ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μελέτη των ποσοτικών σχέσεων δομής - δράσης, για τη δημιουργία μοντέλων πρόβλεψης φυσικοχημικών και φαρμακοκινητικών παραμέτρων φαρμάκων

Βαΐα Αλεξ. Ρουμελιώτου

ΛΑΜΙΑ

ΙΟΥΝΙΟΣ 2019

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μελέτη των ποσοτικών σχέσεων δομής - δράσης, για τη δημιουργία μοντέλων πρόβλεψης φυσικοχημικών και φαρμακοκινητικών παραμέτρων φαρμάκων

Βαΐα Αλεξ. Ρουμελιώτου

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

Βασίλειος Πλαγιανάκος
Καθηγητής
Πανεπιστήμιο Θεσσαλίας

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ

Βασίλειος Πλαγιανάκος
Καθηγητής
Πανεπιστήμιο Θεσσαλίας

Παντελής Μπάγκος
Καθηγητής
Πανεπιστήμιο Θεσσαλίας

Μαρία Αδάμ
Αναπληρώτρια Καθηγήτρια
Πανεπιστήμιο Θεσσαλίας

Ημερομηνία Εξέτασης: 26 Ιουνίου 2019

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία έγινε η προσπάθεια δημιουργίας ενός μοντέλου πρόβλεψης του χρόνου ημίσειας ζωής των φαρμάκων, βάσει διαφόρων φαρμακοκινητικών, φαρμακοδυναμικών, φυσικοχημικών και περιγραφικών μεταβλητών της δομής τους. Για την εκπαίδευση των μοντέλων πρόβλεψης χρησιμοποιήθηκαν δεδομένα από την ευρέως χρησιμοποιούμενη βάση δεδομένων φαρμάκων Drugbank. Για τη βέλτιστη πρόβλεψη του χρόνου ημίσειας ζωής έγινε σύγκριση διαφόρων μοντέλων πρόβλεψης, όπως Support Vector Regressor, Multi - Layer Perceptron Regressor, Gradient Boosting Regressor και Random Forest Regressor. Τα μοντέλα πρόβλεψης δοκιμάστηκαν με τρεις προσεγγίσεις, τη βασική, την προσέγγιση με clustering και την προσέγγιση με voting. Στη συνέχεια οι ίδιες προσεγγίσεις δοκιμάστηκαν για την πρόβλεψη του χρόνου ημίσειας ζωής για φαρμακευτικές ουσίες που είχαν χορηγηθεί αποκλειστικά ενδοφλεβίως. Έγινε αξιολόγηση της ακρίβειας των προβλέψεων που εξήχθησαν σε όλες τις περιπτώσεις με τον υπολογισμό δεικτών σφάλματος. Το βασικό συμπέρασμα που προέκυψε είναι ότι είναι εφικτή ως ένα βαθμό η πρόβλεψη του χρόνου ημίσειας ζωής. Επιπρόσθετα, η εκπαίδευση μοντέλων με αποκλειστική χρήση φαρμάκων με συγκεκριμένη οδό χορήγησης δρα ευεργετικά στην ακρίβεια των προβλέψεων. Μελλοντικός στόχος είναι η δημιουργία μοντέλων τα οποία εκμεταλλευόμενα την ποσοτική σχέση μεταξύ της δομής, των φυσικοχημικών, φαρμακοκινητικών και φαρμακοδυναμικών παραμέτρων των ουσιών θα είναι ικανά να προβλέψουν το χρόνο ημίσειας ζωής, αλλά και άλλες παραμέτρους χρήσιμες για το σχεδιασμό νέων φαρμάκων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Σχεδιασμός Φαρμάκων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: χρόνος ημίσειας ζωής, μοντέλο QSAR, πρόβλεψη, σχεδιασμός φαρμάκων

ABSTRACT

In the present thesis, it was attempted to create a prediction model for drugs' half-life, upon various pharmacokinetic, pharmacodynamic, physicochemical and descriptive variables of their structure. For training purposes, data from the widely used online drug database Drugbank, were employed. In order to achieve the best possible prediction results, various regression methods, namely Support Vector Regressor, Multi - Layer Perceptron Regressor, Gradient Boosting Regressor and Random Forest Regressor, were tested. The prediction models were built according to three different approaches, the basic approach, the clustering approach and the voting approach. Additionally, the same approaches have also been used in the case where drugs have been solely administrated intravenously. The prediction accuracy for all cases has been assessed through the use of error estimators. The main conclusion conducted is that it is feasible, at least to some extent, to predict the half-life value of drugs. Additionally, training models on drugs with the same route of administration, seems to improve the prediction accuracy. As future work, it is planned to train models, which will be able to take advantage of the quantitative relationship between drugs' structure and their pharmacokinetic, pharmacodynamic and physicochemical parameters, in order to accurately predict half-life values along with other useful, for drugs design, parameters.

SUBJECT AREA: Drug design

KEYWORDS: half-life, QSAR model, regression, drug design

Στην οικογένειά μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγήτη Βασίλη Πλαγιανάκο, καθώς και τα μέλη της τριμελούς Επιτροπής για την καθοδήγηση και τη συνεισφορά τους στην εκπόνηση της παρούσας εργασίας.

Β.Ρ.

Λαμία, 2019

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο **Μελέτη των ποσοτικών σχέσεων δομής - δράσης, για τη δημιουργία μοντέλων πρόβλεψης φυσικοχημικών και φαρμακοκινητικών παραμέτρων φαρμάκων** αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δε μου ανήκει διότι είναι προϊόν λογοκλοπής.

Η ΔΗΛΟΥΣΑ

Βαΐα Αλεξ. Ρουμελιώτου

26/06/2019

ΠΕΡΙΕΧΟΜΕΝΑ

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

1. ΕΙΣΑΓΩΓΗ

Το φάρμακο είναι ένα πολύτιμο αγαθό που χρησιμοποιείται τόσο για την πρόληψη όσο και για την αντιμετώπιση διαφόρων ασθενειών. Στην αρχαία Ελλάδα ο Ιπποκράτης δίδασκε «Η τροφή πρέπει είναι το φάρμακό σου και το φάρμακο να είναι η τροφή σου». Τα φυτά ήταν από τα πρώτα θεραπευτικά μέσα που χρησιμοποιήθηκαν για την ανακούφιση διαφόρων συμπτωμάτων, αλλά και ως αναλγητικά. Σταθμό στη φαρμακευτική ιστορία αποτέλεσε η απομόνωση ενός αλκαλοειδούς σε κρυσταλλική μορφή από το όπιο, από έναν ερασιτέχνη φαρμακοποιό τον Sertürner κατά τα έτη 1803-1817. Το αλκαλοειδές αυτό είναι η μορφίνη, που ακόμη και σήμερα χρησιμοποιείται ως ισχυρότατο αναλγητικό φάρμακο.

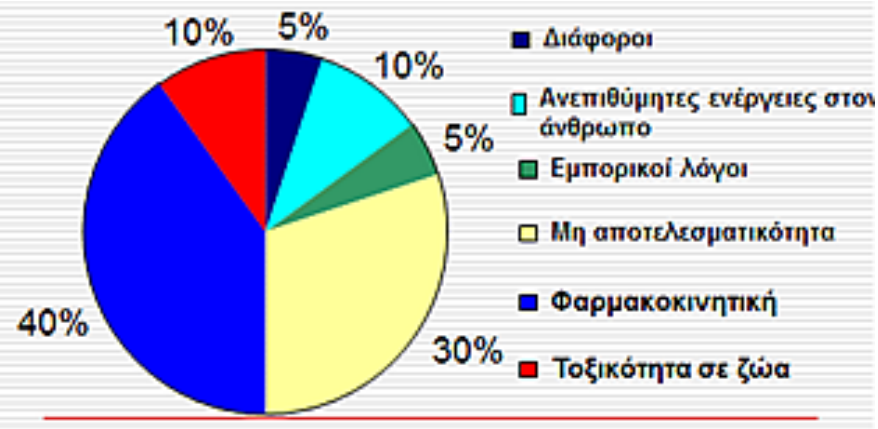
Με την ανάπτυξη της σύγχρονης φαρμακευτικής βιομηχανίας γύρω στο 1930 ξεκίνησε και η σύνθεση οργανικών ενώσεων, με σκοπό την παραγωγή υποψηφίων φαρμάκων. Προκειμένου να ελεγθεί η φαρμακολογική δράση των ουσιών άρχισαν να πραγματοποιούνται πειράματα σε ζώα καθώς και στη συνέχεια με την εξέλιξη της βιολογίας τα πειράματα αυτά επεκτάθηκαν σε απομονωμένα όργανα, κυτταροκαλλιέργειες και μεμβράνες. Η ανάπτυξη της χημικής τεχνολογίας κατά το δεύτερο μισό του εικοστού αιώνα με την εφαρμογή της κρυσταλλογραφίας ακτινών-Χ, της φασματοσκοπίας NMR, της φασματομετρίας μαζών, της υπερφυγοκέντρησης, της υγρής χρωματογραφίας υψηλής απόδοσης κ.ά., καθώς και η ραγδαία εξέλιξη των υπολογιστών και της πληροφορικής, συνέβαλαν αποφασιστικά στην έρευνα για νέα χημικά μόρια με θεραπευτική δράση.

Με την εξέλιξη της φαρμακευτικής επιστήμης μελετάται λεπτομερώς τόσο η πορεία του φαρμάκου στον οργανισμό μέσω της φαρμακοκινητικής (PK), όσο και το αποτέλεσμα δράσης του φαρμάκου στον οργανισμό μέσω της φαρμακοδυναμικής (PD). Η έρευνα σε αυτό το επίπεδο και οι παρεχόμενες γνώσεις, διαμόρφωσαν τις σύγχρονες αντιλήψεις για τις υψηλές προδιαγραφές των φαρμάκων και επηρέασαν αντίστοιχα τον προσανατολισμό της έρευνας. Το φάρμακο πρέπει πρωτίστως να είναι αποτελεσματικό και ασφαλές, ενώ παράλληλα, σύμφωνα με τις αυξανόμενες κοινωνικές απαιτήσεις για καλύτερη ποιότητα ζωής, πρέπει να είναι εύληπτο και φθηνό. Ένα αυστηρό πλέγμα κανονιστικών διατάξεων για την έγκριση νέων φαρμάκων εξασφαλίζει την τήρηση των προδιαγραφών, ωστόσο έχει αυξήσει σημαντικά, τόσο το κόστος όσο και το χρόνο που απαιτείται για την ανάπτυξή τους. Εξ άλλου ακόμη και μετά την κυκλοφορία ενός φαρμάκου στην αγορά, αυτό παραμένει υπό έλεγχο μέσω της φαρμακοεπαγρύπνησης.

Στόχος της φαρμακευτικής βιομηχανίας είναι να μειώσει τόσο το κόστος όσο και το

χρόνο που απαιτείται για την παραγωγή ενός φαρμάκου. Η πρόκληση αυτή είναι μεγάλη και για τους ακαδημαϊκούς ερευνητές. Η προσπάθεια που γίνεται στοχεύει στο να είναι εφικτή η απόρριψη ενός φαρμάκου στα πρώτα στάδια παραγωγής του, στο προκλινικό στάδιο, πριν περάσει τις χρονοβόρες και κοστοβόρες διαδικασίες των κλινικών δοκιμών.

Τη φαρμακευτική έρευνα έχουν απασχολήσει οι λόγοι αποκλεισμού των υποψηφίων φαρμάκων από την κλινική ανάπτυξη. Στατιστικές μελέτες στα τέλη της δεκαετίας του 1990 έδειξαν ότι το 40% των αποτυχιών αποδίδονται σε φαρμακοκινητικούς λόγους, ενώ 30% σε έλλειψη αποτελεσματικότητας, παρόλο που έχει τεκμηριωθεί η συγγένεια με τον υποδοχέα - στόχο. Στο Σχήμα 1.1 παρουσιάζονται οι βασικοί λόγοι διακοπής στην ανάπτυξη φαρμάκων [42].



Σχήμα 1.1: Λόγοι αποκλεισμού από την κλινική ανάπτυξη

Η εκμετάλλευση της τεχνολογικής ανάπτυξης και ο συνδυασμός της πληροφορικής με την ανάπτυξη της φαρμακευτικής χημείας και των γνώσεων της βιολογίας, οδήγησαν στη δημιουργία μοντέλων πρόβλεψης που στηρίζονται κυρίως σε μεθόδους τεχνητής νοημοσύνης και χρησιμοποιούνται για την ανακάλυψη νέων φαρμακομορίων. Οι πιο συχνά χρησιμοποιούμενες μέθοδοι είναι τα Artificial neural networks (ANN), ο Gradient boosting machine (GMB), οι Support vector regressions (SVR), οι Local lazy learnings (LLL), τα KNN, τα LLR (SA, SR, GP) και τα Consensus models (ACM, SCM).

Η διαπίστωση για το σημαντικό ρόλο των φαρμακοκινητικών ιδιοτήτων στην τελική αποτελεσματικότητα των φαρμάκων, επέδρασε καθοριστικά στην προσέγγιση της διαδικασίας ανακάλυψης νέων φαρμακομορίων. Η μελέτη των ποσοτικών σχέσεων δομής - δράσης, για τη δημιουργία μοντέλων πρόβλεψης φυσικοχημικών και φαρμακοκινητικών παραμέτρων των φαρμάκων, αποτελεί πια βασικό μέρος της διαδικασίας ανακάλυψης νέων φαρμακομορίων.

Οι Ποσοτικές Σχέσεις Δομής - Δράσης (Quantitative Structure Activity Relationships, QSAR) αποσκοπούν στην εξαγωγή εξισώσεων ή μοντέλων που συσχετίζουν τη βιολογική δράση με τη δομή, αντλώντας και αξιοποιώντας όσο το δυνατόν περισσότερες πληροφορίες και μειώνοντας αισθητά το πλήθος των απαιτούμενων πειραμάτων. Τα μοντέλα αυτά συμβάλουν σημαντικά στην κατανόηση του μηχανισμού δράσης των φαρμακομορίων. Ο απώτερος στόχος όμως των Ποσοτικών Σχέσεων Δομής - Δράσης και το βασικό κίνητρο για την ανάπτυξή τους, ήταν και παραμένει η πρόβλεψη της δράσης που θα κατευθύνει τον φαρμακοχημικό στη σύνθεση (ή τη μη σύνθεση), νέων παραγώγων.

Μία πολύ σημαντική φαρμακοκινητική παράμετρος που καθορίζει τη δράση ενός φαρμάκου είναι ο χρόνος ημίσειας ζωής. Ο χρόνος ημίσειας ζωής είναι ο χρόνος που απαιτείται ώστε η συγκέντρωση του φαρμάκου να μειωθεί κατα 50%. Τα φάρμακα χορηγούνται σε δόσεις οι οποίες επαναλαμβάνονται ανά τακτά χρονικά διαστήματα, έτσι ώστε η συγκέντρωση του φαρμάκου να παραμένει σταθερή στον οργανισμό για να υπάρξει το αναμενόμενο θεραπευτικό αποτέλεσμα. Ο χρόνος ημίσειας ζωής βοηθάει στον καθορισμό του δοσολογικού σχήματος, ώστε η συγκέντρωση του φάρμακου να είναι σταθερή. Έχει άμεση συσχέτιση με την κυκλοφορία του φαρμάκου στο πλάσμα αίματος, αλλά και με το προφίλ της συγκέντρωσης του φαρμάκου με τον χρόνο, μετά από επαναλαμβανόμενες δόσεις. Είναι αρκετά σημαντικό να μπορούμε να γνωρίζουμε το χρόνο ημίσειας ζωής ενός καινούριου φαρμάκου, ώστε να προσαρμόσουμε τη δοσολογία και τα μεσοδιαστήματα λήψης αυτου.

Η πρόβλεψη του χρόνου ημίσειας ζωής ενός νέου φαρμάκου, θα μπορούσε να μειώσει σημαντικά, τη χρονική διάρκεια και το κόστος των κλινικών δοκιμών. Με την εισαγωγή της μηχανικής μάθησης στη φαρμακευτική επιστήμη έχουν γίνει διάφορες μελέτες για τη δημιουργία μοντέλων πρόβλεψης φαρμακοκινητικών και φαρμακοδυναμικών παραμέτρων.

Στην παρούσα εργασία διερευνήθηκε η ανάπτυξη μοντέλων πρόβλεψης της τιμής του χρόνου ημίσειας ζωής νέων φαρμάκων. Δημιουργήθηκαν διάφορα μοντέλα πρόβλεψης με τη βοήθεια μεθόδων μηχανικής μάθησης. Από τη βάση δεδομένων των φαρμάκων έγινε εξαγωγή αυτών που είχαν δεδομένα για το χρόνο ημίσειας ζωής και στη συνέχεια από τα smiles τους, έγινε περιγραφή των φαρμάκων αυτών με διάφορους descriptors. Το dataset που παρείχθηκε χρησιμοποιήθηκε για την εκπαίδευση αλγορίθμων, με σκοπό τη δημιουργία μοντέλων πρόβλεψης του χρόνου ημίσειας ζωής. Δοκιμάστηκαν τρεις προσεγγίσεις, η βασική προσέγγιση που περιλαμβάνει την εκπαίδευση των αλγορίθμων με το τελικό επεξεργασμένο dataset, η προσέγγιση με τη χρήση clustering με τη δημιουργία επιμέρους clusters όσον αφορά το dataset και η προσέγγιση με voting που ουσιαστικά συνδυάζει τις

τεχνικές μηχανικής μάθησης της βασικής προσέγγισης. Έγινε σύγκριση και αξιολόγηση των αποτελεσμάτων αυτών, με σκοπό την εφαρμογή των μοντέλων για την καλύτερη δυνατή πρόβλεψη του χρόνου ημίσειας ζωής, ενός νεοεισαγόμενου φαρμάκου.

2. ΣΧΕΤΙΚΗ ΕΡΕΥΝΑ

Το 1999 οι S. Agatonovic-Kustrin, R.Beresford, δημοσίευσαν μια εργασία για τα artificial neural network, όπου αναφέρεται πως είναι μια υποσχόμενη τεχνική μοντελοποίησης, ειδικά όταν έχουμε δεδομένα με μη γραμμικη σχέση, που είναι πολύ συχνό φαινόμενο στη φαρμακευτική διαδικασία παραγωγής φαρμάκων. Όσον αφορά τη μοντελοποίηση, δεν έχει τόσο μεγάλη σημασία η πηγή που μας παρέχει τα δεδομένα, αλλά καθώς περιλαμβάνουν πολλά βάρη που πρέπει να εκτιμηθούν για να είναι ακριβή στις προβλέψεις τους, απαιτείται να χρησιμοποιούνται μεγάλα training sets. Ένα επιπλέον πλεονέκτημά τους, είναι πως μπορούν να συνδυάσουν και να επεξεργαστούν δεδομένα που προέρχονται από τη βιβλιογραφία, αλλά και από πειράματα προκειμένου να επιλύσουν κάποιο πρόβλημα [1].

Το 2003 οι Turner et al., ανέπτυξαν ένα artificial neural network για την πρόβλεψη της κάθαρσης, της σύνδεσης του φαρμάκου με τις πρωτεΐνες του πλάσματος και τον όγκο κατανομής του, σε μια ομάδα φαρμάκων με αρκετά διαφορετικές δομές [32]. Οι τιμές εισόδου στο μοντέλο ήταν ένας μεγάλος αριθμός descriptors των ουσιών, τον οποίο εξήγαγαν από τις δομές των φαρμάκων. Το μοντέλο εκπαιδεύτηκε με ενα σετ φαρμάκων κ με ένα άλλο ανεξάρτητο σετ, έγινε validation. Οι τιμές πρόβλεψης της ολικής κάθαρσης, της νεφρικής κάθαρσης και του όγκου κατανομής του φαρμάκου, ήταν αρκετά κοντά με τις πειραματικές τιμές, ενώ οι τιμές πρόβλεψης για τη σύνδεση αυτού με τις πρωτεΐνες του πλάσματος ήταν ενθαρρυντικές. Ο συνδυασμός των descriptors, των ANNs, της ταχύτητας και της επιτυχίας που είχε στην πρόβλεψη αυτή η τεχνική έναντι των παλιών συμβατικών μεθόδων, έδειξε τη δυναμική που έχει για χρήση στην ανάπτυξη φαρμακευτικών προϊόντων.

Το 2003 οι Turner et al., ανέπτυξαν ένα artificial neural network για την πρόβλεψη του χρόνου ημίσειας ζωής, της νεφρικής και ολικής κάθαρσης, του κλάσματος απέκκρισης στα ούρα, της σύνδεσης του φαρμάκου με τις πρωτεΐνες του πλάσματος και τον όγκο κατανομής του, σε μια ομάδα κεφαλοσπορινών [33]. Η ανάπτυξη αυτού του μοντέλου έφτασε στο συμπέρασμα πως μπορεί να επιτευχθεί ακριβής πρόβλεψη διαφόρων φαρμακοκινητικών παραμέτρων, κάνοντας χρήση μόνο βιβλιογραφικών δεδομένων. Το γεγονός αυτό είναι πολύ σημαντικό για τη χρησιμοποιήση του μοντέλου αυτού στα πρώτα στάδια της ανάπτυξης ενός φαρμάκου, παρακάμπτοντας χρονοβόρες και κοστοβόρες πειραματικές διαδικασίες.

Το 2008 οι Obach et al., συνέλλεξαν φαρμακοκινητικά δεδομένα που αφορούσαν 670 φαρμακευτικες ουσίες, των οποίων είχε γίνει αποκλειστικά ενδοφλέβια χορήγηση [23]. Τι-

μές για την κάθαρση, τον όγκο κατανομής σε σταθερή συγκέντρωση, το μέσο χρόνο παραμονής και της τελικής φάσης του χρόνου ημίσειας ζωής, ελήφθησαν αποκλειστικά από αναφορές σε μελέτες όπου έγινε ενδοφλέβια χορήγηση. Από άλλες πηγές έγινε λήψη των δεδομένων της πρωτεϊνικής δέσμευσης του φαρμάκου με το πλάσμα. Οι παράμετροι αυτές αναλύθηκαν περαιτέρω με τη χρήση descriptors και οι παρατηρήσεις αυτές δύναται να χρησιμοποιηθούν σε μελλοντικές προσπάθειες ανάπτυξης σχέσεων **QSPR** ποσοτικών δομών - φαρμακοκινητικών δεδομένων, ώστε να βελτιωθεί και να γίνει περισσότερο κατανοητή η σχέση μεταξύ των δομικών χημικών χαρακτηριστικών και της φαρμακευτικής διάθεσης.

Το 2012 οι Zvetanka Zhivkova et al., χρησιμοποίησαν τις ποσοτικές σχέσεις δομής - φαρμακοκινητικών δεδομένων (QSPKR) για τη δημιουργία μοντέλων για την πρόβλεψη της τιμής του όγκου κατανομής διαφόρων όξινων φαρμάκων [40]. Συλλέχθησαν οι τιμές του όγκου κατανομής 132 όξινων φαρμάκων και περιγράφηκε η δομή τους με τη βοήθεια 178 μοριακών descriptors, ενώ με τη χρήση του γενετικού αλγορίθμου και του stepwise regression δημιουργήθηκαν τα μοντέλα QSPKR. Εξήχθησαν διάφορα συμπεράσματα για τις παραμέτρους που συμβάλλουν περισσότερο στην πρόβλεψη του όγκου κατανομής, ενώ τα μοντέλα είναι χρήσιμα ως curator των διαθέσιμων φαρμακοκινητικών βάσεων δεδομένων.

Το 2013 οι Zandkarimi et al., παρουσίασαν μία μέθοδο που έκανε προβλέψεις για την κάθαρση, για την πρωτεϊνική σύνδεση με το πλάσμα και για τον όγκο κατανομής του φαρμάκου, για μια ομάδα αλκαλικών φαρμάκων [39]. Χρησιμοποιήθηκε ο γενετικός αλγόριθμος σε συνδυασμό με artificial neural networks για να γίνει η επιλογή των πιο σχετικών μοριακών descriptors και να αναπτυχθούν τα μοντέλα ποσοτικών σχέσεων δομής - φαρμακοκινητικών δεδομένων. Τα αποτελέσματα έδειξαν πως οι 3d μοριακοί descriptors είχαν τη μεγαλύτερη επίδραση στα μοντέλα αυτά. Τα αποτελέσματα των τιμών πρόβλεψης φαρμακοκινητικών παραμέτρων ήταν αρκετά ικανοποιητικα.

Το 2016 οι Lu et al., δημιούργησαν ένα μοντέλο πρόβλεψης για το χρόνο ημίσειας ζωής του φαρμάκου [16]. Με τις τιμές του χρόνου ημίσειας ζωής 1105 φαρμάκων και την εξαγωγή των descriptors αυτών, χρησιμοποιήθηκαν διάφορες μέθοδοι μηχανικής μάθησης, όπως gradient boosting machine (GBM), support vector regressions (RBF-SVR and Linear-SVR), local lazy regression (LLR), SA, SR, and GP. Χρησιμοποιήθηκαν επίσης και δύο consensus models, τα οποία ωστόσο δεν κατάφεραν να δώσουν καλύτερη απόδοση στην πρόβλεψη. Ένα ακριβές μοντέλο πρόβλεψης του χρόνου ημίσειας ζωής δημιουργήθηκε με την μέθοδο GBM, δίνοντας ενθαρρυντικά αποτελέσματα για την εφαρμογή του

στα πρώτα στάδια σχεδιασμού των υποψηφίων φαρμάκων.

Το 2011 οι Lowe et al., δοκίμασαν διάφορες μεθόδους μηχανικής μάθησης για την πρόβλεψη του logP [15]. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι artificial neural networks (ANNs), support vector machines (SVM) για regression, και k-nearest neighbor (k-NN). Την καλύτερη προβλεπτική ικανότητα από τις μεμονωμένες μεθόδους είχε η k-NN, ενώ την καλύτερη απόδοση από τα consensus που εξετάστηκαν είχε το ANN/SVM/k-NN model.

Το 2006 οι Chen et al., διαπίστωσαν πως έχουν δημιουργηθεί διάφορα μοντέλα πρόβλεψης της ολικής κάθαρσης του φαρμάκου [38]. Τα μοντέλα αυτά έχουν υψηλή προβλεπτική ικανότητα, αλλά στηρίζονται κυρίως σε περιορισμένο αριθμό συγγενών ουσιών που είναι πολύ λίγες σε αριθμό και σε ποικιλία, σε σχέση με τις ουσίες που αναφέρονται στη βιβλιογραφία με γνωστή τιμή ολικής κάθαρσης. Χρησιμοποιήθηκαν 3 μέθοδοι, οι general regression neural network (GRNN), support vector regression (SVR) και k-nearest neighbour (KNN), προκειμένου να εξεταστεί αν η προβλεπτική ικανότητα αυτών, μπορεί να είναι εξίσου καλή και σε μεγαλύτερο αριθμό διαφορετικών μεταξύ τους φαρμάκων. Δημιουργήθηκαν κάποια σετ descriptors με τα οποία δούλεψαν οι αλγόριθμοι και βρέθηκε αυτό που δίνει τα καλύτερα αποτελέσματα πρόβλεψης QSPKR ιδιοτήτων των φαρμάκων, τόσο στα μεμονωμένα μοντέλα, όσο και στα consensus.

Το 2011, έγινε μια έρευνα για ένα φάρμακο που έφτασε στην κλινική φάση ΙΙ ως μη αναστρέψιμος αναστολέας του FAAH για πόνους οστεοαρθρίτιδας. Βασιζόμενοι στα δεδομένα αυτής της έρευνας, το 2014 οι Benson et al. [18], δημοσίευσαν μια μελέτη για το πως το επίπεδο των ενδοκανναβινοειδών (νευροδιαβιβαστές μικρής διαρκείας), εξαρτάται από τα ένζυμα που τα διασπούν όπως το FAAH (υδρολάση αμιδίου λιπαρού οξέος).

Η μελέτη αυτή χρησιμοποίησε τόσο φαρμακοκινητικά δεδομένα όσο και βιολογικά συστήματα, προκειμένου να δημιουργηθεί ένα μοντέλο που στοχεύει στην αναστολή δράσης του FAAH. Με την αναστολή της δράσης του, διατηρείται σε υψηλή συγκέντρωση το επίπεδο των ενδοκανναβινοειδών κ έτσι μειώνεται ο πόνος. Το μοντέλο που δημιουργήθηκε κατάφερε να διελευκάνει κάποια κενά που υπήρχαν στην έρευνα, αλλά και να εκτιμήσει το ρίσκο της επόμενης φάσης. Συγκεκριμένα έδειξε αν αυτές οι μέθοδοι μπορούν να σχετίσουν αποτελεσματικά το φαρμακολογικό αποτέλεσμα, με το στόχο δράσης. Η αξία του εξετάζεται λεπτομέρως με τα δεδομένα της κλινικής φάσης ΙΙ και μελετάται κατά πόσο μπορεί να χρησιμοποιηθεί σε αποφάσεις που απαιτείται να ληφθούν για την κλινική αξιόγηση του φαρμάκου.

Μελέτη των ποσοτικών σχέσεων δομής - δράσης, για τη δημιουργία μοντέλων πρόβλεψης φυσικοχημικών και φαρμακοκινητικών παραμέτρων φαρμάκων

3. ΚΛΙΝΙΚΗ ΑΝΑΠΤΥΞΗ ΦΑΡΜΑΚΟΥ

3.1 Φάρμακο, ιστορική αναδρομή

Φάρμακο γενικά ονομάζεται κάθε χημική ουσία ή και μίγμα ουσιών, ικανή να επηρεάσει τη λειτουργία κάθε έμβιου όντος ή μικροοργανισμού όταν εισέλθει σε αυτόν. Απλούστερα ως φάρμακο χαρακτηρίζεται κάθε ουσία ή παρασκεύασμα, που χρησιμοποιείται για να ανακουφίσει ή να θεραπεύσει τον ανθρώπινο οργανισμό από ασθένειες ή πόνους και γενικότερα για να αποκαταστήσει την ανθρώπινη υγεία. Το φάρμακο μπορεί να χρησιμοποιηθεί στη διάγνωση, στη θεραπεία ή και στην πρόληψη μιας νόσου. Ο μηχανισμός δράσης του μπορεί να σχετίζεται με την αποκατάσταση, τη διόρθωση ή τη μεταβολή οργανικών λειτουργιών στον άνθρωπο ή στα ζώα.

Τα φάρμακα ταξινομούνται σε φυσικές και σε συνθετικές ενώσεις. Οι φυσικές μπορεί να προέρχονται από φυτά, από ζώα ή και από ορυκτά όπως είναι για παράδειγμα τα αλκαλοειδή (ατροπίνη), οι γλυκοσίδες (διγοξίνη), οι ορμόνες (ινσουλίνη), τα εμβόλια, ο σίδηρος και το ιώδιο. Συνθετικές είναι οι ενώσεις που παράγονται σε χημικά εργαστήρια όπως για παράδειγμα είναι η σιμετιδίνη και η ανιστρεπλάση, ή και τα νεότερα φάρμακα που συνθέτονται με την τεχνολογία του ανασυνδυασμένου DNA όπως είναι η ανθρώπινη ινσουλίνη.

Αρχικά τα φάρμακα είχαν μόνο φυτική προέλευση και η θεραπευτική γνώση είχε στηριχθεί κατά πολύ, στα έργα του Διοσκουρίδη και του Γαληνού που ήταν ανάμεσα στα πρώτα έργα που μεταφράστηκαν στα λατινικά κι εκτυπώθηκαν με την ανάκαλυψη της τυπογραφίας. Κατά τον 19ο αιώνα ξεκίνησε να παραγκωνίζεται η θεραπευτική και τα φάρμακα αυτών, καθώς άρχισε να αναπτύσσεται και να επικρατεί η σύγχρονη ιατρική και τα νέα χημικά φάρμακα.

Για παράδειγμα, το σαλικυλικό οξύ (ο πρόδρομος της ασπιρίνης) απομονώθηκε το 1874 από το φλοιό της ιτιάς. Διάφορα πιο ισχυρά παυσίπονα, όπως η μορφίνη και η κωδεΐνη απομονώθηκαν από την παπαρούνα, από τα φύλλα της κίγχονης απομονώθηκε η κινίνη που χρησιμοποιήθηκε κατά της ελονοσίας και από το φυτό digitalis purpurea απομονώθηκε η δακτυλίτιδα που χρησιμοποιήθηκε για τις καρδιακές παθήσεις.

Μετά την ανακάλυψη των διαφόρων συστατικών των δρογών δημιουργήθηκε το πρόβλημα της απομόνωσής τους σε μεγάλες ποσότητες και η σύνθεση είτε ουσιών όμοιων με τις φυσικές, είτε με παραπλήσια δομή και βελτιωμένες ιδιότητες π.χ. αντί της κοκαΐνης παρασκευάσθηκε συνθετικά η νοβοκαΐνη κλπ. Τα προβλήματα αυτά αντιμετωπίστηκαν από

τις φαρμακευτικές βιομηχανίες, που μέχρι τοτε παρασκεύαζαν μόνον ανόργανα φάρμακα.

Η ανακάλυψη των πρώτων συνθετικών φαρμάκων άρχισε το πρώτο μισό του 19ου αιώνα και αποτέλεσε τη βάση για την ανάπτυξη της ιατρικής ως προς τη συμπτωματική αντιμετώπιση των ασθενειών, αλλά και τη βάση της φαρμακευτικής βιομηχανίας. Στις αρχές του 19ου αιώνα η χημική έρευνα γινόταν κυρίως στα εργαστήρια των φαρμακείων, απ' όπου γεννήθηκαν πολλά γνωστά σήμερα φαρμακευτικά εργοστάσια.

Εξετάζοντας την ιστορία της Φαρμακευτικής Βιομηχανίας διακρίνουμε μια εξελικτική πορεία. Αρχίζοντας από την περίοδο της βοτανικής έρευνας, πέρασε στην περίοδο της τυποποίησης κατά το τελευταίο τέταρτο του 19ου αιώνα και κατέληξε από το τέλος του 19ου αιώνα στη χημική και βιολογική περίοδο.

Κατά τη δεύτερη εικοσαετία του 19ου αιώνα άρχισε στις βιομηχανίες η συνθετική παρασκευή χημικών ουσιών. Από τη λιθανθρακόπισσα παρασκευάσθηκαν το ανθρακένιο, η φαινόλη, η ανιλίνη και το βενζόλιο που αποτέλεσαν τις πρώτες ύλες για τη συνθετική παρασκευή φαρμάκων.

Από το 1920 έως το 1960 έγινε η ανακάλυψη της σημασίας των βιταμινών και της ελλειψής τους σε συγκεκριμένες ασθένειες. Οι περισσότερες βιταμίνες ανακαλύφθηκαν από τις πρώτες δεκαετίες του 20ου αιώνα και μέχρι τον Β' Παγκόσμιο Πόλεμο και η ανακάλυψή τους συνδέθηκε με αρκετά βραβεία Nobel. Η ανακάλυψη, η απομόνωση, η ταυτοποίηση, ο καθορισμός της βιολογικής δράσης και η σύνθεση των βιταμινών υπήρξε αντικείμενο πολυάριθμων μελετών επί πολλά χρόνια από ερευνητικές ομάδες τόσο στην Ευρώπη, όσο και στις ΗΠΑ. Παράλληλα έγινε η τυχαία σχεδόν ανακάλυψη της πενικιλλίνης από τον Φλέμινγκ, η ανακάλυψη της στρεπτομυκίνης, αλλά και του ημισυνθετικού αντιβιοτικού τετρακυκλίνη [46].

Το 1960 ο συνδυασμός γνώσεων από τη βιολογία και την αναλυτική χημεία, η γνώση του γενετικού υλικού και η ενίσχυση της τεχνολογίας της πληροφορικής, βοήθησε στον αυτοματισμό των εργαστηρίων και στην ανάπτυξη της επιστήμης, ως προς την ανακάλυψη νέων φαρμάκων, με έναν περισσότερο επιστημονικό τρόπο.

Το 1970 αναπτύχθηκε η γενετική μηχανική που μεταξύ άλλων έχει συνεισφέρει στην καταπολέμηση του καρκίνου, αλλά και ο κλάδος της βιοτεχνολογίας. Η τεχνολογία του ανασυνδυασμενου DNA και η μοριακή κλωνοποίηση είναι κάποιες από τις σημαντικότερες εξελίξεις της περιόδου.

Το 1980 βάσει των προηγούμενων επιστημονικών αλλά και τεχνολογικών εξελίξεων,

αναπτύχθηκε ιδιαίτερα ο τομέας της ανοσολογίας, με κύριο κίνητρο την εμφάνιση του ιού του AIDS. Στην ίδια δεκαετία η αλυσιδωτή αντίδραση πολυμεράσης οδήγησε σε σημαντικές προόδους στον τομέα της βιοτεχνολογίας, που είχαν αποφασιστικό αντίκτυπο στην ανακάλυψη νέων φαρμάκων. Πολύ σημαντικό ρόλο επίσης στην ανακάλυψη και το σχεδιασμό νεών φαρμάκων, έπαιξε η μοριακή βιολογία και η χρήση της τεχνολογίας των ηλεκτρονικών υπολογιστών. Η ανακάλυψη νέων δραστικών ενώσεων, η μελέτη οργανικών ενώσεων ως προς τη βιολογική τους δράση καθώς και η εμπορευματοποίηση της ανακάλυψης του φαρμάκου, έδωσαν κίνητρο τόσο για τη δημιουργία μικρών εταιρειών βιοτεχνολογίας όσο και για τη μαζική παραγωγή φαρμάκων από τις μεγάλες φαρμακευτικές εταιρείες.

Το 1990 άρχισε να γίνεται χρήση της ρομποτικής και του αυτοματισμού. Επιστήμονες από διάφορες ειδικότητες, συνεισφέρουν στη διαμόρφωση των κύριων χαρακτηριστικών αυτής της δεκαετίας που είναι η ορθολογική οργάνωση της διαδικασίας ανακάλυψης των φαρμάκων και η άμεση σύνδεσή της με οικονομικά οφέλη. Επιπλέον οι νέες τεχνολογίες, όπως η μοριακή βιολογία και η βιοπληροφορική, ενισχύουν την ανακάλυψη και το σχεδιασμό των νέων φαρμάκων.

Η εξέλιξη της χημείας και η σύνθεση νέων φαρμάκων και εμβολίων ήταν πολύ σημαντικοί παράγοντες για την ανάπτυξη της φαρμακευτικής βιομηχανίας. Ο συνδυασμός επίσης των γνώσεων της βιολογίας, της φυσικής και της χημείας συνετέλεσαν στην προώθηση της φαρμακευτικής επιστήμης.

Στη συνέχεια με την ανακάλυψη της δομής του DNA από τους Watson και Crick, οδηγηθήκαμε στην ανακάλυψη νέων φαρμάκων κι εφαρμογών με κυρίαρχο παράδειγμα τα αντιιίκα. Σε αυτή τη φάση αναπτύχθηκε ο τομέας της βιοτεχνολογίας που συνδύασε την εκμετάλλευση του μεγάλου πια όγκου των γνώσεων της βιολογίας και της χημείας με την ανάπτυξη αρκετά εξειδικευμένου τεχνολογικού εξοπλισμού. Εξέλιξη στη θεραπευτική του 20ου αιώνα αποτέλεσε και η χημειοθεραπεία κατά του καρκίνου. Μέχρι και τον 19ο αιώνα για την αντιμετώπιση των νεοπλασιών χρησιμοποιούντο παράγωγα του αρσενικού και μυστήρια φάρμακα χωρίς αποτελεσματικότητα. Κατά τον Α' Παγκόσμιο πόλεμο διαπιστώθηκε ότι ο θειοϋπερίτης (ή αέριο της μουστάρδας), που χρησιμοποιήθηκε ως τοξικό αέριο, είναι δραστικός σε ορισμένες περιπτώσεις καρκίνου του δέρματος. Λόγω όμως της μεγάλης τοξικότητάς του, αποκλείσθηκε από τη θεραπευτική. Στην αρχή του Β' Παγκοσμίου πολέμου στα πλαίσια μελέτης χημικών όπλων, βρέθηκε ότι οι μουστάρδες του αζώτου ή αζωθυπερίτες μπορούν να χρησιμοποιηθούν σε ορισμένες μορφές καρκίνου, διευκρινίσθηκε δε ότι δρουν ως αλκυλιωτικοί παράγοντες.

Μέχρι το τέλος του 19ου αιώνα, ο αριθμός των απομονωμένων αλκαλοειδών ήταν μεγαλύτερος από 100 που είναι κυρίως βιοκατευθυνόμενοι και έχουν οδηγήσει σε φαρμακευτικά σκευάσματα με δραστικά συστατικά είτε φυσικά προϊόντα, είτε ημισυνθετικά παράγωγα, είτε συνθετικά προϊόντα των οποίων το φαρμακοφόρο τμήμα είναι βασισμένο σε κάποιο φυσικό προϊόν. Παρά το γεγονός ότι τα σύγχρονα φάρμακα κινούνται μακριά από τα φυσικά προϊόντα αυτά καθαυτά, φαίνεται ότι κατά τα τελευταία 30 χρόνια ακόμα και στην εποχή της συνδυαστικής χημείας οι δομές κάποιων φυσικών προϊόντων ή οι τροποποιημένες δομές, εξακολουθούν να είναι σημαντικές για την ανακάλυψη φαρμάκων έναντι ποικίλων ασθενειών [20]. Έκτοτε μέχρι και σήμερα, πολλαπλασιάστηκαν οι φυτοχημικές έρευνες.

Σήμερα οι φαρμακοβιομηχανίες παρέχουν πλήθος φαρμακευτικών προϊόντων, των οποίων η έρευνα συμπληρώνεται από φαρμακολογικές, τοξικολογικές μελέτες, καθώς και κλινικές δοκιμασίες. Στόχος η ελαχιστοποίηση των ανεπιθύμητων ενεργειών και η εξειδίκευση της δράσης τους. Κατά τις τελευταίες δεκαετίες παρασκευάζονται φάρμακα (π.χ. ανθρώπινη ινσουλίνη) και εμβόλια (π.χ. της λύσσας), που είναι προϊόντα βιοτεχνολογικών μεθόδων.

Τον 20ο αιώνα η ιατρική επιστήμη, ιδιαίτερα στις δύο τελευταίες δεκαετίες έκανε τεράστια πρόοδο. Ο συνδυασμός πληροφορικής και γνώσης της ανθρώπινης βιολογίας δημιούργησε μία πληθώρα γνώσεων και πόρων, η αξιοποίηση των οποίων έγινε μείζον θέμα για τις φαρμακευτικές εταιρίες.

Η έναρξη της νέας χιλιετίας ήδη χαρακτηρίζεται από τεράστια πρόοδο στην ανακάλυψη νέων φαρμάκων που βασίζονται σε ”state-of-the art” της χημείας και της γενετικής. Σηματοδοτήθηκε και από την εφαρμογή της κυτταρικής και γονιδιακής θεραπείας [31]. Έχει ξεκινήσει να μελετάται και να εφαρμόζεται η εξατομικευμένη ιατρική πρακτική και φαρμακευτική αγωγή. Τα τελευταία χρόνια έως και σήμερα οι παραδοσιακοί επιστημονικοί κλάδοι της φαρμακευτικής εμπλουτίστηκαν από τις εξελίξεις των γονιδιωματικών τεχνολογιών. Αυτό είχε ως αποτέλεσμα την εκμετάλλευση αυτών των γνώσεων, για το σχεδιασμό και την ανάπτυξη των νέων φαρμάκων, αλλά και τη δημιουργία νέων κατευθύνσεων στη φαρμακευτική έρευνα όπως είναι η φαρμακογονιδιωματική και η φαρμακευτική βιοτεχνολογία.

Η φαρμακευτική βιοτεχνολογία για παράδειγμα, έχει προσφέρει τα τελευταία χρόνια νέα δεδομένα σε όλους σχεδόν τους τομείς του φαρμάκου (σχεδίαση, ανάπτυξη, μορφοποίηση, ανάλυση, χορήγηση) με συνέπεια την εκμετάλλευσή τους για πιο πρωτοποριακά,

εξειδικευμένα, αποτελεσματικότερα και ασφαλέστερα φάρμακα. Ιστορικά, τα βιοφάρμακα πρωτεϊνικής φύσης κατέχουν την πρώτη θέση σ' αυτόν τον κατάλογο, ενώ η εξέλιξη που επιτελείται στο επίπεδο της γονιδιακής θεραπείας, της χρήσης των "μικροπλακών DNA" (DNA chips) για τον έλεγχο της γονιδιακής έκφρασης στους διάφορους οργανισμούς, της ανάπτυξης και εφαρμογής πρωτοκόλλων κυτταρικών θεραπειών, καθώς επίσης και της χρησιμοποίησης καινοτόμων θεραπευτικών στοχευμένης φαρμακολογικής δράσης συμπεριλαμβανομένων των ολιγονουκλεοτιδίων και των ριβοζυμών (μικρά μόρια RNA με καταλυτικές ιδιότητες) για θεραπευτικούς σκοπούς, ανοίγει νέους ορίζοντες στο φαρμακευτικό και ιατρικό χώρο.

Επιπρόσθετα, οι βασικές αρχές που διέπουν σήμερα την επιλογή φαρμάκων στη σύγχρονη θεραπευτική στηρίζονται στις υπάρχουσες γνώσεις της μοριακής φαρμακολογίας και παθοφυσιολογίας για τη δράση των φαρμάκων στον οργανισμό ενώ, με την ανάπτυξη της φαρμακογονιδιωματικής, δίνεται η δυνατότητα ανάλυσης του φαινομένου της διαφορετικής φαρμακολογικής απόκρισης μεταξύ των ασθενών, της εμφάνισης ανεπιθύμητων ενεργειών (ADRs) και αλληλεπιδράσεων των φαρμάκων, καθώς επίσης και της συσχέτισης του γενετικού πολυμορφισμού συγκεκριμένων γονιδίων με τη δράση των φαρμάκων στον οργανισμό. Ένα πρόσθετο στοιχείο αφορά τις σύγχρονες τεχνολογικές εξελίξεις στη μορφοποίηση των φαρμάκων σε νέες φαρμακοτεχνικές μορφές ελεγχόμενης αποδέσμευσης, μεταφοράς και κατευθυνόμενης χορήγησης, που συνεισφέρουν σημαντικά στην αποτελεσματικότητα και στην ασφάλεια των φαρμάκων στον οργανισμό.

Αν ληφθούν επίσης υπόψη οι νέες δυνατότητες που παρέχονται από τους τομείς της συνδυαστικής και υπολογιστικής χημείας, της ανάπτυξης των καινοτόμων ελέγχων υψηλής απόδοσης (high-throughputscreening), καθώς επίσης και των νέων μεθόδων μοριακής διάγνωσης (π.χ. γενετικής, απεικονιστικής) και ανάπτυξης βιοδεικτών, τότε μπορεί εύκολα να γίνει αντιληπτό το διαφαινόμενο πλαίσιο της φαρμακευτικής - ιατρικής εκπαίδευσης και έρευνας στον 21ο αιώνα. Σε τελευταία ανάλυση, οι εξελίξεις αυτές σηματοδοτούν τη νέα εποχή και στο χώρο της φαρμακολογίας και της θεραπευτικής γενικότερα, δίνοντας μια νέα διάσταση στη φαρμακευτική αγωγή με τη δυνατότητα εξατομίκευσης σε ευρεία κλίμακα των δοσολογικών σχημάτων στην κλινική πράξη. Η κατεύθυνση - εξέλιξη αυτή στη στιγμιογράφηση φαρμάκων στην καθημερινή ιατρική πρακτική αναφέρεται ως φαρμακοτυπία (pharmacotyping), ("κατά το πρότυπο" genotyping/haplotyping), επιστημονικός όρος που εισήχθη στη διεθνή βιβλιογραφία το 2004. Οι πρόσφατες εξελίξεις στο χώρο της φαρμακευτικής έρευνας με την ενσωμάτωση των γονιδιωματικών και άλλων τεχνολογιών στην ανάπτυξη νέων φαρμάκων επιδρούν θετικά και στη βελτίωση της φαρ-

μακευτικής αγωγής στην κλινική πράξη. Παράλληλα, αναδεικνύονται νέες επιστημονικές κατευθύνσεις στη φαρμακευτική έρευνα, όπως η φαρμακευτική βιοτεχνολογία, και η φαρμακογονιδιωματική, ενώ οι μοριακές μεθοδολογίες σε συνδυασμό με τη νανοτεχνολογία εμπλουτίζουν τη φαρμακευτική τεχνολογία προς καινοτόμους φορείς κατά τη μορφοποίηση φαρμάκων. Η φαρμακογονιδιωματική, συνδυάζοντας νέες τεχνολογίες και επιστημονικές γνώσεις, προσφέρει νέες ευκαιρίες τόσο στην ανάλυση του μηχανισμού δράσης των φαρμάκων σε μοριακό επίπεδο, όσο και στη χορήγηση των φαρμάκων στη θεραπευτική. Ουσιαστικά, συνδέει τη φαρμακολογία με τους νέους τομείς της γονιδιωματικής, της βιοπληροφορικής, και της πρωτεϊνωματικής. Ο τελικός στόχος αυτής της προσέγγισης είναι η εξατομίκευση των δοσολογικών σχημάτων μέσα από τη διευκρίνιση και την κατανόηση των μοριακών μηχανισμών που οδηγούν σε διαφορετικό φαρμακολογικό αποτέλεσμα και την εμφάνιση των ανεπιθύμητων ενεργειών στην κλινική πράξη.

Για να υπάρξει όμως η πλήρης ενσωμάτωση κι εκμετάλλευση των γενετικών γνώσεων που αφορούν τα φάρμακα στη θεραπευτική, θα πρέπει πρώτα να καθιερωθούν κατάλληλες συνθήκες λειτουργίας στο υγειονομικό σύστημα, να διευκρινιστούν νομικά και ηθικά ζητήματα, να εκπαιδευτεί κατάλληλα το ιατρικό, φαρμακευτικό και νοσηλευτικό προσωπικό και τέλος, να ενημερωθεί η κοινή γνώμη για τις επιπτώσεις των γενετικών ελέγχων στη φαρμακευτική αγωγή και στην ιατρική πρακτική. Μέσα σ' αυτό το πλαίσιο, οι εξελίξεις αυτές καθορίζουν, ως ένα βαθμό, και το ρόλο του φαρμακοποιού στο νέο περιβάλλον της φαρμακευτικής αγωγής που διαμορφώνεται για τα χρόνια που έρχονται. Η φαρμακογονιδιωματική συνδυάζοντας νέες τεχνολογίες και επιστημονικές γνώσεις προσφέρει νέες δυνατότητες ανάλυσης του μηχανισμού δράσης των φαρμάκων σε μοριακό επίπεδο. Στοχεύει έτσι στην ασφαλέστερη χορήγηση των φαρμάκων για την επίτευξη του μέγιστου φαρμακολογικού αποτελέσματος και ασφάλειας. Με αυτόν τον τρόπο αναμένεται να ελαχιστοποιηθεί και η πιθανότητα εμφάνισης ανεπιθύμητων ενεργειών (ADRs) λόγω φαρμακευτικής αλληλεπίδρασης, αφού η εμφάνιση των ADRs αποτελεί έναν αρνητικό παράγοντα για την επίτευξη του βέλτιστου θεραπευτικού αποτελέσματος. Το τελευταίο, είναι ιδιαίτερα σημαντικό στις μέρες μας όπου η συγχορήγηση αρκετών φαρμάκων είναι η καθιερωμένη ιατρική πρακτική στη στιγμιογράφηση φαρμάκων, με αποτέλεσμα την αυξημένη πιθανότητα ανάπτυξης ADRs, εξαιτίας των φαρμακευτικών αλληλεπιδράσεων.

Πράγματι, η ανάλυση των δεδομένων φαρμακοεπαγρύπνησης δείχνει ότι αρκετές ADRs αφορούν αλληλεπιδράσεις μεταξύ φαρμάκων, ενώ τα τελευταία χρόνια και οι αλληλεπιδράσεις φαρμάκων - τροφίμων ή φαρμάκων - φυτοθεραπευτικών, αποκτούν ιδιαίτερο ενδιαφέρον. Η φαρμακογονιδιωματική προσέγγιση έχει σημαντικά επηρεάσει τον τομέα της

ανάπτυξης νέων φαρμάκων, ενώ και η εφαρμογή των μοριακών διαγνωστικών επιτρέπει την εξατομίκευση της παρεχόμενης περίθαλψης που αναμφισβήτητα οδηγεί και στη βελτίωση της φαρμακευτικής αγωγής. Βασικό συστατικό όμως επίτευξης αυτού του στόχου αποτελεί η δημιουργία της κατάλληλης υποδομής και η εκπαίδευση του ανθρώπινου δυναμικού μέσα από την αναβάθμιση και την ανάπτυξη νέας εκπαιδευτικής μεθοδολογίας στην οργάνωση των προγραμμάτων σπουδών. Με αυτόν τον τρόπο η κλινική αξιολόγηση και η αξιοποίηση των φαρμακογονιδιωματικών πληροφοριών μπορεί να επιτευχθεί και να οδηγήσει σε βελτιωμένα πρωτόκολλα φαρμακευτικής αγωγής, στοιχείο σημαντικό για το σύστημα περίθαλψης και την ποιότητα ζωής των ασθενών γενικότερα.

Η πρόσφατη έγκριση από τον Αμερικανικό Οργανισμό Φαρμάκων και Τροφίμων (FDA), αλλά και από τον αντίστοιχο Ευρωπαϊκό (EMA), των πρώτων φαρμακογονιδιωματικών τεστ καθώς και η δημοσιοποίηση της οδηγίας για τη βιομηχανία σχετικά με την κατάθεση των φαρμακογονιδιωματικών δεδομένων για την ανάπτυξη νέων φαρμάκων, δείχνουν εύγλωττα αυτήν τη νέα στροφή που έχει ήδη υπάρξει στη φαρμακευτική έρευνα και περίθαλψη. Οι εξελίξεις αυτές σηματοδοτούν μια νέα εποχή στο χώρο της φαρμακολογίας και της θεραπευτικής, δίνοντας μια νέα διάσταση στη φαρμακευτική αγωγή με την εξατομίκευση των δοσολογικών σχημάτων στην κλινική πράξη. Αυτό το γεγονός συνακόλουθα, επηρεάζει τόσο τον τομέα φαρμακοεπαγρύπνησης (καταγραφή, ανάλυση, αξιοποίηση των ADRs μετά την έγκριση κυκλοφορίας των φαρμάκων) και τη δομή - οργάνωση των συστημάτων περίθαλψης, όσο και την εκπαίδευση των υγειονομικών σ’ αυτές τις σύγχρονες τάσεις [41].

Με τη ραγδαία ανάπτυξη του τομέα της πληροφορικής, η χρήση αυτού γίνεται όλο και πιο σημαντική στην ανακάλυψη νέων φαρμακευτικών ενώσεων. Είναι πια εφικτή η έγκαιρη απόρριψη των ακατάλληλων ενώσεων στις πρώτες φάσεις σχεδιασμού, με αποτέλεσμα την πολύ σημαντική μείωση του κόστους καθώς και του χρόνου ανάπτυξης και παραγωγής ενός φαρμάκου.

3.2 Φαρμακοκινητική - Φαρμακοδυναμική

Οι κύριοι παράγοντες που επηρεάζουν τις ενέργειες και κατά συνέπεια τη δράση μιας φαρμακευτικής ουσίας είναι οι εξής :

- Χημική δομή
- Οδός χορήγησης και φαρμακοτεχνική μορφή

- Βάρος του σώματος
- Ηλικία
- Φύλο
- Ατομική ευαισθησία
- Ανάπτυξη αντοχής
- Ανάπτυξη αντίστασης
- Ψυχολογικοί παράγοντες
- Παθολογικες καταστάσεις
- Παρουσία άλλων φαρμάκων

Στην εργασία αυτή εξετάσθηκε και κατά πόσο τα δεδομένα για την οδό χορήγησης ενός φαρμάκου μπορούν να επηρεάσουν την απόδοση των μοντέλων πρόβλεψης, για το χρόνο ημίσειας ζωής αυτού.

Η χορήγηση ενός φαρμάκου μπορεί να γίνει με διάφορους τρόπους όπως από το στόμα(peros), ενδομυϊκά (im), ενδοφλέβια (iv), υποδόρια (sc), διαδερμικά, υπογλώσσια, εισπνεόμενα, από το ορθό, κλπ. Ο τρόπος χορήγησης ενός φαρμάκου παίζει πολύ σημαντικό ρόλο στον τρόπο απελευθέρωσης της δραστικής ουσίας στον οργανισμό, στην κατανομή της, στο μεταβολισμό της καθώς και στη μεταφορά της στο σημείο που αποτελεί το βιολογικό της στόχο.

Ο πλέον αποδεκτός τρόπος λήψης φαρμάκων, είναι η από του στόματος χορήγηση, ο οποίος όμως είναι και ο πιο πολύπλοκος για τη μελέτη ενός νέου φαρμάκου. Κατά τη λήψη από το στόμα, παρατηρείται το φαινόμενο πρώτης διόδου (first-pass effect) όπου μια σημαντική ποσότητα του χορηγούμενου φαρμάκου εξουδετερώνεται από τα ένζυμα του στομάχου, από βακτηρίδια και ένζυμα του εντέρου και από τα ηπατικά κύτταρα. Η απορρόφηση των δραστικών ουσιών των φαρμακευτικών σκευασμάτων που χορηγούνται από το στόμα ή από το ορθό, πραγματοποιείται ως επί το πλείστον στο έντερο για αυτό και ονομάζεται εντερική, ενώ παρεντερική είναι η απορρόφηση όταν πρόκειται για ενδοφλέβια χορήγηση, ενδομυϊκή και γενικά για κάποια οδό χορήγησης που παρακάμπτει το γαστρεντερικό σωλήνα. Κατά την peros χορήγηση τα φάρμακα περνούν από τον οισοφάγο και εισέρχονται στο στομάχι. Η διάσπαση του φαρμάκου και η απελευθέρωση για

την απορρόφηση της δραστικής ουσίας, ξεκινά με τη διαδικασία της πέψης κι εξαρτάται από το pH του στομάχου και την παρουσία ή όχι τροφής. Στο λεπτό έντερο συνεχίζεται η χημική διαδικασία της πέψης που ξεκίνησε στο στομάχι. Το λεπτό έντερο λόγω του πιο υψηλού pH σε σύγκριση με αυτό του στομάχου, της μεγαλύτερης επιφάνειάς του, του μεγαλύτερου χρόνου διέλευσης του φαρμάκου από αυτό και της υφής του, δίνει περισσότερο χρόνο στις ουσίες και κατά συνέπεια και περισσότερες ευκαιρίες για απορρόφηση αυτών από τον οργανισμό.

Στη συνέχεια τα φάρμακα εισέρχονται στο ήπαρ. Η πυλαία φλέβα είναι ένα σύνολο αιμοφόρων αγγείων που συλλέγει το αίμα από το στομάχι, το λεπτό έντερο, το παχύ έντερο, τη σπλήνα και το πάγκρεας και το παραδίδει στο ήπαρ. Από το ήπαρ, το αίμα εισέρχεται στο γενικό κυκλοφορικό σύστημα. Τίποτα δεν μπορεί να εισέλθει στην κυκλοφορία τουαίματος από το πεπτικό σύστημα, αν δε διέλθει πρώτα από το ήπαρ. Ένα φάρμακο μπορεί να απορροφηθεί από το λεπτό έντερο, αλλά να μεταβολιστεί σχεδόν ποσοτικά από το ήπαρ. Στο ήπαρ εντοπίζεται κυρίως το φαινόμενο πρώτης διόδου που ήδη αναφέραμε και φάρμακα επιρρεπή στο μεταβολισμό πρώτης διόδου, εμφανίζουν χαμηλή βιοδιαθεσιμότητα όταν λαμβάνονται από το στόμα. Η χαμηλή βιοδιαθεσιμότητα δείχνει ότι το φάρμακο είτε απορροφάται ελάχιστα από το γαστρεντερικό σωλήνα, είτε υπόκειται σε υψηλό μεταβολισμό πρώτης διόδου.

Η ενέσιμη χορήγηση ενός φαρμάκου είναι μια μέθοδος παρεντερικής χορήγησης που παρακάμπτει τα προβλήματα που συνδέονται με το φαινόμενο πρώτης διόδου και την απορρόφηση μέσω του πεπτικού συστήματος. Η ενέσιμη χορήγηση επιταχύνει την απορρόφηση και την άφιξη του φαρμάκου στα σημεία δράσης, αλλά επίσης παρέχει μεγαλύτερη ακρίβεια στη δοσολογία. Η ενδοφλέβια χορήγηση παρουσιάζει την πλέον άμεση απορρόφηση μέσα στο κυκλοφορικό σύστημα. Όπως είδαμε προηγουμένως υπάρχουν πολλές μεταβλητές που εμπλέκονται στην απορρόφηση και την κατανομή ενός φαρμάκου, και συχνά είναι δύσκολο να εκτιμηθεί η σωστή δοσολογία για ένα φάρμακο, δηλαδή η ποσότητα του φαρμάκου που χορηγείται σε κάθε δόση, όπως και η συχνότητα χορήγησής του. Στην ιδανική περίπτωση, τα επίπεδα κάθε φαρμάκου στο αίμα θα πρέπει να είναι σταθερά κι ελεγχόμενα όπως συμβαίνει στη χορήγηση με συνεχή ενδοφλέβια έγχυση. Ασφαλώς η συνεχής ενδοφλέβια χορήγηση είναι σαφώς ανέφικτη για τα περισσότερα φάρμακα. Έτσι τα φάρμακα συνήθως λαμβάνονται σε τακτά χρονικά διαστήματα και οι δόσεις που λαμβάνονται έχουν σχεδιαστεί έτσι ώστε να διατηρούν τα επίπεδα του φαρμάκου στην κυκλοφορία του αίματος ανάμεσα σε ένα μέγιστο και ένα ελάχιστο επίπεδο. Η συγκέντρωση του φαρμάκου πρέπει να μην είναι πάρα πολύ υψηλή και επομένως τοξική, αλλά ούτε πάρα πολύ

χαμηλή για να μην είναι αναποτελεσματικό, δηλαδή πρέπει να είναι μέσα στα επίπεδα που ορίζει το αντίστοιχο θεραπευτικό παράθυρο. Σε γενικές γραμμές, η συγκέντρωση του ελεύθερου φαρμάκου στο αίμα (δηλαδή του φαρμάκου που δεν είναι δεσμευμένο από τις πρωτεΐνες του πλάσματος) είναι μια καλή ένδειξη της διαθεσιμότητάς του στη θέση στόχου του. Αυτό δε σημαίνει ότι η συγκέντρωση στο αίμα είναι ίδια με τη συγκέντρωση στη θέση στόχο. Ωστόσο, οποιεσδήποτε διακυμάνσεις στη συγκέντρωση στο αίμα θα οδηγήσουν σε παρόμοιες διακυμάνσεις στη θέση στόχο. Έτσι, τα επίπεδα συγκέντρωσης στο αίμα μπορούν να χρησιμοποιηθούν για τον προσδιορισμό θεραπευτικών και ασφαλών επιπέδων δόσης για ένα φάρμακο, [43].

Ο χρόνος ημίσειας ζωής είναι ανάλογος του όγκου κατανομής V του φαρμάκου και αντιστρόφως ανάλογως της κάθαρσης CL αυτού από τον οργανισμό, Εξίσωση 3.1.

$$t_{\frac{1}{2}} = \frac{0.693 * V}{CL} \tag{3.1}$$

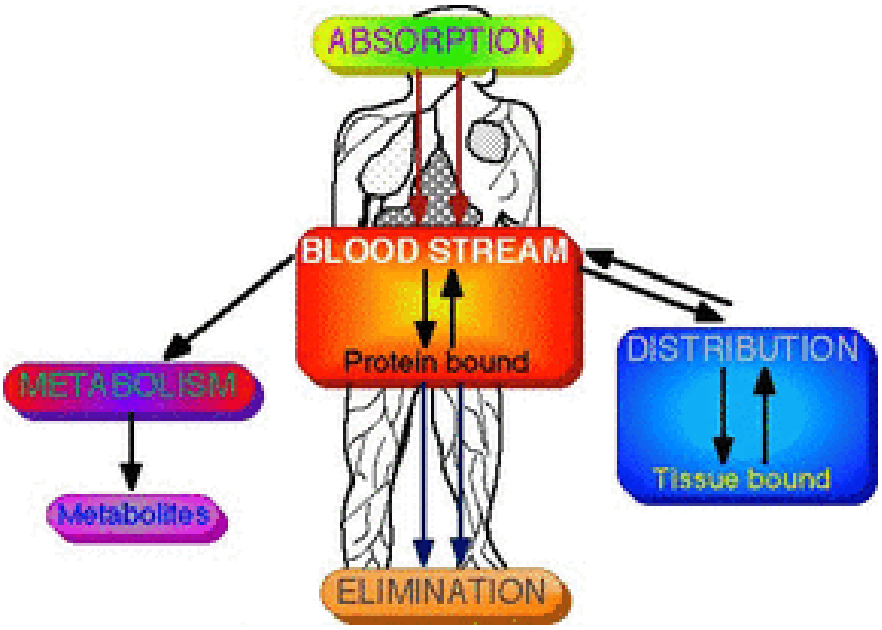
Οι φαρμακοκινητικές αυτές ιδιότητες είναι πιο εύκολο να υπολογιστούν και να εκτιμηθούν σωστά κατά την ενδοφλέβια χορήγηση, καθώς όπως αναφέρθηκε παρακάμπτεται μέσω της παρεντερικής χορήγησης ο μεταβολισμός πρώτης δίοδου του φαρμάκου.

Για να μελετήσουμε τη δράση ενός φαρμάκου στηριζόμαστε στη φαρμακοκινητική και τη φαρμακοδυναμική αυτού.

Η φαρμακοκινητική είναι ο κλάδος της φαρμακολογίας που μελετά τη χρονική εξέλιξη των ποσοτικών και ποιοτικών μεταβολών των φαρμάκων στον οργανισμό και μελετάται συνήθως από τις εξής ιδιότητες (ADME), Σχήμα 3.1 :

- Απορρόφηση (Absorption)
- Κατανομή (Distribution)
- Μεταβολισμός (Metabolism)
- Απομάκρυνση (Excretion)

Τα φάρμακα είναι μόρια που αντιδρούν με άλλα (μακρο)μόρια και το αποτέλεσμα αυτής της αντίδρασης - αλληλεπίδρασης, είναι μια βιολογική απόκριση του οργανισμού. Η φαρμακοδυναμική είναι ο κλάδος της φαρμακολογίας που μελετά τη σχέση ανάμεσα στη



Σχήμα 3.1: ADME

δόση - συγκέντρωση ενός φαρμάκου και τη φαρμακολογική, τοξικολογική ή κλινική απόκριση του οργανισμού. Μελετά την επίδραση του φαρμάκου στον οργανισμό γενικότερα αλλά και ειδικότερα στο μόριο, στο κύτταρο, στον ιστό ή στο όργανο.

Στην κλινική πράξη σημαντικές φαρμακοκινητικές παράμετροι, θεωρούνται η βιοδιαθεσιμότητα, ο όγκος κατανομής, η κάθαρση και ο χρόνος ημίσειας ζωής.

Βιοδιαθεσιμότητα είναι το ποσοστό του χορηγούμενου φαρμάκου που φτάνει αναλλοίωτο στη συστημική κυκλοφορία. Η βιοδιαθεσιμότητα ενός φαρμάκου εξαρτάται τόσο από την απορρόφηση η οποία είναι συνάρτηση της οδού χορήγησης, των φυσικοχημικών ιδιοτήτων του φαρμάκου, των φυσικών και βιοχημικών παραμέτρων όσο και του μεταβολισμού πρώτης διόδου.

Επειδή η ποσότητα του φαρμάκου στο σώμα δεν είναι άμεσα μετρήσιμη, γίνεται μέτρηση της συγκέντρωσής του στο πλάσμα ή στο αίμα. Ο φαινόμενος, **όγκος κατανομής** ισούται με τον όγκο που καταλαμβάνει το φάρμακο με συγκέντρωση, ίση με τη συγκέντρωσή του στο πλάσμα. Είναι πολύ σημαντικός και για τον υπολογισμό της δόσης του φαρμάκου που απαιτείται για την επίτευξη μιας συγκεκριμένης συγκέντρωσης στο πλάσμα.

Κάθαρση καλείται ο όγκος του πλάσματος που καθαρίζεται από το φάρμακο στη μονάδα του χρόνου. Οι μονάδες μέτρησης της κάθαρσης είναι ml/min ή L/h. Η κάθαρση (CL)

είναι μια πολύ χρήσιμη φαρμακοκινητική παράμετρος στη θεραπευτική, καθώς και στην αποτίμηση των μηχανισμών απομάκρυνσης των φαρμάκων από το σώμα. Όπως ο όγκος κατανομής συσχετίζει τη συγκέντρωση του φαρμάκου στο πλάσμα με την ποσότητά του στο σώμα, έτσι και η κάθαρση συσχετίζει τη συγκέντρωση του φαρμάκου στο πλάσμα με το ρυθμό απομάκρυνσής του από το σώμα.

Ο **χρόνος ημίσειας ζωής** είναι ο χρόνος που απαιτείται για να μεταβληθεί η συγκέντρωση του φαρμάκου κατά 50 %. Ο χρόνος ημίσειας ζωή είναι αντιστρόφως ανάλογος προς την κάθαρση και ευθέως ανάλογος προς τον όγκο κατανομής. Υπάρχουν κάποιες κλινικές καταστάσεις που συνεπάγονται αύξηση ή μεταβολή του χρόνου ημίσειας ζωή. Όταν ένας ασθενής έχει μια διαταραχή που μεταβάλλει τον χρόνο ημίσειας ζωής του φαρμάκου, απαιτείται προσαρμογή της δοσολογίας.

Σε κλινικές περιπτώσεις κατά τις οποίες προκαλείται αύξηση του χρόνου ημίσειας ζωής, η ροή του πλάσματος στους νεφρούς είναι ελαττωμένη. Αυτό για παράδειγμα συμβαίνει σε περιπτώσεις καρδιακής ανεπάρκειας, αιμορραγίας ή και νεφρικής νόσου. Αύξηση του χρόνου ημίσειας ζωής μπορεί να παρατηρηθεί και με ταυτόχρονη χορήγηση δεύτερου φαρμάκου που είτε συνδέεται και αυτό με την αλβουμίνη οπότε αυξάνεται η βιοδιαθεσιμότητα του πρώτου φαρμάκου, είτε μπορεί να μειώσει τον μεταβολισμό του φαρμάκου.

Όταν ένα φάρμακο έχει μικρό χρόνο ημίσειας ζωής, αλλά απαιτείται η παράταση της δράσης του, γίνεται συχνή χορήγηση αυτού, με μεσοδιαστήματα πολύ μικρότερα του χρόνου ημιζωής. Απαιτείται ιδιαίτερη προσοχή σε φάρμακα με μεγάλο χρόνο ημίσειας ζωής ή εξαιρετικά λιποδιαλυτά φάρμακα των οποίων η συχνή χορήγηση οδηγεί σε άθροισή τους στον οργανισμό και ίσως σε τοξικές συνέπειες.

Άλλη τεχνική για αύξηση του χρόνου ημίσειας ζωής, είναι η επιβράδυνση της απορρόφησής του από τον οργανισμό. Αυτό μπορεί να γίνει π.χ. με τροποποιήσεις της φαρμακοτεχνικής μορφής, όπως είναι το εντερικό περίβλημα. Η επιβράδυνση της απορρόφησης μπορεί επίσης να επιτευχθεί με ταυτόχρονη χορήγηση ενός δεύτερου φαρμάκου ή ουσίας που ουσιαστικά παρεμποδίζει την απέκκριση του πρώτου φαρμάκου ή αναστέλει τον μεταβολισμό του, με αποτέλεσμα την παράταση της δράσης του.

3.3 Στάδια έρευνας και ανάπτυξης των φαρμακευτικών προϊόντων

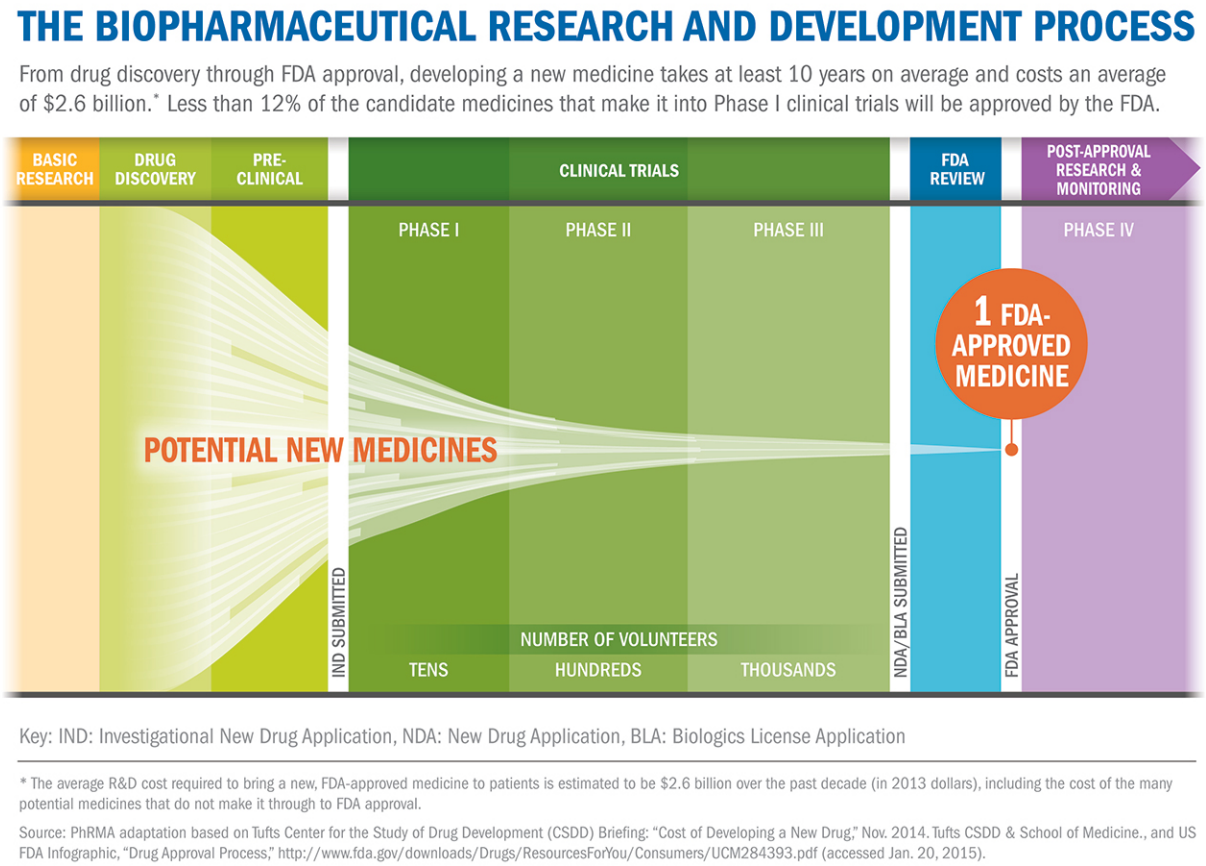
Τα στάδια της έρευνας και ανάπτυξης των φαρμακευτικών προϊόντων είναι τα εξής :

- Πνευματική ιδιοκτησία

- Βασική έρευνα
- Προκλινική έρευνα in vitro/in vivo (Good Laboratory Practice)
- Βιομηχανική παραγωγή δοκιμαστικής παρτίδας (Good Manufacturing Practice)
- Βιομηχανική παραγωγή μεγάλης κλίμακας (Good Manufacturing Practice)
- Κλινική έρευνα: κλινικές μελέτες Φάσεων I-III (Good Clinical Practice)
- Έγκριση
- Marketing
- Εμπόριο
- Ανάπτυξη μετά την έγκριση : μετεγκριτικές κλινικές μελέτες, μελέτες Φάσης IV, παρατήρησης, επιδημιολογικές
- Φαρμακοεπαγρύπνηση

Στο Σχήμα 3.2, απεικονίζονται τα στάδια της διαδικασίας παραγωγής ενός φαρμάκου. Το πρώτο στάδιο είναι η έρευνα που γίνεται στα εργαστήρια για τη σύνθεση του φαρμάκου και την εύρεση του κατάλληλου στόχου. Το δεύτερο στάδιο αφορά στις κλινικές μελέτες και χωρίζεται σε τρεις επιμέρους φάσεις. Το τρίτο στάδιο σχετίζεται με την έγκριση του μοναδικού πια φαρμάκου από τον FDA, ενώ ακολουθεί η τέταρτη φάση που αφορά στον έλεγχο της δράσης του φαρμάκου, αφού αυτό παραχθεί.

Στο Σχήμα 3.3, φαίνεται τόσο ο τεράστιος αριθμός ουσιών με τον οποίο ξεκινά εργαστηριακά η έρευνα για την παρασκευη ενός φαρμάκου, όσο και τα χρόνια που απαιτούνται από το αρχικό στάδιο έρευνας εώς την παραγωγή και εν τέλει την κυκλοφορία του φαρμάκου. Το χρονικό διάστημα αυτό κυμαίνεται από 10 εώς και 15 χρόνια. Πολύ σημάντικη είναι και η φάση μετά την κυκλοφορία του φαρμάκου, κατά την οποία παράλληλα με την κυκλοφορία διεξάγονται μετεγκριτικές κλινικές μελέτες για να επιβεβαιώσουν τόσο την δράση όσο και την ασφάλεια του φαρμάκου (επιδημιολογικές, μελέτες παρατήρησης, μελέτες φάσης IV κτλ.). Στη συνέχεια ακολουθεί η φαρμακοεπαγρύπνηση, διαδικασία που αφορά κι επιβεβαιώνει τη συνεχή παρακολούθηση του προφίλ ασφαλείας κάθε εγκεκριμένου φαρμάκου στην αγορά. Η τελευταία διαδικασία είναι πολύ σημαντική καθώς απο τα αποτελέσματά της, είναι ανά πάσα στιγμή δυνατόν να διακοπεί η κυκλοφορία του φαρμάκου.



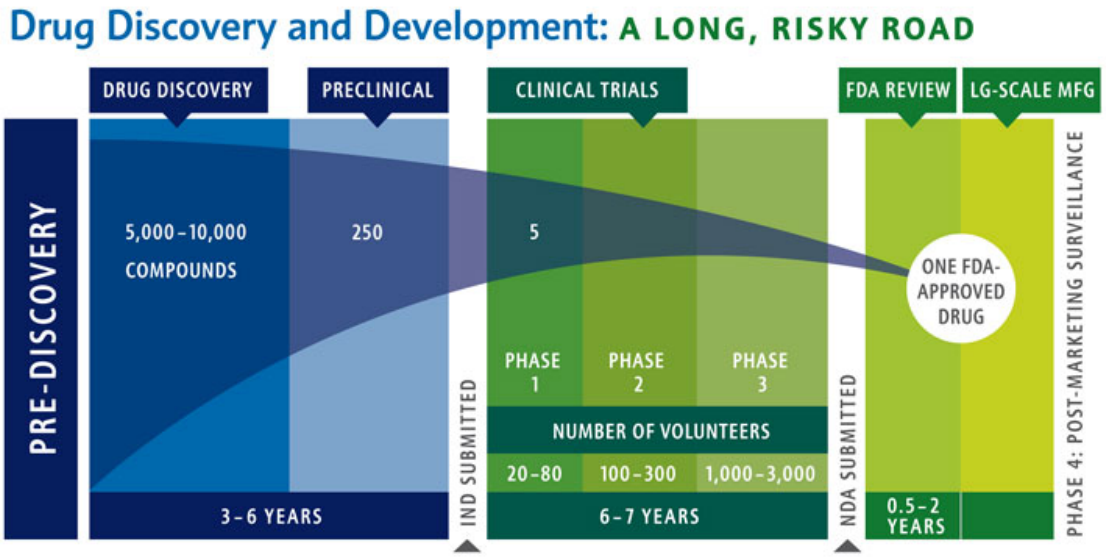
Σχήμα 3.2: Στάδια παραγωγής φαρμάκου

3.4 Ανάπτυξη κλινικών φάσεων και κόστους

Στο Σχήμα 3.4, φαίνεται το κόστος που απαιτείται για την όλη διαδικασία. Το συνολικό κόστος εκτιμάται πως μπορεί να αγγίξει έως και τα 2,6 δισεκατομμύρια δολλάρια.

Ιδιαίτερο ενδιαφέρον προσελκύει η δυσκολία προσπέλασης της παρομοιώδους "κοιλάδας του θανάτου", του χάσματος που υπάρχει ανάμεσα στην ανακάλυψη ενός υποσχόμενου νεοεμφανιζόμενου παράγοντα, μιας νέας χημικής ουσίας, εώς την απόδειξη της αποτελεσματικότητάς και της ασφάλούς χορήγησής της στον ανθρώπινο οργανισμό, Σχήμα 3.5. Η προσπέλαση αυτή μπορεί να είναι μια μεγάλη πρόκληση για τους ακαδημαικούς ερευνητές που δεν έχουν αρκετή εμπερία στην ανάπτυξη και στη διαδικασία παραγωγής ενός φαρμάκου.

Η φαρμακευτική βιομηχανία αντιμετωπίζει έναν συνεχώς αυξανόμενο αριθμό προκλή-



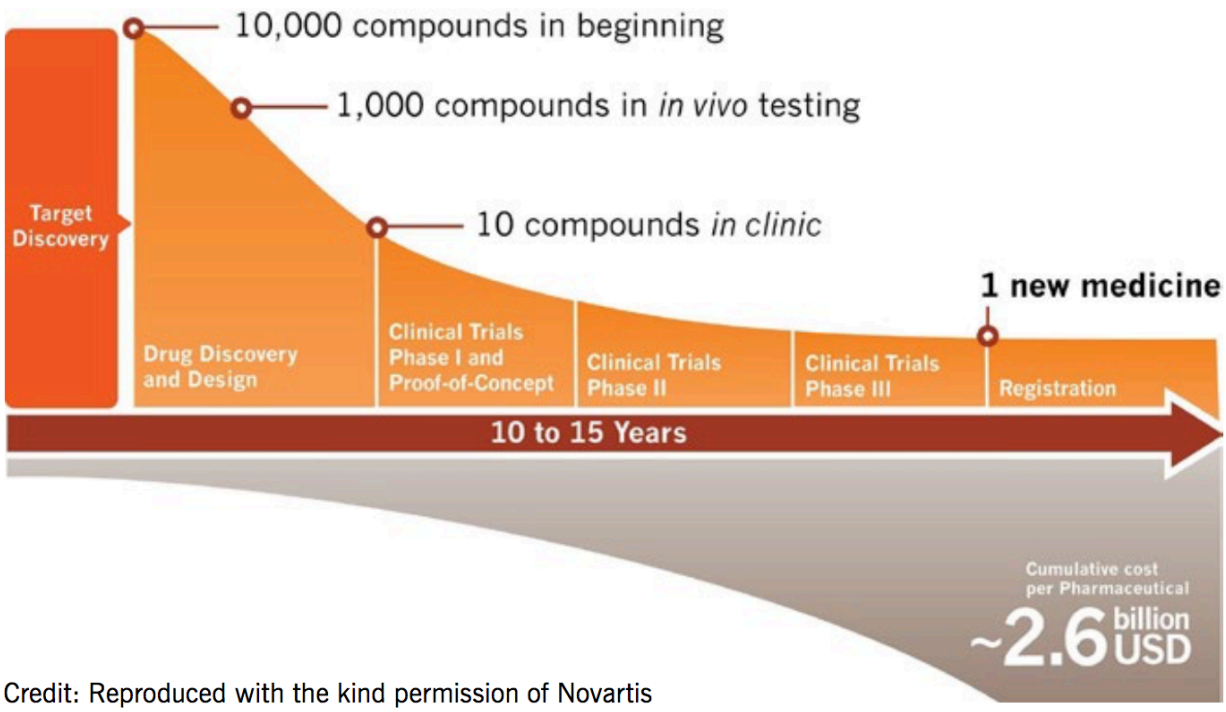
Σχήμα 3.3: Πλήθος ουσιών και χρονική διάρκεια για την παραγωγή φαρμάκου

σεων συμπεριλαμβανομένων της μείωση του κόστους της RD, του χαμηλού αριθμού εγκρίσεων των νέων θεραπευτικών παραγόντων και των πιο αυστηρών προδιαγραφών παραγωγής των φαρμάκων. Μια πιο σωστή αξιολόγηση των εν δυνάμει υποψηφίων παραγόντων στη φάση ανακάλυψής τους, θα οδηγούσε στη μείωση τόσο του κόστους όσο και των ποσοστών αποτυχίας. Παράλληλα θα μείωνε σημαντικά το χρόνο που απαιτείται για την παραγωγή των φαρμάκων. Η δυνατότητα επέμβασης στη διαδικασία παραγωγής και ασφάλειας στο προκλινικό στάδιο, πριν γίνουν οι μεγάλες κοστοβόρες επενδύσεις, επιτρέπει στις εταιρείες να επικεντρωθούν στον πιο ”δυνατό” υποψήφιο, μειώνοντας έτσι κατά πολύ το κόστος της RD διαδικασίας.

Ωστόσο, το πλέον δαπανηρό και πιο χρονοβόρο στάδιο στην έρευνα και ανάπτυξη φαρμάκων είναι οι κλινικές μελέτες (Φάσεις I, II και III), κατά τις οποίες διερευνάται η δράση μιας υπό έρευνα ουσίας στους ανθρώπους. Αυτό καθιστά τη φαρμακοβιομηχανία σε παγκόσμιο επίπεδο τον μεγαλύτερο επενδυτή σε έρευνα και ανάπτυξη από οποιονδήποτε άλλο κλάδο.

Σε γενικές γραμμές, οι κλινικές δοκιμές έχουν τρεις υποχρεωτικές φάσεις : την πρώτη, τη δεύτερη και την τρίτη.

Προκλινικές μελέτες



Σχήμα 3.4: Κόστος για την παραγωγή φαρμάκου

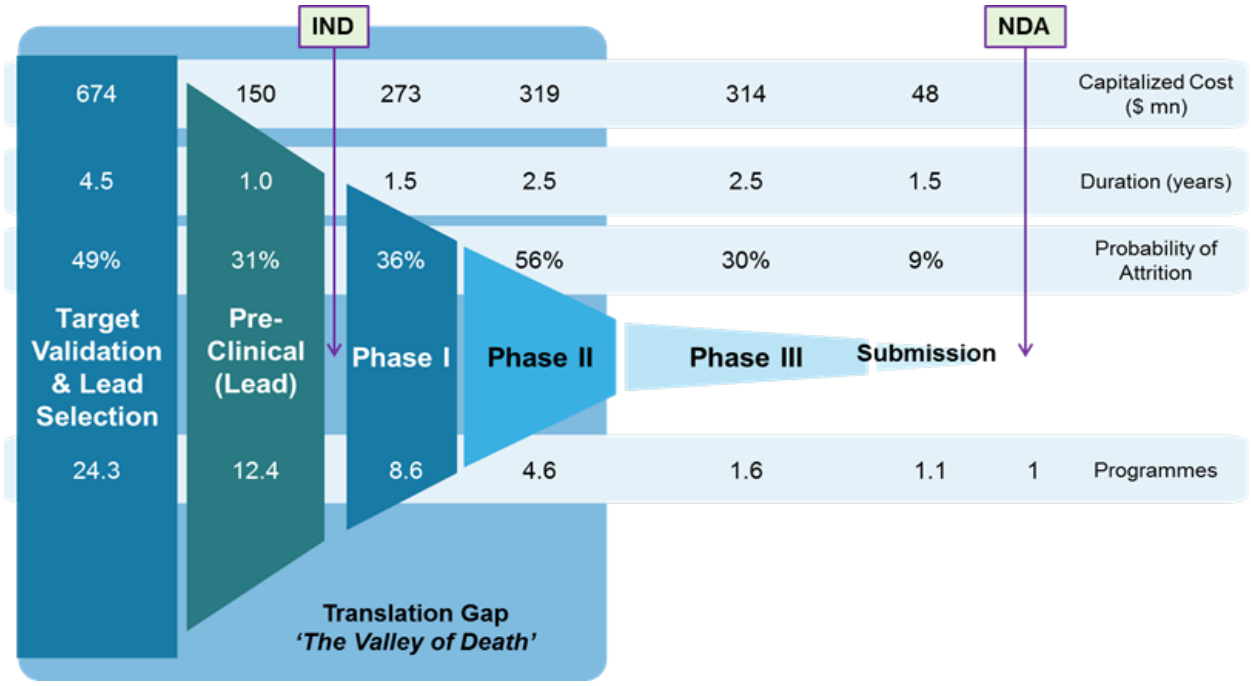
Είναι οι μελέτες που διεξάγονται στα πειραματόζωα και αποτελούν κομμάτι της βασικής έρευνας. Πρόκειται για μελέτες που εκτιμούν τόσο τη φαρμακολογική όσο και την τοξικολογική δράση μιας ουσίας.

Φάση I

Πρόκειται για τη φάση που εκτιμά αρχικά την ανεκτικότητα της ουσίας στους ανθρώπους. Το φάρμακο σε αυτή τη φάση, χορηγείται σε μικρό αριθμό ατόμων απο 20 εως 80 περίπου και συλλέγονται δεδομένα σχετικά με την ασφάλεια και το εύρος της δοσολογίας, ενώ λαμβάνεται και μια αρχική εκτίμηση των ανεπιθύμητων ενεργειών. Έτσι μελετάται η ασφάλεια του φαρμάκου σε διάφορες δόσεις, ενώ ακόμη δεν είναι δυνατή η αξιολόγηση της αποτελεσματικότητάς του. Τόσο ο μικρός αριθμός εθελοντών όσο και και η σύντομη διάρκεια της φάσης είναι κρίσιμες παράμετροι, καθώς υπάρχει πάντα ο κίνδυνος άγνωστων παρενεργειών.

Φάση II

Η φάση αυτή έχει ως σκοπό την αξιολόγηση της αποτελεσματικότητας της σύνθεσης του φαρμάκου, αλλά και της περαιτέρω ασφάλειας της χρήσης του. Η πειραματική θερα-



Σχήμα 3.5: Valley of death

πεία αφορά σε αυτή τη φάση μια μεγαλύτερη ομάδα ασθενών, από 100 εώς 300 άτομα. Η διάρκειά της μπορεί να είναι από μερικούς μήνες έως και κάποια χρόνια. Από τη φάση ΙΙ και μετά η ομάδα των εθελοντών, αποτελείται από άτομα που πάσχουν από τη νόσο για την οποία αναπτύσσεται η φαρμακευτική αγωγή.

Φάση ΙΙΙ

Στη φάση αυτή γίνεται διερεύνηση των πλέον ευνοϊκών συνθηκών, κάτω από τις οποίες αυξάνεται η απόδοση του φαρμάκου. Το υπό μελέτη φάρμακο και / ή το εικονικό φάρμακο (placebo), χορηγούνται πλέον σε μεγάλες ομάδες πασχόντων ατόμων, από 1000 έως και 3000 άτομα, ώστε να επιβεβαιωθεί η αποτελεσματικότητά του σε σχέση με τις υπάρχουσες καθιερωμένες θεραπείες, να παρακολουθηθούν οι ανεπιθύμητες ενέργειες και να συλλεχθούν πληροφορίες που θα επιτρέψουν την ασφαλή χρήση του φαρμάκου ή της θεραπείας. Η φάση αυτή μπορεί να διαρκέσει 2 με 4 χρόνια έως ότου ολοκληρωθεί. Ο μεγάλος αριθμός ατόμων που συμμετέχουν στη μελέτη βοηθάει πολύ στην παρατήρηση δευτερευούσων παρενεργειών. Επίσης πρέπει να ληφθεί υπόψιν πως αυτή η φάση αντιστοιχεί στο 60% του κόστους της όλης διαδικασίας. Οι Μελέτες Φάσης ΙΙΙ, αποτελούν το τελευταίο βήμα για τη σχεδιασμένη αξιολόγηση του νέου φαρμακευτικού προϊόντος, προτού εγκριθεί η κυκλοφορία του από τον αρμόδιο κρατικό μηχανισμό, όπως για παράδειγμα τον Ευρω-

παϊκό Οργανισμό Φαρμάκων και τον ΕΟΦ, στην Ελλάδα. Τα φαρμακευτικά προϊόντα που ολοκληρώνουν επιτυχώς τις κλινικές μελέτες Φάσης ΙΙΙ, εφόσον αξιολογηθούν θετικά από τις Εγκριτικές Αρχές, λαμβάνουν στη συνέχεια την απαραίτητη έγκριση κυκλοφορίας.

Οι φάσεις ΙΙ και ΙΙΙ έχουν ονομαστεί ελεγχόμενες δοκιμές, διότι συγκρίνουν την αποτελεσματικότητα των δοκιμαζόμενων φαρμάκων έναντι κάποιων αγωγών που λειτουργούν ως αναφορά. Ενίοτε, οι αγωγές - αναφορά μπορεί να είναι απλά placebo (ανενεργά χάπια) ή εν μέρει placebo, εάν δεν υπάρχουν άλλες εγκεκριμένες αγωγές. Για την αποφυγή επηρεασμού των κλινικών δοκιμών στις φάσεις ΙΙ και ΙΙΙ, από διάφορους υποκειμενικούς παράγοντες, οι εθελοντές πρέπει να αγνοούν ότι τους χορηγούνται φάρμακα που δοκιμάζονται (τυφλή δοκιμή). Σε ορισμένες δοκιμές, μπορεί ο αρμόδιος γιατρός να αγνοεί και αυτός ποια αγωγή δίνεται (διπλά τυφλή δοκιμή).

Οι λόγοι αποτυχίας στις προ - εγκριτικές μελέτες στη Φάση Ι, είναι κυρίως εμπορικοί, ενώ στις φάσεις ΙΙ και ΙΙΙ, σχετίζονται με την ανεπαρκή αποτελεσματικότητα (efficacy). Πολύ αυξημένο είναι και το ποσοστό απόρριψης στη Φάση ΙΙΙ λόγω ανεπαρκούς ασφάλειας - safety του φαρμάκου.

Γενικά, ο χρόνος που χρειάζεται ώστε ένα δοκιμαζόμενο φάρμακο να συμπληρώσει την οριζόμενη διαδικασία είναι από 7 έως 10 έτη, ανάλογα με το βαθμό δυσκολίας εξαγωγής αδιαμφισβήτητων συμπερασμάτων, όσον αφορά στην αποτελεσματικότητα και στην ανεκτικότητά του.

Φάση IV

Η φάση αυτή περιλαμβάνει τις μετεγκριτικές μελέτες για σκευάσματα που κυκλοφορούν στην αγορά κι έχουν ήδη ενταχθεί στην κλινική πράξη. Γίνεται καταγραφή πληροφοριών από την καθημερινή κλινική εφαρμογή τους σχετικά με τα οφέλη και τις ανεπιθύμητες ενέργειές τους, προκειμένου να γίνει η καλύτερη δυνατή χρήση τους. Επιπλέον με τις κλινικές μελέτες Φάσης IV αξιολογούνται τα μακροπρόθεσμα αποτελέσματα των φαρμακευτικών προϊόντων και αξιολογούνται σπάνιες μεν, αλλά δυνητικά σοβαρές ανεπιθύμητες ενέργειες, [22],[24].

4. ΑΝΑΛΥΣΗ ΜΕΘΟΔΩΝ

Η τεχνολογική ανάπτυξη και ο συνδυασμός της πληροφορικής με την ανάπτυξη της φαρμακευτικής χημείας και των γνώσεων της βιολογίας, οδήγησαν στην ανάπτυξη του τομέα της βιοπληροφορικής. Η εκμετάλλευση όλων των παραπάνω οδήγησε στη δημιουργία μοντέλων πρόβλεψης που χρησιμοποιούνται για την ανακάλυψη νέων φαρμακομορίων. Τα μοντέλα στηρίζονται κυρίως σε διάφορες μεθόδους τεχνητής νοημοσύνης και καθώς η χρήση αυτού του τομέα της πληροφορικής είναι σε ραγδαία ανάπτυξη, υπάρχουν πολλές ερευνητικές εργασίες που στηρίζονται σε αυτό και έχουν αρχίσει να δίνουν και πρακτικά αποτελέσματα.

Οι πιο συχνά χρησιμοποιούμενες μέθοδοι είναι

- Artificial neural networks (ANN)
- Gradient boosting machine (GMB)
- Support vector regressions (SVR)
- Local lazy learnings (LLL)
- KNN
- LLR (SA, SR, GP)
- Consensus models (ACM, SCM)

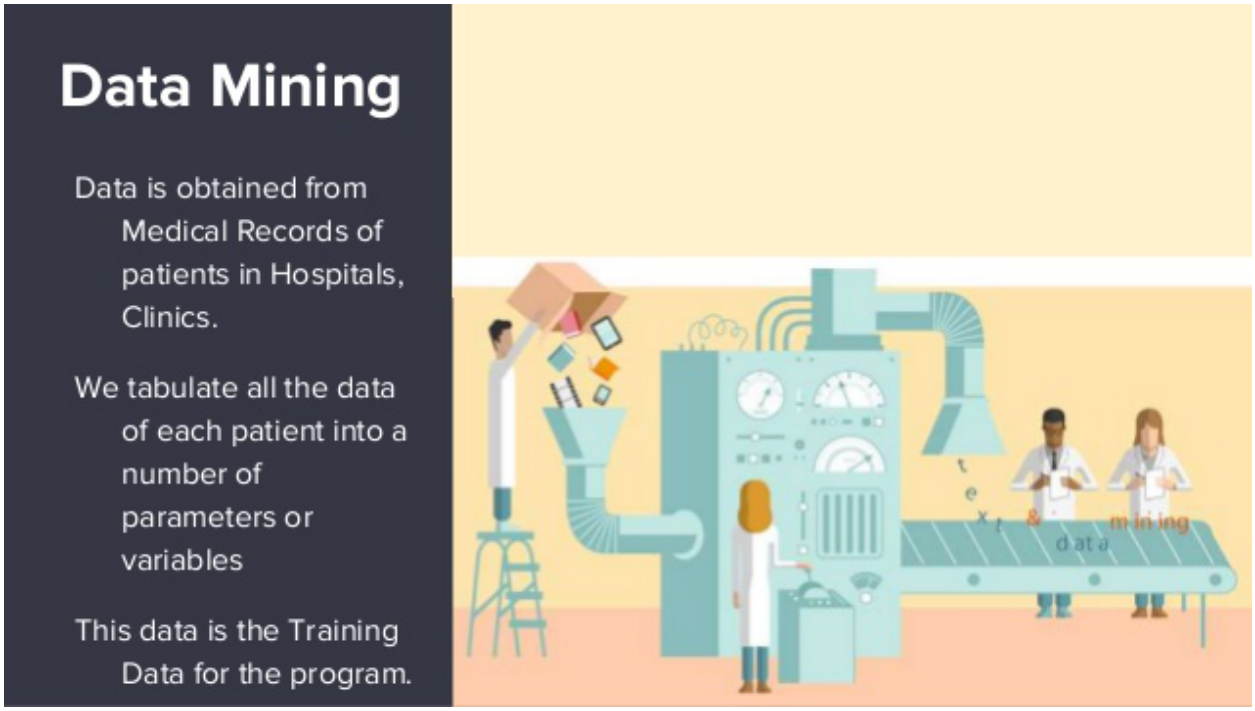
4.1 Artificial Neural Networks

4.1.1 Τεχνητή νοημοσύνη

Η τεχνητή νοημοσύνη (Artificial intelligence, AI) έχει εδραιωθεί ως η επιστημονική περιοχή της πληροφορικής που είναι ικανή να παράγει software, τα οποία μπορούν να λειτουργήσουν σοφά κι έξυπνα μιμούμενα την ανθρώπινη εγκεφαλική λειτουργία. Περιλαμβάνει μεθόδους, εργαλεία και συστήματα ικανά να προσομοιώσουν την ανθρώπινη λογική και την επαγωγική μέθοδο εκμάθησης, καθώς και τη συλλογιστική της εγκεφαλικής λειτουργίας για την επίλυση προβλημάτων. Υπάρχουν δύο βασικές κατηγορίες στα συστήματα τεχνητής νοημοσύνης. Η πρώτη περιλαμβάνει μεθόδους και συστήματα που προσπαθούν μιμούμενα την ανθρώπινη εμπειρία να εξαγουν συμπεράσματα που βασίζονται σε

ενα σετ κανόνων. Η δεύτερη περιλαμβάνει συστήματα που μοντελοποιούν τον τρόπο που το μυαλό λειτουργεί, όπως για παράδειγμα τα νευρωνικά δίκτυα (ANNs) .

Στην πρώτη κατηγορία ανήκουν τα έμπειρα συστήματα (expert systems), που είναι συστήματα βασισμένα στη γνώση, μια προέκταση των συμβατικών υπολογιστικών συστημάτων τα οποία πολύ συχνά τα αποκαλούν ως την πέμπτη γενιά των υπολογιστικών συστημάτων. Αυτή η μέθοδος που βασίζεται στη γνώση, επιτρέπει στον expert να καθορίσει τους κανονές που προσομοιάζουν τη διαδικασία σκέψης και παρέχουν έναν απλό τρόπο εξαγωγής συμπερασμάτων και λύνουν τα προβλήματα ακολουθώντας ενα σετ κανόνων. Τα expert systems βασίζονται στην ιδέα πως μπορεί να γίνει μια μοντελοποίηση της λογικής σκέψης, αφού συγκεντρωθεί μια λίστα λογικών συνθηκών - προτάσεων και εκτελεστούν σε αυτή τη λίστα διάφορες λογικές μετατροπές. Τα συστήματα αυτά είναι χρήσιμα για ιατρικές διαγνώσεις και για τη λύση άλλων διαγνωστικών προβλημάτων. Παρέχουν έναν οδηγό για την πρόβλεψη και τη λήψη αποφάσεων σε διάφορα περιβάλλοντα, όπου υπάρχει αβεβαιότητα ή ασάφεια. Στην ιατρική υπάρχουν πολλά παραδείγματα επιστημονικών μοντέλων που χρησιμοποιούνται για την πρόγνωση ή τη διάγνωση διαφόρων ασθενειών, Σχήμα 4.1.



Σχήμα 4.1: Medical Data mining

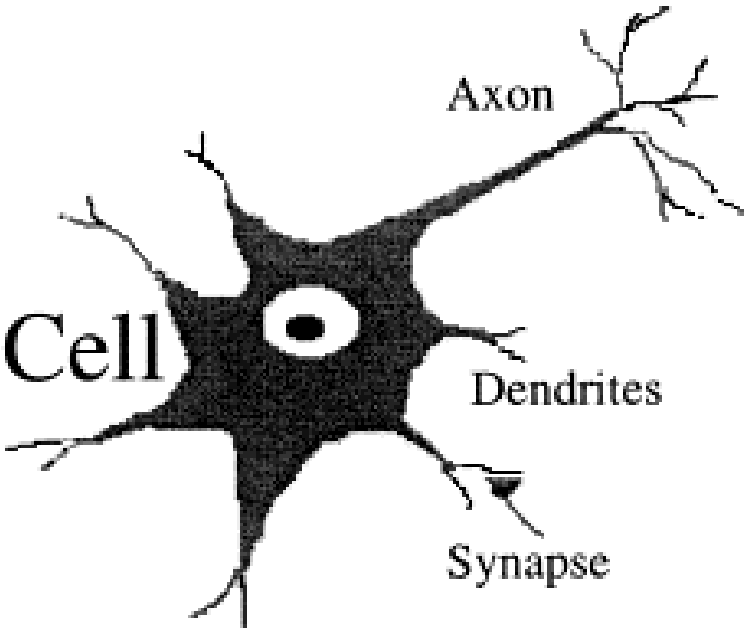
Εντούτοις, είναι αρκετά σύνηθες το μοντέλο να μην μπορεί να ολοκληρωθεί λόγω ανα-

κρίβειας ή μη ολοκλήρωσης της λήψης των δεδομένων του ασθενούς. Για παράδειγμα στην ιατρική, είναι σύνηθες το φαινόμενο των ημιτελών ή μη ακριβών επιστημονικών μοντέλων που αφορούν την ανθρώπινη υγεία ή κάποια ασθένεια.

Τα ANNs είναι ψηφιοποιημένα μοντέλα του ανθρώπινου εγκέφαλου, υπολογιστικά προγράμματα που έχουν σχεδιαστεί για να προσομοιάσουν τον τρόπο που ο ανθρώπινος εγκέφαλος επεξεργάζεται τις πληροφορίες. Τα ANNs μαθαίνουν ή εκπαιδεύονται, μέσω εμπειρίας με τα κατάλληλα εκπαιδευτικά πρότυπα ακριβώς όπως ο άνθρωπος, όχι μέσω υπολογιστικών προγραμμάτων. Τα νευρωνικά δίκτυα συλλέγουν τις γνώσεις τους ανακαλύπτοντας τα πρότυπα και τις σχέσεις τους από υπάρχοντα δεδομένα.

4.1.2 Προσομοίωση ανθρώπινου εγκεφάλου

Ο εγκέφαλος είναι ένα εξαιρετικό εργαλείο αναγνώρισης προτύπων. Όταν κοιτάμε ενα στυλό γνωρίζουμε πως είναι στυλό, επειδή κάποιοι βιολογικοί νευρώνες σε συγκεκριμένη περιοχή του εγκεφάλου έχουν εισάγει ως πρότυπο αυτή την πληροφορία σε προηγούμενες περιπτώσεις και έχουν εκπαιδευτεί να συνδέουν αυτό το συγκεκριμένο πρότυπο με την περιγραφή του αντικειμένου ως στυλό. Καθώς ο εγκέφαλός μας περιέχει δισεκατομμύρια νευρώνες που είναι πλήρως διασυνδεδεμένοι, μπορούμε να εκπαιδευτούμε και να αναγνωρίσουμε μια σχεδόν ατελείωτη ποικιλία από εισαγόμενα πρότυπα.



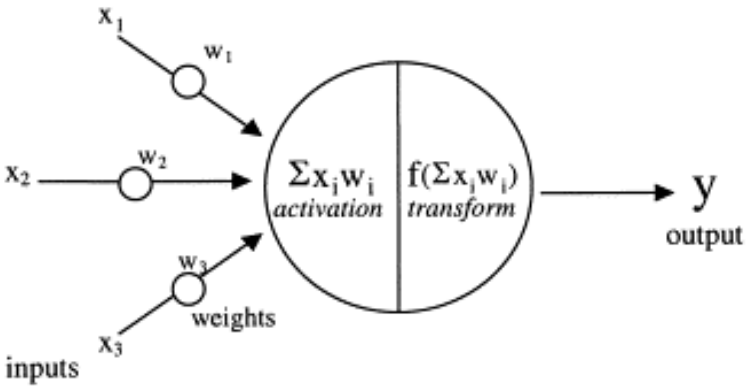
Σχήμα 4.2: Κυτταρικό σώμα νευρώνα

Κατά μέσο όρο ένας εγκέφαλος περιέχει περίπου 100 δισεκατομμύρια νευρώνες, ο καθένας από τους οποίους έχει 1.000 με 10.000 διασυνδέσεις με άλλους νευρώνες. Κάθε νευρώνας αποτελείται από ένα κυτταρικό σώμα που περιέχει τον πυρήνα ο οποίος ελέγχει την κυτταρική λειτουργία, πολλές λεπτές αυλακώσεις και δενδρίτες που μεταφέρουν την πληροφορία στο εσωτερικό του κυττάρου και μια πιο μακριά αυλάκωση που είναι γνωστή ως ο άξονας που μεταφέρει από το εσωτερικό του εγκεφάλου προς το εξωτερικό, το σήμα, Σχήμα 4.2. Οι παλμοί διαχέονται μέσω του άξονα στη σύναψη, την ένωση ενός νευρώνα με τον επόμενο και τα σήματα περνούν διαδοχικά από τον ένα νευρώνα στον άλλο. Οι νευρώνες είναι οργανωμένοι σε ένα πλήρες συνδεδεμένο δίκτυο και λειτουργούν σαν αγγελιοφόροι - παραλήπτες στη λήψη και αποστολή των παλμών. Το αποτέλεσμα αυτών είναι ένας έξυπνος εγκέφαλος ικανός να μαθαίνει, να προβλέπει και να αναγνωρίζει [1].

Τα ΑΝΝ είναι υπολογιστικά μοντέλα εμπνευσμένα από τη βιολογία, αποτελούμενα από εκατοντάδες μονάδες, τους τεχνητούς νευρώνες, που συνδέονται με κάποιους συντελεστές (βάρη), δημιουργώντας έτσι μια δομή νευρωνικού δικτύου. Είναι επίσης γνωστά και ως μοντέλα επεξεργασίας καθώς επεξεργάζονται στην ουσία τις πληροφορίες που τους δίνονται. Κάθε μοντέλο αποτελείται από τα εισερχόμενα δεδομένα, τροποποιημένα με τα κατάλληλα βάρη, από την εξίσωση μετατροπής και από το εξαγόμενο αποτέλεσμα. Ουσιαστικά πρόκειται για μια εξίσωση με ζυγισμένα - κατάλληλα τροποποιημένα από τα βάρη εισερχόμενα δεδομένα και αποτελέσματα. Αν και κάθε νευρώνας μπορεί να εκτελέσει συγκεκριμένες απλές εξισώσεις για την μετατροπή των πληροφοριών, η δυναμικότητα των νευρωνικών υπολογιστικών δικτύων προέρχεται από τη διασύνδεση των νευρώνων σε ένα δίκτυο. Τα τεχνητά νευρωνικά δίκτυα σπανίως αποτελούνται από εκατοντάδες ή μερικές χιλιάδες ”νευρώνες ”, ενώ ο ανθρώπινος εγκέφαλος αποτελείται από περίπου 100 δισεκατομμύρια νευρώνες. Τα τεχνητά δίκτυα συγκρινόμενα σε πολυπλοκότητα με τον ανθρώπινο εγκέφαλο απέχουν πολύ από τη δημιουργική του ικανότητα. Ο ανθρώπινος εγκέφαλος είναι αρκετά πιο πολύπλοκος και δυστυχώς πολλές από τις διανοητικές του λειτουργίες δεν έχουν ακόμη αποκωδικοποιηθεί. Τα ΑΝΝ ωστόσο, είναι ικανά να επεξεργαστούν μεγάλο όγκο δεδομένων και να κάνουν και προβλέψεις που πολλές φορές είναι εκπληκτικά ακριβείς. Υπάρχουν διάφοροι τύποι νευρωνικών δικτύων που έχουν σχεδιαστεί εώς τώρα, αλλά όλοι μπορούν να περιγραφούν από την εξίσωση που μεταφέρει το σήμα από τους νευρώνες, από τον κανόνα εκμάθησης και από τη φόρμουλα σύνδεσης των νευρώνων.

4.1.3 Τεχνητός νευρώνας

Ο τεχνητός νευρώνας είναι το δομικό στοιχείο των ANN που σχεδιάζεται για να μιμηθεί τη λειτουργία του βιολογικού νευρώνα. Τα σήματα που δέχεται, καλούνται ως εισερχόμενα, πολλαπλασιαζόμενα με τα κατάλληλα προσαρμοσμένα βάρη, αρχικά αθροίζονται - συνδυάζονται και στη συνέχεια μέσω της εξίσωσης μετατροπής παράγουν το αποτέλεσμα για αυτόν τον νευρώνα. Η εξίσωση ενεργοποιήσης - μετατροπής είναι ο συνδυασμός των κατάλληλα τροποποιημένων από τα βάρη εισερχόμενων δεδομένων και η πιο συχνά χρησιμοποιούμενη, είναι η σιγμοειδής εξίσωση, Σχήμα 4.3.



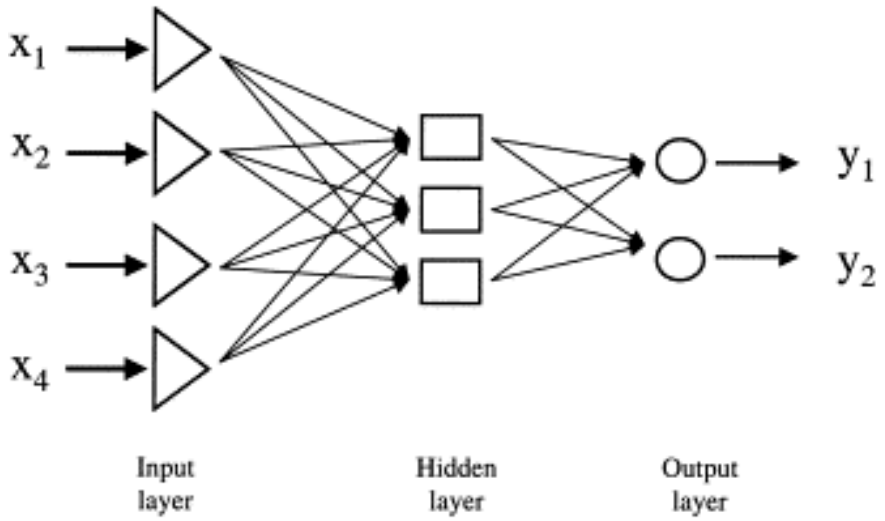
Σχήμα 4.3: Μοντέλο τεχνητού νευρώνα

4.1.4 Διασύνδεση νευρώνων

Ο τρόπος που οι νευρώνες συνδέονται μεταξύ τους έχει πολύ σημαντική επίδραση στη λειτουργία του τεχνητού νευρωνικού δικτύου. Όπως ακριβώς οι βιολογικοί νευρώνες έτσι και οι τεχνητοί μπορούν να δεχτούν σήματα που διεγείρουν ή και αναστέλλουν τη λειτουργία τους.

Τα σήματα που διεγείρουν προκαλούν έναν αθροιστικό μηχανισμό, με τον επόμενο νευρώνα να προσθέτει κάτι στο αποτέλεσμα, ενώ τα σήματα που αναστέλλουν τη λειτουργία προκαλούν τον επόμενο νευρώνα ώστε να αφαιρέσει από το αποτέλεσμα. Ένας νευρώνας μπορεί επίσης να εμποδίσει την δράση κάποιου άλλου στην ίδια ”στοιβάδα”, αυτό καλείται πλευρική αναστολή. Το νευρωνικό δίκτυο επιθυμεί να επιλέξει το πιο δυνατό ενδεχόμενο και να αναστείλει τους υπόλοιπους νευρώνες. Η διαδικασία αυτή ονομάζεται αναμέτρηση. Η προς τα πίσω ανατροφοδότηση είναι ένας άλλος τύπος σύνδεσης όπου το εξερχόμενο δεδομένο μιας στοιβάδας, επιστρέφει ως εισερχόμενο σήμα είτε στην ίδια στοιβάδα, είτε

σε κάποια προηγούμενη. Δύο τύποι αρχιτεκτονικής μπορούν να διακριθούν αναλόγως της παρουσίας ή της απουσίας της ανατροφοδότησης σε ένα νευρωνικό δίκτυο. Η αρχιτεκτονική της προς τα εμπρός τροφοδότησης δεν έχει κάποια σύνδεση των εξερχόμενων με τα εισέρχομενα σήματα των νευρώνων και για αυτό το λόγο δε διατηρεί κάποιο αρχείο των προηγούμενων εξερχόμενων σημάτων, Σχήμα 4.4 .



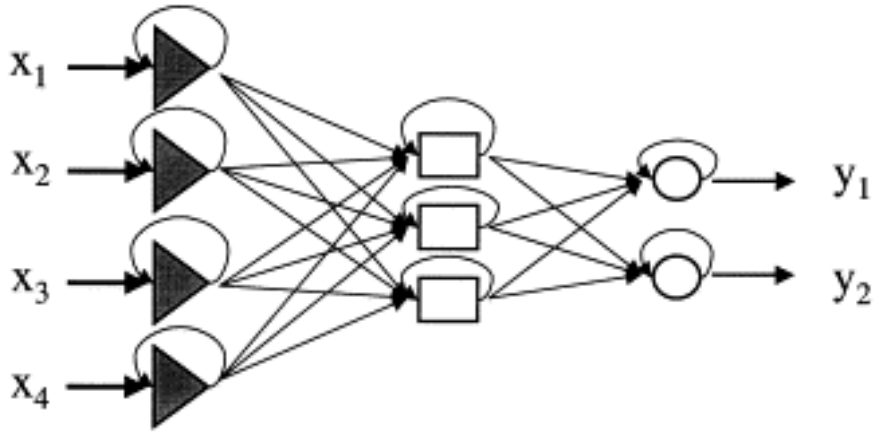
Σχήμα 4.4: Τεχνική τροφοδότησης προς τα εμπρός, feedforward network

Η ανατροφοδότηση έχει συνδέσεις από τους εξερχόμενους προς τους εισερχόμενους νευρώνες. Κάθε νευρώνας έχει ένα επιπρόσθετο βάρος ως εισερχόμενο δεδομένο που επιτρέπει την προσθήκη βαθμών ελευθερίας όταν προσπαθούμε να μειώσουμε το σφάλμα εκπαίδευσης, Σχήμα 4.5. Ένα τέτοιο δίκτυο διατηρεί στη μνήμη του την προηγούμενη κατάσταση, ώστε το επόμενο στάδιο να εξαρτάται όχι μόνο από τα εισερχόμενα σήματα, αλλά και από τις πρότερες καταστάσεις του νευρωνικού δικτύου.

4.1.5 Κανόνες εκμάθησης

Υπάρχουν διάφοροι κανόνες εκμάθησης, αλλά οι πιο συχνά χρησιμοποιούμενοι είναι ο Delta ή ο Back - propagation. Ένα νευρωνικό δίκτυο εκπαιδεύεται ώστε να ταιριάξει όσο το δυνατόν καλύτερα ένα εισερχόμενο σετ δεδομένων μετά από επαναλαμβανόμενες προσαρμογές των βαρών του. Η χρήση των βαρών με τις κατά επανάληψη προσαρμογές αυτών, είναι άκρως απαραίτητη για τις αναγνωριστικές ικανότητες των ANN.

Η πληροφορία από τα εισερχόμενα δεδομένα τροφοδοτείται μπρος τα εμπρός στο δίκτυο, ώστε να βελτιστοποιεί τα βάρη μεταξύ των νευρώνων. Στον backward propagation



Σχήμα 4.5: Τεχνική ανατροφοδότησης, feedback network

η βελτιστοποίηση των βαρών πραγματοποιείται από το σφάλμα που συμβαίνει κατά την εκπαίδευση ή τη φάση εκμάθησης. Το ANN ”διαβάζει” τις εισερχόμενες και εξερχόμενες τιμές του training data και αλλάζει την τιμή των βαρών, ώστε να μειώσει την διαφορά ανάμεσα στην προβλεπόμενη και την πραγματική τιμή. Τα λάθη στην προβλεπτική ικανότητα ελαχιστοποιούνται μετά από κάποιους κύκλους εκπαίδευσης εώς ότου το δίκτυο φτάσει σε ένα υψηλό επίπεδο ακρίβειας.

Αν κάποιο δίκτυο αφεθεί να εκπαιδευτεί για υπερβολικό αριθμό επαναλήψεων, εκτελέσει δηλαδή πολλούς κύκλους εκπαίδευσης, θα υπερεκπαιδευτεί (over-trained). Σε αυτή την περίπτωση το νευρωνικό δίκτυο θα χάσει τη δυνατότητα της γενίκευσης των αποτελεσμάτων του και την προβλεπτική του ικανότητα.

4.1.6 Προσεγγίσεις εκπαίδευσης

Υπάρχουν διάφοροι τύποι ANN με κάποιους να κρίνονται δημοφιλέστεροι έναντι των υπολοίπων. Όταν νευρωνικά δίκτυα χρησιμοποιούνται για την ανάλυση δεδομένων είναι πολύ σημαντικό να διαφοροποιηθεί το ANN μοντέλο (το στήσιμο του νευρωνικού δικτύου) από τον αλγόριθμο του ANN (το υπολογιστικό σύστημα που ουσιαστικά θα παράγει τα αποτελέσματα). Όταν ένα νευρωνικό δίκτυο έχει κατασκευαστεί για μια συγκεκριμένη εφαρμογή, τότε αυτό είναι έτοιμο για να εκπαιδευτεί. Υπάρχουν δύο προσεγγίσεις για την εκπαίδευσή του η supervised και η unsupervised. Το πιο συχνά χρησιμοποιούμενο δίκτυο είναι ένα πλήρως συνδεδέμενο, supervised δίκτυο με τον back - propagation κανόνα εκμάθησης. Αυτός ο τύπος ANN αποδίδει εξαιρετικά σε προβλήματα πρόβλεψης και ταξινόμησης. Ένας άλλος τύπος είναι ο Kohonen ή Self Organizing Map με unsupervised

αλγορίθμους εκμάθησης, που αποδίδει εξαιρετικά στην εύρεση σχέσεων μεταξύ διαφόρων σετ δεδομένων.

4.1.7 Supervised learning ANNs

Ο στόχος μίας supervised μεθόδου είναι να προβλέψει μία ή περισσότερες τιμές στόχους, από ένα ή περισσότερα σετ δεδομένων. Supervised learning είναι μια μέθοδος regression που βασίζεται σε ζευγάρια των δεδομένων, τις εισερχόμενες και τις εξερχόμενες τιμές του training set. Αυτός ο τύπος νευρωνικού δικτύου είναι ένα σύστημα πλήρως συνδεδεμένων μεταξύ τους νευρώνων οργανωμένων σε στοιβάδες, την εισερχόμενη στοιβάδα, την εξερχόμενη και τις κρυμμένες αναμεσά τους στοιβάδες. Οι νευρώνες της εισερχόμενης στοιβάδας λαμβάνουν τα δεδομένα από το dataset. Οι νευρώνες της εξερχόμενης στοιβάδας παρέχουν την απάντηση του ANN μετά την επεξεργασία των εισερχόμενων δεδομένων. Οι κρυμμένοι νευρώνες επικοινωνούν μόνο με άλλους νευρώνες. Είναι μέρος ενός μεγάλου εσωτερικού μοτίβου που ουσιαστικά καθορίζει, βρίσκει την λύση στο πρόβλημα. Η θεωρία λέει πως οι περισσότερες συναρτήσεις μπορούν να προσεγγιστούν χρησιμοποιώντας έστω και μια απλή κρυμμένη στοιβάδα .

Η πληροφορία που μεταφέρεται από το ένα στοιχείο στο άλλο εμπεριέχει κ ένα σετ βαρών. Κάποιες από τις συνδέσεις γίνονται πιο δυνατές λαμβάνοντας μεγαλύτερα βάρη και άλλες πιο αδύναμες ώστε να καταλήξουμε στην πιο σωστή απάντηση. Ο πιο συχνά χρησιμοποιούμενος αλγόριθμος για τα σφάλματα είναι ο back - propagation. Το σφάλμα της πρόβλεψης διοχετεύεται προς τα πίσω στο δίκτυο, ώστε να γίνει η εκ νέου προσαρμογή των βαρών για την ελαχιστοποίηση της τιμής του, αλλά και να εμποδίσει την επανάληψη του ίδιου σφάλματος. Αυτή η διαδικασία συνεχίζεται με διάφορα training sets, ώστε να γίνει ελαχιστοποίηση του σφάλματος μέσω αυτών των σετ. Το αποτέλεσμα αυτών προέρχεται από το συνδυασμό των εισερχόμενων με τα εξερχόμενα δεδομένα μέσω μιας ασαφούς κρυμμένης στοιβάδας.

Ο αριθμός των νευρώνων της κρυμμένης στοιβάδας επηρεάζει τον αριθμό των συνδέσεων. Κατά τη φάση της εκπαίδευσης τα εισερχόμενα δεδομένα προσαρμόζονται, μετατρέπονται σύμφωνα με τα βάρη. Για αυτό το λόγο ο αριθμός των συνδέσεων επιδρά σημαντικά στην απόδοση του δικτύου. Αν ο αριθμός των κρυμμένων νευρώνων είναι μικρός, θα καθυστερήσει την διαδικασία εκμάθησης, ενώ αν είναι μεγάλος θα παρεμποδίσει την προβλεπτική ικανότητα του δικτύου λόγω της υπερεκπαίδευσης. Αυξάνοντας τον αριθμό των κρυμμένων νευρώνων, το δίκτυο πλησιάζει καλύτερα την τοπολογία του training dataset.

Ωστόσο αν υπερβούμε τον μέγιστο αριθμό, το παραγόμενο μοντέλο είναι πολύ κοντά στο training dataset, αλλά ακατάλληλο για να γενικεύσει σε άλλα datasets.

Όταν το ANN παράγει ικανοποιητικά αποτελέσματα (έχει δηλαδή εκπαιδευτεί σε ικανοποιητικό επίπεδο), αποθηκεύονται τα βάρη μεταξύ των νευρώνων. Αυτά είναι τα βάρη που χρησιμοποιούνται για την πρόβλεψη που θα εκτελέσει το δίκτυο, κατά την εισαγωγή νέων δεδομένων.

4.1.8 Εφαρμογές των ANN στην φαρμακευτική έρευνα

Η ANN μεθοδολογία είναι βασισμένη στην προσπάθεια να μοντελοποιηθεί ο τρόπος που ένας εγκέφαλος επεξεργάζεται τα δεδομένα. Για αυτό και διαφέρει σημαντικά από τις υπόλοιπες κλασικές μεθόδους στατιστικής ανάλυσης. Τα ANN αντιπροσωπεύουν ένα τεχνικά υποσχόμενο μοντέλο, ειδικά όταν πρόκειται για επεξεργασία δεδομένων με μη γραμμική εξάρτηση, που απαντάται πολύ συχνά σε φαρμακευτικά δεδομένα. Τα νευρωνικά δίκτυα απαιτούν λιγότερη στατιστική ανάλυση και είναι ικανά να εντοπίσουν αρκετά πολύπλοκες μη γραμμικά εξαρτώμενες σχέσεις μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών, καθώς και των πιθανών μεταξύ τους αλληλεπιδράσεων. Για την επίτευξη αυτού του στόχου δεν απαιτείται ωστόσο η χρήση πολύπλοκων εξισώσεων, αλλά η δοκιμή διαφορετικών αλγορίθμων για την εκπαίδευσή τους.

Σε σχέση με την ειδίκευση του μοντέλου, τα νευρωνικά δίκτυα δεν απαιτούν να γνωρίζουν την πηγή των δεδομένων. Ανάλογα με το πλήθος των βαρών που πρέπει να καθορισθούν απαιτείται και ο αντίστοιχος όγκος δεδομένων εκπαίδευσης. Επιπλέον τα ANN μπορούν να συνδυάσουν και να συνδέσουν περισσότερα από ένα σετ δεδομένων που προέρχονται τόσο από βιβλιογραφικές όσο και από πειραματικές, εργαστηριακές πηγές προκειμένου να λύσουν το πρόβλημα. Η χρήση των ANN είναι μια νέα, αλλά ευρέως αναπτυσσόμενη περιοχή στον τομέα της φαρμακευτικής έρευνας.

Οι διάφορες εφαρμογές των ANN μπορούν να κατανεμηθούν σε classification ή pattern recognition, prediction και modeling. Οι δυνατότητες εφαρμογής της ANN μεθοδολογίας στις φαρμακευτικές επιστήμες είναι ευρείες. Ποικίλουν από την απλή ερμηνεία δεδομένων αναλυτικής φαρμακευτικής χημείας (μοντελοποίηση του ποιοτικού ελέγχου στη φαρμακευτική ανάλυση), το σχεδιασμό φαρμάκου (QSAR και μοντελοποίηση χημικού μορίου) και σχεδιασμό της δοσολογίας εώς και τη μοντελοποίηση της κλινικής φαρμακολογίας μέσω της βιοφαρμακευτικής (φαρμακοκινητική και φαρμακοδυναμική μοντελοποίηση, με in vitro/in vivo συσχέτιση).

4.2 Gradient boosting machine (GBM)

Η ιδέα της τεχνικής boosting προήλθε κατά την προσπάθεια τροποποίησης ενός weak learner, ώστε να γίνει καλύτερος. Ως weak learner ή weak hypothesis, ορίζεται αυτός ο αλγόριθμος του οποίου η απόδοση είναι έστω και ελάχιστα καλύτερη από μια τυχαία επιλογή.

Η βασική ιδέα έγκειται στο φιλτράρισμα των παρατηρήσεων, ωστέ να εξαιρεθούν αυτές τις οποίες ένας weak learner μπορεί να διαχειριστεί και να επικεντρωθεί η εκπαίδευση στο υπόλοιπο πλήθος των πιο δύσκολων παρατηρήσεων αναπτύσοντας νέους weak learner που να μπορούν να τις διαχειριστούν. Η μέθοδος επαναλαμβάνεται αρκετές φορές ώστε να παραχθεί ένα ικανοποιητικό αποτέλεσμα, επικεντρωμένο κάθε φορά στα παραδείγματα - παρατηρήσεις που ο προηγούμενος αλγόριθμος δεν κατάφερε ή δυσκολεύτηκε να κατηγοριοποίησει.

4.2.1 AdaBoost

Ο AdaBoost ήταν ο πρώτος boosting αλγόριθμος. Οι weak learners στη μέθοδο αυτή είναι decision trees με μονό σπλιτ, που λόγω του μικρού τους μεγέθους αποκαλούνται και decision stumps. Η πρώτη υλοποίηση αυτού με μεγάλη επιτυχία σε εφαρμογές, ήταν ο αλγόριθμος Adaptive Boosting ή AdaBoost.

Ο AdaBoost δουλεύει ζυγίζοντας, αξιολογώντας τις παρατηρήσεις, δίνει μεγαλύτερο βάρος στα δεδομένα που δύσκολα κατηγοριοποιούνται και λιγότερο σε αυτά που έχουν ήδη κατηγοριοποιηθεί. Οι νέοι weak learners που προστίθενται διαδοχικά, επικεντρώνουν την εκπαίδευση τους στις πιο δύσκολες παρατηρήσεις που παραμένουν miss - classified.

Αυτό σημαίνει πως οι παρατηρήσεις που δυσκολεύονται να κατηγοριοποιηθούν, λαμβάνουν συνεχώς αυξανόμενα βάρη εώς ότου ο αλγόριθμος εντοπίσει το μοντέλο που καταφέρει να τις κατηγοριοποιήσει σωστά.

4.2.2 Γενίκευση του AdaBoost ως Gradient Boosting

Αυτό το πλαίσιο αναπτύχθηκε από τον Friedman και ονομάστηκε Gradient Boosting Machine (GBM). Αργότερα αποκαλείτο και απλώς gradient boosting ή gradient tree boosting.

Το στατιστικό πλαίσιο του boosting είναι ουσιαστικά μια αριθμητική βελτιστοποίηση

του προβλήματος, όπου ο σκοπός είναι να ελαχιστοποιήσουμε τη ζημιά του μοντέλου, προσθέτοντας νέους weak learners χρησιμοποιώντας τη διαδικασία του gradient descent. Αυτή η τάξη αλγορίθμων περιγράφεται ως ένα μοντέλο όπου σταδιακά προστίθενται σοφά επιλεγμένα στάδια. Αυτό συμβαίνει καθώς ένας νέος weak learner προστίθεται στιγμιαία, ενώ οι ήδη υπάρχοντες παγώνουν και μένουν αναλλοίωτοι.

Ο Gradient boosting αποτελείται από τρία στοιχεία, [2] :

- Από μία loss function βελτιστοποίησης
- Από έναν weak learner για να πραγματοποιεί τις πρόβλεψεις
- Από ένα μοντέλο που σταδιακά προσθέτει weak learners, για να ελαχιστοποιήσει την loss function

4.2.3 Loss Function

Η χρήση της εξαρτάται από τον τύπο του προβλήματος που θέλουμε να λύσουμε. Πρέπει να είναι διαφοροποιημένη, αλλά υπάρχουν διάφορες συνήθως χρησιμοποιούμενες loss functions που υποστηρίζονται, ένω ο καθένας μπορεί να καθορίσει τη δικιά του ανάλογα με το πρόβλημα.

Για παράδειγμα για regression μπορεί να χρησιμοποιηθεί η squared error, ενώ για classification η logarithmic loss. Το πλεονέκτημα του gradient boosting είναι πως δεν υπάρχει περιορισμός σχετικά με τους πιθανούς συνδυασμούς των αλγορίθμων boosting και της χρησιμοποιούμενης loss function. Υπάρχει ένα αρκετά γενικό πλαίσιο όπου κάθε διαφοροποιημένη loss function, μπορεί να χρησιμοποιηθεί με τους επιθυμητούς αλγορίθμους.

4.2.4 Weak Learner

Στο gradient boosting, ως weak learner χρησιμοποιούνται τα decision trees. Χρησιμοποιούνται συγκεκριμένα regression trees που εξάγουν πραγματικές τιμές των σπλιτ και οι οποίες μπορούν να αθροιστούν, επιτρέποντας στα επακόλουθα μοντέλα να εξάγουν τιμές οι οποίες μπορούν να προστεθούν και να διορθώσουν την απόκλιση των προβλέψεων. Τα δέντρα αποφάσεων κατασκευάζονται με έναν greedy τρόπο, επιλέγοντας το βέλτιστο σπλιτ που ελαχιστοποιεί την απώλεια και βασίζεται σε κάποια σκορ.

Αρχικά, όπως στην περίπτωση του AdaBoost, χρησιμοποιούνταν πολύ μικρά δέντρα αποφασέων με ένα μόνο σπλιτ. Μεγαλύτερα δέντρα μπορούν να χρησιμοποιηθούν αποτελούμενα από 4 μέχρι 8 επίπεδα. Είναι σύνηθες να κατασκευάζουμε τους weak learners με συγκεκριμένο τρόπο, όπως για παράδειγμα ορίζοντας εξαρχής έναν μάξιμουμ αριθμό στοιβάδων, κόμβων, σπλιτ ή διακλαδώσεων. Αυτό γίνεται, για να διασφαλιστεί πως θα παραμείνουν weak, ωστόσο όμως μπορούν να κατασκευαστούν με έναν greedy τρόπο.

4.2.5 Additive Model

Σε αυτό το μοντέλο προστίθεται κάθε στιγμή νέο δέντρο, ενώ τα ήδη υπάρχοντα του μοντέλου δεν αλλάζουν. Μια διαδικασία **gradient descent** χρησιμοποιείται ώστε να ελαχιστοποιήσει το σφάλμα που δημιουργείται κατά την προσθήκη των νέων δέντρων.

Παραδοσιακά η **gradient descent** χρησιμοποιείται για να μειώσει το σετ των παραμέτρων, όπως οι συντελεστές μιας εξίσωσης regression ή τα βάρη που υπολογίζονται σε ένα νευρωνικό δίκτυο. Μετά τον υπολογισμό του σφάλματος ή της απώλειας, γίνεται επαναπροσδιορισμός των βαρών ώστε να μειωθεί το σφάλμα. Αντί για παραμέτρους έχουμε μοντέλα από weak learners, ή πιο συγκεκριμένα από δέντρα αποφάσεων. Μετά τον υπολογισμό του σφάλματος, για να εφαρμόσουμε τη διαδικασία του **gradient descent**, πρέπει να προσθέσουμε ένα δέντρο στο μοντέλο που θα μειώνει το σφάλμα. Αυτό το κάνουμε παραμετροποιώντας το δέντρο, μεταβάλλοντας δηλαδή τις παραμέτρους του, ώστε να ελαχιστοποιηθεί το σφάλμα. Γενικότερα αυτή η προσέγγιση, καλείται **functional gradient descent** ή **gradient descent with functions**.

Το αποτέλεσμα που λαμβάνεται από το νέο δέντρο, συνυπολογίζεται στο αποτέλεσμα της προηγούμενης συστάδας δέντρων σε μια προσπάθεια ώστε να διορθωθεί ή να βελτιωθεί το τελικό αποτέλεσμα του μοντέλου. Είτε προστίθεται ένας προκαθορισμένος αριθμός δέντρων, είτε η εκπαίδευση σταματά όταν το σφάλμα φτάσει ένα επιθυμητό επίπεδο ή όταν δεν βελτιώνεται πια σε σχέση με κάποιο σετ δεδομένων επικύρωσης.

4.3 Support vector regressions (SVR)

4.3.1 Maximal-Margin Classifier

Ο Maximal-Margin Classifier είναι ένας υποθετικός classifier που δίνει μια πολύ καλή εξήγηση για το πώς δουλευει πρακτικά ένας SVM.

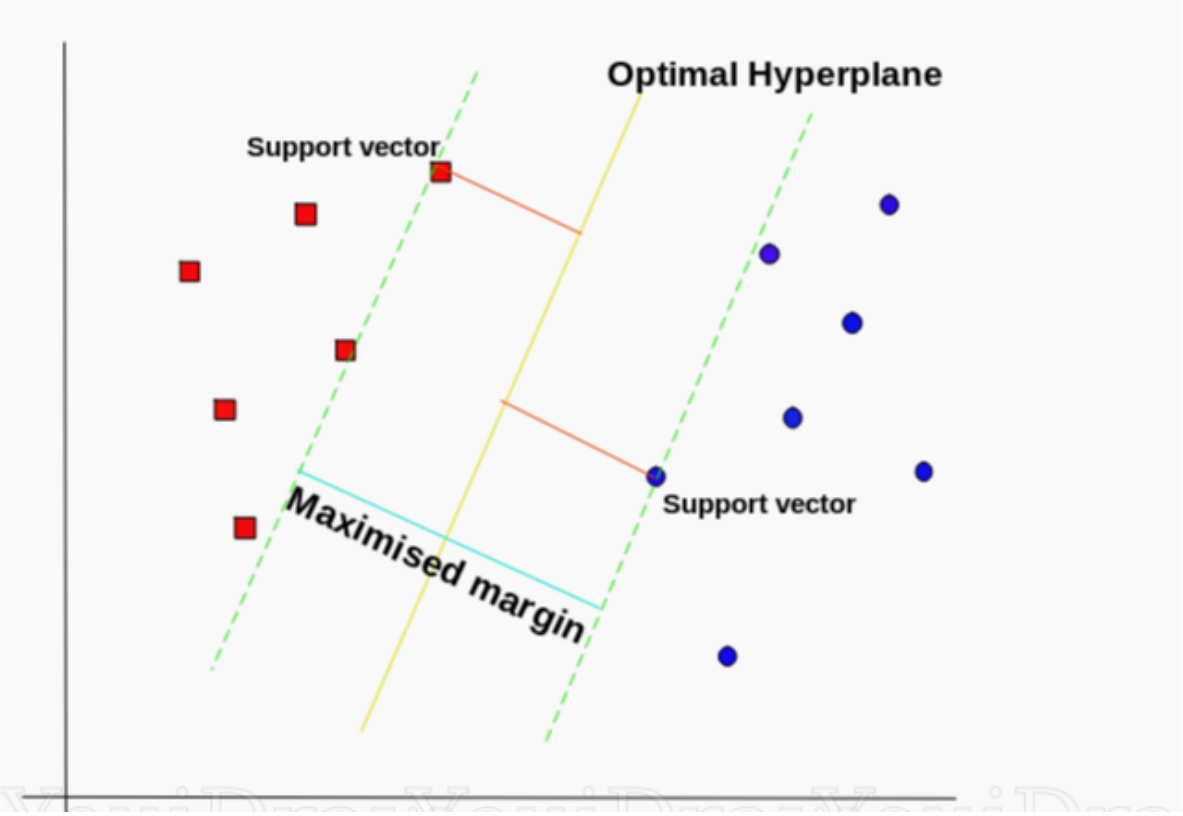
Οι τιμές των εισερχόμενων μεταβλητών (x) των δεδομένων (στήλες), διαμορφώνουν ένα χώρο n - διαστάσεων. Για παράδειγμα αν είχαμε μόνο δύο εισερχόμενες μεταβλητές ο χώρος που θα διαμορφωνόταν θα ήταν δύο διαστάσεων. Το hyperplane (υπερεπίπεδο) είναι μια ”γραμμή”, η οποία στην ουσία διαχωρίζει τον χώρο των εισερχόμενων μεταβλητών. Στον SVM το hyperplane επιλέγεται κατάλληλα, ώστε να διαχωρίσει με τον καλύτερο δυνατό τρόπο τα σημεία των εισερχομένων μεταβλητών βάση της κλάσης. Όταν υπάρχουν μόνο δύο διαστάσεις αυτό οπτικοποιείται εύκολα με μια γραμμή σε ένα δισδιάστατο επίπεδο, όπου υποθέτουμε πως τα εισερχόμενα δεδομένα μπορούν εύκολα να χωριστούν από αυτήν. Ένα τέτοιο παράδειγμα αποτελεί η εξίσωση 4.1 :

$$B_0 + (B_1 * X_1) + (B_2 * X_2) = 0 \tag{4.1}$$

Οι συντελεστές B_1 και B_2 που καθορίζουν την κλίση της ευθείας και το αρχικό σημείο B_0, καθορίζονται από τον learning algorithm, ενώ τα X_1 και X_2 είναι οι τιμές των δύο εισερχομένων μεταβλητών. Με την χρήση της παραπάνω εξίσωσης, μπορεί να γίνει classification νέων σημείων. Με την εισαγωγή νέων τιμών μεταβλητών στη γραμμική εξίσωση, υπολογίζεται αν τα σημεία αυτά είναι πάνω ή κάτω από τη γραμμή. Πάνω από αυτή τη γραμμή, η εξίσωση επιστρέφει μια τιμή που είναι μεγαλύτερη του 0 και το σημείο κατατάσσεται στην πρώτη κλάση. Κάτω από τη γραμμή, η εξίσωση επιστρέφει μια τιμή που είναι μικρότερη του 0 και το σημείο αυτό κατατάσσεται στη δεύτερη κλάση.

Όταν για την εισερχόμενη τιμή επιστρέφεται μια τιμή της εξίσωσης πολύ κοντά στο 0, για το σημείο αυτό είναι πολύ δύσκολο να γίνει classification. Αντίθετα εάν το μέγεθος - magnitude αυτής της τιμής είναι πολύ μεγάλο, το μοντέλο μπορεί να δείξει μεγαλύτερη εμπιστοσύνη στην πρόβλεψη ώστε να το κατηγοριοποιήσει. Η απόσταση μεταξύ της γραμμής αυτής και του κοντινότερου σε αυτή σημείου, αναφέρεται ως margin - περιθώριο. Η καλύτερη ή η βέλτιστη γραμμή που είναι ικανή να διαχωρίσει τις δύο κλάσεις, είναι αυτή με το μεγαλύτερο margin. Αυτό είναι που αποκαλείται ως Maximal-Margin hyperplane.

Το margin υπολογίζεται ως η κάθετη απόσταση από την γραμμή στα πλησιέστερα σημεία. Αυτά τα πλησιέστερα σημεία χρησιμοποιούνται στον καθορισμό της γραμμής και μπορούν να βοηθήσουν στην κατασκευή του classifier. Τα σημεία αυτά είναι που αποκαλούνται support vectors. Υποστηρίζουν ή καλύτερα καθορίζουν το hyperplane βάσει μίας διαδικασίας βελτιστοποίησης που μεγαλώνει το margin, Σχήμα 4.6, [7].



Σχήμα 4.6: Support vectors, hyperplane, max.margin

4.3.2 Support Vector Machines (Kernels)

Όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα, με άλλα λόγια υπάρχει ευθεία στο προηγούμενο παράδειγμα που θα μπορούσε να διαχωρίσει τις δύο κλάσεις, τότε η κατηγοριοποίηση των δεδομένων είναι εύκολη. Όταν όμως τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, τότε εφαρμόζεται η χρήση των kernels. Τα kernels είναι αντιστοιχίσεις από ένα χώρο ισάριθμων διαστάσεων με τα δεδομένα, σε ένα χώρο με περισσότερες διαστάσεις. Ο σκοπός της χρήσης τους είναι να μετασχηματίσουν την απεικόνιση των δεδομένων, έτσι ώστε να υπάρχει ένα hyperplane το οποίο είναι ικανό να κάνει το διαχωρισμό.

Ο αλγόριθμος SVM εφαρμόζεται - υλοποιείται πρακτικά χρησιμοποιώντας ένα kernel. Η δημιουργία του hyperplane στο γραμμικό SVM γίνεται τροποποιώντας το πρόβλημα και χρησιμοποιώντας γραμμική άλγεβρα, όπως εξηγήθηκε στην εισαγωγή του SVM. Μια δυναμική εικόνα είναι πως ο γραμμικός SVM μπορεί να επαναδιατυπωθεί χρησιμοποιώντας το εσωτερικό αποτέλεσμα της εξίσωσης οποιωνδήποτε δύο παρατηρήσεων, πλέον των

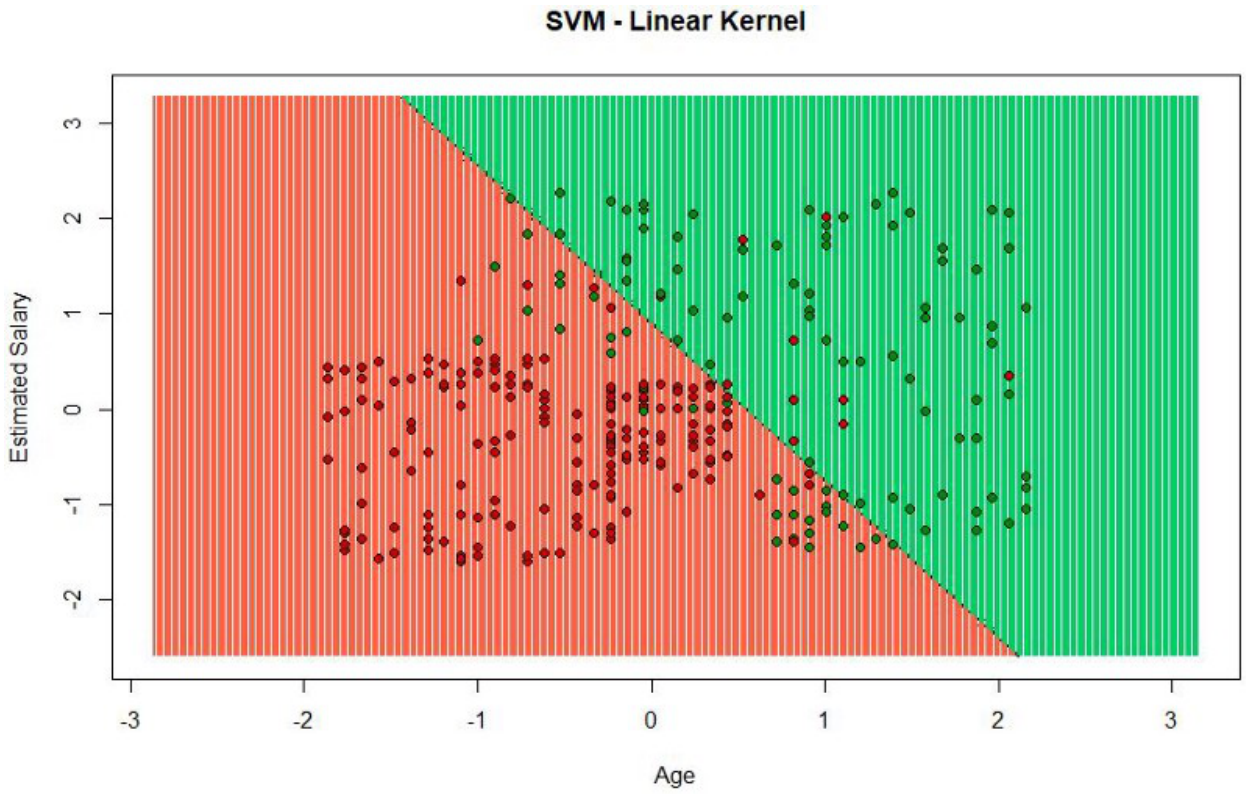
ίδιων των παρατηρήσεων.

4.3.3 Linear Kernel SVM

Το παραγόμενο αποτέλεσμα - σημείο, ονομάζεται kernel και μπορεί να γραφτεί και ως:

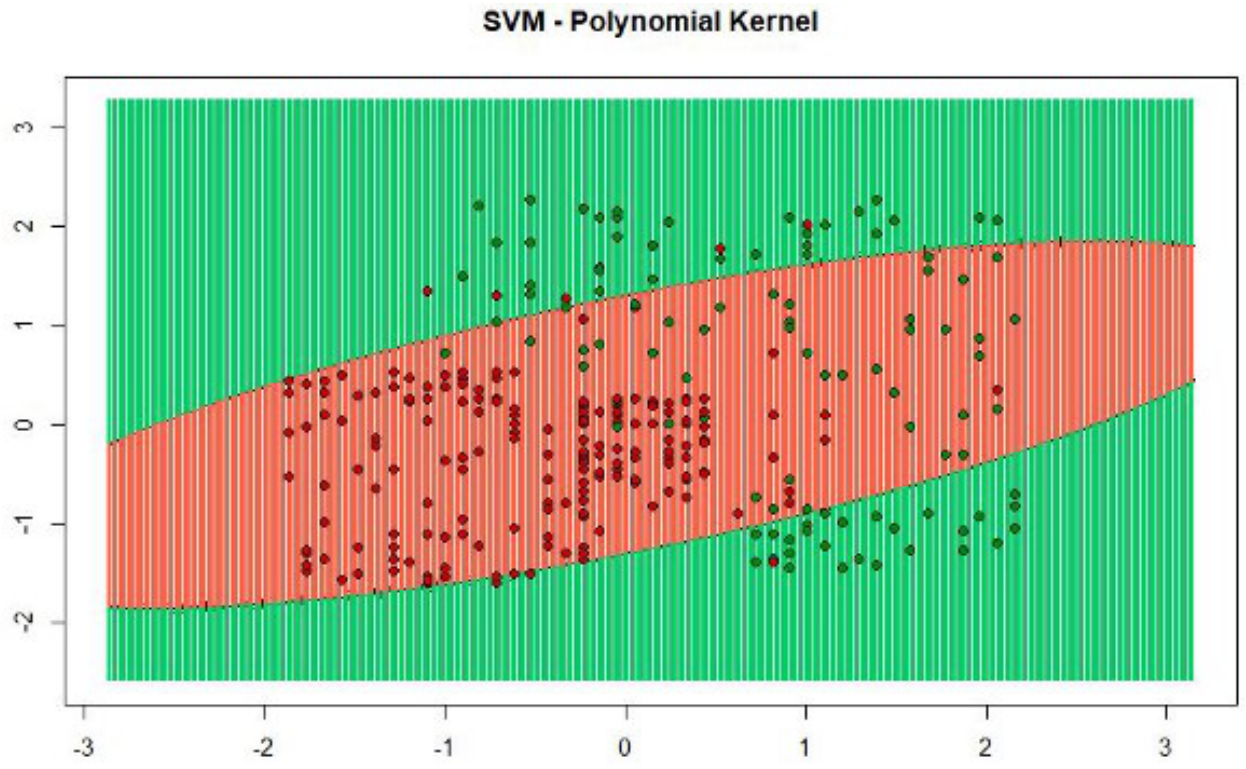
$$K(x, x_i) = sum(x * x_i) \tag{4.2}$$

Το kernel καθορίζει την ομοιότητα και είναι ένα μέτρο της απόστασης μεταξύ του νέου δεδομένου και των support vectors. Το παραγόμενο σημείο είναι το μέτρο ομοιότητας που χρησιμοποιείται για τον linear SVM ή ένα linear kernel, καθώς η απόσταση αυτή είναι ένας γραμμικός συνδυασμός των εισερχομένων δεδομένων, Σχήμα 4.7, [19].



Σχήμα 4.7: Linear SVM

Μπορούν να χρησιμοποιηθούν και άλλα kernels που μετατρέπουν το χώρο των εισερχομένων δεδομένων σε άλλο μεγαλύτερων διαστάσεων, όπως είναι οι Polynomial Kernel, Σχήμα 4.8, [19] και το Radial Kernel, Σχήμα 4.9, [19]. Αυτό αποκαλείται Kernel Trick.



Σχήμα 4.8: Polynomial Kernel

Είναι επιθυμητό να χρησιμοποιούνται τα πιο πολύπλοκα kernels, καθώς αυτά επιτρέπουν καλύτερα τις γραμμές να διαχωρίσουν τις κλάσεις που είναι κυρτές ή ακόμη πιο πολύπλοκες. Με τη σειρά της αυτή η εφαρμογή, μπορεί να οδηγήσει σε πιο ακριβείς classifiers.

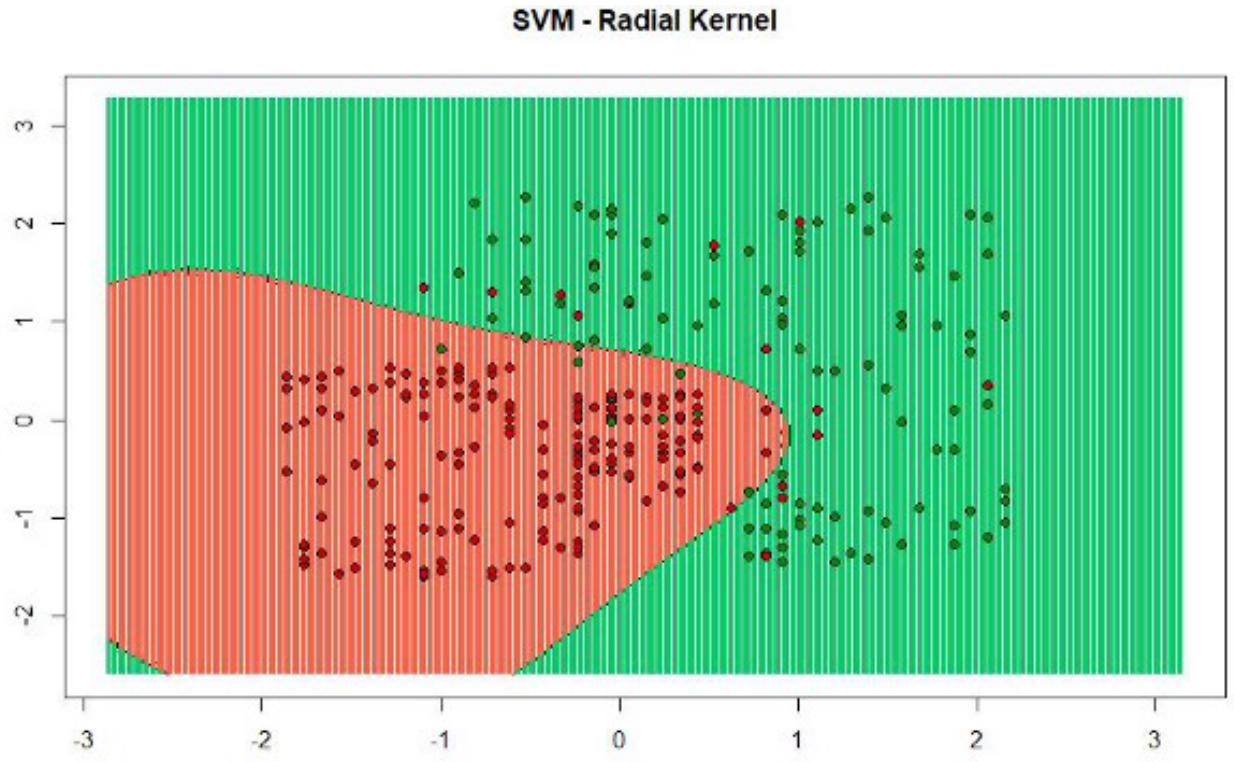
Αντί για το παραγόμενο αποτέλεσμα - σημείο, μπορούμε να κάνουμε χρήση ενός polynomial kernel, όπως για παράδειγμα :

$$K(x, x_i) = 1 + sum(x * x_i)^d \tag{4.3}$$

Όπου ο βαθμός του πολυώνυμου d, θα πρέπει να καθοριστεί με το χέρι στον αλγόριθμο εκμάθησης. Όταν ο βαθμός του πολυωνύμου είναι ίσος με την μονάδα, d=1 τότε είναι το ίδιο με το linear kernel. Τα polynomial kernel εισάγουν με μεγαλύτερη ακρίβεια τις καμπύλες - κυρτές γραμμές στον χώρο.

Τέλος, μπορούμε να κάνουμε χρήση ενός ακόμη πιο πολύπλοκου kernel, του radial kernel. Για παράδειγμα :

$$K(x, x_i) = exp(-gamma * sum((x–x_i^2)) \tag{4.4}$$



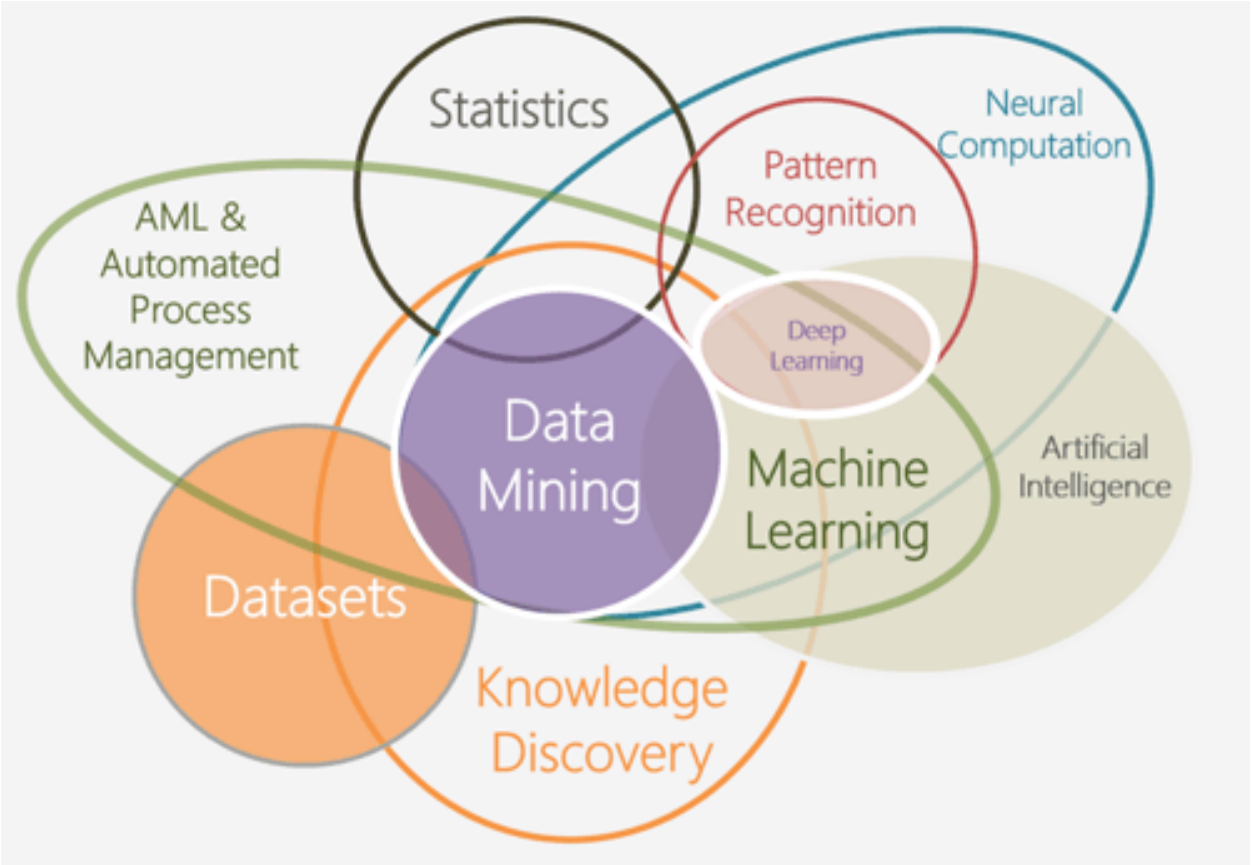
Σχήμα 4.9: Radial Kernel

Όπου το gamma είναι μια παράμετρος που πρέπει να καθοριστεί από τον αλγόριθμο εκμάθησης. Μια καλή προεπιλεγμένη τιμή για την παράμετρο gamma είναι το 0.1, καθώς το gamma είναι συνήθως 0 < gamma < 1. Το radial kernel έχει πολύ τοπικό χαρακτήρα και έτσι είναι ικανό να δημιουργήσει πολύπλοκες περιοχές εντός του επιθυμητού χώρου.

4.4 Dimensionality reduction

Το Data Mining χρησιμοποιείται ως όρος για να περιγράψει το σύνολο της διαδικασίας εξόρυξης γνώσης, από βάσεις δεδομένων. Το ερευνητικό του πεδίο είναι μια τομή μεθόδων και εργαλείων που προέρχονται από τη στατιστική, τη μηχανική μάθηση, βάσεις και αποθήκες δεδομένων. Σκοπός είναι η κατασκευή προγραμμάτων ηλεκτρονικών υπολογιστών τα οποία να είναι ικανά να κατασκευάσουν ισχυρά πρότυπα, που θα χρησιμοποιηθούν για να κάνουν ακριβείς προβλέψεις για μελλοντικά δεδομένα. Για να είναι τα αποτελέσματα πιο ασφαλή σε πρακτικά προβλήματα, πρέπει να γίνει συνδυασμός της μηχανικής μάθησης και της στατιστικής ανάλυσης, Σχήμα 4.10. Η εύρεση ισχυρών προτύπων εκτός της πρόβλεψης βοηθάει και στη γενίκευση από ένα δείγμα στο πλήρες σύνολο, καθώς και για

τη συμπίεση μεγάλων δεδομένων σε μικρότερα με σκοπό να γίνουν πιο κατανοητά και πιο χρήσιμα.



Σχήμα 4.10: Data mining

Τα δεδομένα συνήθως είναι ετερογενή και πολλών διαστάσεων και το data mining βοήθησε στην παραγωγή μεγάλου αριθμού εξειδικευμένων αλγορίθμων. Μια απαίτηση που καλούνται να υποστηρίξουν οι αλγόριθμοι, είναι η μεγάλη ποικιλία των τύπων των δεδομένων, η οποία επιβάλλει τη δημιουργία συγκεκριμένων κάθε φορά αλγορίθμων, μη επιτρέποντας την ύπαρξη ενός μοναδικού και ταυτόχρονα αποτελεσματικού συστήματος data mining για πολλά προβλήματα. Οφείλουν επίσης να είναι αποτελεσματικοί και να έχουν τη δυνατότητα κλιμάκωσης σε μεγάλες βάσεις δεδομένων και σε ικανοποιητικό χρόνο. Επιπλέον θα πρέπει να είναι ικανοί να διαχειριστούν το θόρυβο καθώς και τα **outliers**, ουτώς ώστε να μπορούν με ακρίβεια να δώσουν εικόνα των δεδομένων, βοηθώντας έτσι τον αναλυτή να οδηγηθεί σε ασφαλή συμπεράσματα. Επειδή τα αποτελέσματά τους δεν μπορει να είναι τέλεια, θα πρέπει να μπορούν να τα εκφράσουν με μέτρα αβεβαιότητας σε μορφή προσεγγιστικών ή ποσοτικών κανόνων. Αυτό είναι βασική απαίτηση καθώς έτσι μπορεί να

αξιολογηθεί η ποιότητα της γνώσης που παρέχουν και η αξιοπιστία των αποτελεσμάτων, ώστε να κατασκευαστούν τα κατάλληλα στατιστικά μοντέλα.

Στα περισσότερα προβλήματα τα δεδομένα είναι πολλών διαστάσεων, με πολλές εγγραφες και εκατοντάδες μεταβλητές. Η ύπαρξη τόσων μεταβλητών θα μπορούσε να θεωρηθεί ως πλεονέκτημα αυξάνοντας την απόδοση στη διαδικασία μάθησης. Στην πράξη όμως μεγάλος αριθμός μεταβλητών που περιλαμβάνει και μη σχετικές μεταβλητές, μπορεί να προκαλέσει σύγχυση στους αλγορίθμους και μείωση τελικά της απόδοσής τους. Επιπλέον η χρήση μεγάλου αριθμού μεταβλητών για τη μοντελοποίηση μιας μεταβλητής απόκρισης μπορεί να παραβεί την αρχή της φειδωλότητας, περιπλέκοντας την ερμηνεία της ανάλυσης. Σύνηθες πρόβλημα που δημιουργείται από την ύπαρξη πολλών μεταβλητών είναι και η υπερπροσαρμογή του συστήματος overfitting, η οποία παρεμποδίζει την γενικότητα των αποτελεσμάτων, καθώς τα νέα δεδομένα δεν εχουν την ίδια συμπεριφορά με τα δεδομένα εκπαίδευσης για όλες τις μεταβλητές.

Για τους παραπάνω λόγους λοιπόν, γίνεται εφαρμογή μεθόδων μείωσης των διαστάσεων της βάσης δεδομένων. Οι μέθοδοι αυτές χρησιμοποιούν τις συσχετίσεις μεταξύ των μεταβλητών για τη μείωση του αριθμού τους, επιβεβαιώνοντας πως οι εναπομείνασες είναι ανεξάρτητες και ικανές να ερμηνεύσουν τα αποτελέσματα. Δύο τέτοιες μέθοδοι είναι η ανάλυση κυρίων συνιστωσών (principal component analysis) και η παραγοντική ανάλυση (factor analysis) [45].

4.4.1 Principal Component Analysis (PCA)

Ο κύριος σκοπός του περιορισμού των διαστάσεων (Dimension reduction) των στοιχείων, είναι η δημιουργία μιας μικρότερης ομάδας αντιπροσωπευτικών μεταβλητών. Η μείωση αυτή γίνεται με την αντικατάσταση του συνολικού πλήθους των εμπλεκόμενων μεταβλητών, με ένα μικρότερο αριθμό νέων μεταβλητών, οι οποίες δεν έχουν καθόλου ή έχουν πολύ μικρή συσχέτιση μεταξύ τους, αλλά έχουν τη δυνατότητα να παρέχουν πολύ ισχυρή πληροφόρηση εκμεταλλευόμενες τα χαρακτηριστικά των αρχικών μεταβλητών.

Η ανάλυση κύριων συνιστωσών είναι μία στατιστική διαδικασία, η οποία μετατρέπει μία ομάδα τιμών (παρατηρήσεων) δυνητικά συσχετιζόμενων μεταβλητών, σε μία ομάδα νέων τιμών μη γραμμικά συσχετιζόμενων μεταβλητών, οι οποίες καλούνται κύριες συνιστώσες. Ο αριθμός των νέων μεταβλητών που προκύπτει, είναι ίσος ή και συχνότερα πολύ μικρότερος, από τον αριθμό των αρχικών μεταβλητών. Η μετάβαση αυτή πραγματοποιείται με τέτοιο τρόπο ώστε, η πρώτη συνιστώσα να εξηγεί τη μέγιστη δυνατή διακύμανση που ανα-

πτύσσεται μεταξύ των αρχικών μεταβλητών, η δεύτερη μη συσχετιζόμενη με την πρώτη, να εξηγεί ένα σημαντικό μέρος αυτής, αλλά πάντα μικρότερο της πρώτης κοκ.

Με την ανάλυση κυρίων συνιστωσών πολλές μεταβλητές δύνανται να προσδιοριστούν από κάποιες νέες συνιστώσες που προέρχονται από αυτές, ώστε να έχουν μεγάλη στατιστική αξία. Θα πρέπει δηλαδή οι νέες συνιστώσες να είναι ικανές να περιγράψουν με μεγαλύτερη ακρίβεια το δυνητικό αποτέλεσμα των αρχικών μεταβλητών. Η πρώτη συνιστώσα θα πρέπει να είναι ικανή να περιγράψει ικανοποιητικά στατιστικά τη δράση κάποιων απο τις αρχικές μεταβλητές, η δεύτερη τη δράση διαφορετικών μεταβλητών από την πρώτη κ.ο.κ. Η επιτυχία της μεθόδου είναι πως μπορεί να περιορίσει στο ελάχιστο την αυτοτέλεια κάθε αρχικής μεταβλητής και να δεσμεύσει με αυτόν τον τρόπο το αποτέλεσμά της.

Η ανάλυση κύριων συνιστωσών αποτελεί την απλούστερη και πλέον διαδεδομένη πολυμεταβλητή ανάλυση και στοχεύει στην ανεύρεση από ένα πλήθος p μεταβλητών, ορισμένων νέων ολιγάριθμων μεταβλητών, οι οποίες έχουν την ιδιότητα να είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και παράλληλα να μη συσχετίζονται μεταξύ τους. Το μεγάλο πλεονέκτημά τους έγκειται στην ιδιαιτερότητα που διαθέτουν, λόγω της ανάλυσης, να εξηγούν πολύ μεγάλο ποσοστό της ολικής μεταβλητότητας που αναπτύσσεται μεταξύ των p μεταβλητών, το οποίο τελικά κατανέμεται σε μερικές μόνο νέες μεταβλητές. Έτσι, το μέγιστο μέρος της πληροφόρησης που θα αντλούνταν αν λαμβάνονταν υπόψη οι p μεταβλητές διατηρείται, με τη δημιουργία αυτών των νέων μεταβλητών.

Η διαδικασία της ανάλυσης βασίζεται στην ακόλουθη αρχή:

Από τις p μεταβλητές $X_1, X_2, ..., X_p$, δημιουργούνται p συνδυασμοί αυτών $Z_1, Z_2, ..., Z_p$, με τέτοιο τρόπο ώστε να μη συσχετίζονται μεταξύ τους. Η απουσία συσχετισμού μεταξύ των μεταβλητών Z_i προδιαθέτει ότι αυτές μετρούν διαφορετικές «διαστάσεις» των στοιχείων.

Οι διακυμάνσεις (μεταβλητότητα) που αναπτύσσονται μεταξύ των μεταβλητών Z_i, διαβαθμίζονται με τέτοιο τρόπο ώστε η πρώτη μεταβλητή Z_1 επιλέγεται να εξηγεί ένα όσο το δυνατόν μέγιστο ποσοστό της ολικής μεταβλητότητας, η Z_2 ένα δεύτερο μέγιστο ποσοστό αυτής κοκ., υπακούοντας στη σχέση: $\lambda_1 > \lambda_2 > ... > \lambda_p$, όπου λ_i η i ποσότητα της διακύμανσης. Οι νέες μεταβλητές Z_i καλούνται κύριες συνιστώσες και με τον τρόπο αυτό δημιουργούνται ολιγάριθμες Z συνιστώσες, οι οποίες ωστόσο, είναι κατάληλες να αντιπροσωπεύσουν μεγάλο ποσοστό της συνολικής διακύμανσης $\sum \lambda_1$.

Ταυτόχρονα, πολυάριθμες δευτερεύουσες συνιστώσες εξηγούν μικρό έως ελάχιστο ποσοστό της συνολικής διακύμανσης και συνεπώς το στατιστικό τους αποτέλεσμα μπορεί να

αγνοηθεί, χωρίς την απώλεια ουσιαστικής πληροφόρησης. Η τεχνική των κύριων συνιστωσών έχει ως βάση, κατά τη διαδικασία υπολογισμού της, τη μήτρα των κατά ζεύγη συσχετίσεων (correlation matrix) των μεταβλητών. Κατά συνέπεια, για να θεωρείται η τεχνική επιτυχημένη, να παρέχει δηλαδή ουσιώδη πληροφόρηση, απαραίτητη προϋπόθεση είναι κάποιοι συντελεστές συσχέτισης των αρχικών μεταβλητών της μήτρας συσχετίσεων να φέρουν υψηλές τιμές θετικές ή αρνητικές (π.χ. r≥±0,700).

Έτσι, είναι δυνατό ένα σύνολο 20 έως 30 μεταβλητών να είναι σε θέση να αντιπροσωπευτεί από δύο έως τρεις κύριες συνιστώσες, αρκεί να καλύπτεται η προϋπόθεση της παρουσίας υψηλών συντελεστών στη μήτρα των συσχετίσεων. Από την άλλη πλευρά, αρχικές μεταβλητές με πολύ ισχυρές τιμές συσχετίσεων >±0,990 θεωρούνται πλεονάζουσες και κάποιες από αυτές θα πρέπει να απορρίπτονται, πριν από την εφαρμογή της μεθόδου.

Συνοψίζοντας, τα στάδια της ανάλυσης των κύριων συνιστωσών έχουν ως εξής:

1. Τυποποίηση των αρχικών μεταβλητών $X_1, X_2, ..., X_p$, έτσι ώστε να έχουν μέσο όρο μηδέν και διακύμανση ίση με 1.

2. Υπολογισμός της μήτρας των συνδιακυμάνσεων, η οποία πλέον έχει την έννοια της μήτρας των συσχετίσεων.

3. Εκτίμηση των χαρακτηριστικών ριζών $\lambda_1, \lambda_2, ..., \lambda_p$, και των συντελεστών στάθμισης a_{ij} ή καλύτερα των διανυσμάτων $a_1, a_2...a_p$. Οι συντελεστές της κύριας συνιστώσας i εμφανίζονται με το διάνυσμα a_i και η διακύμανση αυτής με τη χαρακτηριστική ρίζα λ_i.

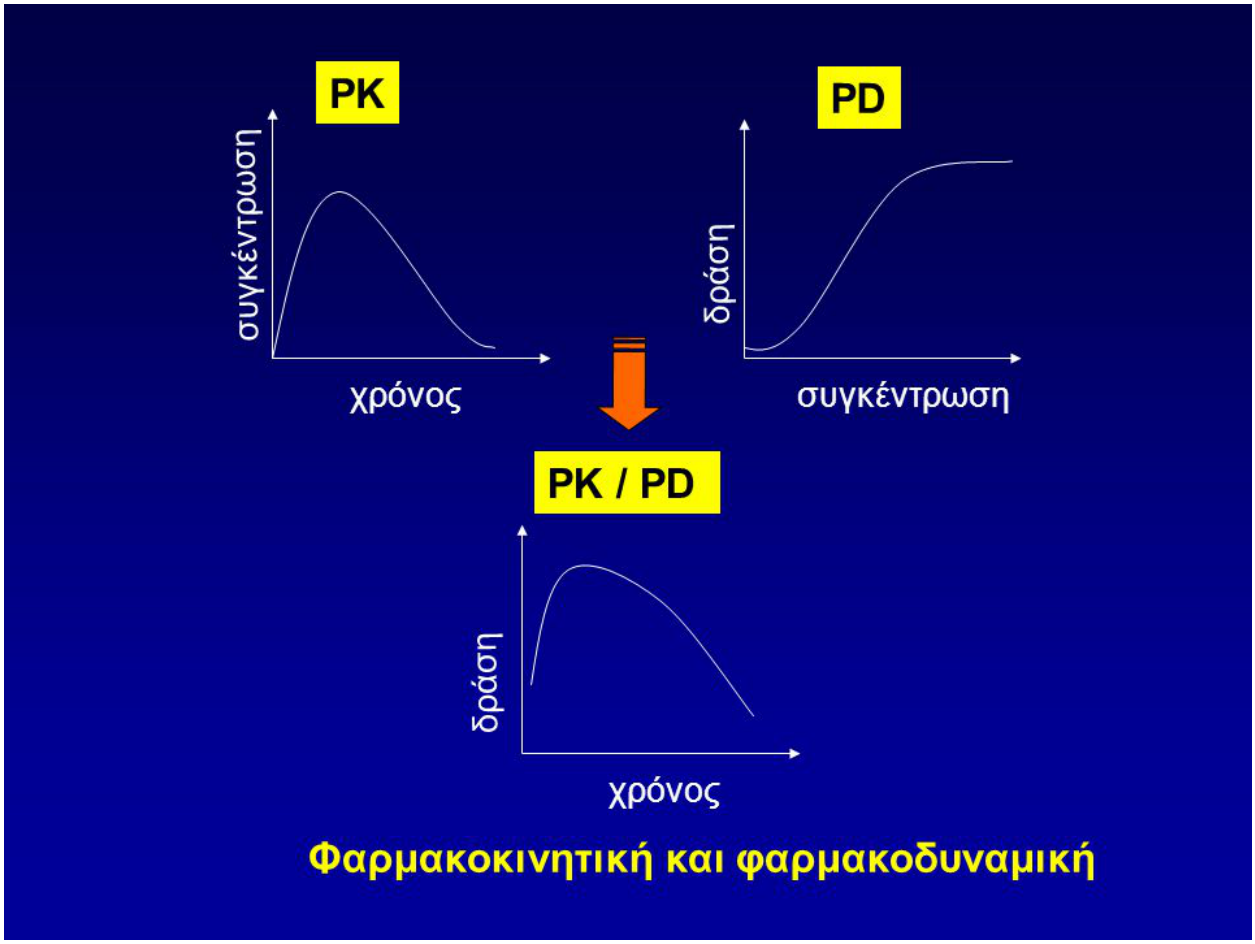
4. Απόρριψη όλων των συνιστωσών που εξηγούν μικρό ποσοστό της ολικής μεταβλητότητας και επιλογή μόνο των πλέον σημαντικών.

5. Χρήση των επιλεγμένων συνιστωσών για την έκφραση των δεδόμενων, σε ένα χώρο λιγότερων διαστάσεων.

4.5 Ορισμός και χρήση των QSP και PK/PD μοντέλων

Η φαρμακοκινητική (PK) δίνει πληροφορίες για την συγκέντρωση της φαρμακευτικής ουσίας συναρτήσει του χρόνου παραμονής της στον οργανισμό, ως αποτέλεσμα της χορήγησης συγκεκριμένης δόσης ενός φαρμάκου. Η φαρμακοδυναμική (PD) περιγράφει το

αποτέλεσμα της δράσης στον οργανισμό, συγκεκριμένης συγκέντρωσης του φαρμάκου [17], Σχήμα 4.11.



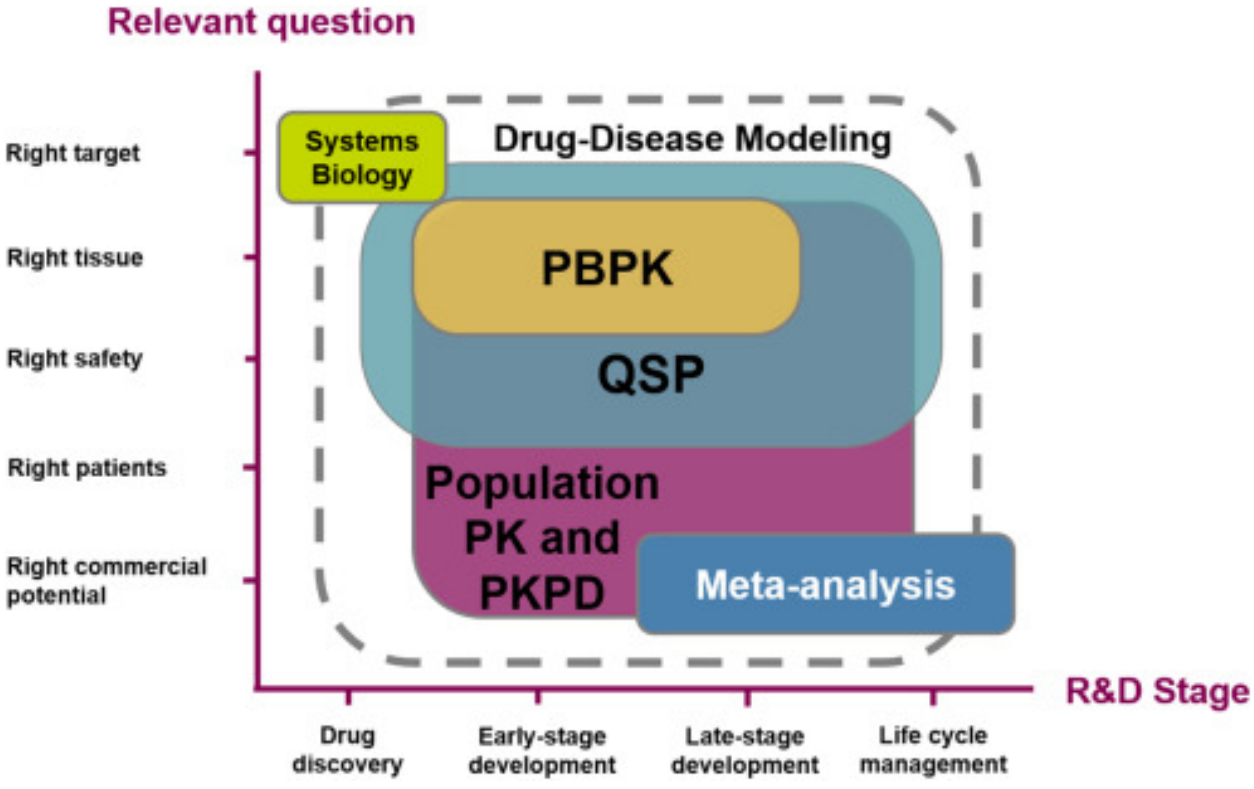
Σχήμα 4.11: Φαρμακοκινητική - Φαρμακοδυναμική

Η λογική για την δημιουργία των PK/PD μοντέλων, είναι να συνδυάσουν τη φαρμακοδυναμική και τη φαρμακοκινητική, ώστε να εδραιωθεί και να αξιολογηθεί μια σχέση δόσης - συγκέντρωσης - φαρμακευτικού αποτελέσματος και ακολούθως να μπορέσει το μοντέλο να περιγράψει και να προβλέψει τον χρόνο επίδρασης που αντιστοιχεί σε μια δόση φαρμάκου. Όταν υπάρχουν σταθέρες φαρμακοκινητικές συνθήκες, οι σχέσεις μεταξύ συγκέντρωσης και φαρμακολογικού αποτελέσματος, μπορούν να περιγραφούν από διάφορα σχετικά απλά φαρμακοδυναμικά μοντέλα. Αυτά μπορεί να είναι τα linear model, long-linear model, το Emax-model και το sigmoid Emax-model. Σε συνθήκες που δεν υπάρχουν σταθερές φαρμακοκινητικές συνθήκες, χρησιμοποιούνται πιο αναβαθμισμένα, πολύπλοκα PK/PD μοντέλα ώστε να μπορέσουν να συνδυάσουν και να υπολογίσουν μια πιθανή συσχέτιση

μεταξύ της συγκέντρωσης στο πλάσμα του αίματος και του παρατηρούμενου φαρμακολογικού αποτελέσματος.

Η Quantitative Systems Pharmacology (QSP) έχει εμφανισθεί προσφάτως ως μία προσέγγιση που μπορεί να συνδυάσει γνώσεις που προέρχονται από διάφορους τομείς συμπεριλαμβανομένων της φαρμακολογίας, των βιολογικών συστημάτων, της φυσιολογίας, των μαθηματικών και της βιοχημείας [13]. Μελετά το μηχανισμό δράσης του φαρμάκου σε σχέση με τους βιολογικούς μηχανισμούς δράσης, προσπαθώντας να βελτιώσει την αντίδραση του ανθρώπινου σώματος στη φαρμακευτική αγωγή, αλλά και στη φαρμακολογία του φαρμάκου. Διαφέρουν από τα PK/PD μοντέλα τόσο στο σκοπό τους, όσο και στα δεδομένα που χρειάζονται για να λειτουργήσουν.

Εμφανίσθηκε τη στιγμή όπου η φαρμακευτική βιομηχανία αντιμετώπιζε πολλές προκλήσεις στην αποδοτικότητα και στην παραγωγή στον R&D τομέα κ έχει τη δυνατότητα να βοηθήσει στην επίλυση πολλών προκλήσεων, Σχήμα 4.12, [9].



Σχήμα 4.12: Σχεδιασμός φαρμάκου με συνδυασμό QSP - PBPK

Αναλογιζόμενοι το αυξανόμενο κόστος της παραγωγής ενός καινούριου φαρμάκου, είναι προφανές πως έπρεπε να χρησιμοποιηθούν καινούριες προσεγγίσεις ώστε να βελ-

τιωθεί η πιθανότητα επιτυχίας ενός νέου σε παραγωγή φαρμάκου, τόσο σε κλινική αποτελεσματικότητα όσο και σε ασφάλεια. Όπως έχει διαπιστωθεί, από όλους τους παράγοντες που επηρεάζουν το κόστος ανάπτυξης - παραγωγής ενός φαρμάκου, ο πιο σημαντικός είναι ο χαμηλός βαθμός επιτυχίας της φάσης II του φαρμάκου.

Μια μελέτη που έγινε το 2014 από το preclinical Pharmacokinetics/Pharmacodynamics (PK/PD) Discussion Group [28], που τους ανατέθηκε από τους Drug Metabolism Leadership Group (DMLG) και το International Consortium (IQ) for Innovation and Quality in Pharmaceutical Development, έδειξε πως τα PK/PD μοντέλα χρησιμοποιούνται ευρέως από τη φαρμακευτική βιομηχανία σε προκλινικές μελέτες καθορίζοντας την επιλογή και βελτιστοποιώντας τις δόσεις και τη δοσολογία, καθώς και την πρόβλέψη της επαρκούς θεραπευτικής δόσης. Ωστόσο παρά το αυξανόμενο ενδιαφέρον στα μοντέλα αυτά στη βιοφαρμακευτικη βιομηχανία με έναν αυξανόμενο αριθμό προσπαθειών από το 2015, ο ρόλος των μοντέλων αυτών δεν έχει φτάσει ακόμη τις πλήρεις δυνατότητες του, συμπεριλαμβανομένων και του ρόλου του στην προκλινική φάση.

Για την καλύτερη κατανόηση της χρήσης του QSP modeling στην προκλινική φάση και της εκτίμησης των γνώσεων που μπορεί να παρέχει στην βιοφαρμακευτική βιομηχανία δημιουργήθηκε, ένα γκρουπ το 2016 αποτελούμενο από αντιπροσώπους από 17 μικρότερες εώς και μεγαλύτερες βιοφαρμακευτικές εταιρείες [21]. Ο σκοπός του γκρουπ για την προκλινική φάση, ήταν τόσο να κατανοήσει τις τρέχουσες προκλήσεις, ευκαιρίες και τα εμπόδια του προκλινικού QSP modeling στην R&D, όσο και να εκτιμήσει την οργανωτική δομή των προκλινικών QSP μοντέλων στην βιομηχανία και τα σημεία επαφής τους με άλλους λειτουργικούς τομείς του R&D τομέα και τους φορείς που είναι αρμόδιοι για την τήρηση του θεσμικού πλαισίου.

Έτσι λοιπόν πραγματοποιήθηκε μία μελέτη ανάμεσα σε 50 φαρμακευτικές εταιρείες το πρώτο εξάμηνο του 2017. Κατά τη διάρκεια της προετοιμασίας για τη μελέτη διαπιστώθηκε πως δεν υπήρχε ξεκάθαρος ορισμός του QSP, πιθανόν λόγω του γεγονότος πως είναι ένας νεοεμφανιζόμενος τομέας και έτσι η ορολογία του προς το παρόν χρησιμοποιείται με μια ευρεία έννοια. Ήταν αναμενόμενο λοιπόν να συνυπάρχει μια ποικιλία ορισμών στον τομέα για το QSP, οπότε αποφασίστηκε από το γκρουπ να συμπεριληφθεί στην μελέτη μια ενότητα ”ορολογίας”, ώστε να συλλεχθούν οι απόψεις των ερευνητών της βιομηχανίας. Το γκρουπ έφτασε στο συμπέρασμα ενός απλού ορισμού, πως το QSP είναι ένα ποσοτικό ή υπολογιστικό πλαίσιο που στηρίζει την ανακάλυψη ενός φαρμάκου και την βιομηχανική ανάπτυξη αυτού, ενσωματώνοντας τις γνώσεις της βιοχημείας, βιολογίας, φυσιολογίας, φαρμακολογίας και κλινικής.

Τα αποτελέσματα αυτής της μελέτης δηλώνουν την σημερινή κατάσταση όσον αφορά την χρήση του QSP modeling στην προκλινική φάση από την φαρμακευτική βιομηχανία, συμπεριλαμβάνοντας τις μελλοντικές δυνάτοτητες, αλλά και τα εμπόδια που μπορεί να εμποδίσουν την ευρεία χρήση του.

Τα QSP μοντέλα δίνουν τη δυνατότητα στους ερευνητές να αξιολογήσουν διάφορες υποθέσεις in-silico που ειδάλλως θα έπρεπε να αξιολογηθούν πειραματικά. Υπάρχει η προσδοκία πως η χρήση των QSP μοντέλων θα βοηθήσει στη μείωση τόσο του κόστους του R&D, όσο και του ρίσκου που σχετίζεται με κάποια κενά ή αβεβαιότητες που εμφανίζονται στις γνώσεις μας, όταν εφαρμόζουμε κάποια καινούρια θεραπεία στους ασθενείς.

Τα QSP μοντέλα τυπικά χρησιμοποιούνται σαν ερευνητικά εργαλεία στην υποθετική παραγωγή και στην ερευνητική κλινική ανάπτυξη φαρμάκου. Ωστόσο προσφάτως ο US FDA χρησιμοποίησε ένα QSP μοντέλο για να εκτιμήσει τη δοσολογία σε μία νέα βιολογική θεραπεία [26]. Σε επαφή με την NPS Pharma, το παράρτημα της κλινικής φαρμακολογίας του FDA χρησιμοποίησε ένα δημοσιευμένο QSP μοντέλο του συστήματος ομοιόστασης του ασβεστίου [25], για να προτείνει ένα εναλλακτικό δοσολογικό σχήμα για το NATPARA, μια ενέσιμη παραθυρεοειδική ορμόνη που χορηγείται ως φάρμακο για να ελέγχει τα χαμηλά ποσοστά ασβεστίου στο πλάσμα του αίματος, σε ασθενείς με υποθυρεοειδισμό. Η χρήση του QSP μοντέλου από τον FDA για το εναλλακτικό δοσολογικό σχήμα είναι μία από τις πιο σημαντικές μελλοντικές εφαρμογές και σημείο ορόσημο, καθώς είναι η πρώτη φορά που ένα τέτοιο μοντέλο χρησιμοποιείται δημόσια από κάποιον ερευνητικό οργανισμό, για να κάνει μια κλινική εκτίμηση για έναν σπόνσορα.

Έχουν γίνει διάφορες έρευνες για το αναπτυσσόμενο πεδίο των QSP με αποτέλεσμα τη δημοσίευση άρθρων όπου περιγράφονται μοντέλα για διαφορετικές ασθένειες, αποδεικνύοντας πώς αυτά τα μοντέλα μπορούν να χρησιμοποιηθούν για να αξιολογήσουν κρίσιμα επιστημονικά ερωτήματα που εμφανίζονται στην έρευνα για τη φαρμακευτική R&D.

Οι Leil and Bertz στην εργασία τους το 2014, περιγράφουν την ιστορία της χρησιμοποίησης των μοντέλων σαν εργαλεία για τη φαρμακολογία και για τις φάσεις ανάπτυξης του φαρμάκου [14]. Τα δύο σημεία που αποδεικνύουν τη σημαντικότητα στην αναπτυσσόμενη QSP προσέγγιση είναι, (i) η δυσκολία που υπάρχει στην ανεύρεση καινούριων θεραπευτικών στόχων και (ii) ο συνεχώς αυξανόμενος χρόνος, αλλά και το αυξανόμενο κόστος που απαιτείται για την ανάπτυξη ενός φαρμάκου και την εισαγωγή αυτού στην κυκλοφορία. Αν τα μοντέλα μπορέσουν να χρησιμοποιηθουν στη διαδικασία ανάπτυξης των φαρμάκων, θα ήταν ένα σημαντικό εργαλείο που θα μπορούσε να συνδυάσει τις γνώσεις μεταξύ των

πειραμάτων που γίνονται π.χ. σε ανθρώπους και σε ζώα και να κάνει πρόβλεψη των αποτελεσμάτων σε συνδυαστικά θεραπευτικά σχήματα, κάτι που θα ήταν αδύνατο να γίνει με κλινικά πειράματα.

Μία πολύ σημαντική εφαρμογή του QSP μοντέλου, είναι η βελτιστοποίηση της θεραπευτικής δόσης και του θεραπευτικού σχήματος. Στον τομέα της ογκολογίας, το θεραπευτικό παράθυρο δεν έχει μεγάλο εύρος και απαιτείται πολύ καλή ρύθμιση της δόσης και του σχήματος. Ενώ απαιτούνται πολύ υψηλές δόσεις αντι-αγγειογενετικών παραγόντων για την μείωση του όγκου και την ελάττωση της αιμάτωσής του, αυτό μπορεί να οδηγήσει παράλληλα και στη χαμηλή διανομή του φαρμάκου στον όγκο με αποτέλεσμα την εκ νεου αγγειογένεση αυτού. Οι Sharan and Woo το 2015 μελέτησαν πως μπορούν να το εμποδίσουν ή να το καθυστερήσουν αυτό να συμβεί, βελτιστοποιώντας το θεραπευτικό σχήμα μέσω ενός QSP μοντέλου για την αγγειογένεση [30].

Άλλη μία εφαρμογή, είναι για την πρόβλεψη σε εξαρτημένο ή αυτόνομο στόχο, της τοξικότητας σε δευτερογενείς ιστούς που μελετάται για πολλά χρόνια καθώς είναι η πιο συχνή αιτία τερματισμού της ανάπτυξης μιας θεραπείας που θεωρούνταν κατά τα αλλά ικανοποιητική. Προκειμένου να γίνει η πρόβλεψη της τοξικότητας με τη χρήση του QSP μοντέλου, είναι ωφέλιμο να συνδυαστεί με ένα PBPK μοντέλο, το οποίο θα προβλέψει τις συγκεντρώσεις του φαρμάκου στο όργανο στόχο. Οι Woodhead et al. και οι Chetty et al., βασιζόμενοι σε PBPK μοντέλα έκαναν πρόβλεψη συγκεντρώσεων φαρμάκων σε ιστούς συνδεόντας τις προβλεπόμενες αυτές τιμές, με φαρμακοδυναμικά αποτελέσματα στον στόχο [36], [5]. Χρησιμοποίησαν τον **PBPK προσομοιωτή Simcyp** [3], που έχει ένα ενσωματωμένο PBPK μοντέλο και επιτρέπει στο χρήστη να εισάγει συγκεκριμένες παραμέτρους του φαρμάκου που έχουν μετρηθεί in vitro, ώστε να γίνει πρόβλεψη των in-vivo τιμών παραμέτρων του φαρμάκου, στο πλάσμα και στον ιστό.

Ένα από τα τεχνικά θέματα που περιορίζουν την εξάπλωση της χρήσης του QSP στη βιοϊατρική έρευνα, είναι η έλλειψη ενός στανταρισμένου μοντέλου που να έχει γίνει αποδεκτό ως εργαλείο από τους ερευνητές. Το εργαλείο αυτό θα έπρεπε να επιτρέπει την αξιολόγηση διαφόρων πειραματικών σεναρίων σε ένα ευέλικτο υπολογιστικό περιβάλλον, ώστε να συνδυάζει την απόδοση μέσω της προσομοίωσης. Μία πιθανή λύση εφαρμόστηκε στην web-based virtual systems pharmacology (ViSP) πλατφόρμα από τους Ermakov et al.[8]. Η βασική λειτουργία της πλατφόρμας έγκειται στη χρήση ενός ευέλικτου εκτελέσιμου αρχείου, το οποίο μπορεί να χρησιμοποιηθεί σε διαφορετικά περιβάλλοντα για την πρόβλεψη διαφόρων παραμέτρων. Επειδή η πλατφόρμα έχει εκπαιδευτεί με αρκετά σετ μοντέλων, δεν εξαρτάται από τη δομή του μοντέλου και από το λογισμικό που χρησιμοποι-

ήθηκε για την ανάπτυξη του μοντέλου, ενώ έχει καταφέρει να λειτουργεί με ευελιξία στις εισερχομένες παραμέτρους.

Αυτά τα χαρακτηριστικά μπορεί να φανούν χρήσιμα στο μέλλον, όταν η χρήση και η ανταλλαγή των QSP μοντέλων θα είναι πλέον ευρέως διαδεδομένα.

Ορισμός QSAR σχέσεων και μοντέλων

Η διαπίστωση για το σημαντικό ρόλο των φαρμακοκινητικών ιδιοτήτων στην τελική αποτελεσματικότητα των φαρμάκων, επέδρασε καθοριστικά στον τρόπο προσέγγισης της διαδικασίας ανακάλυψης νέων φαρμακομορίων. Πρακτικά οδήγησε στο δόγμα ”fail fast, fail cheap”, δηλαδή στον έγκαιρο διαχωρισμό των αποτυχιών πριν προχωρήσει η ανάπτυξη σε δαπανηρές κλινικές δοκιμές. Σύμφωνα με τη νέα αντίληψη οι ιδιότητες Απορρόφηση, Κατανομή, Μεταβολισμός, Απέκκριση, που συνοψίζονται στο ακρωνύμιο ADME (Absorption,Distribution, Metabolism, Elimination), θα πρέπει να εξετάζονται παράλληλα με τη συγγένεια των νέων μορίων προς τον υποδοχέα, από τα πρώιμα στάδια του σχεδιασμού.

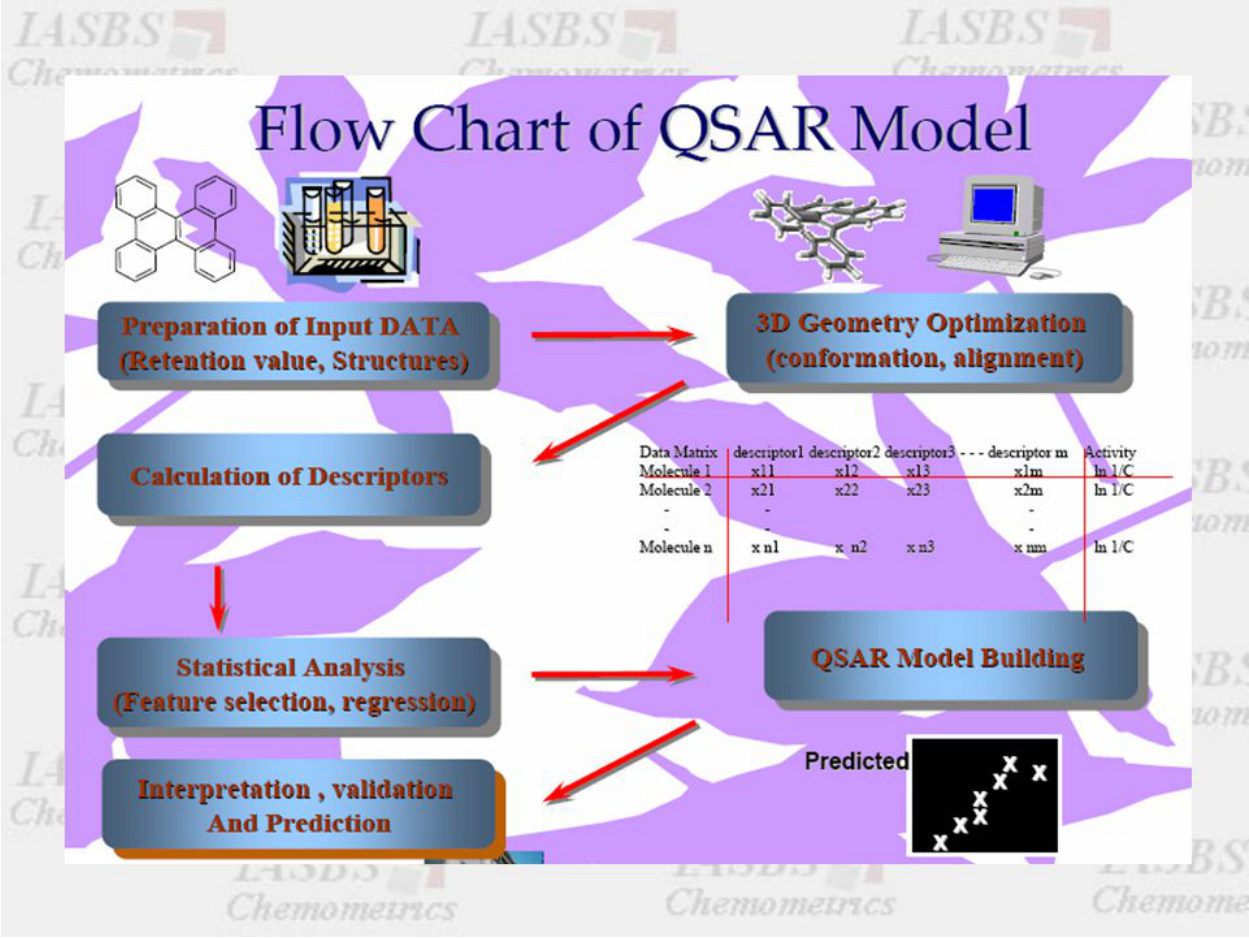
Δεδομένου ότι τόσο οι ιδιότητες ADME όσο και η τοξικότητα επηρεάζονται από τις φυσικοχημικές/μοριακές ιδιότητες, προτάθηκε η έννοια της φαρμακο - ομοιότητας, ενώ αναπτύχθηκαν μέθοδοι υπολογισμού και φιλικές τεχνικές προσδιορισμού συγκεκριμένων ιδιοτήτων που διαμορφώνουν το φυσικοχημικό προφίλ των νέων ενώσεων.

Λόγω των εξελίξεων στη συνθετική χημεία, είναι πολύ πιο εύκολη η σύνθεση νέων ενώσεων και ως εκ τούτου δημιουργήθηκε η ανάγκη και σε αυτό το επίπεδο να εφαρμοσθεί ένας περισσότερο ορθολογικός σχεδιασμός. Αυτό απαιτεί την ποσοτικοποίηση της διαφοροποίησης που επιφέρουν στη δράση οι δομικές αλλαγές των διαφορετικών ενώσεων. Η ποσοτικοποίηση αυτή είναι δυνατή με εφαρμογή της Υπολογιστικής Χημείας, που επιτρέπει την περιγραφή της δομής των μορίων και την αξιοποίηση πληθώρας πληροφοριών μέσω στατιστικής επεξεργασίας των δεδομένων που σχετίζονται με τη δομή.

Με αυτή τη διαδικασία, κατασκευάζονται μοντέλα, γνωστά ως μοντέλα Ποσοτικών Σχέσεων Δομής – Δράσης. Η αξιοποίηση των Ποσοτικών Σχέσεων Δομής - Δράσης (Quantitative Structure Activity Relationships, QSAR), αποσκοπεί στην εξαγωγή εξισώσεων ή μοντέλων που συσχετίζουν τη βιολογική δράση με τη δομή. Τα μοντέλα αντλούν και αξιοποιούν όσο το δυνατόν περισσότερες πληροφορίες, καθιστώντας έτσι λιγότερο αναγκαία τη χρήση κοστοβόρων και χρονοβόρων πειραμάτων. Στοχεύουν στην πρόβλεψη της βιολογικής δραστικότητας, έτσι ώστε οι νέες συνθέσεις να οδηγούνται με περισσότερη ασφάλεια σε βελτιωμένα φαρμακομόρια. Συμβάλλουν σημαντικά στην κατανόηση του μηχανισμού δράσης

των φαρμακομορίων και έχουν προσφέρει καθοριστικά στη δημιουργία των σύγχρονων αντιλήψεων για το σχεδιασμό των φαρμάκων.

Ο απώτερος στόχος όμως των Ποσοτικών Σχέσεων Δομής - Δράσης και το βασικό κίνητρο για την ανάπτυξή τους, ήταν και παραμένει η πρόβλεψη της δράσης που θα κατευθύνει τον φαρμακοχημικό στη σύνθεση (ή τη μη σύνθεση), νέων παραγώγων. Από την ανάλυση μικρών σειρών δομικώς συγγενών ενώσεων και την εξαγωγή μοντέλων με απλή στατιστική επεξεργασία λίγων παραμέτρων, ο τομέας QSAR διευρύνθηκε ώστε να περιλαμβάνει μεγάλες σειρές δεδομένων και εξόρυξη πληροφοριών από πληθώρα περιγραφικών μεταβλητών (descriptors), που μπορούν να αναλυθούν με προηγμένα υπολογιστικά προγράμματα, Σχήμα 4.13.



Σχήμα 4.13: Απεικόνιση μοντέλου QSAR

Οι Ποσοτικές Σχέσεις Δομής - Δράσης αποτελούν σήμερα σημαντικό εργαλείο στο σχεδιασμό νέων υποψήφιων φαρμάκων, στοχεύοντας αφενός στην κατανόηση του μη-

χανισμού δράσης και αφετέρου στη μείωση του ποσοστού αποτυχίας και κατά συνέπεια του κόστους της φαρμακευτικής έρευνας. Είναι δυνατόν να βασίζονται σε δύο διαστάσεις μοριακής απεικόνισης (2-D QSAR) με εξαγωγή εξισώσεων μοντέλων, ή σε τρεις διαστάσεις (3-D QSAR), όπως είναι η μεθοδολογία CoMFA και CoMSIA. Πολύ σημαντικό για να μπορέσει να λειτουργήσει οποιοδήποτε μοντέλο πρόβλεψης, είναι να γίνει σωστή επιλογή κι επεξεργασία των δεδομένων, προκειμένου το μοντέλο πρόβλεψης να είναι ακριβές.

5. ΥΛΟΠΟΙΗΣΗ

5.1 Δημιουργία σετ δεδομένων

5.1.1 Συλλογή δεδομένων

Στο πλαίσιο της παρούσας εργασίας η συλλογή των δεδομένων για τη δημιουργία της βάσης δεδομένων του μοντέλου έγινε από την Drugbank [35]. Η Drugbank είναι μια ανοιχτή και ελεύθερα προσβάσιμη online βάση δεδομένων. Περιέχει πολύ σημαντικές πληροφορίες για τα φάρμακα, τις χημικές δομές και τις φυσικοχημικές / φαρμακοκινητικές ιδιότητες αυτών.

Τα δεδομένα της έχουν χρησιμοποιηθεί ευρέως από το 2006 που δημιουργήθηκε για το σχεδιασμό φαρμάκων και την εύρεση των βιολογικών τους στόχων in silico, για την πρόβλεψη του μεταβολισμού φαρμακευτικών ουσιών, αλληλεπιδράσεων αυτών και γενικότερα στη φαρμακευτική εκπαίδευση και βιομηχανία. Η έκδοση 5.0 που χρησιμοποιήθηκε για την εργασία, περιέχει συνολικά 8261 φάρμακα. Από αυτά τα 2021 είναι εγκεκριμένα από τον FDA φάρμακα μικρής σχετικά δομής (έχει υπολογισθεί πως M.B. ίσο με 1000 είναι το όριο πάνω από το οποίο μία ουσία πολύ δύσκολα μπορεί να διαχυθεί, τα περισσότερα φάρμακα έχουν M.B. μέχρι 500). Περιέχονται επίσης 233 βιοτεχνολογικά φάρμακα (πρωτεΐνες/πεπτίδια) εγκεκριμένα από τον FDA, και 94 nutraceuticals (ενώσεις που είτε είναι τρόφιμα, είτε μέρη τροφίμων και δε συμπληρώνουν απλά τη διατροφή, αλλά επίσης στοχεύουν στην πρόληψη ή/και θεραπεία κάποιας ασθένειας ή/και διαταραχής). Τέλος στη βάση δεδομένων υπάρχουν και πάνω από 6000 φάρμακα τα οποία ακόμη βρίσκονται σε πειραματικό στάδιο. Για κάθε φάρμακο στη βάση, υπάρχουν πάνω από 200 πεδία που το χαρακτηρίζουν. Τα μισά σχεδόν δίνουν πληροφορίες για τη φυσικοχημική δομή και τη φαρμακευτική συμπεριφορά της ουσίας, ενώ τα άλλα μισά για τους βιολογικούς της στόχους.

Στην εργασία χρησιμοποιήθηκαν δεδομένα που υπάρχουν στη βάση για την απορρόφηση, τη διάλυση, τη χημική δομή των φαρμάκων καθώς και άλλα δεδομένα που προσδιορίζουν τη δομή μοριακά, γεωμετρικά, ηλεκτροστατικά κτλ και βοηθούν στην ποσοτικοποίηση της σχέσης Δομής - Δράσης. Συγκεκριμένα έγινε εξαγωγή των εξής παραμέτρων :

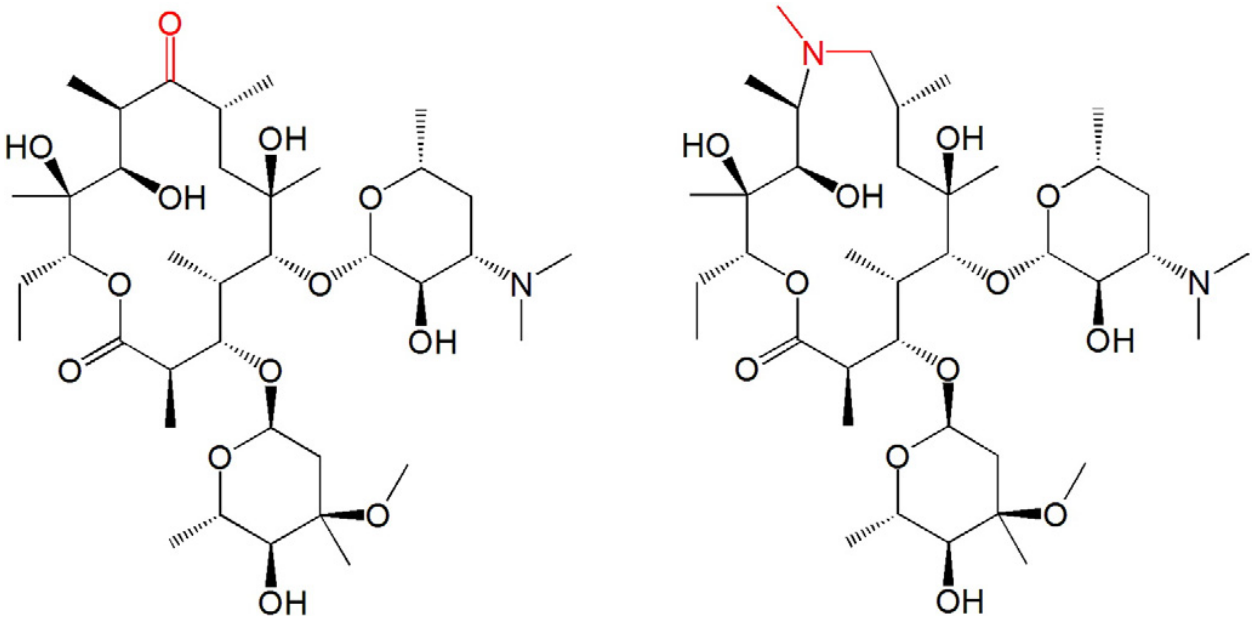
- HLT (half-life)

- logP (measure of lipophilicity)
- logS (measure of solubility)
- psa (polar surface area)
- rbc (rotatable bond count)
- nbr (number of rings rule of five)
- hbac (hydrogen acceptor count)
- hbdc (hydrogen donor count)
- mw (molecular weight)
- ref (molar refractivity)
- pol (polarizability)
- smiles (Simplified Molecular Input Line Entry System)
- pbn (binding protein of drugs)

Στη συνέχεια ακολούθησε επεξεργασία των δεδομένων για αξιοποίησή τους σε περαιτέρω ανάλυση.

Εχει διαπιστωθεί πως ο χρόνος ημίσειας ζωής HLT ενός φαρμάκου, είναι μία πολύ σημαντική φαρμακοκινητική παράμετρος, καθώς μας υποδεικνύει το χρονικό διάστημα κατά την διάρκεια του οποίου η συγκέντρωση του φαρμάκου ελαττώνεται στο 50% της δόσης του. Τα φάρμακα χορηγούνται σε δόσεις οι οποίες επαναλαμβάνονται ανά τακτά χρονικά διαστήματα, έτσι ώστε η συγκέντρωση του φαρμάκου να παραμένει σταθερή στον οργανισμό για να υπάρξει το αναμενόμενο θεραπευτικό αποτέλεσμα. Ο χρόνος ημίσειας ζωής βοηθάει στον καθορισμό του δοσολογικού σχήματος, ώστε η συγκέντρωση του φάρμακου να είναι σταθερή κι έχει άμεση συσχέτιση με την κυκλοφορία του φαρμάκου στο πλάσμα αίματος, αλλά και με το προφίλ της συγκέντρωσης του φαρμάκου με το χρόνο, μετά από επαναλαμβανόμενες δόσεις. Λόγω της σημαντικότητας της παραμέτρου αυτής διατηρήθηκαν μόνο οι εγγραφές της βάσης δεδομένων, οι οποίες περιείχαν πληροφορία σχετικά με το χρόνο ημίσειας ζωής. Έτσι διατηρήθηκαν στο dataset 1138 ενώσεις που ικανοποιούν το παραπάνω κριτήριο.

Κάθε φάρμακο έχει διαφορετική HLT τιμή. Ακόμη και φάρμακα με παρόμοια χημική δομή και ίδια φαρμακολογική δράση μπορεί να έχουν πολύ σημαντικές διαφορές στις τιμές HLT σχήμα 5.1, [16]. Η τιμή της παραμέτρου στη βάση δεν είναι αυστηρά ορισμένη, αλλά δίνεται με εύρος τιμών για κάθε ουσία. Οπότε έγινε επεξεργασία των τιμών HLT του dataset, βάσει των παρακάτω κανόνων:



Σχήμα 5.1: Χρόνος ημίσειας ζωής ερυθρομυκίνης (αριστερά) 0,8 – 3 h και της αζιθρομυκίνης (δεξιά) 35 - 41h.

- όπου αναφέρεται διαφορά της τιμής της παραμέτρου για ενήλικες ή νεογνά, διατηρήθηκε η τιμή των ενηλικών
- στις περιπτώσεις που αναφέρεται διαφοροποίηση της τιμής σε υγιείς και αρρώστους, διατηρήθηκε η τιμή που αντιστοιχεί σε υγιή άτομα
- σε περιπτώσεις που δίνεται η οδός χορήγησης του φαρμάκου, υπολογίστηκε η τιμή αυτής, για την από του στόματος (peros) χορήγηση
- από το εύρος των τιμών της, ελήφθη η μέση τιμή για κάθε ουσία
- έγινε μετατροπή όλων των τιμών σε ώρες.

Ένας ακόμη πολύ σημαντικός παράγοντας, που συνδεέται εν μέρει και με το χρόνο ημίσειας ζωής, είναι το ποσοστό της πρωτεϊνικής σύνδεσης του φαρμάκου. Ο γενικός κανόνας είναι πως ουσίες που είναι ελάχιστα συνδεδεμένες με τις πρωτεΐνες διαπερνούν σε

μεγαλύτερο ποσοστό τον ιστό, από αυτές που είναι ισχυρά συνδεδεμένες, αλλά αποβάλλονται πολύ πιο γρήγορα. Σε γενικές γραμμές όταν το ποσοστό σύνδεσης του φαρμάκου, είναι κάτω από 80-85% δεν υπάρχει σημαντική κλινική επίδραση της σύνδεσης, στις φαρμακοκινητικές ιδιότητές του.

Ωστόσο για τις ουσίες που συνδέονται ισχυρά με τις πρωτεΐνες, μπορεί να υπάρξουν αισθητές διαφορές και διακυμάνσεις σε παραμέτρους που αφορούν τη διαπερατότητα των ουσιών στους ιστούς και το χρόνο ημίσειας ζωής αυτών. Στα δεδομένα που εξήχθησαν έγινε επεξεργασία των τιμών για τη λήψη του ποσοστιαίου μέσου όρου για κάθε φάρμακο και όπου γινόταν αναφορά προτιμήθηκε το ποσοστό σύνδεσής του με την αλβουμίνη. Από το dataset αφαιρέθηκαν οι ουσίες με μηδενικό χρόνο ημίσειας ζωής καθώς και οι ουσίες που δεν είχαν SMILES (το οποίο είναι απαραίτητο για την εξαγωγή των descriptors που περιγράφονται στην επόμενη ενότητα) και απέμειναν για περαιτέρω επεξεργασία 1106 χημικές ουσίες.

5.1.2 Περιγραφή Της Δομής - Περιγραφικές Μεταβλητές Και Υπολογισμός Αυτών (Descriptors)

Η δομή ενός μορίου μπορεί να εκφραστεί με πληθώρα περιγραφικών μεταβλητών (descriptors), οι οποίες υπολογίζονται από κατάλληλα λογισμικά. Το περιεχόμενο της πληροφορίας που κωδικοποιείται σε μια περιγραφική μεταβλητή, εξαρτάται από δύο βασικούς παράγοντες, τον τρόπο αναπαράστασης του μορίου και τον αλγόριθμο ο οποίος χρησιμοποιείται για τον υπολογισμό της περιγραφικής μεταβλητής. Μέσω λογισμικών μπορούν να υπολογιστούν περισσότερες από 4500 περιγραφικές μεταβλητές, που προκύπτουν από διάφορες θεωρίες, βασιζόμενες στη φυσικοχημεία, την οργανική χημεία, την κβαντοχημεία, τα μαθηματικά, τη διαφορική τοπολογία, την αλγεβρική τοπολογία, τη θεωρία των γράφων, τη θεωρία της πληροφορίας και ειδικότερα τη χημειοπληροφορική. Για τον υπολογισμό των περιγραφικών μεταβλητών με τα διάφορα λογισμικά προγράμματα η δομή των μορίων εισάγεται στον υπολογιστή σχεδιαστικά ή με την απλοποιημένη γραφή SMILES [34].

Ανάλογα με το πληροφοριακό τους περιεχόμενο, οι περιγραφικές μεταβλητές κατατάσσονται σε φυσικοχημικές, μοριακές, κβαντομηχανικές, τοπολογικές, γεωμετρικές και δομικές [42].

- Οι φυσικοχημικές ιδιότητες περιλαμβάνουν τη λιποφιλία, τις σταθερές ιοντισμού, τη διαλυτότητα, τη μοριακή διαθλασιμότητα, την πολωσιμότητα κλπ.

- Μοριακές ιδιότητες αποτελούν η ικανότητα σχηματισμού δεσμών υδρογόνου, το μοριακό βάρος, παράμετροι συνολικού όγκου και επιφάνειας, η πολική επιφάνεια / τοπολογική πολική επιφάνεια, ο αριθμός των περιστρεφόμενων δεσμών, η αρωματικότητα κλπ.

- Οι κβαντομηχανικές ιδιότητες περιλαμβάνουν τη διπολική ροπή, παραμέτρους ενέργειας (EHOMO, ELUMO, Ενεργειακό χάσμα), ηλεκτροστατικά δυναμικά, μερικά φορτία κλπ. Για τον υπολογισμό τους απαιτείται η γεωμετρική βελτιστοποίηση του μορίου.

- Γεωμετρικές ιδιότητες θεωρούνται οι αποστάσεις μεταξύ ατόμων, γωνίες και δίεδρες γωνίες, συμμετρίες, ο καθορισμός κεντροειδούς κλπ.

- Οι τοπολογικές παράμετροι αποτελούν μια ευρύτατη κατηγορία περιγραφικών μεταβλητών που περιλαμβάνει δείκτες μοριακής συνδετικότητας χ (connectivity indices) και δείκτες μοριακού σχήματος (shape indices), που αφορούν σε ολόκληρο το μόριο, αλλά και τους ηλεκτροτοπολογικούς δείκτες (electrotopological state indices, Estate) που επικεντρώνονται σε συγκεκριμένα άτομα του μορίου. Οι τοπολογικοί δείκτες προκύπτουν με βάση τη θεωρία των γράφων.

- Οι δομικές μεταβλητές αφορούν απλά στην παρουσία (ή τη συχνότητα παρουσίας) συγκεκριμένων δομικών χαρακτηριστικών στο μόριο, όπως ετεροατόμων, λειτουργικών ομάδων, διπλών, τριπλών δεσμών, κλπ. Ουσιαστικά οι δομικές μεταβλητές εντάσσονται στη λογική της περιγραφής των μορίων, σύμφωνα με την ανάλυση Free-Wilson.

Περιγραφικές παράμετροι είναι επίσης δυνατόν να προκύψουν από Γραμμικές Σχέσεις Ελεύθερης Ενέργειας (Liner Free Energy Relationships, LFER) και Γραμμικές Σχέσεις Ενέργειας Επιδιαλύτωσης (Linear Solvation Energy Relationships, LSER). Με Γραμμικές Σχέσεις Ελεύθερης Ενέργειας έχουν προκύψει οι ηλεκτρονιακές σταθερές σ του Hammett (σm, σp, σo), η επαγωγική σταθερά σI του Charton, οι ηλεκτρονιακές σταθερές F και R των Swain και Lupton, για το επαγωγικό φαινόμενο και το φαινόμενο συντονισμού αντίστοιχα, η στερική σταθερά Es του Taft, η υδρόφοβη σταθερά π του Hansch κ.α.

Οι Γραμμικές Σχέσεις Ενέργειας - Επιδιαλύτωσης βασίζονται στην άποψη ότι, φαινόμενα που αφορούν σε ισορροπία ανάμεσα σε δύο φάσεις, μπορούν να περιγραφούν με 5 θεμελιώδεις παραμέτρους που αφορούν στην οξύτητα σε δεσμούς υδρογόνου (A), στη

βασικότητα σε δεσμούς υδρογόνου (B), στον όγκο (V), στην πολωσιμότητα/διπολικότητα S και στην πλεονάζουσα μοριακή διαθλασιμότητα E.

Ανάλογα με το επίπεδο των διαστάσεων που επηρεάζουν τον υπολογισμό τους, οι μεταβλητές κατατάσσονται σε μεταβλητές 1D, 2D ή 3D. Π.χ. το μοριακό βάρος αποτελεί μεταβλητή 1D, ενώ η λιποφιλία η οποία υπολογίζεται με βάση το συντακτικό τύπο είναι 2D. Οι μεταβλητές 3D υπολογίζονται με βάση τη θερμοδυναμικά ευνοϊκή διαμόρφωση, μετά από ελαχιστοποίηση της ενέργειας και αντλούν πληροφορίες από την τριδιάστατη δομή. Στην κατηγορία αυτή ανήκουν τα κβαντομηχανικά μεγέθη, τα μήκη των δεσμών, οι γωνίες, η διπολική ροπή, τα μερικά φορτία καθώς και ορισμένες παράμετροι όγκου και επιφάνειας. Μεταβλητές 3D αποτελούν επίσης οι μεταβλητές πλέγματος (GRID descriptors), οι οποίες εκτός από την εύρεση της θερμοδυναμικά ευνοϊκής διαμόρφωσης, προϋποθέτουν την υπέρθεση και ευθυγράμμιση (alignment) των ενώσεων εντός ενός θεωρητικού τριδιάστατου πλέγματος.

Ωστόσο, πολλές από τις μεταβλητές αυτές δεν είναι εύκολα ερμηνεύσιμες από τον φαρμακοχημικό. Τα αντίστοιχα μοντέλα είναι χρήσιμα για την πρόβλεψη της δράσης νέων υπό σύνθεση ενώσεων, ωστόσο όμως δεν μπορούν να κατευθύνουν άμεσα τη σύνθεση ή να οδηγήσουν σε υποθέσεις για το μηχανισμό δράσης. Από την άλλη, απλές φυσικοχημικές και μοριακές μεταβλητές που εκφράζουν χαρακτηριστικά φαρμακο - ομοιότητας και που πρακτικά εντάσσονται στο τρίπτυχο λιποφιλία, ηλεκτρονιακές ιδιότητες, στερικές ιδιότητες, σε συνδυασμό και με δομικές ή γεωμετρικές παραμέτρους, οδηγούν συχνά σε εξίσου ικανοποιητικά και παράλληλα εύκολα ερμηνεύσιμα μοντέλα.

Ο υπολογισμός των descriptors στην εργασία αυτή, έγινε με τη χρήση των smiles των 1106 ουσιών, μέσω του λογισμικού ανοιχτού κώδικα Padel [37]. Υπολογίστηκαν για τις παραπάνω ουσίες, 585 descriptors για κάθε χημική ένωση και μαζί με τις 13 παραμέτρους που είχαν ληφθεί από την Drugbank, δημιουργήθηκε ένα dataset για τις 1106 ουσίες με 598 μεταβλητές για την καθεμία από αυτές.

5.2 Επεξεργασία δεδομένων

5.2.1 Χρησιμοποιούμενες μέθοδοι τεχνητής νοημοσύνης

Οι αλγόριθμοι που εκπαιδεύθηκαν από τα dataset για τη συγκεκριμένη εργασία και δημιούργησαν τα μοντέλα τεχνητής νοημοσύνης για την πρόβλεψη του χρόνου ημίσειας ζωής, είναι οι εξής:

Support Vector Regressor SVR

Αρχικά δοκιμάστηκε ο Support Vector Regression. Στόχος του SVM είναι ο εντοπισμός του βέλτιστου υπερεπιπέδου το οποίο μπορεί να ταξινομήσει τα σημεία σε κατηγορίες (classification) και το οποίο παρουσιάζει τη μεγαλύτερη απόσταση μεταξύ των σημείων των διαφορετικών κατηγοριών. Η υποπερίπτωση της Παλινδρόμησης Διανυσμάτων Υποστήριξης (Support Vector Regression, SVR) πηγαίνει τη διαδικασία της κατηγοριοποίησης ένα βήμα παραπέρα, προσπαθώντας να προβλέψει με όσο το δυνατόν μεγαλύτερη ακρίβεια τις εξόδους κάποιας συνάρτησης (εξαρτημένες μεταβλητές), σε συνάρτηση με τις μεταβλητές εισόδου της (ανεξάρτητες μεταβλητές), [47].

Multi-layer Perceptron Regressor MLPRegressor

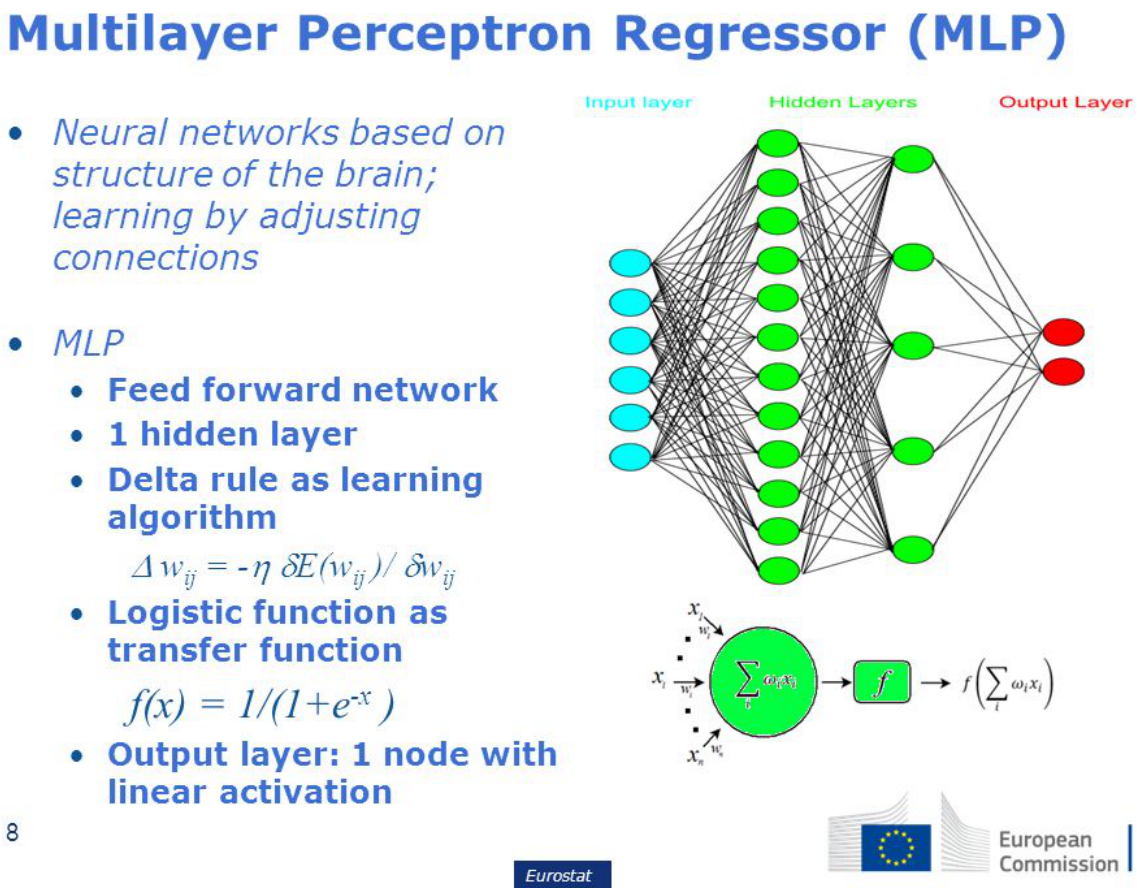
Στη συνέχεια δοκιμάστηκε ο MLPR. Το MLPR είναι ένα Νευρωνικό Δίκτυο που έχει πολύ καλές επιδόσεις και επιλύει προβλήματα που ένα απλό Perceptron δεν μπορεί, Σχήμα 5.2, [27] . Η εκπαίδευση ενός δικτύου MLPR έχει ιδιαίτερο ενδιαφέρον λόγω της ικανότητας του MLPR να συμπεριφέρεται ως ”Καθολικός Προσεγγιστής” (Universal Approximator). Αποδεικνύεται πως εάν έχουμε το κατάλληλο μέγεθος δικτύου, τότε μπορούμε να το εκπαιδεύσουμε να μάθει όποια συνάρτηση θέλουμε και με οποιαδήποτε ακρίβεια θέλουμε. Αυτό αιτιολογεί και την μεγάλη δημοτικότητα των αλγορίθμων εκπαίδευσης MLPR. Ο πιο γνωστός αλγόριθμος εκπαίδευσης είναι ο Back - Propagation.

Gradient Boosting Regressor

Για τη δημιουργία μοντέλου πρόβλεψης δοκιμάστηκε και ο GBR. Ο GBR εκπαιδεύει ένα decision tree και στη συνέχεια κάνει πρόβλεψη μέσω αυτού. Υπολογίζει τα residual αυτού του decision tree, και αποθηκεύει τα residual errors σαν τα καινούρια y. Επαναλαμβάνει την εκπαίδευση νεων decision trees εώς ότου εκπαιδεύσει τόσα, όσα του έχουμε ορίσει. Εκτελεί την τελική πρόβλεψη. Η νέα πρόβλεψη προκύπτει απλά προσθέτοντας τις προβλέψεις κάθε decision tree, Σχήμα 5.3 [4].

Random Forest Regressor

Τέλος, έγινε κι εφαρμογή του Random Forest Regressor. Ο Random Forest Regressor είναι ένα ensemble μοντέλο το οποίο χρησιμοποιεί το bagging ως μέθοδο ensemble και το decision tree ως ανεξάρτητο μοντέλο, Σχήμα 5.4, [4].



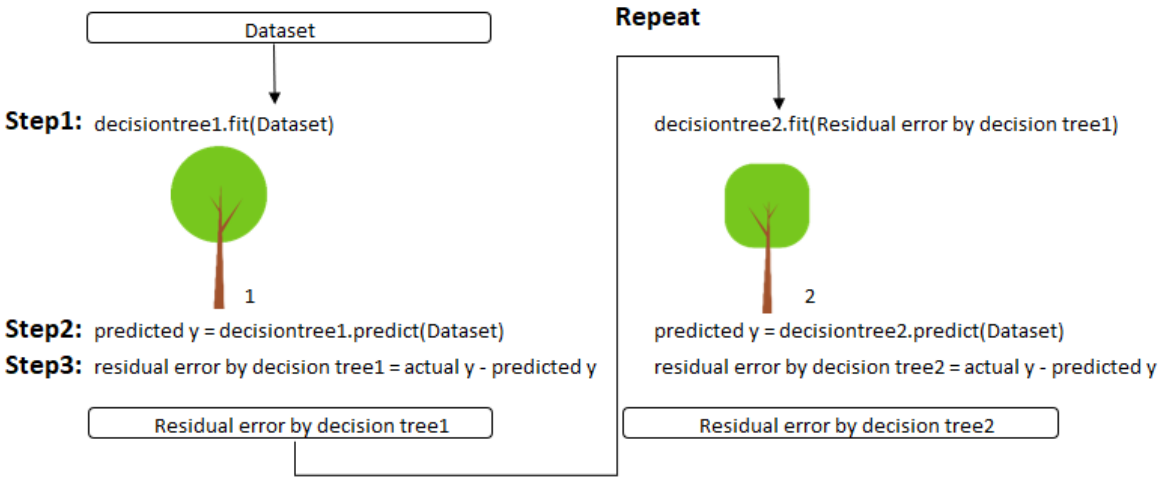
Σχήμα 5.2: Multi-layer Perceptron Regressor

5.2.2 Εκπαίδευση αλγορίθμων και πρόβλεψη

Το dataset που παρείχθηκε χρησιμοποιήθηκε για την εκπαίδευση των αλγορίθμων, με σκοπό τη δημιουργία μοντέλων πρόβλεψης του χρόνου ημίσειας ζωής. Δοκιμάστηκαν τρεις προσεγγίσεις, η βασική προσέγγιση που περιλαμβάνει την εκπαίδευση των αλγορίθμων με το τελικό επεξεργασμένο dataset, η προσέγγιση με τη χρήση clustering με τη δημιουργία επιμέρους clusters όσον αφορά το dataset και η προσέγγιση με voting, που ουσιαστικά συνδυάζει τις τεχνικές μηχανικής μάθησης της βασικής προσέγγισης.

5.2.2.1 Βασική προσέγγιση

- **Ανάκληση δεδομένων**



Σχήμα 5.3: How Gradient Boost learns

Αρχικά έγινε η επεξεργασία του dataset 1106 ουσιών με 598 μεταβλητές. Από αυτές απομακρύνθηκαν οι στήλες που είχαν προβληματικές τιμές οι 11 και 72, καθώς και η στήλη των smiles και έμεινε το πρώτο υποσύνολο του dataset με 1106 ουσίες και 595 μεταβλητές που τις χαρακτηρίζουν. Επίσης έγινε μετατροπή των τιμών της πρωτεϊνικής σύνδεσης (pbn) από ποσοστιαίες σε δεκαδικές τιμές, όπως φαίνεται στο Listing 1.

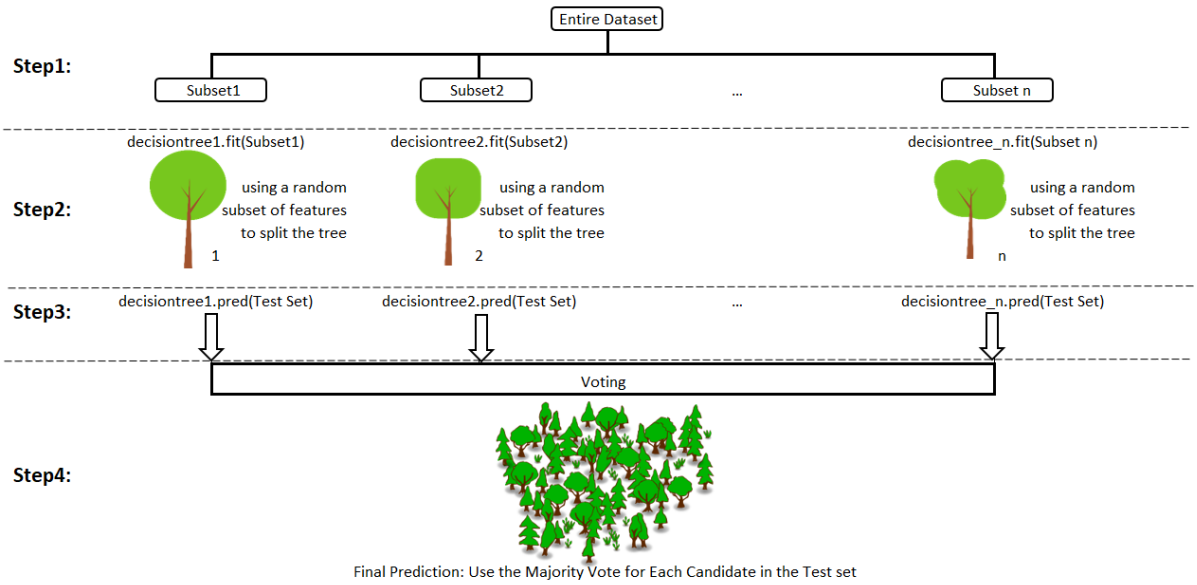
- **Αφαίρεση εγγραφών**

  Στη συνέχεια έγινε απομάκρυνση των ουσιών που είχαν σχετικά πολύ υψηλές τιμές χρόνου ημίσειας ζωής διατηρώντας τις τιμές που ήταν μικρότερες των 50 ωρών και δημιουργήθηκε ένα υποσύνολο με 1007 ουσίες και 595 μεταβλητές αυτών, όπως φαίνεται στο Listing 1.

- **Διαχωρισμός δεδομένων**

  Ακολούθως έγινε διαχωρισμός των δεδομένων σε (input data), που αποτελούνται από 1007 σειρές και 594 στήλες και σε (target value) που αποτελούνται από 1007 σειρές και μία στήλη, το χρόνο ημίσειας ζωής. Τέλος έγινε και λογαριθμοποίηση του χρόνο ημίσειας ζωής Log(hour) για τα στάνταρ της QSAR practice, όπως φαίνεται στο Listing 2.

- **Ελλιπείς τιμές - Missing values**



Σχήμα 5.4: Random Forest Regressor

Στα υποσύνολα αυτά έγινε υπολογισμός των missing values και αντικατάσταση αυτών, με βάση το μέσο όρο της κάθε στήλης, ώστε να υπάρχουν τιμές για όλες τις ουσίες, όπως φαίνεται στο Listing 2.

- **Μηδενική διακύμανση**

  Στα input data έγινε απομάκρυνση των πεδίων με μηδενική διακύμανση. Απέμειναν ως input data 1007 ουσίες με 583 μεταβλητές, όπως φαίνεται στο Listing 2.

- **Απομάκρυνση outliers**

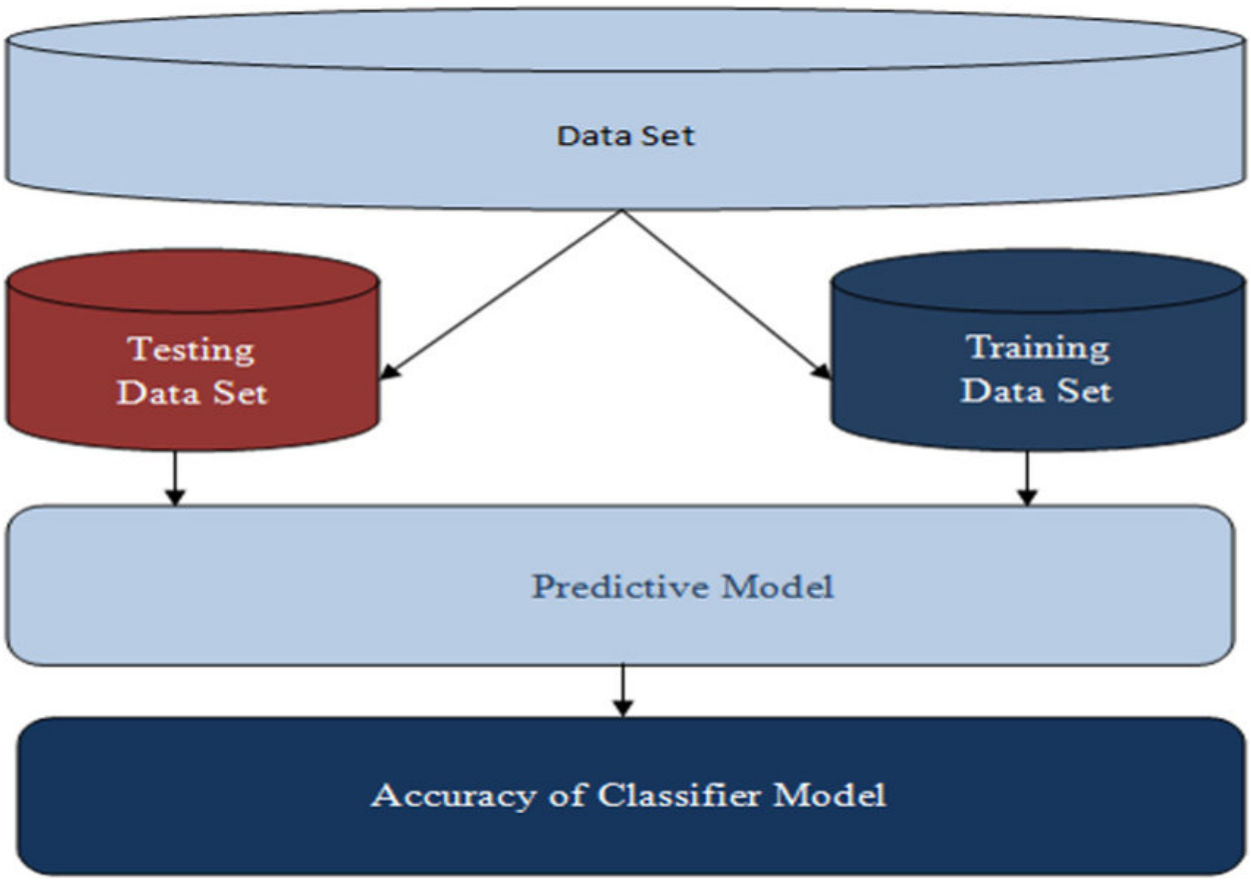
  Στη συνέχεια έγινε έρευνα για ύπαρξη outliers και ακολούθησε η απομάκρυνση αυτών. Χρησιμοποιήθηκε ο αλγόριθμος Local Outlier Factor (LOF) algorithm που βασίζεται στον υπολογισμό αποκλίσεων της πυκνότητας γειτονικών σημείων, όπως φαίνεται στο Listing 2.

- **Κανονικοποίηση δεδομένων**

  Στα δεδομένα υπήρχαν ορισμένες μεταβλητές με πολύ μεγάλες τιμές και άλλες με πολύ μικρές. Στις περιπτώσεις αυτές γίνεται αναγωγή των αριθμητικών τιμών σε άλλες αριθμητικές τιμές, που να κυμαίνονται εντός των ορίων της επιθυμητής περιοχής. Επίσης τα νευρωνικά δίκτυα λειτουργούν καλύτερα όταν οι τιμές που επεξεργάζονται κυμαίνονται στην περιοχή [0...1]. Η διαδικασία αυτή ονομάζεται κανονικοποίηση

(normalization). Έγινε λοιπόν κανονικοποίηση όλων των δεδομένων ώστε να αποκλειστούν τα outliers φαινόμενα και να είναι τα δεδομένα πιο εύκολα προς επεξεργασία από τους αλγόριθμους, όπως φαίνεται στο Listing 2.

- **Δημιουργία training και test set**



Σχήμα 5.5: Data mining

Το υποσύνολο που είχαμε διαχωρίστηκε τυχαία σε training και test set, Σχήμα 5.5 σε ποσοστό 67 - 33 %. Με τον τρόπο αυτό έχουμε ένα training set : data_train : 674 x 583 - hl_train : 674 x 1 κ ένα test set : data_test : 333 x 583 - hl_test : 333 x 1. Με το training set γίνεται η εκπαίδευση των αλγορίθμων και με το test set έλεγχουμε την προβλεπτική ικανότητα του εκπαιδευμένου συστήματος, όπως φαίνεται στο Listing 2.

- **Μείωση των διαστάσεων**

Με εφαρμογή της PCA μεθόδου, προσπαθήσαμε να μειώσουμε τις διαστάσεις των δεδομένων για την καλύτερη επεξεργασία του dataset. Έτσι οι 583 μεταβλητές πλέον

αντικαταστάθηκαν από 30 πιο σημαντικές, αντιπροσωπευτικές του αρχικού συνόλου. Η διαμόρφωση των νέων dataset έχει ως εξής data_train : 674 x 30 - hl_train : 674 x 1, data_test : 333 x 30 - hl_test : 333 x 1, όπως φαίνεται στο Listing 3.

- **Εκπαίδευση μοντέλων**

  Κατά τη βασική προσέγγιση έγινε η εκπαίδευση των αλγορίθμων με τα δεδομένα του training dataset, data_train : 674 x 30 - hl_train : 674 x 1, όπως φαίνεται στο Listing 4.

- **Πρόβλεψη τιμών**

  Στη συνέχεια οι εκπαιδευμένοι αλγόριθμοι με τις τιμές του test set, data_test : 333 x 30 - hl_test : 333 x 1 έκαναν πρόβλεψη του χρόνου ημίσειας ζωής, όπως φαίνεται στο Listing 5.

- **Υπολογισμός MAE, RMSE**

  Στο τέλος, έγινε υπολογισμός του μέσου απόλυτου σφάλματος (MAE) και της τετραγωνικής ρίζας του μέσου όρου των τετραγώνων των σφαλμάτων (RMSE). Αυτό γίνεται ώστε να αξιολογήσουμε την προβλεπτική ικανότητα των μοντέλων που δημιουργήσαμε, αλλά και να συγκρίνουμε τα μοντέλα ώστε να δούμε ποιο μπορεί να μας δώσει την πιο ασφαλή πρόβλεψη, όπως φαίνεται στο listing 6.

5.2.2.2 Προσέγγιση με clustering

Το Clustering είναι μια τεχνική μηχανικής μάθησης που ουσιαστικά οργανώνει τα δεδομένα σε γκρουπ. Λαμβάνοντας το σετ δεδομένων χρησιμοποιείται ένας αλγόριθμος clustering, ώστε κάθε σημείο των δεδομένων να κατηγοριοποιηθεί σε ένα συγκεκριμένο γκρουπ. Θεωρητικά τα σημεία που ανήκουν στο ίδιο γκρουπ θα έπρεπε να έχουν παρόμοιες ιδιότητες ή και χαρακτηριστικά, ενώ τα σημεία που ανήκουν σε διαφορετικά γκρουπ θα έπρεπε να έχουν σημαντικά διαφορετικές ιδιότητες ή και χαρακτηριστικά. Το Clustering είναι μια unsupervised μέθοδος εκμάθησης και είναι μια τεχνική που χρησιμοποιείται συχνά σε στατιστική ανάλυση δεδομένων σε διάφορα πεδία [29].

Στην εργασία το clustering πραγματοποιήθηκε με τη χρήση του αλγόριθμου K - Means. Αρχικά επιλέξαμε να κατηγοριοποιηθούν τα δεδομένα μας σε τρία γκρουπ. Ο αλγόριθμος ουσιαστικά εντοπίζει το κεντρικό σημείο (από το μέσο όρο των διανυσματικών αποστάσεων των σημείων) για κάθε γκρουπ και στη συνέχεια κατηγοριοποιεί τα σημεία ανάλογα

```
### READING DATASET ############################################################

#read data and remove columns 72,111 (problematic) and 0,1 (indeces)
data = pandas.read_csv("../finallist.csv",low_memory=False)
data = data.drop([data.columns[72],data.columns[111]],axis=1)
data = data.drop([data.columns[0],data.columns[1]],axis=1)

#drop smiles and name cols (strings)
data = data.drop(['smiles','Name'],axis=1)

# convert percentage values to float values for column pbn
data['pbn'] = data['pbn'].str.replace(".", "")
data['pbn'] = data['pbn'].str.replace(" ", "")
data['pbn'] = data['pbn'].str.replace(",", ".")
data['pbn'] = data['pbn'].str.rstrip('%').astype('float') / 100.0

# filter out records with high half-life values as this seems wrong
data = data[data.half_life < 50]
```

Listing 1: Reading data

με τις αποστάσεις από αυτά τα κεντρικά σημεία, στα αντίστοιχα γκρουπ. Κάθε σημείο κατηγοριοποίειται στο γκρούπ από το οποίο απέχει τη μικρότερη απόσταση από το κεντρικό του σημείο.

Ένας από τους πιο γνωστούς αλγόριθμους ομαδοποίησης αυτής της κατηγορίας, είναι ο αλγόριθμος των K - μέσων (K-means). Ο αριθμός K των ομάδων καθορίζεται πριν την εκτέλεση του αλγορίθμου. Ο αλγόριθμος ξεκινά διαλέγοντας K τυχαία σημεία από τα δεδομένα ως τα κέντρα των ομάδων. Έπειτα αναθέτει κάθε σημείο στην ομάδα της οποίας το κέντρο είναι πιο κοντά (μικρότερη απόσταση) σε αυτό το σημείο. Στη συνέχεια, υπολογίζει για κάθε ομάδα το μέσο όρο όλων των σημείων της (μέσο διάνυσμα) και ορίζει αυτό ως νέο κέντρο της. Τα δύο τελευταία βήματα επαναλαμβάνονται για ένα προκαθορισμένο αριθμό βημάτων, ή μέχρι να μην υπάρχει αλλαγή στο διαχωρισμό των σημείων σε ομάδες.

- **Ανάκληση δεδομένων** - Όμοια με τη βασική προσέγγιση
- **Αφαίρεση εγγραφών** - Όμοια με τη βασική προσέγγιση
- **Διαχωρισμός δεδομένων** - Όμοια με τη βασική προσέγγιση

- **Ελλιπείς τιμές - Missing values** - Όμοια με τη βασική προσέγγιση
- **Μηδενική διακύμανση** - Όμοια με τη βασική προσέγγιση
- **Απομάκρυνση outliers** - Όμοια με τη βασική προσέγγιση
- **Κανονικοποίηση δεδομένων** - Όμοια με τη βασική προσέγγιση
- **Δημιουργία training και test set** - Όμοια με τη βασική προσέγγιση
- **Αναγνώριση clusters στο training dataset**

  Στη συνέχεια με χρήση του αλγορίθμου K-Means πραγματοποιήθηκε η αναγνώριση των clusters του data_train, όπως φαίνεται στο Listing 7. Συγκεκριμένα το πλήθος των προς εύρεση clusters ορίστηκε σε διαφορετικές τιμές και τα αποτελέσματα ήταν καλύτερα για 3 - 5 clusters. Από αυτές επιλέχθηκε η τιμή 3, ώστε τα παραγόμενα clusters να έχουν το μέγιστο δυνατό μέγεθος.

- **Μείωση των διαστάσεων** - Όμοια με τη βασική προσέγγιση
- **Διαχωρισμός training set σε clusters**

  Ακολούθως έγινε η κατηγοριοποίηση των στοιχείων του data_train στα τρία επιμέρους clusters, με αποτέλεσμα τη δημιουργία τριων σετ δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων, όπως φαίνεται στο Listing 8.

- **Εκπαίδευση μοντέλων**

  Κατά την προσέγγιση με clustering, πραγματοποιήθηκε η εκπαίδευση των τεσσάρων αλγορίθμων τρεις φορές, με τα τρια σετ δεδομένων των clusters, όπως φαίνεται στο Listing 9.

- **Πρόβλεψη τιμών**

  Για κάθε στοιχείο του data_test έγινε η αντιστοίχισή του σε κάποιο από τα τρία clusters. Για την πρόβλεψη του χρόνου ημίσειας ζωής για καθένα από αυτά τα στοιχεία, έγινε χρήση του μοντέλου που εκπαιδεύτηκε με τα στοιχεία του συγκεκριμένου cluster, όπως φαίνεται στο Listing 10.

- **Υπολογισμός MAE, RMSE** - Όμοια με τη βασική προσέγγιση

5.2.2.3 Προσέγγιση με voting

Η voting είναι μια μέθοδος ensemble. Η τεχνική των μεθόδων αυτών βασίζεται στη δημιουργία πολλαπλών μοντέλων, που συνδυάζονται ώστε να παραχθούν πιο ακριβή βελτιστοποιημένα αποτελέσματα. Συνήθως οι ensemble μέθοδοι παράγουν πιο ακριβή αποτελέσματα σε συγκρίση με τα αποτελέσματα ενός απλού, μεμονωμένου μοντέλου. Στην εργασία αφού αρχικά εκπαιδεύτηκαν τα μοντέλα των αλγορίθμων Support Vector Regressor SVR, Gradient Boosting Regressor και Random Forest Regressor με το ίδιο data_train set , συνδυάστηκαν στη συνέχεια για τη δημιουργία του μοντέλου voting. Το νέο voting μοντέλο δημιουργήθηκε με τη μέθοδο Weighted averaging, όπου η πρόβλεψη κάθε μοντέλου πολλαπλασιάζεται με ένα βάρος αναλογώς με το ποιο μοντέλο είχε κάνει ως μεμονωμένο (base model), την καλύτερη πρόβλεψη. Η τελική πρόβλεψη προκύπτει από το μέσο όρο των πολλαπλασιασμένων με βάρη αρχικών προβλέψεων των base models [6], [10] .

- **Ανάκληση δεδομένων** - Όμοια με τη βασική προσέγγιση
- **Αφαίρεση εγγραφών** - Όμοια με τη βασική προσέγγιση
- **Διαχωρισμός δεδομένων** - Όμοια με τη βασική προσέγγιση
- **Ελλιπείς τιμές - Missing values** - Όμοια με τη βασική προσέγγιση
- **Μηδενική διακύμανση** - Όμοια με τη βασική προσέγγιση
- **Απομάκρυνση outliers** - Όμοια με τη βασική προσέγγιση
- **Κανονικοποίηση δεδομένων** - Όμοια με τη βασική προσέγγιση
- **Δημιουργία training και test set** - Όμοια με τη βασικη προσσέγγιση
- **Μείωση των διαστάσεων** - Όμοια με τη βασική προσέγγιση
- **Εκπαίδευση μοντέλων** - Ακολούθως δημιουργήθηκε το voting μοντέλο και έγινε η εκπαίδευσή του με το data_train set. Τα βάρη που χρησιμοποιήθηκαν, ορίστηκαν βάσει των επιδόσεων του κάθε αλγορίθμου ξεχωριστά και συγκεκριμένα 3 για τον SVR, 1 για τον GBR και 2 για τον RFR, όπως φαίνεται στο Listing 11.
- **Πρόβλεψη τιμών** Τέλος το εκπαιδευμένο voting μοντέλο προέβλεψε τους χρόνους ημίσειας ζωής για το data_test set, όπως φαίνεται στο Listing 12.
- **Υπολογισμός MAE, RMSE** - Όμοια με τη βασική προσέγγιση

```
### PREPROCESSING ########################################################################

# All columns except first one (half-life) are used as input data
columns = data.values[:,1:]
# The first column (half-life) is used as target value
half_life = data.values[:,0]
# Target values (half-life) are transformed by using natural logarithm
half_life = np.log(half_life)

# Handle missing values by using the mean of the corresponding column
im = SimpleImputer()
columns = im.fit_transform(columns)

# Remove features with zero variance
selector = VarianceThreshold()
columns = selector.fit_transform(columns)

# Search for ouliers
clf = LocalOutlierFactor(n_neighbors=20, contamination=0.1)
outliers = clf.fit_predict(columns)=
for i in range(outliers.shape[0]-1,-1,-1):
    if outliers[i] == -1:
        columns = np.delete(columns,i,axis=0)
        half_life = np.delete(half_life,i,axis=0)

# Apply Quantile tranformer (Gaussian/Uniform output) to restrict outliers effect
qt = QuantileTransformer(output_distribution='uniform')
columns = qt.fit_transform(columns)

# Split data to training and test sets
data_train, data_test, hl_train, hl_test = train_test_split(columns,half_life, test_size = 0.33)
```

Listing 2: Prepossessing

```
### DIMENSIONALITY REDUCTION ########################################################################

# Fit PCA
pca = PCA(n_components=30)
pca.fit(data_train)
data_train = pca.transform(data_train)
data_test = pca.transform(data_test)
```

Listing 3: Dimensionality reduction

```
### TRAIN ###########################################################################################

svr = svm.SVR(kernel='rbf',gamma='scale')
mlpr = MLPRegressor(max_iter=3000,hidden_layer_sizes=(20,20),activation='logistic')
gbr = GradientBoostingRegressor(loss='huber')
rfr = RandomForestRegressor(n_estimators=100)

mlpr.fit(data_train,hl_train)
svr.fit(data_train,hl_train)
gbr.fit(data_train,hl_train)
rfr.fit(data_train,hl_train)
```

Listing 4: Training

```
### PREDICTING ######################################################################################

res_svr = svr.predict(data_test)
res_mlpr = mlpr.predict(data_test)
res_gbr = gbr.predict(data_test)
res_rfr = rfr.predict(data_test)
```

Listing 5: Prediction

```
### ERROR METRICS ############################################################

def metrics(name,hl_test,res):
    mae = mean_absolute_error(hl_test,res)
    rmse = math.sqrt(mean_squared_error(hl_test,res))
    hl_test = np.exp(hl_test)
    res = np.exp(res)
    errors = np.abs(res - hl_test)
    max_error = np.max(errors,axis=0)
    ave_error = np.average(errors,axis=0)
    return ("{0}\t {1:.2f}\t {2:.2f}\t {3:.2f} \t {4:.2f}"
        .format(name,mae,rmse,max_error,ave_error))

print("\n\t MAE\t RMSE\t ME \t AE")
print(metrics('SVR',hl_test,res_svr))
print(metrics('MLPR',hl_test,res_mlpr))
print(metrics('GBR',hl_test,res_gbr))
print(metrics('RFR',hl_test,res_rfr))
```

Listing 6: Error calculation

```
### FIND CLUSTERS ##############################################################################

pca3 = PCA(n_components=30)
pca3.fit(data_train)
X = pca3.transform(data_train)
num_clusters = 3
km = KMeans(n_clusters=num_clusters, random_state=1)
km.fit(X)
clusters_labels = km.labels_
```

Listing 7: Finding clusters

```
### SPLIT TRAINING SET TO SUBSETS ########################################

cluster_indices = {}
for i in range(num_clusters):
    cluster_indices[i] = []

for i,l in enumerate(clusters_labels):
    cluster_indices[l].append(i)

cluster_data_training = {}
cluster_hl_training = {}
for i in range(num_clusters):
    cluster_data_training[i] = data_train[cluster_indices[i],:]
    cluster_hl_training[i] = hl_train[cluster_indices[i]]
```

Listing 8: Split training dataset to clusters

```
svr = {}
mlpr = {}
gbr = {}
rfr = {}

for i in range(num_clusters):

    svr[i] = svm.SVR(gamma='auto')
    mlpr[i] = MLPRegressor(max_iter=1000,hidden_layer_sizes=(20,20))
    gbr[i] = GradientBoostingRegressor()
    rfr[i] = RandomForestRegressor(n_estimators=100)

    mlpr[i].fit(cluster_data_training[i],cluster_hl_training[i])
    svr[i].fit(cluster_data_training[i],cluster_hl_training[i])
    gbr[i].fit(cluster_data_training[i],cluster_hl_training[i])
    rfr[i].fit(cluster_data_training[i],cluster_hl_training[i])
```

Listing 9: Training regressors

```
### PREDICT ###############################################################################

data_test_cluster_indices = {}
X = pca.transform(data_test)
for i in range(X.shape[0]):
    data_test_cluster_indices[i] = km.predict(X[i,:].reshape(1, -1))[0]

data_test = pca.transform(data_test)

res_svr_partial = {}
res_mlpr_partial = {}
res_gbr_partial = {}
res_rfr_partial = {}

for i in range(num_clusters):
    res_svr_partial[i] = svr[i].predict(data_test)
    res_mlpr_partial[i] = mlpr[i].predict(data_test)
    res_gbr_partial[i] = gbr[i].predict(data_test)
    res_rfr_partial[i] = rfr[i].predict(data_test)

res_svr = np.zeros((data_test.shape[0],1))
res_mlpr = np.zeros((data_test.shape[0],1))
res_gbr = np.zeros((data_test.shape[0],1))
res_rfr = np.zeros((data_test.shape[0],1))

for i in range(data_test.shape[0]):
    cl = data_test_cluster_indices[i]
    res_svr[i,0] = (res_svr_partial[cl])[i]
    res_mlpr[i,0] = (res_mlpr_partial[cl])[i]
    res_gbr[i,0] = (res_gbr_partial[cl])[i]
    res_rfr[i,0] = (res_rfr_partial[cl])[i]
```

Listing 10: Predicting

```
### TRAIN #####################################################

svr = svm.SVR(kernel='rbf',gamma='scale')
mlpr = MLPRegressor(max_iter=3000,hidden_layer_sizes=(20,20),activation='logistic')
gbr = GradientBoostingRegressor(loss='huber')
rfr = RandomForestRegressor(n_estimators=100)
vr = VotingRegressor([('svr',svr),('gbr',gbr),('rfr',rfr)],[3,1,2])

vr.fit(data_train,hl_train)
```

Listing 11: Training (Voting)

```
### PREDICT ###############################

res_vr = vr.predict(data_test)
```

Listing 12: Predicting (Voting)

6. ΑΠΟΤΕΛΕΣΜΑΤΑ - ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Δείκτες σφάλματος

Οι προβλέψεις πάντα εμπεριέχουν κάποιο σφάλμα. Τα σφάλματα διακρίνονται σε στατιστικά και σε τυχαία. Στατιστικά σφάλματα είναι αυτά που οφείλονται στο μοντέλο πρόβλεψης, είτε σε κακή εκτίμησή του, είτε στην παράλειψη κάποιων παραγόντων που επηρεάζουν τη μεταβλητή (π.χ. η οδός χορήγησης του φαρμάκου). Τα τυχαία λάθη είναι αποτέλεσμα μη προβλέψιμων παραγόντων που επηρεάζουν την εξεταζόμενη μεταβλητή. Η προβλεπόμενη τιμή της εξεταζόμενης μεταβλητής θα είναι πάντα μεγαλύτερη ή μικρότερη από την πραγματική απαίτηση και σχεδόν ποτέ ίση με αυτήν. Η διαφορά μεταξύ της προβλεπόμενης και της πραγματικής τιμής, ονομάζεται σφάλμα πρόβλεψης. Ο στόχος της πρόβλεψης είναι να ελαχιστοποιηθεί η τιμή του σφάλματος. Αν το μέγεθος του σφάλματος είναι μεγάλο, αυτό μπορεί να σημαίνει είτε ότι η τεχνική πρόβλεψης είναι λάθος, είτε ότι χρειάζεται τροποποίηση στις παραμέτρους [44].

Το σφάλμα των προβλέψεων μπορεί να υπολογισθεί συγκρίνοντας τις προβλέψεις με τις πραγματικές τιμές της μεταβλητής. Η συνολική επίδοση ενός μοντέλου πρόβλεψης, υπολογίζεται από το μέσο όρο των σφαλμάτων πρόβλεψης του μοντέλου για καθένα στοιχείο του test_set. Αρνητικές τιμές υποδηλώνουν υπερεκτίμηση της τιμής, ενώ θετικές τιμές δείχνουν υποεκτίμηση της τιμής. Για το λόγο αυτό χρησιμοποιούνται κυρίως απόλυτες τιμές σφάλματος, για τον υπολογισμό των δεικτών σφαλμάτων.

Για την αξιολόγηση της ακρίβειας των προβλέψεων που εξάγονται από κάποιο μοντέλο, απαιτείται ο υπολογισμός δεικτών σφάλματος. Παρατίθενται οι δείκτες σφάλματος που χρησιμοποιήθηκαν στην παρούσα εργασία:

Μέσο τετραγωνικό σφάλμα (Mean Squared Error-MSE) :

Το μέσο τετραγωνικό σφάλμα είναι ο μέσος όρος των τετραγώνων των σφαλμάτων. Ο δείκτης αυτός εξουδετερώνει τις μεγάλες αποκλίσεις ανάμεσα στην προβλεπόμενη και την πραγματική τιμή. Όπως είναι αναμενόμενο επηρεάζεται σημαντικά από τα μεγάλα σφάλματα, λόγω τετραγωνισμού και πολύ λιγότερο από τα μικρότερα. Ο τρόπος υπολογισμού απεικονίζεται στην Εξίσωση 6.1, όπου y_j είναι η πραγματική τιμή και x_j είναι η τιμή πρό-

βλεψης από το μοντέλο του χρόνου ημίσειας ζωής, ενώ n είναι το πλήθος των δειγμάτων.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - x_j)^2 \tag{6.1}$$

Ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error - RMSE) :

Ο δείκτης αυτός είναι η ρίζα του μέσου τετραγωνικού σφάλματος. Η ρίζα του μέσου τετραγωνικού σφάλματος έχει τις ίδιες ιδιότητες με το μέσο τετραγωνικό σφάλμα. Ο δείκτης είναι εκφρασμένος στις ίδιες μονάδες με την αρχική τιμή, σε αντίθεση με το μέσο τετραγωνικό σφάλμα που είναι υψωμένο στο τετράγωνο. Ο τρόπος υπολογισμού απεικονίζεται στην Εξίσωση 6.2. [12].

$$RMSE = \frac{1}{n}\sqrt{\sum_{j=1}^{n}(y_j - x_j)^2} \tag{6.2}$$

Μέσο απόλυτο σφάλμα (Mean Absolute Error – MAE) :

Το μέσο απόλυτο σφάλμα υπολογίζεται λαμβάνοντας υπόψη το μέσο όρο της απόλυτης τιμής των σφαλμάτων. Ο δείκτης αυτός είναι απαλλαγμένος από το μειονέκτημα του μέσου σφάλματος. Το μέσο απόλυτο σφάλμα δείχνει το μέγεθος των αποκλίσεων μεταξύ της προβλεπόμενης τιμής και της πραγματικής τιμής, όμως δε δίνει το πρόσημο τους. Όσο μεγαλύτερη είναι η τιμή αυτού του δείκτη, τόσο μικρότερη είναι η ακρίβεια της πρόβλεψης.

Σφάλμα πρόβλεψης = Πραγματική τιμή (y) - Τιμή πρόβλεψης (x)

Απόλυτο σφάλμα πρόβλεψης = |Σφάλμα πρόβλεψης|

Για κάθε στοιχείο του σετ δεδομένων υπολογίζεται η απόλυτη τιμή του σφάλματος πρόβλεψής του από το μοντέλο. Στη συνέχεια υπολογίζεται ο μέσος όρος όλων των απολύτων τιμών των σφαλμάτων των επιμέρους στοιχείων [11].

Ο μέσος όρος των απολύτων τιμών των σφαλμάτων πρόβλεψης συμβολίζεται με MAE και ο τρόπος υπολογισμού του απεικονίζεται στην Εξίσωση 6.3 [12].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - x_j\right| \tag{6.3}$$

	MAE	RMSE	AE
SVR	0.96	1.38	6.12
MLPR	1.13	1.55	7.86
GBR	0.99	1.35	6.33
RFR	0.98	1.38	6.30

Πίνακας 6.1: Αποτελέσματα Βασικής προσέγγισης

6.2 Αποτελέσματα

Στην παρούσα ενότητα αναλύονται τα αποτελέσματα των διαφορετικών μεθοδολογιών που χρησιμοποιήθηκαν για την πρόβλεψη του χρόνου ημίσειας ζωής.

Η αξιολόγηση της ακρίβειας των προβλέψεων των μοντέλων μηχανικής μάθησης, πραγματοποιήθηκε με τον υπολογισμό του μέσου απόλυτου σφάλματος (Mean Absolute Error - MAE) και της τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος (Root Mean Square Error - RMSE).

Επίσης έγινε και ο υπολογισμός του μέσου πραγματικού λάθους (Average Error - AE), που είναι ένας δείκτης του πόσο απέχει ουσιαστικά η προβλεπόμενη τιμή του χρόνου ημίσειας ζωής από την πραγματική, για την ερμηνεία πρακτικά της πρόβλεψης. Καθώς οι χρόνοι ημίσειας ζωής έχουν λογαριθμοποιηθεί κατά τη διάρκεια της διαδικασίας, το λάθος που υπολογίζεται από τους υπόλοιπους δείκτες δεν έχει φυσική ερμηνεία. Ο υπολογισμός του AE έγινε μετά την απολογαριθμοποίηση των τιμών του χρόνου ημίσειας ζωής, ώστε να προκύψει ποιο είναι το λάθος πρόβλεψης σχετικά με τις πραγματικές τιμές του χρόνου ημίσειας ζωής των φαρμάκων.

Στον Πίνακα 6.1, αναγράφονται τα αποτελέσματα που προέκυψαν κατά τη βασική προσέγγιση. Συγκρίνοντας την τιμή του MAE φαίνεται πως καλύτερη απόδοση έχει ο SVR με τιμή 0.96, ακολουθεί ο RFR με 0.98 και ο GBR με 0.99, ενώ έπεται ο MLPR με τη μεγαλύτερη τιμή 1.13. Η τιμή του RMSE για τον SVR και για τον RFR είναι 1.38, για τον MLPR είναι 1.55 και για τον GBR 1.35. Το μέσο πραγματικό λάθος ήταν 6.12 h για τον SVR, 6.30 για τον RFR, 6.33 για τον GBR και 7.86 h για τον MLPR. Συγκρίνοντας μεταξύ τους τις τιμές αυτές, φαίνεται πως η επίδοση του SVR είναι οριακά καλύτερη σε σχέση τους άλλους αλγορίθμους, ενώ ακολουθούν με την ίδια περίπου επίδοση οι GBR και RFR και στη συνέχεια, έπεται το νευρωνικό δίκτυο MLPR.

Στον Πίνακα 6.2, αναφέρονται τα αποτελέσματα που προέκυψαν κατά την προσέγγιση με clustering. Κατά την προσέγγιση αυτή, η τιμή του MAE για τον SVR είναι 0.98, ακολουθεί

	MAE	RMSE	AE
SVR	0.98	1.39	6.17
MLPR	1.27	1.76	10.48
GBR	1.07	1.44	6.74
RFR	1.01	1.39	6.34

Πίνακας 6.2: Αποτελέσματα Προσέγγισης clustering

	MAE	RMSE	AE
VR	0.95	1.35	6.06
VR (noPCA)	0.94	1.34	6.04

Πίνακας 6.3: Αποτελέσματα Προσέγγισης voting

ο RFR με 1.01 και ο GBR με 1.07, ενώ έπεται ο MLPR με τη μεγαλύτερη τιμή 1.27. Η τιμή του RMSE για τον SVR και για τον RFR, είναι 1.39, για τον MLPR είναι 1.76 και για τον GBR είναι ίση με 1.44. Το μέσο πραγματικό λάθος, ήταν 6.17 h για τον SVR, 6.34 για τον RFR, 6.74 για τον GBR και 10.48 h για τον MLPR. Συγκρίνοντας μεταξύ τους τις τιμές αυτές, φαίνεται πως η πρόβλεψη είναι καλύτερη για τον SVR, ακολουθεί ο RFR, τρίτος έρχεται ο GBR και στη συνέχεια ακολουθεί ο MLPR.

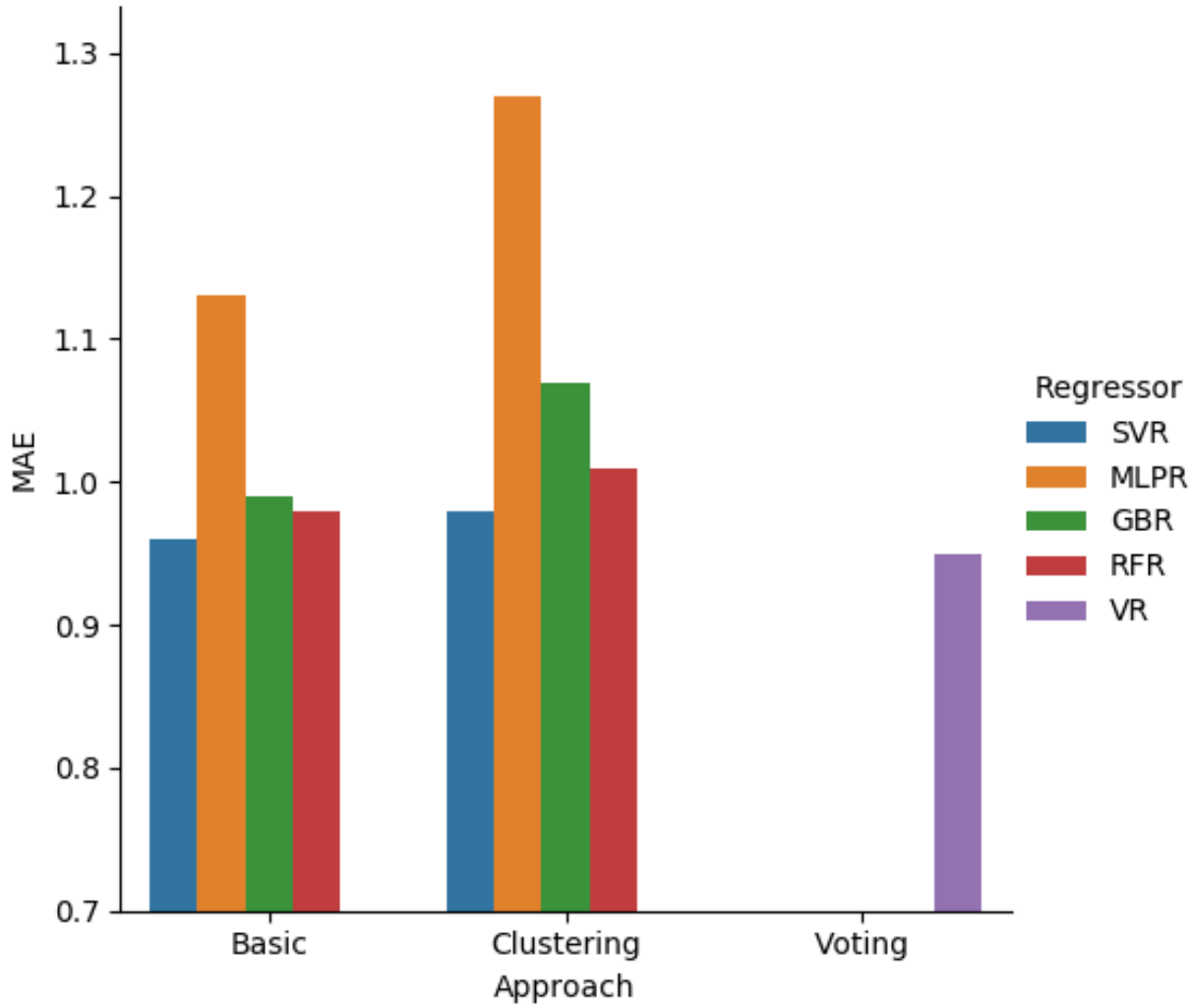
Στον Πίνακα 6.3, αναγράφονται τα αποτελέσματα που προέκυψαν κατά την προσέγγιση με voting. Η τιμή του MAE είναι ίση με 0.95, η τιμή του RMSE είναι ίση με 1.35 και η τιμή του AE 6.06. Τα αποτελέσματα αυτά δείχνουν ότι η προσέγγιση του voting υπερτερεί των υπολοίπων.

Με απώτερο στόχο τη διερεύνηση της επιρροής της μειώσης των διαστάσεων του σετ δεδομένων με τη χρήση της μεθόδου PCA έγινε μία επιπλέον δοκιμή. Επιχειρήθηκε η πρόβλεψη του χρόνου ημίσειας ζωής με την προσέγγιση voting, χωρίς τη χρήση PCA. Η τιμή του MAE είναι ίση με 0.94, η τιμή του RMSE είναι ίση με 1.34 και η τιμή του AE 6.04. Πρακτικά, η εφαρμογή της μεθόδου PCA εξάλειψε ένα μικρό ποσοστό χρήσιμης πληροφορίας αυξάνοντας τις τιμές των δεικτών σφάλματος.

Η διαφορά της τιμής του MAE κατά τη βασική προσέγγιση, την προσέγγιση με clustering και την voting, απεικονίζεται στο Σχήμα 6.1.

Στο Σχήμα 6.2, απεικονίζεται η διαφορά της τιμής του RMSE κατά τη βασική προσέγγιση, την προσέγγιση με clustering και την voting.

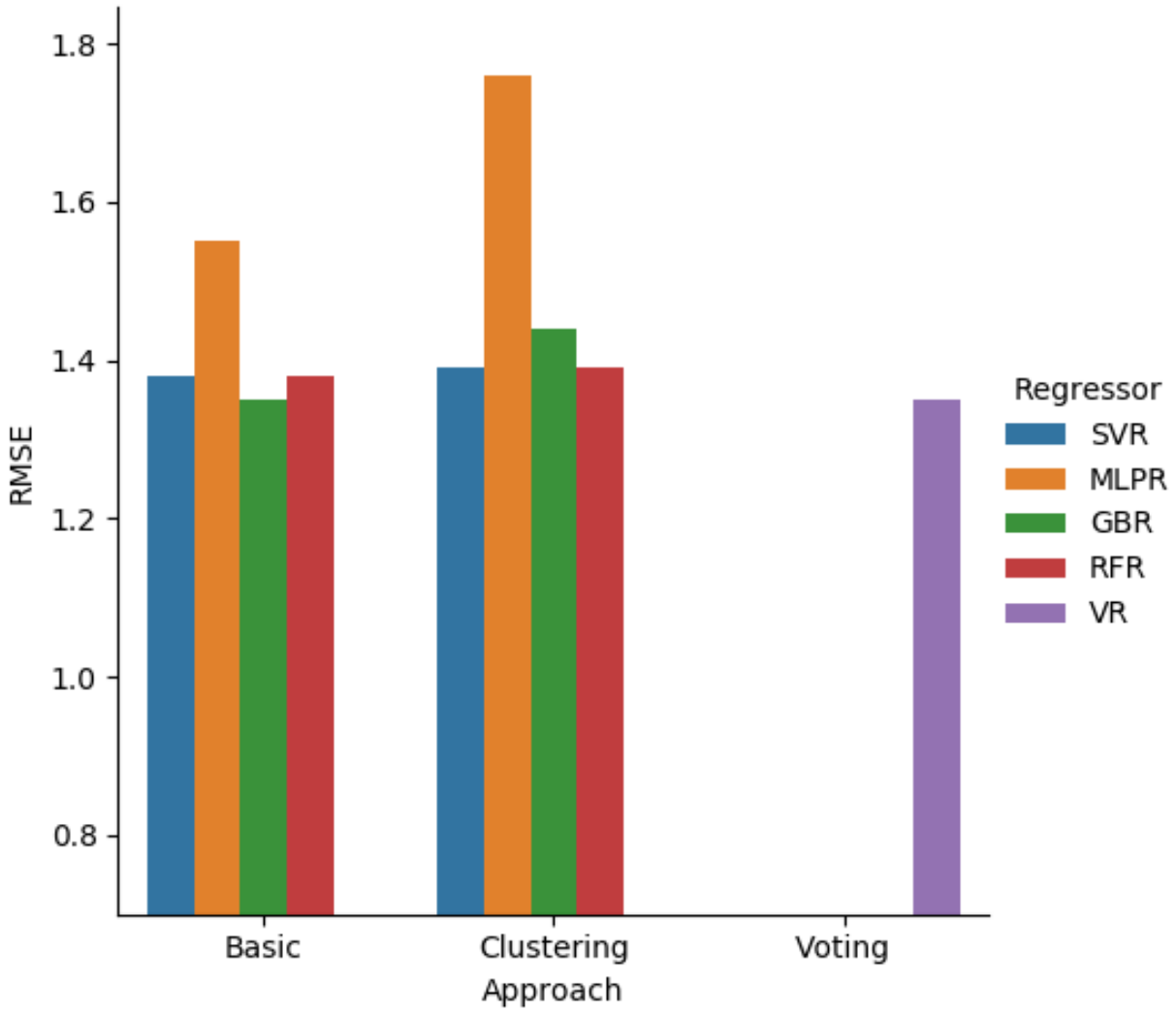
Ομοίως στο σχήμα 6.3, απεικονίζεται η διαφορά της τιμής του AE κατά τη βασική προσέγγιση, την προσέγγιση με clustering και την voting.



Σχήμα 6.1: Απεικόνιση του MAE στις τρεις προσεγγίσεις

Το συμπέρασμα που προκύπτει κατά την ανάλυση των παραπάνω αποτελεσμάτων, είναι πως οι αλγόριθμοι SVR, RFR και GBR αποδίδουν καλύτερα έναντι της εφαρμογής του νευρωνικού δικτύου MLPR, με μικρές διαφορές μεταξύ τους τόσο στην τιμή του MAE και του RMSE, όσο και του AE. Από τις προσεγγίσεις που εφαρμόστηκαν, την καλύτερη απόδοση προβλεπτικής ικανότητας είχε η προσέγγιση Voting (no PCA), με τις χαμηλότερες τιμές στους δείκτες σφαλμάτων. Και πάλι όμως η απόδοσή τους δεν είναι ιδιαίτερα ικανοποιητική για την άμεση εφαρμογή των μοντέλων, για την πρόβλεψη του χρόνου ημίσειας ζωής μιας νέας φαρμακευτικής ουσίας.

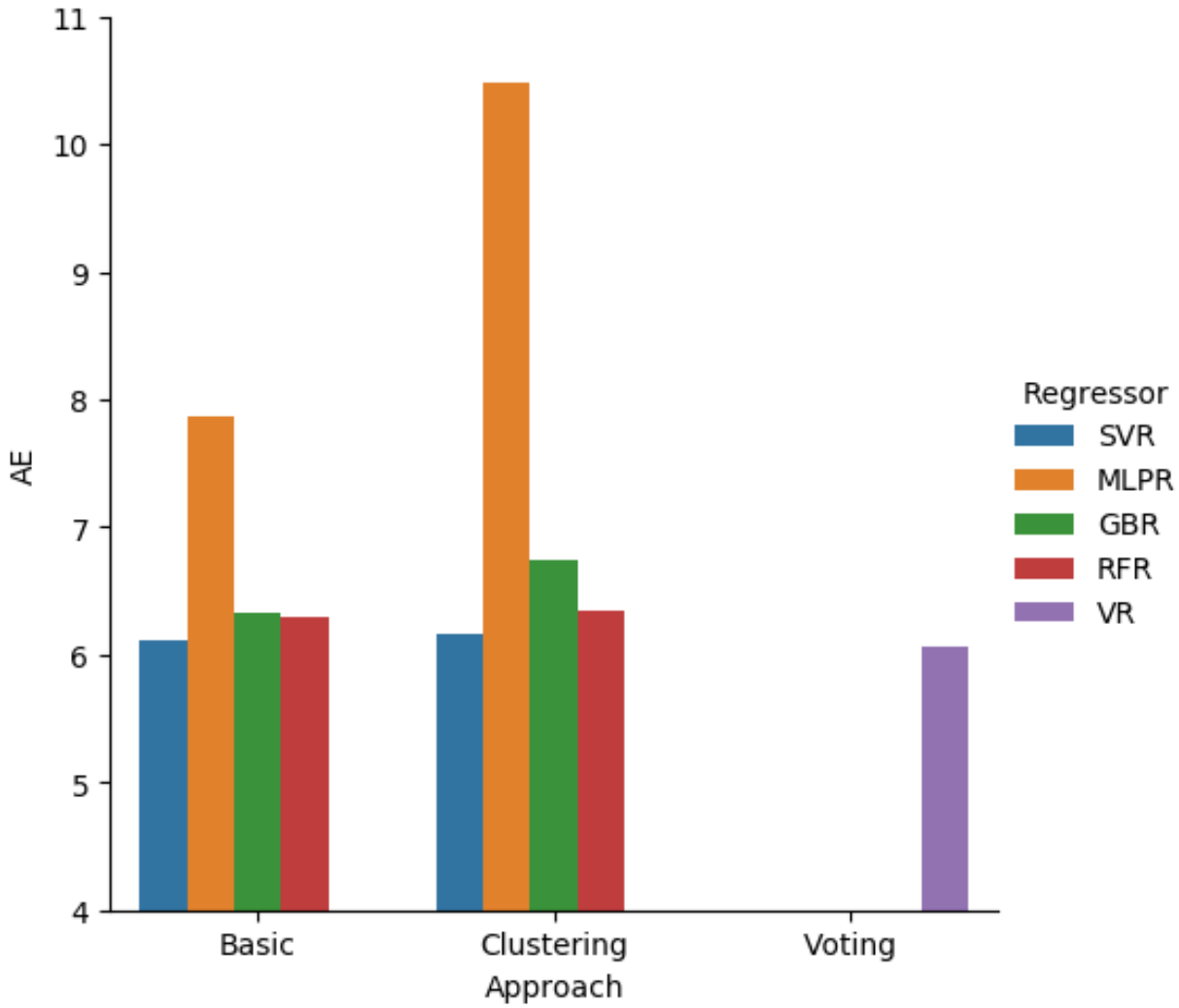
Αυτό μπορεί να οφείλεται στη μεγάλη ετερογένεια των descriptors που χαρακτηρίζουν



Σχήμα 6.2: Απεικόνιση του RMSE στις τρεις προσεγγίσεις

τις ουσίες του σετ δεδομένων. Η απόδοση αυτών, ίσως θα μπορούσε να βελτιωθεί ελάχιστα με αλλαγή διαφόρων παραμέτρων των αλγορίθμων. Έγιναν κάποιες δοκιμές με αλλάγη π.χ. στον αριθμό των clusters και στις στοιβάδες του MLPR, οι οποίες όμως δεν έδωσαν και πάλι ικανοποιητικά αποτελέσματα.

Στην παρούσα εργασία έγινε μια προσπάθεια, ώστε να απομονωθούν παράμετροι φαρμακευτικών ουσιών οι οποίες χορηγούνται ενδοφλέβια. Ο σκοπός της προσπάθειας αυτής ήταν να διερευνηθεί εάν τα μοντέλα πρόβλεψης που κατασκευάστηκαν για ουσίες με ποικίλλους τρόπους χορήγησης, θα είχαν καλύτερες επιδόσεις εάν είχαν κατασκευαστεί μόνο για το υποσύνολο ουσιών που χορηγούνται ενδοφλέφια. Σύνηθες πρόβλημα είναι



Σχήμα 6.3: Απεικόνιση του AE στις τρεις προσεγγίσεις

πως δεν υπάρχουν αναφορές στον τρόπο χορήγησης των ουσιών στις διάφορες έρευνες όπως και στη βάση δεδομένων που χρησιμοποιήθηκε για την εργασία αυτή [35]. Από την εργασία που έγινε το 2008 από τους Obach et al. οι οποίοι συνέλλεξαν φαρμακοκινητικά δεδομένα που αφορούσαν 670 φαρμακευτικές ουσίες, των οποίων είχε γίνει αποκλειστικά ενδοφλέβια χορήγηση [23], έγινε λήψη των 670 φαρμακευτικών ουσιών στις οποίες είχαν αναφερθεί.

Στη συνέχεια έγινε έλεγχος στο συνολικό data_set : 1106 x 601 και βρέθηκε ότι 377 από τις 670 φαρμακευτικές ουσίες υπήρχαν εκεί. Αυτές οι 377 ουσίες απομονώθηκαν σε ένα νέο dataset και ακολουθήθηκε εκ νέου η διαδικασία που περιγράφθηκε πιο πάνω για

	MAE	RMSE	AE
SVR	0.86	1.06	5.72
MLPR	0.95	1.18	5.87
GBR	0.90	1.12	5.73
RFR	0.83	1.01	5.31

Πίνακας 6.4: Αποτελέσματα Βασικής Προσέγγισης για ενδοφλέβια χορήγηση

	MAE	RMSE	AE
SVR	0.80	1.01	5.17
MLPR	1.01	1.33	5.62
GBR	0.95	1.22	5.55
RFR	0.81	1.03	5.19

Πίνακας 6.5: Αποτελέσματα Προσέγγισης clustering για ενδοφλέβια χορήγηση

το συνολικό dataset. Από το νέο data_set : 377 x 601 απομακρύνθηκαν οι στήλες που ήταν προβληματικές και απέμεινε το data_set : 377 x 595. Αφού απομακρύνθηκαν και οι ουσίες με σχετικά υψηλό χρόνο ημίσειας ζωής, απέμεινε το data_set : 352 x 595 το οποίο διαχωρίστηκε σε input_data : 352 x 594 και σε half_life (target value) : 352 x 1. Στη συνέχεια έγινε αντικατάσταση των missing values χρησιμοποιώντας το μέσο όρο κάθε στήλης, αφαιρέθηκαν τα πεδία με μηδενική βαρύτητα και απέμειναν το input_data : 352 x 583 και το half_life (target value) : 352 x 1. Έγινε απομάκρυνση των outliers και διαμορφώθηκαν σε input_data : 316 x 583 και σε half_life (target value) : 316 x 1 κι έγινε κανονικοποίηση των εναπομείναντων δεδομένων. Το υποσύνολο διαχωρίστηκε τυχαία σε training και test set, σε ποσοστά 67 - 33 %. Διαμορφώθηκαν λοιπόν τα data_train : 211 x 583 - hl_train : 211 x 1 και data_test : 105 x 583 - hl_test : 105 x 1. Με εφαρμογή της μεθόδου PCA, οι 583 μεταβλητές αντικαταστάθηκαν από τις 30 πιο σημαντικές συνιστώσες τους, οι οποίες ήταν αντιπροσωπευτικές του αρχικού συνόλου. Τα τελικά dataset που χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων, είναι το data_train : 211 x 30 - hl_train : 211 x 1 και το data_test : 105 x 30 - hl_test : 105 x 1.

Κατά τη βασική προσέγγιση όπου ακολουθήθηκαν τα ίδια βήματα με τα παραπάνω, παρήχθησαν τα εξής αποτελέσματα, Πίνακας 6.4.

Ακολούθως έγινε η εκπαίδευση των αλγορίθμων με τα καινούρια dataset και με την προσέγγιση clustering ακολουθώντας και πάλι τα ίδια βήματα, από όπου προέκυψαν τα αποτελέσματα του Πίνακα 6.5.

Κατά την προσέγγιση με voting όπου ακολουθήθηκε και πάλι η ίδια διαδικασία για τα νέα dataset, παρήχθησαν τα αποτελέσματα του Πίνακα 6.6.

	MAE	RMSE	AE
VR	0.84	1.03	5.54

Πίνακας 6.6: Αποτελέσματα Προσέγγισης voting για ενδοφλέβια χορήγηση

Στο Σχήμα 6.4, απεικονίζεται η τιμή του MAE κατά την υλοποίηση της βασικής προσέγγισης στο αρχικό dataset και στο intravenous subset. Στην δεύτερη περίπτωση οι τιμές είναι αρκετά μικρότερες για όλους τους αλγορίθμους.

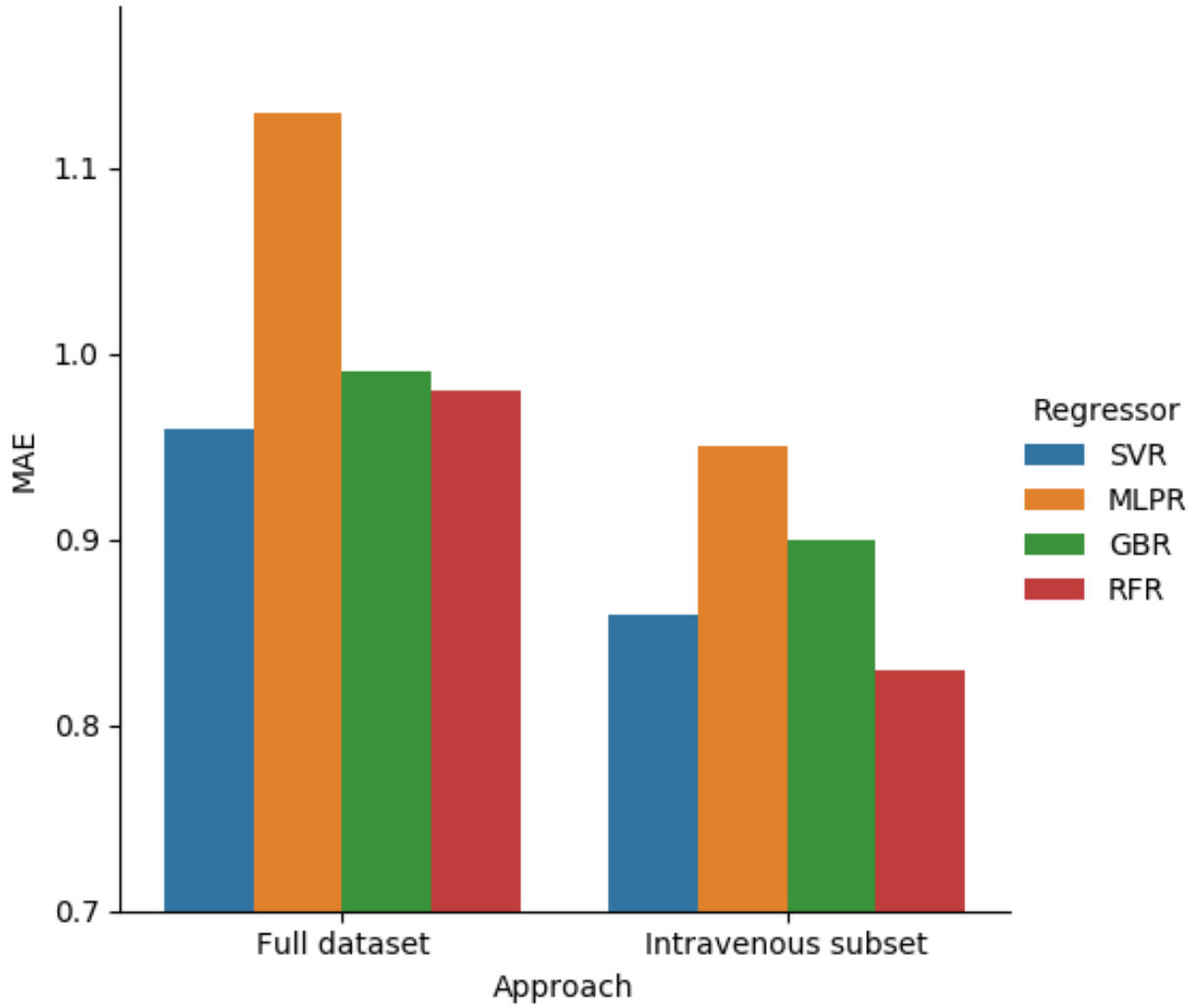
Η απεικόνιση των τιμών του RMSE για την βασική προσέγγιση στα δύο επιμέρους dataset (Full - intravenous subset) στο Σχήμα 6.5, δείχνει επίσης την μείωση της τιμής αυτού κατά την ενδοφλέβια χορήγηση.

Τέλος, στο Σχήμα 6.6, απεικονίζεται και η διαφορά της τιμής του AE στα επιμέρους dataset. Είναι εμφανής και σε αυτό το διάγραμμα η καλύτερη απόδοση του μοντέλου με τη χρήση του intravenous subset.

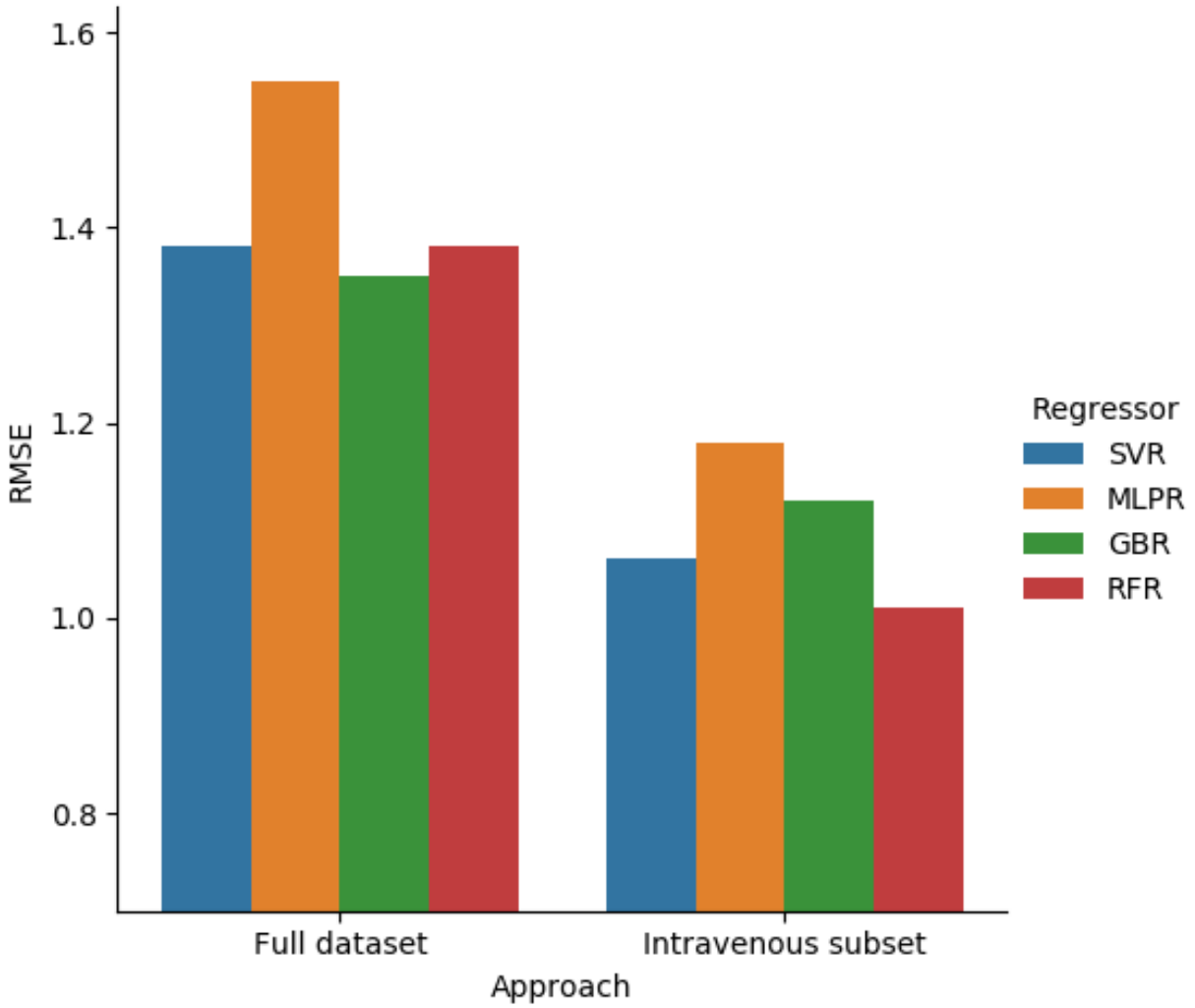
Το συμπέρασμα που προκύπτει κατά την ανάλυση των παραπάνω αποτελεσμάτων είναι πως οι αλγόριθμοι SVR, RFR και GBR αποδίδουν και πάλι καλύτερα στην προβλεπτική τους ικανότητα έναντι της εφαρμογής του νευρωνικού δικτύου MLPR, με μικρές διαφορές μεταξύ τους τόσο στην τιμή του RMSE όσο και του AE. Οι τιμές των δεικτών σφαλμάτων είναι μειωμένες σε σχέση με το αρχικό dataset, γεγονός που αποδεικνύει την καλύτερη προβλεπτική ικανότητα των μοντέλων κατά αυτή την υλοποίηση. Ωστόσο στο νέο dataset η προβλεπτική ικανότητα του νευρωνικού δικτύου MLPR, φαίνεται πως είναι εμφανώς πιο βελτιωμένη και οι τιμές των δεικτών σφαλμάτων είναι σχεδόν ίδιες με αυτές του GBR. Επίσης έχει βελτιωθεί κατά μία ώρα σχεδόν και το μέσο πραγματικό λάθος, για την τιμή του χρόνου ημίσειας ζωής.

Τα δεδομένα που συλλέχθησαν από την Drugbank ήταν από διάφορες πηγές με μεγάλες διακυμάνσεις, γεγονός που μπορεί να μειώσει την ποιότητα των υπολογιστικών μοντέλων που δημιουργήθηκαν. Η επεξεργασία των δεδομένων και ειδικότερα του χρόνου ημίσειας ζωής, έγινε βάση κάποιων κανόνων. Βασικοί παράγοντες που μπορεί να επηρεάσουν την ποιότητα αυτών, είναι η αναφορά της τιμής του χρόνου ημίσειας ζωής σε ενήλικες ή νεογνά, σε υγιείς ή αρρώστους καθώς και η οδός χορήγησης του φαρμάκου. Επίσης για πολλά φάρμακα υπήρχε ένα εύρος τιμών για το χρόνο ημίσειας ζωής αυτών, κατά το οποίο ελήφθη προς υπολογισμό η μέση τιμή αυτού. Σημαντικό ρόλο επίσης παίζει και το ποσοστό της πρωτεϊνικής σύνδεσης του φαρμάκου, που συνδέεται άμεσα με την απορρόφηση και την κατανομή αυτού. Μελλοντικός στόχος είναι η δημιουργία μοντέλων πρόβλεψης του χρόνου ημίσειας ζωής αλλά και άλλων παραμέτρων, που να μπορούν να

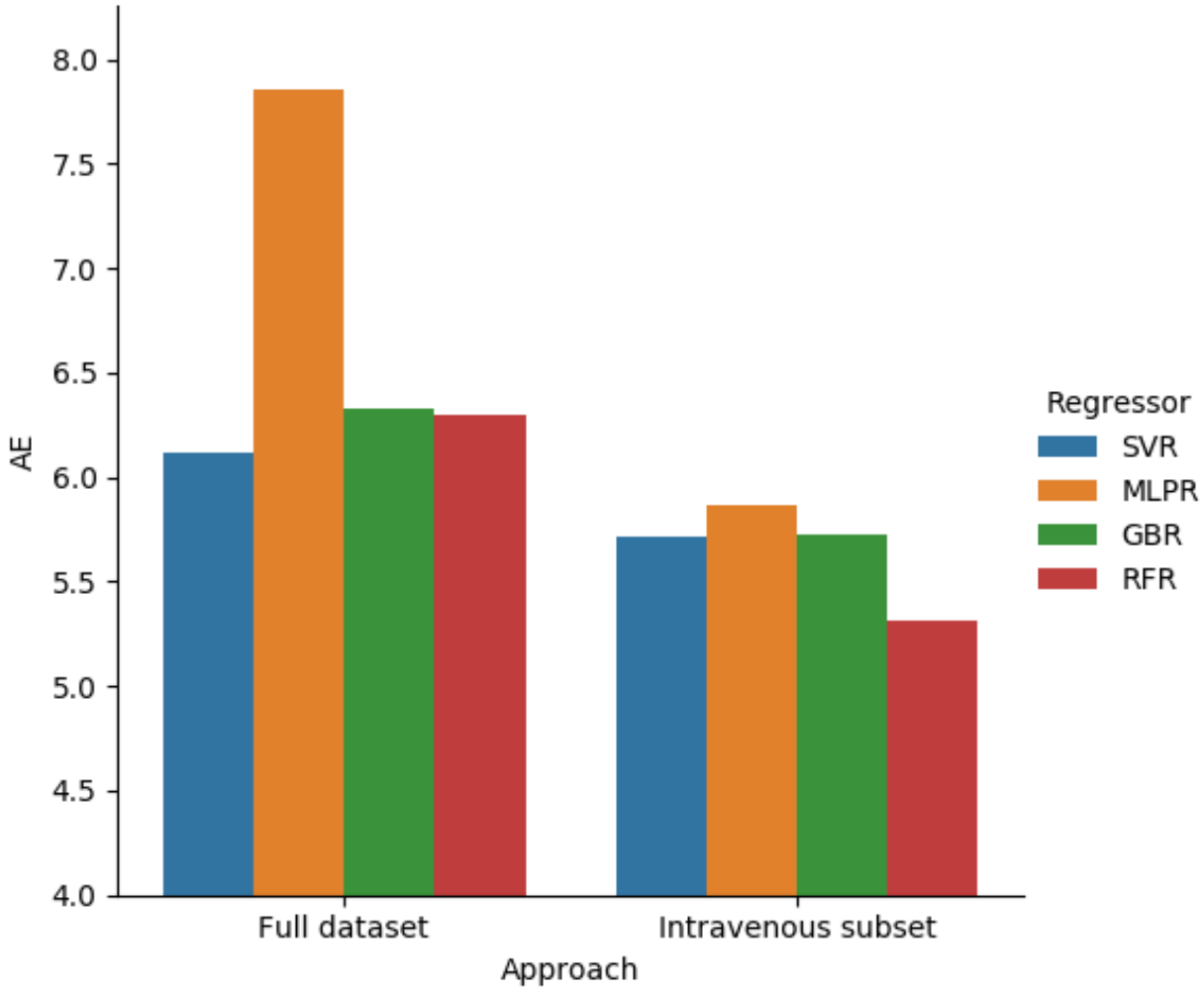
αξιολογήσουν ποσοτικά τη σχέση δομής - φυσικοχημικών - φαρμακοκινητικών- φαρμακοδυναμικών παραμέτρων, ώστε να δώσουν ασφαλείς προβλέψεις για παραμέτρους χρήσιμες για το σχεδιασμό νέων φαρμάκων. Σημαντικό ρόλο στη σωστή δημιουργία αυτών αποτελεί η σωστή συλλογή των δεδομένων με λιγότερες διακυμάνσεις.



Σχήμα 6.4: Απεικόνιση του MAE κατά τη βασική προσέγγιση του αρχικού dataset και του intravenous subset



Σχήμα 6.5: Απεικόνιση του RMSE κατά τη βασική προσέγγιση του αρχικού dataset και του intravenous subset



Σχήμα 6.6: Απεικόνιση του AE κατά τη βασική προσέγγιση του αρχικού dataset και του intravenous subset

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] S Agatonovic-Kustrin and R Beresford. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22(5):717 – 727, 2000.

[2] Jason Brownlee. A gentle introduction to the gradient boosting algorithm for machine learning, 2016.

[3] Certara. Simcyp simulator, 2018.

[4] Lujing Chen. Basic ensemble learning (random forest, adaboost, gradient boosting)-step by step explained.

[5] Manoranjenni Chetty, Rachel H. Rose, Khaled Abduljalil, Nikunjkumar Patel, Gaohua Lu, Theresa Cain, Masoud Jamei, and Amin Rostami-Hodjegan. Applications of linking pbpk and pd models to predict the impact of genotypic variability, formulation differences, differences in target binding capacity and target site drug concentrations on drug responses and variability. *Frontiers in Pharmacology*, 5:258, 2014.

[6] NECATI DEMIR. Ensemble methods: Elegant techniques to produce improved machine learning results, 2019.

[7] Georgios Drakos. Support vector machine vs logistic regression, 2018.

[8] Sergey Ermakov, Peter Forster, Jyotsna Pagidala, Marko Miladinov, Albert Wang, Rebecca Baillie, Derek Bartlett, Mike Reed, and Tarek Leil. Virtual systems pharmacology (visp) software for mechanistic system-level model simulations. *Frontiers in Pharmacology*, 5:232, 2014.

[9] Gabriel Helmlinger, Nidal Al-Huniti, Sergey Aksenov, Kirill Peskov, Karen M. Hallow, Lulu Chu, David Boulton, Ulf Eriksson, Bengt Hamrén, Craig Lambert, Eric Masson, Helen Tomkinson, and Donald Stanski. Drug-disease modeling in the pharmaceutical industry - where mechanistic systems pharmacology and statistical pharmacometrics meet. *European Journal of Pharmaceutical Sciences*, 109:S39 – S46, 2017. Special issue in honour of Professor Meindert Danhof.

[10] Sagar Howal. Ensemble learning in machine learning | getting started, 2017.

[11] Sagar Howal. Mean absolute error mae machine learning(ml), 2018.

[12] http://yahwes.github.io/. Mae and rmse — which metric is better?, 2016.

[13] Tarek Leil and Sergey Ermakov. Editorial: The emerging discipline of quantitative systems pharmacology. *Frontiers in Pharmacology*, 6:129, 2015.

[14] Tarek A. Leil and Richard Bertz. Quantitative systems pharmacology can reduce attrition and improve productivity in pharmaceutical research and development. *Frontiers in Pharmacology*, 5:247, 2014.

[15] E. W. Lowe, M. Butkiewicz, M. Spellings, A. Omlor, and J. Meiler. Comparative analysis of machine learning techniques for the prediction of logp. In *2011 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–6, April 2011.

[16] Jing Lu, Dong Lu, Xiaochen Zhang, Yi Bi, Keguang Cheng, Mingyue Zheng, and Xiaomin Luo. Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1860(11, Part B):2664 – 2671, 2016.

[17] Bernd Meibohm and Hartmut Derendorf. Basic concepts of pharmacokinetic/pharmacodynamic (pk/pd) modeling. *International journal of clinical pharmacology and therapeutics*, 35:401–13, 11 1997.

[18] Eugeniy Metlkin, Natalia Bagrova, Oleg Demin, and Neil Benson. A systems pharmacology model of anandamide dynamics after faah inhibitor administration. 2011.

[19] Pranov Mishra. Comprehensive support vector machines guide - using illusion to solve reality!, 2018.

[20] David J Newman and Gordon M Cragg. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of natural products*, 75(3):311–335, 2012.

[21] Marjoleen J.M.A. Nijsen, Fan Wu, Loveleena Bansal, Erica Bradshaw-Pierce, Jason R. Chan, Bianca M. Liederer, Jerome T. Mettetal, Patricia Schroeder, Edgar Schuck, Alice Tsai, Christine Xu, Anjaneya Chimalakonda, Kha Le, Mark Penney, Brian Topp, Akihiro Yamada, and Mary E. Spilker. Preclinical qsp modeling in the

pharmaceutical industry: An iq consortium survey examining the current landscape. *CPT: Pharmacometrics & Systems Pharmacology*, 7(3):135–146.

[22] novartis. Klinikes meletes.

[23] R. Scott Obach, Franco Lombardo, and Nigel J. Waters. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metabolism and Disposition*, 36(7):1385–1405, 2008.

[24] INSTITOYTO MELETHS OYROLOGIKON PATHISEON. Gnorimia me clinical trials.

[25] Mark C. Peterson and Matthew M. Riggs. A physiologically based mathematical model of integrated calcium homeostasis and bone remodeling. *Bone*, 46(1):49 – 63, 2010.

[26] MC Peterson and MM Riggs. Fda advisory meeting clinical pharmacology review utilizes a quantitative systems pharmacology (qsp) model: A watershed moment? *CPT: Pharmacometrics & Systems Pharmacology*, 4(3):189–192.

[27] Pilar Rey del Castillo. On the use of data mining for imputation. 06 2014.

[28] Edgar Schuck, Tonika Bohnert, Arijit Chakravarty, Valeriu Damian-Iordache, Christopher Gibson, Cheng-Pang Hsu, Tycho Heimbach, Anu Shilpa Krishnatry, Bianca M. Liederer, Jing Lin, Tristan Maurer, Jerome T. Mettetal, Daniel R. Mudra, Marjoleen JMA Nijsen, Joseph Raybon, Patricia Schroeder, Virna Schuck, Satyendra Suryawanshi, Yaming Su, Patrick Trapa, Alice Tsai, Majid Vakilynejad, Shining Wang, and Harvey Wong. Preclinical pharmacokinetic/pharmacodynamic modeling and simulation in the pharmaceutical industry: An iq consortium survey examining the current landscape. *The AAPS Journal*, 17(2):462–473, Mar 2015.

[29] George Seif. The 5 clustering algorithms data scientists need to know, 2018.

[30] Satish Sharan and Sukyung Woo. Systems pharmacology approaches for optimization of antiangiogenic therapies: challenges and opportunities. *Frontiers in Pharmacology*, 6:33, 2015.

[31] Christos Tsinopoulos and IP McCarthy. An evolutionary classification of the strategies for drug discovery. In *Manufacturing Complexity Network Conference, Cambridge*, pages 373–385, 2002.

[32] Joseph V Turner, Desmond J Maddalena, and David J Cutler. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. *International journal of pharmaceutics*, 270(1-2):209–219, 2004.

[33] Joseph V. Turner, Desmond J. Maddalena, David J. Cutler, and Snezana Agatonovic-Kustrin. Multiple pharmacokinetic parameter prediction for a series of cephalosporins. *Journal of Pharmaceutical Sciences*, 92(3):552–559, 2003.

[34] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[35] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2017.

[36] Jeffrey L. Woodhead, Kyunghee Yang, Scott Q. Siler, Paul B. Watkins, Kim L. R. Brouwer, Hugh A. Barton, and Brett A. Howell. Exploring bsep inhibition-mediated toxicity with a mechanistic model of drug-induced liver injury. *Frontiers in Pharmacology*, 5:240, 2014.

[37] Chun Wei Yap. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7):1466–1474, 2011.

[38] C.W. Yap, Z.R. Li, and Y.Z. Chen. Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods. *Journal of Molecular Graphics and Modelling*, 24(5):383 – 395, 2006.

[39] Majid ZANDKARIMI, Mohammad SHAFIEI, Farzin HADIZADEH, Mohammad Ali DARBANDI, and Kaveh TABRIZIAN. Prediction of pharmacokinetic parameters using a genetic algorithm combined with an artificial neural network for a series of alkaloid drugs. *Scientia Pharmaceutica*, 82(1):53–70, 2014.

[40] Zvetanka Zhivkova and Irini Doytchinova. Prediction of steady-state volume of distribution of acidic drugs by quantitative structure–pharmacokinetics relationships. *Journal of Pharmaceutical Sciences*, 101(3):1253 – 1266, 2012.

[41] Ιωάννης Σ. Βιζιριανάκης. Εξατομικευμένη Ιατρική Πρακτική & Φαρμακευτική Αγωγή, 2014.

[42] Άννα Δημόπουλος, Βασίλειος Τσαντίλη-Κακουλίδου. *Βασικές αρχές σχεδιασμού και ανάπτυξης φαρμάκων*. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[43] Βικτωρία Κόκοτος, Γεώργιος-Ισίδωρος Μαγκριώτη. *Φαρμακοχημεία*. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[44] Κωνσταντίνος Σωκράτη Μάνθος. Πρόβλεψη Τιμών στη Λιανική. Master's thesis, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Σχολή Θετικών Επιστημών, Τμήμα Πληροφορικής, Ελλάδα, 2018.

[45] Δημήτριος Πετρίδης. *Ανάλυση πολυμεταβλητών τεχνικών*. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[46] Ε. Σκαλτσά. *Ιστορία της φαρμακευτικής*. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[47] Δημήτριος Τσαρμπόπουλος. Σχεδιασμός και υλοποίηση μεθοδολογίας βραχυπρόθεσμης πρόβλεψης τιμών μετοχών του ελληνικού χρηματιστηρίου με συνδυασμό εξελικτικών αλγορίθμων, μηχανών διανυσμάτων υποστήριξης και τεχνικής κυλιόμενου παραθύρου, 2016.