτήθηκαν από τη βάση δεδομένων UniProt. Η κάθε περιγραφή υποβλήθηκε σε διαδικασία φιλτραρίσματος κειμένου, όπου απομακρύνθηκαν οι άχρηστες πληροφορίες και οι περιττές λέξεις και έγινε κανονικοποίηση του κειμένου. Στη συνέχεια, για κάθε λέξη της περιγραφής υπολογίστηκε ένα διάνυσμα μέσω του μοντέλου BioDocVec. Αυτή η προσέγγιση επιτρέπει την αξιοποίηση λεπτομερών και διαφοροποιημένων πληροφοριών για κάθε πρωτεΐνη, καθώς λαμβάνονται υπόψη όλες οι λέξεις που περιγράφουν την πρωτεΐνη.

Για να προκύψει ένα μοναδικό ενσωμάτωμα για κάθε πρωτεΐνη κόμβο, υπολογίστηκε ο μέσος όρος όλων των ενσωματωμάτων κάθε λέξης της περιγραφής. Αυτή η διαδικασία διασφαλίζει ότι το τελικό διάνυσμα αναπαριστά συνολικά την πρωτεΐνη και περιλαμβάνει πληροφορίες από όλο το εύρος της περιγραφής της.

4.3 Μείωση διάστασης και οπτική αναπαράσταση

Για τη μείωση της διάστασης των δεδομένων μας και την οπτική τους αναπαράσταση, εφαρμόστηκαν δύο κύριες τεχνικές: t-Distributed Stochastic Neighbor Embedding t-SNE) (van der Maarten & Hinton, 2008) και Principal Component Analysis (PCA) (Makiewicz & Ratajczak, 1993). Αυτές οι τεχνικές επιτρέπουν τη μείωση του πλήθους των διαστάσεων των δεδομένων μας, διατηρώντας παράλληλα την ουσία και τη δομή τους.

Αρχικά, χρησιμοποιήθηκε η μέθοδος PCA για τη γραμμική μείωση των διαστάσεων. Η PCA είναι ιδανική για την εξαγωγή των κυριότερων συνιστωσών των δεδομένων και την απεικόνιση της διακύμανσής τους σε χαμηλότερες διαστάσεις. Στη συνέχεια, εφαρμόστηκε η μέθοδος t-SNE , η οποία είναι ιδιαίτερα αποτελεσματική για τη μη γραμμική μείωση της διάστασης και τη διατήρηση των τοπικών σχέσεων των δεδομένων (Cai & Ma, 2022).

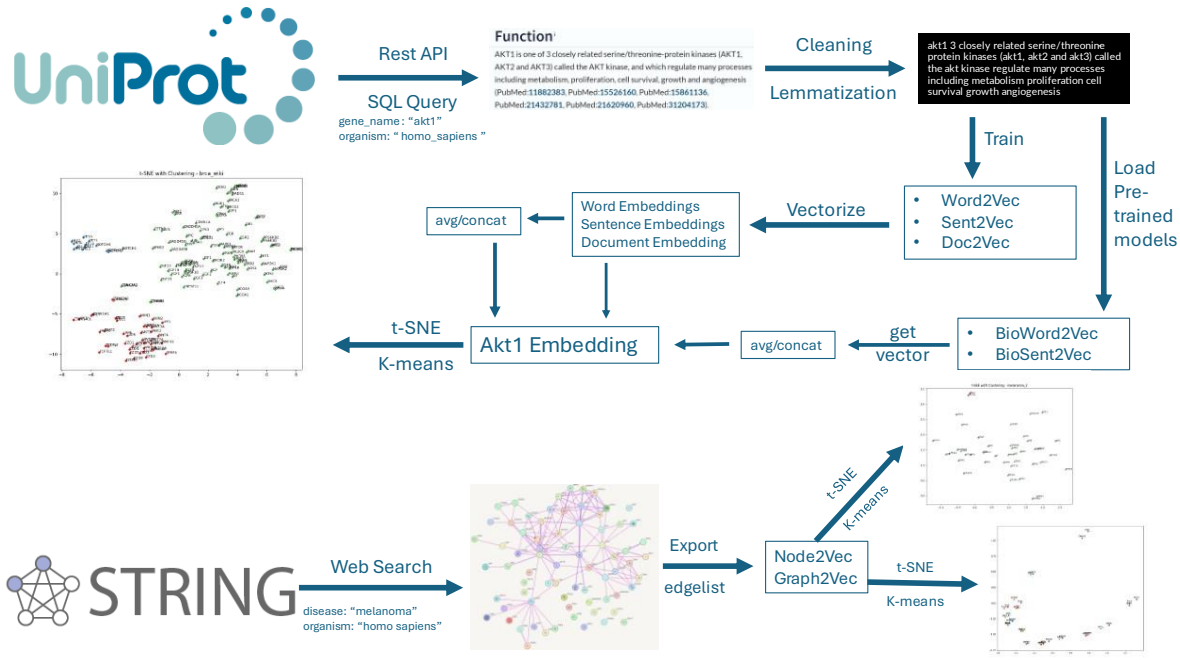
Το επιθυμητό πλήθος διαστάσεων μετά τη μείωση καθορίστηκε να είναι 2 ή 3, ώστε τα αποτελέσματα να μπορούν να αναπαρασταθούν γραφικά σε δισδιάστατες ή τρισδιάστατες αναπαραστάσεις. Αυτή η προσέγγιση επιτρέπει την παρατήρηση των δεδομένων σε έναν πιο διαχειρίσιμο και κατανοητό χώρο, καθιστώντας εμφανείς τις σχέσεις και τις συσχετίσεις μεταξύ των πρωτεϊνών.

Επιπλέον, πραγματοποιήθηκε ομαδοποίηση των δεδομένων για να αποδοθεί μια πρώιμη βιολογική έννοια στα αποτελέσματα. Χρησιμοποιήθηκε ο αλγόριθμος ομαδοποίησης κ-

μέσων (k-means clustering) (Ahmed et al., 2020; Steinhaus, 1957). Η ομαδοποίηση αυτή έχει ως στόχο να διαχωρίσει n παρατηρήσεις σε k ομάδες, έτσι ώστε κάθε παρατήρηση να ανήκει στη συστάδα με το κοντινότερο μέσο, το οποίο χρησιμεύει ως ένα χαρακτηριστικό δείγμα της κάθε συστάδας.

Μέσω της ομαδοποίησης, μπορεί να εξεταστεί αν πρωτεΐνες με γνωστή παρόμοια λειτουργικότητα και έκφραση τοποθετούνται κοντά μεταξύ τους στον μειωμένο χώρο. Αυτό επιτρέπει μια προκαταρκτική αξιολόγηση των δεδομένων μας, διερευνώντας αν τα διανύσματα των πρωτεϊνών, μετά τη μείωση της διάστασης, βρίσκονται κοντά μεταξύ τους στο δισδιάστατο ή τρισδιάστατο χώρο.

Τα αποτελέσματα αυτής της διαδικασίας δίνουν μια αρχική ένδειξη για την ορθότητα και τη χρησιμότητα των αναπαραστάσεων, καθώς και για την ικανότητα των μεθόδων μείωσης της διάστασης να διατηρήσουν τις βιολογικά σημαντικές σχέσεις των δεδομένων.



Εικόνα 42. Συγκεντρωτικό διάγραμμα της ερευνητικής μεθοδολογίας που ακολουθήθηκε για τη διανυσματοποίηση και παραγωγή ενσωματωμάτων των δικτύων αλληλεπίδρασης πρωτεϊνών σε διάφορους τύπους καρκίνου, τόσο με τη χρήση της πληροφορίας των γράφων από τη String, όσο και τις πληροφορίες κειμένου των περιγραφών της λειτουργίας των πρωτεϊνών από τη UniProt.

4.4 Αξιολόγηση Αποτελεσμάτων

Μετά και την οπτική αναπαράσταση των ενσωματωμάτων που παρήχθησαν από τις διαφορετικές μεθόδους διανυσματοποίησης, το επόμενο βήμα είναι η αξιολόγηση της ποιότητας των αποτελεσμάτων. Δεν υπάρχει κάποιος άμεσος τρόπος σύγκρισης των

ενσωματωμάτων που παρήχθησαν από τις διαφορετικές μεθόδους, καθώς ανήκουν σε διαφορετικούς διανυσματικούς χώρους, και επομένως η τοποθέτησή τους στο ίδιο επίπεδο και η μέτρηση αποστάσεων δεν έχει νόημα. Επομένως, η προσέγγιση που εφαρμόστηκε ήταν να διερευνηθεί η απόδοση κάθε μεθόδου διανυσματοποίησης μέσω ομαδοποίησης. Δηλαδή να διερευνηθεί το κατά πόσο τα ενσωματώματα που παράγονται, για την ίδια πρωτεΐνη, από τις διαφορετικές μεθόδους διανυσματοποίησης, τοποθετούνται στην ίδια συστάδα από τον αλγόριθμο ομαδοποίησης κ-μέσων, ώστε να απαντηθεί το ερώτημα: *Πόσο παρόμοιες είναι οι ομαδοποιήσεις που παράγονται από τις μεθόδους word2vec, doc2vec, bioword2vec, biosent2vec, biodoc2vec σε σύγκριση με την ομαδοποίηση που παράγεται από την τεχνική node2vec;*

Για το σκοπό αυτό χρησιμοποιήθηκε ο προσαρμοσμένος δείκτης Rand (Adjusted Rand Index - ARI) (Fisher & Hoffman, 1988; Hubert & Arabie, 1985; Vinh et al., 2010; Warrens & Van Der Hoef, 2022). Η μετρική ARI αποτελεί την διορθωμένη, όσον αφορά την τυχαιότητα, (corrected-for-chance), έκδοση της μετρικής Rand Index (R. Liu et al., 2018). Το ARI εισάγει ένα μέτρο αντικειμενικότητας στη σύγκριση ομαδοποιήσεων /συσταδοποιήσεων (Yeung & Ruzzo, 2001). Η αρχή λειτουργίας του παρουσιάζεται στη συνέχεια:

Δεδομένου ενός συνόλου $S = \{O_1, O_2, \dots, O_n\}$, έστω ότι τα $U = \{u_1, u_2, \dots, u_R\}$ και $V = \{v_1, v_2, \dots, v_c\}$ σύνολα αναπαριστούν δύο διαφορετικές ομάδες αντικειμένων στο S έτσι ώστε $\bigcup_{i=1}^{R} u_i = S = \bigcup_{j=1}^{C} v_j$ και $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ για $1 \leq i \neq i' \leq R$ και $1 \leq j \neq j' \leq C$. Έσω ότι το U είναι ένα εξωτερικό κριτήριο του οποίου η αλήθεια είναι αδιαμφισβήτητη, και το V είναι το αποτέλεσμα μιας ομαδοποίησης/συσταδοποίησης (clustering result). Έστω n_{ij} το πλήθος των αντικειμένων που ανήκουν και στην κλάση u_i και τοποθετήθηκαν στην ομάδα v_j. Έστω n_i και n_j τα πλήθη των αντικειμένων στην κλάση u_i και στην ομάδα v_j αντίστοιχα. Η αλληλοεπικάλυψη μεταξύ κλάσεων και ομάδων συνοψίζεται σε έναν $R \times C$ πίνακα ενδεχομένων $M = \left[n_{ij}\right]_{j=1\dots C}^{i=1\dots R}$ όπως παρουσιάζεται στον Πίνακα 1, όπου το n_{ij} υποδηλώνει το πλήθος των αντικειμένων που είναι κοινά στην κλάση U_i και την ομάδα V_j.

Πίνακας 1. Ο πίνακας ενδεχομένων $n_{ij} = |U_i \cap V_j|$

Κλάση\Ομάδα	v_1	v_2	…	v_C	*Άθροισμα*
U_1	n_{11}	n_{12}	…	n_{1C}	$n_{1.}$

U_2	n_{21}	n_{22}	…	n_{2C}	$n_{2.}$
: .	: .	: .		⋱	: .
U_R	n_{R1}	n_{R2}	…	n_{RC}	$n_{R.}$
Άθροισμα	$n_{.1}$	$n_{.2}$	…	$n_{.C}$	$\sum_{ij} n_{ij} = N$

Οι μετρικές που βασίζονται στην καταμέτρηση ζευγών στηρίζονται στην καταμέτρηση ζευγών αντικειμένων για τα οποία δύο ομαδοποιήσεις συμφωνούν ή διαφωνούν. Συγκεκριμένα, τα $\binom{N}{2}$ ζεύγη στοιχείων στο S μπορεί να ταξινομηθούν σε έναν από τους 4 τύπους:

- N_{11} το πλήθος των ζευγών που ανήκουν στην ίδια κλάση U και τοποθετήθηκαν στην ίδια ομάδα V.
- N_{00} το πλήθος των ζευγών που ανήκουν σε διαφορετικές κλάσεις U και τοποθετήθηκαν σε διαφορετικές ομάδες V.
- N_{01} το πλήθος των ζευγών που ανήκουν σε διαφορετικές κλάσεις U αλλά τοποθετήθηκαν στην ίδια ομάδα V.
- N_{10} το πλήθος των ζευγών που ανήκουν στην ίδια κλάση U και τοποθετήθηκαν σε διαφορετικές ομάδες V.

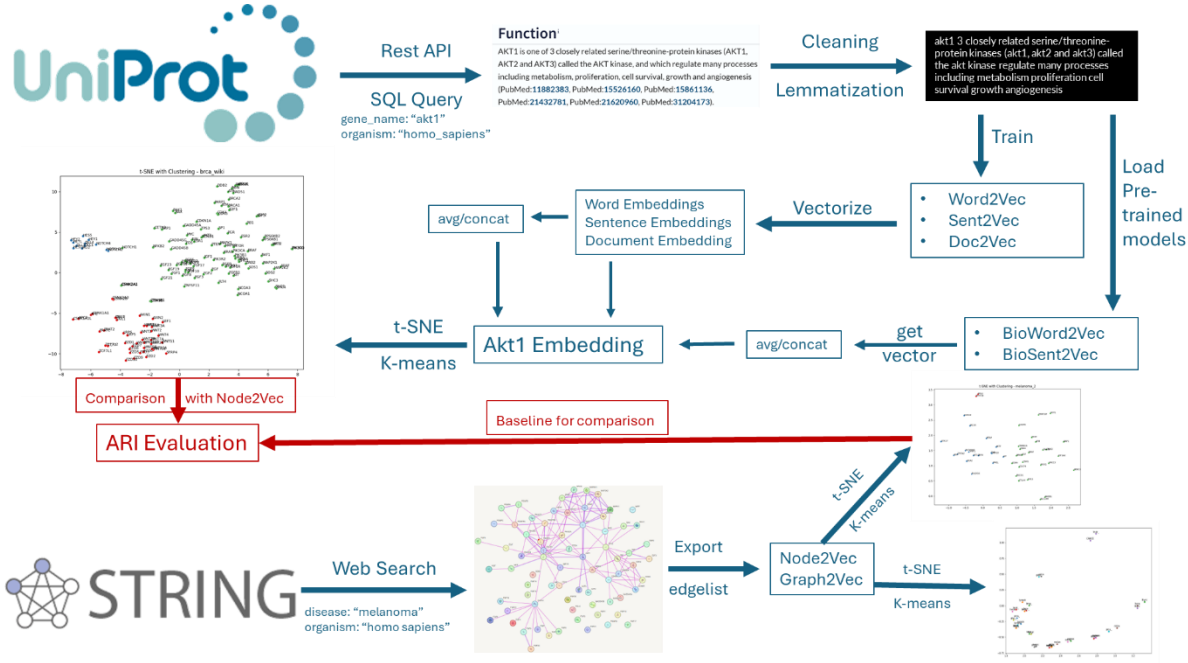
Ο δείκτης Rand (Rand Index, (Rand, 1971)) υπολογίζεται ως $\mathrm{RI}(U,V) = (N_{00} + N_{11})/\binom{N}{2}$ και παίρνει τιμές μεταξύ 0 και 1. Στην πράξη, ωστόσο, ο RI συχνά παίρνει τιμές στο στενότερο εύρος [0.5,1]. Επίσης, η βασική του τιμή μπορεί να είναι υψηλή και να μην παίρνει σταθερή τιμή. Για τους λόγους αυτούς, ο δείκτης Rand χρησιμοποιείται κυρίως στην προσαρμοσμένη μορφή του, γνωστή ως προσαρμοσμένος δείκτης Rand (Adjuster Rand Index - ARI, (Hubert & Arabie, 1985)), η οποία υπολογίζεται ως εξής:

$$ARI(U,V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}$$

Ο ARI δεν μπορεί να πάρει τιμές μεγαλύτερες του 1, και ισούται με 0 όταν ο RI ισούται με την αναμενόμενη τιμή του (σύμφωνα με την υπόθεση της γενικευμένης υπεργεωμετρικής κατανομής για την τυχαιότητα) (Steinley, 2004; Vinh et al., 2010).

Επομένως, στην συγκεκριμένη ερευνητική εργασία, η ομαδοποίηση K-μέσων των ενσωματωμάτων που παρήχθησαν από το Node2Vec χρησιμοποιήθηκαν ως κριτήριο

αληθείας, και οι ομαδοποιήσεις των υπόλοιπων ενσωματωμάτων συγκρίθηκαν με αυτή, για κάθε δίκτυο πρωτεϊνικών αλληλεπιδράσεων, και ο προσαρμοσμένος δείκτης Rand υπολογίστηκε.



Εικόνα 43. Συγκεντρωτικό διάγραμμα της ερευνητικής μεθοδολογίας που ακολουθήθηκε για την παραγωγή και σύγκριση των ενσωματωμάτων που παρήχθησαν με τις διαφορετικές μεθόδους διανυσματοποίησης.

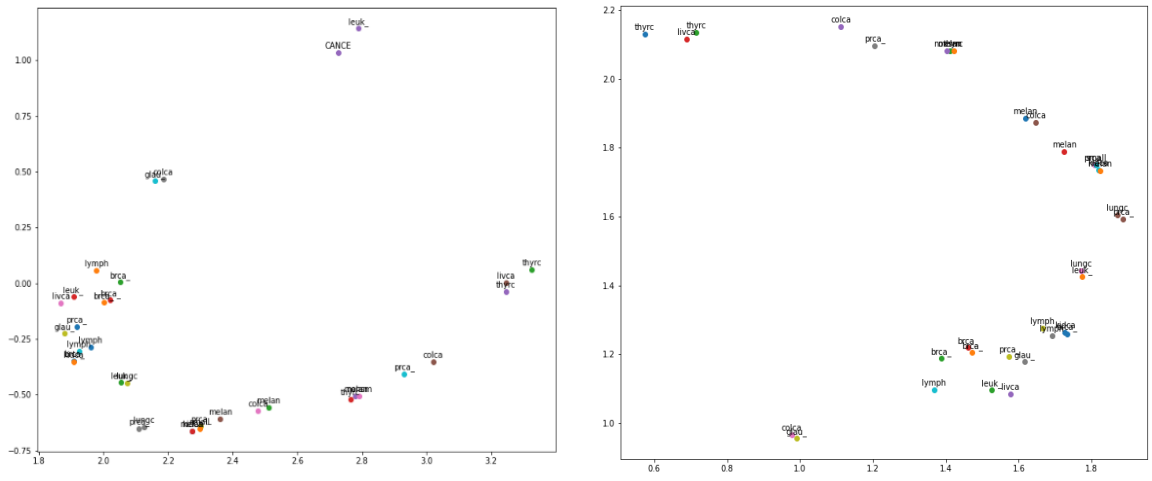
5 Αποτελέσματα

5.1 Διανυσματοποίηση δικτύων πρωτεϊνικών αλληλεπιδράσεων

5.1.1 Graph2Vec

Αρχικά πραγματοποιήθηκε διανυσματοποίηση κάθε συνόλου δεδομένων ως ένα διάνυσμα/ενσωμάτωμα. Τα ενσωματώματα αυτά παρουσιάζονται στην Εικόνα 44. Αρχικά, όπως φαίνεται στο αριστερό τμήμα της εικόνας, υπάρχουν δύο πολύ απομακρυσμένα σημεία, που αντιστοιχούν στα σύνολα δεδομένων για τον καρκίνο, γενικά, και τη λευχαιμία. Πρόκειται για τα δύο μεγαλύτερα, με μεγάλη διαφορά πλήθους κόμβων και συνδέσεων, σύνολα δεδομένων γράφων πρωτεϊνικών αλληλεπιδράσεων που χρησιμοποιήθηκαν.

Δεξιά παρουσιάζονται όλα τα υπόλοιπα ενσωματώματα, πλην των δυο μεγαλύτερων, για να διερευνηθεί αν αυτά τα μεγάλα σύνολα δεδομένων αλλάζουν τη θέση των υπόλοιπων στο επίπεδο. Παρατηρείται ότι, και μετά την αφαίρεσή τους, οι θέσεις των εναπομεινάντων συνόλων δεδομένων δεν επηρεάζονται σημαντικά. Περιστρέφοντας τη δεξιά γραφική παράσταση 180^{o} προς τα δεξιά, τα σημεία που αντιστοιχούν στα ενσωματώματα των γράφων συμπίπτουν, σχεδόν, με τα αντίστοιχα ενσωματώματα της αριστερής γραφικής παράστασης, που εμπεριέχουν και τα δύο πολύ μεγάλα σύνολα δεδομένων.



Εικόνα 44. Αριστερά: Αναπαράσταση στο επίπεδο με t-SNE όλων των συνόλων δεδομένων γράφων πρωτεϊνικών αλληλεπιδράσεων που ανακτήθηκαν από τη String. Δεξιά: t-SNE όλων πλην δύο συνόλων δεδομένων (Cancer – γενικό για τον καρκίνο και leukemia – λευχαιμία).

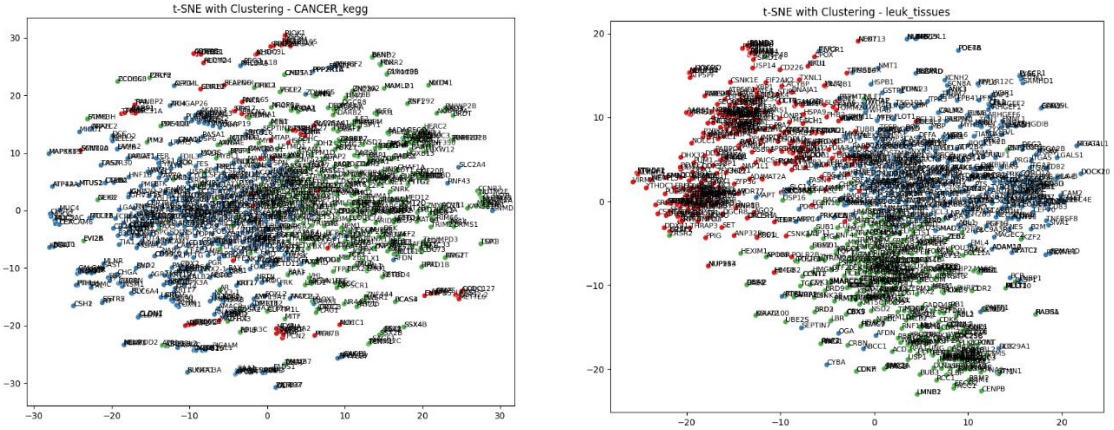
5.1.2 Node2Vec

Όπως αναφέρθηκε στην προηγούμενη ενότητα, το μεγαλύτερο σύνολο δεδομένων είναι ένα συγκεντρωτικό σύνολο δεδομένων γνωστών πρωτεϊνικών αλληλεπιδράσεων στον καρκίνο.

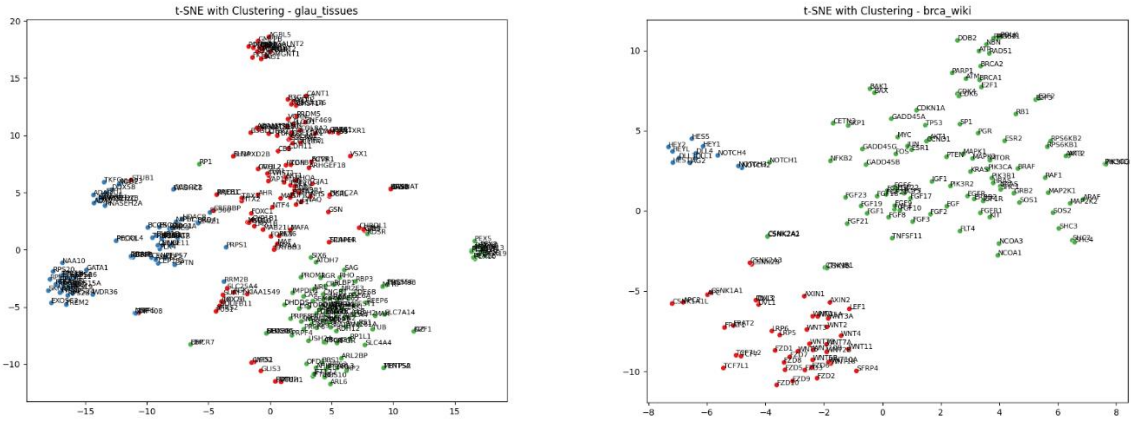
Περιλαμβάνει επομένως διάφορες μορφές καρκίνου και παρουσιάζει όλες τις πρωτεϊνικές αλληλεπιδράσεις που έχουν παρατηρηθεί σε αυτούς. Περιλαμβάνει μεγάλο πλήθος κόμβων-πρωτεϊνών, για κάθε έναν από τους οποίους δημιουργήθηκε ένα ενσωμάτωμα, με την τεχνική Node2Vec. Τα ενσωματώματα, μετά από μείωσης διάστασης με t-SNE, παρουσιάζονται στην Εικόνα 45.

Παρατηρείται μεγαλύτερη συγκέντρωση σημείων, που αντιστοιχούν σε πιο διασυνδεδεμένα ενσωματώματα, στο κέντρο της γραφικής παράστασης, με την πυκνότητα των ενσωματωμάτων να μειώνεται στην περιφέρεια. Η συγκεκριμένη γραφική παράσταση παρουσιάζεται ενδεικτικά, καθώς, λόγω του μεγάλου όγκου του συνόλου δεδομένων από το οποίο προέρχεται, και τη δυσκολία που αντιμετώπιζαν τα μοντέλα μηχανικής μάθησης να το διαχειριστούν, λόγω περιορισμένης διαθέσιμης μνήμης, δεν χρησιμοποιήθηκε στις περαιτέρω αναλύσεις που πραγματοποιήθηκαν.

Στην Εικόνα 46 παρουσιάζονται γραφικά τα ενσωματώματα δύο άλλων μορφών καρκίνου, του γλαυκώματος και του καρκίνου του μαστού. Το πλήθος των κόμβων είναι εμφανώς μικρότερο, γεγονός που καθιστά τα αντίστοιχα σύνολα δεδομένων πιο εύκολα διαχειρίσιμα. Παρατηρείται, επίσης, η ύπαρξη ορισμένων συστάδων (clusters), που είναι εμφανείς οπτικά, και συμβαδίζουν με τα χρώματα των σημείων, που αντιστοιχούν στις συστάδες που προέκυψαν από τον αλγόριθμο ομαδοποίησης κ-μέσων.

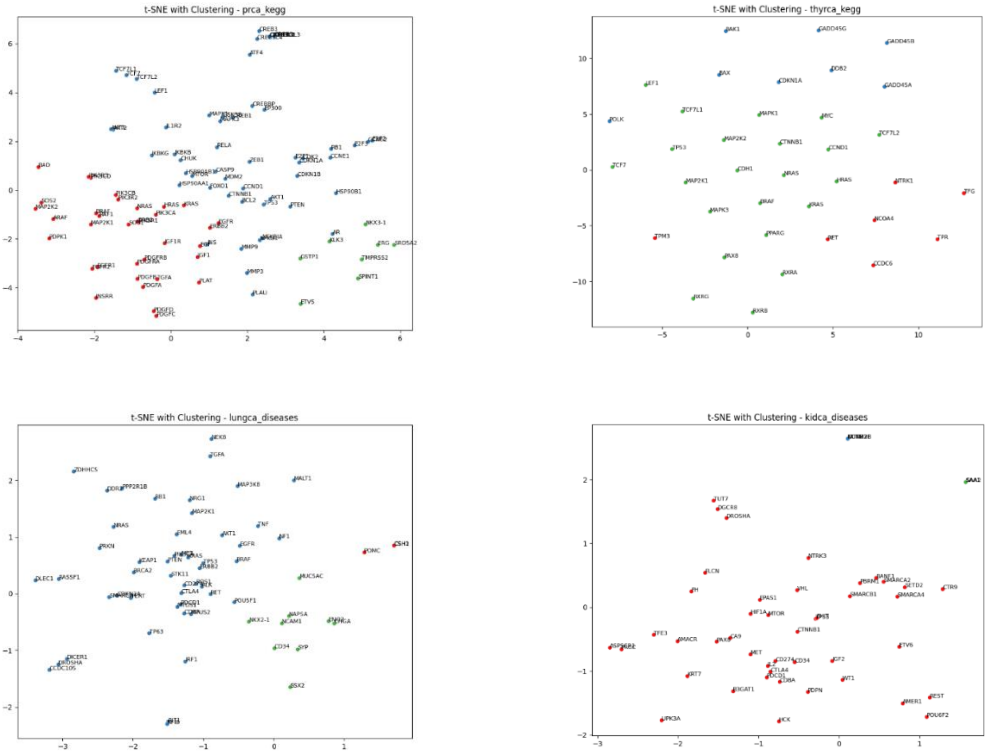


Εικόνα 45. Γραφική αναπαράσταση των ενσωματωμάτων που προέκυψαν από Node2Vec στα σύνολο δεδομένων για τον καρκίνο γενικά, και τη λευχαιμία (από τη βάση δεδομένων tissues), και στη συνέχεια t-SNE για μείωση σε 2 διαστάσεις.



Εικόνα 46. Γραφικές παραστάσεις των ενσωματωμάτων που προέκυψαν από Node2Vec και στη συνέχεια t-SNE για μείωση σε 2 διαστάσεις, από τα σύνολα δεδομένων για το γλαύκωμα (αριστερά) και τον καρκίνο του μαστού (δεξιά).

Τα αποτελέσματα της Εικόνα 46 αποτελούν ένα καλό παράδειγμα οπτικής αναπαράστασης ενσωματωμάτων, όπου η δημιουργία ομάδων είναι εμφανής. Ωστόσο, για κάποια άλλα σύνολα δεδομένων δικτύων πρωτεϊνικών αλληλεπιδράσεων, πιθανώς λόγω μικρού μεγέθους, τα ενσωματώματα δεν ομαδοποιούνται τόσο καλά. Τέτοια παραδείγματα παρουσιάζονται στην Εικόνα 47.

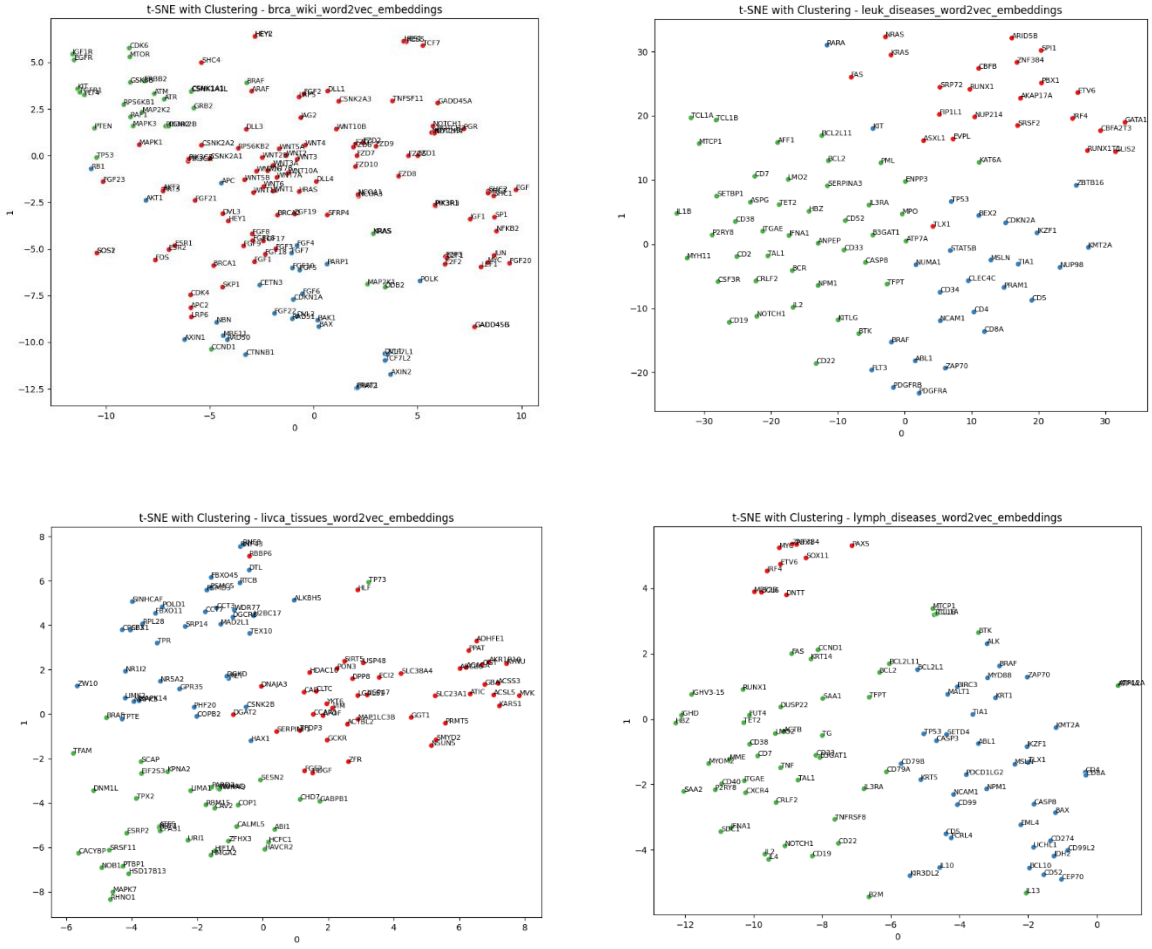


Εικόνα 47. Γραφικές παραστάσεις ενσωματωμάτων που έχουν παραχθεί με τη μέθοδο Node2Vec, για τον καρκίνο του προστάτη (πάνω αριστερά), του θυρεοειδούς (πάνω δεξιά), του πνεύμονα (κάτω αριστερά) και του νεφρού (κάτω δεξιά).

5.2 Διανυσματοποίηση κειμένου λειτουργικότητας πρωτεϊνών

5.2.1 Word2Vec

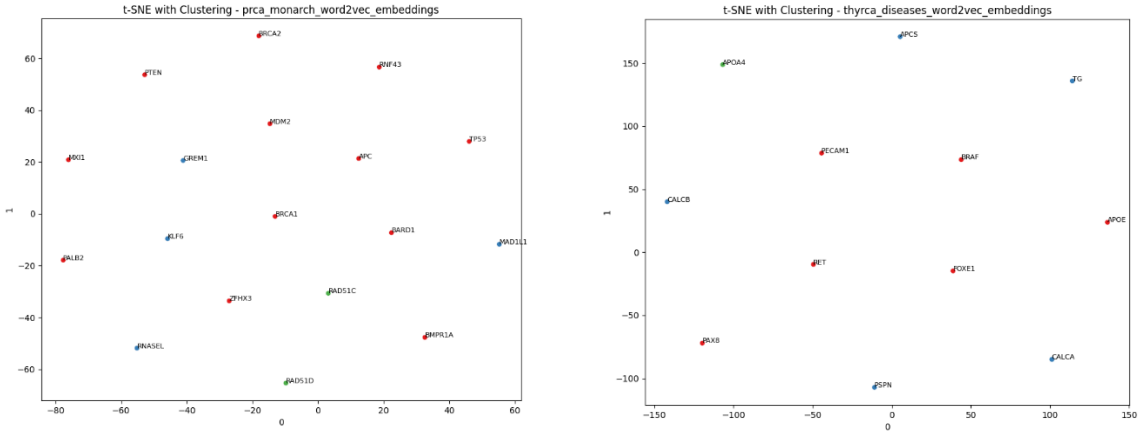
Στη συνέχεια, τα ονόματα των πρωτεϊνών κάθε δικτύου πρωτεϊνικής αλληλεπίδρασης χρησιμοποιήθηκαν για την εκπαίδευση ενός Word2Vec μοντέλου, για κάθε δίκτυο. Τα ενσωματώματα των ονομάτων των πρωτεϊνών παρουσιάζονται ενδεικτικά, για 4 από τα σύνολα δεδομένων, στην Εικόνα 48.



Εικόνα 48. Γραφική αναπαράσταση, μετά από μείωση διαστάσεων με t-SNE, και ομαδοποίηση κ-μέσων, των ενσωματωμάτων που προέκυψαν από τα ονόματα των πρωτεϊνών, με τη μέθοδο Word2Vec για τον καρκίνο του μαστού (πάνω αριστερά), τη λευχαιμία (πάνω δεξιά), τον καρκίνο του ήπατος (κάτω αριστερά) και το λέμφωμα (κάτω δεξιά).

Η Εικόνα 48 αντιστοιχεί σε σύνολα δεδομένων για τα οποία παρατηρήθηκε κάποια ομαδοποίηση και φαίνεται να σχηματίζονται ορισμένες συστάδες πρωτεϊνών. Ωστόσο, για άλλα σύνολα δεδομένων, όπως αυτά που παρουσιάζονται στην Εικόνα 49, τα σημεία που

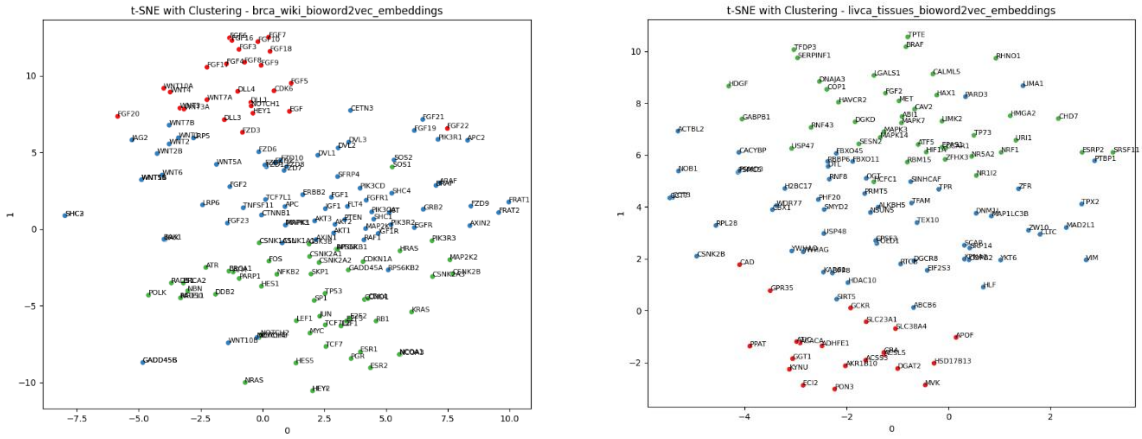
αντιστοιχούν στα επιμέρους ενσωματώματα πρωτεϊνών παρουσιάζονται να ισαπέχουν μεταξύ τους στο επίπεδο, και δεν παρατηρείται κάποια ιδιαίτερη συσταδοποίηση.



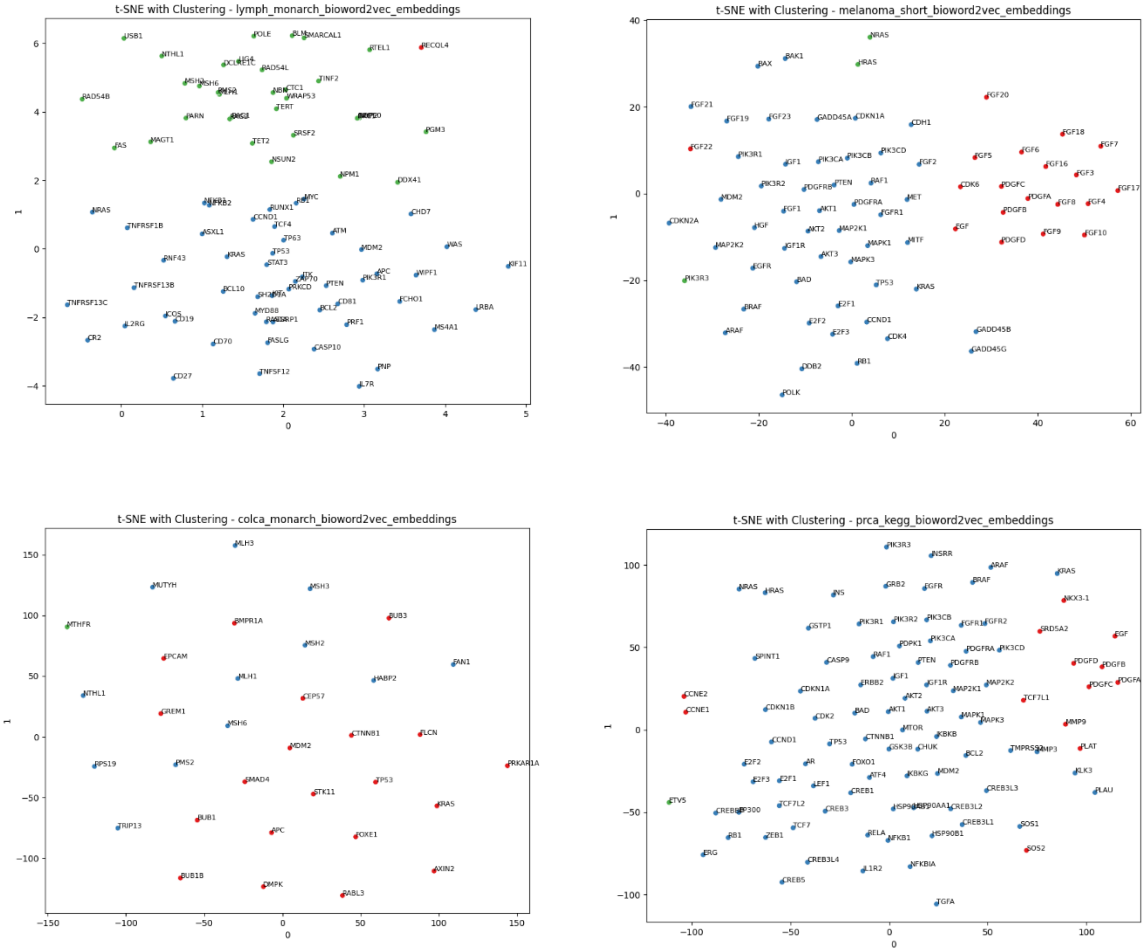
Εικόνα 49. Γραφική αναπαράσταση, μετά από μείωση διαστάσεων με t-SNE, και ομαδοποίηση κ-μέσων, των ενσωματωμάτων που προέκυψαν από τα ονόματα των πρωτεϊνών, με τη μέθοδο Word2Vec για τον καρκίνο του προστάτη (αριστερά) και του θυρεοειδούς (δεξιά).

5.2.2 BioWord2Vec

Στη συνέχεια διερευνήθηκε η χρήση του μοντέλου προεκπαιδευμένων διανυσμάτων BioWordVec για την παραγωγή ενσωματωμάτων για τα ονόματα των πρωτεϊνών που συμμετέχουν στα δίκτυα πρωτεϊνικών αλληλεπιδράσεων των διαφόρων μορφών καρκίνου που μελετήθηκαν. Σε αυτή τη μέθοδο, το μοντέλο δεν χρειάστηκε εκπαίδευση, καθώς ήταν ήδη προεκπαιδευμένο. Τα αποτελέσματα οπτικοποίησης των ενσωματωμάτων σε δύο διαστάσεις, και ομαδοποίησης κ-μέσων, για κάποια από τα διαθέσιμα σύνολα δεδομένων, παρουσιάζονται ενδεικτικά στην Εικόνα 50 και στην Εικόνα 51.



Εικόνα 50. Γραφική αναπαράσταση, μετά από μείωση διαστάσεων με t-SNE, και ομαδοποίηση κ-μέσων, των ενσωματωμάτων που προέκυψαν από τα ονόματα των πρωτεϊνών, με τη μέθοδο BioWord2Vec για τον καρκίνο του μαστού (αριστερά) και του ήπατος (δεξιά).

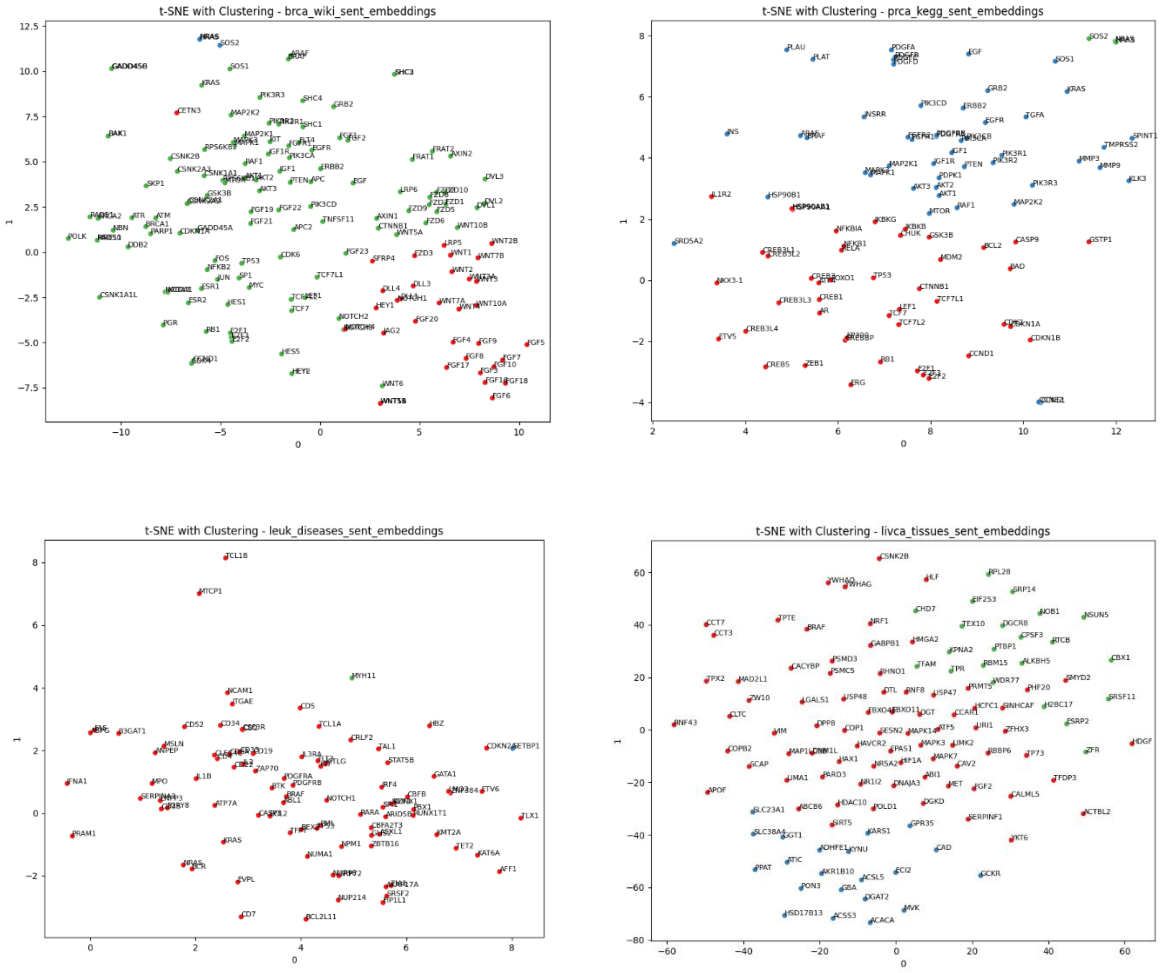


Εικόνα 51. Γραφική αναπαράσταση, μετά από μείωση διαστάσεων με t-SNE, και ομαδοποίηση κ-μέσων, των ενσωματωμάτων που προέκυψαν από τα ονόματα των πρωτεϊνών, με τη μέθοδο BioWord2Vec για το λέμφωμα (πάνω αριστερά), το μελάνωμα(πάνω δεξιά), τον καρκίνο του παχέος εντέρου (κάτω αριστερά) και του προστάτη(κάτω δεξιά).

Στη χρήση της συγκεκριμένης μεθόδου παρουσιάστηκε μια τεχνική δυσκολία, που οφείλεται στη χρήση έτοιμου προεκπαιδευμένου μοντέλου. Όπως αναφέρθηκε στην ενότητα της Μεθοδολογίας, το BioWordVec μοντέλο προεκπαιδευμένων διανυσμάτων που χρησιμοποιήθηκε, είχε εκπαιδευθεί σε δεδομένα της PubMed και του MeSH, και παρείχε έτοιμα διανύσματα για λέξεις που γνώριζε. Ωστόσο, για αρκετές πρωτεΐνες, το μοντέλο δεν γνώριζε κάτι. Δεν υπήρχαν δηλαδή προεκπαιδευμένα διανύσματα για τα ονόματα αρκετών πρωτεϊνών που συμμετέχουν στα δίκτυα πρωτεϊνικών αλληλεπιδράσεων που αναλύθηκαν. Αυτό είχε ως αποτέλεσμα να παράγονται ελλιπή αποτελέσματα σε αρκετές περιπτώσεις.

5.2.3 BioSent2Vec

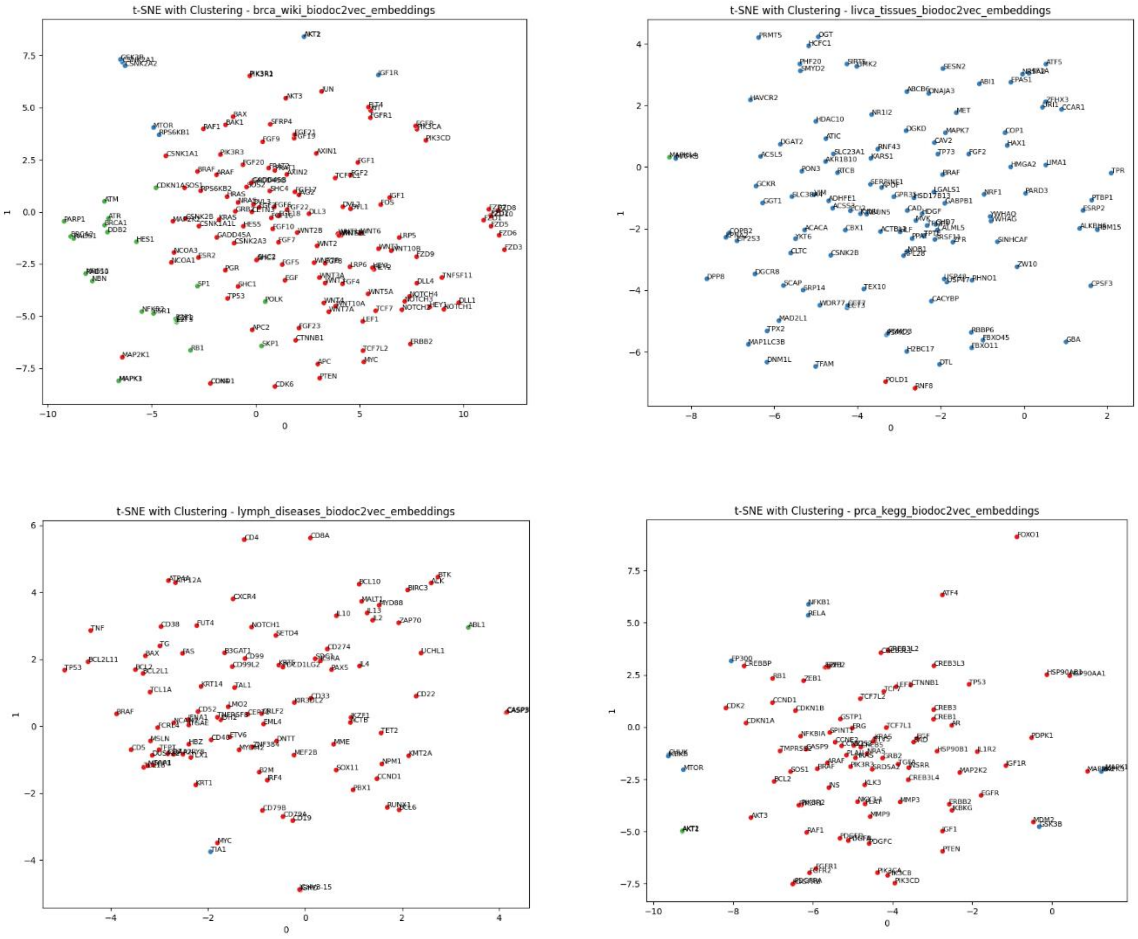
Για την εφαρμογή αυτής της μεθόδου χρησιμοποιήθηκε το προεκπαιδευμένο BioSentVec μοντέλο, επίσης εκπαιδευμένο σε δεδομένα PubMed και MeSH. Εδώ δεν αντιμετωπίστηκαν προβλήματα ελλείψεων του μοντέλου, όπως στην περίπτωση της BioWord2Vec μεθόδου, καθώς το κείμενο περιγραφής της λειτουργικότητας κάθε πρωτεΐνης χρησιμοποιήθηκε ως είσοδος στο μοντέλο για την παραγωγή κάθε αντίστοιχου ενσωματώματος. Επομένως, ακόμη και αν το μοντέλο δεν κατείχε γνώση για το όνομα μιας πρωτεΐνης, ήταν σε θέση να κατανοήσει και να παραστήσει διανυσματικά το νόημα της περιγραφής της λειτουργικότητάς της.



Εικόνα 52. Γραφική αναπαράσταση, μετά από μείωση διαστάσεων με t-SNE, και ομαδοποίηση κ-μέσων, των ενσωματωμάτων που προέκυψαν από τα ονόματα των πρωτεϊνών, με τη μέθοδο BioSent2Vec για τον καρκίνο του μαστού (πάνω αριστερά), του προστάτη (πάνω δεξιά), τη λευχαιμία (κάτω αριστερά) και τον καρκίνο του ήπατος (κάτω δεξιά).

Ενδεικτικά ορισμένα αποτελέσματα οπτικοποίησης των ενσωματωμάτων που παρήχθησαν με τη μέθοδο BioSent2Vec, μετά από μείωση διάστασης, παρουσιάζονται στην Εικόνα 52. Όπως θα αναλυθεί και στην ενότητα της σύγκρισης των αποτελεσμάτων, η μέθοδος BioSent2Vec παρήγαγε τα καλύτερα αποτελέσματα, με ουσιαστικές ομαδοποιήσεις, αρκετά κοντά στις ομαδοποιήσεις των ενσωματωμάτων της Node2Vec μεθόδου.

5.2.4 BioDoc2Vec



Εικόνα 53. Γραφική αναπαράσταση, μετά από μείωση διαστάσεων με t-SNE, και ομαδοποίηση κ-μέσων, των ενσωματωμάτων που προέκυψαν από τα ονόματα των πρωτεϊνών, με τη μέθοδο BioDoc2Vec για τον καρκίνο του μαστού (πάνω αριστερά), του ήπατος (πάνω δεξιά), το λέμφωμα (κάτω αριστερά) και τον καρκίνο του προστάτη (κάτω δεξιά).

Τέλος, εφαρμόστηκε η μέθοδος BioDoc2Vec για τη διανυσματοποίηση του κειμένου περιγραφής της λειτουργίας των πρωτεϊνών των δικτύων πρωτεϊνικών αλληλεπιδράσεων. Η υλοποίησή της στηρίχθηκε στο BioWordVec μοντέλο προεκπαιδευμένων διανυσμάτων. Το διάνυσμα κάθε λέξης της περιγραφής της λειτουργίας κάθε πρωτεΐνης ανακτήθηκε από το

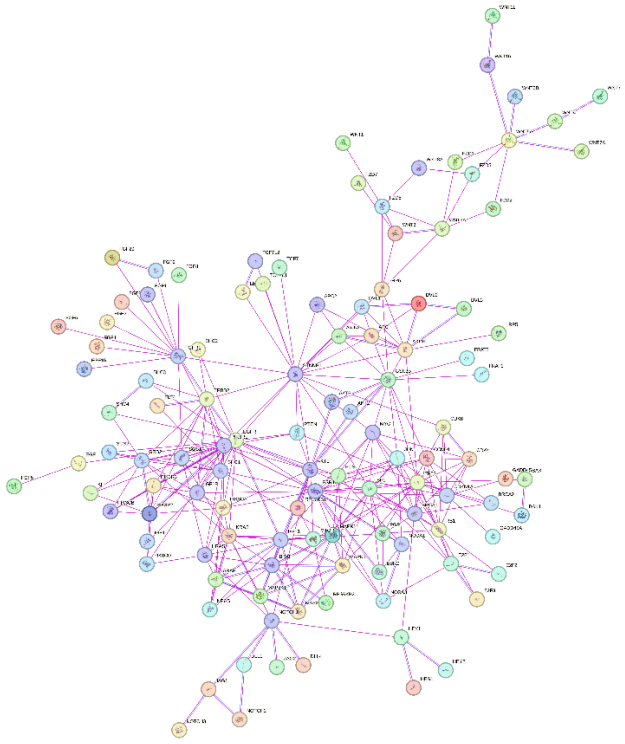
μοντέλο, και υπολογίζοντας το μέσο όρο των επιμέρους διανυσμάτων, προέκυψε ένα τελικό διάνυσμα για όλο το κείμενο της περιγραφής κάθε πρωτεΐνης.

Οι ελλείψεις διανυσμάτων για ορισμένες λέξεις, του BioWordVec μοντέλου, δεν αποτέλεσαν πρόβλημα, καθώς σίγουρα για κάθε κείμενο περιγραφής υπήρχαν λέξεις που γνώριζε το μοντέλο. Ο υπολογισμός του συνολικού διανύσματος έγινε με εξαγωγή μέσου όρου των επιμέρους διανυσμάτων μόνο των λέξεων που εντοπίστηκαν.

Ενδεικτικά ορισμένα αποτελέσματα της εφαρμογής της BioWord2Vec μεθόδου παρουσιάζονται στην Εικόνα 53.

5.3 Σύγκριση μεθόδων διανυσματοποίησης

Στη συνέχεια παρουσιάζονται ορισμένα συγκριτικά αποτελέσματα όλων των μεθόδων διανυσματοποίησης που εφαρμόστηκαν για ένα μόνο σύνολο δεδομένων κάθε φορά. Η Εικόνα 56 παρουσιάζει αυτή τη σύγκριση για το σύνολο δεδομένων του μικροκυτταρικού καρκίνου του πνεύμονα, το δίκτυο πρωτεϊνικής αλληλεπίδρασης του οποίου παρουσιάζεται στην Εικόνα 54.



Εικόνα 54. Δίκτυο πρωτεϊνικών αλληλεπιδράσεων στον μικροκυτταρικό καρκίνο του πνεύμονα, όπως ανακτήθηκε από τη βάση δεδομένων String.

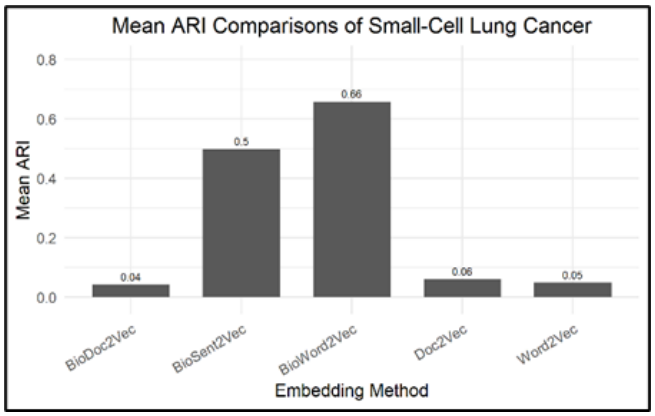
Από την πρώτη γραφική παράσταση, που αντιστοιχεί στη Node2Vec μέθοδο, είναι εμφανές ότι υπάρχουν διακριτές ομάδες πρωτεϊνών στο συγκεκριμένο δίκτυο. Οι ομάδες αυτές δεν είναι ιδιαίτερα εμφανείς στην Εικόνα 54, που παρουσιάζει το δίκτυο σε μορφή γράφου, όπως παρέχεται στην ιστοσελίδα της βάσης δεδομένων String. Αυτό ενδεχομένως να οφείλεται στον τρόπο δημιουργίας των τυχαίων διαδρομών του Node2Vec αλγορίθμου, και τα κριτήρια δημιουργίας των ενσωματωμάτων που παράγει.

Η δεύτερη γραφική παράσταση απεικονίζει τα ενσωματώματα των ονομάτων των πρωτεϊνών, με τη μέθοδο Word2Vec. Εδώ παρατηρείται η ύπαρξη μιας ακραίας τιμής (outlier) που επεκτείνει πολύ το επίπεδο. Ωστόσο, παρατηρείται μια χωρική διαίρεση των υπόλοιπων ενσωματωμάτων, που να συμβαδίζει με τις ομάδες που δημιουργήθηκαν από την ομαδοποίηση κ-μέσων.

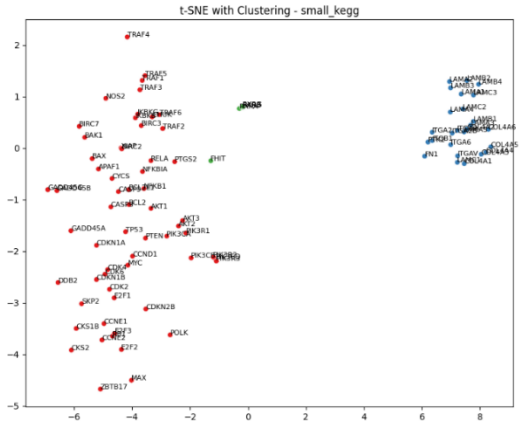
Τα αποτελέσματα της διανυσματοποίησης με την τεχνική Doc2Vec, που παρουσιάζονται στην τρίτη εικόνα, δείχνουν μια συγκέντρωση των ενσωματωμάτων ορισμένων πρωτεϊνών στο κέντρο του επιπέδου, με αραιότερη κατανομή στην περιφέρεια, και κάποιες απομονωμένες πρωτεΐνες.

Η τέταρτη και η πέμπτη γραφική παράσταση απεικονίζουν στο επίπεδο τα ενσωματώματα που παρήχθησαν με τη χρήση των προεκπαιδευμένων μοντέλων BioWordVec και BioSentVec. Μια πρώτη παρατήρηση αφορά την ομοιόμορφη κατανομή και τις μεγαλύτερες αποστάσεις μεταξύ των ενσωματωμάτων, σε σύγκριση με τις υπόλοιπες μεθόδους. Επίσης, παρατηρείται διαχωρισμός των ενσωματωμάτων στο επίπεδο, που να συμβαδίζει σε μεγάλο βαθμό με τις ομάδες που δημιούργησε ο αλγόριθμος ομαδοποίησης κ-μέσων.

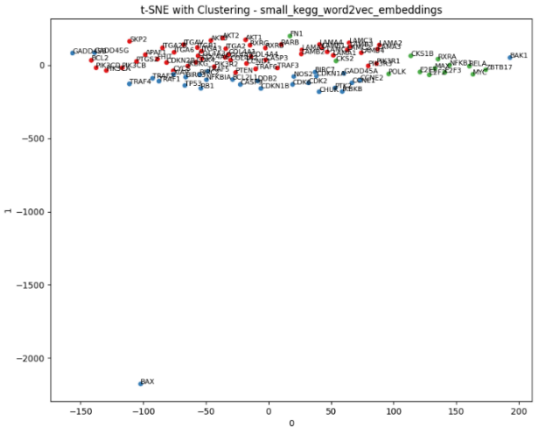
Τέλος, η έκτη γραφική παράσταση, που προέκυψε με τη χρήση της μεθόδου BioDoc2Vec, που στηρίζεται στα BioWordVec ενσωματώματα κάθε λέξης του κειμένου περιγραφής της λειτουργίας κάθε πρωτεΐνης, παρουσιάζει εικόνα που ομοιάζει με τη Doc2Vec μέθοδο. Πιθανώς λόγω χρήσης της ίδιας μεθοδολογίας υπολογισμού μέσου όρου των ενσωματωμάτων, που εδώ όμως προέρχονται από το προεκπαιδευμένο BioDocVec μοντέλο.



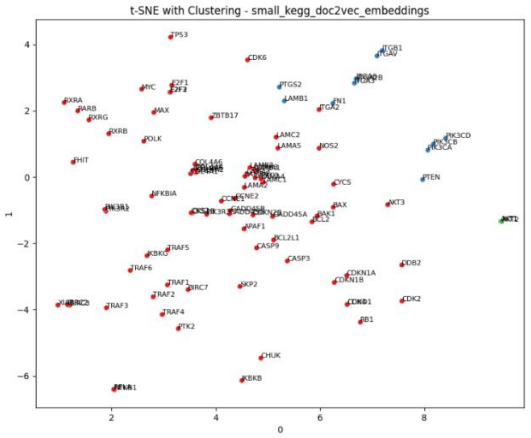
Εικόνα 55. Προσαρμοσμένος δείκτης Rand για την αξιολόγηση της απόδοσης των μεθόδων διανυσματοποίησης στο σύνολο δεδομένων για τον μικροκυτταρικό καρκίνο του πνεύμονα.



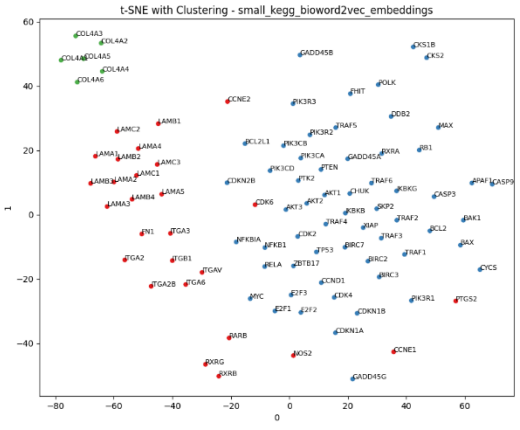
1. Node2Vec



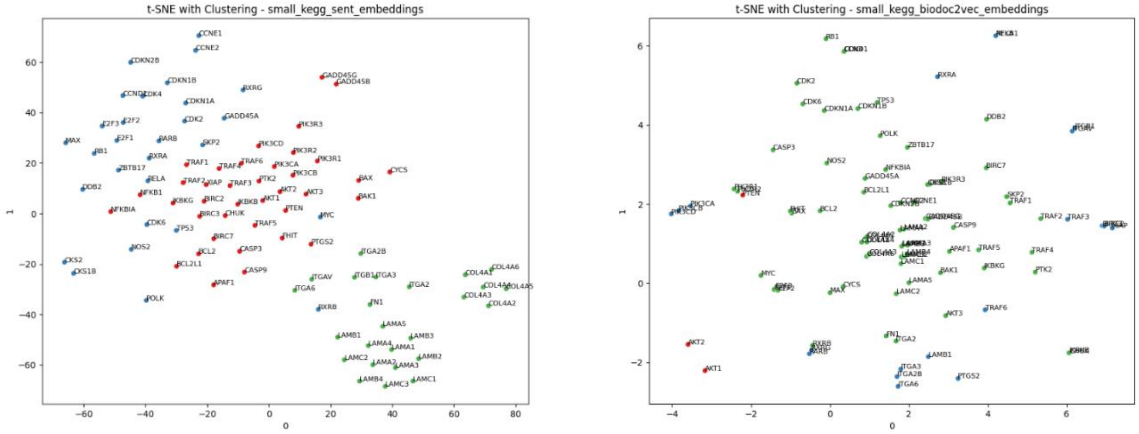
2. Word2Vec



3. Doc2Vec



4. BioWord2Vec



5. BioSent2Vec 6. BioDoc2Vec

Εικόνα 56. Γραφική αναπαράσταση των ενσωματωμάτων που παρήχθησαν με όλες τις μεθόδους που εξετάστηκαν, για το δίκτυο πρωτεϊνικής αλληλεπίδρασης στον μικροκυτταρικό καρκίνο του πνεύμονα.

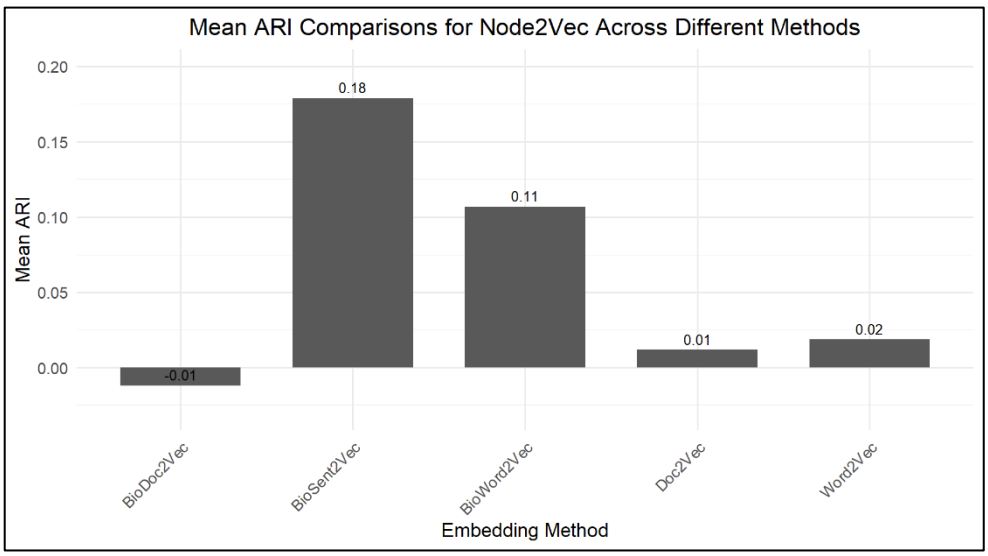
Στο ραβδόγραμμα της Εικόνα 55 φαίνεται ο προσαρμοσμένος δείκτης Rand που χρησιμοποιήθηκε για την αξιολόγηση των επιμέρους μεθόδων διανυσματοποίησης που υλοποιήθηκαν στα πλαίσια της συγκεκριμένης ερευνητικής εργασίας. Παρατηρείται πως, για το συγκεκριμένο, τουλάχιστον, σύνολο δεδομένων, η BioWord2Vec μέθοδος διανυσματοποίησης των κειμένων περιγραφής της λειτουργικότητας των πρωτεϊνών του δικτύου, έδωσε τα καλύτερα αποτελέσματα ομαδοποίησης. Δηλαδή, η ομαδοποίηση κ-μέσων στα ενσωματώματα που παρήχθησαν από το Node2Vec είναι πιο κοντά στην ομαδοποίηση κ-μέσων των ενσωματωμάτων της BioWord2Vec μεθόδου διανυσματοποίησης κειμένου.

5.4 Αξιολόγηση Αποτελεσμάτων

Σε αυτό το κεφάλαιο παρουσιάζεται μια ολική αξιολόγηση των αποτελεσμάτων που παρήχθησαν στα πλαίσια της συγκεκριμένης ερευνητικής εργασίας. Συγκρίνεται, δηλαδή, το πόσο προσεγγίζουν οι ομαδοποιήσεις των ενσωματωμάτων που δημιουργήθηκαν από τις μεθόδους διανυσματοποίησης κειμένου, την ομαδοποίηση των ενσωματωμάτων του Node2Vec. Η μετρική που χρησιμοποιήθηκε για την ποσοτικοποίηση της αξιολόγησης των αποτελεσμάτων είναι το ARI (προσαρμοσμένος δείκτης Rand).

Ο μέσος όρος του ARI όλων των συνόλων δεδομένων υπολογίστηκε για κάθε μέθοδο, και τα αποτελέσματα παρουσιάζονται στην Εικόνα 57. Αρχικά, παρατηρείται ότι οι μέθοδοι BioWord2Vec και BioSent2Vec ξεχωρίζουν, δίνονται σημαντικά μεγαλύτερες τιμές ARI, από τις υπόλοιπες μεθόδους. Επομένως, τα δεδομένα στα οποία έχει εκπαιδευτεί το κάθε

μοντέλο φαίνεται να παίζουν σημαντικό ρόλο στην ποιότητα των παραγόμενων ενσωματωμάτων.



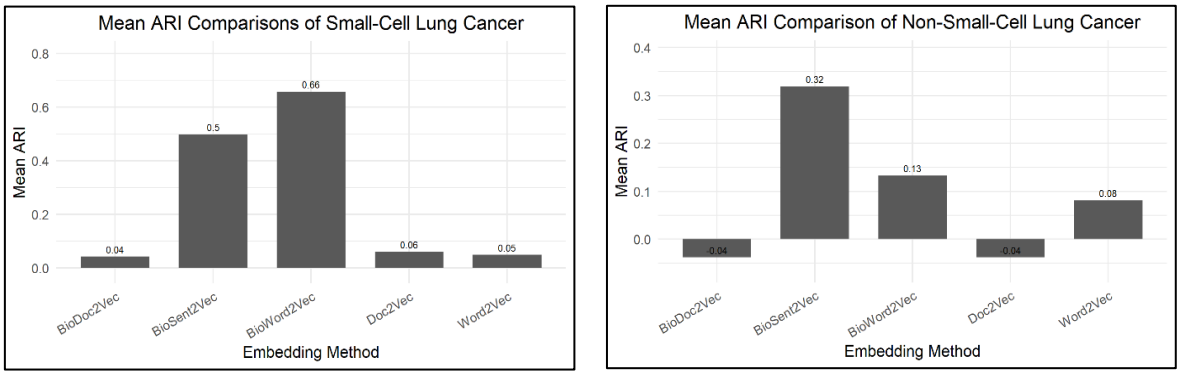
Εικόνα 57. Μέσος όρος ARI όλων των συνόλων δεδομένων, για κάθε μέθοδο διανυσματοποίησης κειμένου.

Ωστόσο, μια ακόμη παρατήρηση στα αποτελέσματα της Εικόνα 57 είναι ότι οι τιμές του ARI είναι πολύ μικρές, για όλες τις μεθόδους. Λαμβάνοντας υπόψιν ότι ο προσαρμοσμένος δείκτης Rand παίρνει την τιμή 1 για απόλυτη συμφωνία μεταξύ των ομαδοποιήσεων, και την τιμή 0 για ομαδοποιήσεις που θα μπορούσαν να προκύψουν τυχαία, λαμβάνοντας δείγματα από μια υπεργεωμετρική κατανομή, είναι σαφές ότι τα συνολικά αποτελέσματα δεν είναι ικανοποιητικά. Τα καλύτερα αποτελέσματα, που προέκυψαν από τις μεθόδους προεκπαιδευμένων μοντέλων σε ιατρικά δεδομένα, δίνουν ARI μεταξύ 0.1 και 0.2, που είναι αρκετά χαμηλές τιμές. Επίσης, αξιοσημείωτο αποτέλεσμα αποτελεί η τιμή ARI της BioDoc2Vec μεθόδου, που είναι αρνητική. Αυτό πρακτικά σημαίνει ότι αν η τοποθέτηση κάθε ενσωματώματος σε ομάδα γινόταν τυχαία, οι ομαδοποιήσεις που θα προέκυπταν θα ήταν πιο κοντά στην ομαδοποίηση αναφοράς, σε σχέση με αυτές που προέκυψαν με τη BioDoc2Vec μέθοδο.

Επίσης, και η τιμή ARI της Doc2Vec μεθόδου είναι πολύ χαμηλή. Αυτό ενδεχομένως να υποδηλώνει ότι, γενικώς, η doc2vec μεθοδολογία δεν είναι κατάλληλη για χρήση σε τέτοιου είδους δεδομένα, ή ότι χρήζει βελτίωσης η προσέγγιση υλοποίησής της.

Ωστόσο, ο μέσος όρος του προσαρμοσμένου δείκτη Rand ίσως να μην είναι ο κατάλληλος τρόπος παρουσίασης της ποιότητας των αποτελεσμάτων, καθώς παρατηρήθηκαν πολύ σημαντικές διακυμάνσεις στην τιμή του μεταξύ των συνόλων δεδομένων. Υπήρξαν δύο

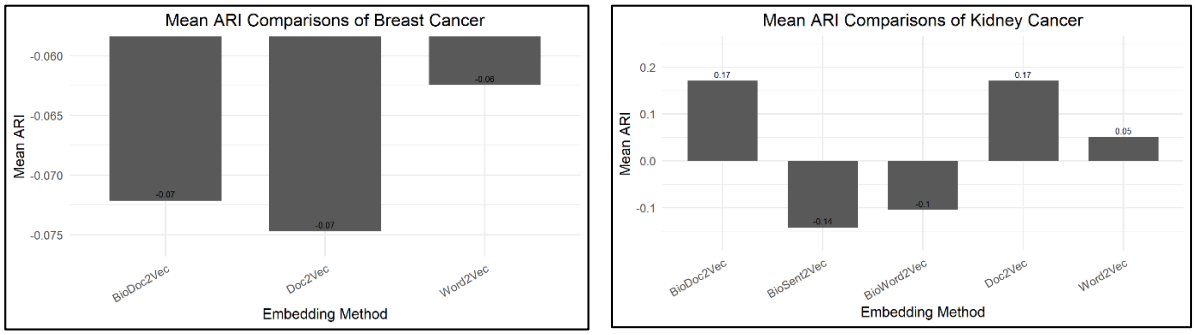
αρκετά καλά δείγματα ομαδοποίησης, από τα σύνολα δεδομένων για τον μικροκυτταρικό και τον μη μικροκυτταρικό καρκίνο του πνεύμονα, τα οποία παρουσιάζονται στην Εικόνα 58.

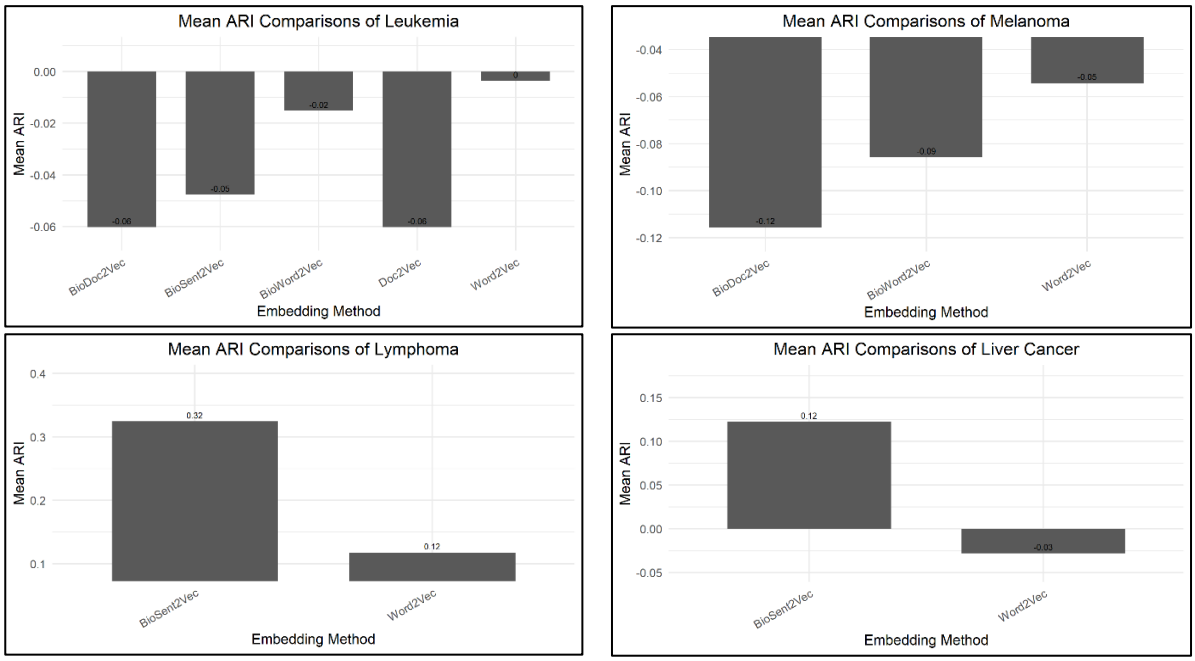


Εικόνα 58. Προσαρμοσμένος δείκτης Rand για την αξιολόγηση των ομαδοποιήσεων κ-μέσων των διάφορων μεθόδων διανυσματοποίησης του κειμένου περιγραφής της λειτουργικότητας των πρωτεϊνών, σε σύγκριση με τις ομαδοποιήσεις του Node2Vec, για τον μικροκυτταρικό και τον μη μικροκυτταρικό καρκίνο του πνεύμονα.

Και για τα δύο σύνολα δεδομένων, είναι εμφανής η καλύτερη απόδοση των προεκπαιδευμένων, σε ιατρικά δεδομένα, μοντέλων, σε σχέση με τα μοντέλα που εκπαιδεύτηκαν μόνο πάνω στο διαθέσιμο σύνολο δεδομένων κάθε φορά. Επίσης, είναι εμφανής χαμηλή απόδοση των doc2vec προσεγγίσεων, και η αδυναμία του να συλλάβουν το συνολικό νόημα των κειμένων με τα οποία τροφοδοτήθηκαν.

Τα αποτελέσματα της Εικόνα 58 αποτελούν τα καλύτερα παραδείγματα, από άποψη απόδοσης, των αποτελεσμάτων της εργασίας. Στην Εικόνα 59 παρουσιάζονται οι προσαρμοσμένοι δείκτες Rand και για τα υπόλοιπα σύνολα δεδομένων, όπου τα αποτελέσματα είναι πολύ χειρότερα.





Εικόνα 59. Προσαρμοσμένος δείκτης Rand για την αξιολόγηση των ομαδοποιήσεων κ-μέσων των διάφορων μεθόδων διανυσματοποίησης του κειμένου περιγραφής της λειτουργικότητας των πρωτεϊνών, σε σύγκριση με τις ομαδοποιήσεις του Node2Vec.

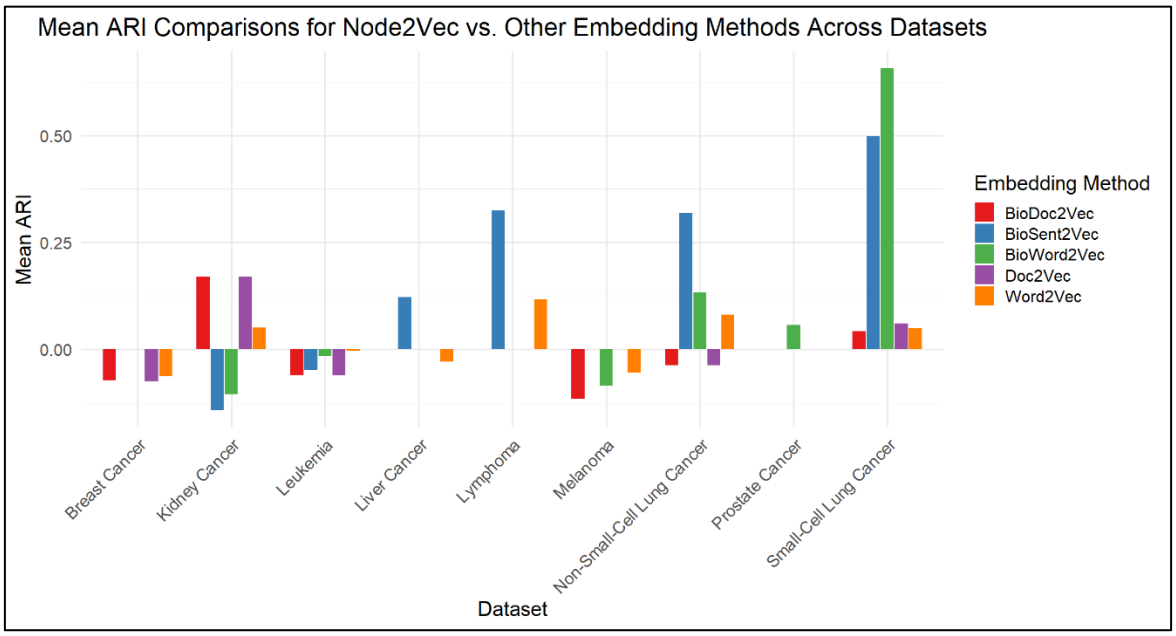
Τα αποτελέσματα της Εικόνα 59 παρουσιάζουν την απόδοση των διάφορων μεθόδων διανυσματικής αναπαράστασης κειμένου, που υλοποιήθηκαν στα πλαίσια της συγκεκριμένης εργασίας. Παρατηρείται ότι για ορισμένα σύνολα δεδομένων (καρκίνου του μαστού, του νεφρού, λευχαιμίας και μελανώματος) ο δείκτης ARI παίρνει αρνητικές τιμές. Αυτό σημαίνει ότι μια τυχαία ομαδοποίηση των ενσωματωμάτων κειμένου θα ήταν πιο κοντά στην ομαδοποίηση των Node2Vec ενσωματωμάτων, σε σχέση με αυτή που παρατηρήθηκε. Ωστόσο, παρατηρώντας προσεκτικά τους άξονες διακρίνεται πως, αν και αρνητική, η τιμή του ARI είναι πολύ κοντά στο μηδέν, για τα περισσότερα από τα σύνολα δεδομένων. Εφόσον πρόκειται για μικρού μεγέθους σύνολα δεδομένων (μικρό πλήθος κόμβων στο δίκτυο πρωτεϊνικών αλληλεπιδράσεων), ενδεχομένως να μπορεί να διατυπωθεί η υπόθεση ότι απαιτούνται μεγάλα σύνολα δεδομένων ώστε να λειτουργήσει σωστά η μεθοδολογία που παρουσιάστηκε.

Επίσης, αξιοσημείωτα είναι τα αποτελέσματα του ARI για τον καρκίνο του νεφρού, όπου παρατηρείται πως η BioWord2Vec και η BioSent2Vec μέθοδοι είναι οι μόνες για τις οποίες παίρνει αρνητικές τιμές. Η παρατήρηση αυτή δεν συνάδει με τα αποτελέσματα της Εικόνα 58, όπου οι δύο αυτές μέθοδοι δίνουν τα καλύτερα αποτελέσματα. Αντίθετα, φαίνεται πως, για το συγκεκριμένο σύνολο δεδομένων, οι doc2vec μέθοδοι αποδίδουν τα βέλτιστα, παρατήρηση που χρήζει περαιτέρω διερεύνησης.

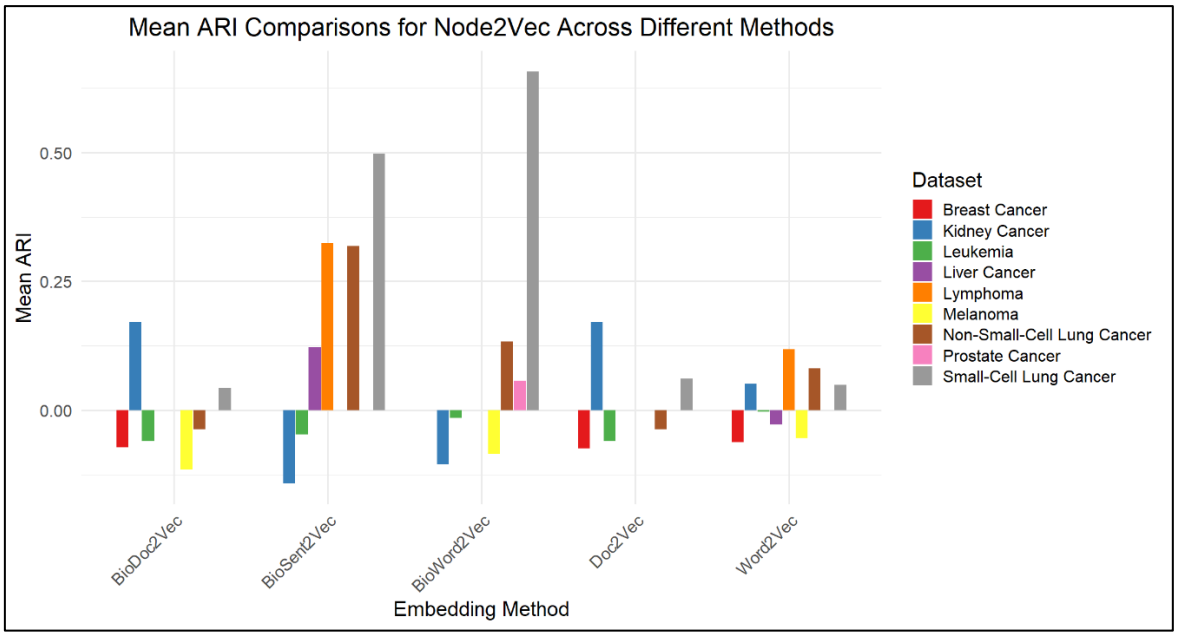
Τέλος, για κάποια σύνολα δεδομένων, δεν υπάρχουν συγκριτικά αποτελέσματα για όλες τις μεθόδους. Αυτό οφείλεται σε τεχνικές δυσκολίες και προκλήσεις που δεν επέτρεπαν την επιτυχή εκτέλεση των αλγορίθμων που αναπτύχθηκαν, για αυτά τα σύνολα δεδομένων. Η ταυτοποίηση των ακριβών αιτιών αδυναμίας υλοποίησης όλων των μεθόδων για όλα τα σύνολα δεδομένων, και ο εντοπισμός και η επίλυση των επιμέρους σφαλμάτων, επίσης χρήζει περαιτέρω διερεύνησης.

Στη συνέχεια παρουσιάζονται συγκριτικά γραφήματα όλων των μεθόδων για όλα τα σύνολα δεδομένων (Εικόνα 60 και Εικόνα 61). Τα γραφήματα αυτά προσφέρουν μια καθολική οπτική για τη συγκριτική αξιολόγηση των επιμέρους μεθόδων, αλλά και για την ποιότητα των συνόλων δεδομένων.

Συνολικά, από τα αποτελέσματα των Εικόνων 61 και 62 είναι εμφανής η επικράτηση των μεθόδων BioWord2Vec και BioSent2Vec ως αυτών με τη μεγαλύτερη ακρίβεια και απόδοση για τη διανυσματική αναπαράσταση γράφων πρωτεϊνικών αλληλεπιδράσεων. Επίσης, τρία σύνολα δεδομένων ξεχωρίζουν ως αυτά που παρήγαγαν τα καλύτερα αποτελέσματα, αυτά του μικροκυτταρικού και του μη μικροκυτταρικού καρκίνου του πνεύμονα, και αυτό του λεμφώματος.



Εικόνα 60. Μέσος όρος προσαρμοσμένου δείκτη Rand κάθε μεθόδου διανυσματοποίησης κειμένου, για κάθε σύνολο δεδομένων.



Εικόνα 61. Μέσος όρος προσαρμοσμένου δείκτη Rand κάθε συνόλου δεδομένων, για κάθε μέθοδο διανυσματοποίησης κειμένου.

6 Συζήτηση

6.1 Σημασία αποτελεσμάτων

Η παρούσα ενότητα επικεντρώνεται στην αξιολόγηση των αποτελεσμάτων των διαφόρων μεθόδων διανυσματοποίησης που εφαρμόστηκαν για την αναπαράσταση των πρωτεϊνικών αλληλεπιδράσεων. Οι μεθοδολογίες που εξετάστηκαν περιλαμβάνουν τις Graph2Vec, Node2Vec, Word2Vec, BioWord2Vec, BioSent2Vec και BioDoc2Vec.

Από τα αποτελέσματα που παρουσιάστηκαν, προκύπτει ότι οι μέθοδοι BioWord2Vec και BioSent2Vec υπερέχουν σημαντικά σε σύγκριση με τις υπόλοιπες. Συγκεκριμένα, οι μέθοδοι αυτές επέδειξαν την υψηλότερη ακρίβεια και συνέπεια στην ομαδοποίηση των πρωτεϊνικών ενσωματωμάτων, με βάση τον προσαρμοσμένο δείκτη Rand (Adjusted Rand Index - ARI). Ο μέσος όρος του ARI όλων των συνόλων δεδομένων υπολογίστηκε για κάθε μέθοδο, και τα αποτελέσματα έδειξαν ότι οι BioWord2Vec και BioSent2Vec παρήγαγαν σημαντικά υψηλότερες τιμές ARI από τις άλλες μεθόδους.

Ειδικότερα, η μέθοδος BioWord2Vec, η οποία χρησιμοποιεί λέξεις-κλειδιά από τις περιγραφές των πρωτεϊνών, φάνηκε να προσφέρει την καλύτερη αναπαράσταση των λειτουργικών σχέσεων μεταξύ των πρωτεϊνών. Αυτό υποδεικνύει ότι η ποιότητα των ενσωματωμάτων που προέρχονται από αυτή τη μέθοδο εξαρτάται σε μεγάλο βαθμό από τα δεδομένα εκπαίδευσης του μοντέλου. Αντίστοιχα, η μέθοδος BioSent2Vec, η οποία αξιοποιεί προτάσεις από τις περιγραφές, προσέφερε επίσης εξαιρετική απόδοση, πιθανώς λόγω της ικανότητάς της να συλλαμβάνει πιο πλούσιες και σύνθετες πληροφορίες από τα κείμενα.

Η σύγκριση των μεθόδων δείχνει ότι, ενώ οι παραδοσιακές τεχνικές διανυσματοποίησης, όπως οι Word2Vec και Doc2Vec, παρέχουν χρήσιμα αποτελέσματα, οι εξειδικευμένες βιοϊατρικές τεχνικές BioWord2Vec και BioSent2Vec είναι πιο κατάλληλες για την αναπαράσταση πρωτεϊνικών αλληλεπιδράσεων σε επίπεδο λειτουργικότητας. Τα ευρήματα αυτά αναδεικνύουν τη σημασία της επιλογής της κατάλληλης μεθόδου ανάλογα με το είδος των δεδομένων και τον στόχο της μελέτης.

Επιπλέον, η ανάλυση των συνόλων δεδομένων κατέδειξε ότι τα σύνολα δεδομένων που περιλαμβάνουν πληροφορίες για συγκεκριμένους τύπους καρκίνου, όπως ο μικροκυτταρικός και μη μικροκυτταρικός καρκίνος του πνεύμονα, καθώς και το λέμφωμα,

απέδωσαν τα καλύτερα αποτελέσματα σε όρους ακρίβειας και ποιότητας των παραγόμενων ενσωματωμάτων. Αυτό υπογραμμίζει τη σημασία της χρήσης κατάλληλων και αντιπροσωπευτικών συνόλων δεδομένων για την εκπαίδευση και αξιολόγηση των μοντέλων .

Παρά τα ενθαρρυντικά αποτελέσματα που προέκυψαν από τη χρήση των BioWord2Vec και BioSent2Vec, είναι σημαντικό να αναγνωρίσουμε ότι οι προτεινόμενες μέθοδοι δεν είναι αλάνθαστες. Σε ορισμένα σύνολα δεδομένων, οι μέθοδοι αυτές δεν απέδωσαν τα αναμενόμενα αποτελέσματα, υποδεικνύοντας την ανάγκη για περαιτέρω βελτίωση και προσαρμογή των μοντέλων. Η πολυπλοκότητα των βιολογικών δεδομένων και οι διαφοροποιήσεις στα χαρακτηριστικά των διαφορετικών συνόλων δεδομένων μπορεί να επηρεάσουν την αποτελεσματικότητα των μεθόδων, καθιστώντας αναγκαία τη συνεχή έρευνα και ανάπτυξη πιο προηγμένων τεχνικών.

Επιπλέον, είναι εμφανές ότι δεν είναι όλα τα σύνολα δεδομένων κατάλληλα για την εκπαίδευση και αξιολόγηση των μοντέλων διανυσματοποίησης. Οι ετερογένειες στα δεδομένα και η ποικιλία των πηγών μπορεί να δημιουργήσουν προκλήσεις στην προσπάθεια αναπαράστασης των πρωτεϊνικών αλληλεπιδράσεων. Συνεπώς, απαιτείται περαιτέρω διερεύνηση για την αναγνώριση των καταλληλότερων δεδομένων και την ανάπτυξη μεθόδων που μπορούν να προσαρμοστούν στις ιδιαιτερότητες κάθε συνόλου δεδομένων, εξασφαλίζοντας έτσι πιο αξιόπιστα και γενικεύσιμα αποτελέσματα.

6.2 Τεχνικές προκλήσεις και περιορισμοί

Η ανάλυση των βιολογικών δεδομένων, ιδιαιτέρως σε επίπεδο πρωτεϊνικών αλληλεπιδράσεων, συνοδεύεται από πληθώρα τεχνικών προκλήσεων και περιορισμών που επηρεάζουν την ακρίβεια και την αξιοπιστία των αποτελεσμάτων. Οι προκλήσεις αυτές σχετίζονται τόσο με την ποιότητα των δεδομένων όσο και με τις μεθόδους επεξεργασίας και ανάλυσης που χρησιμοποιούνται.

Πρώτον, η ποιότητα και η ετερογένεια των δεδομένων αποτελούν σημαντικούς παράγοντες που επηρεάζουν τα αποτελέσματα. Η ταχύτητα με την οποία παράγονται τα βιολογικά δεδομένα και η ποικιλία των πηγών από τις οποίες προέρχονται, δημιουργούν προκλήσεις όσον αφορά την αποθήκευση, τη διαχείριση και την επεξεργασία τους. Οι παραδοσιακές πλατφόρμες και μέθοδοι ανάλυσης συχνά δεν μπορούν να ανταποκριθούν στις απαιτήσεις

που θέτει ο τεράστιος όγκος και η πολυπλοκότητα των βιολογικών δεδομένων, καθιστώντας αναγκαία την ανάπτυξη πιο προηγμένων τεχνικών και υποδομών.

Δεύτερον, οι μέθοδοι διανυσματοποίησης που χρησιμοποιούνται, όπως οι BioWord2Vec και BioSent2Vec, αν και έχουν δείξει υποσχόμενα αποτελέσματα, δεν είναι απαλλαγμένες από περιορισμούς. Για παράδειγμα, η απόδοση των μεθόδων αυτών μπορεί να επηρεαστεί αρνητικά από την ποιότητα των δεδομένων εκπαίδευσης και την παρουσία θορύβου στα δεδομένα. Επιπλέον, οι μέθοδοι αυτές μπορεί να μην αποδίδουν εξίσου καλά για όλα τα σύνολα δεδομένων, όπως φάνηκε από τα αρνητικά αποτελέσματα του ARI για τον καρκίνο του νεφρού.

Ένας σημαντικός περιορισμός αφορά την υλοποίηση και αυτοματοποίηση των αλγορίθμων. Οι μέθοδοι διανυσματοποίησης απαιτούν συνήθως την εκτέλεση πολλών επιμέρους αναλύσεων, υπολογισμών και εκπαιδεύσεων μοντέλων. Αυτές οι διαδικασίες είναι συχνά χρονοβόρες και απαιτούν σημαντικούς υπολογιστικούς πόρους. Για παράδειγμα, η μέθοδος BioSent2Vec στηρίζεται σε παρωχημένες βιβλιοθήκες Python και σε μια παρωχημένη υλοποίηση του Sent2Vec, η οποία απαιτεί συγκεκριμένη έκδοση του Linux για να λειτουργήσει και συγκεκριμένες παλιότερες εκδόσεις πολλών βιβλιοθηκών Python. Αυτοί οι περιορισμοί καθιστούν την υλοποίηση πιο πολύπλοκη και δύσκολη.

Επιπλέον, το τεράστιο μέγεθος των προεκπαιδευμένων μοντέλων BioWord2Vec και BioSent2Vec αποτέλεσε σημαντική τεχνική δυσκολία. Ιδίως το μοντέλο BioSent2Vec, με προεκπαιδευμένα διανύσματα 700 διαστάσεων, είχε τεράστιο μέγεθος. Το γεγονός αυτό, σε συνδυασμό με τη μη βελτιστοποιημένη υλοποίηση της βιβλιοθήκης Sent2Vec, που απαιτούσε να φορτωθεί ολόκληρο στη μνήμη RAM του υπολογιστή, έθεσαν περιορισμούς στο μέγεθος των συνόλων δεδομένων που μπορούσαν να αναλυθούν. Αυτές οι δυσκολίες καθιστούν την επεξεργασία των δεδομένων απαιτητική και συχνά αναποτελεσματική.

Οι προγραμματιστικές προκλήσεις δεν περιορίζονται μόνο στην υλοποίηση των αλγορίθμων, αλλά επεκτείνονται και στην αυτοματοποίηση των διαδικασιών. Για να επιτευχθεί πλήρης αυτοματοποίηση των εργασιών, έπρεπε να γίνουν πολλές επιμέρους αναλύσεις, εκπαιδεύσεις μοντέλων και οπτικοποιήσεις, για πολλά σύνολα δεδομένων. Κάθε φορά παρουσιαζόταν διαφορετικές προκλήσεις και περιορισμοί. Για παράδειγμα, κάθε μοντέλο παράγει τα ενσωματώματα σε διαφορετική μορφή. Κάποια μοντέλα τα αποθηκεύουν σε μορφή dataframe, άλλα σε dictionary, ενώ άλλα αποθηκεύουν τα

ενσωματώματα σε αρχεία τιμών χωρισμένων με κόμμα ή κενό. Επιπλέον, κάποια μοντέλα βάζουν αγκύλη στην αρχή και το τέλος κάθε ενσωματώματος, που έπρεπε να αφαιρεθεί προγραμματιστικά για να διαβαστεί σωστά το ενσωμάτωμα ως διάνυσμα και όχι ως συμβολοσειρά. Αυτές οι διαφοροποιήσεις απαιτούσαν προσαρμογές και επιπλέον προγραμματιστική εργασία για να διασφαλιστεί η ομοιομορφία και η σωστή επεξεργασία των δεδομένων.

Τέλος, η εφαρμογή μεθόδων μηχανικής μάθησης και βαθιάς μάθησης για την ανάλυση βιολογικών μεγάλων δεδομένων φέρει τις δικές της προκλήσεις. Οι τεχνικές αυτές, αν και εξαιρετικά ισχυρές, απαιτούν μεγάλους όγκους δεδομένων για εκπαίδευση και σημαντικούς υπολογιστικούς πόρους. Η επιτυχία τους εξαρτάται από την ποιότητα των δεδομένων εκπαίδευσης και την ικανότητα των αλγορίθμων να γενικεύουν καλά σε νέα, αόρατα δεδομένα.

Οι προκλήσεις αυτές υπογραμμίζουν την ανάγκη για συνεχή έρευνα και ανάπτυξη. Η κατανόηση και η αντιμετώπιση των τεχνικών περιορισμών είναι ζωτικής σημασίας για τη βελτίωση της ακρίβειας και της αξιοπιστίας των μεθόδων ανάλυσης βιολογικών δεδομένων. Η εξέλιξη των τεχνολογιών και η ανάπτυξη νέων αλγορίθμων και τεχνικών θα συμβάλουν σημαντικά στην υπέρβαση αυτών των προκλήσεων και στην προώθηση της επιστήμης των βιολογικών δεδομένων.

6.3 Μελλοντική έρευνα

Η μελλοντική έρευνα στον τομέα της ανάλυσης βιολογικών δεδομένων μπορεί να επωφεληθεί από διάφορες προτάσεις, οι οποίες στοχεύουν στην ενίσχυση της ακρίβειας και της αποδοτικότητας των μεθόδων διανυσματοποίησης και ανάλυσης που αναπτύχθηκαν και αξιολογήθηκαν. Οι προτάσεις αυτές περιλαμβάνουν τη βελτίωση των υπαρχουσών μεθόδων, την ενσωμάτωση νέων προσεγγίσεων και την εκμετάλλευση προηγμένων τεχνικών υπολογιστικής ισχύος, και συνοψίζονται παρακάτω.

Η εφαρμογή των μεθόδων dna2Vec και Protein2Vec, σε συνδυασμό με την πληροφορία κειμένου, που διανυσματοποιούν ακολουθίες dna και αμινοξέων των πρωτεϊνών αντίστοιχα, αναμένεται να προσφέρουν πιο ακριβή και αντιπροσωπευτικά ενσωματώματα. Η προσέγγιση αυτή μπορεί να ενσωματώσει τις ιδιαιτερότητες των πρωτεϊνικών αλληλουχιών

και των γονιδίων από τα οποία προέρχονται, οδηγώντας σε βελτιωμένη ακρίβεια στην ανάλυση των πρωτεϊνικών αλληλεπιδράσεων.

Η χρήση πρόσθετων μεθόδων αξιολόγησης της απόδοσης των ομαδοποιήσεων, όπως οι NMI (Normalized Mutual Information), Homogeneity, Completeness, και V-Measure, θα επιτρέψει μια πιο ολοκληρωμένη και λεπτομερή ανάλυση των αποτελεσμάτων. Οι μεθόδοι αυτές θα παρέχουν πρόσθετες μετρικές που θα βοηθήσουν στην καλύτερη εκτίμηση της ποιότητας των παραγόμενων ομαδοποιήσεων.

Η σύγκριση των ομαδοποιήσεων των ενσωματωμάτων που προκύπτουν από το Node2Vec με τις ομαδοποιήσεις των κόμβων του γράφου σε μορφή γράφου θα βοηθήσει στην αξιολόγηση της απόδοσης του Node2Vec. Αυτή η προσέγγιση θα επιτρέψει τη σύγκριση των αποτελεσμάτων από διαφορετικές προσεγγίσεις και θα βοηθήσει στην κατανόηση των πλεονεκτημάτων και περιορισμών κάθε μεθόδου.

Η σύγκριση των ομαδοποιήσεων των ενσωματωμάτων που παράγονται με μεθόδους διανυσματοποίησης κειμένου με τις ομαδοποιήσεις που προκύπτουν από την ανάλυση των γράφων (π.χ. με τον αλγόριθμο Walktrap) μπορεί να αποτελέσει πιο αυστηρό κριτήριο για την αξιολόγηση της απόδοσης κάθε μεθόδου. Αυτό θα επιτρέψει την κατανόηση του πόσο καλά οι μέθοδοι διανυσματοποίησης κειμένου αντικατοπτρίζουν τις πραγματικές δομές των βιολογικών γράφων.

Η εφαρμογή και δοκιμή διαφορετικών μεθόδων ομαδοποίησης, όπως ο k-NN (k-nearest neighbors), που είναι κατάλληλες για ανάλυση δικτύων, μπορεί να αποκαλύψει νέες πτυχές και να βελτιώσει την ακρίβεια των αποτελεσμάτων. Η εξερεύνηση πολλαπλών μεθόδων θα βοηθήσει στην ανεύρεση της βέλτιστης προσέγγισης για κάθε σύνολο δεδομένων.

Η δυναμική προσαρμογή του αριθμού των ομάδων (clusters) αντί της χειροκίνητης ρύθμισης σε σταθερό αριθμό (π.χ. τρεις ομάδες) μπορεί να βελτιώσει την ακρίβεια των ομαδοποιήσεων. Η προσαρμογή αυτή ανάλογα με το σύνολο δεδομένων και τη μέθοδο διανυσματοποίησης θα επιτρέψει την παραγωγή πιο ρεαλιστικών και ακριβών ομαδοποιήσεων.

Η ενσωμάτωση δεδομένων εικόνων μέσω της μεθόδου Image2Vec μπορεί να προσφέρει πρόσθετες διαστάσεις πληροφορίας, βελτιώνοντας την ακρίβεια και την πολυπλοκότητα των

αναλύσεων. Η προσέγγιση αυτή θα επιτρέψει την ενσωμάτωση οπτικών δεδομένων στην ανάλυση βιολογικών γράφων.

Η χρήση πραγματικών εργαστηριακών δεδομένων, απευθείας από νοσοκομεία, αντί από δημόσιες βάσεις δεδομένων, θα βελτιώσει την ακρίβεια των αποτελεσμάτων. Τα δεδομένα αυτά θα παρέχουν πιο ρεαλιστικές και αξιόπιστες πληροφορίες για την ανάλυση των πρωτεϊνικών αλληλεπιδράσεων.

Οι παραπάνω προτάσεις αποσκοπούν τελικά στη δημιουργία και εκπαίδευση ενός προεκπαιδευμένου μοντέλου που θα βασίζεται ακριβώς στα δεδομένα που ενδιαφέρουν την έρευνα, αντί της χρήσης γενικών μοντέλων όπως το BioWord2Vec και το BioSent2Vec, το οποίο μπορεί να βελτιώσει σημαντικά την ακρίβεια των αποτελεσμάτων. Η εξειδίκευση αυτή θα επιτρέψει την καλύτερη προσαρμογή του μοντέλου στα συγκεκριμένα χαρακτηριστικά και ανάγκες των δεδομένων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Abdolahi, M., & Zahedi, M. (2019). A New Method for Sentence Vector Normalization Using Word2vec. *International Journal of Nonlinear Analysis and Applications*, *10*(2). https://doi.org/10.22075/ijnaa.2019.4177

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare*, *10*(3), 541. https://doi.org/10.3390/healthcare10030541

al-Khateeb, H. M., & Epiphaniou, G. (2016). How technology can mitigate and counteract cyber-stalking and online grooming. *Computer Fraud & Security*, *2016*(1), 14–18. https://doi.org/10.1016/S1361-3723(16)30008-2

Alam, A. (2023). *What is Machine Learning?* https://doi.org/10.5281/ZENODO.8231580

Alom, Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., & Nasrin, M. S. (n.d.). *The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches*.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1), 689. https://doi.org/10.1186/s12909-023-04698-z

Ammari, M. G., Gresham, C. R., McCarthy, F. M., & Nanduri, B. (2016). HPIDB 2.0: A curated database for host–pathogen interactions. *Database*, *2016*, baw103. https://doi.org/10.1093/database/baw103

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., & Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, *13*(3), 229–232. https://doi.org/10.1038/nmeth.3728

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural Machine Translation by Jointly Learning to Align and Translate* (arXiv:1409.0473). arXiv. http://arxiv.org/abs/1409.0473

Bastian, F. B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S. S., de Farias, T. M., Moretti, S., Parmentier, G., de Laval, V. R., Rosikiewicz, M., Wollbrett, J., Echchiki, A., Escoriza, A., Gharib, W. H., Gonzales-Porta, M., Jarosz, Y., Laurenczy, B., Moret, P., Person, E., … Robinson-Rechavi, M. (2021). The Bgee suite: Integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research*, *49*(D1), D831–D847. https://doi.org/10.1093/nar/gkaa793

Blazewicz, J., & Kasprzak, M. (2012). Complexity Issues in Computational Biology. *Fundamenta Informaticae*, *118*(4), 385–401. https://doi.org/10.3233/FI-2012-721

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Brown, A. V., & Brown, K. B. (1984). *Toward a Concise Definition and Perspective for Biology*. *38*.

C A Padmanabha Reddy, Y., Viswanath, P., & Eswara Reddy, B. (2018). Semi-supervised learning: A brief review. *International Journal of Engineering & Technology*, *7*(1.8), 81. https://doi.org/10.14419/ijet.v7i1.8.9977

Cai, T. T., & Ma, R. (2022). *Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data* (arXiv:2105.07536). arXiv. http://arxiv.org/abs/2105.07536

Cantelli, G., Bateman, A., Brooksbank, C., Petrov, A. I., Malik-Sheriff, R. S., Ide-Smith, M., Hermjakob, H., Flicek, P., Apweiler, R., Birney, E., & McEntyre, J. (2022). The European Bioinformatics Institute (EMBL-EBI) in 2021. *Nucleic Acids Research*, *50*(D1), D11–D19. https://doi.org/10.1093/nar/gkab1127

Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., Via, A., & Colombo, T. (2021). AI applications in functional genomics. *Computational and Structural Biotechnology Journal*, *19*, 5762–5790. https://doi.org/10.1016/j.csbj.2021.10.009

Cavnar, W. B., & Trenkle, J. M. (n.d.). *N-Gram-Based Text Categorization*.

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. *Sensors*, *23*(2), 634. https://doi.org/10.3390/s23020634

Chen, Q., Peng, Y., & Lu, Z. (2019). BioSentVec: Creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–5. https://doi.org/10.1109/ICHI.2019.8904728

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., … Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, *15*(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

Cirillo, D., Ponce-de-Leon, M., & Valencia, A. (n.d.). *Algorithmic complexity in computational biology: Basics, challenges and limitations*.

Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egyp, Alfonse, M., M. Aref, M., & M. Salem, A.-B. (2014). An Ontology-Based System for Cancer Diseases Knowledge Management. *International Journal of Information Engineering and Electronic Business*, *6*(6), 55–63. https://doi.org/10.5815/ijieeb.2014.06.07

Deng, X., Den Bakker, H. C., & Hendriksen, R. S. (Eds.). (2017). *Applied Genomics of Foodborne Pathogens*. Springer International Publishing. https://doi.org/10.1007/978-3-319-43751-4

Farooq, Q. U. A., Shaukat, Z., Aiman, S., & Li, C.-H. (2021). Protein-protein interactions: Methods, databases, and applications in virus-host study. *World Journal of Virology*, *10*(6), 288–300. https://doi.org/10.5501/wjv.v10.i6.288

Filkov, V. (2005). Identifying Gene Regulatory Networks from Gene Expression Data. In S. Aluru (Ed.), *Handbook of Computational Molecular Biology* (Vol. 20053851, pp. 27-1-27–29). Chapman and Hall/CRC. https://doi.org/10.1201/9781420036275.ch27

Fisher, D. G., & Hoffman, P. (1988). The adjusted rand statistic: A SAS macro. *Psychometrika*, *53*(3), 417–423. https://doi.org/10.1007/BF02294222

Frank, E., & Olaoye, G. (n.d.). *Privacy and data protection in AI-enabled healthcare systems*.

Fulkerson, B. (1995). Machine Learning, Neural and Statistical Classification. *Technometrics*, *37*(4), 459–459. https://doi.org/10.1080/00401706.1995.10484383

Gaffney, J., Tibebu, R., Bart, R., Beyene, G., Girma, D., Kane, N. A., Mace, E. S., Mockler, T., Nickson, T. E., Taylor, N., & Zastrow-Hayes, G. (2020). Open access to genetic

sequence data maximizes value to scientists, farmers, and society. *Global Food Security*, *26*, 100411. https://doi.org/10.1016/j.gfs.2020.100411

Gehring, J., Auli, M., Grangier, D., & Dauphin, Y. N. (2017). *A Convolutional Encoder Model for Neural Machine Translation* (arXiv:1611.02344). arXiv. http://arxiv.org/abs/1611.02344

Grover, A., & Leskovec, J. (2016). *node2vec: Scalable Feature Learning for Networks* (arXiv:1607.00653). arXiv. http://arxiv.org/abs/1607.00653

Gundersen, T., & Bærøe, K. (2022). The Future Ethics of Artificial Intelligence in Medicine: Making Sense of Collaborative Models. *Science and Engineering Ethics*, *28*(2), 17. https://doi.org/10.1007/s11948-022-00369-2

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., & Beyene, J. (n.d.). Data Integration in Genetics and Genomics: Methods and Challenges. *Human Genomics and Proteomics*.

Hamilton, W. L., Ying, R., & Leskovec, J. (2018). *Inductive Representation Learning on Large Graphs* (arXiv:1706.02216). arXiv. http://arxiv.org/abs/1706.02216

Hassoun, S., Jefferson, F., Shi, X., Stucky, B., Wang, J., & Rosa, E. (2022). Artificial Intelligence for Biology. *Integrative and Comparative Biology*, *61*(6), 2267–2275. https://doi.org/10.1093/icb/icab188

Haykin, S. S., & Haykin, S. S. (2009). *Neural networks and learning machines* (3rd ed). Prentice Hall.

Heidorn, P. B., Palmer, C. L., & Wright, D. (2007). [No title found]. *Journal of Biomedical Discovery and Collaboration*, *2*(1), 1. https://doi.org/10.1186/1747-5333-2-1

High-Level Expert Group on Artificial Intelligence. (2019). A definition of AI: Main capabilities and scientific disciplines. *European Commission*.

Hoda, M. N. (Ed.). (2016). *Proceedings of the 10th INDIACom ; 2016 3rd International Conference on Computing for Sustainable Global Development: (16th-18th March, 2016) : INDIA Com-2016*. Bharati Vidyapeeth's Institute of Computer Applications and Management.

Hood, L., & Rowen, L. (2013). The human genome project: Big science transforms biology and medicine. *Genome Medicine*, *5*(9), 79. https://doi.org/10.1186/gm483

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Ishak, I., & Salim, N. (2006). *Database Integration Approaches for Heterogeneous Biological Data Sources: An overview*.

Jeong, J., Lee, S., & Kwak, N. (2020). Self-Training using Selection Network for Semi-supervised Learning: *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, 23–32. https://doi.org/10.5220/0008940900230032

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with

AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kanehisa, M., & Goto, S. (n.d.). *KEGG: Kyoto Encyclopedia of Genes and Genomes*.

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, *20*(4), 1085–1093. https://doi.org/10.1093/bib/bbx085

Karuppusamy, P., García Márquez, F. P., & Nguyen, T. N. (Eds.). (2022). *Ubiquitous Intelligent Systems: Proceedings of Second ICUIS 2022* (Vol. 302). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-2541-2

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., … Pandey, A. (2009). Human Protein Reference Database—2009 update. *Nucleic Acids Research*, *37*(Database), D767–D772. https://doi.org/10.1093/nar/gkn892

Khanal, J., Tayara, H., Zou, Q., & Chong, K. T. (2021). Identifying DNA N4-methylcytosine sites in the rosaceae genome with a deep learning model relying on distributed feature representation. *Computational and Structural Biotechnology Journal*, *19*, 1612–1619. https://doi.org/10.1016/j.csbj.2021.03.015

Kushwaha, V., Prasad, S., & Kumar, P. (n.d.). *Chapter—1 Applications of Artificial Intelligence in Biological Science*.

Lamarck, J. B. P. A. de M. (1802). *Hydrogeology*.

Larivière, D., Abueg, L., Brajuka, N., Gallardo-Alba, C., Ko, B. J., Ostrovsky, A., Palmada-Flores, M., Pickett, B. D., Rabbani, K., Balacco, J. R., Chaisson, M., Cheng, H., Collins, J., Denisova, A., Fedrigo, O., Gallo, G. R., Giani, A. M., Gooder, M., Jain, N., … Formenti, G. (n.d.). *Scalable, accessible, and reproducible reference genome assembly and evaluation in Galaxy*.

Le, Q. V., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents* (arXiv:1405.4053). arXiv. http://arxiv.org/abs/1405.4053

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

Liu, C., Liu, X., Wu, F., Xie, M., Feng, Y., & Hu, C. (2018). Using Artificial Intelligence (Watson for Oncology) for Treatment Recommendations Amongst Chinese Patients with Lung Cancer: Feasibility Study. *Journal of Medical Internet Research*, *20*(9), e11087. https://doi.org/10.2196/11087

Liu, Q., & Wu, Y. (2012). Supervised Learning. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 3243–3245). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_451

Liu, R., Wang, H., & Yu, X. (2018). Shared-nearest-neighbor-based clustering by fast search and find of density peaks. *Information Sciences*, *450*, 200–226. https://doi.org/10.1016/j.ins.2018.03.031

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology*, *13*(5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457

MacNeil, L. T., & Walhout, A. J. M. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, *21*(5), 645–657. https://doi.org/10.1101/gr.097378.109

Mak, K.-K., & Pichika, M. R. (2019). Artificial intelligence in drug development: Present status and future prospects. *Drug Discovery Today*, *24*(3), 773–780. https://doi.org/10.1016/j.drudis.2018.11.014

Makiewicz, A., & Ratajczak, W. (n.d.). PRINCIPAL COMPONENTS ANALYSIS (PCA). *Principal Components Analysis*.

Mansour, A., Mohammad, J., & Kravchenko, Y. (2022). Text Vectorization Method Based on Concept Mining Using Clustering Techniques. *2022 VI International Conference on Information Technologies in Engineering Education (Inforino)*, 1–10. https://doi.org/10.1109/Inforino53888.2022.9782908

Mcculloch, W. S., & Pitts, W. (n.d.). *A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*.

McRae, C. (1890). *FATHERS OF BIOLOGY*. Percival & Co. King Streed, Covent Garden.

Miltiadous, A., Tzimourta, K. D., Giannakeas, N., Tsipouras, M. G., Afrantou, T., Ioannidis, P., & Tzallas, A. T. (2021). Alzheimer's Disease and Frontotemporal Dementia: A Robust Classification Method of EEG Signals and a Comparison of Validation Methods. *Diagnostics*, *11*(8), 1437. https://doi.org/10.3390/diagnostics11081437

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., & Ping, P. (2019). Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes*, *10*(2), 87. https://doi.org/10.3390/genes10020087

Moghadasi, M. N., & Zhuang, Y. (2020). Sent2Vec: A New Sentence Embedding Representation With Sentimental Semantic. *2020 IEEE International Conference on Big Data (Big Data)*, 4672–4680. https://doi.org/10.1109/BigData50022.2020.9378337

Myneni, S., & Patel, V. L. (2010). Organization of biomedical data for collaborative scientific research: A research information management system. *International Journal of Information Management*, *30*(3), 256–264. https://doi.org/10.1016/j.ijinfomgt.2009.09.005

Naeem, M., Rizvi, S. T. H., & Coronato, A. (2020). A Gentle Introduction to Reinforcement Learning and its Application in Different Fields. *IEEE Access*, *8*, 209320–209344. https://doi.org/10.1109/ACCESS.2020.3038605

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*, *13*(1), 911–921. https://doi.org/10.12785/ijcds/130172

Navarro, F. C. P., Mohsen, H., Yan, C., Li, S., Gu, M., Meyerson, W., & Gerstein, M. (2019). Genomics and data science: An application within an umbrella. *Genome Biology*, *20*(1), 109. https://doi.org/10.1186/s13059-019-1724-1

Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. B. (2019). Machine learning: Applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, *11*(1), 111–118. https://doi.org/10.1007/s12551-018-0449-9

Nightingale, A., Antunes, R., Alpi, E., Bursteinas, B., Gonzales, L., Liu, W., Luo, J., Qi, G., Turner, E., & Martin, M. (2017). The Proteins API: Accessing key integrated protein and genome information. *Nucleic Acids Research*, *45*(W1), W539–W544. https://doi.org/10.1093/nar/gkx237

Noori, S., Al-A'araji, N., & Al-Shamery, E. (2022). Construction of dynamic protein interaction network based on gene expression data and quartile one principle. *Proteins: Structure, Function, and Bioinformatics*, *90*(5), 1219–1228. https://doi.org/10.1002/prot.26304

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). *The complete sequence of a human genome*.

Oladipupo, T. (2010). Machine Learning Overview. In Y. Zhang (Ed.), *New Advances in Machine Learning*. InTech. https://doi.org/10.5772/9374

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., … Hermjakob, H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, *42*(D1), D358–D363. https://doi.org/10.1093/nar/gkt1115

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., & Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, *47*(D1), D529–D541. https://doi.org/10.1093/nar/gky1079

Pagliardini, M., Gupta, P., & Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 528–540. https://doi.org/10.18653/v1/N18-1049

Pal, S., Mondal, S., Das, G., Khatua, S., & Ghosh, Z. (2020). Big data in biology: The hope and present-day challenges in it. *Gene Reports*, *21*, 100869. https://doi.org/10.1016/j.genrep.2020.100869

Pascucci, M., Royer, G., Adamek, J., Aristizabal, D., Blanche, L., Bezzarga, A., Boniface-Chang, G., Brunner, A., Curel, C., Dulac-Arnold, G., Malou, N., Nordon, C., Runge, V., Samson, F., Sebastian, E., Soukieh, D., Vert, J.-P., Ambroise, C., & Madoui, M.-A. (2020). *The first AI-based mobile application for antibiotic resistance testing*. https://doi.org/10.1101/2020.07.23.216929

Pascucci, M., Royer, G., Adamek, J., Asmar, M. A., Aristizabal, D., Blanche, L., Bezzarga, A., Boniface-Chang, G., Brunner, A., Curel, C., Dulac-Arnold, G., Fakhri, R. M., Malou, N., Nordon, C., Runge, V., Samson, F., Sebastian, E., Soukieh, D., Vert, J.-P., … Madoui, M.-A. (2021). AI-based mobile application to fight antibiotic resistance. *Nature Communications*, *12*(1), 1173. https://doi.org/10.1038/s41467-021-21187-3

Patient, S., Wieser, D., Kleen, M., Kretschmann, E., Jesus Martin, M., & Apweiler, R. (2008). UniProtJAPI: A remote API for accessing UniProt data. *Bioinformatics*, *24*(10), 1321–1322. https://doi.org/10.1093/bioinformatics/btn122

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010

Perakakis, N., Yazdani, A., Karniadakis, G. E., & Mantzoros, C. (2018). Omics, big data and machine learning as tools to propel understanding of biological mechanisms and

to discover novel diagnostics and therapeutics. *Metabolism*, *87*, A1–A9. https://doi.org/10.1016/j.metabol.2018.08.002

Piccinini, G. (2004). The First Computational Theory of Mind and Brain: A Close Look at Mcculloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity." *Synthese*, *141*(2), 175–215. https://doi.org/10.1023/B:SYNT.0000043018.52445.3e

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019). An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. *2019 International Engineering Conference (IEC)*, 200–204. https://doi.org/10.1109/IEC47844.2019.8950616

Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering*, *14*, 156–180. https://doi.org/10.1109/RBME.2020.3013489

Ramon, C., & Stelling, J. (2023). Functional comparison of metabolic networks across species. *Nature Communications*, *14*(1), 1699. https://doi.org/10.1038/s41467-023-37429-5

Rand, W. M. (1971). *Objective Criteria for the Evaluation of Clustering Methods*.

Rani, D., Kumar, R., & Chauhan, N. (2022). Study and Comparision of Vectorization Techniques Used in Text Classification. *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT54827.2022.9984608

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, *149*, 106043. https://doi.org/10.1016/j.compbiomed.2022.106043

Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B., & Peyvandi, A. A. (n.d.). *Protein-protein interaction networks (PPI) and complex diseases*.

Sahani, A. (2023). *Natural Language Processing*. *4*(6).

Salwinski, L. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, *32*(90001), 449D – 451. https://doi.org/10.1093/nar/gkh086

Sammut, C., & Webb, G. I. (Eds.). (2010). TF–IDF. In *Encyclopedia of Machine Learning* (pp. 986–987). Springer US. https://doi.org/10.1007/978-0-387-30164-8_832

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Schork, N. J. (2019). Artificial Intelligence and Personalized Medicine. In D. D. Von Hoff & H. Han (Eds.), *Precision Medicine in Cancer Therapy* (Vol. 178, pp. 265–283). Springer International Publishing. https://doi.org/10.1007/978-3-030-16391-4_11

Sellwood, M. A., Ahmed, M., Segler, M. H., & Brown, N. (2018). Artificial Intelligence in Drug Discovery. *Future Medicinal Chemistry*, *10*(17), 2025–2028. https://doi.org/10.4155/fmc-2018-0212

Senders, Y. (n.d.). *THE IMPACT OF STEMMING AND LEMMATIZATION APPLIED TO WORD VECTOR BASED MODELS IN SENTIMENT ANALYSIS*.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The New System Technology*. Springer International Publishing. https://doi.org/10.1007/978-3-031-21448-6

Shortliffe, E. H., & Barnett, G. O. (2006). Biomedical Data: Their Acquisition, Storage, and Use. In E. H. Shortliffe & J. J. Cimino (Eds.), *Biomedical Informatics* (pp. 46–79). Springer New York. https://doi.org/10.1007/0-387-36278-9_2

Singh, A. K., & Shashi, M. (2019). Vectorization of Text Documents for Identifying Unifiable News Articles. *International Journal of Advanced Computer Science and Applications*, *10*(7). https://doi.org/10.14569/IJACSA.2019.0100742

Sodhi, P., Awasthi, N., & Sharma, V. (2019). Introduction to Machine Learning and Its Basic Application in Python. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3323796

Steinhaus, H. (1957). *Sur la division des corps matériels en parties*.

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods*, *9*(3), 386–396. https://doi.org/10.1037/1082-989X.9.3.386

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks* (arXiv:1312.6199). arXiv. http://arxiv.org/abs/1312.6199

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000

The Role of Protein and Amino Acids in Sustaining and Enhancing Performance (p. 9620). (1999). National Academies Press. https://doi.org/10.17226/9620

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., … Teodoro, D. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Tohti, T., Zhao, Y., & Musajan, W. (2017). Word2vec and dictionary based approach for uyghur text filtering. *Journal of Physics: Conference Series*, *887*, 012013. https://doi.org/10.1088/1742-6596/887/1/012013

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

van der Maarten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 9*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks* (arXiv:1710.10903). arXiv. http://arxiv.org/abs/1710.10903

Vinh, N. X., Epps, J., & Bailey, J. (2010). *Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance*.

Vyas, M., Thakur, S., Riyaz, B., Bansal, K. K., Tomar, B., & Mishra, V. (n.d.). Artificial Intelligence: The Beginning of a New Era in Pharmacy Profession. *Artificial Intelligence*.

Wang, L., Ling, Y., Yuan, Z., Shridhar, M., Bao, C., Qin, Y., Wang, B., Xu, H., & Wang, X. (2024). *GenSim: Generating Robotic Simulation Tasks via Large Language Models* (arXiv:2310.01361). arXiv. http://arxiv.org/abs/2310.01361

Wang, X., Zhang, H., & Liu, Y. (2018). Sentence Vector Model Based on Implicit Word Vector Expression. *IEEE Access*, *6*, 17455–17463. https://doi.org/10.1109/ACCESS.2018.2817839

Warrens, M. J., & Van Der Hoef, H. (2022). Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs. *Journal of Classification*, *39*(3), 487–509. https://doi.org/10.1007/s00357-022-09413-z

Watson, D. S. (2023). On the Philosophy of Unsupervised Learning. *Philosophy & Technology*, *36*(2), 28. https://doi.org/10.1007/s13347-023-00635-6

Watson, J., & Crick, F. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*.

Webb, S. (2018). Deep learning for biology. *Nature*, *554*(7693), 555–557. https://doi.org/10.1038/d41586-018-02174-z

Xu, C., & Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biology*, *20*(1), 76, s13059-019-1689–0. https://doi.org/10.1186/s13059-019-1689-0

Yang, X., Yang, K., Cui, T., Chen, M., & He, L. (2022). A Study of Text Vectorization Method Combining Topic Model and Transfer Learning. *Processes*, *10*(2), 350. https://doi.org/10.3390/pr10020350

Yeh, C.-W., Huang, C.-W., Yang, C.-L., & Wang, Y.-T. (2023). A High Performance Computing Platform for Big Biological Data Analysis. *2023 9th International Conference on Applied System Innovation (ICASI)*, 68–70. https://doi.org/10.1109/ICASI57738.2023.10179527

Yeung, K. Y., & Ruzzo, W. L. (2001). *Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in Bioinformatics)*.

Yin, Z., Lan, H., Tan, G., Lu, M., Vasilakos, A. V., & Liu, W. (2017). Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges. *Computational and Structural Biotechnology Journal*, *15*, 403–411. https://doi.org/10.1016/j.csbj.2017.07.004

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, *6*(1), 52. https://doi.org/10.1038/s41597-019-0055-0

Zhao, J., Yu, H., Luo, J., Cao, Z. W., & Li, Y. (2006). Complex networks theory for analyzing metabolic networks. *Chinese Science Bulletin*, *51*(13), 1529–1537. https://doi.org/10.1007/s11434-006-2015-2