



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΧΗΜΙΚΩΝ ΜΗΧΑΝΙΚΩΝ

**ΤΟΜΕΑΣ II: Ανάλυσης, Σχεδιασμού και Ανάπτυξης
Διεργασιών και Συστημάτων**

Πολυτεχνειούπολη Ζωγράφου, Αθήνα 157 80

Διπλωματική Εργασία

**Ανάπτυξη και αξιολόγηση μοντέλου μηχανικής μάθησης για
την πρόβλεψη της κυτταρικής πρόσληψης νανοϋλικών**

Θεοδωρή Αικατερίνη

Επιβλέπων: Χαράλαμπος Σαρίμβεης, Καθηγητής ΕΜΠ

Αθήνα, 2025

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στην δημιουργία ενός εκτενούς συνόλου δεδομένων και την χρήση αυτού για την ανάπτυξη μοντέλου επιβλεπόμενης μηχανικής μάθησης που προβλέπει την πρόσληψη νανοϋλικών σε διάφορες κυτταρικές σειρές. Το μοντέλο βασίζει τις προβλέψεις της κυτταρικής πρόσληψης (Νανოსωματίδια/κύτταρο) σε φυσικοχημικά χαρακτηριστικά νανοςωματιδίων (Σχήμα, Υδροδυναμική διάμετρος, Ζ-δυναμικό, Επικάλυψη), πειραματικές συνθήκες (Συγκέντρωση νανοςωματιδίων, Ιοντική ισχύς μέσου καλλιέργειας, Εφαρμογή υπερήχων) και βιολογικές παραμέτρους των κυττάρων (Μορφολογία).

Έπειτα από ενδελεχή βιβλιογραφική αναζήτηση, επιλέχθηκαν 71 πειραματικές έρευνες σχετικές με την μελέτη της πρόσληψης διαφόρων νανοςωματιδίων σε κυτταρικές σειρές *in vitro* και καταγράφηκαν τα πρωτογενή δεδομένα τους. Μέσω κατάλληλων τεχνικών προεπεξεργασίας δεδομένων και με τη θεώρηση ορισμένων παραδοχών σχετικά με το σχήμα και την πυκνότητα των νανοςωματιδίων, όλες οι μεταβλητές κωδικοποιήθηκαν σε συνεχείς ή διακριτές αριθμητικές μεταβλητές. Αρχικά επιχειρήθηκε η μείωση του αριθμού των κατηγοριών των κατηγορικών μεταβλητών μέσω περαιτέρω ομαδοποίησής τους, ενώ οι ελλειπείς τιμές συμπληρώθηκαν μέσω της επαναληπτικής μεθοδολογίας MICE (με χρήση μοντέλου παλινδρόμησης Random Forest). Οι κατηγορικές μεταβλητές κωδικοποιήθηκαν ως αριθμητικές μέσω της μεθόδου «One-hot encoding», και τα δεδομένα εισόδου κανονικοποιήθηκαν με την μέθοδο Z-score. Επίσης, η μεταβλητή εξόδου υπέστη λογαριθμική μετατροπή με σκοπό την μείωση του εύρους τιμών της. Έτσι, διαμορφώθηκε ένα πρωτότυπο σύνολο δεδομένων κατάλληλο για εφαρμογές μηχανικής μάθησης στην περιοχή της νανοπληροφορικής. Το σύνολο των δεδομένων διαχωρίστηκε σε σύνολα εκπαίδευσης και ελέγχου (ποσοστό 80%/20%). Οι βέλτιστες τιμές των υπερπαραμέτρων του μοντέλου XGBoost προσδιορίστηκαν μέσω της μεθοδολογίας «cross-validation» με 10 επαναλήψεις και το μοντέλο αξιολογήθηκε μέσω στατιστικών μέτρων (R^2 , MAE, MSE, RMSE) και μεθόδων όπως η «leave-one-out cross-validation» και η τυχαία αντικατάσταση των τιμών της μεταβλητής εξόδου (Y-randomization). Τέλος, υπολογίστηκε το πεδίο εφαρμοσιμότητας του μοντέλου με την μέθοδο k-κοντινότερων γειτόνων (k=5) και συζητήθηκαν οι σχέσεις μεταξύ των μεταβλητών μέσω ανάλυσης SHAP.

Κατά την εκπαίδευση και αξιολόγηση του μοντέλου παρατηρήθηκε μία σημαντική εξάρτηση των προβλέψεων του μοντέλου από την μεταβλητή του Χρόνου, η οποία κρίθηκε απαραίτητο να αφαιρεθεί ώστε να διευκολυνθεί η ερμηνευτική ανάλυση της συμμετοχής των υπόλοιπων μεταβλητών. Για τον σκοπό αυτό, επιλέχθηκε η χρονική στιγμή στην οποία είχαν ληφθεί οι περισσότερες μετρήσεις του συνόλου δεδομένων - οι 24 ώρες- και τα υπόλοιπα δεδομένα αφαιρέθηκαν. Ακολούθως, μέσω δοκιμής και σφάλματος μειώθηκε ο αριθμός των μεταβλητών εισόδου, στοχεύοντας στην δημιουργία του πιο απλού μοντέλου που προβλέπει με ακρίβεια τα δεδομένα. Το τελικό μοντέλο XGBoost κατάφερε ικανοποιητική πρόβλεψη στα δεδομένα του συνόλου ελέγχου, καταγράφοντας συντελεστή τιμή της μετρικής $R^2 = 0.668$. Κατά την διαδικασία του Y-randomization καταγράφηκαν αρνητικές τιμές R^2 , γεγονός που αποκλείει την τυχαία εύρεση σχέσεων μεταξύ των μεταβλητών του αρχικού μοντέλου.

Η ανάλυση ερμηνευσιμότητας SHAP του μοντέλου ανέδειξε τη Συγκέντρωση των νανοσωματιδίων ως τον σημαντικότερο παράγοντα θετικής επίδρασης στην κυτταρική πρόσληψη. Επιπλέον, μικρότερες Υδροδυναμικές διαμέτροι και ακραίες τιμές Ζ-δυναμικού φάνηκε πως σχετίζονται με αυξημένη ενδοκυττάρωση, ενώ η απουσία επιφανειακής επικάλυψης ευνοεί επίσης την πρόσληψη νανοσωματιδίων στις κυτταρικές σειρές που μελετήθηκαν. Τέλος, η κυτταρική μορφολογία των ινοβλαστών και η χαμηλή Ιοντική ισχύς του μέσου καλλιέργειας ενισχύουν περαιτέρω την κυτταρική πρόσληψη των νανοσωματιδίων.

Τα θετικά αποτελέσματα της αξιολόγησης του μοντέλου, σε συνδυασμό με τις ερμηνευτικές σχέσεις που ευθυγραμμίζονται σε μεγάλο βαθμό με τα «μοτίβα» που αναδεικνύονται σε αρκετές μελέτες του μηχανισμού της ενδοκυττάρωσης νανοσωματιδίων, επιβεβαιώνουν την αξιοπιστία της προσέγγισης. Το μοντέλο αυτό μπορεί να συμβάλλει στη μείωση του κόστους και του χρόνου πειραματικής ανάλυσης, υποστηρίζοντας την ανάπτυξη βιοϊατρικών εφαρμογών.

Λέξεις-κλειδιά: Νανοσωματίδια, κυτταρική πρόσληψη, επιβλεπόμενη μηχανική μάθηση, XGBoost, ανάλυση ερμηνευσιμότητας, SHAP.

Abstract

The present thesis focuses on the creation of an extensive dataset that is used for the development of a supervised machine learning model that predicts the uptake of nanoparticles by various cell lines. The model predictions of cellular uptake (Nanoparticles/cell) are based on nanoparticles physicochemical properties (Shape, Hydrodynamic diameter, Z-potential, Surface coating), experimental conditions (Nanoparticles concentration, growth medium Ionic strength, Sonication), and cell lines biological characteristics (Cell morphology).

Following thorough bibliographical research, 71 experimental studies investigating the *in vitro* cellular uptake of various nanoparticles were elected and incorporated in the initial dataset as raw data. Through appropriate data preprocessing techniques and by making certain assumptions regarding the shape and density of the nanoparticles, all variables were encoded as continuous or discrete numerical variables. Firstly, the dimensionality of categorical variables was reduced through further grouping, while missing data were imputed using the MICE algorithm (with a Random Forest regression estimator). Categorical variables were numerically encoded via «One-hot encoding», and all input features were normalized using the Z-score method. Additionally, a logarithmic transformation was applied to the target variable to reduce its value range. Thus, an original dataset was constructed, suitable for machine learning applications in the field of nanoinformatics. The data were split into a training and testing subset (ratio 80/20). The optimal hyperparameter values of the XGBoost algorithm were determined by a 10-fold cross-validation, and the fine-tuned model was subsequently evaluated using both standard statistical metrics (R^2 , MAE, MSE, RMSE) and common validation techniques, such as leave-one-out cross-validation and Y-randomization. Finally, the model's applicability domain was defined via the k-nearest neighbours method ($k=5$), while interpretative insights into variable relationships were obtained via a SHAP analysis.

During model training and validation, the model predictions were found to heavily rely on the Time variable, which was removed in order to allow for the remaining variables' relationships interpretation. To this end, only the time point which the majority of the experiments included -the 24-hour time point- was retained in the dataset. Following this major size reduction, the number of input features was reduced through a trial-and-error process, aiming to develop the simplest yet accurate machine learning model. Indeed, the final XGBoost model achieved an R^2 metric value of 0.668. The Y-randomization process recorded negative R^2 values, confirming that the original model did not captures relationships by chance. The SHAP interpretability analysis highlighted nanoparticle Concentration as the most influential factor, positively affecting cellular uptake. Additionally, smaller Hydrodynamic diameters and extreme Z-potential values were found to be associated with increased endocytosis, while the absence of surface coating also promoted nanoparticle uptake in the examined cell lines. Finally, the fibroblast cell morphology and low culture medium Ionic strength were found to enhance nanoparticle internalization.

The positive model validation results, along with the interpretative variable relationships that align with patterns observed in numerous experimental studies on

endocytosis mechanisms, support the robustness of the proposed approach. The developed model can be used as a tool for further improving the accuracy of cellular uptake predictions, contributing to the reduction of experimental costs and supporting the development of biomedical applications.

Key-words: Nanoparticles, cellular uptake, supervised machine learning, XGBoost, interpretative analysis, SHAP.

Πίνακας Περιεχομένων

Περίληψη	2
Abstract.....	4
Κατάλογος σχημάτων.....	8
Κατάλογος εικόνων	9
Κατάλογος πινάκων.....	10
Κατάλογος διαγραμμάτων	12
Πίνακας συντομογραφιών	14
1. Εισαγωγή	15
1.1. Νανοϋλικά και φαρμακευτικές εφαρμογές.....	15
1.2. Πειραματικός προσδιορισμός της κυτταρικής πρόσληψης νανοσωματιδίων ..	16
1.3. Μηχανική μάθηση και μείωση πειραματικού κόστους.....	17
1.4. Προγενέστερες έρευνες	18
1.5. Δομή της διπλωματικής εργασίας	19
2. Βασικές παράμετροι της κυτταρικής πρόσληψης νανοσωματιδίων	22
2.1. Φυσικοχημικές ιδιότητες νανοσωματιδίων	22
2.1.1. Μέγεθος νανοσωματιδίων.....	22
2.1.2. Σχήμα νανοσωματιδίων	23
2.1.3. Επιφανειακό φορτίο νανοσωματιδίων	24
2.1.4. Επιφανειακή χημεία νανοσωματιδίων	25
2.2. Πειραματικές παράμετροι που επηρεάζουν την συσσωμάτωση	26
2.3. Μηχανισμοί κυτταρικής πρόσληψης.....	26
2.3.1. Φαγοκυττάρωση	27
2.3.2. Μακροπινοκύττωση.....	27
2.3.3. Ενδοκυττάρωση εξαρτώμενη από κλαθρίνη	27
2.3.4. Ενδοκυττάρωση εξαρτώμενη από καβεολίνη	28
2.3.5. Ενδοκυττάρωση ανεξάρτητη από κλαθρίνη και καβεολίνη.....	29
3. Θεωρητικές βάσεις και μεθοδολογικές πρακτικές στη μηχανική μάθηση	31
3.1. Προεπεξεργασία δεδομένων.....	31
3.1.1. Διαχωρισμός δεδομένων σε σύνολο εκπαίδευσης και ελέγχου.....	31
3.1.2. Μετατροπή κατηγορικών μεταβλητών σε αριθμητικές.....	32
3.1.3. Κανονικοποίηση μεταβλητών εισόδου	32
3.1.4. Συμπλήρωση ελλিপών τιμών	33
3.2. Μοντελοποίηση δεδομένων και βελτιστοποίηση μοντέλου μηχανικής μάθησης	33

3.2.1. Ο αλγόριθμος XGBoost	33
3.2.2. Υπερπαραμέτροι των μοντέλων XGBoost.....	35
3.2.3. Επιλογή βέλτιστων τιμών υπερπαραμέτρων μέσω «cross-validation»	36
3.3. Αξιολόγηση και πεδίο εφαρμοσιμότητας μοντέλων μηχανικής μάθησης.....	37
3.3.1. Αξιολόγηση μοντέλων παλινδρόμησης	37
3.3.2. Πεδίο εφαρμοσιμότητας μοντέλων μηχανικής μάθησης	39
3.4. Ερμηνεία μοντέλων μέσω της μεθόδου «Shapley Additive exPlanations»	40
4. Συλλογή και διαχείριση δεδομένων	42
4.1. Συλλογή πειραματικών ερευνών.....	42
4.2. Αξιολόγηση και επιλογή δεδομένων	43
4.3. Δημιουργία συνόλου δεδομένων	44
4.4. Επεξεργασία των δεδομένων.....	46
4.4.1. Αφαίρεση μεταβλητών.....	46
4.4.2. Εμπλουτισμός συνόλου δεδομένων με δευτερογενείς μεταβλητές.....	47
4.4.3. Μετατροπή μονάδων μέτρησης.....	50
4.4.4. Διαχείριση ελλιπών τιμών	54
5. Μεθοδολογία μηχανικής μάθησης για την κυτταρική πρόσληψη νανοσωματιδίων: Βήματα ανάπτυξης, αξιολόγησης και ερμηνείας μοντέλων	56
5.1. Προεπεξεργασία συνόλου δεδομένων	56
5.2. Δημιουργία και επιλογή του βέλτιστου μοντέλου μηχανικής μάθησης.....	57
5.3. Επικύρωση, πεδίο εφαρμοσιμότητας και ερμηνεία μοντέλου.....	58
5.4. Προγραμματιστικά εργαλεία	59
6. Αποτελέσματα – Συζήτηση	61
6.1. Το αρχικό σύνολο δεδομένων και η μεγάλη βαρύτητα της μεταβλητής του Χρόνου.....	61
6.2. Το αναθεωρημένο σύνολο δεδομένων	65
6.2.1. Οπτικοποίηση δεδομένων	65
6.2.2. Εκπαίδευση και αξιολόγηση μοντέλου XGBoost	69
6.3. Αφαίρεση μεταβλητών και απλοποίηση μοντέλου XGBoost.....	71
6.4. Αξιολόγηση και Πεδίο Εφαρμοσιμότητας απλοποιημένου μοντέλου	74
6.5. Ερμηνεία απλοποιημένου μοντέλου	77
7. Συμπεράσματα – Προτάσεις για μελλοντική μελέτη	86
Παράρτημα	89
Βιβλιογραφία	90

Κατάλογος σχημάτων

Σχήμα 1. Συμμετοχή στις προβλέψεις του μοντέλου μηχανικής μάθησης μεταβλητών που κωδικοποιούν τις φυσικοχημικές ιδιότητες των νανοσωματιδίων, τις πειραματικές παραμέτρους και τα μορφολογικά χαρακτηριστικά της κυτταρικής σειράς	20
Σχήμα 2. Αναπαράσταση αλγορίθμου XGBoost (τροποποιημένο από Yao et al. (2022) ⁷⁷).....	35
Σχήμα 3. Διάγραμμα ροής της διαδικασίας συλλογής και αξιολόγησης πειραματικών ερευνών για την δημιουργία του συνόλου δεδομένων.	43
Σχήμα 4. Τυπικό διάγραμμα ροής μοντελοποίησης δεδομένων με μεθόδους μηχανικής μάθησης (τροποποιημένο από Ponce-Bobadila et al. (2024) ⁹³).....	56
Σχήμα 5. Πίνακας «Heat map» της συσχέτισης των μεταβλητών εισόδου του συνόλου δεδομένων. Με έντονο κόκκινο χρώμα φαίνεται η υψηλή θετική συσχέτιση των μεταβλητών της Ονομαστική (Nominal size (nm)) και Μέσης υδροδυναμικής (Mean Hydrodynamic size (nm)) διαμέτρου.	62

Κατάλογος εικόνων

Εικόνα 1. Παράγοντες που επηρεάζουν την κυτταρική πρόσληψη των νανοσωματιδίων. A) Επιφανειακό φορτίο B) Σχήμα C) Μέγεθος και D) Επιφανειακή χημεία των νανοσωματιδίων (τροποποιημένο από Foroozandeh et al. (2018) ²⁶).....	22
Εικόνα 2. Σχηματική αναπαράσταση των διαφορετικών μηχανισμών κυτταρικής πρόσληψης (τροποποιημένο από Conner and Schmid (2003) ⁵⁰).....	26
Εικόνα 3. Σύνθετο αίτημα αναζήτησης σχετικών ερευνών στη βάση δεδομένων «PubMed».....	42

Κατάλογος πινάκων

Πίνακας 1. Επεξήγηση δευτερογενών μεταβλητών που προκύπτουν από την ομαδοποίηση των κατηγοριών των μεταβλητών Νανοσωματίδιο, Μέθοδος ποσοτικοποίησης και Επικάλυψη νανοσωματιδίων.....	48
Πίνακας 2. Ιοντική ισχύς των μέσων καλλιέργειας κυττάρων.....	49
Πίνακας 3. Δευτερογενή δεδομένα σχετικά με την κατηγορία, το είδος προέλευσης και την μορφολογία των κυττάρων που περιλαμβάνονται στο πρωτογενές σύνολο δεδομένων.....	50
Πίνακας 4. Επεξήγηση μεταβλητών του συνόλου δεδομένων.....	54
Πίνακας 5. Πιθανές τιμές των παραμέτρων του μοντέλου XGBoost που δοκιμάστηκαν για την εύρεση του βέλτιστου μοντέλου με βάση το στατιστικό μέτρο R^2	58
Πίνακας 6. Παρουσίαση «βιβλιοθηκών» και πακέτων της γλώσσας προγραμματισμού Python που χρησιμοποιήθηκαν σε κάθε υπολογιστική ανάλυση.....	59
Πίνακας 7. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία του «cross-validation» στο αρχικό σύνολο εκπαίδευσης.....	63
Πίνακας 8. Στατιστικά μέτρα στο σύνολο εκπαίδευσης πριν και μετά την λογαριθμική μετατροπή της μεταβλητής εξόδου για το αρχικό μοντέλο XGBoost.....	63
Πίνακας 9. Βέλτιστες τιμές των παραμέτρων του μοντέλου XGboost με βάση το στατιστικό μέτρο R^2	69
Πίνακας 10. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία του «cross-validation» στο αναθεωρημένο σύνολο εκπαίδευσης.....	70
Πίνακας 11. Στατιστικά μέτρα στο αναθεωρημένο σύνολο εκπαίδευσης πριν και μετά την λογαριθμική μετατροπή της μεταβλητής εξόδου για το μοντέλο XGBoost.....	70
Πίνακας 12. Συντελεστής προσδιορισμού για το αναθεωρημένο σύνολο εκπαίδευσης μετά την αφαίρεση των μεταβλητών Z-δυναμικό, Επικάλυψη νανοσωματιδίων, Σχήμα νανοσωματιδίων και Ιοντική ισχύς του μέσου καλλιέργειας και την εκ νέου εκπαίδευση του μοντέλου XGBoost.....	72
Πίνακας 13. Μεταβλητές εισόδου που επιλέχθηκαν για την εκπαίδευση του απλοποιημένου μοντέλου XGBoost.....	73
Πίνακας 14. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία του «cross-validation» στο απλοποιημένο σύνολο εκπαίδευσης.....	74
Πίνακας 15. Στατιστικά μέτρα στο απλοποιημένο σύνολο εκπαίδευσης πριν και μετά την λογαριθμική μετατροπή της μεταβλητής εξόδου για το μοντέλο XGBoost.....	75

Πίνακας 16. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία της τυχαίας αντικατάστασης των τιμών της μεταβλητής εξόδου..... 75

Πίνακας 17. Πεδίο εφαρμοσιμότητας του απλοποιημένου μοντέλου και αξιολόγηση των προβλέψεων στο σύνολο ελέγχου 76

Κατάλογος διαγραμμάτων

Διάγραμμα 1. Παράδειγμα εύρεσης πεδίου εφαρμοσιμότητας με την μέθοδο kNN. Τα σημεία εκείνα των οποίων η απόσταση από τους κοντινότερους γείτονες ξεπερνά το προκαθορισμένο κατώφλι απόστασης, επισημασμένα με κόκκινο, θεωρούνται εκτός του πεδίου εφαρμοσιμότητας, και κατ' επέκταση η πρόβλεψη του μοντέλου για αυτά τα σημεία δεν θεωρείται αξιόπιστη.....	40
Διάγραμμα 2. Αριθμός μελετών ανά έτος που περιλαμβάνονται στο σύνολο δεδομένων.....	44
Διάγραμμα 3. Σημαντικότητα μεταβλητών για την προβλεπτική ικανότητα του αρχικού μοντέλου XGBoost, όπως προκύπτει από την ιδιότητα (attribute) «.feature_importances_» της βιβλιοθήκης XGBoost στην Python. Η μεταβλητή της Χρονικής διάρκειας (Incubation time) φαίνεται να υπερισχύει σημαντικά σε σχέση με τις υπόλοιπες.....	64
Διαγράμματα 4-6. Ιστογράμματα των μεταβλητών 4) Μέση υδροδυναμική διάμετρος 5) Z-δυναμικό 6) Ιοντική ισχύς (mol/L).....	66
Διαγράμματα 7-15. Κατανομή των κατηγοριών των κατηγορικών μεταβλητών 7) Είδος νανοσωματιδίων 8) Σχήμα 9)Είδος επικάλυψης 10) Κατηγορία κυτταρικής σειράς 11) Είδος προέλευσης κυτταρικής σειράς 12) Μορφολογία κυτταρικής σειράς 13) Πενικιλίνη/Στρεπτομυκίνη (Ναι/Όχι) 14) Εφαρμογή υπερήχων 15) Κατηγορία μεθόδων ποσοτικοποίησης κυτταρικής πρόσληψης.....	69
Διάγραμμα 16. Σημαντικότητα μεταβλητών για την προβλεπτική ικανότητα του αναθεωρημένου μοντέλου XGBoost όπως προκύπτει από την ιδιότητα (attribute) «.feature_importances_» της βιβλιοθήκης XGBoost στην Python.....	71
Διάγραμμα 17. Αντιπαραβολή προβλεπόμενων και πραγματικών τιμών (σε λογαριθμική κλίμακα) για το μοντέλο XGBoost. Οι μπλε κουκκίδες αντιστοιχούν σε δείγματα εντός του Πεδίου Εφαρμοσιμότητας (AD), ενώ οι κόκκινες σε δείγματα εκτός. Η εγγύτητα στην ευθεία $y = x$ υποδηλώνει μεγαλύτερη ακρίβεια πρόβλεψης.	77
Διάγραμμα 18. Σημαντικότητα μεταβλητών για την προβλεπτική ικανότητα του απλοποιημένου μοντέλου XGBoost, όπως προκύπτει από την ιδιότητα (attribute) «.feature_importances_» της βιβλιοθήκης XGBoost στην Python.....	77
Διάγραμμα 19. Σημαντικότητα μεταβλητών για την διαμόρφωση των προβλέψεων του απλοποιημένου μοντέλου XGBoost σύμφωνα με τις τιμές SHAP.....	78
Διάγραμμα 20. Διάγραμμα τύπου «beeswarm» από την ανάλυση SHAP. Το γράφημα απεικονίζει το εύρος και την κατεύθυνση της επίδρασης κάθε μεταβλητής στις προβλέψεις του μοντέλου. Η απόσταση των σημείων από την τιμή 0 στον οριζόντιο άξονα αντιστοιχεί στο μέγεθος της επίδρασης (μεγαλύτερη απόσταση σημαίνει ισχυρότερη επίδραση), ενώ η θέση τους δεξιά ή αριστερά από το μηδέν δηλώνει θετική ή αρνητική συνεισφορά, αντίστοιχα. Το χρώμα κάθε σημείου υποδηλώνει την	

τιμή της αντίστοιχης μεταβλητής (μπλε για μικρές, κόκκινο για μεγάλες και μωβ για ενδιάμεσες τιμές)..... 80

Διάγραμμα 21. Διάγραμμα «waterfall» επεξήγησης της επίδρασης των πιο σημαντικών μεταβλητών στην διαμόρφωση μίας μεμονωμένης πρόβλεψης..... 84

Πίνακας συντομογραφιών

Συντομογραφία	Επεξήγηση
AD	Applicability Domain
AI	Artificial Intelligence
BSA	Bovine Serum Albumin
ECGM	Endothelial Cell Growth Medium
ECM	Endothelial Cell Medium
ELS	Electrophoretic Light Scattering
FBS	Fetal Bovine Serum
FEME	Fast Endophilin-Mediated Endocytosis
ICP-AAS	Inductively Coupled Plasma Atomic Absorption Spectrometry
ICP-MS	Inductively Coupled Plasma Mass Spectrometry
ICP-OES	Inductively Coupled Plasma Optical Emission Spectrometry
IS	Ionic Strength
kNN	k-Nearest Neighbours
LDL	Low Density Lipoprotein
MICE	Multiple Imputation by Chained Equations
MS	Mass Spectrometry
NP	Nanoparticle
PBK	Physiologically Based Kinetic
PEG	Polyethylene glycol
PLA	Polylactic Acid
QSAR	Quantitative Structure-Activity Relationship
SEM	Scanning Electron Microscopy
SHAP	Shapley Additive exPlanations
SMILES	Simplified Molecular Input Line Entry System
SPIONS	Superparamagnetic Iron Oxide Nanoparticles
TEM	Transmission Electron Microscopy
XGBoost	extreme Gradient Boosting

1. Εισαγωγή

1.1. Νανοϋλικά και φαρμακευτικές εφαρμογές

Η νανοτεχνολογία είναι ο κλάδος που μελετά τα υλικά που βρίσκονται στην νανοκλίμακα, τα λεγόμενα νανοϋλικά. Τυπικά ως νανοϋλικά αναφέρονται σωματίδια μεγέθους 1-100 nm (αν και συχνά μεγαλύτερα σωματίδια χαρακτηρίζονται επίσης ως νανοϋλικά) διότι σε αυτό το εύρος μεγεθών εμφανίζονται κβαντικές αλληλεπιδράσεις που διαφέρουν από τις χημικές αλληλεπιδράσεις της ύλης σε ατομικό ή μοριακό επίπεδο. Οι ιδιαίτερες ιδιότητες κοινών υλικών στην νανοκλίμακα τα καθιστούν εξαιρετικά χρήσιμα σε τομείς όπως η βιοϊατρική, η κοσμητολογία και η ηλεκτρονική¹.

Η χρήση νανοσωματιδίων στην βιοϊατρική έχει ως βασικό στόχο την στοχευμένη χορήγηση των θεραπευτικών παραγόντων, δηλαδή της εγκλεισμένης ή επιφανειακά προσδεδεμένης φαρμακευτικής ουσίας ή αυτού καθ' αυτού του νανοϋλικού, στον επιθυμητό ιστό ή κύτταρο. Η στοχευμένη μεταφορά νανοσωματιδίων δίνει την δυνατότητα αύξησης της συγκέντρωσης του θεραπευτικού παράγοντα στα κύτταρα ή ιστούς-στόχους χωρίς την ανάλογη αύξηση της χορηγούμενης δόσης, μειώνοντας έτσι τις πιθανές τοξικές παρενέργειες της θεραπείας. Επιπροσθέτως, η χρήση νανοφορέων δίνει την δυνατότητα βελτίωσης της βιοδιαθεσιμότητας μη διαλυτών στο νερό φαρμακευτικών ουσιών. Τέλος, ο συνδυασμός νανοσωματιδίων με σκιαγραφικούς παράγοντες διευκολύνει την απεικόνιση και παρακολούθηση της διαδρομής χορήγησης των νανοσωματιδίων για την διάγνωση ασθενειών².

Όλα αυτά τα πιθανά πλεονεκτήματα της στοχευμένης χορήγησης νανοϋλικών σε βιολογικούς ιστούς ή κύτταρα έχουν οδηγήσει στην προσπάθεια σύνθεσης και βελτίωσης διαφορετικών τύπων νανοσωματιδίων. Οι πιο συνηθισμένες κατηγορίες περιλαμβάνουν: Ανόργανα νανοσωματίδια (σιδήρου, χρυσού, αργύρου, χαλκού, κβαντικές τελείες), οργανικά (χιτοζάνης, λιπιδικά, βιοδιασπώμενα πολυμερή όπως πολυγαλακτικό οξύ) αλλά και δενδριμερή και φουλερένια³. Τα νανοϋλικά αυτά βρίσκουν ευρεία εφαρμογή στην στοχευμένη μεταφορά αντικαρκινικών φαρμάκων κυρίως μέσω συνδυασμού αυτών με πολυμερικά νανοσωματίδια (π.χ. το εμπορικό φάρμακο Doxil®). Επίσης, νανοσωματίδια έχουν χρησιμοποιηθεί για την μεταφορά φαρμάκων για την αντιμετώπιση βακτηριακών λοιμώξεων, αυτοάνοσων νοσημάτων, νευροεκφυλιστικών και οπτικών ασθενειών². Σημαντική είναι η συμβολή της νανοτεχνολογίας στην στοχευμένη μεταφορά αυξητικών παραγόντων για την ανακατασκευή ιστών, αλλά και η αξιοποίηση τους ως φορείς αντιγόνων σε εμβόλια mRNA (π.χ. εμβόλιο κατά του COVID-19)⁴.

Σε όλες τις παραπάνω περιπτώσεις, ο ακριβής σχεδιασμός των θεραπευτικών προσεγγίσεων προϋποθέτει τη σε βάθος κατανόηση της κινητικής πρόσληψης των νανοσωματιδίων από τα κύτταρα-στόχους, ώστε να καθοριστεί η κατάλληλη δόση που θα οδηγήσει στα επιθυμητά θεραπευτικά αποτελέσματα. Η γνώση της συγκέντρωσης του θεραπευτικού παράγοντα εντός των κυττάρων σε κάθε χρονική στιγμή αποτελεί τη βάση για τη βελτιστοποίηση του θεραπευτικού σχεδιασμού με σκοπό τον καθορισμό της ελάχιστης αποτελεσματικής δόσης και την ταυτόχρονη ελαχιστοποίηση των πιθανών τοξικών παρενεργειών.

1.2. Πειραματικός προσδιορισμός της κυτταρικής πρόσληψης νανοσωματιδίων

Η ποσοτικοποίηση των νανοσωματιδίων που εισέρχονται στα κύτταρα, δηλαδή ο ακριβής προσδιορισμός των αριθμητικών τιμών μάζας, όγκου ή σωματιδίων εντός των κυττάρων, αποτελεί μία ιδιαίτερα πολύπλοκη διαδικασία, κυρίως λόγω της δυναμικής φύσης του μικροπεριβάλλοντος των ιστών και των κυττάρων. Παράγοντες όπως η μεταβλητότητα του pH και η παρουσία πρωτεϊνών στον εξωκυττάριο χώρο επηρεάζουν σημαντικά την έκβαση της πρόσληψης νανοσωματιδίων. Κατά συνέπεια, η αποτελεσματικότητα και η αξιοπιστία της ποσοτικοποίησης εξαρτώνται άμεσα από την ακρίβεια και την ευαισθησία της μεθόδου μέτρησης που χρησιμοποιείται⁵.

Μία ευρέως χρησιμοποιούμενη κατηγορία μεθόδων για τη μελέτη της κυτταρικής πρόσληψης νανοσωματιδίων είναι η μικροσκοπία. Η μικροσκοπία φωτός, ιδιαίτερα όταν συνδυάζεται με φθορίζοντα νανοσωματίδια, επιτρέπει την παρακολούθηση της πρόσληψης σε ζωντανά κύτταρα, προσφέροντας τη δυνατότητα μη καταστροφικής και σχετικά ταχείας απεικόνισης. Αυτή η προσέγγιση είναι χρήσιμη για την ποιοτική εκτίμηση της χωρικής κατανομής των νανοσωματιδίων εντός των κυττάρων. Ωστόσο, η μέθοδος αυτή παρουσιάζει περιορισμούς ως προς την ποσοτική της ακρίβεια. Η πιθανότητα διαρροής (leakage) των φθορίζοντων χρωστικών παραγόντων και η φωτολεύκανση (photobleaching) καθιστούν τον ακριβή προσδιορισμό του αριθμού των νανοσωματιδίων αδύνατο. Έτσι, λαμβάνονται ημι-ποσοτικά αποτελέσματα ως προς ένα δείγμα ελέγχου⁶.

Αντιθέτως, οι τεχνικές ηλεκτρονικής μικροσκοπίας, όπως η μικροσκοπία σάρωσης (Scanning Electron Microscopy, SEM) και διαπερατότητας (Transmission electron microscopy, TEM), προσφέρουν εξαιρετικά υψηλή ευκρίνεια και δυνατότητα οπτικοποίησης της ενδοκυτταρικής θέσης των νανοσωματιδίων στη νανοκλίμακα με λεπτομέρεια. Οι μέθοδοι αυτές όμως, περιορίζονται σε νανοσωματίδια με υψηλή ατομική αντίθεση (atomic contrast) σε σχέση με το περιβάλλον του κυττάρου, ενώ απαιτούν εκτενέστερη προετοιμασία των δειγμάτων^{5,6}. Επιπλέον, η δυνατότητα επεξεργασίας μικρού αριθμού δειγμάτων αλλά και μικρών όγκων κυττάρων (low throughput) καθιστά τη χρήση τους περιοριστική για μελέτες μεγάλης κλίμακας^{6,7}.

Η κατηγορία μεθόδων που προσφέρει τη μεγαλύτερη ακρίβεια στην ποσοτικοποίηση νανοσωματιδίων, καθώς και τα χαμηλότερα όρια ανίχνευσης (σε επίπεδο ppt έως ppm), είναι η φασματομετρία μαζών (Mass Spectrometry, MS) -ιδίως όταν συνδυάζεται με τεχνικές όπως η επαγόμενη σύζευξη πλάσματος (Inductively Coupled Plasma Mass Spectrometry, ICP-MS)- αλλά και η φασματοσκοπία εκπομπής με επαγόμενη σύζευξη πλάσματος (Inductively Coupled Plasma Optical Emission Spectrometry, ICP-OES) ή η φασματοσκοπία ατομικής απορρόφησης με επαγόμενη σύζευξη πλάσματος (Inductively Coupled Plasma Atomic Absorption Spectrometry, ICP-AAS). Ωστόσο, οι μέθοδοι αυτές παρουσιάζουν σημαντικούς περιορισμούς. Αφενός, η εφαρμογή τους περιορίζεται αποκλειστικά σε νανοσωματίδια που περιέχουν ανιχνεύσιμα μέταλλα ή στοιχεία με υψηλή ατομική μάζα. Αφετέρου, η προετοιμασία των δειγμάτων, ιδίως όταν αυτά προέρχονται από βιολογικά συστήματα (όπως κύτταρα ή ιστοί), απαιτεί περίπλοκα πρωτόκολλα χώνευσης. Επιπλέον, πρόκειται για καταστροφικές μεθόδους που δεν επιτρέπουν την επαναξιολόγηση του ίδιου δείγματος, ενώ για την εξασφάλιση

της ακρίβειας απαιτείται σχολαστική βαθμονόμηση (calibration) με τη χρήση κατάλληλων στοιχειακών προτύπων. Τέλος, η αναγωγή των μετρήσεων σε αριθμό νανοσωματιδίων ανά κύτταρο απαιτεί την θεώρηση παραδοχών σχετικά με το μέγεθος και το σχήμα των νανοσωματιδίων^{6,7}.

Η εξέλιξη της ICP-MS σε ICP-MS ενός σωματιδίου (single particle ICP-MS) αποτελεί την πλέον ακριβή τεχνική ποσοτικοποίησης νανοσωματιδίων, καθώς επιτρέπει τον διαχωρισμό της σωματιδιακής από τη διαλυμένη μορφή του υλικού. Ωστόσο, η επίδραση της βιολογικής μήτρας στις μετρήσεις παραμένει υπό διερεύνηση, καθώς μπορεί να επηρεάσει τη σταθερότητα και την ανίχνευση των σωματιδίων⁵.

Είναι πλέον σαφές πως οι μέθοδοι ποσοτικοποίησης νανοσωματιδίων σε κύτταρα, παρά την τεχνολογική τους πρόοδο, παρουσιάζουν μειονεκτήματα που οδηγούν σε σημαντική αύξηση του κόστους των πειραμάτων. Πολλές από αυτές είναι καταστροφικές, απαιτώντας νέα δείγματα για κάθε μέτρηση, ενώ συχνά συνοδεύονται από χρονοβόρα προετοιμασία και ανάγκη βαθμονόμησης με πιστοποιημένα πρότυπα. Επιπλέον, η περιορισμένη δυνατότητα ανάλυσης μεγάλου αριθμού δειγμάτων ή διαφορετικών τύπων νανοσωματιδίων επιβάλλει την εφαρμογή πολλαπλών μετρήσεων ή συμπληρωματικών τεχνικών. Οι παράγοντες αυτοί βεβαίως αυξάνουν την κατανάλωση πόρων και επιμηκύνουν τη διάρκεια των πειραματικών σταδίων.

1.3. Μηχανική μάθηση και μείωση πειραματικού κόστους

Ο όρος τεχνητή νοημοσύνη (Artificial Intelligence, AI) χρησιμοποιήθηκε για πρώτη φορά το 1955 για να περιγράψει την «μεταφορά» ανθρωπομορφικής γνωστικής λειτουργίας σε μηχανές έτσι ώστε αυτές να «συμπεριφέρονται» με τρόπο που θα χαρακτηριζόταν «έξυπνος» αν ήταν άνθρωποι⁸. Σήμερα, θα μπορούσε να περιγραφεί πιο απλά ως η αυτοματοποίηση ορισμένων γνωστικών λειτουργιών⁹. Ο πιο ευρέως ίσως χρησιμοποιούμενος κλάδος της τεχνητής νοημοσύνης είναι η μηχανική μάθηση που το 1959 ορίστηκε από τον Arthur Samuel ως «ο κλάδος έρευνας που δίνει την δυνατότητα στους υπολογιστές να μαθαίνουν χωρίς να είναι ρητά προγραμματισμένοι»¹⁰.

Η μηχανική μάθηση περιλαμβάνει διάφορες προσεγγίσεις εκπαίδευσης αλγορίθμων, ανάλογα με τη διαθεσιμότητα και τη φύση των δεδομένων. Η επιβλεπόμενη μάθηση (supervised learning) βασίζεται σε ζεύγη εισόδων-εξόδων, με στόχο την εκμάθηση μιας συνάρτησης που προβλέπει σωστά την έξοδο για νέες εισόδους. Περιλαμβάνει κλασσικούς αλγορίθμους όπως η γραμμική παλινδρόμηση, αλλά και πιο περίπλοκες μεθοδολογίες που βασίζονται σε «δένδρα απόφασης» (π.χ. Random Forest, Extreme Gradient Boosting). Αντίθετα, η μη επιβλεπόμενη μάθηση (unsupervised learning) χρησιμοποιείται όταν δεν υπάρχουν ετικέτες στα δεδομένα, επιχειρώντας την ανάλυση και κατηγοριοποίηση βάσει της εσωτερικής δομής των εισόδων (π.χ. αλγόριθμος K-Means). Η ημι-επιβλεπόμενη μάθηση (semi-supervised learning) συνδυάζει επισημασμένα (labelled) και μη επισημασμένα (unlabelled) δεδομένα για τη δημιουργία πιο ακριβών μοντέλων παλινδρόμησης ή κατηγοριοποίησης. Τέλος, η ενισχυτική μάθηση (reinforcement learning) αφορά την εκμάθηση μέσω αλληλεπίδρασης και ανατροφοδότησης από το περιβάλλον, με τον αλγόριθμο να επιλέγει ενέργειες που μεγιστοποιούν την επιβράβευση¹¹.

Η πρόβλεψη της κυτταρικής πρόσληψης ναοσωματιδίων μπορεί να αντιμετωπιστεί ως ένα πρόβλημα επιβλεπόμενης μηχανικής μάθησης καθώς υπάρχει μεγάλος όγκος πειραματικών δεδομένων με γνωστές τιμές της μεταβλητής εξόδου, δηλαδή της κυτταρικής πρόσληψης, ώστε να εκπαιδευτούν κατάλληλα μοντέλα. Αυτή η προσέγγιση βεβαίως στοχεύει στην αντιμετώπιση των περιορισμών του προσδιορισμού της κυτταρικής πρόσληψης ναοσωματιδίων μέσω κλασικών πειραματικών μεθόδων.

Πιο συγκεκριμένα, έχει αποδειχθεί πως η μοντελοποίηση μέσω μηχανικής μάθησης επιτρέπει την πρόβλεψη της αποδοτικότητας κυτταρικής πρόσληψης ναοσωματιδίων, όπως οι διαπερατές πεπτιδικές αλυσίδες, πριν από οποιοδήποτε πειραματικό στάδιο, μειώνοντας σημαντικά την ανάγκη για εκτεταμένη πειραματική επικύρωση, εξοικονομώντας χρόνο και πόρους¹². Σε μία σχετική μελέτη, η εφαρμογή προσομοιώσεων μέσω μηχανικής μάθησης κατάφερε έως και 7.5 φορές ταχύτερη πρόβλεψη κρίσιμων χαρακτηριστικών των ναοσωματιδίων (όπως η προσβάσιμη επιφάνεια στον διαλύτη) σε σχέση με παραδοσιακές μεθόδους προσομοίωσης¹³. Ακόμη, τα μοντέλα μηχανικής μάθησης μπορούν να λειτουργήσουν ως εργαλεία αποδοτικής αξιολόγησης της κυτταρικής πρόσληψης μεγάλου όγκου πιθανών τύπων ναοσωματιδίων, δίνοντας την δυνατότητα επιλογής μόνο λίγων, πολλά υποσχόμενων διαμορφώσεων ναοσωματιδίων για τον πειραματικό έλεγχο με βελτιστοποιημένες πειραματικές παραμέτρους (π.χ. κυτταρική σειρά, θερμοκρασία)^{14,15}.

1.4. Προγενέστερες έρευνες

Ο όγκος και η διαθεσιμότητα δεδομένων στην εποχή των «Μεγάλων Δεδομένων» (big data) έχουν αποδειχθεί καθοριστικοί παράγοντες για την στροφή προς την μοντελοποίηση με στόχο την μείωση του πειραματικού κόστους σε διάφορους ερευνητικούς τομείς. Δεδομένου του κόστους και του χρόνου που απαιτείται για τη διεξαγωγή πειραμάτων προσδιορισμού της κυτταρικής πρόσληψης ναοσωματιδίων, είναι αναγκαία πλέον η γρήγορη και αποδοτική αξιολόγηση δεκάδων ή και χιλιάδων διαφορετικών συνδυασμών φυσικοχημικών χαρακτηριστικών ναοσωματιδίων και πειραματικών παραμέτρων, ώστε να προταθούν μόνο οι πιο υποσχόμενοι πειραματικοί σχεδιασμοί για τον εργαστηριακό έλεγχο.

Για τον σκοπό αυτό, έχουν δημιουργηθεί διάφορα μοντέλα ποσοτικής σχέσης δομής-δραστικότητας (Quantitative Structure-Activity Relationship, QSAR) που προβλέπουν την κυτταρική πρόσληψη ναοσωματιδίων μέσω της παραμετροποίησης διαφορικών εξισώσεων που περιγράφουν τον μηχανισμό αλληλεπίδρασης ναοσωματιδίων-κυττάρων. Πιο συγκεκριμένα, οι Lu *et al.*¹⁶, δημιούργησαν ένα QSAR μοντέλο αξιοποιώντας δεδομένα επιφανειακής ενέργειας και επιφανειακού φορτίου ενός συνόλου 94 ναοσωματιδίων σιδήρου με οργανική επικάλυψη διαφορετικού φορτίου και 5 κυτταρικών σειρών. Επίσης, μοντέλα QSAR έχουν αποδειχθεί χρήσιμα στην πρόβλεψη της κυτταρικής πρόσληψης φθορίζοντων ναοσωματιδίων σε 2 παγκρεατικές καρκινικές σειρές με υψηλή ακρίβεια (R^2 έως 0.885 στο σύνολο ελέγχου). Τα μοντέλα αυτά αξιοποιούν περιγραφείς (descriptors) που υπολογίζονται βάση των δομών SMILES και περιγράφουν τα ατομιστικά χαρακτηριστικά της επικάλυψης των ναοσωματιδίων¹⁷.

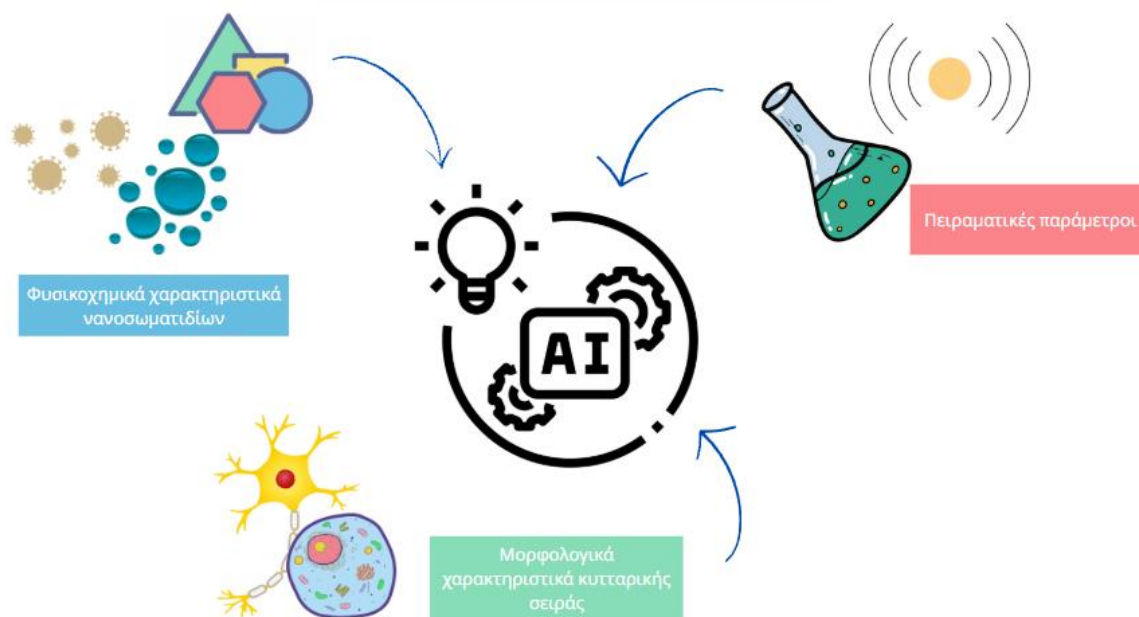
Παράλληλα, έχουν αναπτυχθεί μοντέλα πρόβλεψης της κυτταρικής πρόσληψης νανοϋλικών που βασίζονται σε δεδομένα (data-driven models) μέσω μεθοδολογιών μηχανικής μάθησης. Ένα σύνολο δεδομένων χορήγησης νανοσωματιδίων σε ποντίκια που φέρουν καρκινικούς όγκους κατασκευάστηκε αρχικά για την δημιουργία ενός φαρμακοκινητικού μοντέλου που βασίζεται στη φυσιολογία (Physiologically Based Kinetic, PBK)¹⁸ και προβλέπει την αποδοτικότητα μεταφοράς νανοσωματιδίων στους όγκους βάσει φυσικοχημικών ιδιοτήτων. Το ίδιο σύνολο δεδομένων χρησιμοποιήθηκε και σε συνδυασμό με μεθόδους μηχανικής μάθησης. Οι Chou et al. εκπαίδευσαν βαθιά νευρωνικά δίκτυα για την παραμετροποίηση ενός νέου PBK μοντέλου με πολύ ικανοποιητικά αποτελέσματα¹⁹, ενώ άλλοι ερευνητές αξιοποίησαν το σύνολο δεδομένων για την εκπαίδευση μοντέλων πρόβλεψης που βασίζονται μόνο σε δεδομένα μέσω αλγορίθμων μηχανικής μάθησης (XGBoost, Random Forest κ.α.)²⁰. Ακόμη, η χρήση νευρωνικών δικτύων για την πρόβλεψη της κυτταρικής πρόσληψης φθορίζοντων νανοσωματιδίων σε παγκρεατικές καρκινικές σειρές έχει αποδειχθεί ότι δίνει πολύ ικανοποιητικά στατιστικά μέτρα αξιολόγησης των μοντέλων (R^2 έως 0.934)²¹. Τέλος, μοντέλα μηχανικής μάθησης, όπως το XGBoost, έχουν χρησιμοποιηθεί για την επεξήγηση των σχέσεων σημαντικών ιδιοτήτων νανοσωματιδίων χρυσού (μέγεθος, Z-δυναμικό) με την ικανότητα κυτταρικής τους πρόσληψης²².

Εξαιρετικά ενδιαφέροντα είναι η χρήση μηχανικής μάθησης για την πρόβλεψη της κατηγορίας καρκινικών κυττάρων του μαστού (στάδιο καρκίνου) μέσω μελέτης της διαφορετικής πρόσληψης νανοσωματιδίων άνθρακα σε καρκινικά κύτταρα που βρίσκονται σε διαφορετικά στάδια ανάπτυξης της ασθένειας²³. Τέλος, φωτογραφίες μικροσκοπίου καρκινικών κυττάρων του μαστού που έχουν προσλάβει φθορίζοντα νανοσωματίδια με αντικαρκινικό παράγοντα έχουν χρησιμοποιηθεί για την πρόβλεψη μέσω βαθιάς μηχανικής μάθησης της υψηλής ή χαμηλής κυτταρικής πρόσληψης της φαρμακευτικής αγωγής από τα κύτταρα²⁴.

Παρά την πρόοδο στη χρήση της μηχανικής μάθησης για την πρόβλεψη της κυτταρικής πρόσληψης νανοσωματιδίων, οι έως τώρα μελέτες περιορίζονται συνήθως σε ένα μόνο είδος νανοϋλικών (όπως νανοσωματίδια χρυσού ή σιδήρου) ή σε περιορισμένο αριθμό κυτταρικών σειρών. Συνεπώς, υπάρχει ένα σημαντικό κενό στη διεθνή βιβλιογραφία ως προς την ανάπτυξη μοντέλων μηχανικής μάθησης που να μπορούν να γενικεύσουν τις προβλέψεις τους βασιζόμενα σε φυσικοχημικές ιδιότητες διαφόρων νανοϋλικών, διαφορετικές πειραματικές παραμέτρους και βιολογικά χαρακτηριστικά πολλαπλών κυτταρικών σειρών.

1.5. Δομή της διπλωματικής εργασίας

Στόχος της παρούσας ερευνητικής εργασίας είναι η ενδελεχής μελέτη των βιβλιογραφικών πηγών σχετικά με την πρόσληψη νανοσωματιδίων διαφορετικών τύπων σε διάφορες κυτταρικές σειρές και η δημιουργία ενός εκτενούς συνόλου δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση ενός απλού αλλά ερμηνεύσιμου μοντέλου μηχανικής μάθησης. Ως έξοδος του μοντέλου θα λαμβάνεται η πρόβλεψη της κυτταρικής πρόσληψης των νανοσωματιδίων ως Νανοσωματίδια/κύτταρο (NPs/cell) βάση φυσικοχημικών χαρακτηριστικών των νανοϋλικών, πειραματικών παραμέτρων και βιολογικών χαρακτηριστικών των κυτταρικών σειρών (Σχήμα 1).



Σχήμα 1. Συμμετοχή στις προβλέψεις του μοντέλου μηχανικής μάθησης μεταβλητών που κωδικοποιούν τις φυσικοχημικές ιδιότητες των νανοσωματιδίων, τις πειραματικές παραμέτρους και τα μορφολογικά χαρακτηριστικά της κυτταρικής σειράς

Για τον σκοπό αυτό πραγματοποιήθηκε αρχικά μία βιβλιογραφική μελέτη των παραμέτρων και μηχανισμών που επηρεάζουν την πρόσληψη νανοσωματιδίων από βιολογικά συστήματα και συγκεκριμένα κύτταρα θηλαστικών, αλλά και των μεθοδολογιών μηχανικής μάθησης που αξιοποιούνται για την επίλυση περίπλοκων προβλημάτων. Έτσι, στο δεύτερο κεφάλαιο της παρούσας εργασίας συνοψίζονται βασικές παράμετροι της κυτταρικής πρόσληψης νανοσωματιδίων, όπως το μέγεθος, το σχήμα, η επιφανειακή χημεία και το επιφανειακό φορτίο τους, αλλά και η σχέση αυτών με την αύξηση ή μείωση της αποδοτικότητας της ενδοκυττάρωσης των νανοσωματιδίων. Σημειώνεται, πως στις περισσότερες περιπτώσεις τα «μοτίβα» και η ανάλυση αυτών των σχέσεων βασίζεται σε πειραματικές παρατηρήσεις και δεν μπορεί να γενικευτεί για όλα τα είδη νανοσωματιδίων ή κυττάρων. Στη συνέχεια, αναλύονται οι μηχανισμοί που αξιοποιούνται για ενδοκυττάρωση νανοσωματιδίων, δηλαδή η φαγοκυττάρωση, η μακροπινοκύττωση, η ενδοκυττάρωση εξαρτώμενη από κλαθρίνη ή καβεολίνη και άλλοι μηχανισμοί ανεξάρτητοι της κλαθρίνης/καβεολίνης.

Το τρίτο κεφάλαιο της εργασίας αναφέρεται σε μεθοδολογίες και πρακτικές που ακολουθούνται για την δημιουργία και αξιολόγηση μοντέλων μηχανικής μάθησης. Ξεκινώντας από πρακτικές διαχωρισμού των δεδομένων σε σύνολο εκπαίδευσης και ελέγχου, κωδικοποίησης των κατηγορικών μεταβλητών ως αριθμητικές και κανονικοποίησης των δεδομένων εισόδου, η θεωρητική αυτή ανάλυση καταλήγει στην περιγραφή του αλγόριθμου XGBoost που χρησιμοποιήθηκε για την δημιουργία του μοντέλου. Στη συνέχεια αναλύονται οι βασικές μέθοδοι αξιολόγησης των μοντέλων μηχανικής μάθησης και εύρεσης του πεδίου εφαρμοσιμότητας (applicability domain) (ή πεδίου εμπιστοσύνης) αυτών.

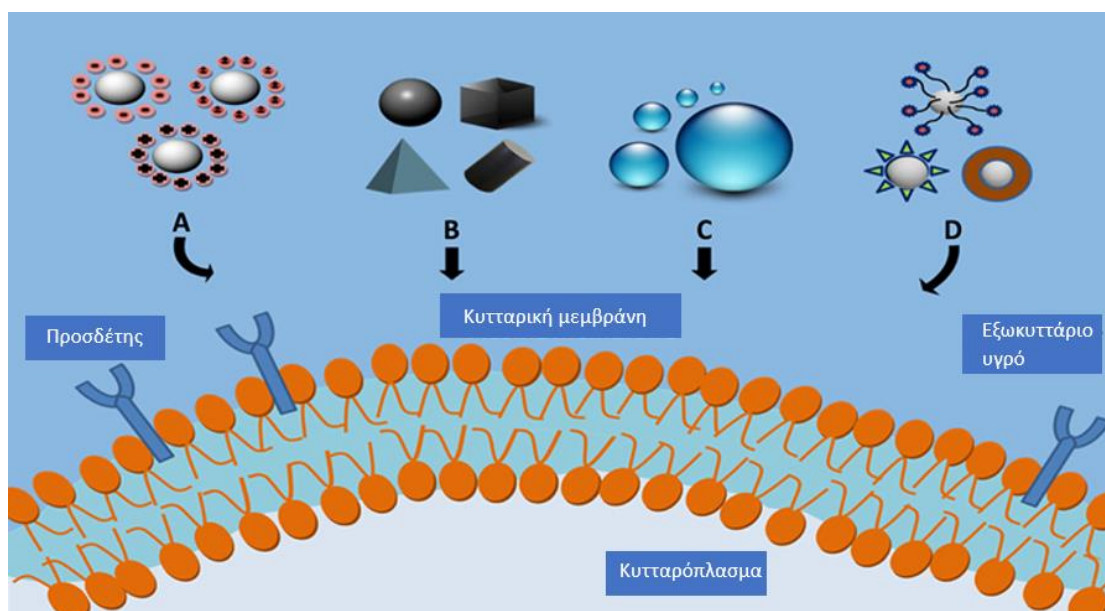
Ακολούθως, αναλύεται η μεθοδολογία που ακολουθήθηκε για την επίτευξη του σκοπού της συγκεκριμένης έρευνας και την εξαγωγή των αποτελεσμάτων. Έτσι, στο τέταρτο κεφάλαιο παρατίθενται τα βήματα προεπεξεργασίας των πρωτογενών δεδομένων, από την συλλογή και αξιολόγηση των πειραματικών ερευνών που είναι διαθέσιμες στην σύγχρονη βιβλιογραφία μέχρι την μετατροπή των μονάδων μέτρησης των μεταβλητών που εφαρμόστηκε για να εξασφαλισθεί η συνέπεια και η συνεκτικότητα του συνόλου δεδομένων. Επίσης, παρουσιάζεται η αντιμετώπιση των ελλিপών τιμών μέσω παλινδρόμησης. Στη συνέχεια, το πέμπτο κεφάλαιο αναλύει τα βήματα προεπεξεργασίας δεδομένων, βελτιστοποίησης, απλοποίησης και αξιολόγησης του μοντέλου XGBoost που ακολουθήθηκαν σε αυτή την έρευνα.

Τέλος, στο έκτο κεφάλαιο παρουσιάζονται και σχολιάζονται τα αποτελέσματα της μοντελοποίησης των δεδομένων αλλά και η αντιμετώπιση των σημαντικότερων προβλημάτων που προέκυψαν σε αυτή τη διαδικασία. Ακολουθεί μία αναλυτική συζήτηση της ποιότητας και του είδους των δεδομένων που συμμετέχουν στην εκπαίδευση του μοντέλου μηχανικής μάθησης. Επιπλέον, αναλύονται και αξιολογούνται τα αποτελέσματα της βελτιστοποίησης των παραμέτρων του μοντέλου και της απλοποίησης αυτού μέσω αφαίρεσης μεταβλητών. Τελικά, προσδιορίζεται το πεδίο εφαρμοσιμότητας του βέλτιστου μοντέλου και ακολουθεί η εκτενής ερμηνεία των σχέσεων μεταξύ των μεταβλητών εισόδου και της κυτταρικής πρόσληψης των νανοϋλικών μέσω της ανάλυσης SHAP (SHapley Additive exPlanations). Η διπλωματική εργασία ολοκληρώνεται με τα συμπεράσματα και τις παρατηρήσεις που προέκυψαν καθ' όλα τα στάδια της έρευνας και της ανάλυσης των αποτελεσμάτων.

2. Βασικές παράμετροι της κυτταρικής πρόσληψης νανοσωματιδίων

Η πρόσληψη νανοσωματιδίων από κύτταρα μπορεί να επηρεάζεται τόσο από τα φυσικοχημικά χαρακτηριστικά των νανοϋλικών όσο και από το μικροπεριβάλλον ή το είδος του κυττάρου²⁵ (Εικόνα 1). Παρά την μεγάλη πρόοδο στην έρευνα για την αλληλεπίδραση διαφόρων ειδών νανοσωματιδίων με κύτταρα θηλαστικών, ακόμη δεν έχουν εξαχθεί συνολικά συμπεράσματα για τα φαινόμενα που λαμβάνουν χώρα κατά την πρόσληψη των νανοσωματιδίων στα κύτταρα και τους παράγοντες από τους οποίους αυτή εξαρτάται. Στο παρόν κεφάλαιο αναλύονται ορισμένα «μοτίβα» που φαίνεται να υπάρχουν σε αρκετές διαφορετικές ερευνητικές μελέτες χωρίς όμως να μπορούν να γενικευτούν για όλα τα είδη των νανοσωματιδίων και κυττάρων.

2.1. Φυσικοχημικές ιδιότητες νανοσωματιδίων



Εικόνα 1. Παράγοντες που επηρεάζουν την κυτταρική πρόσληψη των νανοσωματιδίων. Α) Επιφανειακό φορτίο Β) Σχήμα C) Μέγεθος και D) Επιφανειακή χημεία των νανοσωματιδίων (τροποποιημένο από Foroozandeh et al. (2018)²⁶).

2.1.1. Μέγεθος νανοσωματιδίων

Μία από τις πιο σημαντικές φυσικές ιδιότητες των νανοσωματιδίων που επηρεάζουν την πρόσληψή τους από τα κύτταρα είναι το μέγεθος τους. Το πολύ μικρό μέγεθος των νανοσωματιδίων συχνά τα καθιστά μη αντιληπτά ως ξένα σώματα για τα κύτταρα²⁷. Παρ' όλα αυτά, για να αλληλεπιδράσουν ουσιαστικά με την κυτταρική μεμβράνη, τα νανοσωματίδια πρέπει πρώτα να διαπεράσουν την εξωκυττάρια μήτρα (extracellular matrix), ένα πλέγμα που επιτρέπει να διαπερνούν μόνο νανοσωματίδια μικρότερα του μεγέθους των πόρων του. Εφόσον συμβεί αυτό, τα νανοσωματίδια μπορούν σχετικά

εύκολα να διέλθουν από την κυτταρική μεμβράνη λόγω της μεγάλης ειδικής τους επιφάνειας σε σχέση με τον όγκο τους²⁸.

Σύμφωνα με ορισμένες έρευνες υπάρχει μία βέλτιστη διάμετρος σφαιρικών νανοσωματιδίων, τα 50 nm, η οποία επιτρέπει την μέγιστη πρόσληψή τους στα κύτταρα²⁵⁻³⁰, με την παρατήρηση αυτή να είναι συνεπής για διαφορετικές κυτταρικές σειρές (BEAS-2B, STO, HeLa και SNB19) που καλλιεργούνται παρουσία νανοσωματιδίων χρυσού²⁹. Η βέλτιστη αυτή διάμετρος οφείλεται στην ισορροπία μεταξύ της διασύνδεσης υποδοχέων της μεμβράνης με τα νανοσωματίδια και της διαδικασίας περιτύλιξης της μεμβράνης γύρω από αυτά. Οι Gao et al.³¹ πρότειναν μία θεωρία που προσδιορίζει το βέλτιστο μέγεθος των νανοσωματιδίων ως το αποτέλεσμα του ανταγωνισμού της κινητικής της διάχυσης μέσω των υποδοχέων της κυτταρικής μεμβράνης και θερμοδυναμικών δυνάμεων. Μικρότερα από το ιδανικό μέγεθος σωματίδια πρέπει να δημιουργήσουν συσσωματώματα ώστε να αποκτήσουν επαρκή κινητήριο δύναμη εισαγωγής στο κύτταρο, αφού η αυξημένη ελαστική ενέργεια μειώνει την ικανότητα περιτύλιξης της κυτταρικής μεμβράνης γύρω από μεμονωμένα μικρά σωματίδια²⁸.

Όπως προαναφέρθηκε όμως, αν και οι παρατηρήσεις μεμονωμένων ερευνών συχνά δείχνουν ότι υπάρχει μία τάση στην επίδραση του μεγέθους των νανοσωματιδίων στην κυτταρική τους πρόσληψη, δεν είναι δυνατόν να προκύψει κάποια γενίκευση αυτών των θεωριών. Οι ασυνέπειες αυτές οφείλονται πιθανώς στην πολυπλοκότητα των διαφορετικών παραμέτρων που ελέγχουν αυτά τα φαινόμενα αλλά και στις αλλαγές του μεγέθους των νανοσωματιδίων λόγω συσσωμάτωσης που μπορεί να διαφέρουν μεταξύ των *in vitro* και *in vivo* συνθηκών²⁶. Έτσι, έχει παρατηρηθεί ότι για καρκινικά κύτταρα η κυτταρική πρόσληψη νανοσωματιδίων είναι αντιστρόφως ανάλογη του μεγέθους τους χωρίς την ύπαρξη βέλτιστης τιμής διαμέτρου²⁷. Παράλληλα, οι Yue et al.³² έδειξαν ότι το βέλτιστο μέγεθος των νανοσωματιδίων μπορεί να εξαρτάται από τις μονάδες μέτρησης της εσωτερίκευσης των νανοϋλικών στο κύτταρο (NPs/cell ή Volume/cell) καθιστώντας την εκτίμηση του ανακριβή²⁵.

2.1.2. Σχήμα νανοσωματιδίων

Οι πρώτες έρευνες για την χρήση νανοϋλικών ως πιθανή αντικαρκινική θεραπεία έδιναν βάση στην βελτιστοποίηση του μεγέθους και των επιφανειακών ιδιοτήτων των νανοσωματιδίων αλλά όχι του σχήματος, καθώς οι πρώτες ιδιότητες θεωρούνταν πιο κομβικές για την επιτυχία των πειραμάτων *in vivo*²⁹. Έτσι, χρησιμοποιούνταν κατά κόρον σφαιρικά νανοσωματίδια. Πράγματι οι Chithrani et al.³³ έχουν δείξει ότι σφαιρικά νανοσωματίδια χρυσού προσλαμβάνονται σε κύτταρα θηλαστικών έως και 5 φορές περισσότερο από αντίστοιχες νανοράβδους, με την πιο αποδοτική πρόσληψη σφαιρικών νανοσωματιδίων να επιβεβαιώνεται για 3 διαφορετικές κυτταρικές σειρές (STO, HeLa, SNB19)²⁶.

Μετέπειτα έρευνες όμως παρουσιάζουν αντίθετα αποτελέσματα, υποστηρίζοντας ότι τα επιμήκη νανοσωματίδια προσλαμβάνονται πιο αποδοτικά στα κύτταρα λόγω της μεγαλύτερης ικανότητάς τους να προσροφώνται στις κυτταρικές μεμβράνες. Αυτή η παρατήρηση βασίζεται πιθανώς στο γεγονός ότι τα σφαιρικά σωματίδια, λόγω της

καμπυλωτής τους επιφάνειας, προσφέρουν λιγότερα σημεία πρόσδεσης στην κυτταρική μεμβράνη²⁸. Κυλινδρικά νανοσωματίδια φαίνεται να προσλαμβάνονται πιο εύκολα από την κυτταρική σειρά HeLa σε σχέση με νανοσφαίρες και νανοκύβους³⁴, ενώ σε άλλη μελέτη φάνηκε ότι διαφορετικοί τύποι κυττάρων (επιθηλιακά και ενδοθηλιακά) προτιμούν να προσλαμβάνουν σωματίδια σε σχήμα νανοδίσκου σε σχέση με ραβδόμορφα σωματίδια³⁵.

Ακόμα όμως και μεταξύ νανοσωματιδίων ίδιας γεωμετρίας υπάρχουν αντιφάσεις μεταξύ ερευνών κυτταρικής πρόσληψης. Για παράδειγμα, σε μία σχετική μελέτη, η αύξηση του λόγου διαστάσεων των νανοράβδων πυριτίου οδήγησε σε αύξηση της κυτταρικής πρόσληψης³⁶. Ωστόσο, η αύξηση αυτή φαίνεται να έχει αντίθετα αποτελέσματα στην πρόσληψη νανοράβδων χρυσού, αφού σε αυτή την περίπτωση η μεγαλύτερη επιφάνεια επαφής των νανοσωματιδίων με την κυτταρική μεμβράνη μπορεί να οδηγήσει σε μείωση των πιθανών σημείων πρόσδεσης³³.

Τα αντικρουόμενα συμπεράσματα σχετικά με την επίδραση του σχήματος στην κυτταρική πρόσληψη νανοσωματιδίων οφείλονται σε περίπλοκες αλληλεπιδράσεις με το μικροπεριβάλλον που δεν έχουν μελετηθεί πλήρως, καθώς και στην δυσκολία κατασκευής των διαφορετικών σχημάτων στην νανοκλίμακα με ακρίβεια. Προσομοιώσεις που πραγματοποιήθηκαν για θεωρητικά νανοσωματίδια σχήματος σφαίρας, μικρής και μεγάλης ράβδου και δίσκου σταθερού όγκου και επιφανειακής πυκνότητας προσδετών έδειξαν ότι η κυτταρική πρόσληψη πιθανότατα πραγματοποιείται σε δύο στάδια για όλα τα νανοσωματίδια. Αρχικά, τα νανοσωματίδια προσανατολίζονται προς την μεμβράνη με τέτοιο τρόπο ώστε η επιφάνεια επαφής και πρόσδεσης να μεγιστοποιείται, με αποτέλεσμα η μεμβράνη να τα περικλείει. Στο δεύτερο στάδιο η μεμβράνη τυλίγεται πλήρως γύρω από τα νανοσωματίδια και τα ενσωματώνει. Η επιτυχής ενδοκυττάρωση προϋποθέτει την απελευθέρωση επαρκούς ελεύθερης ενέργειας από την σύνδεση συνδετών-υποδοχέων για την κάλυψη της ενέργειας που απαιτείται για την περιτύλιξη της μεμβράνης³⁷.

2.1.3. Επιφανειακό φορτίο νανοσωματιδίων

Η λιπιδική διπλοστιβάδα που δομεί την κυτταρική μεμβράνη έχει ένα εγγενές αρνητικό φορτίο λόγω των φωσφορικών ομάδων των φωσφολιπιδίων, γεγονός που καθιστά καθοριστικές τις ηλεκτροστατικές αλληλεπιδράσεις μεταξύ μεμβράνης και νανοσωματιδίων για την εσωτερίκευσή τους^{28,29}. Έχει παρατηρηθεί ότι θετικά φορτισμένα νανοσωματίδια προσλαμβάνονται πιο εύκολα από τα κύτταρα σε σχέση με τα αντίστοιχα ανιονικά^{25,26,29,37}. Ωστόσο, λόγω του φορτίου τους, είναι πιθανό να διαταράζουν την δομή της μεμβράνης, οδηγώντας σε αύξηση της κυτταροτοξικότητας²⁶. Όσον αφορά τα αρνητικά φορτισμένα νανοσωματίδια, υπάρχουν αντικρουόμενα στοιχεία σχετικά με την επίδραση της αύξησης του φορτίου τους στην αποδοτικότητα της κυτταρικής πρόσληψης²⁵.

Σύμφωνα με αρκετές έρευνες, τα επιφανειακά φορτισμένα νανοσωματίδια, ανεξαρτήτως φορτίου, έχουν μεγαλύτερη ικανότητα εσωτερίκευσης σε κύτταρα από τα ουδέτερα ή λιγότερα φορτισμένα σωματίδια^{25,26,28}. Αυτό το φαινόμενο μπορεί να εξηγηθεί αφενός λόγω της μικρότερης συγγένειας των μη φορτισμένων σωματιδίων με

την αρνητικά φορτισμένη κυτταρική μεμβράνη, καθώς επίσης και μέσω της ουδέτερης επιφάνειάς τους, η οποία εμποδίζει την προσρόφηση πρωτεϊνών και ίσως οδηγεί στην δημιουργία μία υδατικής στοιβάδας μέσω ηλεκτροστατικών αλληλεπιδράσεων³⁸. Ακόμη, η αύξηση του επιφανειακού φορτίου αποτρέπει την δημιουργία συσσωματωμάτων λόγω των αποστικών δυνάμεων μεταξύ των νανοσωματιδίων³⁹, με αποτέλεσμα την πιο εύκολη διέλευση των μικρότερων σωματιδίων μέσω της κυτταρικής μεμβράνης²⁵.

Σημαντικό είναι να τονιστεί ότι το αρχικό φορτίο των νανοσωματιδίων δεν είναι πάντα ενδεικτικός παράγοντας για την πρόβλεψη της κυτταρική πρόσληψής τους, καθώς η αλληλεπίδραση των νανοσωματιδίων με το εξωκυττάριο υγρό έχει ως αποτέλεσμα τον σχηματισμό μία πρωτεϊνικής «κορόνας» στην επιφάνειά τους που είναι δυνατόν να μεταβάλλει διάφορες φυσικοχημικές ιδιότητες, συμπεριλαμβανομένου του επιφανειακού φορτίου^{28,29}. Σε ορισμένες περιπτώσεις έχει παρατηρηθεί ότι τόσο τα κατιονικά όσο και τα ανιονικά νανοσωματίδια αποκτούν αρνητικό φορτίο παρόμοιο με αυτό των πρωτεϊνών του μέσου καλλιέργειας⁴⁰.

2.1.4. Επιφανειακή χημεία νανοσωματιδίων

Η επιφανειακή τροποποίηση των νανοσωματιδίων μέσω πρόσδεσης μορίων με διάφορες λειτουργικές ομάδες έχει ως στόχο την σταθεροποίησή τους ή/και την στόχευση συγκεκριμένων κυττάρων και τελικά την αύξηση της κυτταρικής πρόσληψης^{25,26,28}. Για παράδειγμα, βιολογικά μόρια όπως πρωτεΐνες, πεπτίδια και αντισώματα χρησιμοποιούνται για την στόχευση των ναούλικών σε συγκεκριμένα κύτταρα ή κυτταρικά οργανίδια⁴¹. Επίσης, η πρόσδεση υδρόφιλων μορίων πολυαιθυλενογλυκόλης (PEG) στην επιφάνεια σωματιδίων οδηγεί σε αυξημένη σταθερότητα και μικρότερη συσσωμάτωση²⁸, άρα πιο αποδοτική εσωτερίκευση σε κύτταρα σε ορισμένες περιπτώσεις. Αντίθετα, η επιφανειακή τροποποίηση των νανοσωματιδίων με PEG έχει φανεί ότι δυσχεραίνει την δημιουργία της πρωτεϊνικής «κορόνας» λόγω της αυξημένης υδροφιλικότητας. Άμεση συνέπεια αυτού είναι η παρεμπόδιση της αλληλεπίδρασης των ναούλικών με τις κυτταρικές μεμβράνες, η αύξηση του χρόνου κυκλοφορίας στο αίμα και η μείωση της κυτταρικής τους πρόσληψης³⁰.

Ακόμη, συχνή είναι η χρήση μορίων με καρβοξυλικές (αρνητικό φορτίο), υδροξυλικές (ουδέτερο φορτίο) ή αμινικές (θετικό φορτίο) λειτουργικές ομάδες. Και οι 3 αυτές περιπτώσεις φαίνεται συνήθως να έχουν ως αποτέλεσμα την αύξηση της κυτταρικής πρόσληψης, λόγω είτε της αύξησης του επιφανειακού φορτίου με την προσθήκη φορτισμένων ομάδων²⁶ ή της αύξησης της προσρόφησης πρωτεϊνών του ορού από τα ουδέτερα μόρια²⁵. Βεβαίως, λόγω της πολυπλοκότητας των συστημάτων που μελετώνται, υπάρχουν εξαιρέσεις στην αύξηση της κυτταρικής πρόσληψης μέσω επιφανειακών τροποποιήσεων, όπως στην περίπτωση πρόσδεσης καρβοξυλικών λειτουργικών ομάδων στην επιφάνεια δενδριμερών⁴².

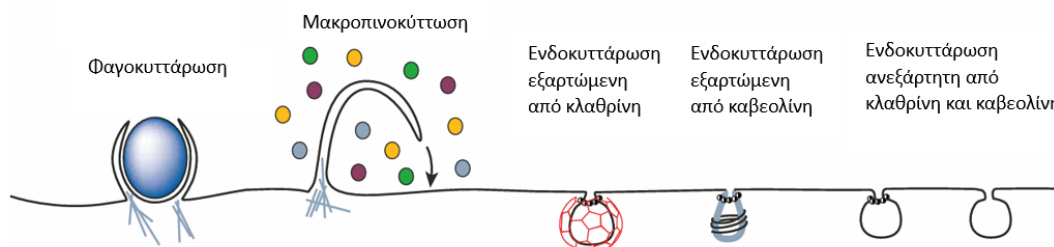
2.2. Πειραματικές παράμετροι που επηρεάζουν την συσσωμάτωση

Η χρήση υπερήχων για την σταθεροποίηση των νανοσωματιδίων είναι μία ευρέως χρησιμοποιούμενη τεχνική⁴³⁻⁴⁵ που έχει ως αποτέλεσμα την διάσπαση των συσσωματωμάτων. Γενικά έχει παρατηρηθεί ότι ως ένα σημείο η αύξηση του χρόνου εφαρμογής των υπερήχων οδηγεί σε μικρότερα συσσωματώματα και άρα πιο αποδοτική κυτταρική πρόσληψη⁴⁶, ενώ και η αύξηση της ισχύς των υπερήχων έχει ως αποτέλεσμα την μείωση της υδροδυναμικής διαμέτρου των νανοσωματιδίων⁴⁷. Παρ' όλα αυτά, η συνεχής αύξηση του χρόνου εφαρμογής υπερήχων μπορεί να οδηγήσει σε επανα-συσσωμάτωση των νανοσωματιδίων⁴⁸.

Επιπροσθέτως, η συσσωμάτωση των νανοσωματιδίων -και συνεπώς η δυνατότητα εσωτερικεύσης στα κύτταρα- επηρεάζεται από την ιοντική ισχύ του μέσου καλλιέργειας. Η υψηλή ιοντική ισχύς έχει ως αποτέλεσμα τον περιορισμό της ηλεκτρικής διπλοστοιβάδας που περιβάλλει κάθε νανοσωματίδιο εμβαπτισμένο σε υγρό μέσο και έτσι το ενεργειακό φράγμα της ηλεκτροστατικής απόθησης δύο ομοίως φορτισμένων σωματιδίων μειώνεται. Συνέπεια αυτού του φαινομένου είναι η εντατικοποίηση της συσσωμάτωσης και η αύξηση της υδροδυναμικής διαμέτρου των νανοσωματιδίων, γεγονός που όπως προαναφέρθηκε, επηρεάζει αρνητικά την ικανότητα κυτταρικής πρόσληψης²⁵.

2.3. Μηχανισμοί κυτταρικής πρόσληψης

Η ενδοκυττάρωση είναι μία διαδικασία ενεργούς μεταφοράς συστατικών του εξοκυττάριου χώρου εντός των κυττάρων μέσω δημιουργίας κυστιδίων κυτταρικής μεμβράνης που τα περικλείουν. Οι γνωστοί μηχανισμοί ενεργού μεταφοράς νανοσωματιδίων είναι κυρίως η φαγοκυττάρωση, η μακροπινοκύττωση, η ενδοκυττάρωση εξαρτώμενη από κλαθρίνη, η ενδοκυττάρωση εξαρτώμενη από καβεολίνη και ειδικοί μηχανισμοί ενδοκυττάρωσης ανεξάρτητης από κλαθρίνη και καβεολίνη⁴⁹ (Εικόνα 2).



Εικόνα 2. Σχηματική αναπαράσταση των διαφορετικών μηχανισμών κυτταρικής πρόσληψης (τροποποιημένο από Conner and Schmid (2003)⁵⁰).

2.3.1. Φαγοκυττάρωση

Η φαγοκυττάρωση πραγματοποιείται από εξειδικευμένα κύτταρα, όπως μακροφάγα ή ουδετερόφιλα, με στόχο την απομάκρυνση παθογόνων μικροοργανισμών (βακτήρια, ζύμες) και υπολειμμάτων νεκρών κυττάρων^{50,51}. Βεβαίως, και άλλα είδη κυττάρων όπως επιθηλιακά, ινοβλάστες και ενδοθηλιακά χρησιμοποιούν αυτόν τον μηχανισμό σε μικρότερο βαθμό⁵¹. Για την φαγοκυττάρωση νανοσωματιδίων συνήθως απαιτείται η πρόσδεση οψωνινών (opsonins) στην επιφάνειά τους για την αναγνώρισή τους από τα φαγοκύτταρα και την πρόσδεση σε ειδικούς υποδοχείς της κυτταρικής μεμβράνης. Στη συνέχεια, ένας «καταρράκτης» σημάτων (signaling cascade) ενεργοποιεί τους μηχανισμούς επέκτασης της κυτταρικής μεμβράνης και περιτύλιξης του νανοσωματιδίου που εσωτερικεύεται σε ένα «φαγόσωμα» (phagosome)^{50,51}.

Αυτό ο μηχανισμός ενδοκυττάρωσης ευνοεί την εσωτερίκευση μεγάλων σωματιδίων διαμέτρου μεγαλύτερης των 750 nm⁴⁹. Παράλληλα, φαίνεται το σχήμα των νανοσωματιδίων να επηρεάζει την ικανότητα πρόσληψής τους από φαγοκύτταρα, με τα σφαιρικά νανοσωματίδια να εμφανίζουν μεγαλύτερη απόδοση κυτταρικής πρόσληψης σε σχέση με τις νανοράβδους. Τέλος, και η επιφανειακή χημεία είναι σημαντική για την φαγοκυττάρωση, καθώς ευνοούνται σωματίδια με φορτισμένη επιφάνεια που προσροφούν πιο εύκολα οψωνίνες⁵¹.

2.3.2. Μακροπυνοκύττωση

Η μακροπυνοκύττωση αναφέρεται στην πρόσληψη εξωκυτταρικής ουσίας ως απάντηση σε διάφορα σήματα που δέχεται το κύτταρο (π.χ. παράγοντες ανάπτυξης του κυττάρου)⁵⁰. Μεγάλες επεκτάσεις ή πτυχώσεις της κυτταρικής μεμβράνης σχηματίζονται με την βοήθεια του κυτταροσκελετού και στη συνέχεια επανενώνονται με την μεμβράνη δημιουργώντας μεγάλα μακροπυνοσώματα μεγέθους 0.2-5 μm⁵². Αυτή η διαδικασία είναι ανεξάρτητη της ύπαρξης ειδικών υποδοχέων στην επιφάνεια των νανοσωματιδίων, καθώς τα μακροπυνοσώματα προσλαμβάνουν όλα τα σωματίδια και τις διαλυτοποιημένες ουσίες του εξωκυττάρου υγρού που περικλείουν⁵¹.

Γενικά, η μακροπυνοκύττωση αξιοποιείται από διάφορα παθογόνα, όπως πρωτόζωα, βακτήρια και ιούς, για την πιο εύκολη είσοδο τους σε κύτταρα θηλαστικών⁵². Όσον αφορά όμως τα νανοσωματίδια, είναι ένας σημαντικός μηχανισμός για την εισαγωγή μεγάλου μεγέθους σωματιδίων μέσω των μεγάλων μακροπυνοσωμάτων⁵¹.

2.3.3. Ενδοκυττάρωση εξαρτώμενη από κλαθρίνη

Η ενδοκυττάρωση που εξαρτάται από την παρουσία κλαθρίνης είναι ο βασικός μηχανισμός πρόσληψης θρεπτικών συστατικών και συστατικών της κυτταρικής μεμβράνης, συμπεριλαμβανομένης της λιποπρωτεΐνης χαμηλής πυκνότητας (Low Density Lipoprotein, LDL) που προσδένεται στους LDL υποδοχείς^{50,51}. Αυτός ο τύπος ενδοκυττάρωσης πραγματοποιείται μόνο σε συγκεκριμένα σημεία της κυτταρικής μεμβράνης που καταλαμβάνουν περίπου το 0.5-2% της επιφάνειάς της και είναι πλούσια σε κλαθρίνη⁵¹. Η πρωτεΐνη κλαθρίνη σχηματίζει μία τρισκελή δομή που αποτελείται από 3 «βαριές» αλυσίδες κλαθρίνης συνδεδεμένες στενά με 1 «ελαφριά»

αλυσίδα έκαστη. Υπό συνθήκες περιβαλλοντικού στρες, αυτές οι δομές οργανώνονται αυθόρμητα σε μορφή κλειστών πολυγώνων, όμως υπό φυσιολογικές συνθήκες, η μεταβολή της δομής τους απαιτεί την ύπαρξη κατάλληλων πρωτεϊνών με δομικό ρόλο (assembly proteins)⁵⁰.

Για την πρόσληψη συστατικών μέσω της εξαρτώμενης από κλαθρίνη ενδοκυττάρωσης απαιτείται η ύπαρξη πρωτεϊνών προσαρμογέων και βοηθητικών πρωτεϊνών που σταθεροποιούν την κυτταρική μεμβράνη ώστε να οργανωθούν τα μόρια κλαθρίνης σε πενταγωνικά και εξαγωνικά πλέγματα. Αυτό οδηγεί στη δημιουργία σταθεροποιημένων εγκολπώσεων της κυτταρικής μεμβράνης, η επιφάνεια των οποίων καλύπτεται από κλαθρίνη και πρωτεΐνες με εξειδικευμένα κέντρα πρόσδεσης για τα κατάλληλα εξωκυτταρικά συστατικά^{51,53}. Τα ανοιχτά κυστιδία αυτά έχουν σχετικά μικρή διάμετρο 100-150 nm και περικλείουν όγκο εξωκυττάρου υγρού ανάλογο με τον εσωτερικό τους διαθέσιμο όγκο⁵¹. Κατά την εσωτερικεύσή τους στο κύτταρο χάνουν την επίστρωση κλαθρίνης και ενώνονται με άλλα κυστιδία για να δημιουργήσουν τα πρόωρα ενδοσώματα⁵⁴. Συχνά, τα νανοσωματίδια που εισέρχονται στο κύτταρο με αυτόν τον μηχανισμό καταλήγουν σε λυσοσώματα που στοχεύουν στην καταστροφή και την απομάκρυνση εξωτερικών παραγόντων και κυτταρικών αποβλήτων⁵¹.

Σημαντικός παράγοντας που επηρεάζει την πρόσληψη νανοϋλικών μέσω της εξαρτώμενης από την κλαθρίνη ενδοκυττάρωσης -εκτός από το μέγεθος που περιορίζεται από τη μέγιστη διάμετρο των κυστιδίων (~150 nm)- είναι το επιφανειακό φορτίο. Σύμφωνα με έρευνα σχετικά με την πρόσληψη πολυμερικών νανοσωματιδίων πολυγαλακτικού οξέος (PLA) σε κύτταρα της καρκινικής σειράς HeLa, τα ανιονικά σωματίδια επέδειξαν μικρό ρυθμό κυτταρικής πρόσληψης χωρίς να αξιοποιούν τον μηχανισμό της εξαρτώμενης από κλαθρίνη ενδοκυττάρωσης, σε αντίθεση με τα θετικά φορτισμένα νανοσωματίδια που προσλαμβάνονται γρήγορα μέσω αυτής της διόδου⁵⁵.

2.3.4. Ενδοκυττάρωση εξαρτώμενη από καβεολίνη

Σε πολλά είδη κυττάρων -και κατά κόρον στα επιθηλιακά κύτταρα- παρατηρούνται ειδικές εγκολπώσεις που ονομάζονται «caveolae». Οι εγκολπώσεις αυτές έχουν διάμετρο 60-80 nm και έναν πιο στενό «λαιμό» διαμέτρου 10-50 nm που τους δίνει το χαρακτηριστικό τους σχήμα⁵⁶ (Εικόνα 2). Το σχήμα και η δομή των εγκολπώσεων εξαρτάται από την ύπαρξη της διμερούς πρωτεΐνης καβεολίνη, η οποία συνδέεται με την χοληστερόλη της κυτταρικής μεμβράνης και δημιουργεί έναν βρόχο που εισέρχεται στην εξωτερική στιβάδα της μεμβράνης δημιουργώντας σταθερά κυστιδία επικαλυπτόμενα με καβεολίνη⁵⁰. Ο συγκεκριμένος μηχανισμός ενδοκυττάρωσης συμμετέχει σε σημαντικές βιολογικές διεργασίες όπως η ρύθμιση της συγκέντρωσης λιπιδίων, λιπαρών οξέων, μεμβρανικών πρωτεϊνών αλλά και της μεμβρανικής πίεσης⁵¹, ενώ θεωρείται ότι συμμετέχει και στην διακυτταρική μεταφορά χοληστερόλης⁵⁰. Κατά την εσωτερικεύση των κυστιδίων στο κύτταρο, η επίστρωση καβεολίνης δεν αποσυνδέεται, αλλά μετά από ένωση με άλλα κυστιδία καβεολίνης οδηγεί στον σχηματισμό των καβεοσωμάτων⁵⁴.

Η πρόσληψη νανοσωματιδίων μέσω του μηχανισμού της εξαρτώμενης από καβεολίνη ενδοκυττάρωσης έχει αποδειχθεί ότι εξαρτάται άμεσα από το μέγεθος τους.

Συγκεκριμένα, κατά την μελέτη πρόσληψης πολυμερικών νανοσωματιδίων με επικάλυψη ορού αλβουμίνης μόσχου (Bovine Serum Albumin, BSA) από επιθηλιακά κύτταρα, φάνηκε ότι νανοσωματίδια μικρότερου μεγέθους σε σχέση με την διάμετρο των εγκολλώσεων (20, 40 nm) έχουν 5-10 φορές μεγαλύτερη απόδοση εσωτερίκευσης σε σχέση με μεγάλα νανοσωματίδια διαμέτρου 100 nm. Παρ' όλα αυτά, παρατηρήθηκε πως ήταν δυνατή και η εσωτερίκευση νανοσωματιδίων μεγαλύτερης διαμέτρου από αυτή των εγκολλώσεων, γεγονός που αποδεικνύει την ικανότητα των στενών «λαιμών» των εγκολλώσεων να διαστέλλονται ώστε να εσωτερικευτούν μεγαλύτερα σωματίδια. Μάλιστα, αυτή η ιδιότητα των κυστιδίων καθιστά δυνατή την πρόσληψη περισσότερων του ενός νανοσωματιδίων (έως 3 νανοσωματίδια 20 nm και έως 2 νανοσωματίδια 40 nm) ανά κυστίδιο⁵⁶. Ομοίως, η πρόσληψη νανοσωματιδίων πολυστυρενίου από επιθηλιακά κύτταρα HUVEC είναι αυξημένη για μικρότερα μεγέθη νανοσωματιδίων⁵⁷.

Τέλος, σημαντική είναι και η επίδραση της επιφανειακής χημείας των νανοσωματιδίων. Στην πρώτη περίπτωση, παρατηρήθηκε ότι νανοσωματίδια χωρίς επικάλυψη BSA αδυνατούν να εσωτερικευτούν σε επιθηλιακά κύτταρα μέσω ενδοκυττάρωσης εξαρτώμενης από καβεολίνη⁵⁶, ενώ η προσθήκη πρωτεϊνικής «κορόνας» βόειου εμβρυικού ορού (FBS) στα νανοσωματίδια πολυστυρενίου είχε ως αποτέλεσμα την αντιστροφή της σχέσης μεγέθους-ικανότητας πρόσληψης, με την μεγαλύτερη ενδοκυττάρια συγκέντρωση νανοσωματιδίων να παρατηρείται για τα μεγαλύτερα νανοσωματίδια διαμέτρου 200 nm⁵⁷.

2.3.5. Ενδοκυττάρωση ανεξάρτητη από κλαθρίνη και καβεολίνη

Τα πρώτα χρόνια μελέτης των μηχανισμών ενδοκυττάρωσης υπερίσχυε η άποψη ότι η πρόσληψη εξωκυττάρων ουσιών με μηχανισμούς ενδοκυττάρωσης με απουσία της πρωτεΐνης κλαθρίνη είναι αδύνατη. Πιο πρόσφατες μελέτες όμως ανέδειξαν την ύπαρξη μηχανισμών ανεξάρτητων τόσο της παρουσίας κλαθρίνης αλλά και καβεολίνης⁵⁸. Ένας εξ αυτών είναι ο μηχανισμός της μακροπινοκύττωσης⁵⁸ που έχει ήδη αναλυθεί, όμως υπάρχουν και άλλες δίοδοι ενδοκυττάρωσης με μικρότερη συνεισφορά στην κυτταρική πρόσληψη νανοσωματιδίων⁵⁴.

Μηχανισμοί ενδοκυττάρωσης που είναι ανεξάρτητη της κλαθρίνης αλλά εξαρτώνται από την παρουσία της πρωτεΐνης δυναμίνη για την δημιουργία των κυστιδίων περιλαμβάνουν την ενδοκυττάρωση εξαρτώμενη από την δυναμίνη και την πρωτεΐνη «RhoA» αλλά και τον μηχανισμό «Fast Endophilin-Mediated Endocytosis» (FEME)⁵⁹. Αυτοί οι μηχανισμοί είναι υπεύθυνοι για την εσωτερίκευση υποδοχέων και τοξινών όπως η τοξίνη «Shiga»⁶⁰. Παράλληλα, μηχανισμοί ανεξάρτητοι της παρουσίας δυναμίνης, όπως η δίοδος που εξαρτάται από την παρουσία της πρωτεΐνης CDC42, είναι υπεύθυνοι για την πρόσληψη εξωκυττάρου υγρού, τοξινών πρωτεϊνών που προσδένονται σε λιπίδια και διαμεμβρανικών πρωτεϊνών (π.χ. της πρωτεΐνης CD44)⁶⁰.

Ανεξάρτητα από την ανάγκη παρουσίας δυναμίνης, η ανεξάρτητη από κλαθρίνη και καβεολίνη ενδοκυττάρωση φαίνεται να σχετίζεται με ειδικές πρωτεΐνες, τις «flotillin 1» και «flotillin 2», οι οποίες απαντώνται σε επίπεδες περιοχές της κυτταρικής μεμβράνης με έντονη παρουσία λιπιδίων (lipid rafts) και σχετίζονται με την συγκέντρωση ειδικών λιπιδίων και υποδοχέων συνδεδεμένων με λιπίδια. Είναι πιθανόν

ότι αυτή η συσσώρευση λιπιδικών μορίων γύρω από εγκολπώσεις της κυτταρικής μεμβράνης να είναι καθοριστική για τον σχηματισμό και την εσωτερίκευση των κυστιδίων⁵⁸.

3. Θεωρητικές βάσεις και μεθοδολογικές πρακτικές στη μηχανική μάθηση

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να δώσουν χρήσιμες πληροφορίες για τις πολύπλοκες σχέσεις μεταξύ των διάφορων μεταβλητών ενός προβλήματος, οδηγώντας σε αξιόπιστες προβλέψεις. Προϋπόθεση, όμως, για την δημιουργία γενικεύσιμων και ερμηνεύσιμων μοντέλων είναι η μεθοδική προεπεξεργασία των δεδομένων για την αποφυγή μεροληψίας στις προβλέψεις, αλλά και η εκλογή του κατάλληλου μοντέλου μηχανικής μάθησης για το εκάστοτε σύνολο δεδομένων⁶¹⁻⁶³.

Στο παρόν κεφάλαιο παρουσιάζονται βασικές τεχνικές προεπεξεργασίας στη μηχανική μάθηση, όπως ο τυχαίος διαχωρισμός μεταβλητών, η κωδικοποίηση κατηγορικών τιμών και η κανονικοποίηση με τη μέθοδο μηδενικής μέσης τιμής. Επίσης, περιγράφεται ο αλγόριθμος XGBoost και τα μέτρα αξιολόγησης μοντέλων, ενώ αναφέρεται η χρήση τυχαιοποίησης των τιμών της μεταβλητής εξόδου. Τέλος, παρουσιάζεται η εύρεση πεδίου εφαρμογής με k πλησιέστερους γείτονες και η ανάλυση SHAP.

3.1. Προεπεξεργασία δεδομένων

3.1.1. Διαχωρισμός δεδομένων σε σύνολο εκπαίδευσης και ελέγχου

Απαραίτητη για την αντικειμενική αξιολόγηση της προβλεπτικής ικανότητας ενός μοντέλου μηχανικής μάθησης είναι η ύπαρξη ενός «άγνωστου» συνόλου δεδομένων που δεν συμμετέχει στην μοντελοποίηση, ώστε να ελεγχθεί η ακρίβεια και η δυνατότητα γενίκευσης των προβλέψεων του μοντέλου. Για τον λόγο αυτό, αποτελεί καθιερωμένη πρακτική ο διαχωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης (training) και ελέγχου (testing). Το σύνολο εκπαίδευσης περιέχει το μεγαλύτερο ποσοστό των συνολικών δεδομένων (80%, 75% ή 70% συνήθως) για την εκμάθηση από το μοντέλο, ενώ το σύνολο ελέγχου αξιοποιείται μόνο για τον έλεγχο των προβλέψεων σε «άγνωστα» δεδομένα^{64,65}.

Έχουν προταθεί πολλές διαφορετικές μέθοδοι για έναν αποδοτικό διαχωρισμό του συνόλου δεδομένων, όμως η πιο απλή μέθοδος είναι ο τυχαίος διαχωρισμός σύμφωνα με το ζητούμενο ποσοστό παρατηρήσεων στα δύο σύνολα. Παρά την απλότητα της, η μέθοδος μπορεί να μειώσει σημαντικά τον κίνδυνο υπερπροσαρμογής του μοντέλου. Συγκεκριμένα περιορίζει τον κίνδυνο «διαφυγής πληροφορίας» (data leakage)⁶⁶ από το σύνολο ελέγχου στο σύνολο εκπαίδευσης, εξασφαλίζοντας ότι τα δεδομένα του συνόλου ελέγχου δεν συμμετέχουν στην εκπαίδευση. Σε αντίθετη περίπτωση, δημιουργούνται μοντέλα μηχανικής μάθησης που προβλέπουν με φαινομενικά μεγάλη ακρίβεια το σύνολο ελέγχου αλλά αποτυγχάνουν να γενικεύσουν τις προβλέψεις τους για άγνωστα ή εξωτερικά δεδομένα.

3.1.2. Μετατροπή κατηγορικών μεταβλητών σε αριθμητικές

Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης απαιτούν την μετατροπή των κατηγορικών μεταβλητών σε αριθμητικές. Είτε πρόκειται για γραμμικά μοντέλα και μοντέλα «βαθιάς μάθησης» (deep learning) που δημιουργούν γραμμικές σχέσεις μεταξύ των μεταβλητών, είτε μοντέλα που ανήκουν σε διαφορετικές υποκατηγορίες της μηχανικής μάθησης, όπως οι συνδυαστικές μέθοδοι (ensemble methods), είναι απαραίτητο όλες οι μεταβλητές εισόδου του μοντέλου να έχουν αριθμητικές τιμές⁶⁷. Μία από τις πιο ευρέως χρησιμοποιούμενες μεθόδους είναι η μέθοδος «One-hot encoding» σύμφωνα με την οποία κάθε κατηγορία μιας κατηγορικής μεταβλητής μετατρέπεται σε ένα μοναδικό δυαδικό διάνυσμα. Για κάθε παρατήρηση, το αντίστοιχο διάνυσμα λαμβάνει την τιμή 1 στη στήλη που αντιπροσωπεύει την κατηγορία της, ενώ όλες οι υπόλοιπες στήλες παίρνουν την τιμή 0. Αυτό σημαίνει ότι για μια μεταβλητή με k μοναδικές κατηγορίες, δημιουργούνται k νέες δυαδικές στήλες (ή χαρακτηριστικά)⁶⁸.

Πλεονέκτημα αυτής της μεθόδου είναι η διατήρηση της σημαντικότητας κάθε κατηγορίας για την πρόβλεψη αλλά και το γεγονός ότι, σε αντίθεση με άλλες μεθόδους, μπορεί να κωδικοποιήσει και μεταβλητές των οποίων οι κατηγορίες δεν είναι σειριακές (ordinal variables). Παρ' όλα αυτά, αυξάνει σημαντικά τον αριθμό των μεταβλητών εισόδου αλλά και των μοναδικών τιμών⁶⁹. Τα μειονεκτήματα αυτά, βεβαίως, μπορούν να αντιμετωπιστούν στη συνέχεια μέσω επιλογής μεταβλητών.

3.1.3. Κανονικοποίηση μεταβλητών εισόδου

Δεδομένου ότι οι διαφορετικές μονάδες μέτρησης κάθε μεταβλητής επηρεάζουν σημαντικά το εύρος τιμών της, είναι πιθανόν μεταβλητές με τιμές μεγαλύτερης τάξης μεγέθους να έχουν μεγαλύτερη βαρύτητα στην τελική πρόβλεψη, χωρίς όμως αυτό να αντικατοπτρίζει τις πραγματικές σχέσεις των μεταβλητών. Για να μειωθεί η εξάρτηση των προβλέψεων από το εύρος τιμών των μεταβλητών και να εξισοροπηθεί η συμμετοχή των διαφόρων μεταβλητών στην τελική πρόβλεψη, συχνά πραγματοποιείται κανονικοποίηση, δηλαδή μετατροπή των δεδομένων ώστε να αποκτήσουν μικρότερο εύρος^{70,71}. Μία από τις πιο γνωστές μεθόδους κανονικοποίησης δεδομένων είναι η κανονικοποίηση μηδενικής μέσης τιμής (Z-score normalization), η οποία είναι ιδιαίτερα χρήσιμη όταν υπάρχουν ακραίες τιμές στο σύνολο δεδομένων. Για τις τιμές κάθε μεταβλητής η Z-score κανονικοποίηση εφαρμόζεται σύμφωνα με την Σχέση 1⁷¹:

$$X_{i,norm} = \frac{X_i - \mu_A}{\sigma_A} \quad (1)$$

Όπου $X_{i,norm}$ οι κανονικοποιημένες τιμές της μεταβλητής A, X_i οι αρχικές τιμές της μεταβλητής A, μ_A η μέση τιμή των δεδομένων της μεταβλητής A και σ_A η τυπική απόκλιση των δεδομένων της μεταβλητής A.

3.1.4. Συμπλήρωση ελλιπών τιμών

Η «Multiple Imputation by Chained Equations» (MICE)⁷² αποτελεί μια επαναληπτική μεθοδολογία η οποία δημιουργεί μέσω μοντέλων παλινδρόμησης πολλαπλές προβλέψεις για κάθε ελλιπή τιμή λαμβάνοντας υπ' όψιν την αβεβαιότητα της πρόβλεψης. Τα αποτελέσματα κάθε επανάληψης ενώνονται σε ένα τελικό σύνολο προβλέψεων που εξασφαλίζει την μείωση των σφαλμάτων λόγω μεροληψίας. Σημαντικό πλεονέκτημα αυτής της μεθόδου αποτελεί η διατήρηση των σχέσεων μεταξύ των μεταβλητών. Τονίζεται ότι απαραίτητη προϋπόθεση για την εφαρμογή της MICE όπως παρουσιάζεται παρακάτω είναι η παραδοχή «Missing-at-Random» (MAR), σύμφωνα με την οποία η πιθανότητα μία τιμή να είναι ελλιπής εξαρτάται μόνο από τα δεδομένα που έχουν συμπεριληφθεί στο σύνολο δεδομένων και όχι από εξωτερικούς παράγοντες που δεν μπορούν να προβλεφθούν^{73,74}.

Δεδομένου ότι ισχύει η παραδοχή MAR, ένας τυπικός αλγόριθμος MICE που μπορεί να εφαρμοστεί είναι ο εξής⁷³.

1^ο Βήμα: Απλή πρόβλεψη των ελλιπών τιμών, π.χ. μέσω της μέσης τιμής.

2^ο Βήμα: Οι αρχικές προβλέψεις για την πρώτη μεταβλητή (A) με ελλιπείς τιμές αντικαθίστανται εκ νέου από ελλιπείς τιμές.

3^ο Βήμα: Η μεταβλητή A χρησιμοποιείται ως μεταβλητή εξόδου σε μοντέλο παλινδρόμησης, ενώ όλες οι υπόλοιπες μεταβλητές (συμπεριλαμβανομένων των μεταβλητών που περιλαμβάνουν ελλιπείς τιμές που έχουν αντικατασταθεί από την μέση τιμή) αποτελούν τις ανεξάρτητες μεταβλητές του μοντέλου.

4^ο Βήμα: Οι ελλιπείς τιμές της μεταβλητής A αντικαθίστανται από προβλέψεις του μοντέλου.

5^ο Βήμα: Τα βήματα 2-4 επαναλαμβάνονται για όλες τις μεταβλητές που περιέχουν ελλιπείς τιμές, ενώ κάθε φορά οι προηγούμενες μεταβλητές για τις οποίες έχουν πραγματοποιηθεί ήδη προβλέψεις χρησιμοποιούνται ως ανεξάρτητες μεταβλητές για τα νέα μοντέλα. Αφού έχει ολοκληρωθεί ένα «κύκλος» προβλέψεων για όλες τις μεταβλητές με ελλιπείς τιμές, όλες οι ελλιπείς τιμές έχουν αντικατασταθεί από προβλέψεις που αντικατοπτρίζουν τις σχέσεις μεταξύ των μεταβλητών.

6^ο Βήμα: Τα βήματα 2-5 επαναλαμβάνονται για τον αριθμό των επαναλήψεων που έχουν οριστεί από τον χρήστη και διατηρούνται οι τιμές του τελευταίου «κύκλου» προβλέψεων. Οι υπάρχουσες τιμές των πρωτογενών δεδομένων παραμένουν ίδιες καθ' όλη την διάρκεια των επαναλήψεων.

3.2. Μοντελοποίηση δεδομένων και βελτιστοποίηση μοντέλου μηχανικής μάθησης

3.2.1. Ο αλγόριθμος XGBoost

Ο αλγόριθμος «extreme Gradient Boosting» (XGBoost) πρόκειται για μία βελτιωμένη πρόταση μοντελοποίησης στον χώρο της επιτηρούμενης μηχανικής μάθησης (supervised machine learning) που βασίζεται στην κλασική μέθοδο «Ενίσχυσης Κλίσης

Δέντρων Απόφασης» (Gradient Tree Boosting) και μπορεί να δώσει ικανοποιητικά αποτελέσματα έως και 10 φορές πιο γρήγορα από άλλες μεθόδους μηχανικής μάθησης⁷⁵.

Σύμφωνα με τη «Gradient Tree Boosting» μεθοδολογία, το μοντέλο βασίζεται στην δημιουργία και τον συνδυασμό διαφορετικών απλών «δέντρων απόφασης» (decision trees), με κάθε νέο «δέντρο» που δημιουργείται να έχει ως στόχο την διόρθωση των σφαλμάτων των προηγούμενων. Η διαδοχική δημιουργία των νέων μοντέλων βασίζεται στην βελτιστοποίηση μίας συνάρτησης απωλειών (loss function). Πιο συγκεκριμένα, σε κάθε στάδιο το νέο μοντέλο προσαρμόζεται στην αρνητική κλίση (negative gradient) της συνάρτησης απωλειών του συνόλου των μοντέλων που έχουν δημιουργηθεί μέχρι εκείνο το σημείο⁷⁶.

Ο αλγόριθμος XGBoost ακολουθεί την ίδια λογική ελαχιστοποίησης της συνάρτησης απωλειών, προσφέροντας επιπροσθέτως την δυνατότητα ελέγχου της πολυπλοκότητας του μοντέλου μέσω κανονικοποίησης (regularisation) με στόχο την αποφυγή της υπερπροσαρμογής του μοντέλου. Έτσι, η αντικειμενική συνάρτηση ($Obj^{(t)}$) αποκτά τη μορφή:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (2)$$

Όπου l είναι μια διαφορίσιμη κυρτή συνάρτηση απώλειας (π.χ. μέσο τετραγωνικό σφάλμα ή λογιστική απώλεια), y_i η πραγματική τιμή της i παρατήρησης, \hat{y}_i η προβλεπόμενη τιμή της i παρατήρησης και f_t το δέντρο απόφασης που έχει δημιουργηθεί μετά την t -επανάληψη. Η $\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2$ είναι μία συνάρτηση που τιμωρεί την πολυπλοκότητα (γ : υπερπαραμέτρος, T : αριθμός φύλλων, ω : βάρη φύλλων)⁷⁵.

Για την απλούστευση και την αύξηση της ταχύτητας των υπολογισμών χρησιμοποιείται η ανάπτυξη Taylor της συνάρτησης απωλειών και προκύπτει μία προσεγγιστική αντικειμενική συνάρτηση (Σχέση 3) που περιλαμβάνει τόσο την πρώτη όσο και την δεύτερη (Hessian) παράγωγο της Σχέσης 2:

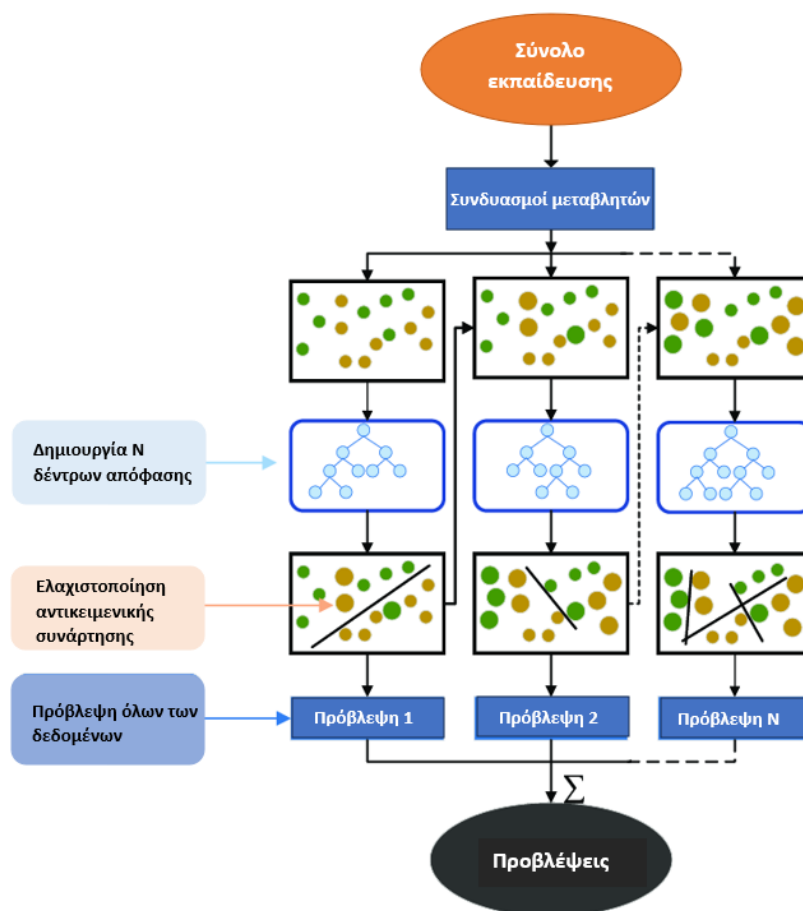
$$Obj^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \quad (3)$$

Όπου g_i η πρώτη παράγωγος (gradient) ως προς την προηγούμενη πρόβλεψη και h_i η δεύτερη παράγωγος (Hessian)⁷⁵.

Παράλληλα, η βελτιωμένη αυτή μέθοδος καθιστά δυνατή την περαιτέρω απλοποίηση του μοντέλου μέσω «κλαδέματος δέντρου» (Tree pruning), δηλαδή του εντοπισμού και της αφαίρεσης κλάδων που δεν βελτιώνουν την απόδοση του μοντέλου. Επίσης, σε αντίθεση με την κλασική «Boosting» μεθοδολογία, μπορεί να διαχειριστεί αραιά ή ελλιπή δεδομένα (sparsity awareness) και επιπλέον επιτρέπει την παραλληλοποίηση της κατασκευής των δέντρων με υποσύνολα των αρχικών μεταβλητών, επιτρέποντας ταχύτερο διαχωρισμό κόμβων⁷⁵.

Έτσι, ένας συνοπτικός αλγόριθμος XGBoost (όπως φαίνεται και στο Σχήμα 2) είναι ο εξής:

- 1^ο Βήμα: Αρχικοποίηση προβλέψεων (π.χ. σταθερή τιμή).
- 2^ο Βήμα: Υπολογισμός κλίσης και δεύτερης παραγώγου.
- 3^ο Βήμα: Προσαρμογή ενός «δέντρου απόφασης» με σκοπό την ελαχιστοποίηση της προσεγγιστικής αντικειμενικής συνάρτησης.
- 4^ο Βήμα: Προσθήκη του νέου «δέντρου απόφασης» στο σύνολο των προηγούμενων.
- 5^ο Βήμα: Ενημέρωση προβλέψεων.
- 6^ο Βήμα: Επανάληψη βημάτων 2-5 μέχρι την σύγκλιση ή την ικανοποίηση των κριτηρίων τερματισμού.



Σχήμα 2. Αναπαράσταση αλγορίθμου XGBoost (τροποποιημένο από Yao et al. (2022)⁷⁷).

3.2.2. Υπερπαραμέτροι των μοντέλων XGBoost

Η εκπαίδευση ενός μοντέλου XGBoost περιλαμβάνει την εύρεση των κατάλληλων τιμών των υπερπαραμέτρων για την γρήγορη και αποτελεσματική σύγκλιση του μοντέλου και την βελτίωση της προβλεπτικής του ικανότητας. Μερικές από τις πιο

σημαντικές υπερπαραμέτρους που είναι απαραίτητο να καθοριστούν πριν την ανάπτυξη του μοντέλου επεξηγούνται παρακάτω⁷⁸:

- Ρυθμός εκμάθησης (learning rate): Αριθμός (από 0 έως 1) που χρησιμοποιείται για την μείωση της τιμής των βαρών των μεταβλητών που προκύπτουν μετά από κάθε γύρο εφαρμογής της μεθοδολογίας. Όσο μικρότερη η τιμή της παραμέτρου, τόσο πιο αργή και συντηρητική είναι η σύγκλιση.
- Ελάχιστη μείωση της συνάρτησης απώλειας για περαιτέρω διαχωρισμό σε έναν κόμβο φύλλων του «δέντρου απόφασης»: Όσο μεγαλύτερη η τιμή, τόσο πιο συντηρητικός ο αλγόριθμος.
- Μέγιστο βάθος δένδρου (maximum depth): Όσο μεγαλύτερη (θετική) τιμή λαμβάνει αυτή η παράμετρος, τόσο πιο περίπλοκο θα είναι το μοντέλο, αυξάνοντας τον κίνδυνο υπερπροσαρμογής στα δεδομένα.
- Ελάχιστο βάρος «παιδιού» (minimum child weight): Το ελάχιστο άθροισμα βαρών των παρατηρήσεων που απαιτείται για να πραγματοποιηθεί διαχωρισμός σε έναν κόμβο του «δέντρου». Αυξάνοντας την τιμή της παραμέτρου το μοντέλο γίνεται πιο συντηρητικό.
- Λόγος υποσυνόλου παρατηρήσεων εκπαίδευσης (subsampling): Ποσοστό παρατηρήσεων από το σύνολο εκπαίδευσης που χρησιμοποιεί ο αλγόριθμος σε κάθε γύρο εκμάθησης. Μικρότερες τιμές της υπερπαραμέτρου οδηγούν σε πιο συντηρητικά μοντέλα.
- Λόγος υποσυνόλου στηλών σε κάθε δένδρο απόφασης: Ένα υποσύνολο των στηλών του συνόλου δεδομένων που χρησιμοποιείται για την δημιουργία κάθε δένδρου απόφασης. Μικρότερες τιμές της υπερπαραμέτρου οδηγούν σε πιο συντηρητικά μοντέλα.
- Παράμετροι της συνάρτησης κανονικοποίησης (lambda, alpha): Αύξηση της τιμής των υπερπαραμέτρων αυτών μειώνει τον κίνδυνο υπερπροσαρμογής των μοντέλων.

Σημειώνεται πως, σύμφωνα με τα παραπάνω, ένα μοντέλο XGBoost χαρακτηρίζεται πιο συντηρητικό όταν οι αλλαγές στις προβλέψεις σε κάθε επανάληψη είναι μικρές (μικρός ρυθμός εκμάθησης) και τα «δέντρα» απόφασης που δημιουργούνται είναι πιο απλά (μικρό βάθος και περισσότερα «φύλλα»).

3.2.3. Επιλογή βέλτιστων τιμών υπερπαραμέτρων μέσω «cross-validation»

Μία από τις πιο γνωστές και ευρέως χρησιμοποιούμενες μεθόδους για την εύρεση των βέλτιστων τιμών των υπερπαραμέτρων ενός μοντέλου ονομάζεται «Διασταυρωμένη Επικύρωση k-Διαίρεσεων» (k-fold cross-validation). Στην μέθοδο αυτή, το σύνολο δεδομένων διαχωρίζεται σε k μη-επικαλυπτόμενα υποσύνολα ίδιου μεγέθους μέσω τυχαίας επιλογής παρατηρήσεων χωρίς αντικατάσταση (without replacement). Στη συνέχεια, το μοντέλο εκπαιδεύεται σε όλους τους συνδυασμούς k-1 υποσυνόλων που αποτελούν το νέο σύνολο εκπαίδευσης και αξιολογείται μέσω στατιστικών μέτρων σχετικά με την ακρίβεια των προβλέψεων στο εκάστοτε k υποσύνολο. Αυτή η διαδικασία επαναλαμβάνεται ώστε κάθε k υποσύνολο να λειτουργήσει μία φορά ως

σύνολο ελέγχου. Ο μέσος όρος των στατιστικών μέτρων μεταξύ k διαφορετικών μοντέλων που ελέγχονται στα k υποσύνολα ελέγχου είναι ένα δείγμα για την συνολική απόδοση των μοντέλων κατά την διαδικασία του «cross-validation»⁷⁹. Η επιλογή του αριθμού υποσυνόλων, k , είναι σημαντική για την αξιολόγηση της προβλεπτικής ικανότητας των μοντέλων καθώς όσο μεγαλύτερη είναι η τιμή k τόσο μειώνεται η μεροληψία για ορισμένα μοντέλα (όπως τα γραμμικά). Στις περισσότερες εφαρμογές χρησιμοποιούνται συνήθως τιμές k μεταξύ 5 και 10⁸⁰.

Όταν πρόκειται για την επιλογή των βέλτιστων τιμών υπερπαραμέτρων, είναι δυνατόν να επιλεγούν ορισμένες τυπικές τιμές για κάθε υπερπαραμέτρο του μοντέλου και να εφαρμοστεί η μέθοδος «cross-validation». Πιο συγκεκριμένα, κάθε πιθανός συνδυασμός των υπερπαραμέτρων χρησιμοποιείται για την δημιουργία m μοντέλων μηχανικής μάθησης που δοκιμάζονται σε k υποσύνολα ελέγχου αντίστοιχα και αξιολογούνται με βάση κατάλληλα στατιστικά μέτρα. Τελικά, επιλέγεται το μοντέλο που για συγκεκριμένο συνδυασμό υπερπαραμέτρων έχει την βέλτιστη μέση τιμή των στατιστικών μέτρων κατά την διαδικασία του «cross-validation». Βεβαίως, η δημιουργία και εκπαίδευση διαφορετικών μοντέλων για πολλαπλούς συνδυασμούς υπερπαραμέτρων και η αξιολόγηση τους σε μεγάλο αριθμό συνόλων ελέγχου αυξάνει σημαντικά το υπολογιστικό κόστος αλλά είναι ένας σχετικά γρήγορος και αποδοτικός τρόπος για την βελτιστοποίηση περίπλοκων μοντέλων μηχανικής μάθησης⁸¹.

3.3. Αξιολόγηση και πεδίο εφαρμοσιμότητας μοντέλων μηχανικής μάθησης

3.3.1. Αξιολόγηση μοντέλων παλινδρόμησης

Η αξιολόγηση του μοντέλου στο σύνολο ελέγχου, το οποίο περιέχει δεδομένα που δεν συμμετείχαν στην εκπαίδευση, είναι απαραίτητη για την επιβεβαίωση της υψηλής προβλεπτικής ικανότητας των μοντέλων. Για τον σκοπό αυτό, τα μοντέλα παλινδρόμησης αξιολογούνται με βάση τα παρακάτω στατιστικά μέτρα⁸².

- Συντελεστής προσδιορισμού (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \text{mean}(y_i))^2} \quad (4)$$

Όπου y_i η πραγματική τιμή και \hat{y}_i η προβλεπόμενη τιμή.

Ο συντελεστής προσδιορισμού καθορίζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που μπορεί να προβλεφθεί από τις ανεξάρτητες μεταβλητές. Όσο πιο κοντά στο 1 είναι ο συντελεστής προσδιορισμού τόσο πιο ικανοποιητική είναι η εξήγηση της διακύμανσης από το μοντέλο.

- Μέσο απόλυτο σφάλμα (Mean Absolut Error, MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

Όπου n ο αριθμός των παρατηρήσεων.

Το μέτρο MAE αποτελεί μία απλή μέτρηση του μέσου μεγέθους των σφαλμάτων μεταξύ πραγματικών και προβλεπόμενων τιμών χωρίς να λαμβάνει υπ' όψιν την κατεύθυνση τους.

- Μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

Το MSE, λόγω της τετραγωνικής φύσης του, «τιμωρεί» περισσότερο τα μεγαλύτερα σφάλματα σε σχέση με το MAE. Είναι επιθυμητό η τιμή του να είναι κοντά στο 0 για μία ικανοποιητική πρόβλεψη.

- Ριζικό μέσο τετραγωνικό σφάλμα (Root Mean Squared Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

Το μέτρο RMSE αποτελεί την τυπική απόκλιση του σφάλματος και είναι ένας δείκτης της προσέγγισης των πραγματικών τιμών από τις προβλέψεις του μοντέλου.

Τα στατιστικά μέτρα MAE, MSE και RMSE είναι μέτρα απόστασης⁸³ και εξαρτώνται από την τάξη μεγέθους και το εύρος της μεταβλητής εξόδου. Έτσι, για μεταβλητές εξόδου με πολύ μεγάλο εύρος τιμών μπορεί να χρησιμοποιηθεί ένα κανονικοποιημένο RMSE διαιρώντας την αρχική τιμή με το εύρος τιμών της μεταβλητής εξόδου (Y).

$$\text{Κανονικοποιημένο RMSE} = \frac{RMSE}{\max(Y) - \min(Y)} \quad (8)$$

Εκτός από τα στατιστικά μέτρα αξιολόγησης έχει προταθεί και μία μέθοδος τυχαίας αντικατάστασης των τιμών της μεταβλητής εξόδου (Y -randomization) για την αξιολόγηση των μοντέλων μηχανικής μάθησης. Σύμφωνα με αυτή τη μέθοδο, που βρίσκει συχνή εφαρμογή στην αξιολόγηση μοντέλων QSAR⁸⁴, οι τιμές της μεταβλητής εξόδου Y αρχικά αναδιατάσσονται τυχαία και ένα νέο μοντέλο παλινδρόμησης προσαρμόζεται στις αρχικές μεταβλητές εισόδου X και την αναδιατεταγμένη μεταβλητή εξόδου. Αυτή η διαδικασία αναμένεται να επηρεάσει αρνητικά τις σχέσεις μεταξύ των μεταβλητών, μειώνοντας δραστικά την τιμή του μέτρου R^2 . Μάλιστα, προκειμένου να επιβεβαιωθεί ότι δεν προκύπτει κάποια τυχαία σχέση συσχέτισης των

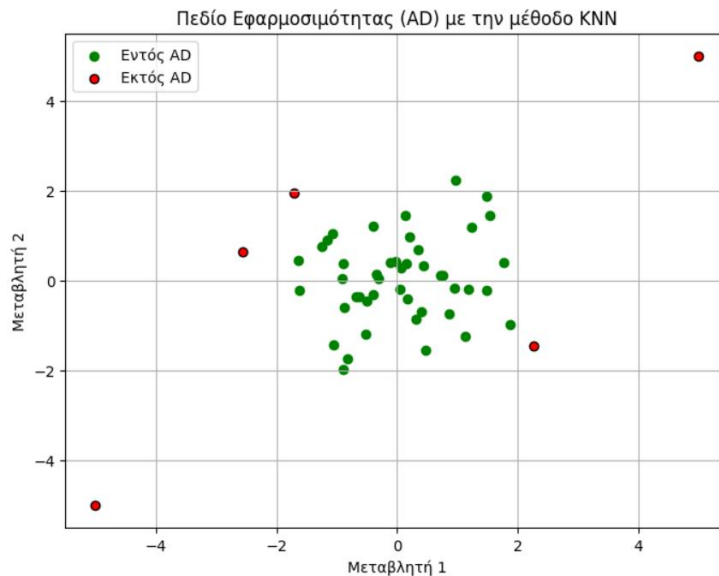
μεταβλητών με την αναδιατεταγμένη μεταβλητή εξόδου, η διαδικασία αυτή επαναλαμβάνεται πολλές φορές για διαφορετικές τυχαίες αναδιατάξεις του διανύσματος απόκρισης Y . Αν σε όλες τις περιπτώσεις το R^2 λαμβάνει χαμηλές τιμές, εξάγεται με ασφάλεια το συμπέρασμα ότι το αρχικό μοντέλο δεν έχει παράξει ικανοποιητικές τιμές μετρικών αποτίμησης εξαιτίας τυχαίας συσχέτισης⁸⁵.

3.3.2. Πεδίο εφαρμοσιμότητας μοντέλων μηχανικής μάθησης

Εξίσου σημαντικός με την επικύρωση του μοντέλου στο σύνολο δεδομένων ελέγχου μέσω στατιστικών και άλλων μέτρων είναι ο χαρακτηρισμός των προβλέψεων με βάση την αξιοπιστία τους. Αυτός μπορεί να πραγματοποιηθεί μέσω του προσδιορισμού ενός πεδίου εφαρμοσιμότητας (Applicability Domain, AD) του μοντέλου, δηλαδή ενός υποχώρου δεδομένων εντός του οποίου το μοντέλο δίνει αξιόπιστες προβλέψεις. Γενικά, η δυνατότητα γενίκευσης ενός μοντέλου μηχανικής μάθησης είναι ανάλογη με το εύρος του πεδίου εφαρμοσιμότητας⁸⁶.

Διάφορες μέθοδοι έχουν χρησιμοποιηθεί για τον προσδιορισμό του πεδίου εφαρμοσιμότητας μοντέλων μηχανικής μάθησης, με τις πιο απλές να βασίζονται στην απόσταση των προς πρόβλεψη δεδομένων από τα δεδομένα του συνόλου εκπαίδευσης. Αυτές, οι μέθοδοι υπερέχουν έναντι πιο περίπλοκων μεθόδων που χρησιμοποιούνται κυρίως για QSAR μοντέλα (π.χ. μέθοδος Convex Hull και Leverage) λόγω της απλότητας και της ευκολίας διαισθητικής κατανόησης, καθώς φαίνεται λογικό ότι όσο αυξάνεται η απόσταση από τα δεδομένα εκπαίδευσης τόσο μειώνεται η ικανότητα πρόβλεψης του μοντέλου και άρα η πιθανότητα αξιόπιστης πρόβλεψης^{86,87}.

Μεταξύ των μεθόδων που βασίζονται στην απόσταση, συχνά χρησιμοποιείται η μέθοδος k -κοντινότερων γειτόνων (k -Nearest Neighbours, k NN) σύμφωνα με την οποία ορίζεται από τον χρήστη ένα ανώτατο όριο απόστασης των δειγμάτων προς πρόβλεψη από τους k κοντινότερους γείτονές τους στο σύνολο εκπαίδευσης. Σε περίπτωση που ο μέσος όρος των k αποστάσεων ξεπερνά το όριο αυτό, η πρόβλεψη κρίνεται μη αξιόπιστη⁸⁸. Βεβαίως, ο προσδιορισμός του πεδίου εφαρμοσιμότητας σε αυτή την περίπτωση εξαρτάται άμεσα από την επιλογή του αριθμού k των κοντινότερων γειτόνων, του ανώτατου ορίου απόστασης αλλά και του τρόπου υπολογισμού της απόστασης (π.χ. Ευκλείδεια)^{87,88}. Στο Διάγραμμα 1 παρουσιάζεται μία τυπική περίπτωση εύρεσης του πεδίου εφαρμοσιμότητας με αυτή την μέθοδο.



Διάγραμμα 1. Παράδειγμα εύρεσης πεδίου εφαρμοσιμότητας με την μέθοδο kNN. Τα σημεία εκείνα των οποίων η απόσταση από τους κοντινότερους γείτονες ξεπερνά το προκαθορισμένο κατώφλι απόστασης, επισημασμένα με κόκκινο, θεωρούνται εκτός του πεδίου εφαρμοσιμότητας, και κατ' επέκταση η πρόβλεψη του μοντέλου για αυτά τα σημεία δεν θεωρείται αξιόπιστη.

3.4. Ερμηνεία μοντέλων μέσω της μεθόδου «Shapley Additive exPlanations»

Στοχεύοντας στην κατανόηση του τρόπου με τον οποίο «παίρνουν αποφάσεις» τα μοντέλα τεχνητής νοημοσύνης, έχει αναπτυχθεί η μέθοδος «Shapley Additive exPlanations» (SHAP) που χρησιμοποιεί τις τιμές «Shapley» της θεωρίας παιγνίων⁸⁹ με σκοπό την αξιολόγηση της συμμετοχής κάθε μεταβλητής στην προβλεπτική διαδικασία των μοντέλων μηχανικής μάθησης. Σύμφωνα με την θεωρία, η δίκαιη συνεισφορά όλων των παικτών ενός συνεργατικού παιχνιδιού διασφαλίζεται όταν ισχύουν οι παρακάτω προϋποθέσεις:

1^η προϋπόθεση: Αποδοτικότητα. Η τελική ανταμοιβή του παιχνιδιού πρέπει να ισούται με το άθροισμα της συνεισφοράς όλων των παικτών.

2^η προϋπόθεση: Συμμετρία. Αν δύο παίκτες συνεισφέρουν το ίδιο σε όλα τα υποσύνολα παικτών τότε θα πρέπει να λάβουν ίδια ανταμοιβή.

3^η προϋπόθεση: Προσθετικότητα. Σε παιχνίδι με υποπαιχνίδια ξεχωριστής ανταμοιβής, το άθροισμα των ανταμοιβών ενός παίκτη σε αυτά πρέπει να ισούται με την συνολική του ανταμοιβή στο παιχνίδι.

4^η προϋπόθεση: Μηδενικός παίκτης. Αν ένας παίκτης δεν συνεισφέρει σε κανένα υποσύνολο παικτών τότε η ανταμοιβή του είναι 0.

Έχει αποδειχθεί ότι υπάρχει μόνο μία σχέση που ικανοποιεί όλες τις προϋποθέσεις ενός δίκαιου παιχνιδιού και έτσι οι τιμές «Shapley» (φ_j) υπολογίζονται σύμφωνα με την Σχέση 9⁹⁰:

$$\varphi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (V(S \cup \{j\}) - V(S)) \quad (9)$$

Όπου S ένα υποσύνολο παικτών, N το σύνολο παικτών, $(V(S \cup \{j\}) - V(S))$ η συνεισφορά του παίκτη j στο υποσύνολο S και $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ το βάρος της συνεισφοράς του παίκτη j στο υποσύνολο S . Η άθροιση γίνεται σε όλα τα υποσύνολα S χωρίς τον παίκτη j .

Λόγω της πολυπλοκότητας της Σχέσης 9, έχουν προταθεί νέοι τρόποι υπολογισμού των τιμών «Shapley» των μεταβλητών ενός μοντέλου μηχανικής μάθησης με στόχο την ελαχιστοποίηση του υπολογιστικού κόστους. Συγκεκριμένα για μοντέλα που βασίζονται στην μεθοδολογία των «δέντρων απόφασης», όπως το XGBoost, οι ακριβείς τιμές «Shapley» μπορούν να υπολογιστούν προσδιορίζοντας «τοπικές εξηγήσεις» (local explanations) και ανάγοντάς αυτές σε γενικές (global) εξηγήσεις της συμπεριφοράς του μοντέλου⁹¹.

Η απλούστευση αυτή των υπολογισμών καθιστά δυνατή την ευρεία χρήση της ανάλυσης SHAP για την ερμηνεία των μοντέλων μηχανικής μάθησης. Μάλιστα, έχουν δημιουργηθεί εύχρηστα «πακέτα» λογισμικού της γλώσσας προγραμματισμού Python που επιτρέπουν την δημιουργία επεξηγηματικών διαγραμμάτων βάση των τιμών «Shapley». Συγκεκριμένα, μπορεί να δημιουργηθεί ένα ραβδόγραμμα με τον μέσο όρο των απόλυτων τιμών «Shapley» κάθε μεταβλητής (στις μονάδες μέτρησης της ίδιας της μεταβλητής) που αντικατοπτρίζει την συνεισφορά τους στις τελικές προβλέψεις. Καθώς αυτό το διάγραμμα δεν δείχνει την κατεύθυνση της συνεισφοράς κάθε μεταβλητής, εξαιρετικά χρήσιμο είναι το διάγραμμα «σμήνους» (beeswarm) στο οποίο παρουσιάζονται οι τιμές SHAP κάθε παρατήρησης του συνόλου ελέγχου για τις μεταβλητές με την πιο ισχυρή συνεισφορά (μία μεταβλητή ανά γραμμή στον άξονα y'). Το χρώμα των σημείων είναι ένα δείγμα της τιμής κάθε παρατήρησης (μπλε για μικρές τιμές, κόκκινο για μεγάλες τιμές), ενώ η θέση τους στον άξονα x' δείχνει την κατεύθυνση της συνεισφοράς (όσο πιο αριστερά τόσο πιο αρνητική είναι η συνεισφορά της παρατήρησης και αντίστροφα). Τέλος, η συνεισφορά των μεταβλητών σε μεμονωμένες προβλέψεις μπορεί να μελετηθεί μέσω ενός διαγράμματος «καταρράκτη» (waterfall plot) στο οποίο παρουσιάζεται τόσο η απόλυτη τιμή της συνεισφοράς όσο και η κατεύθυνση της για κάθε μεταβλητή⁹².

4. Συλλογή και διαχείριση δεδομένων

Στο παρόν κεφάλαιο συζητείται η διαδικασία συλλογής και αξιολόγησης πειραματικών ερευνών που σχετίζονται με την κυτταρική πρόσληψη νανοσωματιδίων. Στη συνέχεια περιγράφεται η καταγραφή και η επεξεργασία των δεδομένων ώστε να αποκτήσουν κοινές μονάδες μέτρησης, να μειωθεί ο αριθμός κατηγορικών μεταβλητών και να συμπληρωθούν οι ελλιπείς τιμές.

4.1. Συλλογή πειραματικών ερευνών

Η συλλογή πειραματικών δεδομένων από διαφορετικές ερευνητικές μελέτες με σκοπό την καταγραφή τους σε ένα κοινό σύνολο δεδομένων απαιτεί μεθοδικότητα. Το διάγραμμα ροής που περιγράφει τη διαδικασία της συλλογής δεδομένων στη συγκεκριμένη εργασία παρουσιάζεται στο Σχήμα 3. Αρχικά δημιουργήθηκε ένα σύνθετο αίτημα αναζήτησης (query) που περιλαμβάνει βασικές λέξεις-κλειδιά για την αναζήτηση επιστημονικών άρθρων με θέμα την πρόσληψη νανοϋλικών σε κύτταρα, όπως «nanoparticle», «NP», «cellular uptake», «cell internalization», «endocytosis», «uptake kinetics», «quantification», «measurement». Το πλήρες αίτημα αναζήτησης παρουσιάζεται στην Εικόνα 3.

```
(nanoparticle OR nanoparticles OR NP OR NPs OR nanomaterial OR nanomaterials OR nanostructure OR nanostructures OR nanocapsule OR nanocapsules OR nanocrystal OR nanocrystals OR nanofiber OR nanofibers OR nanorod OR nanorods OR nanoshell OR nanoshells OR nanosphere OR nanospheres OR nanotube OR nanotubes OR nanopowder OR nanocomposite OR nanocomposites OR nanophase OR nanophases OR nanocluster OR nanoclusters OR "gold nanoparticle" OR "gold nanoparticles" OR "silver nanoparticle" OR "quantum dot" OR "quantum dots" OR "carbon nanotube" OR "carbon nanotubes" OR "liposome" OR "liposomes" OR "dendrimer" OR "dendrimers" OR "polymeric nanoparticle" OR "polymeric nanoparticles" OR "magnetic nanoparticle" OR "magnetic nanoparticles" OR "metal oxide nanoparticle" OR "metal oxide nanoparticles" OR "mesoporous nanoparticle" OR "mesoporous nanoparticles")
```

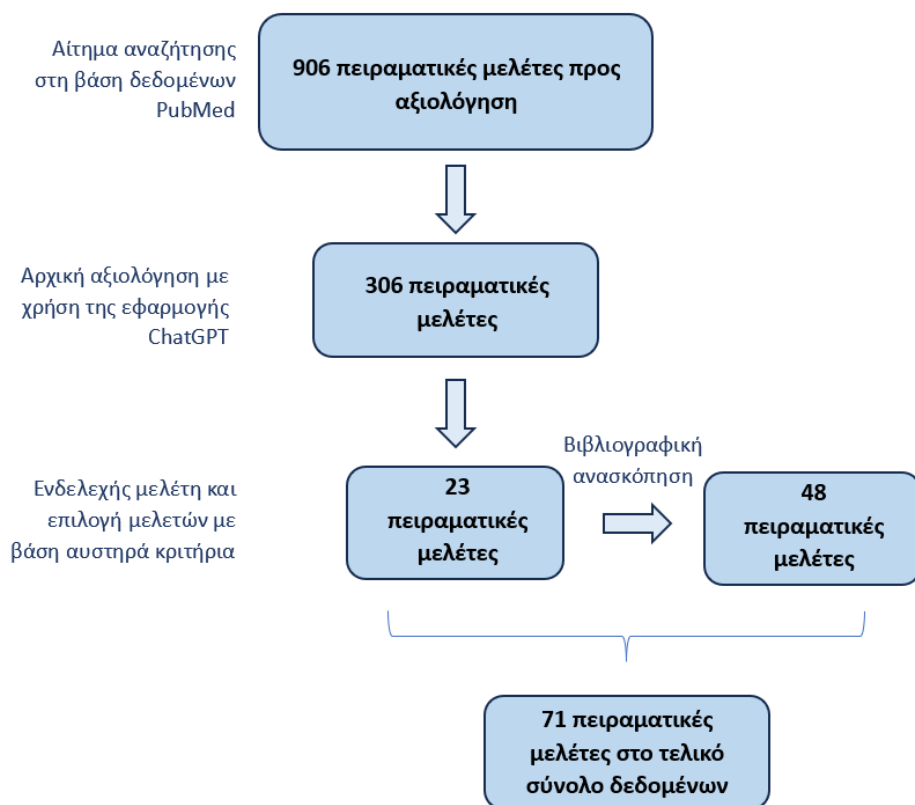
```
AND (cellular uptake OR cell uptake OR cellular internalization OR cell internalization OR cellular absorption OR cell absorption OR cellular ingestion OR cell ingestion OR endocytosis OR phagocytosis OR pinocytosis OR cellular transport OR cell transport OR intracellular uptake OR intracellular transport OR cytoplasmic uptake OR cytoplasmic transport OR cellular assimilation OR cell assimilation OR "receptor-mediated uptake" OR "receptor-mediated endocytosis" OR macropinocytosis OR "clathrin-mediated endocytosis" OR "caveolae-mediated uptake" OR "membrane transport")
```

```
AND (rate OR rate constant OR "rate of reaction" OR "rate parameter" OR "temporal rate" OR "reaction kinetics" OR "uptake kinetics" OR "absorption rate" OR "internalization rate" OR "ingestion rate" OR "uptake efficiency" OR "uptake capacity" OR "kinetic analysis" OR "time-dependent uptake")
```

```
AND ("quantitative analysis" OR quantification OR measurement OR "quantitative assessment" OR "quantitative evaluation")
```

Εικόνα 3. Σύνθετο αίτημα αναζήτησης σχετικών ερευνών στη βάση δεδομένων «PubMed».

Στη συνέχεια, αναπτύχθηκε ένα απλό πρόγραμμα σε Python για την αποστολή του αιτήματος στη βάση δεδομένων «PubMed»⁹³ και την καταγραφή των αποτελεσμάτων της αναζήτησης. Συγκεκριμένα, επιλέχθηκε να εμφανιστούν έως 1500 δημοσιεύσεις από 01/01/2000 έως 31/12/2024 σε φθίνουσα σειρά σχετικότητας με τις λέξεις-κλειδιά και να καταγραφούν τα στοιχεία τους (DOI, τίτλος, περίληψη και έτος δημοσίευσης) σε ένα αρχείο «Excel». Παράλληλα, οι περιλήψεις όλων των επιστημονικών άρθρων αριθμήθηκαν και καταγράφηκαν σε ένα ξεχωριστό αρχείο. Μέσω αυτής της αναζήτησης καταγράφηκαν συνολικά 906 άρθρα προς αξιολόγηση.



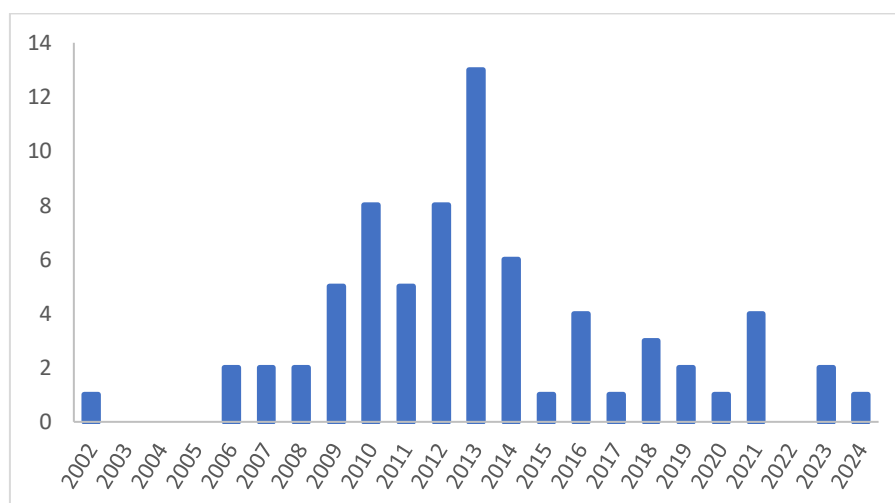
Σχήμα 3. Διάγραμμα ροής της διαδικασίας συλλογής και αξιολόγησης πειραματικών ερευνών για την δημιουργία του συνόλου δεδομένων.

4.2. Αξιολόγηση και επιλογή δεδομένων

Για τον ταχύ και αποδοτικό αποκλεισμό μη σχετικών μελετών οι 906 περιλήψεις χωρίστηκαν ανά περίπου 50 και δημιουργήθηκε μία προτροπή (prompt) σύμφωνα με την οποία η εφαρμογή «ChatGPT» της OpenAI⁹⁴ κλήθηκε να επιλέξει σε κάθε περίπτωση τις 20 πιο σχετικές με την πρόσληψη νανοϋλικών σε κύτταρα περιλήψεις. Βασικό κριτήριο αυτής της επιλογής ζητήθηκε να είναι η ύπαρξη ποσοτικών δεδομένων της πρόσληψης των νανοϋλικών. Η ακριβής προτροπή που χρησιμοποιήθηκε ήταν η εξής: «Here is a PDF containing 50 abstracts. Please manually search the PDF and, depending on the abstracts, give me the 20 most likely articles to contain quantitative information about nanoparticles uptake rates or nanoparticle concentration with time.».

Δεδομένης της φθίνουσας σχετικότητας των άρθρων με τις λέξεις-κλειδιά, συχνά η εφαρμογή επέλεγε λιγότερες από 20 περιλήψεις με αποτέλεσμα ο τελικός αριθμός επιστημονικών άρθρων να μειωθεί στα 306.

Οι μελέτες αυτές αναλύθηκαν ενδελεχώς και αξιολογήθηκαν με γνώμονα την ύπαρξη ρητών αναφορών σχετικά με το είδος και το μέγεθος των νανοσωματιδίων, την αρχική συγκέντρωση αυτών στο μέσο καλλιέργειας των κυττάρων, τον χρόνο διεξαγωγής του πειράματος, το είδος της κυτταρικής σειράς και τις συνθήκες (θερμοκρασία, ποσοστό CO₂, θρεπτικό μέσο) της καλλιέργειας. Τέλος, καθοριστικής σημασίας ήταν η διαθεσιμότητα ποσοτικών δεδομένων για την ύπαρξη νανοσωματιδίων στα κύτταρα (π.χ. NPs/cell) καθώς αρκετοί μελετητές αρκούνται στον ποιοτικό σχολιασμό της πρόσληψης νανοϋλικών στα κύτταρα^{95,96}. Έτσι, μόλις 23 έρευνες πληρούσαν όλες τις προϋποθέσεις για την δημιουργία του συνόλου δεδομένων. Οι βιβλιογραφικές αναφορές αυτών των άρθρων παρέπεμψαν σε ακόμη 48 έρευνες που πληρούσαν τις απαραίτητες προϋποθέσεις επιλογής, αυξάνοντας τελικά τον τελικό αριθμό σε 71 άρθρα που δημοσιεύτηκαν μεταξύ των ετών 2002 και 2024 όπως φαίνεται στο Διάγραμμα 2.



Διάγραμμα 2. Αριθμός μελετών ανά έτος που περιλαμβάνονται στο σύνολο δεδομένων.

4.3. Δημιουργία συνόλου δεδομένων

Για την δημιουργία του συνόλου δεδομένων επιλέχθηκαν ως μεταβλητές εισόδου βασικές φυσικοχημικές παράμετροι των νανοσωματιδίων όπως το Είδος, το Σχήμα, η Ονομαστική και η Υδροδυναμική διάμετρος, η Ειδική επιφάνεια, ο Δείκτης πολυδιασποράς, το Ζ-δυναμικό και η Επικάλυψη των νανοϋλικών. Επίσης, μεταβλητές εισόδου αποτελούν και οι πειραματικές συνθήκες, όπως δηλαδή η Χρονική διάρκεια του πειράματος, η Συγκέντρωση των νανοϋλικών στην καλλιέργεια των κυττάρων, το Θρεπτικό μέσο της καλλιέργειας, η Θερμοκρασία, η Εφαρμογή υπερήχων (Ναι/Όχι), το Είδος κυτταρικής σειράς και ο Αριθμός των κυττάρων. Βεβαίως, μεταβλητή εξόδου του προβλήματος αποτελεί η Συγκέντρωση των νανοϋλικών ανά κύτταρο.

Εφόσον στόχος ήταν η εξασφάλιση της δημιουργίας ενός εκτενούς συνόλου δεδομένων υψηλής ποιότητας, καταγράφηκαν αρχικά οι πρωτογενείς τιμές συνοδευόμενες από τις μονάδες μέτρησης που επέλεξε η κάθε ερευνητική ομάδα, ώστε να αποφευχθούν λάθη στην κωδικοποίηση των δεδομένων. Τα πρωτογενή πειραματικά δεδομένα δεν ήταν πάντοτε διαθέσιμα σε μορφή πίνακα και έτσι οι ακριβείς τιμές των σημείων των διαγραμμάτων προσδιορίστηκαν με χρήση του ελεύθερου λογισμικού «WebPlotDigitizer»⁹⁷. Επίσης, έγινε η παραδοχή ότι στα δείγματα ελέγχου (μηδενική χρονική διάρκεια πειράματος ή αρχική συγκέντρωση νανοϋλικών) αντιστοιχεί συγκέντρωση νανοσωματιδίων ανά κύτταρο ίση με 0.

Ως Ονομαστική διάμετρος ορίστηκε η διάμετρος που δηλώνεται από τον κατασκευαστή ή η διάμετρος που μετρήθηκε μέσω TEM, ενώ η Υδροδυναμική διάμετρος αναφερόταν αποκλειστικά σε μετρήσεις DLS. Σε περιπτώσεις που δινόταν εύρος αριθμού κυττάρων καλλιέργειας ή μεγέθους νανοϋλικών καταγραφόταν ο μέσος όρος. Τέλος, ως πυκνότητα των νανοσωματιδίων θεωρήθηκε η πυκνότητα του υλικού από το οποίο είναι κατασκευασμένα, θεωρώντας αμελητέα την πυκνότητα της επικάλυψης ή των επιφανειακά συνδεδεμένων μορίων. Η τυπική πυκνότητα για το υλικό κατασκευής βρέθηκε από τις βάσεις δεδομένων «PubChem»⁹⁸ και «MatWeb»⁹⁹, αλλά και από τις πληροφορίες του κατασκευαστή, όταν αυτές ήταν διαθέσιμες. Σημειώνεται ότι η πυκνότητα νανοϋλικών που αποτελούνταν από δύο διαφορετικά υλικά, π.χ. Superparamagnetic Iron Oxide (SPIONS), Fe₃O₄ και Fe₂O₃¹⁰⁰, υπολογίστηκε με βάση δεδομένη αναλογία mole των δύο υλικών. Επίσης, η πυκνότητα κενών σφαιρών με εσωτερική (r_{in}) και εξωτερική (r_{out}) διάμετρο υπολογίστηκε με βάση την ονομαστική πυκνότητα του υλικού κατασκευής (d_{bulk}) ως εξής:

$$d_{hollow} = \frac{m_{hollow}}{V_{total}} = \frac{d_{bulk} * \frac{4}{3}\pi(r_{out}^3 - r_{in}^3)}{\frac{4}{3}\pi r_{out}^3} \approx d_{bulk} \left(1 - \frac{r_{in}^3}{r_{out}^3}\right) \quad (10)$$

Έτσι, δημιουργήθηκε ένα σύνολο πρωτογενών δεδομένων με τις εξής μεταβλητές εισόδου:

- Μεταδεδομένα (2 στήλες): Κωδικοποίηση της έρευνας και DOI.
- Φυσικοχημικές ιδιότητες νανοϋλικών (17 στήλες): Νανοσωματίδιο, Σχήμα, Πυκνότητα, Ονομαστική διάμετρος, Μήκος και Διάμετρος (για μη σφαιρικά σωματίδια), Λόγος διαστάσεων, Ονομαστική ειδική επιφάνεια, Μέση υδροδυναμική διάμετρος, Συγκέντρωση νανοσωματιδίων κατά την μέτρηση της μέσης υδροδυναμικής διαμέτρου, Μονάδες μέτρησης συγκέντρωσης νανοσωματιδίων κατά την μέτρηση της μέσης υδροδυναμικής διαμέτρου, Μέσο μέτρησης μέσης υδροδυναμικής διαμέτρου, Μέθοδος μέτρησης μέσης υδροδυναμικής διαμέτρου, Z-δυναμικό, Μέσο μέτρησης Z-δυναμικού, Επικάλυψη νανοσωματιδίων.
- Πειραματικές συνθήκες (31 στήλες): Χρονική διάρκεια πειράματος και πρωτογενείς μονάδες μέτρησης, Συγκέντρωση νανοσωματιδίων στην καλλιέργεια κυττάρων και πρωτογενείς μονάδες μέτρησης, Είδος μέσου καλλιέργειας, Θερμοκρασία, Ποσοστό CO₂ και O₂, Ποσοστό Ορού Βόειου Εμβρύου (Fetal

Bovine Serum, FBS), Πενικιλίνη, Στρεπτομυκίνη και L-Γλουταμίνη με τις αντίστοιχες μονάδες μέτρησης συγκέντρωσης, Ποσοστό ορού αλόγου, Συγκέντρωση BSA, Διασπορά νανοσωματιδίων σε διάλυμα (Ναι/Όχι), Συγκέντρωση διασποράς και αντίστοιχες μονάδες μέτρησης, Εφαρμογή υπερήχων (Ναι/Όχι), Χρόνος εφαρμογής υπερήχων με τις πρωτογενείς μονάδες μέτρησης, Ισχύς και Συχνότητα υπερήχων, Είδος κυτταρικής σειράς, Συγκέντρωση κυττάρων και μονάδες μέτρησης συγκέντρωσης κυττάρων, αλλά και Μέθοδος ποσοτικοποίησης νανοσωματιδίων στα κύτταρα.

Η μοναδική μεταβλητή εξόδου είναι η Συγκέντρωση των νανοϋλικών ανά κύτταρο με τις αντίστοιχες μονάδες μέτρησης.

4.4. Επεξεργασία των δεδομένων

Σκοπός της επεξεργασίας των πρωτογενών δεδομένων ήταν αρχικά η μετατροπή τους σε συγκρίσιμη μορφή μέσω της επιλογής κοινών μονάδων μέτρησης για κάθε μεταβλητή του συνόλου δεδομένων. Επίσης, σε αυτό το στάδιο πραγματοποιήθηκε μία πρωταρχική απομάκρυνση μεταβλητών που δεν προσφέρουν σημαντικές πληροφορίες.

4.4.1. Αφαίρεση μεταβλητών

Αρχικά, αφαιρέθηκαν μεταβλητές που περιέχουν δεδομένα μόνο από μία έρευνα, δηλαδή το Ποσοστό O₂ στο μέσο της καλλιέργειας, η Ισχύς και η Συχνότητα υπερήχων. Ομοίως, αφαιρέθηκαν στήλες με πολύ μεγάλο ποσοστό έλλειψης δεδομένων (>60%), όπως η Διασπορά νανοσωματιδίων σε διάλυμα, η Συγκέντρωση της διασποράς, ο Χρόνος εφαρμογής υπερήχων, ο Λόγος των διαστάσεων των νανοσωματιδίων, ο Δείκτης πολυδιασποράς, η Ονομαστική ειδική επιφάνεια, η Συγκέντρωση νανοσωματιδίων κατά την μέτρηση της μέσης υδροδυναμικής διαμέτρου, το Μέσο μέτρησης μέσης υδροδυναμικής διαμέτρου, το Μέσο μέτρησης Z-δυναμικού, η Συγκέντρωση L-Γλουταμίνης, BSA και ορού αλόγου, όπως και οι αντίστοιχες μονάδες μέτρησης. Η Μέθοδος μέτρησης μέσης υδροδυναμικής διαμέτρου αφαιρέθηκε επίσης καθώς σε όλες τις μελέτες χρησιμοποιήθηκε η μέθοδος DLS. Τέλος, αφαιρέθηκαν οι μεταβλητές που εμφανίζουν μηδενική ή σχεδόν μηδενική διασπορά, δηλαδή η Θερμοκρασία και το Ποσοστό διοξειδίου του άνθρακα, %CO₂.

Για τα νανοσωματίδια μη σφαιρικού σχήματος έγιναν ορισμένες παραδοχές ώστε να υπολογιστεί μία φαινόμενη Ονομαστική διάμετρος. Συγκεκριμένα, τα ραβδόμορφα και σωληνοειδή σωματίδια θεωρήθηκε ότι έχουν Ονομαστική διάμετρο ίση με τον μέσο όρο του Μήκους και της Διαμέτρου τους (Σχέση 11). Η παραδοχή αυτή εξασφαλίζει την κάλυψη όλων των διαφορετικών περιπτώσεων αρχικού προσανατολισμού του νανοσωματιδίου κατά την προσέγγιση της κυτταρικής μεμβράνης.

$$\text{Nominal Size} = \frac{\text{Length} + \text{Diameter}}{2} \quad (11)$$

Τα κυβικά νανοσωματίδια θεωρήθηκε ότι έχουν Ονομαστική διάμετρο περίπου ίση με την ακμή τους. Έτσι, αφού υπολογίστηκε η Ονομαστική διάμετρος και των μη σφαιρικών σωματιδίων αφαιρέθηκαν οι στήλες Μήκους και Διαμέτρου.

Στη συνέχεια, η προσθήκη Πενικιλίνης και Στρεπτομυκίνης στο μέσο καλλιέργειας κωδικοποιήθηκε σε μία κοινή στήλη ως 1 ή 0 (δυαδική κωδικοποίηση 1 = Ναι, 0 = Όχι) καθώς τα δύο αυτά αντιβιοτικά χρησιμοποιούνταν σε όλες τις μελέτες από κοινού, ενώ συχνά δεν υπήρχαν ποσοτικά δεδομένα για την συγκέντρωσή τους. Σημειώνεται πως τόσο για την κοινή πλέον μεταβλητή Πενικιλίνη/Στρεπτομυκίνη (Ναι/Όχι), όσο και για την Εφαρμογή υπερήχων (Ναι/Όχι) και την Επικάλυψη των νανοσωματιδίων, θεωρήθηκε ότι δεν έχουν χρησιμοποιηθεί σε όσα πειράματα δεν αναφέρονται ρητά, οπότε η αντίστοιχη τιμή καταγραφής ήταν το μηδέν.

Τέλος, οι μεταβλητές της Πυκνότητας και της Συγκέντρωσης κυττάρων χρησιμοποιήθηκαν για τις απαραίτητες μετατροπές μονάδων μέτρησης αλλά αφαιρέθηκαν από το τελικό σύνολο δεδομένων καθώς η πρώτη περιείχε αρκετές παραδοχές, ενώ για την δεύτερη συχνά δεν ήταν διαθέσιμες οι απαραίτητες πληροφορίες για την μετατροπή της σε συνεπείς μονάδες μέτρησης.

4.4.2. Εμπλουτισμός συνόλου δεδομένων με δευτερογενείς μεταβλητές

Το πρωτογενές σύνολο δεδομένων περιέχει τόσο αριθμητικές όσο και κατηγορικές μεταβλητές οι οποίες, λόγω της ποικιλομορφίας των δεδομένων που συλλέχθηκαν από τις πειραματικές μελέτες, περιέχουν πολλαπλές κατηγορίες. Προκειμένου να αποφευχθεί ο κίνδυνος υπερπροσαρμογής («overfitting»)¹⁰¹ των μοντέλων μηχανικής μάθησης που θα χρησιμοποιηθούν στη συνέχεια, αλλά και να απλουστευθεί το περίπλοκο σύνολο δεδομένων χωρίς να χαθούν σημαντικές πληροφορίες, κρίθηκε απαραίτητη η μείωση του αριθμού των διαφορετικών κατηγοριών κάθε κατηγορικής μεταβλητής ή η μετατροπή της σε αριθμητική μεταβλητή.

Πιο συγκεκριμένα, για τις κατηγορικές μεταβλητές με τις περισσότερες διαφορετικές παρατηρήσεις, δηλαδή τη μεταβλητή Νανοσωματίδιο και την Μέθοδο ποσοτικοποίησης νανοσωματιδίων στα κύτταρα, πραγματοποιήθηκε περαιτέρω ομαδοποίηση των διαφορετικών ειδών νανοϋλικών και μεθόδων ποσοτικοποίησης αντίστοιχα. Τα νανοσωματίδια κατηγοριοποιήθηκαν σύμφωνα με το είδος τους ως Πυριτίου, Άνθρακα, Μεταλλικά, Ημιαγώγιμα, Οργανικά και Ανιούσας μετατροπής (urconversion), ενώ οι μέθοδοι ποσοτικοποίησης ως Φασματοσκοπίες μάζας, Μαγνητικές, Μικροσκοπικές και Οπτικές όπως φαίνεται πιο αναλυτικά στον Πίνακα 1. Ομοίως, τα διαφορετικά μόρια επικάλυψης που έχουν καταγραφεί στην μεταβλητή Επικάλυψη νανοσωματιδίων κατηγοριοποιήθηκαν με βάση την χαρακτηριστική τους ομάδα ως Καμία, Αμίνη, Καρβοξύλιο, Βιολογικό μόριο ή Άλλη.

Πίνακας 1. Επεξήγηση δευτερογενών μεταβλητών που προκύπτουν από την ομαδοποίηση των κατηγοριών των μεταβλητών Νανοδομικό, Μέθοδος ποσοτικοποίησης και Επικάλυψη νανοδομικών.

Πρωτογενής μεταβλητή	Κατηγορίες ομαδοποίησης	Παραδείγματα κατηγορικών τιμών
Είδος νανοδομικών	Πυριτίου	SiNP-R6G, SiO ₂ , Au/Si
	Άνθρακα	Φθορίζοντα νανοδιαμάντια (FND), Μονοστρωματικοί σωλήνες άνθρακα (SWNT)
	Μεταλλικά	Μαγνητικά νανοδομικά (MNP), MNPs@SiO ₂ (RITC), Au, Fe ₃ O ₄ @PS, SPION, TiO ₂ , Ag, Fe ₂ O ₃ , CeO ₂
	Ημιαγώγιμα	CdSe/ZnS κβαντικές τελείες (Qd), CdSe/CdS QD, Si QD
	Οργανικά	PS, HSPC λιποσώματα, C20-5 λιποσώματα, PMMA, PCL, PLGA
	Ανιούσας μετατροπής	UCNPs
Μέθοδος ποσοτικοποίησης νανοδομικών στα κύτταρα	Φασματοσκοπίες μάζας	ICP-MS, SP-ICP-MS, SC-ICP-MS, Digestion ICP-MS
	Μαγνητικές	MRI, Magnetophoresis, Ferromagnetic resonance
	Μικροσκοπικές	TEM, SEM, Confocal scanning microscopy, Dark Field Microscopy
	Οπτικές	dSTORM imaging, ICP-AES, 3D live-cell imaging, AAS, XRF, UV-vis, Fluorescence spectroscopy, ICP-OES
Επικάλυψη νανοδομικών	Καμία	-
	Αμίνη	Π.χ. PEG-NH ₃ , PEG-NH ₂ , αλλυλαμίνη
	Καρβοξύλιο	Π.χ. κιτρικό οξύ, PEG-COO,
	Βιολογικό μόριο	Π.χ. PEG-TAT, τρανσφερίνη, αλβουμίνη ορού μόσχου (BSA), dsRNA, DNA
	Άλλη	CTAB, PSS-CTAB, PEG, FITC, δεξτράνη, άμυλο

Παράλληλα, η κατηγορική μεταβλητή που αφορά το Μέσο καλλιέργειας των κυττάρων αντικαταστάθηκε από την αριθμητική μεταβλητή Ιοντική ισχύς (mol/L)¹⁰², θεωρώντας ότι αυτή η ιδιότητα είναι η πιο αντιπροσωπευτική για την σύσταση και τα χαρακτηριστικά του κάθε μέσου. Καθώς δεν ήταν δυνατόν να βρεθούν πληροφορίες για την ιοντική ισχύ (IS) των μέσων καλλιέργειας κυττάρων, συλλέχθηκαν πληροφορίες για την σύσταση των μέσων από τον κατασκευαστή¹⁰³ και η επιθυμητή ιδιότητα υπολογίστηκε ως εξής:

$$IS = \frac{1}{2} \sum C_i Z_i^2 \quad (12)$$

Όπου C_i η μοριακή συγκέντρωση και Z_i το καθαρό φορτίο του ιόντος i αντίστοιχα.

Από την Σχέση 12, υπολογίζοντας ξεχωριστά την συνεισφορά των κατιόντων και των ανιόντων κάθε ιοντικής ένωσης που περιέχει το μίγμα (1 έως n), η ιοντική ισχύς υπολογίζεται τελικά ως:

$$IS = \frac{1}{2} \sum_1^n \left[(\text{αριθμός}_{\text{κατιόντων}}) * (\text{φορτίο}_{\text{κατιόντος}})^2 + (\text{αριθμός}_{\text{ανιόντων}}) * (\text{φορτίο}_{\text{ανιόντος}})^2 \right] * \text{moles}_{\text{ιοντικής ένωσης}} \quad (13)$$

Στον Πίνακα 2 παρουσιάζονται οι τιμές της ιοντικής ισχύος που υπολογίστηκαν για κάθε μέσο κυτταρικής καλλιέργειας που περιλαμβάνεται στο σύνολο δεδομένων. Σημειώνεται ότι για τα μέσα καλλιέργειας ενδοθηλιακών κυττάρων (ECM, ECGM) η σύσταση δεν ήταν διαθέσιμη και έτσι οι υπολογισμοί της ιοντικής ισχύος βασίστηκαν στη σύσταση του Ενδοθηλιακού βασικού μέσου (Endothelial basal medium)¹⁰⁴ το οποίο αποτελεί κύριο συστατικό τους.

Πίνακας 2. Ιοντική ισχύς των μέσων καλλιέργειας κυττάρων.

Μέσο καλλιέργειας κυττάρων	Ιοντική ισχύς (mol/L)
DMEM	0.1684829
RPMI 1640	0.1516472
EMEM	0.1300679
MEM	0.1575791
F12	0.1497732
DMEM/F12	0.159128
B27	0.09164053
McCoy's	0.152246
MCDB131	0.1743051
TCM199	0.1575895
Basal	0.1603358

Τέλος, από την μεταβλητή Κυτταρική σειρά προέκυψαν 4 δευτερογενείς μεταβλητές που κωδικοποιούν την Κυτταρική κατηγορία (π.χ. καρκινικό, πρωτογενές), το Είδος προέλευσης του κυττάρου (π.χ. ανθρώπινο, ποντικού), τον Ιστό προέλευσης του κυττάρου (π.χ. εγκέφαλος, μήτρα) και την Μορφολογία του κυττάρου (π.χ. μακροφάγο, επιθηλιακό). Εξ αυτών, η μεταβλητή Ιστός προέλευσης κυττάρου που περιέχει πολλαπλές κατηγορίες δεν συμπεριλήφθηκε στο τελικό σύνολο δεδομένων καθώς θεωρείται ότι οι υπόλοιπες μεταβλητές, δηλαδή η Κυτταρική κατηγορία, το Είδος και η Μορφολογία του κυττάρου, κωδικοποιούν με απλό τρόπο σχεδόν το σύνολο της πληροφορίας της αρχικής μεταβλητής. Οι παραπάνω πληροφορίες, όπως φαίνονται στον Πίνακα 3, συλλέχθηκαν από τη βάση δεδομένων «Cellosaurus»¹⁰⁵ και όπου ήταν δυνατόν διασταυρώθηκαν με πληροφορίες του επίσημου ιστοτόπου του κατασκευαστή «ATCC»¹⁰⁶.

Πίνακας 3. Δευτερογενή δεδομένα σχετικά με την κατηγορία, το είδος προέλευσης και την μορφολογία των κυττάρων που περιλαμβάνονται στο πρωτογενές σύνολο δεδομένων.

Παραδείγματα κυτταρικών σειρών		
Κατηγορία κυττάρων	Καρκινικά	Π.χ. RAW 264.7, HeLa, PC12, MCF-7, A549, HEPG2, MiaPaCa2, FaDu
	Αθάνατα	Π.χ. NIH3T3, C17.2, hCMEC/D3, MDCK, CHO-K1
	Πρωτογενή	Π.χ. HUVEC, coAEC, 1-BEC, RAECM, MPMC, Hippocampal neurons, MSC, BMDC
	Μετασηματισμένα	Π.χ. HEK293, BMEC, 16HBE14o, HMEC-1
Είδος προέλευσης κυττάρων	Αρουραίος	Π.χ. PC12, RAECM, NRK
	Άνθρωπος	Π.χ. HEK293, HeLa, HUVEC, hCMEC/D3, coAEC, 1-BEC, BMEC, A549
	Ποντίκι	Π.χ. RAW 264.7, NIH3T3, C17.2, Mouse macrophages
	Άλλο	Π.χ. Pk15, CHO-K1, MDCK
Μορφολογία κυττάρων	Νευρικά	Π.χ. PC12, C17.2, Hippocampal neurons, Neuro-2a
	Ενδοθηλιακά	Π.χ. HUVEC, coAEC, BMEC
	Μακροφάγα	Π.χ. RAW 264.7, Mouse macrophages, J774 A1
	Ινοβλάστες	Π.χ. NIH3T3, BMDC, 3T3, CF-31, 3T3-L1
	Επιθηλιακά	Π.χ. HEK293, HeLa, MCF-7, A549, DU145, MCF7, HEPG2, MiaPaCa2, RAECM
	Άλλη	Π.χ. PC3, THP-1 monocytes, MPMC, cumulus-oocyte

4.4.3. Μετατροπή μονάδων μέτρησης

Όπως προαναφέρθηκε, απαραίτητη προϋπόθεση για την συνοχή του συνόλου δεδομένων είναι η καταγραφή των πειραματικών μετρήσεων χρησιμοποιώντας κοινές μονάδες μέτρησης. Συνεπώς, οι μεταβλητές που παρουσιάζονται με περισσότερες από μία μονάδες μέτρησης, δηλαδή η Χρονική διάρκεια του πειράματος, η Συγκέντρωση νανοϋλικών στο μέσο καλλιέργειας και η μεταβλητή εξόδου (Συγκέντρωση νανοϋλικών στα κύτταρα), επιλέχθηκε να μετατραπούν στις πιο κοινές μονάδες μέτρησης μεταξύ των δεδομένων, hours, nM και NPs/cell αντίστοιχα.

Αρχικά, έγιναν χειροκίνητα ορισμένες μετατροπές για μεμονωμένες περιπτώσεις μονάδων μέτρησης. Συγκεκριμένα, για την Συγκέντρωση νανοϋλικών στο μέσο καλλιέργειας ζητούμενο ήταν η έκφραση της ποσότητας ανά όγκο μέσου καλλιέργειας. Έτσι, όπου δινόταν ο αριθμός των νανοσωματιδίων στο μέσο καλλιέργειας αυτός διαιρέθηκε με τον δεδομένο όγκο, ενώ ο αριθμός νανοσωματιδίων ανά επιφάνεια φρεατίου κυτταρικής καλλιέργειας («well») ανάχθηκε σε αριθμό νανοσωματιδίων ανά όγκο θεωρώντας μία τυπική τιμή επιφάνειας φρεατίου $9.6 \text{ cm}^2/\text{well}^{107}$ (Σχέση 14).

$$\frac{\frac{\text{Αριθμός NPs}}{\text{cm}^2} * 9.6 \frac{\text{cm}^2}{\text{well}}}{\text{Όγκος μέσου καλλιέργειας} \frac{\text{ml}}{\text{well}}} \rightarrow \frac{\text{NPs}}{\text{ml}} \quad (14)$$

Όσον αφορά την Συγκέντρωση νανοσωματιδίων στα κύτταρα, επιλέχθηκε να εκφραστεί η ποσότητα νανοσωματιδίων ανά ένα κύτταρο. Η ποσότητα (αριθμός ή μάζα) νανοσωματιδίων ανά η κύτταρα διαιρέθηκε με τον αριθμό (n) των κυττάρων που μετρήθηκαν για να βρεθεί κατά μέσο όρο η ποσότητα νανοσωματιδίων ανά ένα κύτταρο. Επίσης, η ποσότητα νανοϋλικών διαιρέθηκε με τον αριθμό κυττάρων καλλιέργειας, ενώ ο αριθμός των νανοσωματιδίων ανά φρεάτιο καλλιέργειας διαιρέθηκε με τον δεδομένο αριθμό κυττάρων ανά φρεάτιο (Σχέση 15). Ακόμη, στην περίπτωση που διαθέσιμος ήταν μόνο ο αριθμός νανοσωματιδίων ανά κενοτόπιο (vacuole), αυτός πολλαπλασιάστηκε με τον αριθμό κενοτοπίων ανά κύτταρο (Σχέση 16). Στη συνέχεια, όσον αφορά τα νανοϋλικά τύπου SPIONS, η μάζα σιδήρου ανά κύτταρο που μετρήθηκε κατά την επώαση κυττάρων παρουσία των SPIONS, ανάχθηκε σε μάζα νανοσωματιδίων ανά κύτταρο με βάση δεδομένη αναλογία. Τέλος, η συγκέντρωση νανοσωματιδίων σε nM διαιρέθηκε με τον αριθμό κυττάρων ανά ml ώστε να προκύψουν τα mol νανοσωματιδίων ανά κύτταρο (Σχέση 17), ενώ στην περίπτωση που διαθέσιμος ήταν ο ρυθμός πρόσληψης νανοσωματιδίων, αυτός ανάχθηκε σε μάζα νανοσωματιδίων ανά κύτταρο κατά τον τελικό χρόνο διεξαγωγής του πειράματος επώασης (Σχέση 18).

$$\frac{\frac{\text{NPs}}{\text{well}}}{\frac{\text{cells}}{\text{well}}} \rightarrow \frac{\text{NPs}}{\text{cell}} \quad (15)$$

$$\frac{\text{NPs}}{\text{vacuole}} * \frac{\text{vacuole}}{\text{cell}} \rightarrow \frac{\text{NPs}}{\text{cell}} \quad (16)$$

$$\frac{\frac{\text{nM} * 10^{-3}}{\text{ml}}}{\frac{\text{cells}}{\text{ml}}} \rightarrow \frac{\text{nmol}}{\text{cell}} \quad (17)$$

$$\frac{\frac{\text{pg}}{\text{cm}^2 * \text{sec}}}{\frac{\text{cells}}{\text{cm}^2}} * \text{Time}(\text{sec}) \rightarrow \frac{\text{pg}}{\text{cell}} \quad (18)$$

Το δεύτερο στάδιο της μετατροπής μονάδων μέτρησης περιελάμβανε την δημιουργία ενός απλού προγράμματος σε γλώσσα προγραμματίσμου Python με σκοπό την αυτοματοποίηση των αριθμητικών πράξεων που απαιτούνταν. Παρακάτω παρουσιάζονται οι μαθηματικές σχέσεις μετατροπής που χρησιμοποιήθηκαν σε κάθε περίπτωση, αναλόγως τις αρχικές μονάδες μέτρησης των πρωτογενών δεδομένων.

- Χρονική διάρκεια διεξαγωγής πειράματος [h]

$$t[days] * 24 \rightarrow t[h] \quad (19)$$

$$\frac{t[minutes]}{60} \rightarrow t[h] \quad (20)$$

- Συγκέντρωση νανοσωματιδίων στο μέσο καλλιέργειας [nM]

$$\frac{\frac{NPs}{ml}}{N_{Avogadro}} 10^{12} \rightarrow C[nM] \quad (21)$$

$$C[\mu M] * 10^3 \rightarrow C[nM] \quad (22)$$

$$\frac{C[pM]}{10^3} \rightarrow C[nM] \quad (23)$$

$$C[mM] * 10^6 \rightarrow C[nM] \quad (24)$$

Όπου $N_{Avogadro} = 6.022 * 10^{23} \text{ mol}^{-1}$ (ο αριθμός Avogadro).

Στην περίπτωση που η συγκέντρωση είχε μετρηθεί σε μάζα νανοσωματιδίων ανά όγκο μέσου καλλιέργειας η μετατροπή έγινε ως εξής:

$$\begin{aligned} \text{Συγκέντρωση σε Molarity} &= \frac{\frac{\text{αριθμός νανοσωματιδίων}}{\text{όγκο καλλιέργειας}}}{N_{Avogadro}} \\ &= \frac{\frac{\text{μάζα νανοσωματιδίων}}{\text{όγκο καλλιέργειας}}}{\text{μάζα ενός νανοσωματιδίου} * N_{Avogadro}} \\ &= \frac{\frac{\text{μάζα νανοσωματιδίων}}{\text{όγκο καλλιέργειας}}}{\text{Όγκος ενός νανοσωματιδίου} * \text{Πυκνότητα νανοσωματιδίου} * N_{Avogadro}} \quad (25) \end{aligned}$$

Για τον υπολογισμό του όγκου ενός νανοσωματιδίου (V) έγινε η παραδοχή τέλειου σφαιρικού σχήματος διαμέτρου (D).

$$V = \frac{4}{3} \pi \left(\frac{D}{2}\right)^3 \quad (26)$$

Αυτή η παραδοχή εμπεριέχει σφάλμα καθώς, αφενός είναι αδύνατον να παρασκευαστούν ομοιόμορφα νανοσωματίδια σε σχήμα τέλειας σφαίρας και αφετέρου τα νανοσωματίδια έχουν την τάση να σχηματίζουν συσσωματώματα κατά την παραμονή τους σε διαλύματα όπως τα κοινά μέσα καλλιέργειας. Έτσι, η πιο

χαρακτηριστική διάμετρος για τους υπολογισμούς σύμφωνα με την Σχέση 25 είναι η Υδροδυναμική διάμετρος που αναφέρεται ακριβώς στην φαινόμενη διάμετρο ενός σφαιρικού συσσωματώματος νανοσωματιδίων¹⁰⁸. Βεβαίως, ακόμη και στις περιπτώσεις που είναι διαθέσιμη η Υδροδυναμική διάμετρος, θα υπάρχει απόκλιση από την πραγματική τιμή NPs/cell λόγω χρήσης της ονομαστικής πυκνότητας του υλικού κατασκευής των νανοϋλικών η οποία είναι κατά κανόνα μεγαλύτερη από την πραγματική πυκνότητα των συσσωματωμάτων που περιέχουν κενά. Πράγματι, όπως έδειξαν οι Hsiao et al.¹⁰⁹, ο πραγματικός αριθμός νανοσωματιδίων που μετρήθηκε μέσω φασματομετρίας μάζας ενός σωματιδίου, single particle ICP-MS, ήταν μικρότερος από αυτόν που υπολογίστηκε μέσω μέτρησης με την κλασική μέθοδο φασματομετρίας μάζας, ICP-MS, και αναγωγής του αποτελέσματος σε NPs/cell χρησιμοποιώντας τη μάζα και την Ονομαστική ή Υδροδυναμική διάμετρο. Στην δεύτερη περίπτωση όμως, ο υπολογισμός του αριθμού νανοσωματιδίων μέσω της Υδροδυναμικής διαμέτρου δίνει καλύτερες προβλέψεις σε σχέση με την χρήση της Ονομαστικής διαμέτρου αφού λαμβάνει υπ' όψιν την πραγματική διάσταση των συσσωματωμάτων. Για αυτόν τον λόγο προτιμήθηκε ο υπολογισμός της Συγκέντρωσης νανοσωματιδίων με βάση την Υδροδυναμική διάμετρο, ενώ η Ονομαστική διάμετρος χρησιμοποιήθηκε μόνο σε περιπτώσεις που η πρώτη δεν ήταν διαθέσιμη. Τελικά, η σχέση 25 χρησιμοποιήθηκε, αναλόγως των αρχικών μονάδων μέτρησης των πρωτογενών δεδομένων, με την μορφή της Σχέσης 27 ή 28.

$$\frac{C \left[\frac{\mu g}{ml} \text{ ή } \frac{mg}{L} \text{ ή } ppm \right]}{\frac{4}{3} \pi \left(\frac{D}{2} 10^{-7} cm \right)^3 d_{bulk} \left(\frac{g}{cm^3} \right) N_{Avogadro}} 10^6 \rightarrow C [nM] \quad (27)$$

$$\frac{C \left[\frac{mg}{ml} \right]}{\frac{4}{3} \pi \left(\frac{D}{2} 10^{-7} cm \right)^3 d_{bulk} \left(\frac{g}{cm^3} \right) N_{Avogadro}} 10^9 \rightarrow C [nM] \quad (28)$$

- Μεταβλητή εξόδου – Αριθμός νανοσωματιδίων ανά κύτταρο [NPs/cell]

Για την μετατροπή της μεταβλητής εξόδου στις επιθυμητές μονάδες μέτρησης έγιναν οι προαναφερθείσες παραδοχές σχετικά με την πυκνότητα, το σχήμα και τον όγκο των νανοσωματιδίων.

$$\frac{attomol}{cell} * 10^{-18} * N_{Avogadro} \rightarrow \frac{NPs}{cell} \quad (29)$$

$$\frac{nmol}{cell} * 10^{-9} * N_{Avogadro} \rightarrow \frac{NPs}{cell} \quad (30)$$

$$\frac{\frac{\text{μάζα νανοσωματιδίων}}{\text{κύτταρο}}}{\text{μάζα ενός νανοσωματιδίου}} = \frac{\frac{\mu g}{cell}}{\frac{4}{3} \pi \left(\frac{D}{2} 10^{-7} cm \right)^3 d_{bulk} \left(\frac{g}{cm^3} \right) 10^6} \rightarrow \frac{NPs}{cell} \quad (31)$$

$$\frac{\frac{pg}{cell}}{\frac{4}{3}\pi\left(\frac{D}{2}10^{-7}cm\right)^3 d_{bulk}\left(\frac{g}{cm^3}\right)10^{12}} \rightarrow \frac{NPs}{cell} \quad (32)$$

$$\frac{\frac{ng}{cell}}{\frac{4}{3}\pi\left(\frac{D}{2}10^{-7}cm\right)^3 d_{bulk}\left(\frac{g}{cm^3}\right)10^9} \rightarrow \frac{NPs}{cell} \quad (33)$$

$$\frac{\frac{mg}{cell}}{\frac{4}{3}\pi\left(\frac{D}{2}10^{-7}cm\right)^3 d_{bulk}\left(\frac{g}{cm^3}\right)10^3} \rightarrow \frac{NPs}{cell} \quad (34)$$

Εφόσον εξασφαλίστηκε η συνέπεια των μονάδων μέτρησης όλων των μεγεθών, προέκυψε ένα σύνολο δεδομένων με 19 στήλες όπως συνοψίζονται στον Πίνακα 4.

Πίνακας 4. Επεξήγηση μεταβλητών του συνόλου δεδομένων

Είδος Μεταβλητής	Όνομα Μεταβλητής	Μονάδες Μέτρησης
Μεταδεδομένα	Κωδικοποίηση έρευνας	-
	DOI έρευνας	-
Φυσικοχημικές Ιδιότητες Νανοσωματιδίων	Είδος νανοσωματιδίων	-
	Σχήμα	-
	Ονομαστική διάμετρος	nm
	Μέση υδροδυναμική διάμετρος	nm
	Z-δυναμικό	mV
	Επικάλυψη νανοσωματιδίων	-
Πειραματικές Παράμετροι	Χρονική διάρκεια πειράματος	hours
	Συγκέντρωση νανοσωματιδίων στο μέσο καλλιέργειας	nM
	Ιοντική ισχύς μέσου καλλιέργειας	mol/L
	%FBS	%
	Πενικιλίνη/Στρεπτομυκίνη	1/0
	Εφαρμογή υπερήχων	1/0
	Μέθοδος ποσοτικοποίησης νανοσωματιδίων στα κύτταρα	-
Χαρακτηριστικά κυτταρικής σειράς	Κατηγορία κυτταρικής σειράς	-
	Είδος προέλευσης κυτταρικής σειράς	-
	Μορφολογία κυτταρικής σειράς	-
Μεταβλητή Εξόδου	Αριθμός νανοσωματιδίων ανά κύτταρο	NPs/cell

4.4.4. Διαχείριση ελλিপών τιμών

Επόμενο βήμα της επεξεργασίας δεδομένων αποτέλεσε η διαχείριση των ελλিপών τιμών. Συγκεκριμένα, σε πολλές πειραματικές μελέτες παρατηρήθηκε ελλιπής καταγραφή δεδομένων σχετικά με την Ονομαστική διάμετρο (ποσοστό ελλিপών δεδομένων 25.44%), την Υδροδυναμική διάμετρο (ποσοστό ελλিপών δεδομένων 51.47%), το Z-δυναμικό (ποσοστό ελλিপών δεδομένων 41.10%) και το Ποσοστό FBS (ποσοστό ελλিপών δεδομένων 4.60%). Δεδομένου ότι οι περισσότερες μεθοδολογίες

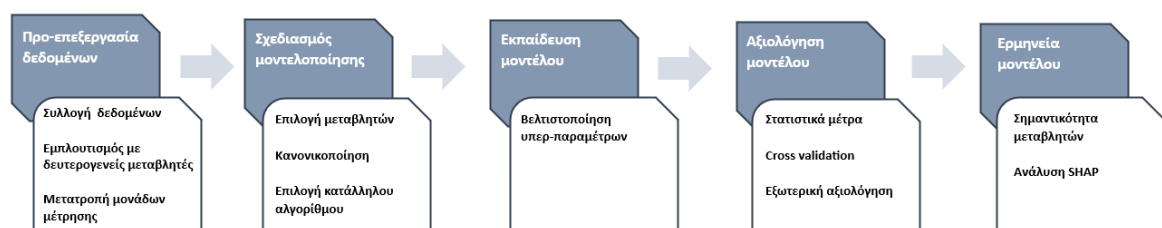
Μηχανικής Μάθησης απαιτούν την παντελή έλλειψη κενών τιμών στο σύνολο δεδομένων, επιλέχθηκε να εφαρμοστεί η μέθοδος MICE προκειμένου να προβλεφθούν οι ελλιπείς τιμές. Σημειώνεται ότι στην συγκεκριμένη περίπτωση θεωρείται ότι ισχύει η παραδοχή MAR.

Η εφαρμογή της μεθόδου MICE στο σύνολο δεδομένων που κατασκευάστηκε πραγματοποιήθηκε μέσω προγράμματος σε Python. Αρχικά, δημιουργήθηκε ένα αντίγραφο (mask) των ελλιπών τιμών και στη συνέχεια διαχωρίστηκαν οι κατηγορικές μεταβλητές από αυτές που περιέχουν αριθμητικά στοιχεία. Οι κατηγορικές μεταβλητές που μπορεί να σχετίζονται με τις μεταβλητές που περιέχουν ελλιπείς τιμές (Είδος νανοσωματιδίων, Σχήμα, Επικάλυψη νανοσωματιδίων) μετατράπηκαν σε αριθμητικές με τη μέθοδο «One-hot encoding». Παρ' όλο που οι μεταβλητές με ελλιπείς τιμές ήταν αριθμητικές, κρίθηκε απαραίτητη η κωδικοποίηση των κατηγορικών μεταβλητών καθώς τα μοντέλα παλινδρόμησης της μεθοδολογίας MICE απαιτούν αριθμητικές μεταβλητές. Έτσι, εξασφαλίστηκε η συμμετοχή όλων των σχετικών μεταβλητών στην επαναληπτική διαδικασία πρόβλεψης των ελλιπών τιμών και η διατήρηση όλων των σχέσεων μεταξύ των μεταβλητών. Σημειώνεται ότι οι μεταβλητές που σχετίζονται με το Είδος της κυτταρικής σειράς δεν συμπεριλήφθησαν στην επαναληπτική διαδικασία καθώς η Ονομαστική, η Υδροδυναμική διάμετρος αλλά και το Z-δυναμικό των νανοσωματιδίων είναι ανεξάρτητα από τα χαρακτηριστικά της κυτταρικής σειράς.

Στη συνέχεια, το σύνολο δεδομένων ενώθηκε εκ νέου και εφαρμόστηκε η μέθοδος MICE με χρήση της συνάρτησης «IterativeImputer» της βιβλιοθήκης «sklearn», επιλέγοντας ένα «Random Forest» μοντέλο για την εφαρμογή της παλινδρόμησης. Σύμφωνα με προηγούμενες μελέτες επιλέχθηκε να πραγματοποιηθούν 10 επαναλήψεις για την πρόβλεψη των ελλιπών τιμών αφού έχει αποδειχθεί ότι στις περισσότερες περιπτώσεις λαμβάνονται ικανοποιητικά αποτελέσματα με 5-10 επαναλήψεις⁷³. Έτσι, προέκυψε το τελικό σύνολο δεδομένων που περιλαμβάνει 19 στήλες και 1022 γραμμές.

5. Μεθοδολογία μηχανικής μάθησης για την κυτταρική πρόσληψη νανοσωματιδίων: Βήματα ανάπτυξης, αξιολόγησης και ερμηνείας μοντέλων

Στο παρόν κεφάλαιο παρατίθενται οι τεχνικές και τα βήματα που ακολουθήθηκαν για την προεπεξεργασία των δεδομένων, τη μοντελοποίηση και τη βελτιστοποίηση του αλγορίθμου XGBoost. Επίσης, αναγράφονται τα στατιστικά μέτρα αξιολόγησης που αξιοποιήθηκαν, η εφαρμογή της μεθόδου τυχαιοποίησης της μεταβλητής εξόδου, καθώς και η ερμηνεία των μοντέλων με χρήση της μεθόδου SHAP (Σχήμα 4).



Σχήμα 4. Τυπικό διάγραμμα ροής μοντελοποίησης δεδομένων με μεθόδους μηχανικής μάθησης (τροποποιημένο από Ponce-Bobadilla et al. (2024)⁹²)

5.1. Προεπεξεργασία συνόλου δεδομένων

Αρχικά, συμμετείχαν στην μοντελοποίηση όλες οι μεταβλητές που παρουσιάζονται στον Πίνακα 4, με εξαίρεση τις 2 μεταβλητές που αναφέρονται σε μεταδεδομένα σχετικά με τις επιστημονικές μελέτες από τις οποίες αντλήθηκαν τα δεδομένα. Ακολουθώντας μία ανάλυση συσχέτισης των μεταβλητών εισόδου, αφαιρέθηκε μία εκ των μεταβλητών του κάθε ζεύγους με υψηλή συσχέτιση. Για τον σκοπό αυτό χρησιμοποιήθηκε ο συντελεστής συσχέτισης «Pearson», ο οποίος συγκρίνει όλες τις τιμές δύο μεταβλητών και υπολογίζει μία τιμή (score) μεταξύ του -1 και του 1 για κάθε ζεύγος. Μεταβλητές με τιμή κοντά στο -1 έχουν έντονα αρνητική συσχέτιση, ενώ οι τιμές που τείνουν στο 1 δείχνουν την ύπαρξη έντονα θετικής συσχέτισης¹¹⁰.

Στη συνέχεια ορίστηκαν οι μεταβλητές εισόδου ως X (όλες εκτός από την μεταβλητή εξόδου, δηλαδή Αριθμός νανοσωματιδίων ανά κύτταρο) και η μεταβλητή εξόδου, Y, υπέστη λογαριθμική μετατροπή με σκοπό την μείωση του πολύ μεγάλου εύρους τιμών των πρωτογενών δεδομένων (από 0 έως 10^{10} NPs/cell)¹¹¹. Τονίζεται ότι η λογαριθμική μετατροπή εφαρμόζεται μόνο σε θετικές τιμές και έτσι σε κάθε παρατήρηση της μεταβλητής εξόδου προστέθηκε μία μικρή θετική τιμή, 10^{-10} , ώστε να μην προκύψουν τιμές $-\infty$ κατά την μετατροπή μηδενικών τιμών πρόσληψης νανοϋλικών σε κύτταρα.

Έπειτα, οι κατηγορικές μεταβλητές εισόδου διαχωρίστηκαν και μετατράπηκαν σε αριθμητικές μέσω της μεθόδου «One-hot encoding» (χρησιμοποιήθηκε η εσωτερική συνάρτηση «OneHotEncoder») προτού ενωθούν εκ νέου με τις υπόλοιπες αριθμητικές

μεταβλητές σε ένα ενιαίο σύνολο μεταβλητών εισόδου. Μετά από αυτή την διαδικασία, οι αρχικές κατηγορικές μεταβλητές αφαιρέθηκαν από το σύνολο δεδομένων.

Το σύνολο δεδομένων που αποτελείται πλέον μόνο από αριθμητικές τιμές, χωρίστηκε με τυχαίο τρόπο σε σύνολο εκπαίδευσης και ελέγχου μέσω της εσωτερικής συνάρτησης «train_test_split», ορίζοντας το ποσοστό των συνολικών δεδομένων που συμμετέχουν στο σύνολο εκπαίδευσης στο 20%. Με στόχο την αποφυγή της πιθανής μεροληψίας των μοντέλων μηχανικής μάθησης λόγω του διαφορετικού εύρους τιμών των μεταβλητών, επιλέχθηκε οι μεταβλητές του συνόλου εκπαίδευσης να κανονικοποιηθούν με την μέθοδο Z-score μέσω της συνάρτησης «StandardScaler». Χρησιμοποιώντας την μέση τιμή (μ) και την τυπική απόκλιση (σ) των μεταβλητών του συνόλου εκπαίδευσης, η μέθοδος Z-score εφαρμόστηκε και στο σύνολο ελέγχου με τις ίδιες παραμέτρους μ και σ ώστε να αποφευχθεί ο κίνδυνος διαφυγής δεδομένων από το σύνολο ελέγχου στο σύνολο δεδομένων (data leakage).

5.2. Δημιουργία και επιλογή του βέλτιστου μοντέλου μηχανικής μάθησης

Για την μοντελοποίηση των δεδομένων επιλέχθηκε, όπως προαναφέρθηκε, ένας αλγόριθμος που έχει δοκιμαστεί στην διεθνή βιβλιογραφία και έχει αποδειχτεί αποδοτικός και γρήγορος για μεγάλο όγκο δεδομένων και περίπλοκα προβλήματα: ο αλγόριθμος XGBoost. Η εύρεση των βέλτιστων παραμέτρων για την δημιουργία του μοντέλου πραγματοποιήθηκε μέσω της μεθόδου «parameter grid search», δίνοντας δύο πιθανές τιμές σε κάθε μία από τις 8 πιο βασικές παραμέτρους (Πίνακας 5) ώστε να δημιουργηθούν και να αξιολογηθούν διαφορετικά μοντέλα μηχανικής μάθησης. Παράλληλα, εφαρμόστηκε η μέθοδος «cross-validation» με 10 επαναλήψεις για κάθε συνδυασμό πιθανών παραμέτρων και μέτρο αξιολόγησης των μοντέλων την μέση τιμή του στατιστικού μέτρου R^2 των 10 διαφορετικών, εσωτερικών συνόλων εκπαίδευσης για κάθε μοντέλο (κάθε συνδυασμό παραμέτρων). Εύκολα γίνεται κατανοητό πως λόγω του μεγάλου αριθμού πιθανών συνδυασμών παραμέτρων αλλά και του δεκαπλού «cross-validation», η υπολογιστική ισχύς που απαιτείται για την δημιουργία $10 \times 2^8 = 2560$ μοντέλων είναι πολύ μεγάλη. Έτσι, για 4 από τις παραμέτρους επιλέχθηκε να δοκιμαστεί μία μόνο τιμή. Πιο συγκεκριμένα, για τον αριθμό των εκτιμητών ($n_estimators$) επιλέχθηκε η τιμή 200, για τον ρυθμό εκμάθησης (learning rate) η τιμή 0.1, ενώ για τις παραμέτρους «subsample» και «colsample_bytree» η τιμή 0.7. Τονίζεται ότι οι μικρότερες τιμές των παραμέτρων «subsample» και «colsample_bytree» σε σχέση με τις προεπιλεγμένες τιμές στοχεύουν στην μείωση του κινδύνου υπερπροσαρμογής των μοντέλων.

Πίνακας 5. Πιθανές τιμές των παραμέτρων του μοντέλου XGBoost που δοκιμάστηκαν για την εύρεση του βέλτιστου μοντέλου με βάση το στατιστικό μέτρο R^2 .

Παράμετρος	Πιθανές τιμές
n_estimators	200
max_depth	4, 5
learning_rate	0.1
subsample	0.7
colsample_bytree	0.7
reg_alpha	0.1, 5
reg_lambda	5, 10
min_child_weight	1, 5

Τελικά, το μοντέλο με τον κατάλληλο συνδυασμό παραμέτρων που οδήγησε στην υψηλότερη μέση τιμή R^2 των εσωτερικών συνόλων εκπαίδευσης κατά την μέθοδο «cross-validation» επιλέχθηκε να εφαρμοστεί στο εξωτερικό σύνολο ελέγχου (20% των συνολικών δεδομένων), για να επικυρωθεί η προβλεπτική ικανότητα του μοντέλου.

5.3. Επικύρωση, πεδίο εφαρμοσιμότητας και ερμηνεία μοντέλου

Το βέλτιστο μοντέλο XGBoost αξιολογήθηκε ως προς την ικανότητα πρόβλεψης των τιμών της μεταβλητής εξόδου του εξωτερικού συνόλου ελέγχου μέσω των στατιστικών μέτρων R^2 , MAE, MSE και RMSE. Τα στατιστικά μετρά που εξαρτώνται από την τάξη μεγέθους των προβλεπόμενων τιμών, δηλαδή τα MAE, MSE και RMSE, ήταν αναμενόμενο να έχουν αρκετά μεγάλες τιμές και έτσι υπολογίστηκε και το κανονικοποιημένο RMSE διαιρώντας την αρχική τιμή με το εύρος τιμών της μεταβλητής εξόδου του εξωτερικού συνόλου δεδομένων. Όλα τα στατιστικά μετρά στο σύνολο ελέγχου υπολογίστηκαν μετά από αντίστροφη λογαριθμική μετατροπή της μεταβλητής Y ώστε να αποτυπώνουν την ικανότητα πρόβλεψης του μοντέλου στον πραγματικό χώρο των δεδομένων.

Παράλληλα, αξιοποιήθηκε η ιδιότητα (attribute) «.feature_importances_» του μοντέλου XGBoost στην Python ώστε να εμφανιστεί ένα διάγραμμα με την σημαντικότητα κάθε μεταβλητής στην πρόβλεψη. Από αυτό το διάγραμμα είναι δυνατόν να αναγνωριστούν προβλήματα υπερπροσαρμογής του μοντέλου σε μία μόνο μεταβλητή.

Ακόμη, για την αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου εφαρμόστηκε η μέθοδος «leave-one-out cross-validation». Στην συγκεκριμένη περίπτωση το μοντέλο εκπαιδεύτηκε σε όλα τα δεδομένα πλην μίας εκ των 71 ερευνών και τα στατιστικά μετρά αξιολόγησης υπολογίστηκαν με βάση την «άγνωστη» για το μοντέλο έρευνα. Πλέον το σύνολο εκπαίδευσης δεν περιελάμβανε καμία από τις χρονικές στιγμές που μελετώνται σε κάθε πείραμα στα πλαίσια της ίδια έρευνας ώστε να ελεγχθεί η ικανότητα γενίκευσης των προβλέψεων σε πραγματικά «άγνωστα» δεδομένα.

Την στατιστική αξιολόγηση του μοντέλου συμπλήρωσε η ανάλυση τυχαίας αντικατάστασης των τιμών της μεταβλητής εξόδου (Y -randomization) και η δημιουργία νέων μοντέλων XGBoost με σκοπό τον έλεγχο της αξιοπιστίας του αρχικού μοντέλου.

Σε αυτή την ανάλυση αναμένεται βεβαίως υποβάθμιση των στατιστικών μέτρων των μοντέλων μετά την τυχαία αντικατάσταση των τιμών Y .

Αφού επιβεβαιώθηκε η αξιοπιστία του μοντέλου, υπολογίστηκε το πεδίο εφαρμοσιμότητας (AD) του μοντέλου, σύμφωνα με το οποίο αποφασίζεται αν μία πρόβλεψη είναι έμπιστη. Για τον σκοπό αυτό, εφαρμόστηκε μία μεθοδολογία k -κοντινότερων γειτόνων (k -NN) με $k=5$. Πιο συγκεκριμένα, δημιουργήθηκε μία συνάρτηση που υπολογίζει την μέση Ευκλείδεια απόσταση κάθε παρατήρησης του συνόλου ελέγχου από τους 5 κοντινότερους γείτονες του συνόλου εκπαίδευσης και δέχεται την πρόβλεψη ως έμπιστη μόνο αν η υπολογισμένη απόσταση είναι μικρότερη του 95% των αποστάσεων μεταξύ των παρατηρήσεων του συνόλου εκπαίδευσης.

Για την ερμηνεία του μοντέλου επιλέχθηκε η μέθοδος SHAP, η οποία παρέχει χρήσιμες πληροφορίες σχετικά με την συμμετοχή κάθε μεταβλητής στην τελική πρόβλεψη αλλά και τον τρόπο με τον οποίο υψηλές ή χαμηλές τιμές των μεταβλητών συμβάλλουν στην πρόβλεψη μεγαλύτερων τιμών πρόσληψης ναυοσωματιδίων στα κύτταρα.

5.4. Προγραμματιστικά εργαλεία

Για την προεπεξεργασία και μοντελοποίηση των δεδομένων, όπως και την ερμηνεία των αποτελεσμάτων αξιοποιήθηκαν οι διαθέσιμες «βιβλιοθήκες» της γλώσσας προγραμματισμού Python. Στον Πίνακα 6 παρουσιάζονται αναλυτικά οι «βιβλιοθήκες», τα «πακέτα» και τα «εργαλεία» που χρησιμοποιήθηκαν σε κάθε υπολογιστική ανάλυση.

Σημειώνεται ότι σε κάθε περίπτωση ανάλυσης δεδομένων και αριθμητικών πράξεων, χρησιμοποιήθηκαν οι «βιβλιοθήκες» NumPy και Pandas. Επίσης, για την κατασκευή διαγραμμάτων αξιοποιήθηκε η βιβλιοθήκη matplotlib.

Πίνακας 6. Παρουσίαση «βιβλιοθηκών» και πακέτων της γλώσσας προγραμματισμού Python που χρησιμοποιήθηκαν σε κάθε υπολογιστική ανάλυση.

Ανάλυση	Βιβλιοθήκη	Πακέτα	Εργαλεία
MICE	sklearn	experimental, impute, ensemble	enable_iterative_imputer, Iterative_Imputer, RandomForestRegressor
One-hot encoding	sklearn	preprocessing	OneHotEncoder
Διαχωρισμός συνόλου εκπαίδευσης και ελέγχου	sklearn	model_selection	train_test_split
Κανονικοποίηση μεταβλητής εξόδου	sklearn	preprocessing	StandardScaler
Cross-validation	sklearn	model_selection	GridSearchCV, cross_val_score
Μοντελοποίηση	xgboost		XGBoostRegressor

Στατιστική αξιολόγηση	sklearn	metrics	mean_absolute_error, mean_squared_error, r2_score
Ανάλυση συσχέτισης	seaborn		Heatmap
Οπτικοποίηση δεδομένων	ydata_profiling		ProfileReport
Πεδίο εφαρμοσιμότητας	sklearn	neighbors	NearestNeighbors
Ανάλυση SHAP	shap		

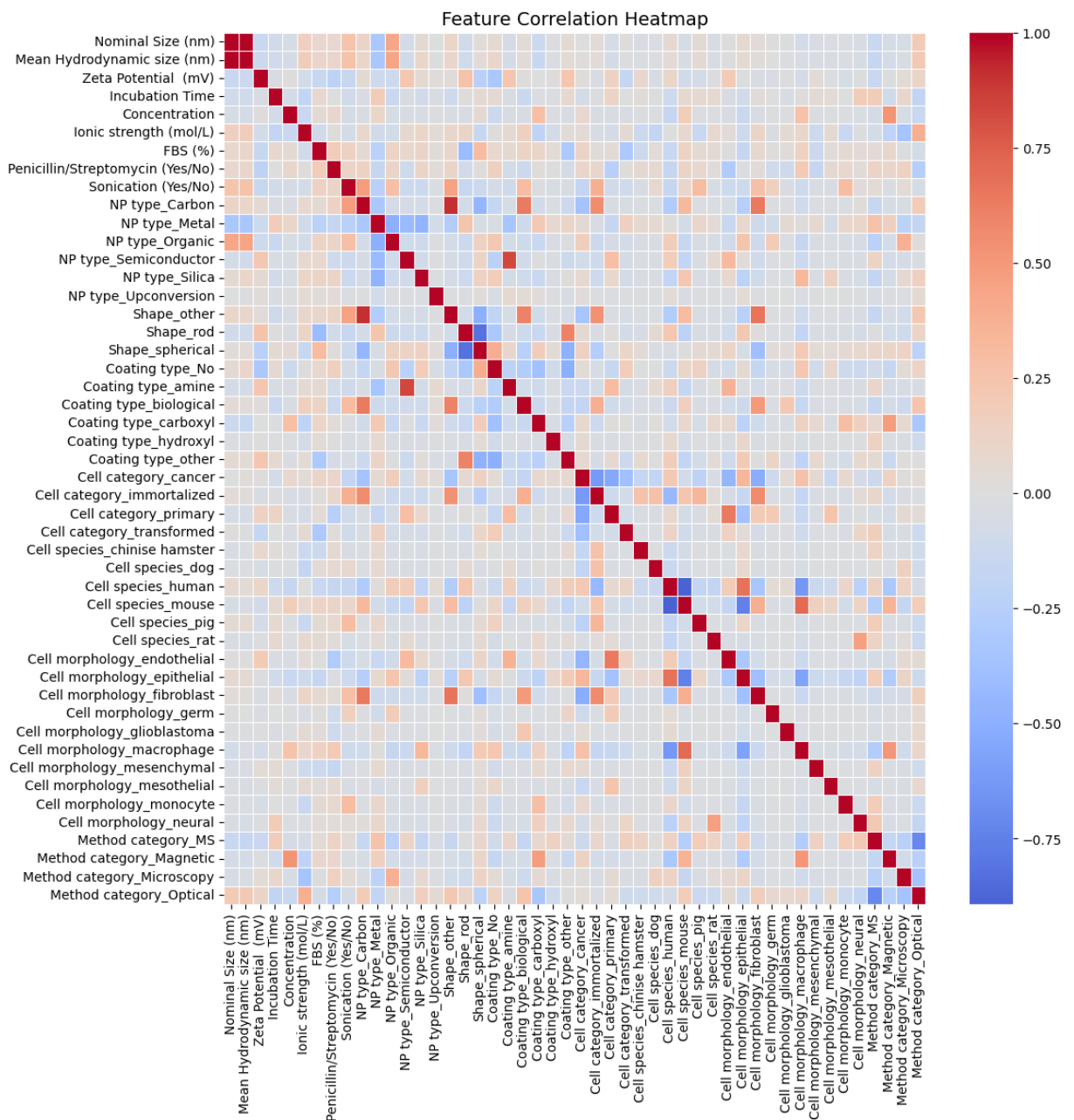
6. Αποτελέσματα – Συζήτηση

Στο παρόν κεφάλαιο αναλύονται και αξιολογούνται τα αποτελέσματα τόσο της συλλογής δεδομένων, όσο και της επεξεργασίας και μοντελοποίησης. Αρχικά παρουσιάζεται η προσπάθεια αντιμετώπισης του σημαντικού προβλήματος της υπερπροσαρμογής που εντοπίστηκε κατά τη διαδικασία μοντελοποίησης. Σε αυτό το στάδιο η αναθεώρηση του συνόλου δεδομένων αποδείχθηκε καθοριστική, επιτρέποντας την επιτυχημένη μοντελοποίηση της κυτταρικής πρόσληψης νανοσωματιδίων. Έπειτα από την συζήτηση του είδους των δεδομένων, ακολουθούν τα αποτελέσματα της βελτιστοποίησης, απλοποίησης και αξιολόγησης του μοντέλου XGBoost. Τελικά, η ερμηνεία του μοντέλου πραγματοποιείται μέσω της ανάλυσης SHAP.

6.1. Το αρχικό σύνολο δεδομένων και η μεγάλη βαρύτητα της μεταβλητής του Χρόνου

Το αρχικό σύνολο δεδομένων περιλαμβάνει 18 μεταβλητές εισόδου (2 εκ των οποίων δεν συμμετέχουν στην μοντελοποίηση ως μεταδεδομένα) και μία μεταβλητή εξόδου τον Αριθμό νανοσωματιδίων ανά κύτταρο (NPs/cell). Έπειτα από την προεπεξεργασία των δεδομένων, και συγκεκριμένα την μετατροπή των κατηγορικών μεταβλητών σε αριθμητικές, το σύνολο δεδομένων αποτελείται από 48 στήλες και 1022 γραμμές. Όπως ήταν αναμενόμενο, η μετατροπή των κατηγορικών μεταβλητών σε αριθμητικές μέσω της μεθόδου «One-hot Encoding» αύξησε σημαντικά τον αριθμό των στηλών, ο οποίος όμως παραμένει αρκετά μικρός σε σχέση με τον αριθμό των παρατηρήσεων έτσι ώστε να μην υπάρχει σημαντικός κίνδυνος υπερπροσαρμογής των μοντέλων μηχανικής μάθησης, καθώς αυτά έχουν επαρκή αριθμό παρατηρήσεων για να γενικεύσουν, χωρίς να εξαρτώνται υπερβολικά από τα διαθέσιμα χαρακτηριστικά.

Ο πίνακας «Heat map» χρησιμοποιήθηκε για την οπτικοποίηση των αποτελεσμάτων. Όσο πιο σκούρο κόκκινο ή μπλε είναι το τετράγωνο που αντιστοιχεί σε ένα ζεύγος μεταβλητών τόσο πιο έντονα θετική ή αρνητική αντίστοιχα είναι η συσχέτισή τους. Από αυτή την ανάλυση αναδείχθηκε μία έντονη θετική συσχέτιση μεταξύ των μεταβλητών της Ονομαστικής και της Μέσης υδροδυναμικής διαμέτρου των νανοσωματιδίων, όπως φαίνεται και στο Σχήμα 5. Αυτή η σχέση είναι εν μέρη αναμενόμενη καθώς πολλές από τις τιμές της μεταβλητής της Μέσης υδροδυναμικής διαμέτρου συμπληρώθηκαν με την μεθοδολογία διαχείρισης ελλειπών τιμών MICE, πιθανώς λαμβάνοντας υπ' όψιν κυρίως τις τιμές της Ονομαστικής διαμέτρου και αντίστροφα. Θεωρώντας λοιπόν πως μέρος της πληροφορίας που κωδικοποιεί η Ονομαστική διάμετρος έχει χρησιμοποιηθεί για τον υπολογισμό της Μέσης υδροδυναμικής διαμέτρου, επιλέχθηκε να αφαιρεθεί εξ ολοκλήρου η μεταβλητή Ονομαστική διάμετρος από το σύνολο δεδομένων. Με την απομάκρυνση των έντονα συσχετισμένων μεταβλητών εξασφαλίζει ότι η προβλεπτική ικανότητα των μοντέλων μηχανικής μάθησης δεν θα επηρεαστεί από την ύπαρξη μεταβλητών με υψηλή συσχέτιση.



Σχήμα 5. Πίνακας «Heat map» της συσχέτισης των μεταβλητών εισόδου του συνόλου δεδομένων. Με έντονο κόκκινο χρώμα φαίνεται η υψηλή θετική συσχέτιση των μεταβλητών της Ονομαστική (Nominal size (nm)) και Μέσης υδροδυναμικής (Mean Hydrodynamic size (nm)) διαμέτρου.

Αφού αφαιρέθηκε μία εκ των μεταβλητών του ζεύγους με υψηλή τιμή συντελεστή «Pearson», το σύνολο δεδομένων χρησιμοποιήθηκε για την δημιουργία ενός αρχικού μοντέλου μηχανικής μάθησης χρησιμοποιώντας τυπικές τιμές των υπερπαραμέτρων της μεθόδου XGBoost ($n_estimators = 200$, $learning_rate = 0.1$, $max_depth = 5$, $reg_alpha = 0.1$, $reg_lambda = 5$), ώστε να ελεγχθεί η ποιότητα των δεδομένων και η ικανότητα δημιουργίας γενικεύσιμων μοντέλων προτού πραγματοποιηθεί η βελτιστοποίηση των υπερπαραμέτρων. Σύμφωνα με τα στατιστικά μέτρα της διαδικασίας του «cross-validation» (Πίνακας 7) αλλά και της αξιολόγησης του

μοντέλου στο σύνολο ελέγχου (Πίνακας 8) φάνηκε αρχικά μία πολύ ικανοποιητική προσαρμογή του μοντέλου XGBoost στα δεδομένα.

Πίνακας 7. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία του «cross-validation» στο αρχικό σύνολο εκπαίδευσης.

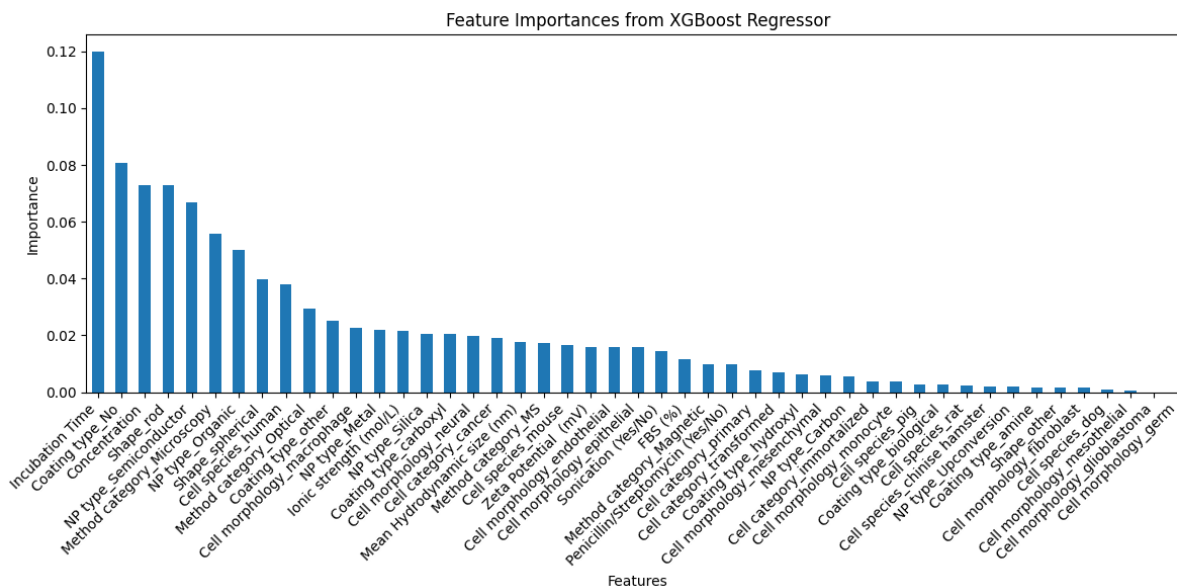
Αριθμός υποδιαίρεσης συνόλου δεδομένων	R^2
1	0.949
2	0.909
3	0.860
4	0.788
5	0.868
6	0.905
7	0.828
8	0.509
9	0.722
10	0.806
Μέσος Όρος	0.814

Πίνακας 8. Στατιστικά μέτρα στο σύνολο εκπαίδευσης πριν και μετά την λογαριθμική μετατροπή της μεταβλητής εξόδου για το αρχικό μοντέλο XGBoost.

	Σύνολο ελέγχου (λογαριθμική κλίμακα μεταβλητής εξόδου)	Σύνολο ελέγχου (κανονική κλίμακα μεταβλητής εξόδου)
R^2	0.936	0.772
MAE	0.976	1.61E+08
MSE	3.800	1.37E+18
RMSE	1.949	1.17E+09
Κανονικοποιημένο RMSE	-	0.0457

Στους Πίνακες 7 και 8 φαίνεται ότι τόσο στο «cross-validation» όσο και στο σύνολο ελέγχου, το αρχικό μοντέλο παρουσιάζει μία πολύ καλή προβλεπτική συμπεριφορά. Βεβαίως, οι μεγάλες τιμές των μέτρων MAE, MSE και RMSE στο σύνολο ελέγχου που περιλαμβάνει τις πραγματικές τιμές της μεταβλητής εξόδου (πριν την λογαριθμική μετατροπή) είναι αναμενόμενες λόγω της μεγάλης κλίμακας τιμών της μεταβλητής εξόδου (NPs/cell).

Παρά τα φαινομενικά ικανοποιητικά στατιστικά μέτρα όμως, η μελέτη της σημαντικότητας των μεταβλητών για την πρόβλεψη του μοντέλου XGBoost ανέδειξε ένα πιθανό πρόβλημα υπερπροσαρμογής του μοντέλου στα δεδομένα. Συγκεκριμένα, όπως φαίνεται στο Διάγραμμα 3, το μοντέλο βασίζεται κυρίως στις τιμές της Χρονικής διάρκειας του πειράματος για να προβλέψει την κυτταρική πρόσληψη των νανοσωματιδίων, γεγονός που αναμένεται να επηρεάσει την δυνατότητα ερμηνευτικής ανάλυσης των σχέσεων των μεταβλητών μικρότερης σημαντικότητας.



Διάγραμμα 3. Σημαντικότητα μεταβλητών για την προβλεπτική ικανότητα του αρχικού μοντέλου XGBoost, όπως προκύπτει από την ιδιότητα (attribute) «feature_importances_» της βιβλιοθήκης XGBoost στην Python. Η μεταβλητή της Χρονικής διάρκειας (Incubation time) φαίνεται να υπερισχύει σημαντικά σε σχέση με τις υπόλοιπες.

Εφόσον δεν ήταν δυνατός ο εμπλουτισμός του συνόλου δεδομένων με περισσότερα παραδείγματα που πιθανώς θα μπορούσαν να μειώσουν τον κίνδυνο υπερπροσαρμογής, κρίθηκε απαραίτητη η αφαίρεση της μεταβλητής της Χρονικής διάρκειας. Γνωρίζοντας βεβαίως την μεγάλη σημαντικότητα αυτής της μεταβλητής για τον προσδιορισμό της κυτταρικής πρόσληψης νανοσωματιδίων, επιλέχθηκε να απλοποιηθεί το μοντέλο και να μην μελετηθεί η μεταβλητή του Χρόνου στοχεύοντας στην πιο εύκολη ανάδειξη και παρατήρηση των σχέσεων των υπολοίπων μεταβλητών. Αυτή η απλοποίηση, λοιπόν, αναμένεται να ενισχυθεί η κατανόηση των σχέσεων μεταξύ σημαντικών φυσικοχημικών ιδιοτήτων και πειραματικών παραμέτρων που επηρεάζουν την κυτταρική πρόσληψη.

Έτσι, προκειμένου να διατηρηθεί ο μέγιστος αριθμός παρατηρήσεων στο σύνολο δεδομένων, επιλέχθηκε η χρονική στιγμή που είναι κοινή στις περισσότερες πειραματικές μελέτες, δηλαδή οι 24 ώρες, και αφαιρέθηκαν όλες οι υπόλοιπες παρατηρήσεις. Το αναθεωρημένο σύνολο δεδομένων πλέον περιλαμβάνει μόνο το 27% των αρχικών παρατηρήσεων αλλά είναι ανεξάρτητο της μεταβλητής της Χρονικής διάρκειας των πειραμάτων, η οποία αφαιρέθηκε εξ ολοκλήρου. Σημαντικό θεωρείται ότι μετά από αυτή την αφαίρεση διατηρείται μέρος των δεδομένων της πλειοψηφίας των ερευνών (44 εκ των 71 αρχικών ερευνών) ώστε να μην χαθεί η πολυμορφία του συνόλου δεδομένων.

6.2. Το αναθεωρημένο σύνολο δεδομένων

Έχοντας αφαιρέσει την μεταβλητή του Χρόνου, τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν στη συνέχεια προβλέπουν την κυτταρική πρόσληψη νανοσωματιδίων μετά από 24 ώρες κυτταρικής επώασης. Αυτή η απλοποίηση του συνόλου δεδομένων δίνει την δυνατότητα δημιουργίας ερμηνεύσιμων μοντέλων που μπορούν να διερευνήσουν τις σχέσεις μεταξύ των υπόλοιπων παραμέτρων ενός πειράματος κυτταρικής πρόσληψης νανοσωματιδίων με μεγαλύτερη ακρίβεια.

6.2.1. Οπτικοποίηση δεδομένων

Το απλοποιημένο σύνολο δεδομένων αποτελείται από 17 μεταβλητές εισόδου (συμπεριλαμβανομένων δύο μεταβλητών που κωδικοποιούν μεταδεδομένα), την μεταβλητή εξόδου και 283 παρατηρήσεις. Μετά την μετατροπή των κατηγορικών μεταβλητών σε αριθμητικές με χρήση της μεθόδου «One-hot Encoding» ο αριθμός των παρατηρήσεων παραμένει σταθερός, αλλά ο αριθμός των μεταβλητών εισόδου αυξάνεται σε 37. Παρατηρείται ότι οι στήλες των αριθμητικών μεταβλητών εισόδου είναι μειωμένες σε σχέση με αυτές του αρχικού συνόλου δεδομένων (48) αφενός διότι έχουν αφαιρεθεί οι μεταβλητές της Ονομαστικής διαμέτρου και της Χρονικής διάρκειας, και αφετέρου επειδή η αφαίρεση μεγάλου μέρους των παρατηρήσεων είχε ως αποτέλεσμα την μη εμφάνιση ορισμένων κατηγοριών των κατηγορικών μεταβλητών. Παρακάτω παρουσιάζονται περιγραφικά στατιστικά στοιχεία για τις βασικές μεταβλητές του αναθεωρημένου συνόλου δεδομένων.

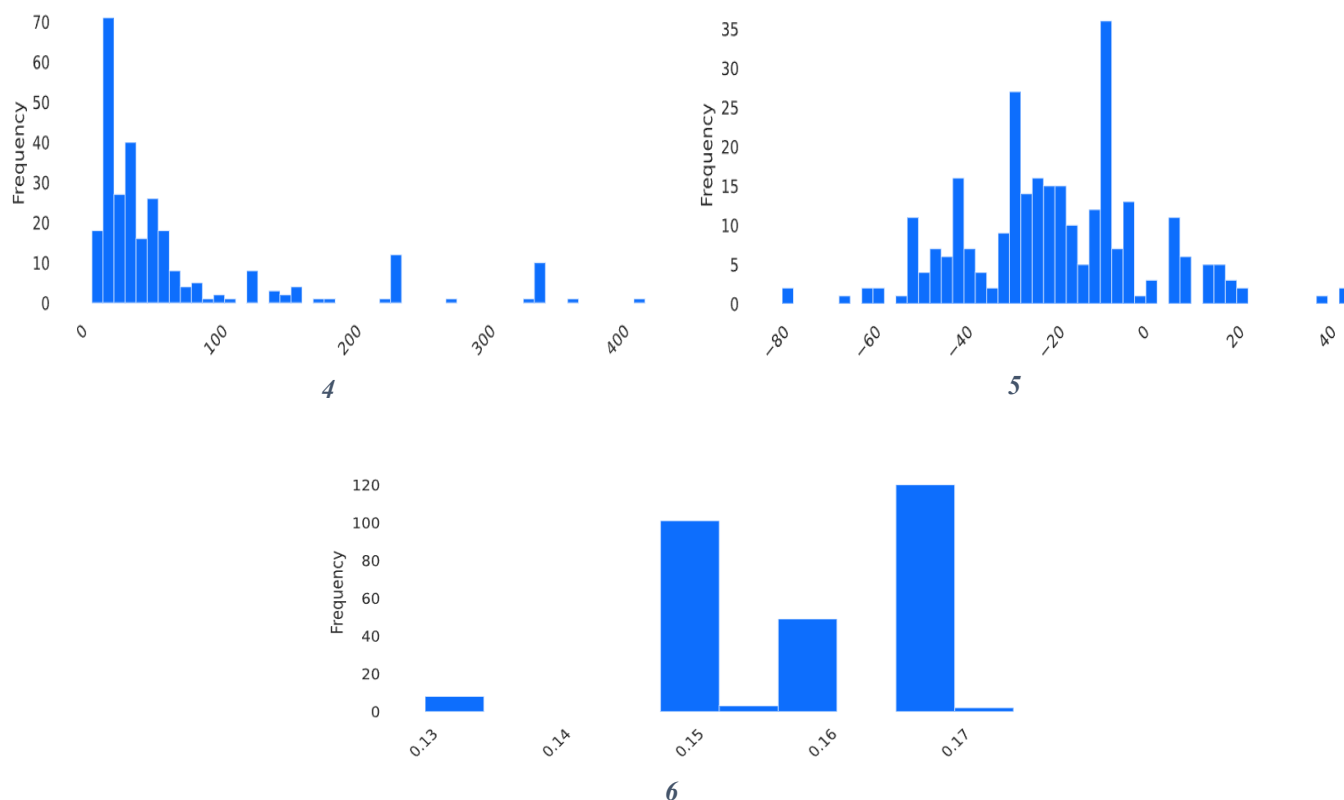
- Αριθμητικές μεταβλητές

Η Μέση υδροδυναμική διάμετρος λαμβάνει τιμές από 7.1 έως 420 nm με την κατανομή των τιμών να είναι έντονα μετατοπισμένη προς τις μικρότερες τιμές διαμέτρων (Διάγραμμα 4). Μάλιστα το 75% των παρατηρήσεων (Q3, τρίτο τεταρτημόριο) βρίσκεται κάτω από το όριο των 62 nm. Φαίνεται λοιπόν πως, παρά το σχετικά μεγάλο εύρος τιμών Μέσης υδροδυναμικής διαμέτρου, το μέγεθος της πλειοψηφίας των νανοσωματιδίων του συνόλου δεδομένων κυμαίνεται σε μικρές τιμές.

Όσον αφορά το επιφανειακό φορτίο, που στο συγκεκριμένο σύνολο δεδομένων αντιπροσωπεύεται από το Z-δυναμικό, η κατανομή των τιμών είναι πιο ομοιόμορφη σε σχέση με αυτή της Υδροδυναμικής διαμέτρου παρά την ύπαρξη ορισμένων τιμών που εμφανίζονται πιο συχνά (Διάγραμμα 5). Επίσης, παρατηρείται μία μετατόπιση της κατανομής προς αρνητικές τιμές με τον μέσο όρο των τιμών του Z-δυναμικού να είναι -19.42 mV και το 75% των τιμών να βρίσκεται μεταξύ των -78.81 (ελάχιστη τιμή) και -7.6 mV (Q3, τρίτο τεταρτημόριο). Είναι σαφές πως τα περισσότερα νανοσωματίδια που μελετώνται είναι αρνητικά φορτισμένα.

Τέλος, η Ιοντική ισχύς του μέσου καλλιέργειας, παρ' όλο που έχει κωδικοποιηθεί ως αριθμητική μεταβλητή, λαμβάνει μόνο 10 διακριτές τιμές που αντιστοιχούν στα 10 διαφορετικά μέσα καλλιέργειας που χρησιμοποιήθηκαν στις έρευνες που συμπεριλαμβάνονται στο αναθεωρημένο σύνολο δεδομένων (Διάγραμμα 6). Από το

ιστόγραμμα φαίνεται ότι στις περισσότερες μελέτες χρησιμοποιούνται κυρίως 2 μέσα καλλιέργειας: το DMEM σε ποσοστό 42.4% και το RPMI 1640 σε ποσοστό 33.6%.



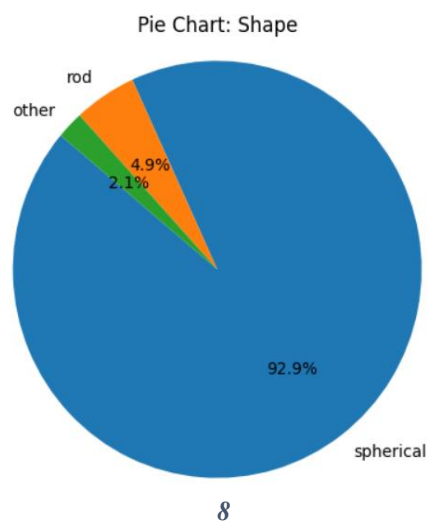
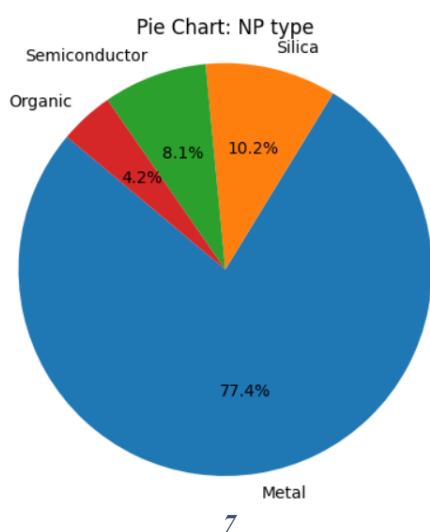
Διαγράμματα 4-6. Ιστογράμματα των μεταβλητών 4) Μέση υδροδυναμική διάμετρος 5) Ζ-δυναμικό 6) Ιοντική ισχύς (mol/L)

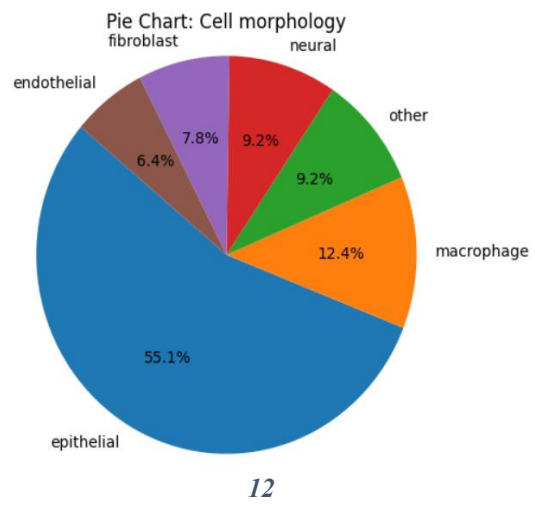
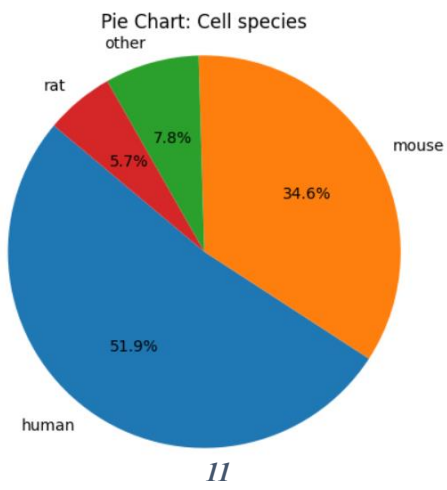
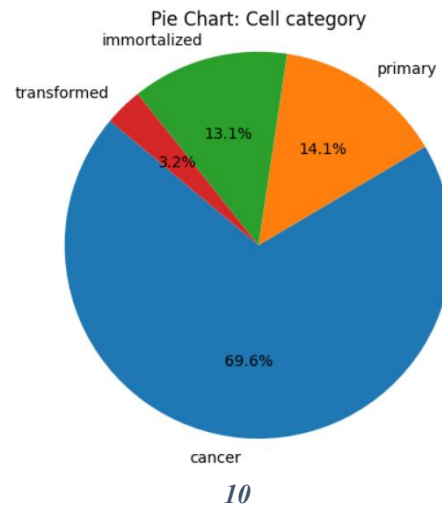
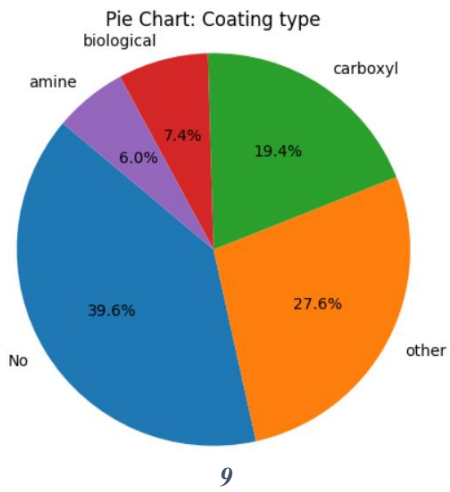
- Κατηγορικές μεταβλητές

Οι κατηγορικές μεταβλητές κωδικοποιούν σημαντικές πληροφορίες για το είδος των νανοσωματιδίων και των κυτταρικών σειρών αλλά και διάφορες πειραματικές παραμέτρους. Όπως φαίνεται από το Διάγραμμα 7, το 77.4% των παρατηρήσεων αφορά μεταλλικά νανοσωματίδια, ακολουθούμενα από νανοσωματίδια πυριτίας (10.2%) και κβαντικές νανοτελείες άνθρακα (quantum dots) ή ημιαγώγιμα νανοσωματίδια (8.1%). Επίσης, ένα μικρό ποσοστό (4.2%) νανοσωματιδίων είναι πολυμερικά ή οργανικά. Όσον αφορά το σχήμα των νανοσωματιδίων, αυτό είναι σχεδόν σε όλες τις περιπτώσεις σφαιρικό (92.9%) και μόνο το 4.9% αφορά νανοράβδους (Διάγραμμα 8). Τέλος, σύμφωνα με το Διάγραμμα 9, το 39.6% των νανοσωματιδίων δεν είχε κάποια επιφανειακή επικάλυψη, ενώ τα νανοσωματίδια με επιφανειακούς προσδέτες που περιείχαν καρβοξυλικές, αμυνικές ομάδες ή βιολογικά μόρια αντιστοιχούν στο 19.4%, 7.4% και 6.0% αντίστοιχα. Συμπερασματικά, η πλειοψηφία των νανοσωματιδίων του συνόλου δεδομένων είναι μεταλλικής φύσεως, σφαιρικού σχήματος και δεν έχουν επιφανειακή επικάλυψη.

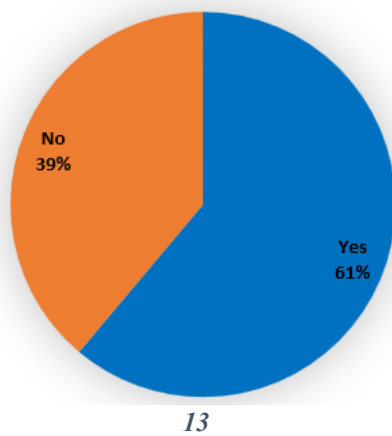
Όπως φαίνεται στο Διάγραμμα 10, οι κυτταρικές σειρές που χρησιμοποιούνται στα πειράματα πρόσληψης νανοσωματιδίων είναι σε ποσοστό 69.6% καρκινικές, ενώ σημαντικό ποσοστό καταλαμβάνουν τα πρωτογενή (14.1%) και τα αθάνατα (13.1%) κύτταρα. Τα πιο κοινά είδη προέλευσης των κυττάρων είναι -όπως είναι αναμενόμενο σε έρευνες με στόχο την εξαγωγή κλινικών συμπερασμάτων- ο άνθρωπος (51.9%) και τα ποντίκια (34.6%) (Διάγραμμα 11). Παράλληλα, σύμφωνα με το διάγραμμα σχετικά με την μορφολογία των κυττάρων, το 55.1% είναι επιθηλιακά κύτταρα ενώ οι υπόλοιπες παρατηρήσεις του συνόλου δεδομένων περιλαμβάνουν μακροφάγα (12.4%), νευρικά (9.2%), ινοβλάστες (7.8%) και ενδοθηλιακά (6.4%) κύτταρα (Διάγραμμα 12).

Η ανάλυση διαφόρων πειραματικών παραμέτρων ανέδειξε πως σε ποσοστό μεγαλύτερο του 60% των πειραμάτων κυτταρικής πρόσληψης πραγματοποιείται προσθήκη αντιβιοτικών πενικιλίνης και στρεπτομυκίνης στο μέσο καλλιέργειας κυττάρων (Διάγραμμα 13). Ακόμη, η εφαρμογή υπερήχων πριν την χρήση των νανοσωματιδίων παρατηρείται μόνο στο 17% των πειραμάτων (Διάγραμμα 14). Τέλος, όσον αφορά τις μεθόδους μέτρησης και ποσοτικοποίησης της κυτταρικής πρόσληψης, φαίνεται να αξιοποιούνται κυρίως μέθοδοι φασματομετρίας μάζας (58.0%), ακολουθούμενες από οπτικές (32.2%) και μικροσκοπικές (7.1%) μεθόδους (Διάγραμμα 15).

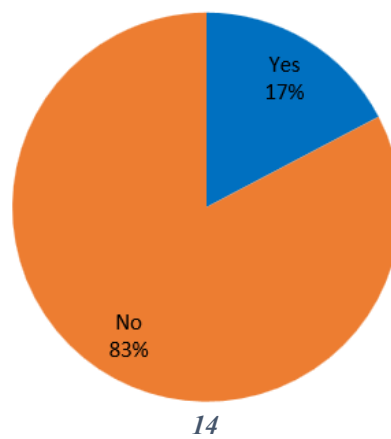


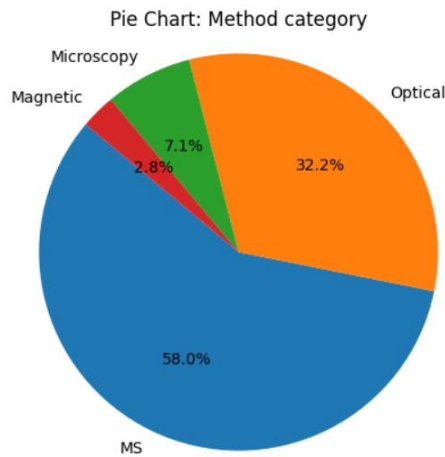


Pie Chart: Penicillin/Streptomycin (Yes/No)



Pie Chart: Sonication (Yes/No)





15

Διαγράμματα 7-15. Κατανομή των κατηγοριών των κατηγορικών μεταβλητών 7) Είδος ναυσοματιδίων 8) Σχήμα 9) Είδος επικάλυψης 10) Κατηγορία κυτταρικής σειράς 11) Είδος προέλευσης κυτταρικής σειράς 12) Μορφολογία κυτταρικής σειράς 13) Πενικιλίνη/Στρεπτομυκίνη (Ναι/Όχι) 14) Εφαρμογή υπερήχων 15) Κατηγορία μεθόδων ποσοτικοποίησης κυτταρικής πρόσληψης.

6.2.2. Εκπαίδευση και αξιολόγηση μοντέλου XGBoost

Το αναθεωρημένο σύνολο δεδομένων χρησιμοποιήθηκε για την εκπαίδευση ενός νέου μοντέλου XGBoost μετά την εκ νέου προεπεξεργασία των δεδομένων. Η διαδικασία της βελτιστοποίησης των υπερπαραμέτρων του αλγορίθμου μέσω αξιολόγησης του μοντέλου με χρήση του «cross-validation» είχε ως αποτέλεσμα την επιλογή των τιμών των υπερπαραμέτρων που παρουσιάζονται στον Πίνακα 9.

Πίνακας 9. Βέλτιστες τιμές των παραμέτρων του μοντέλου XGboost με βάση το στατιστικό μέτρο R^2 .

Παράμετρος	Βέλτιστες τιμές
n_estimators	200
max_depth	5
learning_rate	0.1
subsample	0.7
colsample_bytree	0.7
reg_alpha	0.1
reg_lambda	5
min_child_weight	1

Χρησιμοποιώντας αυτές τις τιμές των παραμέτρων, το μοντέλο XGBoost που εκπαιδεύτηκε μέσω «cross-validation» 10 υποδιαίρεσεων παρουσίασε ικανοποιητική προβλεπτική ικανότητα σύμφωνα με το μέτρο R^2 που υπολογίστηκε σε κάθε μία από τις 10 υποδιαίρεσεις του συνόλου εκπαίδευσης, όπως φαίνεται στον Πίνακα 10. Συγκεκριμένα, ο συντελεστής προσδιορισμού κυμαίνεται μεταξύ των τιμών 0.716 και 0.955, με εξαίρεση την 2^η υποδιαίρεση στην οποία παρουσιάζει πολύ χαμηλή τιμή. Αυτό πιθανόν να οφείλεται στον διαχωρισμό του συνόλου εκπαίδευσης με τέτοιο τρόπο ώστε στην συγκεκριμένη υποδιαίρεση που χρησιμοποιείται ως εσωτερικό σύνολο ελέγχου κατά την διαδικασία του «cross-validation» να περιλαμβάνονται κυρίως

παρατηρήσεις που διαφέρουν σημαντικά από αυτές στις οποίες έχει εκπαιδευτεί το μοντέλο. Φαίνεται ότι η πρόβλεψη της κυτταρικής πρόσληψης βάσει δεδομένων που αφορούν είδη νανοσωματιδίων ή κυττάρων που δεν περιλαμβάνονται στο σύνολο εκπαίδευσης οδηγεί σε ανακριβείς και μη έμπιστες προβλέψεις.

Πίνακας 10. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία του «cross-validation» στο αναθεωρημένο σύνολο εκπαίδευσης.

Αριθμός υποδιαίρεσης συνόλου δεδομένων	R^2
1	0.883
2	0.314
3	0.716
4	0.839
5	0.885
6	0.900
7	0.692
8	0.824
9	0.955
10	0.848
Μέσος Όρος	0.786

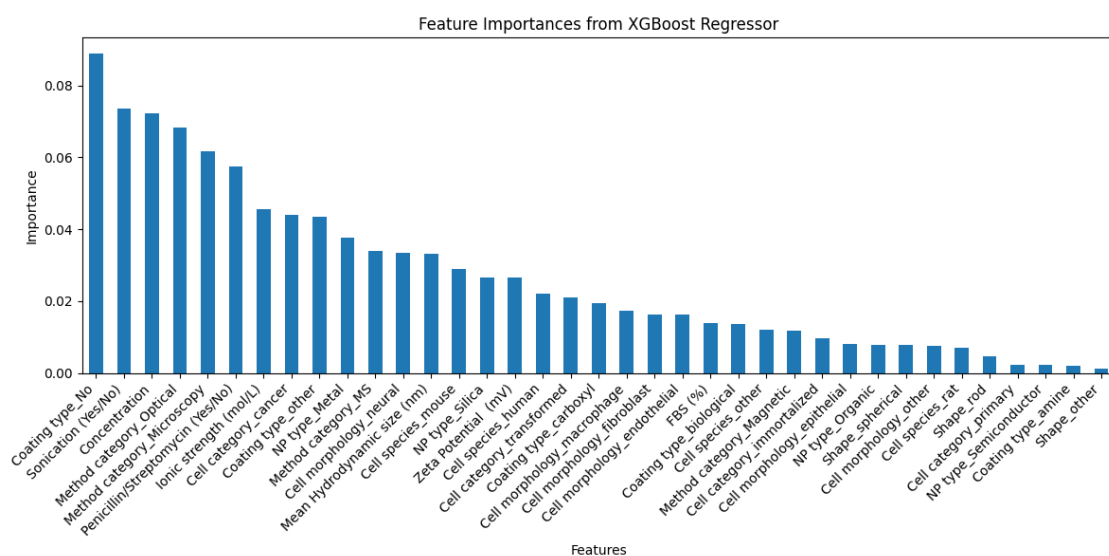
Η εφαρμογή του βελτιστοποιημένου μοντέλου στο σύνολο ελέγχου αξιολογείται ως εξαιρετικά ικανοποιητική, αφού ο συντελεστής προσδιορισμού υπολογίζεται αρκετά κοντά στις μέγιστες τιμές του R^2 στο σύνολο εκπαίδευσης κατά το «cross-validation», ενώ ταυτόχρονα οι δείκτες MAE, MSE και RMSE έχουν μικρές τιμές, γεγονός που υποδεικνύει την ύπαρξη σχετικά μικρών σφαλμάτων και την ικανοποιητική προσέγγιση των πραγματικών τιμών (Πίνακας 11). Η αξιολόγηση του μοντέλου με βάση τις προβλέψεις στο πραγματικό εύρος τιμών της μεταβλητής εξόδου οδηγεί σε σημαντική μείωση του συντελεστή προσδιορισμού στο 0.657 και σε πολύ μεγαλύτερες τιμές των μέτρων MAE, MSE και RMSE λόγω της μεγάλης κλίμακας μέτρησης της κυτταρικής πρόσληψης. Παρ' όλα αυτά, η συγκεκριμένη τιμή του R^2 μπορεί να θεωρηθεί ικανοποιητική για ένα τόσο περίπλοκο πρόβλημα αφού το μοντέλο φαίνεται να εξηγεί το 65.7% της διακύμανσης της μεταβλητής εξόδου στον πραγματικό χώρο τιμών.

Πίνακας 11. Στατιστικά μέτρα στο αναθεωρημένο σύνολο εκπαίδευσης πριν και μετά την λογαριθμική μετατροπή της μεταβλητής εξόδου για το μοντέλο XGBoost.

	Σύνολο ελέγχου μετά τη λογαριθμική μετατροπή της μεταβλητής εξόδου	Σύνολο ελέγχου στο πραγματικό εύρος τιμών της μεταβλητής εξόδου
R^2	0.929	0.657
MAE	1.219	5.63E+08
MSE	2.806	5.80E+18
RMSE	1.675	2.41E+09
Κανονικοποιημένο RMSE	-	0.0939

6.3. Αφαίρεση μεταβλητών και απλοποίηση μοντέλου XGBoost

Παρά την υψηλή προβλεπτική ικανότητα του μοντέλου που δημιουργήθηκε, η απεικόνιση της σημαντικότητας των μεταβλητών εισόδου για την λήψη αποφάσεων με βάση τον αλγόριθμο XGBoost ανέδειξε το πιθανό πρόβλημα της μείωσης της ατομικής επιρροής κάθε μεταβλητής λόγω της υψηλής διαστασιμότητας του χώρου των χαρακτηριστικών (Διάγραμμα 16). Καθώς το «βάρος» των αποφάσεων διαμοιράζεται σε 37 μεταβλητές, δυσχεραίνεται η αξιολόγηση της σημαντικότητάς τους και η ερμηνεία της σχέσης τους με την μεταβλητή εξόδου. Έτσι, φαίνεται πως πολλές από τις κατηγορίες των κατηγορικών μεταβλητών που μετατράπηκαν σε αριθμητικές έχουν σχεδόν μηδενική συνεισφορά στο μοντέλο.



Διάγραμμα 16. Σημαντικότητα μεταβλητών για την προβλεπτική ικανότητα του αναθεωρημένου μοντέλου XGBoost όπως προκύπτει από την ιδιότητα (attribute) «feature_importances_» της βιβλιοθήκης XGBoost στην Python.

Παρατηρώντας το διάγραμμα σημαντικότητας των μεταβλητών προκύπτει πως οι κατηγορίες με την μικρότερη συμμετοχή στις αποφάσεις του μοντέλου σχετίζονται με την Κατηγορία των κυτταρικών σειρών (πρωτογενή και αθάνατα), το Είδος προέλευσης των κυττάρων (αρουραίος και άλλο), το Είδος των νανοσωματιδίων (ημιάγωγιμα, οργανικά), το Ποσοστό FBS που προστίθεται στο μέσο καλλιέργειας και την Κατηγορία μεθόδου ποσοτικοποίησης της κυτταρικής πρόσληψης (μαγνητικές μέθοδοι). Ακόμη, μερικές υποκατηγορίες που ανήκουν στο Είδους της επικάλυψης των νανοσωματιδίων και του Σχήματος φαίνεται να έχουν μικρή συμμετοχή στην διαμόρφωση των προβλέψεων.

Στοχεύοντας στην δημιουργία ενός όσο το δυνατόν πιο απλού μοντέλου που διατηρεί τόσο την ικανότητα πρόβλεψης της κυτταρικής πρόσληψης νανοσωματιδίων με ικανοποιητική ακρίβεια όσο και τις σχέσεις μεταξύ των μεταβλητών, επιλέχθηκε η αφαίρεση ορισμένων χαρακτηριστικών που δεν συμβάλλουν σημαντικά στην προβλεπτική απόδοση του μοντέλου XGBoost. Η εύρεση των μεταβλητών που

μπορούν να αφαιρεθούν καθοδηγείται από το διάγραμμα σημαντικότητας (Διάγραμμα 16) και πραγματοποιείται μέσω δοκιμής και σφάλματος, ελέγχοντας δηλαδή τον αντίκτυπο της διαδοχικής αφαίρεσης μεταβλητών στα στατιστικά μέτρα των μοντέλων που εκπαιδεύονται στα νέα σύνολα δεδομένων που προκύπτουν σε κάθε περίπτωση. Βεβαίως, από αυτή την διαδικασία εξαιρούνται μεταβλητές που στις περισσότερες έρευνες αποτελούν την μοναδική διαφοροποίηση μεταξύ των διαφορετικών πειραμάτων, δηλαδή η Μέση υδροδυναμική διάμετρος και η Συγκέντρωση των νανοσωματιδίων στο μέσο καλλιέργειας.

Αρχικά, ελέγχθηκε η συμμετοχή στο μοντέλο σημαντικών σύμφωνα με την βιβλιογραφία παραμέτρων, όπως το Z-δυναμικό, η Επικάλυψη των νανοσωματιδίων, το Σχήμα τους αλλά και η Ιοντική ισχύς του μέσου καλλιέργειας. Για τον σκοπό αυτό μία εκ των μεταβλητών αφαιρέθηκε από το σύνολο δεδομένων σε κάθε περίπτωση, το μοντέλο εκπαιδεύτηκε στο νέο σύνολο εκπαίδευσης και τελικά αξιολογήθηκε με βάση τον συντελεστή προσδιορισμού για το σύνολο ελέγχου (στο πραγματικό εύρος τιμών της μεταβλητής εξόδου). Όπως εύκολα γίνεται αντιληπτό από τον Πίνακα 12, οι προαναφερθείσες μεταβλητές είναι σημαντικές για την επιτυχή πρόβλεψη της κυτταρικής πρόσληψης αφού η αφαίρεσή τους οδηγεί σε αισθητή μείωση του δείκτη R^2 (αρχική τιμή 0.657).

Πίνακας 12. Συντελεστής προσδιορισμού για το αναθεωρημένο σύνολο εκπαίδευσης μετά την αφαίρεση των μεταβλητών Z-δυναμικό, Επικάλυψη νανοσωματιδίων, Σχήμα νανοσωματιδίων και Ιοντική ισχύς του μέσου καλλιέργειας και την εκ νέου εκπαίδευση του μοντέλου XGBoost.

Μεταβλητή που αφαιρέθηκε	R^2 στο σύνολο ελέγχου για το πραγματικό εύρος τιμών της μεταβλητής εξόδου
Αρχικό μοντέλο	0.657
Z-δυναμικό (mV)	0.490
Επικάλυψη νανοσωματιδίων	0.355
Σχήμα νανοσωματιδίων	0.576
Ιοντική ισχύς (mol/L)	0.540

Έπειτα, δοκιμάστηκε η διαδοχική αφαίρεση ορισμένων εκ των μεταβλητών με τον χαμηλότερο δείκτη σημαντικότητας σύμφωνα με το Διάγραμμα 16 ή και μεταβλητών που δεν αναφέρονται στην βιβλιογραφία ως σημαντικοί παράγοντες για την κυτταρική πρόσληψη (π.χ. Πενικιλίνη/Στρεπτομυκίνη), ώστε να βρεθεί ο συνδυασμός μεταβλητών που όταν αφαιρεθούν οδηγούν στην ελάχιστη μείωση των στατιστικών μέτρων στο σύνολο ελέγχου. Συγκεκριμένα, αφαιρέθηκαν κατά σειρά οι μεταβλητές Κατηγορία μεθόδου ποσοτικοποίησης, Πενικιλίνη/Στρεπτομυκίνη, Είδος προέλευσης κυττάρων, Κατηγορία κυτταρικής σειράς, ποσοστό FBS και Κατηγορία νανοσωματιδίων. Ο δείκτης R^2 βελτιώθηκε μετά την αφαίρεση μεταβλητών και ανήλθε στην τιμή 0.668 για τις προβλέψεις στο κανονικό εύρος τιμών της μεταβλητής εξόδου.

Αφού βρέθηκε λοιπόν ο κατάλληλος συνδυασμός μεταβλητών που μπορεί να αφαιρεθεί χωρίς να επηρεάζει την προβλεπτική ικανότητα του μοντέλου, το σύνολο των

μεταβλητών εισόδου του απλοποιημένου πλέον μοντέλου διαμορφώνεται όπως φαίνεται στον Πίνακα 13.

Πίνακας 13. Μεταβλητές εισόδου που επιλέχθηκαν για την εκπαίδευση του απλοποιημένου μοντέλου XGBoost.

Είδος Μεταβλητής	Όνομα Μεταβλητής	Μονάδες Μέτρησης
Φυσικοχημικές Ιδιότητες νανοϋλικών	Σχήμα	-
	Μέση υδροδυναμική διάμετρος	nm
	Z-δυναμικό	mV
	Επικάλυψη νανοσωματιδίων	-
Πειραματικές Παράμετροι	Συγκέντρωση νανοσωματιδίων στο μέσο καλλιέργειας	nM
	Ιοντική ισχύς μέσου καλλιέργειας	mol/L
	Εφαρμογή υπερήχων	1/0
Χαρακτηριστικά κυτταρικής σειράς	Μορφολογία κυτταρικής σειράς	-

Είναι αξιοσημείωτο ότι οι επιλεγμένες μεταβλητές παρέχουν πληροφορίες τόσο για το ίδιο το νανοϋλικό μέσω των φυσικοχημικών του ιδιοτήτων, όσο και για τις πειραματικές συνθήκες και τα χαρακτηριστικά της κυτταρικής σειράς που χρησιμοποιήθηκε σε κάθε μελέτη κυτταρικής πρόσληψης. Με τον τρόπο αυτό, καθίσταται εφικτή η εξαγωγή ολοκληρωμένων συμπερασμάτων αναφορικά με τη σχέση των σημαντικότερων παραμέτρων με την κυτταρική πρόσληψη.

Επιπρόσθετα, ένα ιδιαίτερα σημαντικό πλεονέκτημα του προτεινόμενου απλοποιημένου μοντέλου είναι ότι επιτρέπει την εξαγωγή προβλέψεων σχετικά με την κυτταρική πρόσληψη νανοϋλικών χωρίς να απαιτείται εκτεταμένος ή πολύπλοκος πειραματικός σχεδιασμός. Η πλειονότητα των απαιτούμενων παραμέτρων του μοντέλου μπορούν να προσδιοριστούν στη φάση σχεδιασμού του πειράματος, βάσει βιβλιογραφικών δεδομένων ή ήδη διαθέσιμων πληροφοριών. Η μόνη απαραίτητη πειραματική διαδικασία για την εξαγωγή των προβλέψεων είναι η μέτρηση της υδροδυναμικής διαμέτρου και του επιφανειακού φορτίου (Z-δυναμικού) των νανοσωματιδίων. Οι τιμές αυτές αποκτώνται μέσω των τεχνικών DLS¹¹² και ηλεκτροφορητικής σκέδασης του φωτός (Electrophoretic Light Scattering, ELS) αντίστοιχα, οι οποίες συνήθως μετρώνται με τα ίδια όργανα (π.χ. Malvern Zetasizer)¹¹³. Αυτές οι τεχνικές αποτελούν μία απλή και καθιερωμένη μεθοδολογία στον τομέα του χαρακτηρισμού νανοϋλικών. Λόγω της ευρείας εφαρμογής και της χαμηλής τεχνικής απαίτησης των μεθόδων μέτρησης, το συγκεκριμένο πειραματικό βήμα δεν επιβαρύνει σημαντικά τον συνολικό πειραματικό φόρτο.

6.4. Αξιολόγηση και Πεδίο Εφαρμοσιμότητας απλοποιημένου μοντέλου

Το απλοποιημένο μοντέλο εκπαιδεύτηκε τελικά σε 8 μεταβλητές εισόδου, οι οποίες, μετά την μετατροπή των κατηγορικών μεταβλητών σε αριθμητικές, ανέρχονται σε 19 στήλες για 283 παρατηρήσεις.

Η εκ νέου βελτιστοποίηση των υπερπαραμέτρων του μοντέλου XGBoost μέσω της διαδικασίας του «cross-validation» οδήγησε στις ίδιες τιμές με το προηγούμενο μοντέλο, όπως αυτές αναγράφονται στον Πίνακα 9. Παράλληλα, η αξιολόγηση του μοντέλου στις 10 υποδιαίρεσεις του συνόλου εκπαίδευσης (Πίνακας 14) κρίθηκε ικανοποιητική με εύρος τιμών του συντελεστή προσδιορισμού από 0.641 έως 0.981. Εξάιρεση αποτελεί και πάλι η 2^η υποδιαίρεση για την οποία η πολύ χαμηλή τιμή του R^2 πιθανόν να οφείλεται σε μη αντιπροσωπευτικό διαχωρισμό του συνόλου εκπαίδευσης.

Πίνακας 14. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία του «cross-validation» στο απλοποιημένο σύνολο εκπαίδευσης.

Αριθμός υποδιαίρεσης συνόλου δεδομένων	R^2
1	0.871
2	0.250
3	0.641
4	0.875
5	0.921
6	0.981
7	0.694
8	0.801
9	0.947
10	0.841
Μέσος Όρος	0.782

Παράλληλα, το απλοποιημένο μοντέλο εξηγεί εξαιρετικά την διακύμανση της μεταβλητής εξόδου στον χώρο δεδομένων που ανήκουν οι τιμές της κυτταρικής πρόσληψης ναοσωματιδίων υπολογισμένες σε λογαριθμική κλίμακα, όπως αποδεικνύει η υψηλή τιμή του συντελεστή προσδιορισμού ($R^2 = 0.921$) που προσεγγίζει την μέγιστη τιμή του δείκτη κατά την διαδικασία του «cross-validation». Επιπροσθέτως, οι τιμές των δεικτών MAE, MSE και RMSE υποδηλώνουν την παρουσία μικρών μόνο σφαλμάτων στις προβλέψεις του μοντέλου (Πίνακας 15). Τέλος, όσον αφορά τις προβλέψεις στον πραγματικό χώρο των δεδομένων, το μοντέλο εξηγεί το 66.8% της διακύμανσης στον κανονικό χώρο του πεδίου τιμών μεταβλητής με έναν πολύ χαμηλό δείκτη σφαλμάτων (κανονικοποιημένο RMSE).

Πίνακας 15. Στατιστικά μέτρα στο απλοποιημένο σύνολο εκπαίδευσης πριν και μετά την λογαριθμική μετατροπή της μεταβλητής εξόδου για το μοντέλο XGBoost.

	Σύνολο ελέγχου μετά τη λογαριθμική μετατροπή της μεταβλητής εξόδου	Σύνολο ελέγχου στο πραγματικό εύρος τιμών της μεταβλητής εξόδου
R^2	0.921	0.668
MAE	1.184	5.62E+08
MSE	3.125	5.61E+18
RMSE	1.768	2.37E+09
Κανονικοποιημένο RMSE	-	0.0137

Εκτός από τα ευρέως χρησιμοποιούμενα στατιστικά μέτρα, η αξιολόγηση του μοντέλου πραγματοποιήθηκε και μέσω της ανάλυσης τυχαίας αντικατάστασης των τιμών της μεταβλητής εξόδου, σύμφωνα με την οποία οι τιμές της κυτταρικής πρόσληψης νανοσωματιδίων αναδιατάχθηκαν τυχαία 10 φορές και δημιουργήθηκαν ισάριθμα μοντέλα XGBoost με βάση τα νέα σύνολα δεδομένων. Τα μοντέλα αυτά εκπαιδεύτηκαν χρησιμοποιώντας τις βέλτιστες υπερπαραμέτρους του Πίνακα 9 και αξιολογήθηκαν με βάση τον δείκτη R^2 στο σύνολο ελέγχου.

Τα αποτελέσματα της διαδικασίας αυτής ανέδειξαν αρνητικές ή πολύ μικρές θετικές τιμές του συντελεστή προσδιορισμού (Πίνακας 16), γεγονός που επιβεβαιώνει την απουσία συστηματικής σχέσης μεταξύ των τυχαία αναδιαταγμένων τιμών της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών του μοντέλου. Ένα R^2 μικρότερο του 0.5 υποδηλώνει ότι το τυχαίο μοντέλο έχει χειρότερη προγνωστική ικανότητα ακόμη και από τη χρήση του απλού μέσου όρου των τιμών ως πρόβλεψη. Το εύρημα αυτό ενισχύει την εγκυρότητα και την μη τυχειότητα του αρχικού μοντέλου, καθώς αποδεικνύεται ότι οι επιδόσεις του δεν μπορούν να αναπαραχθούν όταν διαταραχθεί η συσχέτιση μεταξύ των μεταβλητών εισόδου και εξόδου.

Πίνακας 16. Τιμές του δείκτη R^2 του μοντέλου XGBoost κατά την διαδικασία της τυχαίας αντικατάστασης των τιμών της μεταβλητής εξόδου.

Αριθμός τυχαίας αντικατάστασης	R^2
1	-0.057
2	0.037
3	0.067
4	-5.495
5	-0.061
6	-0.061
7	-0.062
8	0.022
9	-0.061
10	-0.060

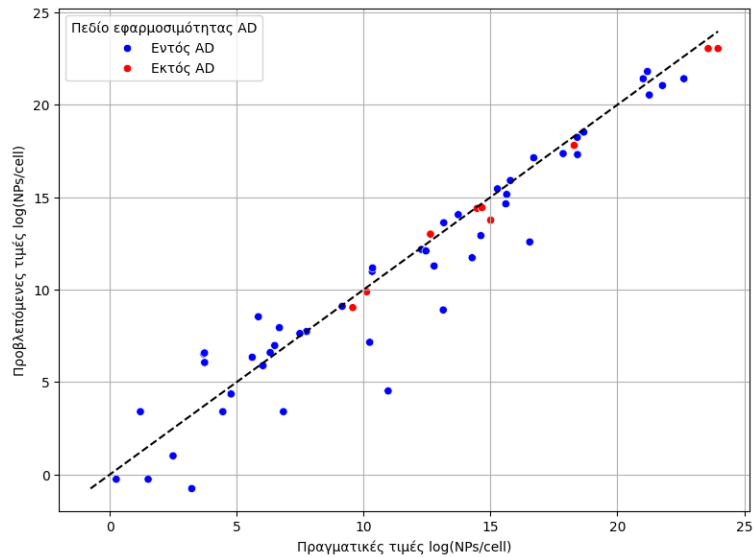
Αφού επιβεβαιώθηκε η εγκυρότητα των προβλέψεων του απλοποιημένου μοντέλου, προσδιορίστηκε το πεδίο εφαρμοσιμότητας, ώστε να είναι δυνατός ο εντοπισμός των μη έγκυρων ή επίφοβων προβλέψεων. Χρησιμοποιώντας την μέθοδο των k -κοντινότερων γειτόνων, με $k = 5$, το όριο της Ευκλείδειας απόστασης των νέων δεδομένων από του 5 εγγύτερους «γείτονες» τους στο σύνολο εκπαίδευσης ανέρχεται στην τιμή 7446.28 (Πίνακας 17). Με βάση αυτό το όριο, 48 από τις 57 προβλέψεις του μοντέλου στο σύνολο ελέγχου κρίθηκαν ως έγκυρες αφού βρίσκονται εντός του πεδίου εφαρμοσιμότητας. Το υψηλό ποσοστό (84.2%) προβλέψεων εντός του πεδίου εφαρμοσιμότητας δείχνει πως το μοντέλο εκπαιδεύτηκε σε αρκετά αντιπροσωπευτικό σύνολο δεδομένων ώστε να καλύπτει επαρκώς τα «άγνωστα» δεδομένα.

Πίνακας 17. Πεδίο εφαρμοσιμότητας του απλοποιημένου μοντέλου και αξιολόγηση των προβλέψεων στο σύνολο ελέγχου

Όριο Ευκλείδειας απόστασης	7446.28
Προβλέψεις εντός του Πεδίου Εφαρμοσιμότητας	48
Προβλέψεις εκτός του Πεδίου Εφαρμοσιμότητας	9

Η αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου στο σύνολο ελέγχου συμπληρώνεται από την γραφική αναπαράσταση των προβλεπόμενων τιμών του μοντέλου σε σχέση με τις πραγματικές τιμές, κατόπιν λογαριθμικής μετατροπής (Διάγραμμα 17). Στο διάγραμμα αυτό, τα ζεύγη πραγματικών (άξονας x') – προβλεπόμενων (άξονας y') τιμών παρουσιάζονται ως σημεία κατανομημένα γύρω από την διαγώνιο ευθεία $y = x$. Όσο πιο κοντά βρίσκονται τα σημεία σε αυτή τη διαγώνιο, τόσο πιο ακριβείς θεωρούνται οι προβλέψεις του μοντέλου, καθώς η προβλεπόμενη τιμή προσεγγίζει την πραγματική.

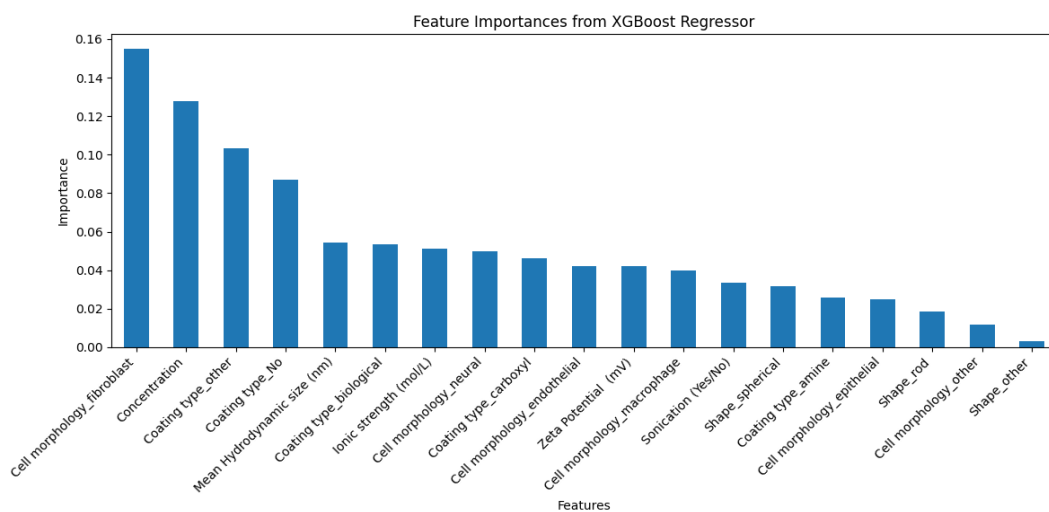
Επιπλέον, στο διάγραμμα παρουσιάζονται με μπλε χρώμα τα σημεία που αντιστοιχούν σε προβλέψεις που βρίσκονται εντός του πεδίου εφαρμοσιμότητας του μοντέλου με βάση το όριο της Ευκλείδειας απόστασης που υπολογίστηκε με την μέθοδο των k -κοντινότερων γειτόνων. Αντιθέτως, οι προβλέψεις που αντιστοιχούν σε σημεία κόκκινου χρώματος βρίσκονται εκτός του πεδίου εφαρμοσιμότητας και κρίνονται ως μη αξιόπιστες. Σημειώνεται πως, παρά το γεγονός ότι ορισμένα σημεία εκτός του πεδίου εφαρμοσιμότητας ενδέχεται να βρίσκονται πλησιέστερα στην διαγώνιο $y = x$ σε σύγκριση με κάποια σημεία μπλε χρώματος, το μοντέλο δεν έχει εκπαιδευτεί σε παρόμοια δεδομένα εισόδου και συνεπώς η πρόβλεψη χαρακτηρίζεται ως επισφαλής.



Διάγραμμα 17. Αντιπαραβολή προβλεπόμενων και πραγματικών τιμών (σε λογαριθμική κλίμακα) για το μοντέλο XGBoost. Οι μπλε κουκκίδες αντιστοιχούν σε δείγματα εντός του Πεδίου Εφαρμοσιμότητας (AD), ενώ οι κόκκινες σε δείγματα εκτός. Η εγγύτητα στην ευθεία $y = x$ υποδηλώνει μεγαλύτερη ακρίβεια πρόβλεψης.

6.5. Ερμηνεία απλοποιημένου μοντέλου

Η μελέτη της σημαντικότητας των μεταβλητών για τις αποφάσεις του απλοποιημένου μοντέλου XGBoost αποδεικνύει την χρησιμότητα της μείωσης του αριθμού των μεταβλητών εισόδου για την ευκολότερη αξιολόγηση των κύριων παραμέτρων του μοντέλου (Διάγραμμα 18). Πράγματι, παρατηρείται πως η αφαίρεση μεταβλητών με μικρή συμμετοχή στο μοντέλο μηχανικής μάθησης ανέδειξε την σημαντικότητα των υπόλοιπων παραμέτρων, αφού πλέον όλες οι αριθμητικές μεταβλητές έχουν μη μηδενική τιμή στο διάγραμμα σημαντικότητας.

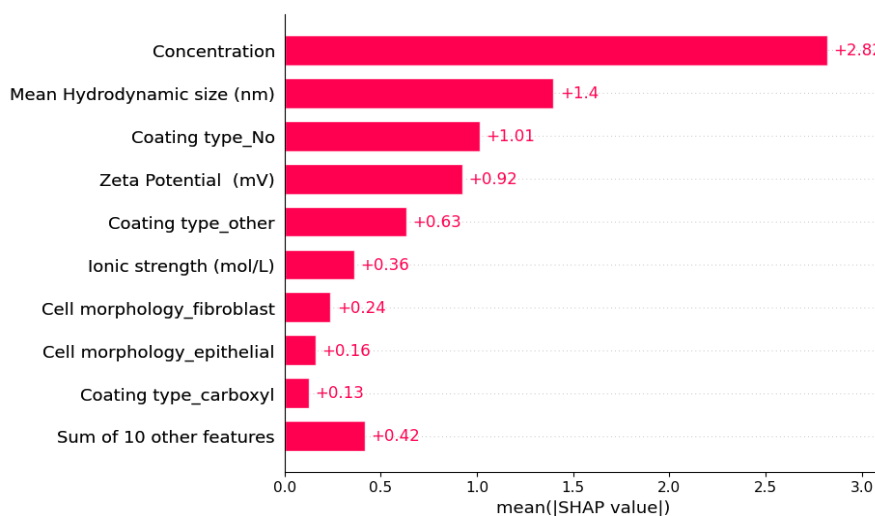


Διάγραμμα 18. Σημαντικότητα μεταβλητών για την προβλεπτική ικανότητα του απλοποιημένου μοντέλου XGBoost, όπως προκύπτει από την ιδιότητα (attribute) «feature_importances_» της βιβλιοθήκης XGBoost στην Python.

Μεγαλύτερη σημαντικότητα για την απόδοση του μοντέλου έχει η κατηγορία των ινοβλαστών (fibroblast), που ανήκει στην μεταβλητή Μορφολογία της κυτταρικής σειράς. Ακολουθεί η Συγκέντρωση των νανοσωματιδίων στο μέσο καλλιέργειας που, όπως είναι αναμενόμενο, επηρεάζει σημαντικά τις αποφάσεις του μοντέλου. Ακόμη, η Επικάλυψη των νανοσωματιδίων -και συγκεκριμένα οι κατηγορίες Καμία και Άλλη- επηρεάζουν σε μεγάλο ποσοστό τις αποφάσεις του μοντέλου, ξεπερνώντας την σημαντικότητα της Μέσης υδροδυναμικής διαμέτρου και του Ζ-δυναμικού, οι οποίες έχουν κοντινές τιμές στο Διάγραμμα σημαντικότητας. Βεβαίως, βάση της βιβλιογραφίας, οι δύο αυτές φυσικοχημικές ιδιότητες αναμένεται να αποτελούν καθοριστικές παραμέτρους της πρόσληψης των νανοσωματιδίων στα κύτταρα. Τέλος, μικρότερη αλλά μη αμελητέα συμμετοχή έχουν πειραματικές παράμετροι όπως η Ιοντική ισχύς του μέσου καλλιέργειας και η Εφαρμογή υπερήχων, ενώ η μόνη κατηγορία που φαίνεται να έχει πρακτικά μηδενική συμμετοχή στο μοντέλο είναι η κατηγορία Άλλο της μεταβλητής Σχήμα νανοσωματιδίων.

Σε κάθε περίπτωση, πρέπει να τονιστεί πως η ιδιότητα «.feature_importances_» της βιβλιοθήκης XGBoost της γλώσσας προγραμματισμού Python, η οποία αξιοποιήθηκε για την κατασκευή του Διαγράμματος 18, αποτελεί ένα χρήσιμο εργαλείο για την εύκολη και γρήγορη αποτύπωση της συνεισφοράς των μεταβλητών στην επίδοση του μοντέλου. Ωστόσο, δεν είναι σε θέση να αποκαλύψει τις περίπλοκες σχέσεις και αλληλεπιδράσεις μεταξύ των μεταβλητών. Παρότι μπορεί να αξιοποιηθεί για την ανάδειξη των σημαντικότερων χαρακτηριστικών σε σχέση με τη συνολική προβλεπτική ικανότητα του μοντέλου, η ερμηνευτική της αξία είναι περιορισμένη όταν πρόκειται για την εις βάθος κατανόηση του τρόπου με τον οποίο οι μεταβλητές επηρεάζουν τις επιμέρους προβλέψεις.

Όταν το ζητούμενο είναι η κατανόηση των σύνθετων αλληλεπιδράσεων και της συμβολής κάθε μεταβλητής στη διαμόρφωση των προβλέψεων του μοντέλου, η ανάλυση SHAP προσφέρει σαφέστερη εικόνα. Οι τιμές SHAP βασίζονται στις μεμονωμένες προβλέψεις και λαμβάνουν υπόψη τη συμβολή κάθε χαρακτηριστικού στο εκάστοτε αποτέλεσμα, παρέχοντας έτσι μια πληρέστερη και πιο διαφανή ερμηνεία του μοντέλου.



Διάγραμμα 19. Σημαντικότητα μεταβλητών για την διαμόρφωση των προβλέψεων του απλοποιημένου μοντέλου XGBoost σύμφωνα με τις τιμές SHAP.

Στο Διάγραμμα 19 παρουσιάζεται η ανάλυση των τιμών SHAP, για την ερμηνεία της σημασίας των μεταβλητών στις προβλέψεις του μοντέλου. Ο κατακόρυφος άξονας απεικονίζει τις επιμέρους μεταβλητές που εισήχθησαν στο μοντέλο, ενώ ο οριζόντιος άξονας δείχνει τη μέση απόλυτη τιμή SHAP για κάθε μεταβλητή. Η τιμή αυτή αντιπροσωπεύει το μέσο μέγεθος της επίδρασης κάθε χαρακτηριστικού στην έξοδο του μοντέλου, ανεξαρτήτως της κατεύθυνσης (θετικής ή αρνητικής). Όσο μεγαλύτερη είναι η τιμή στον οριζόντιο άξονα, τόσο μεγαλύτερη είναι και η σημασία της αντίστοιχης μεταβλητής στη διαμόρφωση των προβλέψεων.

Σύμφωνα με την ανάλυση, η μεγαλύτερη συμβολή στη διαμόρφωση των προβλέψεων του μοντέλου αποδίδεται στη μεταβλητή της Συγκέντρωσης ναυσοσωματιδίων στο μέσο καλλιέργειας των κυττάρων. Η αμέσως επόμενη σημαντική μεταβλητή είναι η Μέση υδροδυναμική διάμετρος των ναυσοσωματιδίων. Το εύρημα αυτό είναι αναμενόμενο, καθώς και οι δύο αυτές παράμετροι αποτελούν βασικά φυσικοχημικά χαρακτηριστικά που, όπως καταγράφεται εκτενώς στη βιβλιογραφία και στο παρόν σύνολο δεδομένων, επηρεάζουν άμεσα την κυτταρική πρόσληψη ναυσοσωματιδίων.

Αξιοσημείωτη είναι επίσης η υψηλή συμμετοχή της μεταβλητής που αντιστοιχεί στην απουσία Επιφανειακής επικάλυψης των ναυσοσωματιδίων, η οποία παρουσιάζει μεγαλύτερη σημασία από άλλες κατηγορίες επικάλυψης, όπως η καρβοξυλική επικάλυψη και η κατηγορία Άλλη επικάλυψη. Η παρατήρηση αυτή υποδηλώνει ότι η έλλειψη επικάλυψης μπορεί να αποτελεί καθοριστικό παράγοντα για την πρόβλεψη της πρόσληψης, πιθανώς λόγω της άμεσης αλληλεπίδρασης της επιφάνειας των ναυσοσωματιδίων με την κυτταρική μεμβράνη.

Επιπλέον, σημαντική είναι και η συμβολή του Z-δυναμικού, το οποίο, σύμφωνα με τη βιβλιογραφία, μπορεί να έχει αμφίσημη επίδραση στην κυτταρική πρόσληψη, ανάλογα με τις συνθήκες του πειράματος, το είδος μηχανισμού πρόσληψης που αξιοποιείται από τα διαφορετικά κύτταρα και το πρόσημο του επιφανειακού φορτίου του ναυσοσωματιδίου.

Τέλος, αν και με μικρότερη σχετική επίδραση, παρατηρείται συμμετοχή και άλλων μεταβλητών όπως η Ιοντική ισχύς του μέσου κυτταρικής καλλιέργειας, η Μορφολογία των κυττάρων -κυρίως συμμετέχουν οι κατηγορίες των ινοβλαστών και των επιθηλιακών κυττάρων- και το υπόλοιπο σύνολο χαρακτηριστικών, γεγονός που αναδεικνύει την πολυπαραγοντική φύση της κυτταρικής πρόσληψης ναυσοσωματιδίων.

Οι μέσες τιμές SHAP μπορούν να αναδείξουν τη σχετική συμβολή των μεταβλητών στη διαμόρφωση των προβλέψεων. Ωστόσο, εξίσου χρήσιμη είναι και η ανάλυση της κατεύθυνσης της επίδρασης κάθε μεταβλητής. Συγκεκριμένα, το διάγραμμα τύπου "beeswarm" απεικονίζει το εύρος και την κατεύθυνση της επίδρασης κάθε μίας εκ των μεταβλητών που συμμετέχουν στη λήψη αποφάσεων του μοντέλου XGBoost (Διάγραμμα 20).

Όσο μεγαλύτερη είναι η απόσταση των σημείων από το 0 στον οριζόντιο άξονα, τόσο ισχυρότερη είναι και η επίδραση της μεταβλητής στην πρόβλεψη του μοντέλου -είτε προς την αρνητική κατεύθυνση (αριστερά) είτε προς τη θετική (δεξιά). Παράλληλα, το χρώμα των σημείων κωδικοποιεί την αριθμητική τιμή κάθε μεταβλητής: οι χαμηλές τιμές εμφανίζονται με μπλε, οι υψηλές με κόκκινο και οι ενδιάμεσες με μωβ.



Διάγραμμα 20. Διάγραμμα τύπου «beeswarm» από την ανάλυση SHAP. Το γράφημα απεικονίζει το εύρος και την κατεύθυνση της επίδρασης κάθε μεταβλητής στις προβλέψεις του μοντέλου. Η απόσταση των σημείων από την τιμή 0 στον οριζόντιο άξονα αντιστοιχεί στο μέγεθος της επίδρασης (μεγαλύτερη απόσταση σημαίνει ισχυρότερη επίδραση), ενώ η θέση τους δεξιά ή αριστερά από το μηδέν δηλώνει θετική ή αρνητική συνεισφορά, αντίστοιχα. Το χρώμα κάθε σημείου υποδηλώνει την τιμή της αντίστοιχης μεταβλητής (μπλε για μικρές, κόκκινο για μεγάλες και μωβ για ενδιάμεσες τιμές).

- Συγκέντρωση νανοσωματιδίων στο μέσο καλλιέργειας κυττάρων

Στο διάγραμμα «beeswarm» που προέκυψε από την ανάλυση SHAP παρατηρείται ότι η μεταβλητή της Συγκέντρωσης νανοσωματιδίων έχει ένα μεγάλο εύρος τιμών «Shapley», γεγονός που αποδίδεται στην μεγάλη επίδραση της συγκεκριμένης μεταβλητής στην διαμόρφωση των προβλέψεων. Σημαντικό εύρημα της ανάλυσης

SHAP είναι η κατεύθυνση της επιρροής της μεταβλητής της Συγκέντρωσης νανοσωματιδίων. Συγκεκριμένα, στο Διάγραμμα 20 φαίνεται πως τα σημεία που βρίσκονται στα δεξιά της γραμμής μηδενικής τιμής «Shapley» του άξονα x, δηλαδή τα σημεία που έχουν την μεγαλύτερη θετική επίδραση στην κυτταρική πρόσληψη νανοσωματιδίων, έχουν μωβ και κόκκινο χρώμα. Έτσι, συμπεραίνεται πως μεγαλύτερες τιμές αρχικής συγκέντρωσης νανοσωματιδίων στο μέσο καλλιέργειας των κυττάρων οδηγούν σε μεγαλύτερη κυτταρική πρόσληψη.

Αυτό είναι εν μέρη αναμενόμενο καθώς στις περισσότερες πειραματικές μελέτες φαίνεται να υπάρχει μία σχεδόν αναλογική σχέση μεταξύ της αρχικής συγκέντρωσης των νανοσωματιδίων και της κυτταρικής πρόσληψης^{114,115}. Σε αρκετές περιπτώσεις όμως εμφανίζεται ένα πλατό στον αριθμό νανοσωματιδίων που εσωτερικεύονται στα κύτταρα πέρα από το οποίο οποιαδήποτε αύξηση της συγκέντρωσης δεν έχει σημαντικά αποτελέσματα στην κυτταρική πρόσληψη νανοϋλικών^{33,116}.

- Μέση υδροδυναμική διάμετρος νανοσωματιδίων

Μία από τις πιο σημαντικές φυσικοχημικές ιδιότητες των νανοσωματιδίων κατά την μελέτη της κυτταρικής τους πρόσληψης είναι, όπως προαναφέρθηκε, το μέγεθος τους και συγκεκριμένα η Υδροδυναμική τους διάμετρος που αποτελεί το λειτουργικό μέγεθος των νανοϋλικών που αλληλεπιδρούν με την κυτταρικές μεμβράνες. Η ανάλυση SHAP φανερώνει πως μεσαίες και μεγάλες τιμές Υδροδυναμικής διαμέτρου (σημεία με μωβ και κόκκινο χρώμα αντίστοιχα) έχουν αρνητική επίδραση στον αριθμό των εσωτερικευμένων νανοσωματιδίων καθώς εντοπίζονται κυρίως στην αρνητική πλευρά του οριζόντιου άξονα, γεγονός που συνδέεται με χαμηλότερη προβλεπόμενη κυτταρική πρόσληψη.

Αν και συχνά αναφέρεται στη βιβλιογραφία η ύπαρξη ενός βέλτιστου μεγέθους νανοσωματιδίων, περίπου 50 nm, για την αποδοτικότερη ενδοκυττάρωσή τους, πρέπει να τονιστεί πως η μελέτη και ο προσδιορισμός του ακριβούς μεγέθους των νανοϋλικών κατά την παραμονή τους σε βιολογικά υγρά είναι εξαιρετικά δύσκολος λόγω των αλληλεπιδράσεων αυτών με τις πρωτεΐνες του εξωκυττάρου υγρού αλλά και μεταξύ τους. Οι παρατηρήσεις, λοιπόν, σχετικά με την βέλτιστο μέγεθος των νανοσωματιδίων είναι αμφίβολες σε αρκετές περιπτώσεις. Έτσι, το γενικό συμπέρασμα πως μεγαλύτερα νανοσωματίδια τείνουν συνήθως να μειώνουν την δυνατότητα κυτταρικής πρόσληψης μπορεί να θεωρηθεί ασφαλές. Αυτό βεβαίως στηρίζεται στο γεγονός πως νανοσωματίδια ή συσσωματώματα νανοσωματιδίων διαμέτρου μεγαλύτερης των 150 nm περίπου δεν μπορούν να αξιοποιήσουν αποδοτικά τις πιο εξειδικευμένες οδούς ενδοκυττάρωσης, και κυρίως την εξαρτώμενη από κλαθρίνη ενδοκυττάρωση, με αποτέλεσμα να εξαρτώνται από λιγότερο αποδοτικούς μηχανισμούς που είτε δεν είναι επαρκτοί σε όλα τα είδη κυττάρων, όπως η φαγοκυττάρωση, είτε δεν είναι εκλεκτικοί, όπως η μακροπινοκύττωση.

- Επιφανειακή επικάλυψη νανοσωματιδίων

Όπως αναφέρθηκε στο Κεφάλαιο 2, η επιφανειακή επικάλυψη των νανοϋλικών έχει ως στόχο την σταθεροποίησή τους και συχνά την στόχευσή τους σε επιθυμητά κύτταρα ή ιστούς. Έτσι, όλα σχεδόν τα είδη επικάλυψης που μελετώνται σε πειραματικές έρευνες έχουν ως αποτέλεσμα την αύξηση της κυτταρικής πρόσληψης των νανοσωματιδίων. Ιδιαίτερο ενδιαφέρον παρουσιάζει λοιπόν το γεγονός πως η ανάλυση SHAP ανέδειξε την απουσία επιφανειακής επικάλυψης ως σημαντική παράμετρο που οδηγεί στην αύξηση της κυτταρικής πρόσληψης νανοσωματιδίων με τις μεγαλύτερες τιμές (τιμή 1 της δυαδικά κωδικοποιημένης μεταβλητής) να τοποθετούνται στην δεξιά πλευρά του οριζόντιου άξονα. Μάλιστα, φαίνεται ότι οι άλλοι τύποι επιφανειακής επικάλυψης και η επικάλυψη με μόρια που περιλαμβάνουν καρβοξυλικές ομάδες έχουν ως συνέπεια την μείωση της τιμής των εσωτερικευμένων νανοσωματιδίων. Αντίθετα, η επικάλυψη με αμυνικές ομάδες ή βιολογικά μόρια, παρά την μικρότερη συμμετοχή τους στην διαμόρφωση των προβλέψεων (μικρότερη απόσταση από την τιμή 0 του οριζόντιου άξονα των τιμών SHAP), οδηγούν σε αύξηση της κυτταρικής πρόσληψης.

Σημειώνεται πως τα μη αναμενόμενα αποτελέσματα σχετικά με την επίδραση της επικάλυψης νανοσωματιδίων στην εσωτερίκευση των νανοσωματιδίων είναι πιθανόν να οφείλονται εν μέρη στην ανισορροπία των κατηγοριών επικάλυψης των παρατηρήσεων που συμμετέχουν στο σύνολο δεδομένων (περίπου 40% των παρατηρήσεων αφορά νανοσωματίδια χωρίς κάποια επικάλυψη). Αξίζει όμως να σημειωθεί ότι οι περισσότερες ομάδες της κατηγορίας Άλλη επικάλυψη περιέχουν PEG, το οποίο, όπως έχει ήδη αναφερθεί, μπορεί να μειώσει την κυτταρική πρόσληψη νανοσωματιδίων. Αυτό οφείλεται στην έντονη υδροφιλικότητά του, η οποία παρεμποδίζει τον σχηματισμό πρωτεϊνικής κορόνας. Ακόμη, και άλλοι τύποι επικαλυψών, όπως αυτές που περιέχουν καρβοξυλικές ομάδες, δεν έχουν πάντα το αναμενόμενο αποτέλεσμα ενίσχυσης της κυτταρικής πρόσληψης.

- Z-δυναμικό νανοσωματιδίων

Το επιφανειακό φορτίο των νανοσωματιδίων, που αντιπροσωπεύεται στο συγκεκριμένο σύνολο δεδομένων από το Z-δυναμικό, είναι η τέταρτη κατά σειρά σημαντικότητα μεταβλητή σύμφωνα με την ανάλυση SHAP. Όπως είναι αναμενόμενο, το επιφανειακό φορτίο είναι ένας παράγοντας που δεν επηρεάζει με σταθερό τρόπο την κυτταρική πρόσληψη των νανοσωματιδίων παρά την ανάδειξη ορισμένων «μοτίβων» από διάφορες πειραματικές μελέτες. Στην συγκεκριμένη περίπτωση, φαίνεται ότι η επιρροή των μεσαίων τιμών Z-δυναμικού είναι κυρίως ουδέτερη, καθώς τα περισσότερα σημεία με μωβ χρώμα βρίσκονται κοντά στην γραμμή μηδενικών τιμών SHAP. Ταυτόχρονα, πιο δεξιά τοποθετούνται κυρίως οι παρατηρήσεις που χαρακτηρίζονται από πολύ μικρές (μπλε) ή πολύ μεγάλες (κόκκινες) τιμές Z-δυναμικού. Ακόμη, φαίνεται ότι μεγαλύτερη θετική επιρροή στις τιμές της κυτταρικής πρόσληψης έχουν μόνο οι μεγαλύτερες τιμές του Z-δυναμικού.

Οι παρατηρήσεις αυτές βρίσκονται σε συμφωνία με τις περισσότερες πειραματικές μελέτες, σύμφωνα με τις οποίες τα πιο έντονα φορτισμένα σωματίδια αλληλεπιδρούν πιο εύκολα με τις κυτταρικές μεμβράνες και εσωτερικεύονται πιο αποτελεσματικά.

Μάλιστα, παρ' όλο που το αρχικό φορτίο των νανοσωματιδίων είναι πιθανό να μεταβληθεί κατά την παρουσία τους στο εξωκυττάριο υγρό λόγω της δημιουργίας της πρωτεϊνικής κορόνας, έχει αποδειχθεί ότι γενικά τα πιο έντονα θετικά φορτισμένα σωματίδια εμφανίζουν πιο εκτεταμένη κυτταρική πρόσληψη.

- Ιοντική ισχύς του μέσου κυτταρικής καλλιέργειας

Η ανάλυση SHAP ανέδειξε την σημασία της Ιοντικής ισχύος του μέσου καλλιέργεια για την κυτταρική πρόσληψη των νανοσωματιδίων. Συγκεκριμένα, στο Διάγραμμα 20 φαίνεται πως τα μέσα καλλιέργειας με μεγαλύτερες τιμές ιοντικής ισχύος οδηγούν σε μικρότερες τιμές πρόσληψης νανοσωματιδίων στα κύτταρα. Το εύρημα αυτό -όπως ήδη εξηγήθηκε- στηρίζεται θεωρητικά στο φαινόμενο της μείωσης της ηλεκτρικής διπλοστοιβάδας των νανοσωματιδίων που βρίσκονται σε μέσα υψηλής ιοντικής ισχύος, με αποτέλεσμα την μείωση των αποστικών δυνάμεων που αναπτύσσονται μεταξύ τους και την εντατικοποίηση του φαινομένου της συσσωμάτωσης. Συνεπώς, τα νανοσωματίδια που βρίσκονται σε μέσα καλλιέργειας υψηλής ιοντικής ισχύος σχηματίζουν μεγαλύτερα συσσωματώματα που δυσχεραίνουν την κυτταρική τους πρόσληψη.

- Μορφολογία των κυττάρων

Η πιο σημαντική κατηγορία κυττάρων για την τιμή της πρόσληψης νανοσωματιδίων είναι, σύμφωνα με την ανάλυση SHAP, η κατηγορία των ινοβλαστών, των κυττάρων των συνδετικών ιστών¹¹⁷, που φαίνεται να σχετίζονται με πιο αποδοτική κυτταρική πρόσληψη σε σχέση με τα επιθηλιακά και ενδοθηλιακά κύτταρα. Επίσης, τα μακροφάγα κύτταρα, παρά τον ρόλο τους ως «επαγγελματίες» μακροφάγοι, έχουν μικρή συμμετοχή στην διαμόρφωση των προβλέψεων και δεν φαίνεται να επηρεάζουν με σταθερό τρόπο την κυτταρική πρόσληψη, γεγονός που πιθανώς οφείλεται στην ελλιπή τους εκπροσώπηση στο σύνολο δεδομένων. Τέλος, τα νευρικά κύτταρα έχουν σχεδόν μηδενική σημαντικότητα για την τιμή των προβλέψεων.

Πράγματι σε πειραματικές μελέτες έχει αποδειχθεί ότι, υπό συγκεκριμένες συνθήκες, η κυτταρική πρόσληψη νανοσωματιδίων πυριτίας μπορεί να είναι έως και 3 φορές μεγαλύτερη σε ινοβλάστες σε σχέση με επιθηλιακά κύτταρα¹¹⁸. Ακόμη, μελέτη της πρόσληψης νανοσωματιδίων χρυσού σε μακροφάγα κύτταρα κατέληξε στο συμπέρασμα πως η ικανότητα φαγοκυττάρωσής τους δεν είναι σταθερά υψηλή και εξαρτάται από τον φαινότυπό τους¹¹⁹.

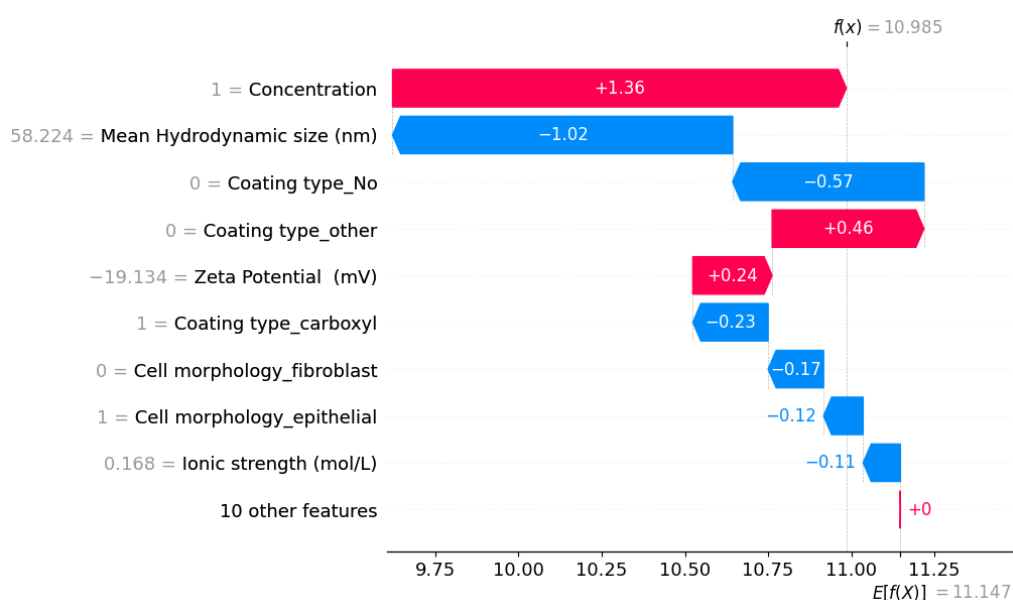
- Εφαρμογή υπερήχων κατά την παρασκευή των νανοσωματιδίων

Η μεταβλητή Εφαρμογή υπερήχων δεν έχει μεγάλη τιμή σημαντικότητας SHAP όμως φαίνεται πως μεγαλύτερες τιμές (δηλαδή η τιμή 1 της δυαδικά κωδικοποιημένης μεταβλητής) σχετίζονται με μεγαλύτερες τιμές κυτταρικής πρόσληψης. Η αύξηση της κυτταρικής πρόσληψης λόγω εφαρμογής υπερήχων κατά την παρασκευή των νανοσωματιδίων στηρίζεται βιβλιογραφικά στο γεγονός ότι οι υπέρηχοι σταθεροποιούν

τα νανοσωματίδια εμποδίζοντας την δημιουργία συσσωματωμάτων. Τελικά, η μικρότερη υδροδυναμική διάμετρος διευκολύνει την εισαγωγή των ναοϋλικών στα κύτταρα.

- Σχήμα των νανοσωματιδίων

Η μεταβλητή Σχήμα νανοσωματιδίων δεν εμφανίζει υψηλή σημαντικότητα στη διαμόρφωση των προβλέψεων του μοντέλου μηχανικής μάθησης. Ωστόσο, η ανάλυση SHAP ανέδειξε ότι τα ραβδοειδή νανοσωματίδια σχετίζονται με χαμηλότερα επίπεδα κυτταρικής πρόσληψης. Βιβλιογραφικά, υπάρχουν αρκετές μελέτες που υποστηρίζουν ότι τα σφαιρικά νανοσωματίδια ενσωματώνονται πιο εύκολα στα κύτταρα σε σύγκριση με τα επιμήκη. Παρ' όλα αυτά, νεότερα ερευνητικά δεδομένα αμφισβητούν αυτό το εύρημα, αναδεικνύοντας την πολυπλοκότητα της επίδρασης του σχήματος.



Διάγραμμα 21. Διάγραμμα «waterfall» επεξήγησης της επίδρασης των πιο σημαντικών μεταβλητών στην διαμόρφωση μίας μεμονωμένης πρόβλεψης.

Ως ένα παράδειγμα της διαμόρφωσης των προβλέψεων του μοντέλου, το γράφημα «waterfall» παρέχει μία αναλυτική ερμηνεία του τρόπου με τον οποίο κάθε χαρακτηριστικό συμβάλλει στην τελική πρόβλεψη του μοντέλου για μία συγκεκριμένη καταχώριση, αναδεικνύοντας τον μηχανισμό λήψης απόφασης του αλγορίθμου. Έτσι, το Διάγραμμα 21 απεικονίζει αναλυτικά τη συμβολή των 10 σημαντικότερων μεταβλητών στη διαμόρφωση μίας συγκεκριμένης πρόβλεψης του μοντέλου XGBoost (οι κόκκινες μπάρες δηλώνουν μεταβλητές που συνέβαλαν θετικά, αυξάνοντας την τιμή της πρόβλεψης, ενώ οι μπλε μπάρες αντιπροσωπεύουν μεταβλητές με αρνητική συνεισφορά, μειώνοντας την προβλεπόμενη τιμή).

Η σημαντικότερη θετική συνεισφορά προέρχεται από τη μεταβλητή Συγκέντρωση νανοσωματιδίων, η οποία αυξάνει την πρόβλεψη κατά +1.36. Αντίθετα, η Μέση

υδροδυναμική Διάμετρος μειώνει σημαντικά την τιμή της πρόβλεψης κατά -1.02, γεγονός που υποδηλώνει ότι μεγαλύτερα μεγέθη νανοσωματιδίων σχετίζονται με χαμηλότερη κυτταρική πρόσληψη. Επιπλέον, αρνητική συμβολή έχουν και άλλες μεταβλητές, όπως η απουσία επικάλυψης, η οποία επηρεάζει την πρόβλεψη κατά -0.57, καθώς και η κατηγορία των ινοβλαστών, που μειώνει την τιμή κατά 0.17. Αν και σύμφωνα με την ανάλυση SHAP τύπου «*beeswarm*» οι εν λόγω κατηγορίες συνδέονται με θετική επίδραση στην κυτταρική πρόσληψη όταν βρίσκονται σε υψηλές τιμές (τιμή 1, δηλαδή παρουσία της κατηγορίας), στην παρούσα περίπτωση έχουν τιμή 0, γεγονός που σημαίνει απουσία επικάλυψης και απουσία ινοβλαστικής μορφολογίας. Κατά συνέπεια, η συγκεκριμένη πρόβλεψη επηρεάζεται αρνητικά, καθώς απουσιάζουν χαρακτηριστικά που έχουν συσχετιστεί με αυξημένη ενδοκυττάρωση νανοσωματιδίων. Τέλος, μικρότερη αλλά θετική επίδραση καταγράφεται από το Z-δυναμικό που έχει έντονα αρνητική μέση τιμή -19.134 mV, ενώ η μεγάλη τιμή Ιοντικής ισχύος του μέσου καλλιέργειας «τιμωρείται» με μείωση της πρόβλεψης πρόσληψης νανοσωματιδίων κατά 0.11.

7. Συμπεράσματα – Προτάσεις για μελλοντική μελέτη

Στόχος της παρούσας μελέτης ήταν η δημιουργία και η αξιοποίηση ενός εκτενούς συνόλου δεδομένων που καθιστά δυνατή την πρόβλεψη της πρόσληψης ναοσωματιδίων από διάφορες κυτταρικές σειρές με μεγάλη ακρίβεια. Έπειτα από ενδελεχή μελέτη των διαθέσιμων βιβλιογραφικών πηγών, καταγραφή των πρωτογενών δεδομένων και επεξεργασία τους ώστε να προκύψει ένα συνεκτικό σύνολο δεδομένων, ακολούθησε η μοντελοποίηση και η βελτίωση των μοντέλων μηχανικής μάθησης με στόχο την δημιουργία ενός όσο το δυνατόν πιο απλού αλλά ερμηνεύσιμου μοντέλου.

Το βελτιστοποιημένο μοντέλο αξιοποιεί δεδομένα του Σχήματος, της Υδροδυναμικής διαμέτρου, του Z-δυναμικού και της Επικάλυψης των ναοσωματιδίων. Ακόμη, μεταβλητές εισόδου αποτελούν πειραματικές παράμετροι όπως η Συγκέντρωση των ναοσωματιδίων, η Ιοντική ισχύς του μέσου καλλιέργειας των κυττάρων και η Εφαρμογή υπερήχων, ενώ και η Μορφολογία της κυτταρικής σειράς είναι απαραίτητη είσοδος για το μοντέλο. Το μοντέλο XGBoost που εκπαιδεύτηκε με αυτά τα δεδομένα φαίνεται να προβλέπει με επιτυχία την κυτταρική πρόσληψη ναοσωματιδίων (ως Αριθμό ναοσωματιδίων ανά κύτταρο) καθώς πέτυχε πολύ ικανοποιητικά στατιστικά μέτρα αξιολόγησης. Συγκεκριμένα, ο δείκτης R^2 για τις προβλέψεις στο σύνολο ελέγχου ανέρχεται στην τιμή 0.921 μετά από λογαριθμική μετατροπή της μεταβλητής εξόδου και 0.668 στο πραγματικό εύρος της μεταβλητής εξόδου αντίστοιχα.

Η ερμηνευτική ανάλυση των σχέσεων μεταξύ των μεταβλητών ανέδειξε την Συγκέντρωση των ναοσωματιδίων ως την μεταβλητή με την μεγαλύτερη συμμετοχή στην δημιουργία των προβλέψεων, με την αύξηση αυτής να οδηγεί σε αύξηση της κυτταρικής πρόσληψης. Επίσης, φάνηκε ότι μικρότερες τιμές Υδροδυναμικής διαμέτρου και έντονα θετικές ή αρνητικές τιμές Z-δυναμικού έχουν ως αποτέλεσμα την αύξηση της ενδοκυττάρωσης ναοσωματιδίων. Εξαιρετικά σημαντική είναι και η Επικάλυψη των ναοϋλικών, η απουσία της οποίας φαίνεται να έχει θετική επίδραση στην κυτταρική πρόσληψη. Τέλος, με μικρότερη συμμετοχή στις προβλέψεις, η κυτταρική κατηγορία των ινοβλαστών φαίνεται να σχετίζεται με πιο αποδοτική εσωτερίκευση των ναοσωματιδίων, ενώ η αύξηση της Ιοντικής ισχύος του μέσου καλλιέργειας έχει το αντίθετο αποτέλεσμα.

Το μοντέλο παρουσιάζει πολύ ικανοποιητική προβλεπτική ικανότητα, αναδεικνύοντας σχέσεις μεταξύ φυσικοχημικών, βιολογικών και πειραματικών παραμέτρων, οι οποίες συμφωνούν με τις θεωρητικές βάσεις της κυτταρικής πρόσληψης ναοσωματιδίων. Ωστόσο, είναι σημαντικό να τονιστεί ότι ορισμένες παραδοχές και προσεγγίσεις που εφαρμόστηκαν ενδέχεται να επηρεάζουν τα αποτελέσματα της μοντελοποίησης. Παρά την αρχική καταγραφή των δεδομένων χωρίς κάποια επεξεργασία των πρωτογενών τιμών, ήταν απαραίτητη η κωδικοποίησή τους με συνεπή τρόπο. Έτσι, ναοσωματίδια διαφορετικού σχήματος θεωρήθηκαν προσεγγιστικά ως σφαίρες ώστε να συμπληρωθεί η απαραίτητη για το μοντέλο τιμή της διαμέτρου τους. Η πιο σημαντική όμως παραδοχή, που ήταν απαραίτητη για την μετατροπή μονάδων (από μάζα σε αριθμό ή συγκέντρωση και αντίστροφα), αφορά το πραγματικό μέγεθος και την πυκνότητα των ναοσωματιδίων. Συγκεκριμένα, τα ναοσωματίδια αντιμετωπίστηκαν ως τέλειες

σφαίρες συμπαγούς υλικού και έτσι η πυκνότητά τους θεωρήθηκε ίση με την πυκνότητα του υλικού κατασκευής τους. Βεβαίως, είναι γνωστό πως η κατασκευή τέλειων σφαιρικών σωματιδίων στην ναοκλίμακα είναι αδύνατη, αλλά και ότι τα νανοσωματίδια σχηματίζουν συσσωματώματα κατά την παραμονή τους σε υγρά μέσα καλλιέργειας. Στις περιπτώσεις που ήταν διαθέσιμη, αξιοποιήθηκε η τιμή της Υδροδυναμικής διαμέτρου ως πιο αντιπροσωπευτική ιδιότητα των συσσωματωμάτων, όμως το σφάλμα που υπεισέρχεται λόγω της προσεγγιστικής τιμής της πυκνότητας των νανοϋλικών παραμένει σημαντικό. Αυτό οφείλεται στο γεγονός ότι η πραγματική τιμή της πυκνότητας αναμένεται να είναι μικρότερη της ονομαστικής λόγω των κενών που δημιουργούνται μεταξύ των σωματιδίων κατά την συσσωμάτωσή τους. Τέλος, σημειώνεται ότι η συμπλήρωση ελλιπών τιμών (κυρίως της Υδροδυναμικής διαμέτρου) μέσω μεθοδολογίας παλινδρόμησης εισάγει επιπρόσθετο σφάλμα στα δεδομένα που χρησιμοποιούνται για την ανάπτυξη των μοντέλων μηχανικής μάθησης.

Δεδομένου ότι ένα μοντέλο μηχανικής μάθησης είναι τόσο αξιόπιστο όσο τα δεδομένα στα οποία βασίζεται, το μοντέλο XGBoost που δημιουργήθηκε θα μπορούσε να βελτιωθεί μέσω της ενίσχυσης της ποιότητας του συνόλου δεδομένων. Ένα βασικό ζητούμενο στην εποχή των «μεγάλων δεδομένων» είναι η τυποποίηση της καταγραφής πειραματικών δεδομένων. Στην προκειμένη περίπτωση, η υιοθέτηση ενός κοινού προτύπου πειραματικών διαδικασιών και ενιαίας μορφής καταγραφής θα συνέβαλλε στη δημιουργία ενός συνεκτικού και ομοιογενούς συνόλου δεδομένων, μειώνοντας την ανάγκη για εκτενή προεπεξεργασία. Με αυτόν τον τρόπο, θα εξαλείφονταν σημαντικές πηγές σφάλματος, κυρίως εκείνες που προκύπτουν από αναγκαίες παραδοχές κατά τη μετατροπή μονάδων, τόσο για τις μεταβλητές εισόδου όσο και για τη μεταβλητή εξόδου. Ακόμη, η εξέλιξη και εδραίωση της χρήσης πιο αξιόπιστων τεχνικών για τον προσδιορισμό του αριθμού νανοσωματιδίων ανά κύτταρο (π.χ. single particle ICP-MS) θα ήταν δυνατόν να αυξήσει την ακρίβεια των μετρήσεων των δεδομένων εκπαίδευσης και άρα την ποιότητα των προβλέψεων.

Η παρούσα μελέτη θέτει τις βάσεις για την μελλοντική εξέλιξη του μοντέλου πρόβλεψης της κυτταρικής πρόσληψης νανοσωματιδίων. Ο εμπλουτισμός του συνόλου δεδομένων με επιπλέον πειραματικά αποτελέσματα αποτελεί ένα πιθανό επόμενο βήμα. Αυτό θα επέτρεπε ενδεχομένως, επιπλέον, τη μελέτη της επίδρασης του χρόνου -μίας σημαντικής παραμέτρου στην ενδοκυττάρωση νανοϋλικών. Η αύξηση του όγκου των δεδομένων θα καθιστούσε δυνατή την μελέτη παραμέτρων που αφαιρέθηκαν ή απλοποιήθηκαν για την αποφυγή της υπέρμετρης αύξησης των μεταβλητών εισόδου σε σχέση με τον αριθμό παρατηρήσεων. Πιο συγκεκριμένα, θα μπορούσε να μελετηθεί η επίδραση όλων των πιθανών μορίων επιφανειακής κάλυψης των νανοσωματιδίων ή και των διαφορετικών κυτταρικών σειρών. Τέλος, μέσω της μελέτης του Χρόνου, σημαντική θα μπορούσε να είναι η συνεισφορά του μοντέλου στην κατανόηση των περίπλοκων μηχανισμών ενδοκυττάρωσης των νανοϋλικών μέσω συνδυασμού με μοντέλα κινητικής που βασίζονται στην φυσιολογία (Physiologically Based Kinetic, PBK). Στόχος αυτού θα ήταν η αξιοποίηση τόσο των θεωρητικών γνώσεων της κινητικής και της φυσιολογίας, όσο και των σχέσεων που μπορεί να αναδειχθεί ο μεγάλος όγκος των διαθέσιμων πειραματικών δεδομένων.

Καταληκτικά, το μοντέλο που αναπτύχθηκε αποτελεί ένα χρήσιμο εργαλείο για την πρόβλεψη της κυτταρικής πρόσληψης νανοσωματιδίων, με άμεσες εφαρμογές στον

σχεδιασμό στοχευμένων νανοϋλικών για φαρμακευτική και διαγνωστική χρήση. Η κατανόηση της συμπεριφοράς νέων νανοσωματιδίων κατά την μεταφορά τους στα κύτταρα-στόχους μπορεί να συμβάλει στη μείωση του πειραματικού κόστους και την επιτάχυνση της ανάπτυξης καινοτόμων νανοϊατρικών λύσεων.

Παράρτημα

Το σύνολο δεδομένων και ο κώδικας που αναπτύχθηκε για την δημιουργία και αξιολόγηση του μοντέλου XGBoost μπορούν να βρεθούν στον παρακάτω σύνδεσμο:

<https://github.com/ntua-unit-of-control-and-informatics/nm-cellular-uptake>

Βιβλιογραφία

- (1) Hobson, D. W.; Guy, R. C. Nanotoxicology. In *Encyclopedia of Toxicology (Third Edition)*; Wexler, P., Ed.; Academic Press: Oxford, 2014; pp 434–436. <https://doi.org/10.1016/B978-0-12-386454-3.01045-9>.
- (2) Yetisgin, A. A.; Cetinel, S.; Zuvin, M.; Kosar, A.; Kutlu, O. Therapeutic Nanoparticles and Their Targeted Delivery Applications. *Molecules* **2020**, *25* (9), 2193. <https://doi.org/10.3390/molecules25092193>.
- (3) Dutt, Y.; Pandey, R. P.; Dutt, M.; Gupta, A.; Vibhuti, A.; Vidic, J.; Raj, V. S.; Chang, C.-M.; Priyadarshini, A. Therapeutic Applications of Nanobiotechnology. *J. Nanobiotechnology* **2023**, *21* (1), 148. <https://doi.org/10.1186/s12951-023-01909-z>.
- (4) Rehan, F.; Zhang, M.; Fang, J.; Greish, K. Therapeutic Applications of Nanomedicine: Recent Developments and Future Perspectives. *Molecules* **2024**, *29* (9), 2073. <https://doi.org/10.3390/molecules29092073>.
- (5) Drasler, B.; Vanhecke, Dimitri; Rodriguez-Lorenzo, Laura; Petri-Fink, Alke; and Rothen-Rutishauser, B. Quantifying Nanoparticle Cellular Uptake: Which Method Is Best? *Nanomed.* **2017**, *12* (10), 1095–1099. <https://doi.org/10.2217/nnm-2017-0071>.
- (6) Elsaesser, A.; Taylor, A.; De Yanés, G. S.; McKerr, G.; Kim, E.-M.; O’Hare, E.; Howard, C. V. Quantification of Nanoparticle Uptake by Cells Using Microscopical and Analytical Techniques. *Nanomed.* **2010**, *5* (9), 1447–1457. <https://doi.org/10.2217/nnm.10.118>.
- (7) J. Richards, C.; Martinez, P. M.; H. Roos, W.; Åberg, C. High-Throughput Approach to Measure Number of Nanoparticles Associated with Cells: Size Dependence and Kinetic Parameters. *Nanoscale Adv.* **2025**, *7* (1), 185–195. <https://doi.org/10.1039/D4NA00589A>.
- (8) Kaplan, J. *Artificial Intelligence: What Everyone Needs to Know*; Oxford University Press, 2016.
- (9) Abbass, H. Editorial: What Is Artificial Intelligence? *IEEE Trans. Artif. Intell.* **2021**, *2* (2), 94–95. <https://doi.org/10.1109/TAI.2021.3096243>.
- (10) Bell, J. What Is Machine Learning? In *Machine Learning and the City*; John Wiley & Sons, Ltd, 2022; pp 207–216. <https://doi.org/10.1002/9781119815075.ch18>.
- (11) Zhang, Y. *New Advances in Machine Learning*; BoD – Books on Demand, 2010.
- (12) Hasanzadeh, A.; Hamblin, M. R.; Kiani, J.; Noori, H.; Hardie, J. M.; Karimi, M.; Shafiee, H. Could Artificial Intelligence Revolutionize the Development of Nanovectors for Gene Therapy and mRNA Vaccines? *Nano Today* **2022**, *47*, 101665. <https://doi.org/10.1016/j.nantod.2022.101665>.
- (13) Kibria, M. R.; Akbar, R. I.; Nidadavolu, P.; Havryliuk, O.; Lafond, S.; Azimi, S. Predicting Efficacy of Drug-Carrier Nanoparticle Designs for Cancer Treatment: A Machine Learning-Based Solution. *Sci. Rep.* **2023**, *13*, 547. <https://doi.org/10.1038/s41598-023-27729-7>.
- (14) Ding, D. Y.; Zhang, Y.; Jia, Y.; Sun, J. Machine Learning-Guided Lipid Nanoparticle Design for mRNA Delivery. arXiv August 29, 2023. <https://doi.org/10.48550/arXiv.2308.01402>.
- (15) Mi, K.; Chou, W.-C.; Chen, Q.; Yuan, L.; Kamineni, V. N.; Kuchimanchi, Y.; He, C.; Monteiro-Riviere, N. A.; Riviere, J. E.; Lin, Z. Predicting Tissue Distribution and Tumor Delivery of Nanoparticles in Mice Using Machine

- Learning Models. *J. Controlled Release* **2024**, *374*, 219–229. <https://doi.org/10.1016/j.jconrel.2024.08.015>.
- (16) Lu, B.; Jan. Hendriks, A.; Nolte, T. M. A Generic Model Based on the Properties of Nanoparticles and Cells for Predicting Cellular Uptake. *Colloids Surf. B Biointerfaces* **2022**, *209*, 112155. <https://doi.org/10.1016/j.colsurfb.2021.112155>.
- (17) Qi, R.; Pan, Y.; Cao, J.; Jia, Z.; Jiang, J. The Cytotoxicity of Nanomaterials: Modeling Multiple Human Cells Uptake of Functionalized Magneto-Fluorescent Nanoparticles via Nano-QSAR. *Chemosphere* **2020**, *249*, 126175. <https://doi.org/10.1016/j.chemosphere.2020.126175>.
- (18) Cheng, Y.-H.; He, C.; Riviere, J. E.; Monteiro-Riviere, N. A.; Lin, Z. Meta-Analysis of Nanoparticle Delivery to Tumors Using a Physiologically Based Pharmacokinetic Modeling and Simulation Approach. *ACS Nano* **2020**, *14* (3), 3075–3095. <https://doi.org/10.1021/acsnano.9b08142>.
- (19) Chou, W.-C.; Chen, Q.; Yuan, L.; Cheng, Y.-H.; He, C.; Monteiro-Riviere, N. A.; Riviere, J. E.; Lin, Z. An Artificial Intelligence-Assisted Physiologically-Based Pharmacokinetic Model to Predict Nanoparticle Delivery to Tumors in Mice. *J. Controlled Release* **2023**, *361*, 53–63. <https://doi.org/10.1016/j.jconrel.2023.07.040>.
- (20) Lin, Z.; Chou, W.-C.; Cheng, Y.-H.; He, C.; Monteiro-Riviere, N. A.; Riviere, J. E. Predicting Nanoparticle Delivery to Tumors Using Machine Learning and Artificial Intelligence Approaches. *Int. J. Nanomedicine* **2022**, *Volume 17*, 1365–1379. <https://doi.org/10.2147/IJN.S344208>.
- (21) Ghorbanzadeh, M.; Fatemi, M. H.; Karimpour, M. Modeling the Cellular Uptake of Magnetofluorescent Nanoparticles in Pancreatic Cancer Cells: A Quantitative Structure Activity Relationship Study. *Ind. Eng. Chem. Res.* **2012**, *51* (32), 10712–10718. <https://doi.org/10.1021/ie3006947>.
- (22) Bilgi, E.; Winkler, David A.; and Oksel Karakus, C. Identifying Factors Controlling Cellular Uptake of Gold Nanoparticles by Machine Learning. *J. Drug Target.* **2024**, *32* (1), 66–73. <https://doi.org/10.1080/1061186X.2023.2288995>.
- (23) Alafeef, M.; Srivastava, I.; Pan, D. Machine Learning for Precision Breast Cancer Diagnosis and Prediction of the Nanoparticle Cellular Internalization. *ACS Sens.* **2020**, *5* (6), 1689–1698. <https://doi.org/10.1021/acssensors.0c00329>.
- (24) Ali, R.; Balamurali, M.; Varamini, P. Deep Learning-Based Artificial Intelligence to Investigate Targeted Nanoparticles' Uptake in TNBC Cells. *Int. J. Mol. Sci.* **2022**, *23* (24), 16070. <https://doi.org/10.3390/ijms232416070>.
- (25) Kettler, K.; Veltman, K.; Van De Meent, D.; Van Wezel, A.; Hendriks, A. J. Cellular Uptake of Nanoparticles as Determined by Particle Properties, Experimental Conditions, and Cell Type. *Environ. Toxicol. Chem.* **2013**, *33* (3), 481–492. <https://doi.org/10.1002/etc.2470>.
- (26) Foroozandeh, P.; Aziz, A. A. Insight into Cellular Uptake and Intracellular Trafficking of Nanoparticles. *Nanoscale Res. Lett.* **2018**, *13* (1), 339. <https://doi.org/10.1186/s11671-018-2728-6>.
- (27) Salatin, S.; Maleki Dizaj, S.; Yari Khosroushahi, A. Effect of the Surface Modification, Size, and Shape on Cellular Uptake of Nanoparticles. *Cell Biol. Int.* **2015**, *39* (8), 881–890. <https://doi.org/10.1002/cbin.10459>.
- (28) Augustine, R.; Hasan, A.; Primavera, R.; Wilson, R. J.; Thakor, A. S.; Kevadiya, B. D. Cellular Uptake and Retention of Nanoparticles: Insights on

- Particle Properties and Interaction with Cellular Components. *Mater. Today Commun.* **2020**, *25*, 101692. <https://doi.org/10.1016/j.mtcomm.2020.101692>.
- (29) Ngake, T.; Nqayi, S.; Gulumian, M.; Cronjé, S.; Harris, R. A. Recent Developments in Computational and Experimental Studies of Physicochemical Properties of Au and Ag Nanostructures on Cellular Uptake and Nanostructure Toxicity. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2022**, *1866* (8), 130170. <https://doi.org/10.1016/j.bbagen.2022.130170>.
- (30) Duan, X.; Li, Y. Physicochemical Characteristics of Nanoparticles Affect Circulation, Biodistribution, Cellular Internalization, and Trafficking. *Small* **2013**, *9* (9–10), 1521–1532. <https://doi.org/10.1002/sml.201201390>.
- (31) Gao, H.; Shi, W.; Freund, L. B. Mechanics of Receptor-Mediated Endocytosis. *Proc. Natl. Acad. Sci.* **2005**, *102* (27), 9469–9474. <https://doi.org/10.1073/pnas.0503879102>.
- (32) Yue, H.; Wei, W.; Yue, Z.; Lv, P.; Wang, L.; Ma, G.; Su, Z. Particle Size Affects the Cellular Response in Macrophages. *Eur. J. Pharm. Sci.* **2010**, *41* (5), 650–657. <https://doi.org/10.1016/j.ejps.2010.09.006>.
- (33) Chithrani, B. D.; Ghazani, A. A.; Chan, W. C. W. Determining the Size and Shape Dependence of Gold Nanoparticle Uptake into Mammalian Cells. *Nano Lett.* **2006**, *6* (4), 662–668. <https://doi.org/10.1021/nl052396o>.
- (34) Gratton, S. E. A.; Ropp, P. A.; Pohlhaus, P. D.; Luft, J. C.; Madden, V. J.; Napier, M. E.; DeSimone, J. M. The Effect of Particle Design on Cellular Internalization Pathways. *Proc. Natl. Acad. Sci.* **2008**, *105* (33), 11613–11618. <https://doi.org/10.1073/pnas.0801763105>.
- (35) Agarwal, R.; Singh, V.; Journey, P.; Shi, L.; Sreenivasan, S. V.; Roy, K. Mammalian Cells Preferentially Internalize Hydrogel Nanodiscs over Nanorods and Use Shape-Specific Uptake Mechanisms. *Proc. Natl. Acad. Sci.* **2013**, *110* (43), 17247–17252. <https://doi.org/10.1073/pnas.1305000110>.
- (36) Huang, X.; Teng, X.; Chen, D.; Tang, F.; He, J. The Effect of the Shape of Mesoporous Silica Nanoparticles on Cellular Uptake and Cell Function. *Biomaterials* **2010**, *31* (3), 438–448. <https://doi.org/10.1016/j.biomaterials.2009.09.060>.
- (37) Li, Y.; Yue, T.; Yang, K.; Zhang, X. Molecular Modeling of the Relationship between Nanoparticle Shape Anisotropy and Endocytosis Kinetics. *Biomaterials* **2012**, *33* (19), 4965–4973. <https://doi.org/10.1016/j.biomaterials.2012.03.044>.
- (38) Jin, Q.; Xu, J.-P.; Ji, J.; Shen, J.-C. Zwitterionic Phosphorylcholine as a Better Ligand for Stabilizing Large Biocompatible Gold Nanoparticles. *Chem. Commun.* **2008**, No. 26, 3058. <https://doi.org/10.1039/b801959b>.
- (39) Hanaor, D.; Michelazzi, M.; Leonelli, C.; Sorrell, C. C. The Effects of Carboxylic Acids on the Aqueous Dispersion and Electrophoretic Deposition of ZrO₂. *J. Eur. Ceram. Soc.* **2012**, *32* (1), 235–244. <https://doi.org/10.1016/j.jeurceramsoc.2011.08.015>.
- (40) Alkilany, A. M.; Nagaria, P. K.; Hexel, C. R.; Shaw, T. J.; Murphy, C. J.; Wyatt, M. D. Cellular Uptake and Cytotoxicity of Gold Nanorods: Molecular Origin of Cytotoxicity and Surface Effects. *Small* **2009**, *5* (6), 701–708. <https://doi.org/10.1002/sml.200801546>.
- (41) Rizzo, L. Y.; Theek, B.; Storm, G.; Kiessling, F.; Lammers, T. Recent Progress in Nanomedicine: Therapeutic, Diagnostic and Theranostic Applications. *Curr. Opin. Biotechnol.* **2013**, *24* (6), 1159–1166. <https://doi.org/10.1016/j.copbio.2013.02.020>.

- (42) Vandamme, Th. F.; Brobeck, L. Poly(Amidoamine) Dendrimers as Ophthalmic Vehicles for Ocular Delivery of Pilocarpine Nitrate and Tropicamide. *J. Controlled Release* **2005**, *102* (1), 23–38. <https://doi.org/10.1016/j.jconrel.2004.09.015>.
- (43) Zhang, L.; Liu, H.; Xin, Q.; Tang, L.; Tang, J.; Liu, Y.; Hu, L. A Quantitative Study of Nanoplastics within Cells Using Magnetic Resonance Imaging. *Sci. Total Environ.* **2023**, *886*, 164033. <https://doi.org/10.1016/j.scitotenv.2023.164033>.
- (44) Milić, M.; Leitinger, G.; Pavičić, I.; Zebić Avdičević, M.; Dobrović, S.; Goessler, W.; Vinković Vrček, I. Cellular Uptake and Toxicity Effects of Silver Nanoparticles in Mammalian Kidney Cells. *J. Appl. Toxicol.* **2015**, *35* (6), 581–592. <https://doi.org/10.1002/jat.3081>.
- (45) Gonçalves, D. R.; Leroy, J. L. M. R.; Van Hees, S.; Xhonneux, I.; Bols, P. E. J.; Kiekens, F.; Marei, W. F. A. Cellular Uptake of Polymeric Nanoparticles by Bovine Cumulus-Oocyte Complexes and Their Effect on in Vitro Developmental Competence. *Eur. J. Pharm. Biopharm.* **2021**, *158*, 143–155. <https://doi.org/10.1016/j.ejpb.2020.11.011>.
- (46) Ulusoy, E.; Derman, S.; Erişen, S. The Cellular Uptake, Distribution and Toxicity of Poly (Lactic-Co-Glycolic) Acid Nanoparticles in Medicago Sativa Suspension Culture. *Romanian Biotechnol. Lett.* **2020**, *25* (3). <https://doi.org/10.25083/rbl/25.3/1572.1580>.
- (47) Cho, C.-W.; Keum, C.-G.; Noh, Y.-W.; Shin, S.-C. Practical Preparation Procedures for Docetaxel-Loaded Nanoparticles Using Polylactic Acid-Co-Glycolic Acid. *Int. J. Nanomedicine* **2011**, 2225. <https://doi.org/10.2147/IJN.S24547>.
- (48) Delgado, A.; Matijević, E. Particle Size Distribution of Inorganic Colloidal Dispersions: A Comparison of Different Techniques. *Part. Part. Syst. Charact.* **1991**, *8* (1–4), 128–135. <https://doi.org/10.1002/ppsc.19910080124>.
- (49) Zhao, F.; Zhao, Y.; Liu, Y.; Chang, X.; Chen, C.; Zhao, Y. Cellular Uptake, Intracellular Trafficking, and Cytotoxicity of Nanomaterials. *Small* **2011**, *7* (10), 1322–1337. <https://doi.org/10.1002/sml.201100001>.
- (50) Conner, S. D.; Schmid, S. L. Regulated Portals of Entry into the Cell. *Nature* **2003**, *422* (6927), 37–44. <https://doi.org/10.1038/nature01451>.
- (51) Behzadi, S.; Serpooshan, V.; Tao, W.; Hamaly, M. A.; Alkawareek, M. Y.; Dreaden, E. C.; Brown, D.; Alkilany, A. M.; Farokhzad, O. C.; Mahmoudi, M. Cellular Uptake of Nanoparticles: Journey inside the Cell. *Chem. Soc. Rev.* **2017**, *46* (14), 4218–4244. <https://doi.org/10.1039/C6CS00636A>.
- (52) Lim, J. P.; Gleeson, P. A. Macropinocytosis: An Endocytic Pathway for Internalising Large Gulps. *Immunol. Cell Biol.* **2011**, *89* (8), 836–843. <https://doi.org/10.1038/icb.2011.20>.
- (53) Mosquera, J.; García, I.; Liz-Marzán, L. M. Cellular Uptake of Nanoparticles versus Small Molecules: A Matter of Size. *Acc. Chem. Res.* **2018**, *51* (9), 2305–2313. <https://doi.org/10.1021/acs.accounts.8b00292>.
- (54) Manzanares, D.; Ceña, V. Endocytosis: The Nanoparticle and Submicron Nanocompounds Gateway into the Cell. *Pharmaceutics* **2020**, *12* (4), 371. <https://doi.org/10.3390/pharmaceutics12040371>.
- (55) Harush-Frenkel, O.; Debotton, N.; Benita, S.; Altschuler, Y. Targeting of Nanoparticles to the Clathrin-Mediated Endocytic Pathway. *Biochem. Biophys. Res. Commun.* **2007**, *353* (1), 26–32. <https://doi.org/10.1016/j.bbrc.2006.11.135>.

- (56) Wang, Z.; Tiruppathi, C.; Minshall, R. D.; Malik, A. B. Size and Dynamics of Caveolae Studied Using Nanoparticles in Living Endothelial Cells. *ACS Nano* **2009**, *3* (12), 4110–4116. <https://doi.org/10.1021/nn9012274>.
- (57) Ho, Y. T.; Kamm, R. D.; Kah, J. C. Y. Influence of Protein Corona and Caveolae-Mediated Endocytosis on Nanoparticle Uptake and Transcytosis. *Nanoscale* **2018**, *10* (26), 12386–12397. <https://doi.org/10.1039/C8NR02393J>.
- (58) Sandvig, K.; Pust, S.; Skotland, T.; van Deurs, B. Clathrin-Independent Endocytosis: Mechanisms and Function. *Curr. Opin. Cell Biol.* **2011**, *23* (4), 413–420. <https://doi.org/10.1016/j.ceb.2011.03.007>.
- (59) Boucrot, E.; Ferreira, A. P. A.; Almeida-Souza, L.; Debard, S.; Vallis, Y.; Howard, G.; Bertot, L.; Sauvonnnet, N.; McMahon, H. T. Endophilin Marks and Controls a Clathrin-Independent Endocytic Pathway. *Nature* **2015**, *517* (7535), 460–465. <https://doi.org/10.1038/nature14067>.
- (60) Sandvig, K.; Kavaliauskiene, S.; Skotland, T. Clathrin-Independent Endocytosis: An Increasing Degree of Complexity. *Histochem. Cell Biol.* **2018**, *150* (2), 107–118. <https://doi.org/10.1007/s00418-018-1678-5>.
- (61) Maharana, K.; Mondal, S.; Nemade, B. A Review: Data Pre-Processing and Data Augmentation Techniques. *Glob. Transit. Proc.* **2022**, *3* (1), 91–99. <https://doi.org/10.1016/j.gltip.2022.04.020>.
- (62) Westphal, M.; Brannath, W. Improving Model Selection by Employing the Test Data. In *Proceedings of the 36th International Conference on Machine Learning*; PMLR, 2019; pp 6747–6756.
- (63) Hidayaturrohman, Q. A.; Hanada, E. Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure. *BioMedInformatics* **2024**, *4* (4), 2201–2212. <https://doi.org/10.3390/biomedinformatics4040118>.
- (64) Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2; IJCAI'95*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; pp 1137–1143.
- (65) Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2* (3), 249–262. <https://doi.org/10.1007/s41664-018-0068-2>.
- (66) Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4* (9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>.
- (67) Poslavskaia, E.; Korolev, A. Encoding Categorical Data: Is There yet Anything “hotter” than One-Hot Encoding? arXiv December 28, 2023. <https://doi.org/10.48550/arXiv.2312.16930>.
- (68) Jamell Ivor Samuels. One-Hot Encoding and Two-Hot Encoding: An Introduction. **2024**. <https://doi.org/10.13140/RG.2.2.21459.76327>.
- (69) Kosaraju, N.; Sankepally, S. R.; Mallikharjuna Rao, K. Categorical Data: Need, Encoding, Selection of Encoding Method and Its Emergence in Machine Learning Models—A Practical Review Study on Heart Disease Prediction Dataset Using Pearson Correlation. In *Proceedings of International Conference on Data Science and Applications*; Saraswat, M., Chowdhury, C., Kumar Mandal, C., Gandomi, A. H., Eds.; Springer Nature: Singapore, 2023; pp 369–382. https://doi.org/10.1007/978-981-19-6631-6_26.

- (70) Mahmud Sujon, K.; Binti Hassan, R.; Tusnia Towshi, Z.; Othman, M. A.; Abdus Samad, M.; Choi, K. When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI. *IEEE Access* **2024**, *12*, 135300–135314. <https://doi.org/10.1109/ACCESS.2024.3462434>.
- (71) Han, J.; Kamber, M.; Pei, J. 3 - Data Preprocessing. In *Data Mining (Third Edition)*; Han, J., Kamber, M., Pei, J., Eds.; The Morgan Kaufmann Series in Data Management Systems; Morgan Kaufmann: Boston, 2012; pp 83–124. <https://doi.org/10.1016/B978-0-12-381479-1.00003-4>.
- (72) Buuren, S. V.; Groothuis-Oudshoorn, K. **Mice** : Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45* (3). <https://doi.org/10.18637/jss.v045.i03>.
- (73) Azur, M. J.; Stuart, E. A.; Frangakis, C.; Leaf, P. J. Multiple Imputation by Chained Equations: What Is It and How Does It Work? *Int. J. Methods Psychiatr. Res.* **2011**, *20* (1), 40–49. <https://doi.org/10.1002/mpr.329>.
- (74) White, I. R.; Royston, P.; Wood, A. M. Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Stat. Med.* **2011**, *30* (4), 377–399. <https://doi.org/10.1002/sim.4067>.
- (75) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016; pp 785–794. <https://doi.org/10.1145/2939672.2939785>.
- (76) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5). <https://doi.org/10.1214/aos/1013203451>.
- (77) Yao, X.; Fu, X.; Zong, C. Short-Term Load Forecasting Method Based on Feature Preference Strategy and LightGBM-XGboost. *IEEE Access* **2022**, *10*, 75257–75268. <https://doi.org/10.1109/ACCESS.2022.3192011>.
- (78) *XGBoost Parameters* — *xgboost 3.0.0 documentation*. https://xgboost.readthedocs.io/en/release_3.0.0/parameter.html (accessed 2025-04-28).
- (79) Berrar, D. Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology*; Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C., Eds.; Academic Press: Oxford, 2019; pp 542–545. <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>.
- (80) Yates, L. A.; Aandahl, Z.; Richards, S. A.; Brook, B. W. Cross Validation for Model Selection: A Review with Examples from Ecology. *Ecol. Monogr.* **2023**, *93* (1), e1557. <https://doi.org/10.1002/ecm.1557>.
- (81) Soper, D. S. Greed Is Good: Rapid Hyperparameter Optimization and Model Selection Using Greedy k-Fold Cross Validation. *Electronics* **2021**, *10* (16), 1973. <https://doi.org/10.3390/electronics10161973>.
- (82) Tatachar, A. V. Comparative Assessment of Regression Models Based On Model Evaluation Metrics. **2021**, *08* (09).
- (83) Botchkarev, A. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *Interdiscip. J. Inf. Knowl. Manag.* **2019**, *14*, 045–076. <https://doi.org/10.28945/4184>.
- (84) Rücker, C.; Rücker, G.; Meringer, M. Y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47* (6), 2345–2357. <https://doi.org/10.1021/ci700157b>.
- (85) Kaneko, H. Estimation of Predictive Performance for Test Data in Applicability Domains Using Y-randomization. *J. Chemom.* **2019**, *33* (9). <https://doi.org/10.1002/cem.3171>.

- (86) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17* (5), 4791–4810. <https://doi.org/10.3390/molecules17054791>.
- (87) Schultz, L. E.; Wang, Y.; Jacobs, R.; Morgan, D. A General Approach for Determining Applicability Domain of Machine Learning Models. *Npj Comput. Mater.* **2025**, *11* (1), 1–22. <https://doi.org/10.1038/s41524-025-01573-x>.
- (88) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a Novel K-Nearest Neighbours Approach to Assess the Applicability Domain of a QSAR Model for Reliable Predictions. *J. Cheminformatics* **2013**, *5* (1), 27. <https://doi.org/10.1186/1758-2946-5-27>.
- (89) 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307–317. In *Classics in Game Theory*; Princeton University Press, 1997; pp 69–79. <https://doi.org/10.1515/9781400829156-012>.
- (90) Štrumbelj, E.; Kononenko, I. An Efficient Explanation of Individual Classifications Using Game Theory. *J. Mach. Learn. Res.* **2010**, *11* (1), 1–18.
- (91) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2* (1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
- (92) Ponce-Bobadilla, A. V.; Schmitt, V.; Maier, C. S.; Mensing, S.; Stodtmann, S. Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development. *Clin. Transl. Sci.* **2024**, *17* (11), e70056. <https://doi.org/10.1111/cts.70056>.
- (93) *PubMed*. PubMed. <https://pubmed.ncbi.nlm.nih.gov/> (accessed 2025-04-26).
- (94) *ChatGPT*. <https://openai.com/chatgpt/overview/> (accessed 2025-04-26).
- (95) Castellani, C.; Radu, C. M.; Morillas-Becerril, L.; Barison, I.; Menato, F.; Do Nascimento, T. M.; Fedrigo, M.; Giarraputo, A.; Virzì, G. M.; Simioni, P.; Basso, C.; Papini, E.; Tavano, R.; Mancin, F.; Vescovo, G.; Angelini, A. Poly(Lipoic Acid)-Based Nanoparticles as a New Therapeutic Tool for Delivering Active Molecules. *Nanomedicine Nanotechnol. Biol. Med.* **2022**, *45*, 102593. <https://doi.org/10.1016/j.nano.2022.102593>.
- (96) Zhu, X.; Tao, W.; Liu, D.; Wu, J.; Guo, Z.; Ji, X.; Bharwani, Z.; Zhao, L.; Zhao, X.; Farokhzad, O. C.; Shi, J. Surface De-PEGylation Controls Nanoparticle-Mediated siRNA Delivery *In Vitro* and *In Vivo*. *Theranostics* **2017**, *7* (7), 1990–2002. <https://doi.org/10.7150/thno.18136>.
- (97) *automeris.io: Computer vision assisted data extraction from charts using WebPlotDigitizer*. <https://automeris.io/> (accessed 2025-04-26).
- (98) *PubChem*. *PubChem*. <https://pubchem.ncbi.nlm.nih.gov/> (accessed 2025-04-26).
- (99) *Online Materials Information Resource - MatWeb*. <https://www.matweb.com/> (accessed 2025-04-26).
- (100) Min, K. A.; Shin, M. C.; Yu, F.; Yang, M.; David, A. E.; Yang, V. C.; Rosania, G. R. Pulsed Magnetic Field Improves the Transport of Iron Oxide Nanoparticles through Cell Barriers. *ACS Nano* **2013**, *7* (3), 2161–2171. <https://doi.org/10.1021/nl3057565>.
- (101) Bashir, D.; Montanez, G. D.; Sehra, S.; Segura, P. S.; Lauw, J. An Information-Theoretic Perspective on Overfitting and Underfitting. arXiv 2020. <https://doi.org/10.48550/ARXIV.2010.06076>.

- (102) Truscott, S. M. Laboratory Calculations. In *Contemporary Practice in Clinical Chemistry*; Elsevier, 2020; pp 97–117. <https://doi.org/10.1016/B978-0-12-815499-1.00006-5>.
- (103) *Life Technologies - GR*. <https://www.thermofisher.com/tr/en/home.html> (accessed 2025-04-26).
- (104) *Home*. Olaf Pharmaceuticals. <https://www.olafpharmaceuticals.com/> (accessed 2025-04-26).
- (105) *Cellosaurus - Cell line encyclopedia*. <https://www.cellosaurus.org/index.html> (accessed 2025-04-26).
- (106) *Cell Products | ATCC*. <https://www.atcc.org/cell-products#t=productTab&numberOfResults=24> (accessed 2025-04-26).
- (107) *Useful Numbers for Cell Culture - GR*. <https://www.thermofisher.com/tr/en/home/references/gibco-cell-culture-basics/cell-culture-protocols/cell-culture-useful-numbers.html> (accessed 2025-04-26).
- (108) Maguire, C. M.; Rösslein, M.; Wick, P.; Prina-Mello, A. Characterisation of Particles in Solution – a Perspective on Light Scattering and Comparative Technologies. *Sci. Technol. Adv. Mater.* **2018**, *19* (1), 732–745. <https://doi.org/10.1080/14686996.2018.1517587>.
- (109) Hsiao, I.-L.; Bierkandt, F. S.; Reichardt, P.; Luch, A.; Huang, Y.-J.; Jakubowski, N.; Tentschert, J.; Haase, A. Quantification and Visualization of Cellular Uptake of TiO₂ and Ag Nanoparticles: Comparison of Different ICP-MS Techniques. *J. Nanobiotechnology* **2016**, *14* (1), 50. <https://doi.org/10.1186/s12951-016-0203-z>.
- (110) Berman, J. J. Chapter 4 - Understanding Your Data. In *Data Simplification*; Berman, J. J., Ed.; Morgan Kaufmann: Boston, 2016; pp 135–187. <https://doi.org/10.1016/B978-0-12-803781-2.00004-7>.
- (111) Lee, D. K. Data Transformation: A Focus on the Interpretation. *Korean J. Anesthesiol.* **2020**, *73* (6), 503–508. <https://doi.org/10.4097/kja.20137>.
- (112) Carvalho, P. M.; Felício, M. R.; Santos, N. C.; Gonçalves, S.; Domingues, M. M. Application of Light Scattering Techniques to Nanoparticle Characterization and Development. *Front. Chem.* **2018**, *6*, 237. <https://doi.org/10.3389/fchem.2018.00237>.
- (113) Ramaye, Y.; Dabrio, M.; Roebben, G.; Kestens, V. Development and Validation of Optical Methods for Zeta Potential Determination of Silica and Polystyrene Particles in Aqueous Suspensions. *Materials* **2021**, *14* (2), 290. <https://doi.org/10.3390/ma14020290>.
- (114) Zhang, L.; Liu, H.; Xin, Q.; Tang, L.; Tang, J.; Liu, Y.; Hu, L. A Quantitative Study of Nanoplastics within Cells Using Magnetic Resonance Imaging. *Sci. Total Environ.* **2023**, *886*, 164033. <https://doi.org/10.1016/j.scitotenv.2023.164033>.
- (115) Peng, L.; He, M.; Chen, B.; Wu, Q.; Zhang, Z.; Pang, D.; Zhu, Y.; Hu, B. Cellular Uptake, Elimination and Toxicity of CdSe/ZnS Quantum Dots in HepG2 Cells. *Biomaterials* **2013**, *34* (37), 9545–9558. <https://doi.org/10.1016/j.biomaterials.2013.08.038>.
- (116) Chou, H.-C.; Chiu, S.-J.; Hu, T.-M. Quantitative Analysis of Macrophage Uptake and Retention of Fluorescent Organosilica Nanoparticles: Implications for Nanoparticle Delivery and Therapeutics. *ACS Appl. Nano Mater.* **2024**, *7* (4), 3656–3667. <https://doi.org/10.1021/acsanm.3c05058>.

- (117) Alfonso García, S. L.; Parada-Sanchez, M. T.; Arboleda Toro, D. The Phenotype of Gingival Fibroblasts and Their Potential Use in Advanced Therapies. *Eur. J. Cell Biol.* **2020**, *99* (7), 151123. <https://doi.org/10.1016/j.ejcb.2020.151123>.
- (118) Sousa de Almeida, M.; Lee, A.; Itef, F.; Maniura-Weber, K.; Petri-Fink, A.; Rothen-Rutishauser, B. The Effect of Substrate Properties on Cellular Behavior and Nanoparticle Uptake in Human Fibroblasts and Epithelial Cells. *Nanomaterials* **2024**, *14* (4), 342. <https://doi.org/10.3390/nano14040342>.
- (119) Lee, H.; Vanhecke, D.; Balog, S.; Taladriz-Blanco, P.; Petri-Fink, A.; Rothen-Rutishauser, B. The Impact of Macrophage Phenotype and Heterogeneity on the Total Internalized Gold Nanoparticle Counts. *Nanoscale Adv.* **2024**, *6* (18), 4572–4582. <https://doi.org/10.1039/D4NA00104D>.