Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Συστημάτων Μετάδοσης Πληροφορίας
Και Τεχνολογιας Υλικων

Πρόβλεψη Διαλείψεων Μεγάλης και Μικρής Κλίμακας στις Ασύρματες Επικοινωνίες με τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αχιλλεύς Ι. Θεοχαρόπουλος

Επιβλέπων: **Αθανάσιος Δ. Παναγόπουλος**
Καθηγητής

Αθήνα, Ιούλιος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Συστημάτων Μετάδοσης Πληροφορίας
και Τεχνολογίας Υλικών

Πρόβλεψη Διαλείψεων Μεγάλης και Μικρής Κλίμακας στις Ασύρματες Επικοινωνίες με τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αχιλλεύς Ι. Θεοχαρόπουλος

Επιβλέπων: **Αθανάσιος Δ. Παναγόπουλος**
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 4η Ιουλίου 2025.

.....	.....	.....
Αθανάσιος Παναγόπουλος	Γεώργιος Φικιώρης	Γεώργιος Ματσόπουλος
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2025

.....

Αχιλλεύς Ι. Θεοχαρόπουλος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη εύρωστων συστημάτων για την πρόβλεψη των απωλειών διάδοσης μικρής και μεγάλης κλίμακας, οι οποίες υπεισέρχονται σε ζεύξη μεταξύ αεροπλοίου χαμηλής τροχιάς και επίγειου τερματικού. Υιοθετώντας μια προσέγγιση οδηγούμενη από τα δεδομένα (data driven), τα εν λόγω συστήματα υλοποιούνται από μοντέλα μηχανικής μάθησης, η σχεδίαση των οποίων πραγματοποιείται με ιδιαίτερη έμφαση στην αξιοπιστία των προβλέψεων. Η πρόβλεψη της κατάστασης του ασύρματου διαύλου με υψηλή ακρίβεια και χαμηλή διασπορά σφάλματος επιτρέπει τη δυναμική αξιοποίηση των τηλεπικοινωνιακών πόρων, με τρόπο που να διασφαλίζει αφενός τη βέλτιστη διαχείρισή τους και αφετέρου τη συνδεσιμότητα μεταξύ των δύο άκρων της ζεύξης.

Προκειμένου να καταστεί σαφής η πορεία διαχείρισης του προβλήματος, από το άκρο της διατύπωσής του και της συλλογής των δεδομένων, έως και το τελικό άκρο της διανομής του συστήματος πρόβλεψης σε λειτουργική μορφή, παρουσιάζεται κατ' αρχάς η τυπική ροή σχεδίασης για ένα σύστημα μηχανικής μάθησης. Δίνεται έμφαση στον συλλογισμό με τον οποίο λαμβάνονται οι σχεδιαστικές αποφάσεις που ανακύπτουν, καθώς και στην επαρκή θεωρητική τεκμηρίωση των μεθόδων που εφαρμόζονται. Επιπλέον, αναλύεται με λεπτομέρεια η διαδικασία λήψης μετρήσεων από το πραγματικό περιβάλλον διάδοσης, η οποία στοχεύει στη διαμόρφωση ενός συνόλου δεδομένων κατά το δυνατόν αντιπροσωπευτικού των συνθηκών που θα συναντήσει το σύστημα σε πραγματική λειτουργία.

Τα αποτελέσματα της μελέτης καταδεικνύουν ότι τα μοντέλα μηχανικής μάθησης έχουν τη δυνατότητα να επιτύχουν ιδιαίτερα υψηλές επιδόσεις στην πρόβλεψη των απωλειών διάδοσης, τόσο σε όρους μέσου σφάλματος, όσο και διασποράς. Δίνοντας έμφαση στην ανάλυση της διασποράς, διαπιστώνεται ότι οι ακραίες τιμές σφαλμάτων εμφανίζονται αρκετά περιορισμένες, στοιχείο ιδιαίτερα σημαντικό, καθότι είναι αυτές που δυνητικά μπορούν να οδηγήσουν τη ζεύξη σε διακοπή. Το γεγονός, μάλιστα, ότι οι επιδόσεις αυτές καταγράφονται επάνω σε πραγματικά δεδομένα υποδηλώνει την καταρχήν δυνατότητα ενσωμάτωσης των συστημάτων αυτών στην τηλεπικοινωνιακή υποδομή του μέλλοντος.

Λέξεις κλειδιά: Ασύρματες επικοινωνίες, Διαλείψεις μικρής και μεγάλης κλίμακας, Πρόβλεψη διαύλου, Μηχανική Μάθηση, Πρόβλεψη χρονοσειρών, DP MIMO, UAV, B5G, XGB, LSTM

Abstract

The objective of this thesis is the development of a robust system capable of predicting both large-scale and small-scale propagation losses in the context of a communication link between a low altitude UAV and a terrestrial terminal. The proposed system adopts a data-driven approach and is implemented with machine learning models, placing particular emphasis on the reliability and consistency of its predictions. Accurate and low-variance prediction of the wireless channel state enables dynamic allocation of communication resources, ensuring both optimal utilization and sustained communication connectivity across the link.

To clearly illustrate the problem-solving workflow— starting from the initial problem formulation and data acquisition up to the final deployment of the prediction system—the standard machine learning system design pipeline is initially presented. Emphasis is placed on both the reasoning behind key design choices and on the sound theoretical justification of the adopted methodologies. Additionally, the thesis provides a detailed account of the real-world propagation measurements, aiming to produce a dataset that closely reflects the operational conditions the system is expected to encounter.

The study's findings demonstrate that machine learning models can achieve highly accurate predictions of propagation losses, both in terms of average error and error variance. Particular attention is paid to the analysis of error dispersion, revealing that extreme prediction errors are notably constrained—a critical aspect, given that such extremes could potentially lead to link outages. Moreover, the fact that these results are obtained using real-world data suggests the practical viability of integrating such systems into future communication infrastructure.

Key Words: Wireless communications, Large and small scale fading, Channel prediction, Machine Learning, Time Series Forecasting, DP MIMO, UAV, B5G, XGB, LSTM

Ευχαριστίες

Θα ήθελα πρώτα από όλα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της εργασίας μου, κ. Αθανάσιο Παναγόπουλο, για τη δυνατότητα που μου έδωσε να ασχοληθώ με ένα θέμα τόσο υψηλού θεωρητικού και πρακτικού ενδιαφέροντος, αλλά και για την ουσιαστική και πολύτιμη καθοδήγηση που μου παρείχε κατά τη διάρκεια της πολύμηνης, αυτής, εκπόνησης. Ιδιαίτερες ευχαριστίες θα ήθελα, επιπλέον, να απευθύνω στο μέλος Ε.ΔΙ.Π. κ. Νεκτάριο Μωραΐτη για την άριστη επικοινωνία και τις εύστοχες παρατηρήσεις, καθώς και στα μέλη της τριμελούς επιτροπής, τους Καθηγητές κ. Γεώργιο Φικιώρη και κ. Γεώργιο Ματσόπουλο.

Τέλος, δεν θα μπορούσα να παραλείψω να ευχαριστήσω την οικογένειά μου για την αδιάλειπτη και πολυεπίπεδη υποστήριξη που μου παρείχε σε όλες τις φάσεις των σπουδών μου.

Πίνακας περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Το θεωρητικό πλαίσιο σχεδίασης ενός συστήματος πρόβλεψης μηχανικής μάθησης

1.1 Εισαγωγή

Στόχος της παρούσας εργασίας είναι η σχεδίαση εύρωστων συστημάτων για την πρόβλεψη των απωλειών διάδοσης μικρής και μεγάλης κλίμακας που υπεισέρχονται σε ασύρματη ζεύξη μεταξύ μη επανδρωμένου αεροπλοίου (UAV) και επίγειου τερματικού. Οι εν λόγω μεταβλητές - στόχοι παράγονται από ένα άγνωστο, στοχαστικό περιβάλλον Ω, το οποίο και οφείλουμε να βολιδοσκοπήσουμε προκειμένου να αποκτήσουμε την επιθυμητή προβλεπτική ικανότητα: Συγκεκριμένα, για την εκάστοτε μεταβλητή απώλειας, έστω Y, θεωρούμε ένα διάνυσμα μεταβλητών X που προέρχεται από το ίδιο περιβάλλον, είναι διαθέσιμο προς μέτρηση και εικάζουμε πως περιέχει πληροφορία για τη στατιστική συμπεριφορά της Y. Κατόπιν, εγκαθιστούμε ένα σύστημα λήψης μετρήσεων - δειγματοληψίας της από κοινού κατανομής $D(X, Y)$ και αναπτύσσουμε τεχνικές μηχανικής μάθησης (ΜΜ) προκειμένου να μοντελοποιήσουμε τη μεταξύ τους σχέση εξάρτησης. Προτού όμως αναλύσουμε με λεπτομέρεια το περιβάλλον διάδοσης, την πειραματική διάταξη και τη συμπεριφορά των προβλεπτών ΜΜ επάνω στα πραγματικά δεδομένα, κρίνουμε σκόπιμο να παρουσιάσουμε το τυπικό μαθηματικό πλαίσιο του προβλήματος, καθώς και τη θεωρητική ανάλυση των μεθόδων ΜΜ που θα αξιοποιήσουμε. Επισημαίνεται ότι, εφόσον οι απώλειες διάδοσης είναι εκ φύσεως συνεχείς μεταβλητές, οι διάφορες θεωρητικές τεχνικές θα αναπτυχθούν με αναφορά σε *προβλήματα παλινδρόμησης*, πρόβλεψης, δηλαδή, συνεχών μεταβλητών από τα δεδομένα.

1.2 Τυπική διατύπωση του προβλήματος της παλινδρόμησης

Υποθέτουμε μια συνεχή μεταβλητή - στόχο y, την οποία μας ενδιαφέρει να προβλέψουμε, και ένα διάνυσμα μεταβλητών x, το οποίο θεωρούμε πως περιέχει την απαιτούμενη πληροφορία ώστε η y να προβλεφθεί ντετερμινιστικά και με μηδενικό σφάλμα, μέσω μιας συνάρτησης $y = f(x)$. Παρότι κάτι τέτοιο θα στοιχειοθετούσε ένα ιδανικό σχεδιαστικά σενάριο, στην πλειονότητα των πρακτικών περιπτώσεων ισχύουν οι εξής περιορισμοί: Αφενός, είναι πιθανό να μας είναι γνωστό και διαθέσιμο προς μέτρηση μόνο ένα υποδιάνυσμα $\bar{x}$ του x, το οποίο μπορεί να αξιοποιηθεί για μια εκτίμηση της y, η οποία εν γένει θα εμπεριέχει σφάλμα (*σφάλμα μερικής παρατηρησιμότητας*). Αφετέρου, με δεδομένο πως η f δεν μας είναι εκ των προτέρων γνωστή, η υπόθεση πως μπορεί να αναπαρασταθεί ως στοιχείο ενός επιλεγμένου χώρου συναρτήσεων H (*χώρος υποθέσεων*)

δεν είναι κατ' ανάγκη απόλυτα έγκυρη, με αποτέλεσμα να εισάγει ένα σφάλμα (*σφάλμα μοντελοποίησης*). Διατυπώνοντας μαθηματικά τις παραπάνω ιδέες, προκύπτουν οι ακόλουθες σχέσεις, (1.1) και (1.2):

$$Y = \hat{Y} + E \tag{1.1}$$

όπου E είναι η τυχαία μεταβλητή σφάλματος μεταξύ της εκτιμήτριας $\hat{Y}$ και της πραγματικής μεταβλητής Y. Το E αποτυπώνει τη συνδυαστική επίδραση των σφαλμάτων μερικής παρατηρησιμότητας και μοντελοποίησης και στη βιβλιογραφία συχνά αναφέρεται ως *σφάλμα επεξήγησης* (explanation error).

$$\hat{Y} = \varphi(X) \tag{1.2}$$

όπου $\varphi(\cdot)$, μια ντετερμινιστική συνάρτηση εκτίμησης, η οποία ανήκει στον χώρο υποθέσεων H, είναι στην ευχέρειά μας να τη σχεδιάσουμε. Τόσο, η επιλογή του H αυτού καθεαυτό, όσο και η σχεδίαση της διαδικασίας επιλογής ενός στοιχείου $\varphi \in H$ για την εκτίμηση της f (*αλγόριθμος μάθησης*) είναι βαρύνουσας σημασίας για την τελική επίδοση του παλινδρομητή, όπως θα εξηγηθεί και παρακάτω.

Θεωρώντας επιπλέον μια συνάρτηση απωλειών $L(Y, \varphi(\mathrm{X}))$ και λαμβάνοντας τη μέση της τιμή, μπορούμε να διατυπώσουμε το πρόβλημα παλινδρόμησης ως το εξής πρόβλημα ελαχιστοποίησης:

$$\varphi^{*} = argmin_{\varphi \in \mathrm{H}} \mathbb{E}_{\sim D(X,Y)}(L(Y, \varphi(\mathrm{X}))) \tag{1.3}$$

Έχοντας πλέον ορίσει το πρόβλημα σε αφηρημένη μορφή, καλούμαστε να πραγματοποιήσουμε τις απαραίτητες σχεδιαστικές επιλογές, ούτως ώστε αφενός να το καταστήσουμε υπολογιστικά επιλύσιμο και αφετέρου να διασφαλίσουμε ότι η επίλυσή του αυτή θα παράγει έναν προβλέπτη υψηλής ακρίβειας.

1. *Επιλογή του τυχαίου διανύσματος X*: Η βελτιστότητα της φ^{*} δεν μας εγγυάται ότι οι εκτιμήσεις που θα παρέχει θα είναι αποδεκτής ακρίβειας. Αν επιλέξουμε ένα διάνυσμα X, το οποίο εγγενώς δεν περιέχει πληροφορία για την Y, τότε αυτό δεν δύναται να αντισταθμιστεί με κανένα μαθηματικό τέχνασμα. Προτού, επομένως, εκκινήσουμε τη διαδικασία προσαρμογής συναρτήσεων στα δεδομένα, θα πρέπει κατά το δυνατόν να διασφαλίσουμε ότι το διάνυσμα X περιέχει πράγματι πληροφορία για την Y και μάλιστα ότι είναι το ελάχιστης διάστασης διάνυσμα που περιέχει μια δεδομένη "ποσότητα πληροφορίας" για την Y. Η ελαχιστοποίηση της διάστασης είναι επιθυμητή, τόσο για την κατανόηση των θεωρητικών ορίων του προβλήματος, όσο και για τη βελτιστοποίηση της επίδοσης του συστήματος, όπως θα εξηγηθεί στη συνέχεια του κεφαλαίου.

2. *Επιλογή του χώρου υποθέσεων H*: Ο H θα πρέπει να επιλεγεί ώστε η *χωρητικότητά του,* οι δυνατές συναρτήσεις, δηλαδή, που μπορεί να αναπαραστήσει, να συμβαδίζει με την "πολυπλοκότητα" της άγνωστης συνάρτησης f. Με δεδομένο ότι η πολυπλοκότητα της f δεν μας είναι εκ των προτέρων γνωστή, αλλά ούτε και ορίζεται με κάποιον αυστηρό τρόπο, μπορούμε σχεδιαστικά να πειραματιστούμε με διαφορετικούς χώρους υποθέσεων. Οι διάφορες αρχιτεκτονικές *μοντέλων μηχανικής μάθησης* που θα διερευνηθούν στην εργασία πράγματι επάγουν διαφορετικούς χώρους υποθέσεων.
3. *Επιλογή της μορφής στην οποία θα αναχθεί το πρόβλημα ελαχιστοποίησης:* Η επίλυση του προβλήματος (1.3) στην αρχική του μορφή απαιτεί γνώση της κατανομής $D(X, Y)$, κάτι το οποίο εν γένει δεν είναι εφικτό. Για τον λόγο αυτό, είναι συχνά απαραίτητος ο μετασχηματισμός του σε κάποιο εναλλακτικό, υπολογίσιμο πρόβλημα βελτιστοποίησης, η λύση του οποίου θα προσεγγίζει δυνητικά τη φ^*.
4. *Επιλογή του αλγορίθμου μάθησης*: Ο αλγόριθμος για την επίλυση του προβλήματος βελτιστοποίησης που διατυπώνεται στη σχέση (1.3), ή κάποιας ανηγμένης μορφής του, είναι, φυσικά, καθοριστικής σημασίας. Ακόμα και αν υπάρχει μια συνάρτηση $\varphi^* \in H$, η οποία να επιτυγχάνει ένα αποδεκτό σφάλμα προσέγγισης, αυτό μας είναι ανώφελο εάν δεν διαθέτουμε έναν συστηματικό τρόπο ώστε να την προσδιορίσουμε, έστω και προσεγγιστικά, ανάμεσα στα υπόλοιπα στοιχεία του χώρου.

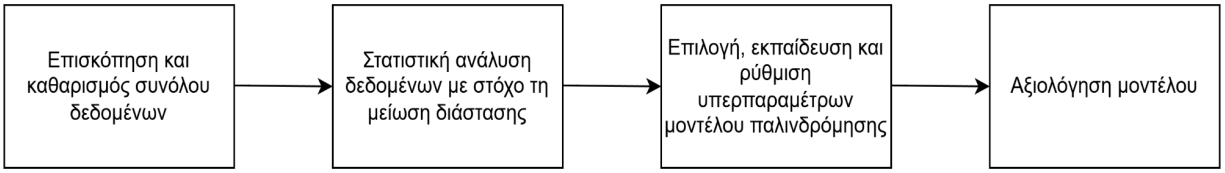
Όλες οι παραπάνω σχεδιαστικές αποφάσεις ενσωματώνονται στην τυπική ροή εργασιών που χαρακτηρίζει ένα πρόβλημα παλινδρόμησης και κατ' επέκταση *επιβλεπόμενης μηχανικής μάθησης*, μάθησης, δηλαδή, μιας συνάρτησης από τα δεδομένα της. Προτού, όμως, εμβαθύνουμε στην εν λόγω ροή εργασιών και στις μεθόδους που τη συνιστούν, θα διακρίνουμε τα προβλήματα παλινδρόμησης σε *στατικά* και *δυναμικά*. Η διάκριση αυτή είναι αναγκαία, εφόσον οι τεχνικές και τα μοντέλα που χρησιμοποιούνται στις δύο περιπτώσεις διαφέρουν αρκετά μεταξύ τους.

1.3 Στατικά και δυναμικά προβλήματα μάθησης

Ένα πρόβλημα παλινδρόμησης θα ονομάζεται στατικό, εφόσον η πληροφορία που φέρει το τυχαίο διάνυσμα εισόδου $X = (X_1, X_2, ..., X_d)$ για την έξοδο δεν έχει ακολουθιακή δομή. Με άλλα λόγια, οι μεταβλητές - στοιχεία του X είναι εννοιολογικά (αν και όχι κατ' ανάγκη στατιστικά) ανεξάρτητες και άρα δεν έχει νόημα να εξεταστούν σε διαδοχή από ένα μοντέλο, παρά μόνο συνδυαστικά. Αντιθέτως, σε ένα δυναμικό πρόβλημα παλινδρόμησης, το τυχαίο διάνυσμα εισόδου αποτελεί μια χρονοσειρά: Τα δείγματα εμφανίζονται σε χρονική διαδοχή, είναι συσχετισμένα μεταξύ τους και η πληροφορία εμπεριέχεται στα διάφορα δυναμικά μοτίβα που παρουσιάζει η ακολουθία, παρά στις μεμονωμένες τιμές των δειγμάτων. Απαιτούνται, λοιπόν, μοντέλα που να μπορούν να επεξεργαστούν το διάνυσμα εισόδου με τον τρόπο που αυτό παράγεται, δηλαδή ακολουθιακά.

1.4 Ροή εργασιών σε ένα πρόβλημα παλινδρόμησης και θεωρητική ανάλυση μεθόδων

Παρακάτω, παρουσιάζεται το λειτουργικό διάγραμμα της ροής εργασιών σε ένα γενικό πρόβλημα παλινδρόμησης. Επισημαίνεται ότι, παρότι οι μέθοδοι που συνιστούν κάθε δομική μονάδα του διαγράμματος είναι διαφορετικές για στατικά και δυναμικά προβλήματα, η ροή εργασιών καθαυτή είναι πρακτικά όμοια. Επιπλέον, η θεωρητική ανάλυση των μεθόδων δεν είναι εξαντλητική, αλλά περιορίζεται σε όσες θα χρησιμοποιηθούν στην παρούσα εργασία.



***Εικόνα 1.1**: Ροή εργασιών σε πρόβλημα παλινδρόμησης.*

1.4.1 Επισκόπηση και καθαρισμός συνόλου δεδομένων

Το πρώτο στάδιο της εν λόγω αλυσίδας διαδικασιών εστιάζει στην ποιότητα της δειγματοληψίας των δεδομένων. Διερευνά το κατά πόσο υπάρχουν ελλιπείς, ακραίες (outliers) ή εσφαλμένες τιμές δειγμάτων, με σκοπό τη διόρθωση ή αφαίρεσή τους. Στα σύνολα δεδομένων που θα χρησιμοποιηθούν στην παρούσα εργασία, τα δείγματα είναι πλήρη και στη σωστή μορφή για περαιτέρω επεξεργασία, οπότε και δεν απαιτείται η χρήση των παραπάνω μεθόδων.

1.4.2 Στατιστική ανάλυση δεδομένων με στόχο τη μείωση διάστασης

- **Στατικά προβλήματα**

Σε στατικά προβλήματα, η μείωση της διάστασης του χώρου εισόδου μπορεί να επιτευχθεί μέσω διαδικασιών *επιλογής* και *μετασχηματισμού χαρακτηριστικών*.

Η επιλογή χαρακτηριστικών στοχεύει στη διατήρηση των στοιχείων του τυχαίου διανύσματος $X = (X_1, X_2, ..., X_d)$ που παρέχουν ισχυρή και κατά το δυνατόν ανεξάρτητη πληροφορία μεταξύ τους για την έξοδο Y. Εν γένει, το κατά πόσο μια τυχαία μεταβλητή X "περιέχει πληροφορία" ή "συσχετίζεται" με μια τυχαία μεταβλητή Y μπορεί να μελετηθεί σε δύο πλαίσια: Σε αυτό της θεωρίας πληροφορίας (I) και σε αυτό της θεώρησης των χώρων τυχαίων μεταβλητών ως χώρων εσωτερικού γινομένου (II).

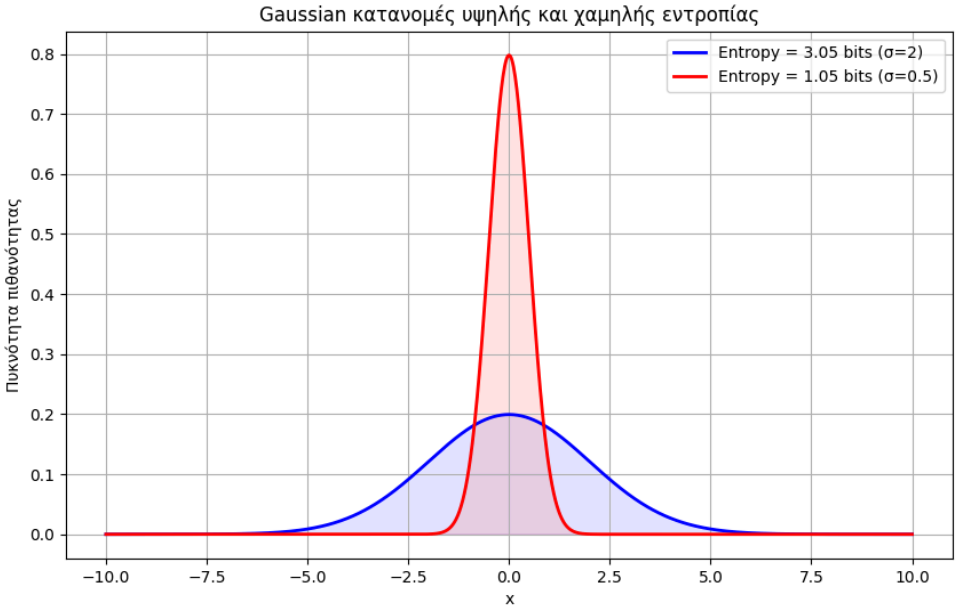
(I) Εν προκειμένω, το μέτρο συσχέτισης που θα χρησιμοποιήσουμε θα είναι η *αμοιβαία πληροφορία* μεταξύ δύο τυχαίων μεταβλητών. Ακολουθούν τρεις βασικοί ορισμοί:

Ορισμός 1.1: Η *εντροπία* μιας συνεχούς τυχαίας μεταβλητής ορίζεται μέσω της ακόλουθης σχέσης:

$$H(X) = \int_{S_X} - \ ln(x) p_X(x) dx \qquad (1.4)$$

όπου S_X είναι το σύνολο τιμών της τυχαίας μεταβλητής X και $p_X(x)$ είναι η συνάρτηση πυκνότητας πιθανότητάς της.

Διαισθητικά, η εντροπία αποτελεί ένα μέτρο της "αβεβαιότητας" που χαρακτηρίζει την κατανομή της X. Η ιδέα αυτή αποτυπώνεται στο ακόλουθο γράφημα, όπου απεικονίζονται από κοινού δύο μηδενικού μέσου Gaussian κατανομές, υψηλής και χαμηλής εντροπίας αντίστοιχα:



***Εικόνα 1.2**: Γραφική απεικόνιση Gaussian κατανομών υψηλής και χαμηλής εντροπίας.*

Παρατηρούμε ότι η κατανομή χαμηλής εντροπίας είναι αρκετά πιο συγκεντρωμένη, εν προκειμένω γύρω από τη μέση της τιμή. Επομένως, αν θεωρήσουμε πως υπάρχει μια άγνωστη μεταβλητή την οποία θέλουμε να εκτιμήσουμε και ότι οι παραπάνω κατανομές αποτυπώνουν τις πεποιθήσεις μας αναφορικά με τη μεταβλητή αυτή, στην περίπτωση της χαμηλής εντροπίας αντιλαμβανόμαστε πως η εκτίμησή μας μπορεί να πραγματοποιηθεί με μεγαλύτερη "βεβαιότητα". Αναμένουμε, άρα, η χαμηλότερη εντροπία να καθιστά δυνατές πιο "βέβαιες" αποφάσεις, χαμηλότερου, δηλαδή, σφάλματος εκτίμησης.

Ορισμός 1.2: Η *υπό συνθήκη εντροπία* μιας τυχαίας μεταβλητής Y με δεδομένη την τυχαία μεταβλητή X ορίζεται ως ακόλουθα:

$$H(Y|X) = \mathbb{E}_{\sim D(X)}(H(Y|X = x) \qquad (1.5)$$

Ουσιαστικά, αποτελεί τη μέση εντροπία της κατανομής της Y, με δεδομένο ότι γνωρίζουμε τη X. Σε προβλήματα απόφασης, όπου τυπικά καλούμαστε να εκτιμήσουμε την τιμή μιας μεταβλητής εξόδου Y υπό τη συνθήκη κάποιων γνωστών μεταβλητών (τις οποίες αποκαλούμε *δεδομένα εισόδου*), η υπό συνθήκη εντροπία αποτελεί ένα πολύ φυσικό μέτρο αποτίμησης της τρέχουσας αβεβαιότητάς μας σχετικά με την Y.

Ορισμός 1.3: Η *αμοιβαία πληροφορία* μεταξύ δύο τυχαίων μεταβλητών X και Y εκφράζει το κέρδος εντροπίας στην κατανομή της Y όταν μας είναι γνωστή η X, και ορίζεται μαθηματικά μέσω της ακόλουθης σχέσης:

$$I(X,Y) = H(Y) - H(Y|X) \qquad (1.6)$$

Από τον παραπάνω ορισμό, είναι προφανές ότι η $I(X,Y)$ είναι ένα θετικό και αύξον μέτρο συσχέτισης: Όσο μεγαλύτερη είναι η τιμή του, τόσο περισσότερο η γνώση της X μειώνει την εντροπία και άρα την αβεβαιότητά μας για την Y, υποδηλώνοντας έτσι την ύπαρξη μιας σχέσης εξάρτησης μεταξύ των δύο. Σε μια τέτοια περίπτωση, έχει νόημα να θέσουμε τη X ως ένα εκ των ορισμάτων της συνάρτησης εκτίμησης $\varphi(\cdot)$ (βλ. Σχέση (1.2)) που επιδιώκουμε να κατασκευάσουμε.

(II) Κατεξοχήν μέτρα γραμμικής συσχέτισης σε διανυσματικούς χώρους (όπως είναι ο χώρος των τυχαίων μεταβλητών, αν εφοδιαστεί με την πράξη της πρόσθεσης και του βαθμωτού πολλαπλασιασμού) αποτελούν οι *πράξεις εσωτερικού γινομένου*, οι οποίες ορίζονται ως ακολούθως:

Ορισμός 1.4: Μια διμερής πράξη $<\cdot,\cdot>$ επάνω σε έναν διανυσματικό χώρο V αποτελεί *εσωτερικό γινόμενο,* εφόσον ικανοποιεί τις παρακάτω ιδιότητες:

(α) Θετικότητα : $< u, u > \geq 0$ και $< u, u > = 0$ αν και μόνο εάν $u = 0$
(β) Συμμετρία : $< v, u > = < u, v >$
(γ) Γραμμικότητα ως προς το πρώτο όρισμα:
$< av + bw, u > = a < v, u > + b < w, u >$, όπου $a, b \in \mathbb{R}$

Συνέπεια του παραπάνω ορισμού αποτελεί η παρακάτω θεμελιώδης ανισότητα:

Λήμμα 1.1 (Ανισότητα Cauchy - Schwarz): Αν u, v είναι στοιχεία ενός χώρου εσωτερικού γινομένου, τότε ισχύει η ακόλουθη ανισότητα:

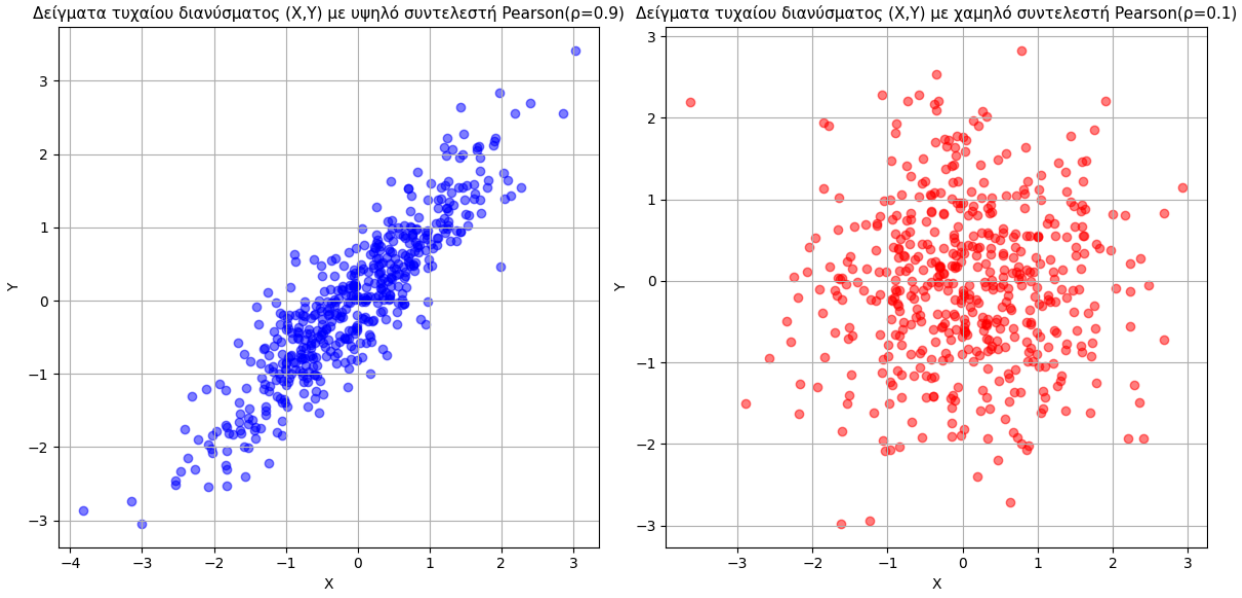
$$| < u, v > | \leq \sqrt{| < u, u > |} \cdot \sqrt{| < v, v > |}$$

Επιπλέον στα παραπάνω, θεωρούμε την έννοια του *κανονικοποιημένου εσωτερικού γινομένου*, το οποίο ορίζεται μέσω της εξής σχέσης: $\rho(u, v) = \frac{<u,v>}{\sqrt{|<u,u>|}\cdot\sqrt{|<v,v>|}}$ (1.7)

Από την (1.7), προκύπτει άμεσα ότι:

- $-1 \leq \rho(u, v) \leq 1$
- Το $\rho(u, v)$ μεγιστοποιείται κατ' απόλυτη τιμή, όταν u και v είναι συγγραμμικά

Εξειδικεύοντας, τώρα, σε διανυσματικούς χώρους τυχαίων μεταβλητών, είναι εύκολο να αποδειχθεί ότι η *ετεροσυσχέτιση* $r_{xy} = \mathbb{E}[XY]$ ορίζει μια πράξη εσωτερικού γινομένου. Μετασχηματίζοντας τις X, Y ως $\overline{X} = X - m_X$ και $\overline{Y} = Y - m_Y$ και θεωρώντας το κανονικοποιημένο εσωτερικό γινόμενο των $\overline{X}, \overline{Y}$, έχουμε ότι $\rho(\overline{X}, \overline{Y}) = \frac{\mathbb{E}[\overline{XY}]}{\sqrt{\mathbb{E}[\overline{XX}]}\cdot\sqrt{\mathbb{E}[YY]}}$. Το μέγεθος αυτό ορίζεται ως ο *συντελεστής συσχέτισης Pearson* των τυχαίων μεταβλητών X, Y. Όσο υψηλότερη είναι η τιμή του συντελεστή αυτού, τόσο οι δύο μεταβλητές τείνουν στατιστικά να συνδιακυμαίνονται. Η ιδέα αυτή αποτυπώνεται στα δύο παρακάτω γραφήματα, όπου απεικονίζονται οι κατανομές δειγμάτων δύο Gaussian τυχαίων μεταβλητών X, Y με υψηλό και χαμηλό συντελεστή Pearson αντίστοιχα.



***Εικόνα 1.3**: Γραφική απεικόνιση της από κοινού κατανομής τυχαίων μεταβλητών με υψηλό και χαμηλό συντελεστή Pearson.*

Εξετάζοντας προβλήματα παλινδρόμησης, τα βασικά πλεονεκτήματα του συντελεστή Pearson συνίστανται αφενός στο χαμηλό κόστος υπολογισμού του, το οποίο μας επιτρέπει, εκτός από τη συσχέτιση μεταξύ χαρακτηριστικών και στόχου, να διερευνήσουμε συσχετίσεις και μεταξύ των χαρακτηριστικών, και αφετέρου στο γεγονός ότι αποτυπώνει

καλά σχέσεις αναλογίας, οι οποίες εμφανίζονται αρκετά συχνά μεταξύ μεταβλητών. Βέβαια, σε πρακτικά προβλήματα ενδέχεται να παρουσιάζονται και αρκετά πολυπλοκότερες, μη γραμμικές εξαρτήσεις, τις οποίες ο συντελεστής Pearson αποτυγχάνει να εντοπίσει. Για τον λόγο αυτό, προκειμένου να αποκτήσουμε πληρέστερη εικόνα και να πραγματοποιήσουμε πιο εύστοχη επιλογή χαρακτηριστικών, θα αξιολογήσουμε τις συσχετίσεις συνεκτιμώντας τον συντελεστή Pearson και την αμοιβαία πληροφορία μεταξύ των μεταβλητών.

Η επιλογή χαρακτηριστικών μπορεί να θεωρηθεί ως μια ειδική περίπτωση *μετασχηματισμού χαρακτηριστικών*. Ο μετασχηματισμός χαρακτηριστικών εδράζεται στην *υπόθεση πολλαπλότητας* (manifold hypothesis) [1], σύμφωνα με την οποία η κατανομή ενός τυχαίου διανύσματος X που συναντάται σε μια πραγματική εφαρμογή είναι πιθανό να εντοπίζεται σε γεωμετρικό τόπο (πολλαπλότητα) $\mathcal{M}$ χαμηλότερης διάστασης (m) από αυτήν του αρχικού διανύσματος (d). Ως εκ τούτου, μπορούμε δυνητικά να κατασκευάσουμε μια απεικόνιση F: $\mathbb{R}^d \to \mathbb{R}^m$, γραμμική ή και μη γραμμική, ανάλογα με το είδος της πολλαπλότητας που υποθέτουμε, η οποία να απεικονίζει με "1-1" τρόπο τη $\mathcal{M}$ στον χώρο $\mathbb{R}^m$. Στην παρούσα εργασία, θα περιοριστούμε σε *γραμμικές πολλαπλότητες* και συγκεκριμένα σε διανυσματικούς υποχώρους του $\mathbb{R}^d$.

Προκειμένου να θέσουμε το νοητικό πλαίσιο για την κατασκευή της F, θα θεωρήσουμε κατ' αρχάς τους ακόλουθους ορισμούς και προτάσεις:

Ορισμός 1.5: Έστω $y, \hat{y} \in \mathbb{R}^d$. Το διάνυσμα $e = y - \hat{y}$ καλείται *διάνυσμα σφάλματος* μεταξύ των y και $\hat{y}$. Η ποσότητα $< e, e > \triangleq ||e||^2$ ονομάζεται *ενέργεια* του σφάλματος.

Ορισμός 1.6: Έστω ο διανυσματικός χώρος εσωτερικού γινομένου $\mathbb{R}^d$ και U ένας υπόχωρός του. Ο γραμμικός μετασχηματισμός P: $\mathbb{R}^d \to U$ ονομάζεται *μετασχηματισμός ορθής προβολής,* εάν το διάνυσμα σφάλματος $e = y - P_U(y)$ είναι κάθετο στον χώρο U και κατ' επέκταση στο $P_U(y)$. Το διάνυσμα $P_U(y)$ ονομάζεται *ορθή προβολή* του y στον U, είναι μοναδικό και ισχύει $< y - P_U(y), u >= 0,\ \forall u \in U$.

Πρόταση 1.1: Έστω $y \in \mathbb{R}^d$ και U ένας υπόχωρος του $\mathbb{R}^d$. Το διάνυσμα $P_U(y)$ αποτελεί το στοιχείο του U που *προσεγγίζει βέλτιστα* το y, υπό την έννοια ότι ελαχιστοποιεί την ενέργεια του σφάλματος.

Έστω τώρα ένα τυχαίο διάνυσμα $\mathcal{X}$, το οποίο αναπαρίσταται από έναν πίνακα X. Πραγματοποιώντας επιλογή χαρακτηριστικών, διατηρούμε k από τα d αρχικά χαρακτηριστικά, τα οποία αντιστοιχούν με τη σειρά τους σε k διανύσματα βάσης της μορφής $u_i = [0,..,1,0,...0]$. Το μη μηδενικό στοιχείο του εκάστοτε u_i αντιστοιχεί στον δείκτη (index) του αντίστοιχου χαρακτηριστικού στο $\mathcal{X}$. Κατ' αυτόν τον τρόπο, ουσιαστικά προβάλλουμε το σύνολο των N διανυσμάτων στον υπόχωρο που παράγεται από τα u_i, έστω

$\tilde{U}$, επάγοντας ένα μέσο σφάλμα προσέγγισης $E = \frac{1}{N}\sum_{i=1}^{N} ||x_i - P_{\tilde{U}}(x_i)||^2$. Εφόσον πρόκειται για μετασχηματισμό προβολής, προκύπτει άμεσα από την Πρόταση 1.1 ότι είναι βέλτιστος ανάμεσα σε όλους τους μετασχηματισμούς που απεικονίζουν τα N διανύσματα στον υπόχωρο $\tilde{U}$. Για να είναι όμως αυτός ο μετασχηματισμός βέλτιστος εν γένει, θα πρέπει ο $\tilde{U}$ να επιλεγεί έτσι ώστε ανάμεσα σε όλους τους (άπειρους το πλήθος) υποχώρους διάστασης k , να είναι αυτός για τον οποίο ελαχιστοποιείται το E. Η διαδικασία επιλογής χαρακτηριστικών εξετάζει μόλις d ανά k εκ των υποχώρων αυτών και άρα είναι εξαιρετικά πιθανό να οδηγεί σε υποβέλτιστη προσέγγιση. Ο ζητούμενος βέλτιστος μετασχηματισμός προκύπτει από την εσωτερική δομή του πίνακα X, όπως καταδεικνύουν τα ακόλουθα θεωρήματα, τα οποία και παρατίθενται χωρίς απόδειξη.

Θεώρημα 1.1 (Ανάλυση Ιδιαζουσών Τιμών - SVD): Αν X είναι ένας πίνακας διαστάσεων $n \times d$ και βαθμού r, τότε μπορεί να γραφεί ως $X = U\Sigma V^T$, όπου $U \in \mathbb{R}^{d\times r}$ και ορθογώνιος, $V \in \mathbb{R}^{n\times r}$ και ορθογώνιος, ενώ $\Sigma \in \mathbb{R}^{r\times r}$ και διαγώνιος. Τα στοιχεία του Σ ονομάζονται *ιδιάζουσες τιμές* του X.

Θεώρημα 1.2: Έστω $X \in \mathbb{R}^{n\times d}$ και $U\Sigma V^T$ η ανάλυσή του σε ιδιάζουσες τιμές. Οι πίνακες V, U θα αποτελούνται από τα ιδιοδιανύσματα των X^TX και XX^T αντίστοιχα, τα οποία αντιστοιχούν σε r μη μηδενικές ιδιοτιμές $\lambda_1,\ldots,\lambda_r$. Για τον πίνακα Σ θα ισχύει $\Sigma = diag(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_r})$.

Θεώρημα 1.3 (Ανάλυση σε κυρίαρχες συνιστώσες - PCA): Έστω σύνολο N διανυσμάτων $x_1,\ldots,x_N$, τα οποία καταχωρίζονται ως γραμμές σε πίνακα X. Ο διανυσματικός υπόχωρος $\mathcal{U} \subseteq \mathbb{R}^d$, διάστασης k, για τον οποίο το σφάλμα προσέγγισης των N διανυσμάτων, $E = \frac{1}{N}\sum_{i=1}^{N} ||x_i - P_{\mathcal{U}}(x_i)||^2$, λαμβάνει ελάχιστη τιμή, έχει ως βάση τα ιδιοδιανύσματα $v_1,\ldots,v_k$ του πίνακα X^TX που αντιστοιχούν στις k μέγιστες ιδιοτιμές. Τα διανύσματα αυτά ονομάζονται *διευθύνσεις των k κυρίαρχων συνιστωσών* του πίνακα X και υπολογίζονται απευθείας από την SVD ανάλυσή του. Το μετασχηματισμένο σύνολο διανυσμάτων προκύπτει ως $\overline{X} = XU$.

Η μείωση της διάστασης σε ένα πρόβλημα επιβλεπόμενης μάθησης είναι σαφώς επιθυμητή, υπό την προϋπόθεση βέβαια ότι η απώλεια πληροφορίας που συνεπάγεται δεν έχει αισθητά αρνητική επίδραση στην επίδοση των μοντέλων. Ελαχιστοποιώντας το πλήθος των μεταβλητών εισόδου στο απολύτως αναγκαίο, εξασθενεί η πολυπλοκότητα του προβλήματος, γεγονός που ευνοεί τόσο τη βαθύτερη κατανόησή του από τον σχεδιαστή, όσο και την ανάπτυξη εύρωστων μοντέλων, ιδίως σε προβλήματα με λιγοστά δεδομένα εκπαίδευσης. Αν, δε, υποθέσουμε ότι το μοντέλο πρόκειται να χρησιμοποιηθεί για

πρόβλεψη σε πραγματικές συνθήκες, η ελαχιστοποίηση της διάστασης μέσω επιλογής χαρακτηριστικών συνεπάγεται ανάγκη μέτρησης λιγότερων φυσικών ποσοτήτων, το οποίο σίγουρα είναι προτιμητέο από πλευράς πολυπλοκότητας υλοποίησης. Σημειώνεται, βέβαια, πως όταν ο όγκος των δεδομένων N είναι αρκετά "μεγάλος" (σε σχέση με τη διάσταση εισόδου), η κατανομή των διανυσμάτων εκπαίδευσης στον αρχικό υψηλοδιάστατο χώρο παύει εν γένει να είναι αραιή (όπως συμβαίνει όταν το N είναι μικρό), κι έτσι ένα μοντέλο μπορεί εξαρχής να ανιχνεύσει πυκνές και πλούσιες σε πληροφορία δομές στα δεδομένα. Στην περίπτωση, λοιπόν, αυτή, θα πρέπει να ληφθεί υπ' όψιν ότι η μείωση της διάστασης έρχεται με συνέπεια την απώλεια πληροφορίας και κατ' επέκταση την πιθανή υποβάθμιση της επίδοσης των υποψήφιων μοντέλων. Προφανώς, όσο λιγότερη ή περιττή πληροφορία περιελάμβαναν τα χαρακτηριστικά που αφαιρέθηκαν, τόσο μικρότερη θα είναι και η επίδραση της αφαίρεσής τους.

- **Δυναμικά προβλήματα**

Στην περίπτωση όπου το διάνυσμα εισόδου $X = (X_1, X_2, ..., X_d)$ αποτελεί μια χρονοσειρά, το σύνολο των επιμέρους μεταβλητών X_i αποτυπώνει την εξέλιξη ενός φυσικού μεγέθους στον χρόνο. Ως εκ τούτου, η πληροφορία του έγκειται στα διάφορα μικροδυναμικά και μακροδυναμικά μοτίβα που περιλαμβάνει, τα οποία αναπαρίστανται μαθηματικά ως υποδιανύσματα του X και κατά κανόνα περιέχουν στοιχεία ισχυρά συσχετισμένα μεταξύ τους. Η ύπαρξη συσχετίσεων και εν γένει στατιστικών εξαρτήσεων μεταξύ των μεταβλητών εισόδου υποδηλώνει περίσσεια πληροφορίας (information redundancy), την οποία μπορούμε δυνητικά να άρουμε μέσω μιας διαδικασίας *μηχανικής χαρακτηριστικών* (feature engineering).

Η μηχανική χαρακτηριστικών επικεντρώνεται στην κατασκευή ενός μετασχηματισμού $T(X)$, ο οποίος απεικονίζει το X σε ένα διάνυσμα X' χαμηλότερης διάστασης, με αποσυσχετισμένα χαρακτηριστικά. Ωστόσο, ο τρόπος κατασκευής του μετασχηματισμού T δεν είναι καθόλου προφανής: Θα πρέπει να λαμβάνει υπόψη την ακολουθιακή δομή των δεδομένων (*sequence awareness* - προϋπόθεση την οποία δεν πληροί ο μετασχηματισμός PCA), την πιθανή *μη στασιμότητα* της χρονοσειράς (μεταβολή των στατιστικών της χαρακτηριστικών με τον χρόνο), ενώ συχνά απαιτεί εξειδικευμένη γνώση του πεδίου όπου εντάσσεται το πρόβλημα (εν προκειμένω, της ασύρματης διάδοσης ισχύος).

Οι παραπάνω σχεδιαστικές δυσκολίες, αλλά και το γεγονός πως δεν υφίσταται ένας καλά ορισμένος τρόπος ώστε να αξιολογήσουμε τη βελτιστότητα ενός τέτοιου μετασχηματισμού, έχουν αναδείξει αρχιτεκτονικές μοντέλων προοριζόμενων να λαμβάνουν απευθείας τη χρονοσειρά ως είσοδο. Στην παρούσα εργασία, η έμφαση θα δοθεί σε αυτά τα μοντέλα, οπότε η τεχνική μείωσης διάστασης που θα εφαρμόσουμε θα είναι στοιχειώδης: Συγκεκριμένα, έστω ότι μας ενδιαφέρει να πραγματοποιήσουμε προβλέψεις σε σχέση με το $X(n)$ ή και μελλοντικών του δειγμάτων. Μέσω της *συνάρτησης αυτοσυσχέτισης* (ACF) $\gamma_X(h)$ της χρονοσειράς, η οποία ορίζεται ως ο συντελεστής συσχέτισης Pearson μεταξύ των τυχαίων μεταβλητών $X(n)$ και $X(n-h)$, διερευνούμε ποιο είναι το μήκος h_c (context

length) του χρονικού παραθύρου για το οποίο τα δείγματα $X(n)$, $X(n-h)$ αποσυσχετίζονται. Έτσι, για $h > h_c$ θεωρούμε ότι τα παρελθόντα δείγματα παύουν να περιέχουν πληροφορία για το $X(n)$ και τα μελλοντικά του, οπότε το διάνυσμα εισόδου διαμορφώνεται ως $X = (X_{n-h_c}, X_{n-h_c+1}, ..., X_{n-1})$ αντί του αρχικού, υψηλότερης διάστασης, $X = (X_0, X_1, ..., X_{n-1})$.

1.4.3 Επιλογή, εκπαίδευση και ρύθμιση υπερπαραμέτρων μοντέλων

Εφόσον έχουμε προεπεξεργαστεί το αρχικό σύνολο δεδομένων και έχουμε αποφασίσει τις συνιστώσες του τυχαίου διανύσματος $X \in \mathbb{R}^d$, το επόμενο σχεδιαστικό βήμα που καλούμαστε να υλοποιήσουμε είναι η *επιλογή αρχιτεκτονικής μοντέλου.* Μια αρχιτεκτονική μοντέλου αναφέρεται κατ' ουσίαν σε ένα πρότυπο υπολογιστικής δομής, το οποίο είναι κατάλληλα παραμετροποιημένο ώστε να παράγει ένα σύνολο συναρτήσεων της μορφής $F(x; w, \theta)$. Στον εν λόγω συμβολισμό, με x αναπαριστούμε το διάνυσμα χαρακτηριστικών εισόδου (υλοποίηση του X), με w ένα διάνυσμα *εκμαθήσιμων παραμέτρων*, το οποίο προσδιορίζεται από τον αλγόριθμο μάθησης και θ ένα *διάνυσμα υπερπαραμέτρων*, το οποίο ορίζεται από τον σχεδιαστή. Για κάθε τιμή του θ, επάγεται και ένας διαφορετικός χώρος υποθέσεων $H(\theta)$.

Η θεμελιώδης ιδέα που διέπει το σύνολο των αρχιτεκτονικών είναι αυτή ενός γενικευμένου "διαίρει και βασίλευε". Σε ένα τέτοιο πλαίσιο, το διαφοροποιητικό στοιχείο της εκάστοτε αρχιτεκτονικής συνίσταται στον τρόπο με τον οποίο *διασπά* τη συνολική εργασία του υπολογισμού της εξόδου y από την είσοδο x σε επιμέρους, απλούστερους υπολογισμούς και *συνδυάζει* τα αποτελέσματά τους ώστε να παράγει την y. Οι ιδέες αυτές θα αναδειχθούν πληρέστερα στη συνέχεια της ενότητας, όπου και θα εξετάσουμε συγκεκριμένες υλοποιήσεις μοντέλων.

Η παραμετροποίηση ως προς τα w, θ προσδίδει στην αρχιτεκτονική την απαραίτητη ευελιξία και ικανότητα προσαρμογής. Η προσαρμογή αυτή μπορεί να νοηθεί σε δύο άξονες: *Προσαρμογή στην εγγενή πολυπλοκότητα του προβλήματος* και *προσαρμογή στα δεδομένα.*

Το πρώτο σκέλος ελέγχεται κατά κύριο λόγο από το θ και στοχεύει στην κλιμάκωση της αρχιτεκτονικής (αύξηση της χωρητικότητάς της) με τρόπο τέτοιο ώστε ο $H(\theta)$ να μπορεί να αναπαραστήσει τη ζητούμενη συνάρτηση με πιστότητα. Από την άλλη βέβαια, εάν ο $H(\theta)$ είναι χώρος υψηλής χωρητικότητας και το πρόβλημα είναι εγγενώς χαμηλής πολυπλοκότητας (π.χ. η ζητούμενη συνάρτηση έχει γραμμική φύση), τότε εγκυμονεί ο κίνδυνος *υπερπροσαρμογής*, "υπερβολικής προσαρμογής", δηλαδή, στα δεδομένα εκπαίδευσης. Η έννοια της υπερπροσαρμογής θα αναλυθεί με λεπτομέρεια στη συνέχεια του κεφαλαίου.

Δεδομένου ότι δεν υπάρχει κάποιος καλά ορισμένος τρόπος ώστε a priori να αποτιμήσουμε την πολυπλοκότητα ενός προβλήματος, η επιλογή των υπερπαραμέτρων γίνεται βάσει δοκιμών: Από το αρχικό σύνολο δεδομένων, απομονώνουμε ένα ποσοστό δειγμάτων, το οποίο θα συγκροτεί το *σύνολο επικύρωσης* (validation set). Κατόπιν,

εξετάζουμε κάποιους πιθανούς συνδυασμούς $\theta_1, \theta_2, \ldots, \theta_n$ από τυπικές τιμές των υπερπαραμέτρων, οι οποίες έχουν προκύψει από την πειραματική εμπειρία, και για καθένα εκ των θ_i, *εκπαιδεύουμε* το μοντέλο μέσω του αλγορίθμου μάθησης. Τελικά, διατηρούμε το θ^* για το οποίο το εκπαιδευμένο μοντέλο επιτυγχάνει τη μέγιστη "ακρίβεια" πρόβλεψης στο σύνολο επικύρωσης. Η έννοια της ακρίβειας μπορεί να οριστεί με διάφορους τρόπους, όπως θα αναλυθεί στην ενότητα για την αξιολόγηση των μοντέλων.

Αξίζει, εδώ, να σημειωθεί πως, για την αποφυγή της υπερπροσαρμογής, εκτός του ελέγχου της χωρητικότητας του χώρου υποθέσεων, μας ενδιαφέρει και ο έλεγχος της διαδικασίας μάθησης, ώστε το τελικό εκπαιδευμένο μοντέλο να παρουσιάζει την ελάχιστη πολυπλοκότητα που απαιτείται για ένα δεδομένο επίπεδο ακρίβειας. Η ποσοτικοποίηση της πολυπλοκότητας σε επίπεδο παραμέτρων αναλύεται στη συνέχεια του κεφαλαίου. Η διαδικασία, όμως, του ελέγχου αυτού συνδέεται με κάποιες επιπλέον υπερπαραμέτρους, τις υπερπαραμέτρους εκπαίδευσης, οι οποίες προστίθενται στο αρχικό διάνυσμα θ, διαμορφώνοντας το επαυξημένο διάνυσμα υπερπαραμέτρων $\overline{\theta}$. Εν γένει, όταν αναφερόμαστε στη ρύθμιση των υπερπαραμέτρων, αναφερόμαστε στην επιλογή του $\overline{\theta}$.

Τέλος, αναφορικά με το δεύτερο σκέλος, την προσαρμογή δηλαδή της αρχιτεκτονικής στα δεδομένα, αυτό επιτυγχάνεται μέσω του διανύσματος παραμέτρων, όπως περιγράφεται ακολούθως: Λαμβάνοντας υπ' όψιν τη σχέση (1.3) και το γεγονός πως μια αρχιτεκτονική παραμετροποιείται ως προς w, θ, το πρόβλημα της παλινδρόμησης μπορεί να διατυπωθεί ως το εξής γενικό πρόβλημα ελαχιστοποίησης:

$$\varphi^* = argmin_{\varphi \in H} \mathbb{E}_{\sim D(X,Y)}(L(Y, \varphi(X; w, \theta))) \qquad (1.10)$$

Δεδομένου, όμως, ότι το διάνυσμα υπερπαραμέτρων έχει ήδη καθοριστεί από τον σχεδιαστή πριν τη διαδικασία της εκπαίδευσης, ο χώρος υποθέσεων ανάγεται στον $H(\theta)$, ο οποίος είναι ισόμορφος με τον χώρο $\mathbb{R}^{dim(w)}$, όπου $dim(w)$ η διάσταση του διανύσματος παραμέτρων w. Συνεπώς, μπορούμε ισοδύναμα να επαναδιατυπώσουμε το αρχικό πρόβλημα συναρτησιακής βελτιστοποίησης ως ένα πρόβλημα διανυσματικής βελτιστοποίησης στον $\mathbb{R}^{dim(w)}$:

$$w^* = argmin_{w \in \mathbb{R}^{dim(w)}} E_{\sim D(X,Y)}(L(Y, \varphi(X; w, \theta)) \qquad (1.11)$$

Από τα παραπάνω, είναι προφανές ότι $\varphi^*(\theta) = \varphi(w^*, \theta)$. Όμως, όπως έχει ήδη αναφερθεί στην Παράγραφο 1.2, το (1.11) είναι εν γένει μη υπολογίσιμο, εφόσον η κατανομή $D(X, Y)$ μας είναι άγνωστη. Ένας αρκετά δημοφιλής μετασχηματισμός του προβλήματος (όχι όμως και ο μοναδικός, όπως θα δούμε για παράδειγμα στα μοντέλα SVR) προκύπτει μέσω της προσέγγισης της μέσης τιμής της συνάρτησης απωλειών από τον εμπειρικό της μέσο:

$$w^{*} = argmin_{w\in\mathbb{R}^{dim(w)}} \frac{1}{N} \sum_{i=1}^{N} (L(y_i, \varphi(x_i; w, \theta)) \qquad (1.12)$$

Η μορφή (1.12) είναι επιλύσιμη, μέσω αλγορίθμων που θα εξεταστούν στη συνέχεια. Παρατηρούμε, λοιπόν, ότι, παραμετροποιώντας την αρχιτεκτονική ως προς το w, έχουμε τη δυνατότητα να την προσαρμόσουμε κατά βέλτιστο τρόπο στα δεδομένα $\{x_i, y_i\}_{i=1}^{N}$. Αυτό είναι σαφώς σημαντικό, εφόσον η επίδοση ενός μοντέλου ελεγχόμενης πολυπλοκότητας στα διαθέσιμα δεδομένα της $Y = f(X)$ δύναται να είναι ισχυρά συσχετισμένη με τη γενική στατιστική του επίδοση στην προσέγγιση της $f(\cdot)$. Με κατάλληλη επιλογή του $\bar{\theta}$, φιλοδοξούμε ότι η βέλτιστη αυτή επίδοση θα είναι και αποδεκτής ακρίβειας.

Ακολούθως, θα εξειδικεύσουμε το κρίσιμο σχεδιαστικό ζήτημα της επιλογής, της εκπαίδευσης και της ρύθμισης υπερπαραμέτρων μοντέλων σε στατικά και δυναμικά προβλήματα.

- **Στατικά προβλήματα**

Οι αρχιτεκτονικές μοντέλων τις οποίες θα εξετάσουμε πηγάζουν από δύο θεμελιώδη σχετικά υποδείγματα (paradigms): Τα *νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης* (feedforward neural networks) και τις *μηχανές boosting*.

Ι) *Νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης (Feedforward neural networks - FFNNs)*

Σε υψηλό επίπεδο αφαίρεσης, η αρχιτεκτονική των νευρωνικών δικτύων βασίζεται στην εξής ιδέα: Έστω ότι θέλουμε να υπολογίσουμε μια αυθαίρετα πολύπλοκη, μη γραμμική συνάρτηση $f: \mathbb{R}^d \to \mathbb{R}$, μεταξύ των τυχαίων μεταβλητών X και Y. Πραγματοποιούμε την εικασία ότι η συνάρτηση αυτή μπορεί να αναπαρασταθεί ως $f(X) = g(h(X))$, ως η σύνθεση, δηλαδή, μιας μη γραμμικής συνάρτησης $h: \mathbb{R}^d \to \mathbb{R}^m$ και μιας γραμμικής συνάρτησης $g: \mathbb{R}^m \to \mathbb{R}$. Κατ' αυτόν τον τρόπο, ο υπολογισμός της f απαιτεί τον μετασχηματισμό του X σε ένα διάνυσμα $\overline{X}$, το οποίο έχει γραμμική σχέση με τη μεταβλητή εξόδου Y.

Το ερώτημα που προκύπτει εύλογα από τα παραπάνω είναι το πώς θα κατασκευάσουμε έναν μετασχηματισμό h, ο οποίος θα έχει τη χωρητικότητα ώστε να μετατρέψει μια μη γραμμική σχέση σε γραμμική.

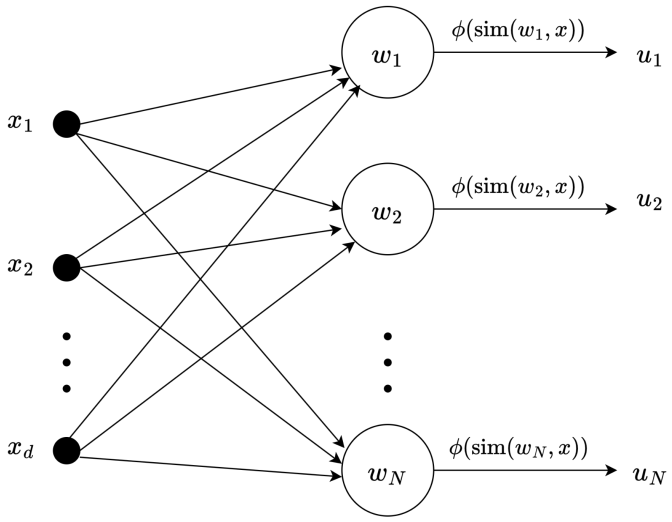
Λήμμα 1.2: Μια απεικόνιση $g: \mathbb{R}^d \to \mathbb{R}$ είναι γραμμική αν και μόνο εάν οι ισοδυναμικές της επιφάνειες είναι παράλληλα υπερεπίπεδα της μορφής $w^T x - c = 0$, όπου $w \in \mathbb{R}^d, c \in \mathbb{R}$.

Σε μια γενική, μη γραμμική συνάρτηση, οι ισοδυναμικές υπερεπιφάνειες είναι αυθαίρετης μορφής και άρα όχι κατ' ανάγκη επίπεδες. Επομένως, από το παραπάνω λήμμα, συνάγουμε ότι μια αναγκαία ιδιότητα που καλείται να φέρει ο μετασχηματισμός h είναι

αυτή της δυνατότητας απεικόνισης μη γραμμικών περιοχών σε επίπεδες επιφάνειες. Βασιζόμενοι στην απαίτηση αυτή, θα θεμελιώσουμε τη γενική αρχιτεκτονική των FFNNs.

Έστω, λοιπόν, μια διαμέριση του χώρου $\mathbb{R}^d$ σε επιμέρους, μη γραμμικές περιοχές R_i, $1 \leq i \leq C$. Προκειμένου να σχεδιάσουμε έναν μετασχηματισμό h, ο οποίος να απεικονίζει έκαστη των R_i σε έναν επίπεδο χώρο H_i, θα αναπαραστήσουμε τα στοιχεία του $\mathbb{R}^d$ μέσω της "ομοιότητάς" τους με ένα σύνολο N *πρότυπων διανυσμάτων* $w_1,\ldots,w_N \in \mathbb{R}^d$, η επιλογή των οποίων είναι στην ευχέρεια του σχεδιαστή. Η "ομοιότητα" $sim(x, w_i)$ μεταξύ ενός διανύσματος $x \in R_i$ και ενός πρότυπου διανύσματος w_i εκφράζεται εν γένει μέσω του εσωτερικού γινομένου ή της ευκλείδειας απόστασης και καταχωρίζεται σε ένα *διάνυσμα ομοιότητας* $u = [sim(x, w_1),\ldots, sim(x, w_N)]^T$. Εάν, τώρα, εφαρμόσουμε έναν σημειακό (pointwise) μη γραμμικό μετασχηματισμό $\varphi(\cdot)$ στο διάνυσμα u, ο οποίος να συμπιέζει στο 0 τα στοιχεία που υποδηλώνουν ανομοιότητα (αρνητικό εσωτερικό γινόμενο ή μεγάλη ευκλείδεια απόσταση), τότε τα διανύσματα του $\mathbb{R}^d$ που παρουσιάζουν ομοιότητα (δηλαδή, μη ανομοιότητα) με το ίδιο υποσύνολο των w_i, έχουν τις ίδιες ενεργές (μη μηδενικές) διαστάσεις και άρα *απεικονίζονται στον ίδιο διανυσματικό υπόχωρο του* $\mathbb{R}^m$. Αν λοιπόν επιλέξουμε κατάλληλα τα πρότυπα διανύσματα, τον μη γραμμικό μετασχηματισμό και το μέτρο ομοιότητας, τότε είναι δυνητικά εφικτό τα στοιχεία της εκάστοτε R_i να είναι όμοια με το ίδιο υποσύνολο πρότυπων διανυσμάτων, και άρα να απεικονίζονται στον ίδιο επίπεδο χώρο H_i, όπως ήταν και το αρχικό ζητούμενο. Επισημαίνεται ότι η ανάλυση αυτή δεν συνιστά τυπική απόδειξη της αρχής λειτουργίας των FFNNs, παρά μόνο παρέχει τη γεωμετρική διαίσθηση πίσω από την αρχιτεκτονική τους.

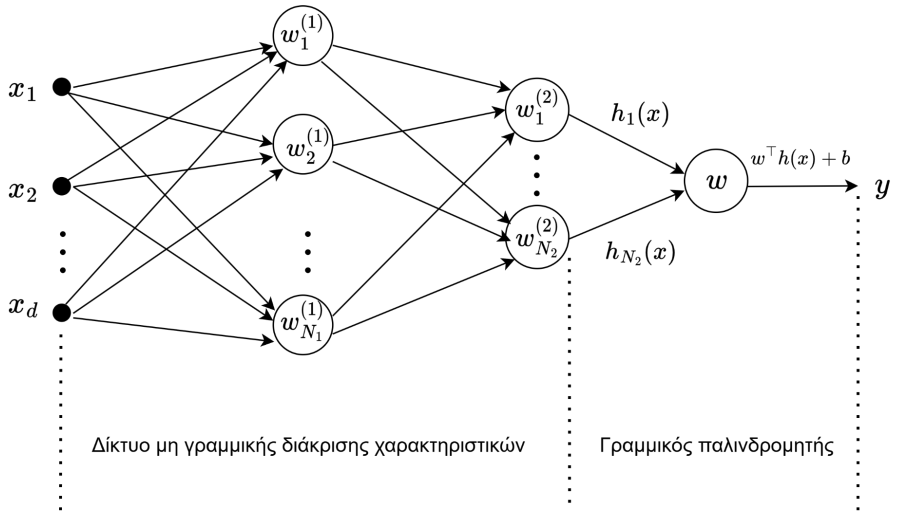
Καθένα εκ των w_i αντιστοιχεί σε έναν *νευρώνα* (ή *πυρήνα)*, μια στοιχειώδη υπολογιστική μονάδα η οποία υλοποιεί την απεικόνιση $\varphi(sim(x, w_i))$. Ο μετασχηματισμός του διανύσματος εισόδου x διά μέσου του συνόλου των νευρώνων (*επίπεδο νευρώνων*) απεικονίζεται γραφικά παρακάτω:



***Εικόνα 1.4**: Γραφική απεικόνιση ενός επιπέδου νευρώνων ενός FFNN*

Διασυνδέοντας με ακολουθιακό τρόπο επίπεδα νευρώνων, έχουμε εν γένει τη δυνατότητα να πετύχουμε πολυπλοκότερες διακρίσεις του αρχικού χώρου χαρακτηριστικών. Η δομή που προκύπτει αποτελεί ένα *δίκτυο* μη γραμμικής διάκρισης χαρακτηριστικών και υλοποιεί την h μέσω ενός σύνθετου μετασχηματισμού της μορφής $\varphi(sim(W^{(N_l)}, \varphi(\dots \varphi(sim(W^{(1)}, x)))$, όπου $W^{(i)}$ ο πίνακας με γραμμές τα πρότυπα διανύσματα του i-οστού επιπέδου και N_l το πλήθος των επιπέδων.

Τοποθετώντας, τέλος, έναν γραμμικό μετασχηματισμό της μορφής $y = w^T h(x) + b$ ως στάδιο εξόδου, έχουμε πλέον διαμορφώσει πλήρως μια αρχιτεκτονική FFNN, η οποία και παρουσιάζεται στο ακόλουθο σχήμα:



***Εικόνα 1.5**: Γραφική απεικόνιση της αρχιτεκτονικής ενός FFNN*

Τα μοντέλα νευρωνικών δικτύων για στατικά προβλήματα παλινδρόμησης διακρίνονται περαιτέρω σε δύο βασικές κατηγορίες, τα *πολυστρωματικά perceptrons (MLPs)* [2] και τους *παλινδρομητές διανυσμάτων υποστήριξης (SVRs)* [3]. Για καθένα εκ των δύο, θα αναφερθούμε λεπτομερώς στα διαφοροποιητικά τους στοιχεία, τόσο σε επίπεδο αρχιτεκτονικής, όσο και σε επίπεδο εκπαίδευσης.

Ι.1) *Πολυστρωματικά perceptrons (Multilayer perceptrons - MLPs)*

Το βασικό γνώρισμα που διακρίνει τα MLPs από τις υπόλοιπες αρχιτεκτονικές νευρωνικών δικτύων (όπως τα SVRs ή άλλες *μέθοδοι πυρήνα*) είναι το γεγονός πως είναι *πλήρως εκπαιδεύσιμα*. Όπως θα εξηγηθεί και στην παράγραφο για τον αλγόριθμο μάθησης, τόσο τα πρότυπα διανύσματα, όσο και οι παράμετροι του γραμμικού παλινδρομητή εξόδου προσδιορίζονται από κοινού από τον αλγόριθμο μάθησης του MLP και μάλιστα με τρόπο που να βελτιστοποιεί τη συνολική επίδοση του συστήματος.

Όσον αφορά τα χαρακτηριστικά των νευρωνικών υπολογισμών που εκτελούνται στα κρυφά επίπεδα του μοντέλου, η συνάρτηση ομοιότητας που επιλέγεται είναι αυτή του πολωμένου εσωτερικού γινομένου, ενώ ο μη γραμμικός μετασχηματισμός είναι συνήθως αυτός της $ReLU(\cdot)$. Ισχύει δηλαδή:

$$sim(w_i, x) = w_i^T x + b \tag{1.13}$$

$$\varphi(x) = ReLU(x) = max(0, x) \tag{1.14}$$

Από τα παραπάνω, προκύπτει ότι ένα MLP διαχωρίζει τον χώρο εισόδου σε *κυρτά πολύεδρα*, τα οποία, κατά τα γνωστά, απεικονίζει σε επίπεδες επιφάνειες.

Έως τώρα, έχουμε εστιάσει στα εγγενή χαρακτηριστικά της αρχιτεκτονικής των FFNNs (και κατ' επέκταση των MLPs), τα οποία τους επιτρέπουν να υπολογίζουν μη γραμμικές συναρτήσεις εν γένει. Ωστόσο, όπως αναφέρθηκε και στην εισαγωγή της ενότητας 1.3.3, στο πλαίσιο της επιβλεπόμενης μάθησης, ο στόχος είναι να προσαρμόσουμε το μοντέλο στην πολυπλοκότητα και τα δεδομένα ενός συγκεκριμένου προβλήματος πρόβλεψης. Η προσαρμογή του μοντέλου πραγματοποιείται, κατά τα γνωστά, σε επίπεδο παραμέτρων και υπερπαραμέτρων και, ως εκ τούτου, θεωρώντας εν προκειμένω το MLP, είναι σκόπιμο να αναφερθούμε στον αλγόριθμο μάθησης που χρησιμοποιεί, καθώς και στις υπερπαραμέτρους που διαθέτει.

- *Αλγόριθμος μάθησης*

Το πρόβλημα μάθησης για ένα MLP διατυπώνεται στη μορφή της ελαχιστοποίησης του μέσου εμπειρικού σφάλματος, όπως αυτή παρουσιάζεται στη σχέση (1.12). Για την επίλυση της (1.12) καλούμαστε να λάβουμε δύο επιπλέον σχεδιαστικές αποφάσεις, οι οποίες αφορούν στην επιλογή της μορφής της συνάρτησης απωλειών L και του αλγορίθμου βελτιστοποίησης. Ο αλγόριθμος βελτιστοποίησης που θα επιλέξουμε, όντας κατ' ουσίαν μια συστηματική μέθοδος αναζήτησης στον χώρο υποθέσεων $H(\theta)$, θα αποτελεί και τον αλγόριθμο μάθησης.

- *Επιλογή συνάρτησης απωλειών*

Η συνάρτηση απωλειών L προφανώς θα αποτελεί μια συνάρτηση απόστασης, η οποία συνήθως επάγεται από την L1 $(L_i = d(y_i, \varphi(x_i)) = |y_i - \varphi(x_i)|)$, ή την L2 νόρμα $(L_i = d(y_i, \varphi(x_i)) = (y_i - \varphi(x_i))^2)$.

Μολονότι η διαδικασία της επιλογής της συνάρτησης απωλειών φαίνεται αρκετά απλή, υπάρχει ένα λεπτό σημείο που θα πρέπει να λάβουμε σχεδιαστικά υπ' όψιν: Παρατηρούμε ότι η μαθηματική διατύπωση του προβλήματος ελαχιστοποίησης δεν θα διέφερε αν αυτοσκοπός του βέλτιστου μοντέλου ήταν να ελαχιστοποιήσει το μέσο εμπειρικό σφάλμα στο σύνολο εκπαίδευσης, το οποίο ορίζεται ως $\frac{1}{N}\sum_{i=1}^{N}(L(y_i, \varphi(x_i; w, \theta))$. Η σχεδίαση, όμως, θα πρέπει να αποσκοπεί στο να εξαάγει ένα μοντέλο που να παράγει χαμηλό μέσο σφάλμα E_{out} *σε οποιοδήποτε τυχαίο σύνολο* N *δειγμάτων*. Το ερώτημα, λοιπόν, που προκύπτει είναι εάν η ελαχιστοποίηση του E_{in} επαρκεί ώστε να κατευθύνει τη σχεδίαση σε ένα μοντέλο φ^* που να επιτυγχάνει επίσης χαμηλό E_{out}.

Για να απαντήσουμε στο ερώτημα αυτό, θα πρέπει να κατανοήσουμε τις συνθήκες υπό τις οποίες τα E_{in} και E_{out} εμφανίζουν απόκλιση. Αν αποκλείσουμε το ενδεχόμενο το δείγμα εκπαίδευσης να μην είναι αντιπροσωπευτικό της κατανομής $D(X, Y)$, τότε η απόκλιση αυτή υποδηλώνει κατ' ανάγκη ότι το μοντέλο δεν μαθαίνει μια συνάρτηση που να προσεγγίζει την τυχαία μεταβλητή Y, αλλά τείνει απλώς να *απομνημονεύει* τα δεδομένα εκπαίδευσης. Η συμπεριφορά αυτή περιγράφεται με τον όρο *υπερπροσαρμογή*, όπως αναφέρεται και στην αρχή της ενότητας.

Πράγματι, εφόσον υπάρχει μια απειρία συναρτήσεων που αναπαράγει τα ζεύγη εκπαίδευσης (x_i, y_i), ανάμεσα σε αυτές θα συγκαταλέγονται εύρωστες υλοποιήσεις, οι οποίες προσεγγίζουν την Y σε οποιοδήποτε τυχαίο δείγμα, αλλά και *ασταθείς* υλοποιήσεις, οι οποίες αποτυγχάνουν να γενικεύσουν εκτός του δείγματος εκπαίδευσης. Η ελαχιστοποίηση του εμπειρικού σφάλματος αποτελεί ένα διαφανές (transparent) κριτήριο ως προς αυτές τις δύο κατηγορίες μοντέλων, υπό την έννοια πως δεν αποδίδει προτίμηση σε καμία εκ των δύο κατηγοριών.

Ο προσανατολισμός της σχεδίασης σε εύρωστα μοντέλα με ικανότητα γενίκευσης επιτυγχάνεται ελέγχοντας την πολυπλοκότητά τους. Η πολυπλοκότητα, όπως έχει αναφερθεί, συναρτάται τόσο με τις υπερπαραμέτρους, όσο και με τις παραμέτρους: Στην πρώτη περίπτωση, είναι προφανές πως σε έναν χώρο υποθέσεων με πολύ μεγάλη χωρητικότητα (εν προκειμένω, MLPs με πολλούς νευρώνες), είναι υπαρκτό το ενδεχόμενο να συγκαταλέγονται και ασταθείς συναρτήσεις. Για την ακρίβεια, η πολυπλοκότητα (σε συνδυασμό με τον χαμηλό όγκο δεδομένων) είναι *αναγκαία συνθήκη* ώστε να είναι δυνατή η αστάθεια. Η ιδέα αυτή αποτυπώνεται στην παρακάτω γενικευμένη ανισότητα της θεωρίας μάθησης [4]:

$$Pr(E_{in}(\varphi) - E_{out}(\varphi) \leq C(\delta; R(H), N)) \geq 1 - \delta \qquad (1.15)$$

Η συνάρτηση $R(H)$ αποτελεί κάποιο μέτρο της πολυπλοκότητας του χώρου υποθέσεων (π.χ. η *πολυπλοκότητα Rademacher*). Η συνάρτηση C "τιμωρεί" την πολυπλοκότητα του H (είναι αύξουσα συνάρτηση του $R(H)$. Ακόμη, εξαρτάται με αύξοντα τρόπο από την παράμετρο δ και το μέγεθος του συνόλου δεδομένων N. Στην ουσία, η ανισότητα αυτή δηλώνει πως, για ένα επίπεδο εμπιστοσύνης $1 - \delta$, τα σφάλματα $E_{in}(\varphi), E_{out}(\varphi)$ για ένα μοντέλο φ θα απέχουν το πολύ κατά C. Όσο πιο πολύπλοκος είναι ο χώρος υποθέσεων ή/και όσο πιο λίγα δεδομένα διαθέτουμε, τόσο μεγαλύτερη θα είναι η απόκλιση για την οποία είναι βέβαιο με πιθανότητα $1 - \delta$ ότι δεν υπερβαίνεται.

Σε επίπεδο παραμέτρων, η πολυπλοκότητα ενός μοντέλου μπορεί να ποσοτικοποιηθεί από την L2 νόρμα $||w||_2 = \sum_{j=1}^{dim(w)} w_j^2$, όπου $dim(w)$ η διάσταση του w. Διαισθητικά, η ιδέα αυτή επεξηγείται ως εξής: Από την (1.13), παρατηρούμε ότι για ένα σήμα ομοιότητας σε κάποιο εσωτερικό επίπεδο του μοντέλου (έστω το i-οστό επίπεδο νευρώνων), ισχύει $\frac{\partial sim(w_i x_j)}{\partial x_j} = w_i$. Επομένως, όσο μεγαλύτερες είναι οι τιμές των w_i, τόσο μεγαλύτερες είναι οι μεταβολές των σημάτων εντός του δικτύου σε σχέση με τις εισόδους τους. Αυτή είναι μια συμπεριφορά που δυνητικά υποδηλώνει αστάθεια και για τον λόγο αυτό, *επαυξάνουμε* την αρχική συνάρτηση απωλειών με έναν όρο *ποινικοποίησης της πολυπλοκότητας.*

$$\overline{L} = \frac{1}{N}\sum_{i=1}^{N}(L(y_i, \varphi(x_i; w, \theta)) + C||w||^2,\ C \in \mathbb{R}^+ \qquad (1.16)$$

Μέσω της C, η οποία αποτελεί υπερπαράμετρο εκπαίδευσης, ελέγχουμε τον βαθμό στον οποίο ο αλγόριθμος ελαχιστοποίησης θα λαμβάνει υπ' όψιν την παραμετρική πολυπλοκότητα ενός μοντέλου σε σχέση με την ακρίβεια που αυτό επιτυγχάνει στο σύνολο εκπαίδευσης.

Συμπερασματικά, ο στόχος του παραμετρικού ελέγχου της πολυπλοκότητας είναι, από ένα σύνολο μοντέλων του χώρου υποθέσεων $H(\theta)$ τα οποία παράγουν ένα παρόμοιο E_{in}, να εξάγει το "απλούστερο", θεωρώντας ότι θα είναι και το πιο εύρωστο. Η ιδέα αυτή αποτελεί μια έκφραση της αρχής του Occam:

Όταν δύο θεωρίες παρέχουν εξίσου ακριβείς προβλέψεις, πάντα επιλέγουμε την απλούστερη.

- *Επιλογή αλγορίθμου βελτιστοποίησης*

Θα εστιάσουμε σε αλγορίθμους βελτιστοποίησης οι οποίοι βασίζονται στην ιδέα της *επαναληπτικής τοπικής μείωσης* της συνάρτησης, έως ότου αυτή συγκλίνει στο/στα σημεία ελαχίστου.

Ορισμός 1.7: Ένας αλγόριθμος ελαχιστοποίησης για μια υποκείμενη συνάρτηση L θα ονομάζεται *μέθοδος καθόδου (descent method),* εφόσον συγκλίνει στο/στα βέλτιστα σημεία μέσω μιας αναδρομικής διαδικασίας $w_{k+1} = w_k - \eta \cdot p_k$, $\eta > 0$. Η παράμετρος η ονομάζεται *ρυθμός μάθησης*, ενώ το διάνυσμα p_k ονομάζεται *κατεύθυνση καθόδου* και φέρει τις εξής ιδιότητες:

(1) $p_k^T \nabla L(w_k) < 0$, αν $\nabla L(w_k) \neq 0$
(2) $p_k = 0$, αν $\nabla L(w_k) = 0$

Η πρώτη συνθήκη δηλώνει ότι, στο τρέχον σημείο w_k, η παράγωγος προς την κατεύθυνση του p_k είναι αρνητική. Επομένως, αν κινηθούμε τοπικά στην κατεύθυνση αυτή, αναμένουμε να παρατηρήσουμε μείωση της τιμής της L ($L(w_{k+1}) < L(w_k)$). Η δεύτερη συνθήκη δηλώνει ότι η κατεύθυνση καθόδου μηδενίζεται στα *στάσιμα σημεία* της L, τα οποία αποτελούν και υποψήφια σημεία ελαχίστου. Αποδεικνύεται εύκολα ότι η αρνητική κλίση $-\nabla L(w_k)$ της υποκείμενη συνάρτησης αποτελεί κατεύθυνση καθόδου.

Για τη συγκεκριμένη επιλογή της κατεύθυνσης καθόδου, η μέθοδος ονομάζεται *μέθοδος κατάβασης κλίσης* (gradient descent) και αποτελεί μια δημοφιλή επιλογή για προβλήματα μάθησης. Ωστόσο, σε περιπτώσεις όπου η συνάρτηση L εμφανίζει πολύπλοκη "μορφολογία" (πολλά τοπικά μέγιστα/ελάχιστα, υψηλή μεταβλητότητα κλίσης), είναι πιθανό η τοπική - μικρής κλίμακας πληροφορία που περιέχει το διάνυσμα κλίσης να μην επαρκεί ώστε να αποτυπώσει τη γενική - μεγάλης κλίμακας κατεύθυνση κατά την οποία η L ελαχιστοποιείται. Στην περίπτωση του MLP, η πολυπλοκότητα της αρχιτεκτονικής θεωρούμε ότι επάγει και αντίστοιχη πολυπλοκότητα στην L.

Λαμβάνοντας υπόψη τα παραπάνω, θα επιλέξουμε ως αλγόριθμο τον *Adam* [5], ο οποίος είναι μεν βασιζόμενος στην κλίση (gradient based), αλλά διαμορφώνει την κατεύθυνση βελτιστοποίησης p_k βάσει των *ροπών* (moments) της κλίσης - μέση τιμή και μη κεντραρισμένη διασπορά. Σημειώνεται ότι, στον εν λόγω αλγόριθμο, η p_k εν γένει δεν είναι κατεύθυνση καθόδου κατά την αυστηρή έννοια που ορίστηκε παραπάνω.

Algorithm 1 Αλγόριθμος Adam

Είσοδος: συνάρτηση κόστους $L(w)$, αρχικές παράμετροι w_0
αρχικοποίηση $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $k \leftarrow 0$
επιλέγουμε υπερπαραμέτρους: η, β_1, β_2, ϵ
while όχι σύγκλιση **do**
    $k \leftarrow k + 1$
    υπολογισμός κλίσης: $g_k \leftarrow \nabla_w L(w_{k-1})$
    $m_k \leftarrow \beta_1 \cdot m_{k-1} + (1 - \beta_1) \cdot g_k$
    $v_k \leftarrow \beta_2 \cdot v_{k-1} + (1 - \beta_2) \cdot g_k^2$
    $\hat{m}_k \leftarrow \frac{m_k}{1-\beta_1^k}$
    $\hat{v}_k \leftarrow \frac{v_k}{1-\beta_2^k}$
    ενημέρωση παραμέτρων: $w_k \leftarrow w_{k-1} - \eta \cdot \frac{\hat{m}_k}{\sqrt{\hat{v}_k}+\epsilon}$
end while
Επιστροφή: τελικές παράμετροι w_k

Επισημαίνεται ότι, για λόγους μείωσης του υπολογιστικού κόστους, η συνάρτηση σφάλματος *L* δεν υπολογίζεται επάνω σε όλα τα δείγματα του συνόλου εκπαίδευσης, αλλά σε μικρότερα υποσύνολα αυτού, τα οποία επιλέγονται τυχαία και καλούνται *δέσμες (batches)*. Το μήκος κάθε δέσμης αποτελεί υπερπαράμετρο εκπαίδευσης του MLP. Όταν ο αλγόριθμος ελαχιστοποίησης έχει εξετάσει όλες τις δέσμες που συνιστούν το σύνολο εκπαίδευσης, τότε λέμε ότι έχει παρέλθει μια *εποχή*. Το πλήθος των εποχών αποτελεί επίσης υπερπαράμετρο εκπαίδευσης.

- *Υπερπαράμετροι του MLP*

Στον ακόλουθο πίνακα, παρουσιάζονται οι βασικές υπερπαράμετροι αρχιτεκτονικής και εκπαίδευσης ενός MLP. Συμβατικά, επιλέγουμε ένα δίκτυο δύο κρυφών επιπέδων, με το πλήθος νευρώνων στο πρώτο επίπεδο να είναι μεγαλύτερο σε σχέση με του δευτέρου.

Υπερπαράμετρος	**Επεξήγηση**	**Τυπικές Τιμές**
Νευρώνες 1ου επιπέδου	Πλήθος νευρώνων στο πρώτο κρυφό επίπεδο του MLP.	32, 64, 128
Νευρώνες 2ου επιπέδου	Πλήθος νευρώνων στο δεύτερο κρυφό επίπεδο.	16, 32, 64, 128
Weight decay	Συντελεστής κανονικοποίησης L2, συμβάλλει στην αποφυγή υπερπροσαρμογής (overfitting).	10^{-5}, 10^{-4}, 10^{-3}
Learning rate	Ρυθμός μάθησης, καθορίζει το μέγεθος του βήματος κατά τη βελτιστοποίηση.	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
Αριθμός εποχών	Ορίζει το πλήθος των ολοκληρωμένων διελεύσεων του συνόλου εκπαίδευσης από τον αλγόριθμο μάθησης	20, 50, 100
Μέγεθος δέσμης	Πλήθος δειγμάτων που υποβάλλονται σε επεξεργασία πριν από την ενημέρωση των βαρών. Μικρότερα μεγέθη επιτρέπουν πιο συχνές ενημερώσεις, αλλά είναι πιο θορυβώδη.	64, 128, 256, 512

***Πίνακας 1.1**: Πίνακας υπερπαραμέτρων του MLP*

Ι.2) *Παλινδρομητές διανυσμάτων υποστήριξης (Support Vector Regressors - SVRs)*

Η αρχιτεκτονική των SVRs θεμελιώνεται βάσει της απαίτησης το πρόβλημα μάθησης να αποτελεί ένα δομημένο και καλώς ορισμένο πρόβλημα ελαχιστοποίησης. Σε αντίθεση με τα MLPs, όπου η υποκείμενη συνάρτηση απωλειών δύναται να διαθέτει πολλαπλά τοπικά βέλτιστα, στα SVRs το πρόβλημα ελαχιστοποίησης σχεδιάζεται ώστε να είναι *κυρτό*.

Ορισμός 1.8: Ένα σύνολο $S \subseteq \mathbb{R}^n$ λέγεται *κυρτό* αν $\forall x, y \in S$ και κάθε $\lambda \in [0,1]$, το διάνυσμα $\lambda x + (1-\lambda)y$ ανήκει επίσης στο S.

Ορισμός 1.9: Μια συνάρτηση $f: S \to \mathbb{R}$, όπου $S \subseteq \mathbb{R}^n$ ένα κυρτό σύνολο, λέγεται *κυρτή* αν $\forall x, y \in S$ και κάθε $\lambda \in [0,1]$, ισχύει $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. Όταν η παραπάνω ανισότητα ισχύει ως ισότητα, τότε λέμε ότι η συνάρτηση είναι *αυστηρά κυρτή*.

Ορισμός 1.10: Ένα πρόβλημα εύρεσης ελαχίστου $x^* = argmin_{x \in S} f(x)$ ονομάζεται *κυρτό,* εφόσον η συνάρτηση $f(x)$ και το σύνολο S είναι κυρτά.

Πρόταση 1.2: Για μια κυρτή συνάρτηση f, ένα σημείο τοπικού ελαχίστου θα είναι και σημείο ολικού ελαχίστου. Αν η f είναι αυστηρά κυρτή, τότε υπάρχει σημείο τοπικού ελαχίστου.

Για προβλήματα ελαχιστοποίησης με κυρτή δομή χρησιμοποιούνται ειδικοί αλγόριθμοι, οι οποίοι, αξιοποιώντας τη δομή αυτή, εξασφαλίζουν ισχυρότερες θεωρητικές εγγυήσεις σύγκλισης σε ελάχιστο σε σχέση με τους γενικούς αλγορίθμους τύπου Adam. Επιπλέον σε αυτό, αν η υποκείμενη συνάρτηση είναι αυστηρά κυρτή, τότε έχει και μοναδικό ελάχιστο, καθιστώντας το πρόβλημα βελτιστοποίησης καλώς ορισμένο. Από μαθηματικής άποψης, αυτές οι δύο ιδιότητες είναι ιδιαίτερα επιθυμητές.

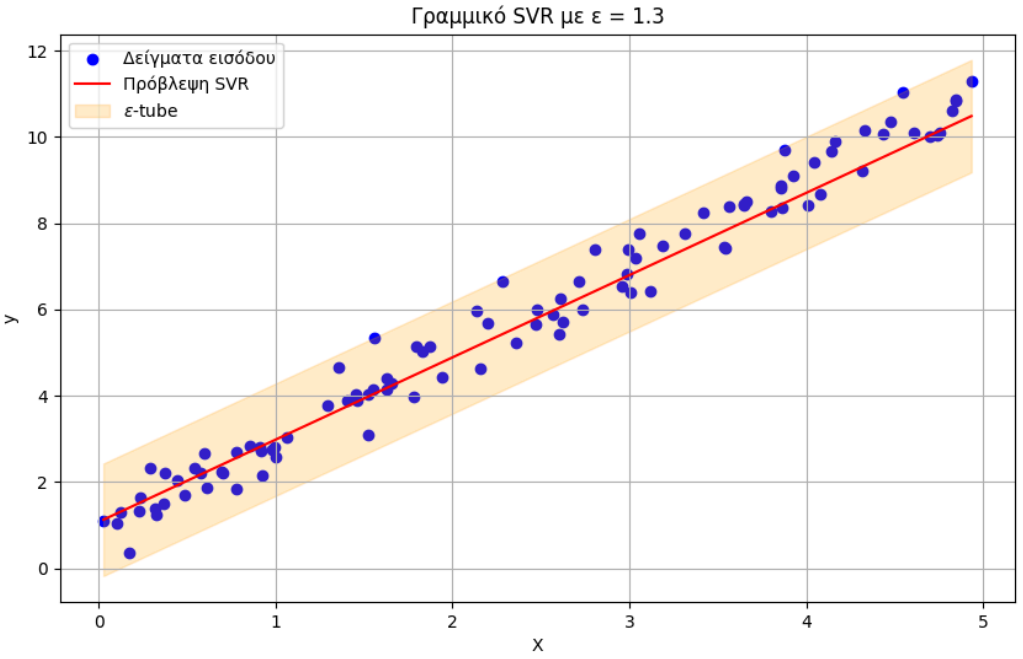
Ακολούθως, παρουσιάζεται η ανάλυση της διαδικασίας σχεδίασης του γραμμικού παλινδρομητή και του δικτύου διάκρισης. Σε αντίθεση με τα MLPs, όπου η σχεδίαση των δύο αυτών συνιστωσών πραγματοποιείται με συζευγμένο τρόπο μέσω μιας κοινής διαδικασίας βελτιστοποίησης, στα SVRs πραγματοποιείται ανεξάρτητα.

- *Σχεδίαση γραμμικού παλινδρομητή*

Η σχεδίαση του γραμμικού παλινδρομητή συνίσταται στη διατύπωση και επίλυση του αντίστοιχου προβλήματος μάθησης. Υποθέτουμε ότι έχουμε υλοποιήσει τη μη γραμμική απεικόνιση $h(\cdot)$ των δεδομένων σε έναν χώρο όπου η σχέση μεταξύ των μεταβλητών Y, $h(X)$ είναι "γραμμικά προσεγγίσιμη". Η γραμμική προσεγγισιμότητα διατυπώνεται μαθηματικά μέσω της έννοιας της ε-γραμμικότητας:

Ορισμός 1.11: Ένα σύνολο δεδομένων (x_i, y_i), $i = 1,\dots,N$, όπου $x_i \in \mathbb{R}^l$, $y \in \mathbb{R}$ ονομάζεται ε-γραμμικά προσεγγίσιμο, εφόσον $\exists \varepsilon > 0, b \in \mathbb{R}, w \in \mathbb{R}^l$, τέτοιο ώστε $\forall (x_i, y_i)$, $1 \le i \le N$ να ισχύει $|w_i^T x_i + b - y| < \varepsilon$.

Ουσιαστικά, τα μετασχηματισμένα δεδομένα θεωρούνται ε - γραμμικά, εφόσον υπάρχει γραμμικό μοντέλο παλινδρόμησης τέτοιο ώστε να πετυχαίνει απόλυτο σφάλμα πρόβλεψης το πολύ ίσο με ε. Προφανώς, όσο μικρότερη είναι η τιμή του ε για την οποία τα δεδομένα παραμένουν ε - γραμμικά, τόσο πιο ακριβείς θα είναι οι δυνητικές προβλέψεις ενός γραμμικού μοντέλου. Η ιδέα αυτή αποτυπώνεται γραφικά στο ακόλουθο μονοδιάστατο παράδειγμα:



Εικόνα 1.6: Γραφική απεικόνιση της ε-γραμμικότητας ενός συνόλου δεδομένων.

Προτού διατυπώσουμε το πρόβλημα μάθησης για ένα σύνολο μετασχηματισμένων δεδομένων που εικάζουμε πως είναι ε - γραμμικό, θα θεωρήσουμε τις ακόλουθες παρατηρήσεις: Σε πρακτικά προβλήματα, η ελάχιστη τιμή του ε δεν είναι εκ των προτέρων γνωστή. Επιπλέον, δεν μπορούμε να αποκλείσουμε την ύπαρξη ακραίων τιμών, οι οποίες θα επάγουν υψηλή τιμή του ε, χωρίς όμως αυτή να αντανακλά την πραγματική δομή των δεδομένων. Τέλος, αν υποθέσουμε πως υπάρχει ένα σύνολο γραμμικών μοντέλων για τον οποίο πληρείται ο Ορισμός 1.9, θα θέλαμε μεταξύ αυτών να επιλέξουμε το απλούστερο, για τους λόγους που αναλύθηκαν νωρίτερα στο κεφάλαιο.

Λαμβάνοντας υπόψη τα παραπάνω, θα διατυπώσουμε το πρόβλημα μάθησης ως το ακόλουθο πρόβλημα ελαχιστοποίησης με περιορισμούς:

$$min_{w,b,\xi,\xi^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (1.17)$$

$$s.t.: y_i - w^T h(x_i) - b \leq \varepsilon + \xi_i$$
$$w^T h(x_i) + b - y_i \leq \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0$$

Ο όρος $\frac{1}{2}||w||^2$ ποινικοποιεί την πολυπλοκότητα του μοντέλου σε επίπεδο παραμέτρων. Οι μεταβλητές ξ_i, ξ_i^* ονομάζονται μεταβλητές χαλάρωσης και επιτρέπουν στο μοντέλο να παράγει σφάλματα μεγαλύτερα του ε, για παράδειγμα σε περιπτώσεις ακραίων σημείων. Προφανώς, δεν θέλουμε τα σφάλματα αυτά να είναι μεγαλύτερα από ότι πραγματικά χρειάζεται, γι' αυτό και οι ξ_i εισάγονται στη συνάρτηση ελαχιστοποίησης. Οι C, ε αποτελούν υπερπαραμέτρους εκπαίδευσης.

Αποδεικνύεται εύκολα ότι το (1.17) είναι ένα αυστηρά κυρτό πρόβλημα και άρα έχει μοναδική λύση (w^*, b^*), η οποία θα προκύψει μέσω ενός *δυικού προβλήματος*. Η κατασκευή του δυικού προβλήματος και η μορφή του τελικού γραμμικού παλινδρομητή προκύπτουν μέσω των ακόλουθων ορισμών και προτάσεων [6]:

Ορισμός 1.13: Έστω ένα πρόβλημα ελαχιστοποίησης με m περιορισμούς, όπου η υποκείμενη συνάρτηση είναι η $J(w)$ και οι περιορισμοί είναι της μορφής $f_i(w) \geq 0,\ i = 1, 2, \ldots, m$. Η *συνάρτηση Lagrange* $\mathcal{L}(w, \lambda)$ για το πρόβλημα ορίζεται ως

$\mathcal{L}(w, \lambda) = J(w) - \sum_{i=1}^{m} \lambda_i f_i(w)$. Τα λ_i ονομάζονται *πολλαπλασιαστές Lagrange*.

Θεώρημα 1.4 (Karush - Kuhn - Tucker): Έστω ένα πρόβλημα ελαχιστοποίησης, όπως αυτό παρουσιάστηκε στον Ορισμό 1.13. Αναγκαία συνθήκη ώστε το w^* να είναι σημείο ελαχίστου είναι να ικανοποιεί τα ακόλουθα:

(1) $\frac{\partial L(w^*, \lambda)}{\partial \lambda} = 0$
(2) $\lambda_i \geq 0$
(3) $\lambda_i f_i(w^*) \geq 0$

Με απλή εφαρμογή του Ορισμού 1.13, προκύπτει ότι η συνάρτηση Lagrange για το (1.17) είναι η ακόλουθη:

$$\mathcal{L}(w, a, a^*, \eta, \eta^*) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n} a_i(\varepsilon + \xi_i - y_i + w^T h(x_i) + b) -$$

$$- \sum_{i=1}^{n} a_i^*(\varepsilon + \xi_i^* + y_i - w^T h(x_i) - b) - \sum_{i=1}^{n} \eta_i \xi_i - \sum_{i=1}^{n} \eta_i^* \xi_i^* \qquad (1.18)$$

Εφαρμόζοντας τις συνθήκες Karush-Kuhn-Tucker στο πρόβλημα (1.17), προκύπτει ότι $w^* = \sum_{i=1}^{n}(a_i - a_i^*)h(x_i)$. Ο υπολογισμός του b^* είναι κάπως πολυπλοκότερος οπότε δεν θα περιγραφεί αναλυτικά, αλλά, όπως και για το w^*, ισχύει ότι το b^* εξαρτάται αποκλειστικά από τους πολλ/στές Lagrange και τους πυρήνες $K(x, x_i)$. Συνεπώς, εάν προσδιορίσουμε το διάνυσμα των πολλ/στών Lagrange, μπορούμε να υπολογίσουμε και τα w^*, b^* και άρα τον ζητούμενο γραμμικό μετασχηματισμό

$$y = w^{*T} h(x) + b^* = \sum_{i=1}^{n}(a_i - a_i^*)K(x, x_i) + b \qquad (1.19)$$

Ο υπολογισμός των πολλαπλασιαστών πραγματοποιείται μέσω του δυϊκού προβλήματος

Θεώρημα 1.5 (Δυϊκή αναπαράσταση κατά Wolfe): Ένα πρόβλημα κυρτής βελτιστοποίησης είναι ισοδύναμο με το

$$max_{\lambda \geq 0} \mathcal{L}(w, \lambda)$$

$$s.t.: \frac{\partial \mathcal{L}(w,\lambda)}{\partial \lambda} = 0$$

Το δυικό πρόβλημα είναι επίσης κυρτό, με ισοτικούς περιορισμούς. Η επίλυσή του, επομένως, πραγματοποιείται με αλγορίθμους κυρτής βελτιστοποίησης όπως ο SMO (Sequential Minimal Optimization). Ο SMO αποτελεί τον αλγόριθμο μάθησης για ένα SVR και αναλύεται με λεπτομέρεια στο [7].

- *Σχεδίαση δικτύου μη γραμμικής διάκρισης χαρακτηριστικών*

Για τη σχεδίαση του δικτύου που υλοποιεί τον μη γραμμικό μετασχηματισμό h, τον οποίο έως τώρα είχαμε θεωρήσει ως δεδομένο, θα πρέπει να λάβουμε υπόψη την απαίτηση που είχαμε θέσει στην εισαγωγική παράγραφο των FFNNs: Ο h καλείται να μπορεί να απεικονίζει μη γραμμικές σε γραμμικές επιφάνειες. Ωστόσο, από τη μορφή της (1.19), μπορούμε να πραγματοποιήσουμε την εξής παρατήρηση: Εκπαιδεύοντας έναν γραμμικό μοντέλο με παραμέτρους (w, b) στον μετασχηματισμένο χώρο κατά h, ουσιαστικά εκπαιδεύουμε και ένα γραμμικό μοντέλο στον χώρο που προκύπτει από τον μετασχηματισμό $x_i \rightarrow K(x, x_i)$, $i = 1, 2, \ldots, n$. Τα δύο μοντέλα παράγουν ίδιες προβλέψεις

για ίδιες εισόδους. Επομένως, αν μια συνάρτηση είναι γραμμική στον έναν μετασχηματισμένο χώρο, τότε είναι και στον άλλο.

Επιλέγουμε, τώρα, να θέσουμε $K(x, x_i) = exp(-\gamma||x - x_i||^2)$. Ο πυρήνας αυτός ονομάζεται RBF και πληροί τα χαρακτηριστικά που αναφέραμε στην αρχή του κεφαλαίου για μετασχηματισμούς της μορφής $\varphi(sim(w_i, x))$. Εν προκειμένω, θα είναι:

$$sim(x_i, x) = ||x - x_i||^2 \quad (1.13)$$

$$\varphi(x) = exp(-\gamma x) \quad (1.14)$$

Τα πρότυπα διανύσματα προκύπτουν από τα δείγματα εκπαίδευσης x_i για τα οποία $(a_i - a_i^*) \neq 0$. Παρατηρούμε ότι το πλήθος των πρότυπων διανυσμάτων (και άρα των νευρώνων του δικτύου διάκρισης) δεν αποτελεί υπερπαράμετρο όπως στα MLPs, αλλά προκύπτει από τα δεδομένα. Στο πλαίσιο των SVRs, τα διανύσματα αυτά ονομάζονται *διανύσματα υποστήριξης*, εξου και η ονομασία της αρχιτεκτονικής. Επιπλέον, και σε εκ νέου αντίθεση με τα MLPs, τα κρυφά επίπεδα νευρώνων δεν είναι πολλαπλά, αλλά μόλις ένα.

Επισημαίνεται ότι η ανάλυση παρέχει μια γεωμετρική διαίσθηση για τους μετασχηματισμούς στο εσωτερικό των SVR, αλλά δεν είναι σε καμία περίπτωση θεωρητικά πλήρης. Επισταμένη θεώρηση των μετασχηματισμών σε χώρους πυρήνα πραγματοποιείται στο [8].

- *Υπερπαράμετροι του SVR*

Στον ακόλουθο πίνακα, παρουσιάζονται οι βασικές υπερπαράμετροι αρχιτεκτονικής και εκπαίδευσης ενός SVR.

Υπερπαράμετρος	**Επεξήγηση**	**Τυπικές Τιμές**
C	Συντελεστής ποινής. Ελέγχει τον συμβιβασμό μεταξύ απλότητας του μοντέλου και ανοχής στα σφάλματα. Υψηλές τιμές περιορίζουν τα λάθη αλλά μπορεί να οδηγήσουν σε υπερπροσαρμογή.	0.1, 1, 10, 100
γ	Παράμετρος του RBF πυρήνα. Ορίζει την «ακτίνα επιρροής» των δειγμάτων: μικρές τιμές οδηγούν σε πιο ομαλά μοντέλα, ενώ μεγάλες τιμές προκαλούν αυξημένη πολυπλοκότητα. Στον RBF kernel ισχύει: $K(x, x') = \exp(-\gamma\|x - x'\|^2)$.	`"scale"`, `"auto"`, 10^{-3}, 10^{-2}
ϵ	Περιθώριο ανοχής. Δηλώνει τη μέγιστη απόκλιση μεταξύ προβλεπόμενων και πραγματικών τιμών που δεν θεωρείται σφάλμα. Ελέγχει την ευαισθησία του μοντέλου.	0.01, 0.1, 0.2

***Πίνακας 1.2**: Πίνακας υπερπαραμέτρων του SVR*

II) *Μηχανές Gradient Boosting*

Η σχεδιαστική προσέγγιση που έχουμε υιοθετήσει έως τώρα για την επίλυση του προβλήματος της παλινδρόμησης αφορούσε στη δόμηση μιας αρχιτεκτονικής υψηλής χωρητικότητας (εν προκειμένω των FFNNs) και κατόπιν στη σχεδίασή της, μέσω ενός αλγορίθμου ελαχιστοποίησης στον χώρο παραμέτρων. Σε ένα τέτοιο πλαίσιο, οι προβλέψεις πραγματοποιούνται από ένα και μοναδικό, ισχυρό μοντέλο που να προσεγγίζει την $\varphi^* = \varphi(w^*)$ (βλ. 1.10 και 1.11).

Μια πιθανή εναλλακτική ιδέα αφορά στο κατά πόσο μπορούμε να κατασκευάσουμε έναν εύρωστο προβλέπτη όχι μέσω ενός μοναδικού ισχυρού μοντέλου, αλλά *συνδυάζοντας* τις επιμέρους προβλέψεις ενός *συνόλου* από *ασθενή* μοντέλα [9]. Μια τέτοια δομή είναι εν γένει πιο ευέλικτη, εφόσον τα ασθενή μοντέλα εκπαιδεύονται ταχύτερα και είναι λιγότερο ευάλωτα στην υπερπροσαρμογή. Τα ασθενή μοντέλα θα εκπαιδεύονται με τον συνήθη τρόπο, με βέλτιστη προσαρμογή, δηλαδή, στα δεδομένα εκπαίδευσης. Ένα πλαίσιο υλοποίησης της ιδέας του συνδυασμού μοντέλων είναι το ακόλουθο:

Έστω ένα σύνολο δεδομένων εκπαίδευσης $\{x_i, y_i\}_{i=1}^{N}$ και ένα μοντέλο F_0 ενός χώρου υποθέσεων H, χαμηλής χωρητικότητας. Το F_0 εκπαιδεύεται μέσω ενός αλγορίθμου μάθησης και κατόπιν παράγει ένα διάνυσμα προβλέψεων $\hat{y} = [F_0(x_1),..., F(x_N)]^T$. Θεωρώντας, τώρα, ότι μας ενδιαφέρει να ελαχιστοποιήσουμε τη συνάρτηση μέσου εμπειρικού σφάλματος $\frac{1}{N}\sum_{i=1}^{N}(L(y_i, F_0(x_i))$, ένας τρόπος να βελτιώσουμε την πρόβλεψη του F_0 είναι να τροποποιήσουμε το διάνυσμα προβλέψεων $[F_0(x_1),..., F(x_N)]^T$ προς μια κατεύθυνση κατα την οποία η L μειώνεται. Η τροποποίηση αυτή μπορεί να διατυπωθεί ως το βήμα ανανέωσης μιας μεθόδου καθόδου, όπως αυτή παρουσιάζεται στον Ορισμό 1.7:

$$F_1(X) = F_0(X) - \eta \cdot p_1 \tag{1.15}$$

Αν, λοιπόν, σε κάθε βήμα εκπαιδεύουμε ένα ασθενές μοντέλο f_k ώστε να παράγει μια εκτίμηση της επιλεχθείσας κατεύθυνσης καθόδου p_k, *τότε ο συνδυασμός των προβλέψεων θα υλοποιεί έναν αλγόριθμο καθόδου στον χώρο των διανυσμάτων πρόβλεψης*:

$$F_k(X) = F_{k-1}(X) - \eta \cdot f_k(X), \tag{1.16}$$

$$\text{οπότε } F_k(X) = F_0(X) - \eta \cdot \sum_{i=1}^{k} f_i(X)$$

Στην (1.16), με $F_k(X)$ συμβολίζουμε την πρόβλεψη του συνδυαστικού μοντέλου για είσοδο το τυχαίο διάνυσμα X και με $f_k(X)$ την πρόβλεψη του ασθενούς μοντέλου βάσης στο k-οστό βήμα. Κάθε επιπλέον βήμα αντιστοιχεί και σε ένα επιπλέον ασθενές μοντέλο

και κατ' επέκταση μια επιπλέον μετάβαση στον χώρο των διανυσμάτων πρόβλεψης. Η επιλογή της κατεύθυνσης καθόδου πραγματοποιείται βάσει των ακόλουθων προτάσεων:

Πρόταση 1.3: Έστω ότι Εσσιανός πίνακας $H_f(x)$ μιας διανυσματικής συνάρτησης $f: \mathbb{R}^n \to \mathbb{R}$ στο σημείο x είναι θετικά ορισμένος, ισχύει δηλαδή $x^T H_f x > 0,\ \forall x \in \mathbb{R}^n$. Τότε, το διάνυσμα $H_f(x)^{-1}\nabla f(x)$ θα αποτελεί κατεύθυνση καθόδου.

Πρόταση 1.4: Θεωρούμε διπλά παραγωγίσιμη διανυσματική συνάρτηση $f: \mathbb{R}^n \to \mathbb{R}$ και την δεύτερης τάξης προσέγγισή της κατά Taylor γύρω από το σημείο x:

$$f(x + a) = f(x) + \nabla f(x)^T a + \frac{1}{2} a^T H_f(x) a$$

Η διαφορά $f(x + a) - f(x)$ ελαχιστοποιείται για $h = - H_f(x)^{-1}\nabla f(x)$

Ο αλγόριθμος ελαχιστοποίησης για τον οποίο η $= - 1$ και $P_k = - H_f(x_{k-1})^{-1}\nabla f(x_{k-1})$ ονομάζεται *αλγόριθμος καθόδου Newton* (Newton Descent). Εν γένει, θεωρείται πιο εύρωστος από τον αλγόριθμο καθόδου κλίσης, καθότι, μέσω της προσέγγισης 2ης τάξης, ενσωματώνει πληροφορία και για την καμπυλότητα της συνάρτησης. Η αντιστροφή, ωστόσο, του Εσσιανού πίνακα επάγει υψηλό υπολογιστικό κόστος και γι' αυτό θα θεωρήσουμε τη διαγώνια προσέγγισή του:

$$H_f(x_{k-1}) \simeq diag(h_1(x_{k-1}),..., h_n(x_{k-1})) \qquad (1.17)$$

Για να είναι έγκυρη η παραπάνω προσέγγιση, θα πρέπει τα μη διαγώνια στοιχεία του $H_f(x)$ να τείνουν στο 0. Στην περίπτωσή μας, θεωρώντας $f = \bar{L} = \frac{1}{N}\sum_{i=1}^{N} (L(y_i, F_k(x_i))$ και ως διάνυσμα ανεξάρτητων μεταβλητών το διάνυσμα προβλέψεων $[F_k(x_1),..., F_k(x_N)]^T$, η υπόθεση της διαγωνιότητας H απαιτεί οι απώλειες $L(y_i, F_k(x_i))$, $L(y_j, F_k(x_j))$ που επάγουν οι προβλέψεις $F_k(x_i), F_k(x_j)$ για δύο δείγματα x_i, x_j να είναι μεταξύ τους ανεξάρτητες. Αυτό προφανώς ισχύει, εφόσον η ίδια η συνάρτηση μέσου εμπειρικού σφάλματος συνίσταται σε ένα άθροισμα N ανεξάρτητων όρων. Συνεπώς, για το βήμα καθόδου της μεθόδου Newton, έχουμε ότι:

$$F_k(x_i) = F_{k-1}(x_i) - \frac{g_i(k)}{h_i(k)},\ i = 1, 2,..., N \qquad (1.18)$$

Με g_i, h_i συμβολίζουμε τις ποσότητες $\frac{\partial \bar{L}}{\partial F_{k-1}(x_i)}$ και $\frac{\partial^2 \bar{L}}{\partial F_{k-1}(x_i)^2}$ αντίστοιχα. Παρατηρώντας, επιπλέον, την Πρόταση 1.4, διαπιστώνουμε ότι η κατεύθυνση Newton μπορεί να υπολογιστεί επιλύοντας το πρόβλημα ελαχιστοποίησης της διαφοράς

$f(x + a) - f(x)$, για τετραγωνική προσέγγιση της f. Για διαγώνιο Εσσιανό πίνακα, η λύση στο εν λόγω πρόβλημα προκύπτει ως εξής:

$$a^* = argmin_{a \in \mathbb{R}^n} \sum_{i=1}^{n} g_i a_i + \frac{1}{2} a_i^2 h_i \tag{1.19}$$

Εξειδικεύοντας την (1.19) στην περίπτωση των Μηχανών Boosting, παίρνουμε ότι:

$$f_k^* = argmin_{f_k \in H(\theta)} \sum_{i=1}^{n} g_i f_k(x_i) + \frac{1}{2} f_k^2(x_i) h_i \tag{1.20}$$

Με $H(\theta)$ συμβολίζουμε τον χώρο υποθέσεων του ασθενούς μοντέλου βάσης. Με επισκόπηση της (1.16), συμπεραίνουμε ότι για να προσδιορίσουμε την $F_k(x)$ σε οποιοδήποτε βήμα k, αρκεί να υπολογίσουμε τις g_i, h_i για κάθε δείγμα εκπαίδευσης και κατόπιν να εκπαιδεύσουμε ένα μοντέλο f_k^* ώστε να ελαχιστοποιεί την έκφραση της (1.20). Αντικαθιστώντας στην (1.16), έχουμε την τιμή της $F_k(x)$. Η διαδικασία αυτή περιγράφεται φορμαλιστικά στον παρακάτω αλγόριθμο (*Extreme Gradient Boosting - XGB* [10]):

Algorithm 2 Αλγόριθμος εκπαίδευσης XGBoost

1: **Είσοδος:** δεδομένα εκπαίδευσης $D = \{(x_i, y_i)\}_{i=1}^N$, συνάρτηση απώλειας L, πλήθος ασθενών μοντέλων M, ρυθμός μάθησης η
2: **Αρχικοποίηση:** $f_0(x) = F_0(x) = \arg\min_\theta \sum_i L(y_i, \theta)$
3: **for** $k = 1$ έως M **do**
4: **for** κάθε δείγμα $i = 1$ έως N **do**
5: Υπολογισμός gradient και hessian:

$$g_i = \left.\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right|_{F=F_{k-1}}, \quad h_i = \left.\frac{\partial^2 L(y_i, F(x_i))}{\partial F(x_i)^2}\right|_{F=F_{k-1}}$$

6: **end for**
7: Εκπαίδευση ασθενούς μοντέλου βάσης $f_k(x)$ με στόχο την ελαχιστοποίηση της:

$$\mathcal{L}^{(k)} = \sum_{i=1}^{N} \left[g_i f_k(x_i) + \frac{1}{2} h_i f_k(x_i)^2 \right] + \Omega(f_k)$$

8: Ενημέρωση: $F_k(x) = F_{k-1}(x) + \eta \cdot f_k(x)$
9: **end for**
10: **Έξοδος:** τελική πρόβλεψη $F_M(x)$

Έχοντας ορίσει σε υψηλό επίπεδο τον αλγόριθμο κατασκευής ενός XGB παλινδρομητή, η κύρια σχεδιαστική απόφαση που εκκρεμεί, πέραν της επιλογής των υπερπαραμέτρων, αφορά την επιλογή της αρχιτεκτονικής των ασθενών μοντέλων. Εδώ, η πλέον δημοφιλής αρχιτεκτονική είναι αυτή των *δυαδικών δένδρων παλινδρόμησης (binary regression trees)*.

- *Δυαδικά δένδρα παλινδρόμησης*

Ένα δένδρο παλινδρόμησης αποτελεί ένα αλγοριθμικό μοντέλο πεπερασμένων καταστάσεων, το οποίο χαρακτηρίζεται από:

- ένα σύνολο διακριτών καταστάσεων $Q = \{q_1,..., q_m\}$, οι οποίες διακρίνονται σε ενδιάμεσες και τερματικές καταστάσεις - *φύλλα*.
- μια αρχική κατάσταση $q^{(0)}$
- μια συνάρτηση μετάβασης $g(q_i, x) = q_j$. Όταν η q_j είναι μια κατάσταση φύλλο, το μοντέλο παράγει ως έξοδο την τιμή που αντιστοιχεί στο φύλλο. Η απεικόνιση g έχει την ακόλουθη μορφή:

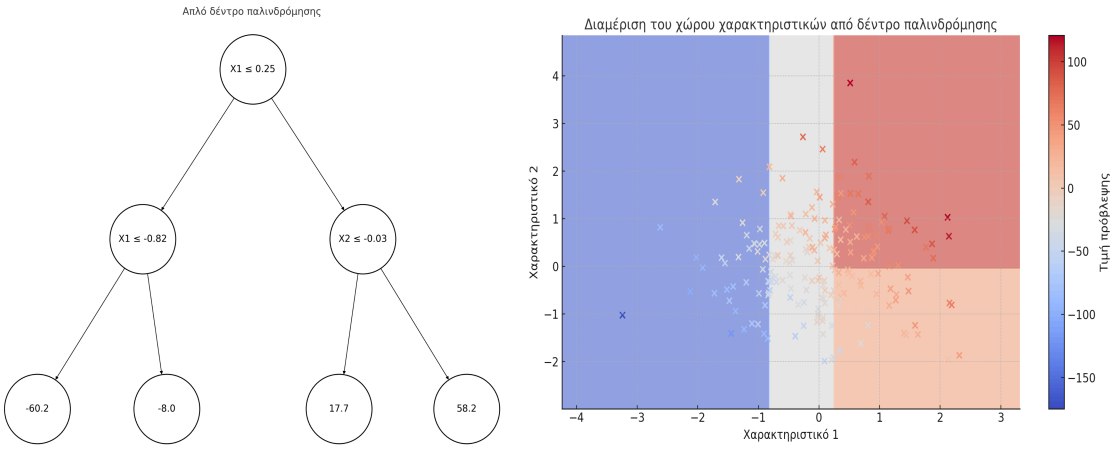
$$g(q_i, x) = q_j,\ x_l > a \qquad (1.21)$$
$$g(q_i, x) = q_m,\ x_l < a$$

Το a ονομάζεται *τιμή διαχωρισμού (split value)*. Παρατηρούμε ότι η μετάβαση στην επόμενη κατάσταση δεν εξαρτάται συνολικά από το διάνυσμα εισόδου x, αλλά από κάποιο χαρακτηριστικό x_l, το οποίο συναρτάται με την τρέχουσα κατάσταση q_i. Η τιμή του a, το ποιό χαρακτηριστικό x_l εξετάζεται σε κάθε κατάσταση, καθώς και το αν ένας κόμβος θα είναι ενδιάμεσος ή φύλλο είναι παράγοντες που προσδιορίζονται κατά την εκπαίδευση του δένδρου.

Ένα διάνυσμα εισόδου $x \in \mathbb{R}^d$ διεγείρει μια *μοναδική* και *πεπερασμένη* ακολουθία μεταβάσεων $(q^{(0)}, q^{(1)},..., q^{(k)})$ η οποία οδηγεί το μοντέλο σε μια κατάσταση - φύλλο και άρα στο να πραγματοποιήσει μια πρόβλεψη. Αν υποθέσουμε ότι τα φύλλα είναι J το πλήθος, τότε είναι προφανές από τη (1.21) ότι ένα φύλλο l_i θα αντιστοιχεί σε μια παραλληλεπίπεδη περιοχή του $\mathbb{R}^d$, η οποία θα απεικονίζεται σε μια τιμή $w_i \in \mathbb{R}$. Ένα δένδρο παλινδρόμησης, λοιπόν, απεικονίζει τον $\mathbb{R}^d$ διαχωρίζοντάς τον σε J παραλληλεπίπεδα. Μέσω της διαδοχικής εναλλαγής καταστάσεων, καταλήγει σε μια κατάσταση - φύλλο, προσδιορίζοντας, έτσι, σε ποια εκ των J περιοχών ανήκει το διάνυσμα εισόδου $x \in \mathbb{R}^d$.

Το δένδρο παλινδρόμησης περιγράφεται πλήρως από τον *γράφο μεταβάσεών* του, G_{tr}. Ο G_{tr} αποτυπώνει γραφικά όλες τις πιθανές ακολουθίες καταστάσεων του μοντέλου μέχρι να οδηγηθεί σε κάποια κατάσταση - φύλλο.

Παρακάτω, παρουσιάζεται ο γράφος μεταβάσεων και η διάκριση του χώρου χαρακτηριστικών $\mathbb{R}^2$ που επιτυγχάνει ένα απλό δυαδικό δένδρο παλινδρόμησης.



***Εικόνα 1.7**: Γράφος μεταβάσεων και διαμέριση του χώρου χαρακτηριστικών από δέντρο παλινδρόμησης*

Στην περίπτωση των μηχανών XGB, ένα δένδρο παλινδρόμησης εκπαιδεύεται ώστε να ελαχιστοποιεί την ομαλοποιημένη εκδοχή της (1.20). Συγκεκριμένα, είναι:

$$f_k^* = argmin_{f_k \in H(\theta)} \sum_{i=1}^{n} [g_i f_k(x_i) + \frac{1}{2} f_k^2(x_i) h_i] + \Omega(f_k) \qquad (1.22)$$

Ο όρος ομαλοποίησης $\Omega(f_k)$ ποινικοποιεί αφενός το πλήθος J των φύλλων του δένδρου, και κατ΄επέκταση την πολυπλοκότητα της διαμέρισης του χώρου εισόδου, και αφετέρου τις υψηλές τιμές εξόδου w_i στα φύλλα. Αναλυτικά, προκύπτει μέσω της ακόλουθης σχέσης:

$$\Omega(f_k) = \gamma J + \frac{1}{2} \lambda \sum_{i=1}^{J} w_j^2 \qquad (1.23)$$

Η διαδικασία κατασκευής ενός δένδρου παλινδρόμησης περιγράφεται τυπικά στον αλγόριθμο που ακολουθεί παρακάτω. Διαισθητικά, η ιδέα είναι ότι το δένδρο κατασκευάζεται δυναμικά, κόμβο προς κόμβο. Για κάθε κόμβο, ο οποίος αντιστοιχεί μοναδικά σε μια κατάσταση q_i, εξετάζουμε όλα τα πιθανά χαρακτηριστικά και όλες τις πιθανές τιμές διαχωρισμού βάσει του εκάστοτε χαρακτηριστικού. Κάθε χαρακτηριστικό και τιμή διαχωρισμού επάγουν έναν διαφορετικό διαχωρισμό του κόμβου σε δύο παιδιά, ο οποίος χαρακτηρίζεται από ένα *κέρδος διαχωρισμού (split gain)* G_{split}. Το κέρδος διαχωρισμού εκφράζει τη μείωση της συνάρτησης απώλειας

$\sum_{i=1}^{n}[g_i f_k(x_i) + \frac{1}{2} f_k^2(x_i) h_i] + \Omega(f_k)$ που επιτυγχάνεται από τον διαχωρισμό του κόμβου. Εάν το μέγιστο G_{split} είναι θετικό, τότε ο διαχωρισμός πράγματι έχει νόημα και ο κόμβος διασπάται σε δύο παιδιά. Εάν όχι, τότε ο κόμβος καθίσταται φύλλο, με τιμή εξόδου

$$w^* = argmin_w \sum_{i=1}^{I}[g_i w + \frac{1}{2} w^2 h_i] + \gamma J + \frac{1}{2}\lambda \sum_{i=1}^{J} w_j^2 = -\frac{\sum_{i \in I} g_i}{\sum_{i \in I} h_i + \lambda} \quad (1.24)$$

Algorithm 3 Αλγόριθμος κατασκευής ενός δένδρου παλινδρόμησης

Είσοδος: gradients g_i, hessians h_i, παράμετροι ομαλοποίησης: γ, λ, ελάχιστη τιμή hessian ανά φύλλο $h_{\min}$
Αρχικοποίηση: Δημιούργησε ρίζα με όλα τα δείγματα
while υπάρχουν κόμβοι προς διαίρεση **do**
  for κάθε κόμβο προς διαίρεση **do**
    for κάθε πιθανό χαρακτηριστικό x_j και split value s **do**
      Χώρισε τα δείγματα σε δύο ομάδες: $I_L = \{i : x_{ij} \leq s\}$, $I_R = \{i : x_{ij} > s\}$
      if $\sum_{i \in I_L} h_i \geq h_{\min}$ και $\sum_{i \in I_R} h_i \geq h_{\min}$ **then**
        Υπολόγισε το Gain της διάσπασης:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma$$

      end if
    end for
    Διάλεξε τη διάσπαση με το μέγιστο Gain
    if μέγιστο Gain > 0 **then**
      Διαίρεσε τον κόμβο σε δύο κόμβους - παιδιά
    else
      Σταμάτα τη διαίρεση (δημιουργία φύλλου)
      Υπολόγισε την τελική τιμή πρόβλεψης του φύλλου:

$$w^* = -\frac{\sum_{i \in I} g_i}{\sum_{i \in I} h_i + \lambda}$$

    end if
  end for
end while
Έξοδος: regression tree με φύλλα που περιέχουν τις τελικές τιμές πρόβλεψης w^*

- *Υπερπαράμετροι του XGB*

Στον πίνακα που ακολουθεί, παρουσιάζονται οι βασικές υπερπαράμετροι εκπαίδευσης και αρχιτεκτονικής ενός XGB, στις οποίες θεωρούμε ότι συμπεριλαμβάνονται και οι αντίστοιχες υπερπαράμετροι των μοντέλων βάσης - εν προκειμένω των δέντρων παλινδρόμησης.

Υπερπαράμετρος	Επεξήγηση	Τυπικές Τιμές
`n_estimators`	Πλήθος ασθενών μοντέλων (δέντρων) που εκπαιδεύονται διαδοχικά. Επηρεάζει τη συνολική ισχύ του συνδυαστικού μοντέλου.	100–500
`learning_rate`	Ρυθμός μάθησης (η). Ελέγχει τη συνεισφορά κάθε νέου δέντρου στη συνολική πρόβλεψη. Μικρότερες τιμές απαιτούν περισσότερα δέντρα.	0.01–0.2
`max_depth`	Μέγιστο βάθος κάθε δέντρου. Ελέγχει την πολυπλοκότητα του δέντρου με στόχο την αποφυγή υπερπροσαρμογής.	3–8
`subsample`	Ποσοστό δειγμάτων που χρησιμοποιούνται για την εκπαίδευση κάθε δέντρου. Συμβάλλει στην αποφυγή της υπερπροσαρμογής σε συγκεκριμένο σύνολο δεδομένων.	0.5–1.0
`colsample_bytree`	Ποσοστό χαρακτηριστικών που χρησιμοποιούνται σε κάθε δέντρο. Στοχεύει στη διαφοροποίηση των δέντρων.	0.5–1.0
`gamma`	Ελάχιστο απαραίτητο μη κανονικοποιημένο split gain (μη κανονικοποιημένο) για να επιτραπεί ο διαχωρισμός ενός κόμβου	0–5
`reg_alpha`	Συντελεστής L1 κανονικοποίησης στις τιμές των φύλλων. Προστίθεται προαιρετικά στον όρο ομαλοποίησης και ενθαρρύνει την αραιότητα (sparsity) του w.	0–10
`reg_lambda`	Συντελεστής L2 κανονικοποίησης. Συμβάλλει στη σταθεροποίηση των τιμών πρόβλεψης.	1–10

***Πίνακας 1.3**: Πίνακας υπερπαραμέτρων του XGB*

- ***Δυναμικά προβλήματα***

Στην περίπτωση των δυναμικών προβλημάτων, οι αρχιτεκτονικές μοντέλων οι οποίες θα εξεταστούν πηγάζουν από το βασικό υπόδειγμα των *αναδρομικών νευρωνικών δικτύων (recurrent neural networks)*.

I) *Απλά αναδρομικά νευρωνικά δίκτυα (Vanilla Recurrent Neural Networks - RNNs)*

Η αρχιτεκτονική των αναδρομικών δικτύων πηγάζει κατ' αρχήν από τη θεμελιώδη υπόθεση ότι τα ακολουθιακά δεδομένα στη φύση παράγονται και επεξεργάζονται από *δυναμικά συστήματα*. Το βασικό χαρακτηριστικό των δυναμικών συστημάτων είναι ότι εξελίσσονται στον χρόνο μέσω της επαναληπτικής εφαρμογής ενός *τοπικού υπολογιστικού μοντέλου*. Το τοπικό μοντέλο αυτό εμπλέκει το *διάνυσμα κατάστασης* (ή απλά κατάσταση) του συστήματος και τις χρονοσειρές εισόδου και εξόδου.

Για ένα σύστημα που επεξεργάζεται, εξάγει πληροφορία και βάσει αυτής απεικονίζει την ακολουθία εισόδου, η κατάσταση h_k θεωρούμε ότι αποτυπώνει την πληροφορία (state of information) που έχει εξάγει το σύστημα για την είσοδο έως και τη χρονική στιγμή k. Από τη διαισθητική αυτή περιγραφή, επάγεται άμεσα ότι η μετάβαση του συστήματος στην επόμενη κατάσταση h_{k+1} (και άρα στο επόμενο state of information) αρκεί να λαμβάνει υπόψη την τρέχουσα κατάσταση h_k (τρέχον state of information) και την είσοδο x_k. Αντίστοιχα, είναι εύλογο να θεωρήσουμε ότι η έξοδος y_k προκύπτει ως συνάρτηση της

εισόδου x_k και της πληροφορίας που έχουμε για την είσοδο έως τη χρονική στιγμή k. Από τα παραπάνω, προκύπτει ο ακόλουθος τυπικός ορισμός του δυναμικού συστήματος:

Ορισμός 1.14: Ένα δυναμικό σύστημα (διακριτού χρόνου) ορίζεται μέσω των δύο ακόλουθων μετασχηματισμών

$$h_{k+1} = f(h_k, x_k) \qquad \text{(Εξίσωση κατάστασης)}$$
$$y_k = g(x_k, u_k) \qquad \text{(Εξίσωση εξόδου)}$$

Έστω, τώρα, ότι μας ενδιαφέρει να σχεδιάσουμε ένα δυναμικό σύστημα για ένα πρόβλημα παλινδρόμησης, όπου στόχος είναι η εκτίμηση μια μεταβλητής εξόδου Y βάσει μιας χρονοσειράς εισόδου $(X_1,..., X_d)$. Στην περίπτωση αυτή, καλούμαστε να προσεγγίσουμε στατιστικά μια συνάρτηση της μορφής $Y = f(X_1,..., X_d)$. Όμως, λαμβάνοντας υπόψη πως ένα δυναμικό σύστημα κωδικοποιεί την πληροφορία της εισόδου $(X_1,..., X_d)$ στην κατάσταση H_{d+1}, μπορούμε να προσεγγίσουμε τη μεταβλητή Y ως $Y = \hat{f}(H_{d+1})$. Η παράμετρος $d \in \mathbb{N}$ ονομάζεται *μήκος συμφραζομένου* (context length) και προσδιορίζεται βάσει της συνάρτησης αυτοσυσχέτισης της χρονοσειράς, όπως αναλύθηκε στην Παράγραφο 1.3.2.

Βάσει, λοιπόν, των παραπάνω, το σύστημα που καλούμαστε να σχεδιάσουμε είναι το ακόλουθο:

$$h_{k+1} = f(h_k, x_k; w_f, \theta_f),\ k = 0, 1,..., d \qquad (1.25)$$
$$y = g(h_d; w_g, \theta_g)$$

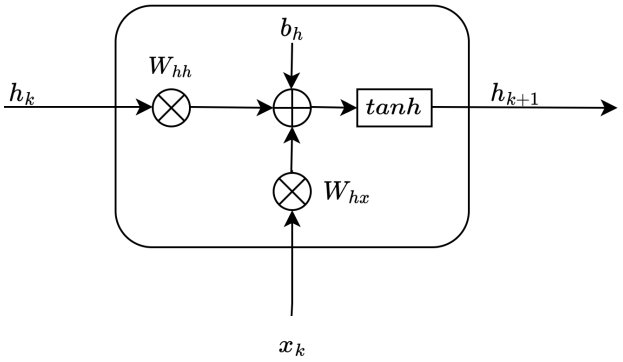
Με w_f, θ_f και w_g, θ_g συμβολίζουμε τα ζεύγη παραμέτρων, υπερπαραμέτρων των f και g αντίστοιχα. Οι f και g, οι οποίες συγκροτούν το τοπικό υπολογιστικό μοντέλο του συστήματος, μπορούν να υλοποιηθούν μέσω μιας στατικής αρχιτεκτονικής και συγκεκριμένα ενός MLP μοναδικού κρυφού επιπέδου. Κατ' αυτόν τον τρόπο, διαμορφώνεται το μοντέλο ενός αναδρομικού νευρωνικού δικτύου, το οποίο συνίσταται στο ακόλουθο ζεύγος εξισώσεων:

$$h_{k+1} = tanh(\left[W_{hh}\ W_{hx}\right][h_k\ x_k]^T + b_h),\ k = 0, 1,..., d \qquad (1.26)$$
$$y = W_{yh} h_d + b_y$$

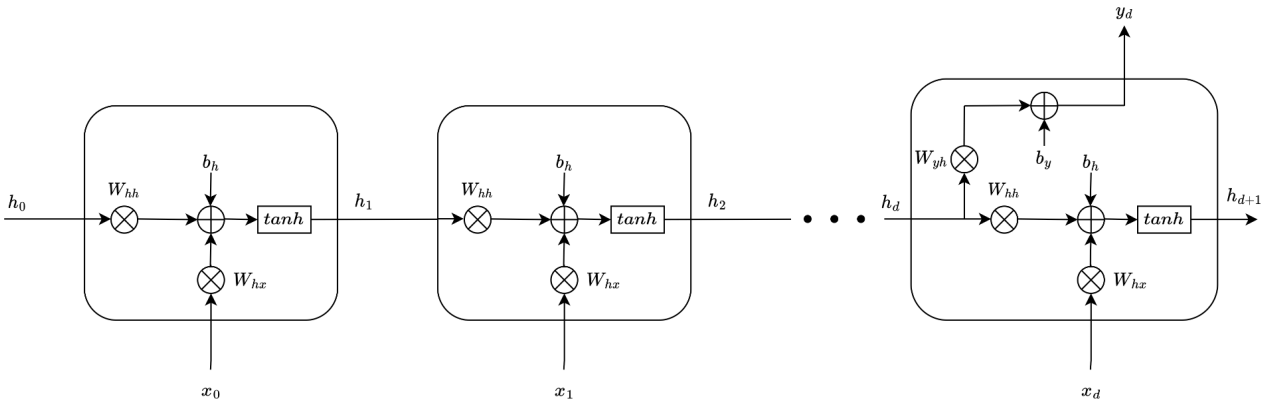
Παρατηρούμε ότι, εφόσον η έξοδος παράγεται ανά d χρονικά βήματα, στα ενδιάμεσα βήματα χρησιμοποιείται αποκλειστικά το δίκτυο διάκρισης του MLP, το οποίο και υλοποιεί την αναδρομική ανανέωση της κατάστασης. Το γραμμικό στάδιο εξόδου ενσωματώνεται στο d-οστό βήμα, όπου πραγματοποιείται και η πρόβλεψη. Φυσικά, εάν θέλαμε η πρόβλεψη

να πραγματοποιείται σε κάθε χρονικό βήμα (π.χ. σε προβλήματα απεικόνισης χρονοσειράς σε χρονοσειρά), τότε θα ίσχυε $y_k = W_{yh} h_k + b_y$, οπότε το στάδιο εξόδου θα ενσωματωνόταν σε κάθε βήμα. Επισημαίνεται, επιπλέον, ότι αντίθεση με τον ορισμό, ως συνάρτηση ενεργοποίησης χρησιμοποιείται η $tanh$ και όχι η $ReLU$.

Παρακάτω, απεικονίζεται το τοπικό υπολογιστικό μοντέλο ή *κύτταρο* (cell) ενός RNN για παλινδρόμηση, καθώς και η ακολουθία των μετασχηματισμών που εκτελεί το σύστημα έως ότου πραγματοποιήσει την τελική του πρόβλεψη.



***Εικόνα 1.8**: Κύτταρο αναδρομικού νευρωνικού δικτύου*



***Εικόνα 1.9**: Ακολουθία μετασχηματισμών μέχρι το βήμα πρόβλεψης (unfolded RNN)*

Προκειμένου να ενισχύσουμε τη χωρητικότητα ενός αναδρομικού δικτύου, έχουμε τη δυνατότητα να "στιβάξουμε" (stack) δύο επίπεδα RNN, με τη χρονοσειρά καταστάσεων του πρώτου επιπέδου να αποτελεί τη χρονοσειρά εισόδου για το δεύτερο. Η τοπολογία αυτή

καλείται *Stacked RNN* και επεκτείνεται με όμοιο τρόπο στις διάφορες παραλλαγές αναδρομικών δικτύων που θα εξετάσουμε στη συνέχεια της ενότητας.

- *Επιλογή συνάρτησης απωλειών και αλγορίθμου μάθησης*

Όπως και στην περίπτωση των στατικών MLPs, το πρόβλημα μάθησης διατυπώνεται ως πρόβλημα ελαχιστοποίησης του εμπειρικού σφάλματος (βλ. 1.12) και ως αλγόριθμος ελαχιστοποίησης επιλέγεται ο Adam. Αναφορικά όμως με τον έλεγχο της πολυπλοκότητας του μοντέλου, επιλέγουμε να μην εφαρμόσουμε $L2$ ομαλοποίηση, αλλά την τεχνική *Dropout* [11], η οποία είναι αρκετά δημοφιλής στα νευρωνικά δίκτυα με μεγάλο βάθος.

- *Έλεγχος πολυπλοκότητας του μοντέλου μέσω Dropout*

Έστω ότι ένα αναδρομικό δίκτυο έχει υπερπροσαρμοστεί στα δεδομένα εκπαίδευσης. Αυτό συνεπάγεται ότι τα πρότυπα διανύσματα στα εσωτερικά επίπεδα του δικτύου έχουν διαμορφωθεί από κοινού, με τρόπο τέτοιο ώστε να διασφαλίζουν μεν υψηλή επίδοση στο σύνολο εκπαίδευσης, παρουσιάζοντας όμως υψηλή αστάθεια σε άγνωστα σύνολα δεδομένων. Προκειμένου να αποτρέψουμε αυτήν τη συμπεριφορά, για κάθε δείγμα εκπαίδευσης του δικτύου απενεργοποιούμε με τυχαίο τρόπο ένα ποσοστό των νευρώνων των εσωτερικών επιπέδων. Ως αποτέλεσμα, το δίκτυο εκπαιδεύεται ώστε να επιτυγχάνει χαμηλά σφάλματα εκπαίδευσης για διάφορους συνδυασμούς νευρώνων, χωρίς να βασίζεται σε μια συγκεκριμένη, ασταθή διαμόρφωση των νευρωνικών παραμέτρων, όπως στην περίπτωση της υπερπροσαρμογής. Μαθηματικά, ο μηχανισμός του dropout αποτυπώνεται ως ακολούθως:

Έστω $h_{k+1} = \left[W_{hh}\, W_{hx}\right][h_k\, x_k]^T + b_h$ η έξοδος ενός εσωτερικού επιπέδου. Θεωρούμε το τυχαίο διάνυσμα $r = [r_1,\ldots, r_m]$, όπου r_i ανεξάρτητες και ισόνομες Bernoulli τυχαίες μεταβλητές με παράμετρο $1 - p$ και m η διάσταση του χώρου κατάστασης. Η παράμετρος p αναφέρεται και ως *dropout rate.* Το διάνυσμα r αποτελεί μια μάσκα, η οποία διαφοροποιείται για κάθε δείγμα εκπαίδευσης και εφαρμόζεται στο h_{k+1} ως εξής

$$\overline{h}_{k+1} = r \odot h_{k+1} \tag{1.27}$$

Με $\odot$ συμβολίζουμε το γινόμενο Hadamard μεταξύ δύο διανυσμάτων. Επισημαίνεται ότι η μάσκα r παραμένει αναλλοίωτη σε όλα τα χρονικά βήματα επεξεργασίας της εισόδου x.

- *Υπερπαράμετροι RNN*

Υπερπαράμετρος	Επεξήγηση	Τυπικές Τιμές
Διάσταση 1ου κρυφού επιπέδου	Πλήθος μονάδων στο πρώτο επίπεδο του RNN. Επηρεάζει άμεσα την εκφραστική ικανότητα του δικτύου.	32, 64, 128
Διάσταση 2ου κρυφού επιπέδου	Πλήθος μονάδων στο δεύτερο (προαιρετικό) RNN επίπεδο, αν χρησιμοποιείται stacking.	16, 32, 64
Learning rate	Ρυθμός μάθησης του αλγορίθμου βελτιστοποίησης. Μικρές τιμές οδηγούν σε σταθερότερη, αλλά αργή σύγκλιση.	10^{-4}, 10^{-3}, 10^{-2}
Dropout rate	Ποσοστό τυχαίας απενεργοποίησης μονάδων κατά την εκπαίδευση για αποφυγή υπερπροσαρμογής.	0.1, 0.2, 0.3, 0.5

***Πίνακας 1.4**: Πίνακας υπερπαραμέτρων του RNN*

II) Δίκτυα Long Short Term Memory (LSTM Networks)

Η απλή εκδοχή των αναδρομικών νευρωνικών δικτύων χαρακτηρίζεται από έναν θεμελιώδη θεωρητικό περιορισμό, ο οποίος είναι δυνατό κατά περίπτωση να υπονομεύει τη δυνατότητα μάθησης του δικτύου. Ο περιορισμός αυτός αναφέρεται στην τάση των παραγώγων $\frac{\partial L}{\partial h_k}$, όπου h_k η κατάσταση σε κάποια παρελθούσα χρονική στιγμή k, να παρουσιάζουν εκθετική αύξηση ή μείωση προς το 0 , καθώς το k τείνει προς το 1 (αρχική χρονική στιγμή). Αν υποθέσουμε ότι εφαρμόζουμε το αλγόριθμο μάθησης Adam και με δεδομένο ότι η παράγωγος της συνάρτησης απωλειών ως προς τον πίνακα W_{hh} προκύπτει ως $\frac{\partial L}{\partial W_{hh}} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_k} \cdot \frac{\partial h_k}{\partial W_{hh}}$, διαπιστώνουμε ότι η συμπεριφορά αυτή είναι ιδιαίτερα δυσμενής, καθότι μπορεί να οδηγήσει είτε σε αστάθεια του αλγορίθμου μάθησης, είτε σε μηδενική συνεισφορά των παλαιότερων δειγμάτων της χρονοσειράς στον υπολογισμό της $\frac{\partial L}{\partial W_{hh}}$ και άρα συνολικά στη διαδικασία της μάθησης. Η συμπεριφορά των $\frac{\partial L}{\partial h_k}$ εξαρτάται από τις ιδιοτιμές του πίνακα W_{hh}. Συγκεκριμένα, εαν $|\lambda_{max}| > 1$, τότε για $T - k \to \infty$ ισχύει:

$$||\frac{\partial L}{\partial h_k}|| \to \infty \tag{1.28}$$

Ενώ, για $|\lambda_{max}| < 1$ και $T - k \to \infty$, θα είναι:

$$||\frac{\partial L}{\partial h_k}|| \to 0 \tag{1.29}$$

Αναλυτική παρουσίαση του φαινομένου της *εξασθένησης και εκρηκτικής αύξησης των παραγώγων (vanishing and exploding gradients)* πραγματοποιείται εδώ [12]

Προκειμένου να αντιμετωπιστούν οι δυσκολίες που παρουσιάζονται στην εκπαίδευση των κλασικών αναδρομικών δικτύων, στην πράξη χρησιμοποιείται συχνά μια εναλλακτική αρχιτεκτονική, η οποία, διαθέτοντες περισσότερες εσωτερικές μεταβλητές πέραν της κατάστασης h, επιτρέπει την αδιάλειπτη ροή των παραγώγων προς παλαιότερα χρονικά βήματα. Η αρχιτεκτονική αυτή καλείται LSTM [13], αποτελεί επίσης ένα αναδρομικό δίκτυο και βασίζεται στην επαναληπτική εφαρμογή του παρακάτω τοπικού μοντέλου:

$$f_k = \sigma(W_f x_k + U_f h_{k-1} + b_f) \quad \text{(forget gate)} \quad (1.30)$$

$$i_k = \sigma(W_i x_k + U_i h_{k-1} + b_i) \quad \text{(input gate)} \quad (1.31)$$

$$o_k = \sigma(W_o x_k + U_o h_{k-1} + b_o) \quad \text{(output gate)} \quad (1.32)$$

$$\bar{c}_k = tanh(W_c x_k + U_c h_{k-1} + b_c) \quad \text{(candidate state)} \quad (1.33)$$

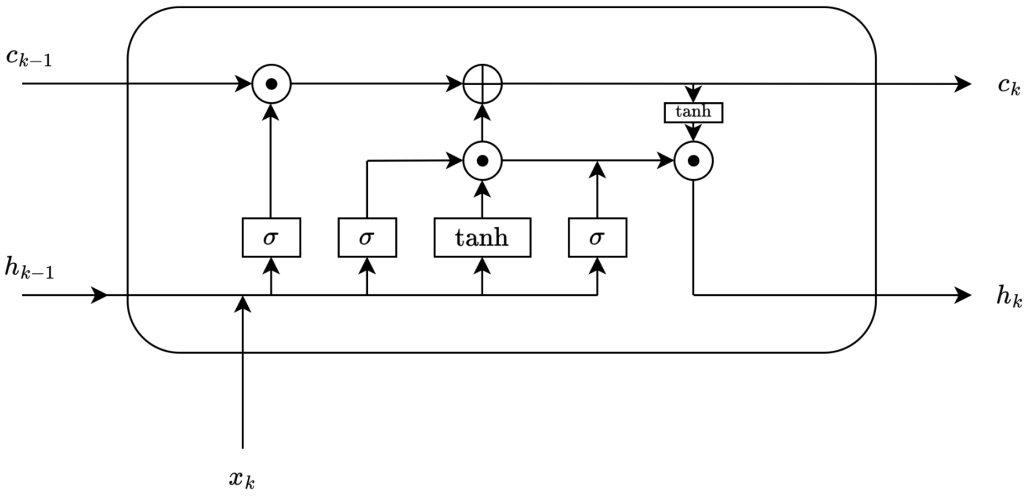
$$c_k = f_k \odot c_{k-1} + i_k \odot \bar{c}_k \quad \text{(cell state update)} \quad (1.34)$$

$$h_k = o_k \odot tanh(c_k) \quad \text{(hidden state)} \quad (1.35)$$

$$y_k = W_{yh} h_k + b_y \quad \text{(output)} \quad (1.36)$$

Παραπάνω, με σ συμβολίζεται η σιγμοειδής συνάρτηση $\sigma(x) = \frac{1}{1+e^{-x}}$

Το εν λόγω τοπικό μοντέλο απεικονίζεται και γραφικά, ως ακόλουθα:



***Εικόνα 1.10**: Κύτταρο LSTM*

Τα LSTMs διασφαλίζουν μια ευσταθή ροή των παραγώγων ως εξής: Θεωρώντας την αναδρομική εξίσωση της μεταβλητής cell (1.34), προκύπτει ότι η $\frac{\partial L}{\partial c_k}$ υπολογίζεται μέσω της σχέσης

$$\frac{\partial L}{\partial c_k} = \frac{\partial L}{\partial c_T} \prod_{j=k+1}^{T} \frac{\partial c_j}{\partial c_{j-1}} = \frac{\partial L}{\partial c_T} \prod_{j=k+1}^{T} f_j \tag{1.37}$$

Παρατηρούμε ότι αν $f_j \simeq 1$ για το διάστημα $[k+1, T]$, τότε η $\frac{\partial L}{\partial c_T}$ μεταφέρεται σχεδόν αυτούσια προς παρελθόντα δείγματα. Με δεδομένο ότι οι τιμές εξόδου των forget gates μπορούν να διαμορφωθούν μέσω της διαδικασία μάθησης, είναι σαφές ότι το μοντέλο μπορεί να εκπαιδευτεί ώστε να μεταφέρει πληροφορία αυτούσια πίσω στον χρόνο, γεγονός που ευνοεί την ανίχνευση μακροπρόθεσμων εξαρτήσεων.

Τέλος, όσον αφορά τον αλγόριθμο μάθησης και τις υπερπαραμέτρους των LSTMs, αυτές είναι όμοιες με τις αντίστοιχες των RNNs.

II) Δίκτυα GRU (Gate Recurrent Unit Networks)

Τα GRUs [14] αποτελούν επίσης μια παραλλαγή αναδρομικού δικτύου, με ελαφρώς χαμηλότερη πολυπλοκότητα από τα LSTMs. Το τοπικό μοντέλο που υλοποιούν συνίσταται στις παρακάτω εξισώσεις:

$$z_k = \sigma(W_z x_k + U_z h_{k-1} + b_z) \quad \text{(update gate)} \tag{1.38}$$

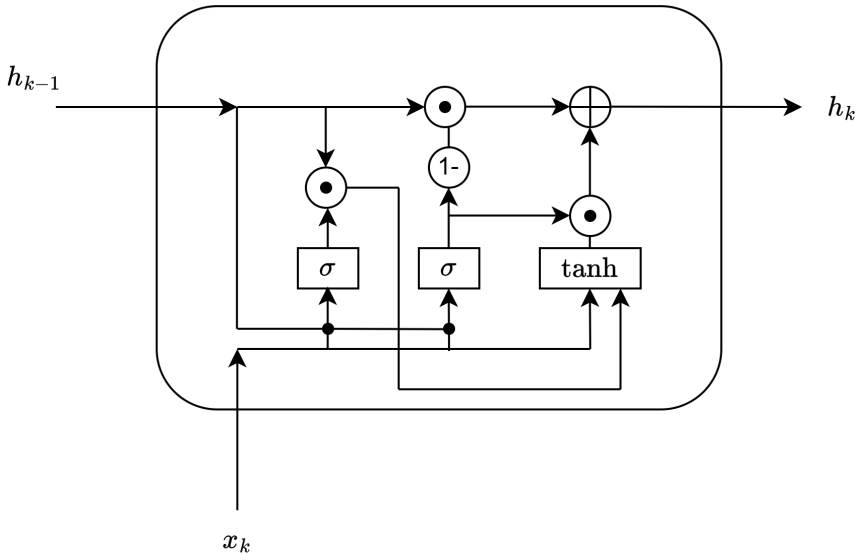
$$r_k = \sigma(W_r x_k + U_r h_{k-1} + b_r) \quad \text{(reset gate)} \tag{1.39}$$

$$\bar{h}_k = tanh(W_h x_k + U_h(r_k \odot h_{k-1}) + b_h) \quad \text{(candidate state)} \tag{1.40}$$

$$h_k = (1 - z_k) \odot h_{k-1} + z_k \odot \bar{h}_k \quad \text{(hidden state update)} \tag{1.41}$$

$$y_k = W_{yh} h_k + b_y \quad \text{(output)} \tag{1.42}$$

Γραφικά, το κύτταρο ενός GRU απεικονίζεται παρακάτω:



***Εικόνα 1.11**: Κύτταρο GRU*

Η εξασφάλιση της ευσταθούς οπισθοδιάδοσης των παραγώγων επιτυγχάνεται κυρίως μέσω της πύλης update και αποδεικνύεται με παρόμοια επιχειρήματα όπως και στα LSTMs. Αντίστοιχα, ο αλγόριθμος μάθησης και οι υπερπαράμετροι είναι όμοιες με των LSTMs και RNNs.

1.4.4 Αξιολόγηση μοντέλων

Η αξιολόγηση των μοντέλων παλινδρόμησης πραγματοποιείται αποτιμώντας το σφάλμα πρόβλεψής τους σε ένα υποσύνολο δεδομένων στο οποίο δεν έχουν εκτεθεί κατά τη διάρκεια της εκπαίδευσης (*σύνολο ελέγχου*). Το σύνολο ελέγχου προκύπτει με δειγματοληψία του αρχικού συνόλου δεδομένων και θα πρέπει να είναι κατά το δυνατόν αντιπροσωπευτικό της κατανομής που θα συναντήσει το μοντέλο σε πραγματικές συνθήκες. Κατ' αυτόν τον τρόπο, θα είμαστε σε θέση να εκτιμήσουμε και τα σφάλματα πρόβλεψης που θα παράγονται σε πραγματικές συνθήκες, το οποίο είναι και αυτό που τελικά μας ενδιαφέρει.

Η αποτίμηση του σφάλματος πρόβλεψης μπορεί να νοηθεί μέσω διαφόρων πρισμάτων, για καθένα εκ των οποίων υφίστανται και κατάλληλες μετρικές. Εν προκειμένω, θα δώσουμε έμφαση στη μέση στατιστική συμπεριφορά του σφάλματος, η οποία υποδηλώνει το πόσο καλά αποκρίνεται το μοντέλο στη μέση περίπτωση, αλλά και στη διασπορά του σφάλματος. Η ανάλυση της διασποράς και των ακραίων τιμών των σφαλμάτων είναι κρίσιμες σε πολλές εφαρμογές, ανάμεσά τους και οι ασύρματες επικοινωνίες, καθότι, όπως θα εξηγηθεί και στα Κεφάλαια 3 και 4, οι αυστηρές προδιαγραφές για τη διαθεσιμότητα της επικοινωνίας θέτουν την ανάγκη για διαστήματα εμπιστοσύνης πολύ υψηλής πιθανότητας γύρω από την εκάστοτε πρόβλεψη. Ως εκ τούτου, ένα μοντέλο που επιτυγχάνει υψηλή μέση ακρίβεια πρόβλεψης, αλλά τείνει να παράγει ακραίες τιμές σφάλματος, θα απαιτεί όλο και

ευρύτερα διαστήματα εμπιστοσύνης για το ίδιο επίπεδο εμπιστοσύνης, γεγονός που συνεπάγεται αυξημένη κατανάλωση τηλεπικοινωνιακών πόρων. Επιπλέον στις παραπάνω μετρικές, ένα εποπτικό μέτρο αποτίμησης της ποιότητας των προβλέψεων το οποίο και θα εξετάσουμε είναι η απόκλιση μεταξύ της κατανομής των πραγματικών τιμών και της κατανομής των προβλέψεων, η οποία και ποσοτικοποιείται από την απόσταση Kullback - Leibler. Οι μετρικές που αποτυπώνουν τις ιδέες που αναφέρθηκαν παρουσιάζονται αναλυτικά στον ακόλουθο πίνακα.

Μετρική	**Μαθηματικός τύπος**	**Επεξήγηση**
RMSE	$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$	Ριζικό μέσο τετραγωνικό σφάλμα. Μετρά το *μέσο σφάλμα*· λόγω τετραγώνου τιμωρεί έντονα τα μεγάλα σφάλματα.
MAE	$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$	Μέσο απόλυτο σφάλμα. Εκτιμά το *μέσο σφάλμα* χωρίς να υπερεκτιμά τις ακραίες τιμές.
Διασπορά απόλυτου σφάλματος	$\sigma^2_{\lvert e\rvert} = \frac{1}{N}\sum_{i=1}^{N}\left(\lvert e_i\rvert - \overline{\lvert e\rvert}\right)^2, \quad e_i = y_i - \hat{y}_i$	Μετρά τη *διασπορά* (variance) των απόλυτων σφαλμάτων γύρω από τη μέση τιμή τους.
90th, 95th, 99th percentiles	$P_q(\lvert e\rvert) = \inf\{\, z : \Pr(\lvert e\rvert \le z) \ge q\}, \; q \in \{0.90, 0.95, 0.99\}$	Ποσοστιαίες τιμές (δείκτες ουράς) του απόλυτου σφάλματος. Εστιάζουν στη *διασπορά*—ιδίως στα ακραία σφάλματα (tail risk).
Απόσταση KL	$D_{\mathrm{KL}}(P\|Q) = \sum_x P(x)\,\log\frac{P(x)}{Q(x)}$	Απόκλιση Kullback–Leibler μεταξύ πραγματικής κατανομής P και εκτιμημένης Q. Μετρά τη διαφορά δύο **κατανομών**

Πίνακας 1.5: Πίνακας μετρικών αξιολόγησης μοντέλων παλινδρόμησης

Κεφάλαιο 2

Ανάλυση του περιβάλλοντος διάδοσης και της διαδικασίας λήψης μετρήσεων

2.1 Εισαγωγή

Η ανάπτυξη αξιόπιστων συστημάτων πρόβλεψης θέτει ως αναγκαία προϋπόθεση την επισταμένη δειγματοληψία των μεταβλητών που επιδιώκουμε να συσχετίσουμε. Μια ποιοτική διαδικασία δειγματοληψίας καλείται:

(α) Να αποφέρει έναν ικανό όγκο δεδομένων. Ένα πυκνό σύνολο δεδομένων ενός ζεύγους μεταβλητών εισόδου - εξόδου (X, Y) καθιστά εφικτή την παρατήρηση δομών στην κατανομή του, οι οποίες κατόπιν μπορούν να αξιοποιηθούν από μοντέλα μάθησης με στόχο την πρόβλεψη.

(β) Να λαμβάνει χώρα υπό αντιπροσωπευτικές συνθήκες. Με αυτό, εννοούμε οι συνθήκες στις οποίες πραγματοποιούνται οι μετρήσεις να προσομοιώνουν τις πραγματικές συνθήκες στις οποίες θα κληθεί να λειτουργήσει το σύστημα πρόβλεψης. Μαθηματικά, η απαίτηση αυτή διασφαλίζει ότι η κατανομή την οποία θα εκπαιδευτεί να προβλέπει το σύστημα προσεγγίζει την κατανομή την οποία θα κληθεί πράγματι να προβλέψει.

Εν προκειμένω, το πρόβλημα ενδιαφέροντος αφορά την πρόβλεψη των απωλειών διάδοσης που υπεισέρχονται σε μια ζεύξη μεταξύ UAV χαμηλής τροχιάς και επίγειου τερματικού, η οποία πραγματοποιείται υπεράνω αστικής περιοχής. Το σκέλος (α) ικανοποιείται λαμβάνοντας μετρήσεις της ισχύος λήψης με υψηλή χρονική πυκνότητα, για παρατεταμένο χρονικό διάστημα. Το σκέλος (β) απαιτεί την εγκατάσταση της διάταξης μετρήσεων σε ένα αστικού τύπου περιβάλλον, καθώς και τη σχεδίαση της τροχιάς του UAV ώστε να ενσωματώνει μια ποικιλία συνθηκών διάδοσης, τις οποίες είναι πιθανό να συναντήσει σε πραγματική λειτουργία.

2.2 Ανάλυση του περιβάλλοντος διάδοσης και της τροχιάς του UAV

Η διαδικασία των μετρήσεων πραγματοποιήθηκε στην Πράγα, σε ένα τυπικό Ευρωπαϊκό αστικό περιβάλλον, το οποίο χαρακτηρίζεται από την παρουσία πολλαπλών σκεδαστών και εμποδίων, ανάμεσά τους βλάστηση και υψηλά κτίρια. Σε ένα τέτοιο περιβάλλον, υπεισέρχεται μια ποικιλία φαινομένων διάδοσης, τα οποία υποβαθμίζουν την ποιότητα της ζεύξης και δυσχεραίνουν την επικοινωνία: Οι απώλειες ισχύος λόγω απόστασης (*απώλειες ελευθέρου χώρου*), οι οποίες είναι εκ προοιμίου παρούσες σε κάθε ασύρματη ζεύξη, εντείνονται από τα εμπόδια μεγάλης κλίμακας (όπως τα δένδρα και τα κτίρια), τα οποία επιφέρουν και αντίστοιχες *απώλειες (ή διαλείψεις) μεγάλης κλίμακας*, είτε διακόπτοντας πλήρως την ευθεία οπτικής επαφής (LoS) μεταξύ πομπού και δέκτη, είτε

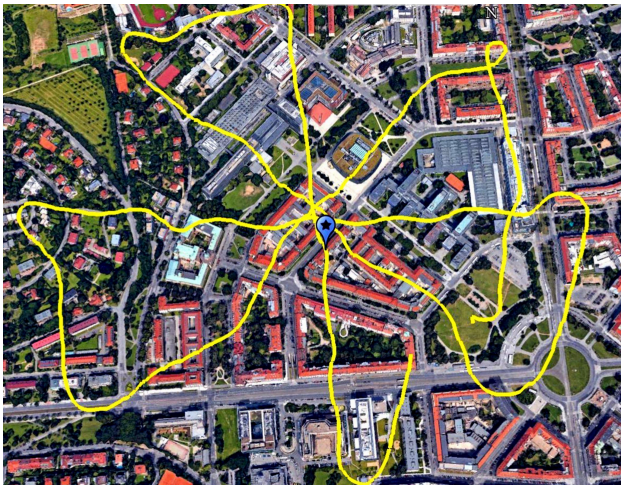
μέσω κυματικών φαινομένων όπως η περίθλαση. Οι διαλείψεις μεγάλης κλίμακας (ή *απώλειες διαδρομής*) φέρουν στοχαστικότητα, αλλά χαρακτηρίζονται από το γεγονός ότι παραμένουν κατά προσέγγιση σταθερές εντός μιας περιοχής διάστασης αρκετών μηκών κύματος. Επιπρόσθετα στις απώλειες διαδρομής, η πολυπλοκότητα του αστικού περιβάλλοντος, η οποία σε όρους ασύρματης διάδοσης μεταφράζεται στην παρουσία πολλαπλών σκεδαστών, επιφέρει και *απώλειες (ή διαλείψεις) μικρής κλίμακας*. Συγκεκριμένα, οι ανακλάσεις που υφίσταται το ηλεκτρομαγνητικό κύμα εκπομπής κατά τη διάδοσή του, έχουν ως αποτέλεσμα τη λήψη πολλαπλών εκδοχών του αρχικού σήματος (*πολυδιαδρομικές συνιστώσες*), οι οποίες καταφθάνουν στον δέκτη με διαφορετικό πλάτος και φάση, όπως αποτυπώνεται στην ακόλουθη σχέση[15]:

$$r(t) = Re\{[\sum_{n=0}^{N(t)} a_n(t)exp(-j\varphi_n(t))u(t-\tau_n(t))]exp(j2\pi f_c t + \psi)]\} \quad (2.1)$$

όπου $Re[u(t)exp(j2\pi f_c t + \psi)]$ είναι το σήμα εκπομπής.

Ανάλογα τη συμφωνία φάσης των συνιστωσών, η υπέρθεσή τους είναι άλλοτε ενισχυτική, άλλοτε ακυρωτική. Στη δεύτερη περίπτωση, η οποία είναι και η συνηθέστερη, το σήμα μπορεί υποστεί ισχυρότατες απώλειες, τάξεων αρκετών dB, οι οποίες είναι πρόσθετες στις απώλειες μεγάλης κλίμακας. Οι απώλειες αυτές μεταβάλλονται σε μικρή κλίμακα (τάξης μερικών μηκών κύματος) και με έντονα στοχαστικό τρόπο, ιδίως σε περιπτώσεις ζεύξεων μη οπτικής επαφής.

Η ένταση των παραπάνω φαινομένων μικρής και μεγάλης κλίμακας συναρτάται με διάφορες παραμέτρους της ζεύξης, όπως η απόσταση, η γωνία ανύψωσης, η γωνία αζιμουθίου. Προκειμένου, λοιπόν, να καταστήσουμε τη δειγματοληψία αντιπροσωπευτική, θα ορίσουμε μια τροχιά, η οποία να εμπλέκει LoS και NLoS συνθήκες διάδοσης, διαφορετικές θέσεις, αποστάσεις και γωνίες μεταξύ πομπού και δέκτη. Η εν λόγω τροχιά απεικονίζεται ακολούθως:



***Εικόνα 2.1**: Τροχιά αεροπλοίου και θέση τερματικού*

Το αερόπλοιο κινείται με περίπου σταθερή ταχύτητα, μέσης τιμής ίσης με 5.9 m/s, σε αποστάσεις μεταξύ 200 και 520 μέτρων. Η γωνία ανύψωσης κυμαίνεται μεταξύ 10 και 90 μοιρών, ενώ η γωνία αζιμουθίου μεταξύ 0 και 360 μοιρών. Το τερματικό τοποθετείται σε σταθερή θέση, ύψους περίπου 1.7 μέτρων από το έδαφος και περίπου στο κέντρο της τροχιάς, όπως απεικονίζεται στην Εικόνα 2.1.

2.3 Η διάταξη εκπομπής και λήψης

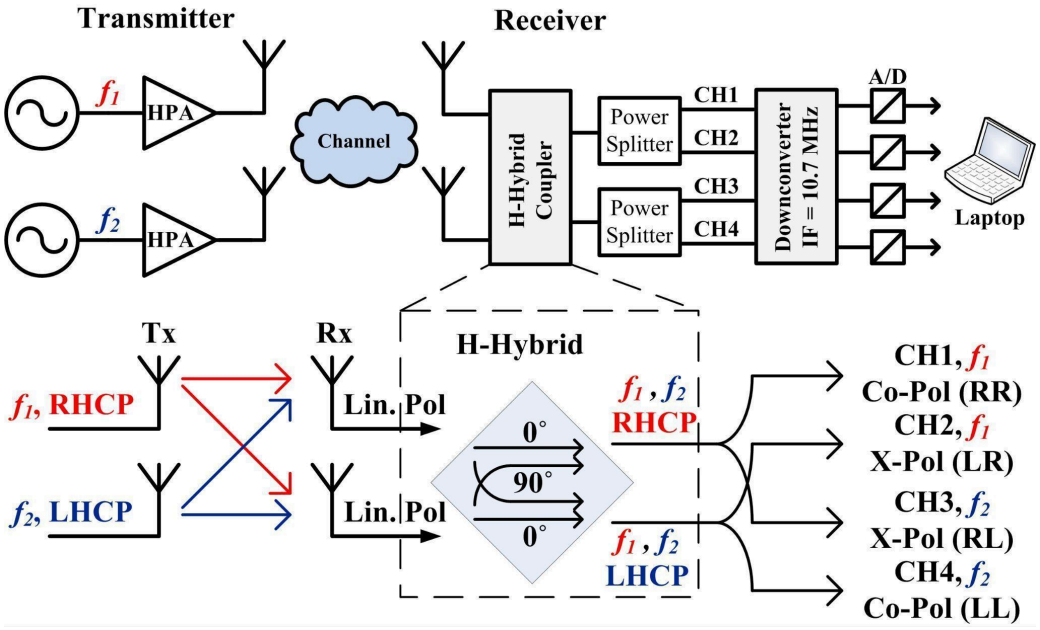
Για την εκπομπή και λήψη του σήματος, θα χρησιμοποιήσουμε μια διάταξη δύο κεραιών, διπλής πόλωσης (DP MIMO 2x2). Η MIMO (Multiple Input Multiple Output) τεχνολογία, η εκπομπή και λήψη, δηλαδή, του σήματος μέσω πολλαπλών κεραιών, εδραιώνεται όλο και περισσότερο στις ασύρματες επικοινωνίες, εξαιτίας των δυνατοτήτων που προσφέρει για βελτίωση της χωρητικότητας και της φασματικής αποδοτικότητας της ζεύξης, καθώς και της άμβλυνσης των διαλείψεων [16]. Ως εκ τούτου, είναι χρήσιμο η σχεδίαση του συστήματος πρόβλεψης να πραγματοποιηθεί σε ένα τέτοιο τηλεπικοινωνιακό πλαίσιο.

Η εκπομπή του ηλεκτρομαγνητικού κύματος - φορέα του σήματος πραγματοποιείται από ένα ζεύγος σπειροειδών κεραιών (spiral antennas), μιας αριστερόστροφα (RHCP) και μιας δεξιόστροφα κυκλικά πολωμένης (LHCP). Καθεμία εκπέμπει από ένα πειραματικό σήμα CW (continuous wave) της μορφής $s_i(t) = Acos(2\pi f_i t + \varphi_i),\ i = 1, 2$, σταθερής ισχύος ίσης με 27dBm. Για την πρώτη κεραία, η φέρουσα συχνότητα f_1 είναι ίση με 2.00106 GHz, ενώ για τη δεύτερη η φέρουσα f_2 ισούται με 2.00086 GHz. Οι δύο κεραίες είναι τοποθετημένες σε απόσταση 12cm (0.8λ).

Τα σήματα εκπομπής λαμβάνονται από μια μικροταινιακή κεραία διπλής γραμμικής πόλωσης (dual-linearly polarized patch antenna). Η κεραία αυτή απομονώνει τις ορθογώνιες συνιστώσες του πεδίου λήψης (E_x και E_y) και τροφοδοτεί με είσοδο έναν H-υβριδικό συζεύκτη (H-hybrid coupler). Από τις ορθογώνιες συνιστώσες, ο συζεύκτης διαχωρίζει τις συνιστώσες ($E_{x,RHCP}$, $E_{y,RHCP}$), ($E_{x,LHCP}$, $E_{y,LHCP}$) που αντιστοιχούν στα δύο αρχικά σήματα βάσει της συχνότητας και για καθένα εξ αυτών, ανακατασκευάζει ένα RHCP και ένα LHCP σήμα. Η ύπαρξη ενός σήματος όμοιας (copolar) και ενός διασταυρωμένης (crosspolar) πόλωσης για κάθε σήμα εκπομπής οφείλεται στα προαναφερθέντα φαινόμενα διάδοσης μικρής και μεγάλης κλίμακας. Ανάλυση της αποπόλωσης βάσει μετρήσεων σε δίαυλο μεταξύ UAV και επίγειου τερματικού πραγματοποιείται εδώ [17].

Τα σήματα αυτά δίνονται ως είσοδος σε έναν διαιρέτη ισχύος (power splitter), ο οποίος παράγει 4 κανάλια ισχύος εξόδου ($RHCP \rightarrow RHCP,\ RHCP \rightarrow LHCP,\ LHCP \rightarrow RHCP,\ LHCP \rightarrow LHCP$). Τα κανάλια αυτά τροφοδοτούνται σε έναν αντίστοιχο δέκτη 4 καναλιών και δειγματοληπτούνται με συχνότητα 10kHz (1 δείγμα / 0.1ms). Θεωρώντας την ταχύτητα του UAV σταθερή στα 5.9m/s, αυτό αντιστοιχεί σε περίπου 258 δείγματα/μήκος κύματος. Τα δείγματα αυτά

κατόπιν ψηφιοποιούνται και αποθηκεύονται σε ηλεκτρονικό υπολογιστή για περαιτέρω επεξεργασία. Επισημαίνεται ότι οι τελικές καταγεγραμμένες αριθμητικές τιμές της ισχύος λήψης είναι κανονικοποιημένες, καθότι έχουν αφαιρεθεί τα κέρδη των κεραιών εκπομπής και λήψης.



Εικόνα 2.2: Λειτουργικό διάγραμμα της διάταξης εκπομπής και λήψης

2.4 Δημιουργία συνόλου δεδομένων

Επιπλέον του συστήματος εκπομπής, το αερόπλοιο είναι εφοδιασμένο με αισθητήριες διατάξεις, με στόχο τη διευκόλυνση του υπολογισμού αποστάσεων και την επίτευξη βέλτιστου προσανατολισμού προς την κεραία λήψης. Οι μετρήσεις που προκύπτουν, και οι οποίες πραγματοποιούνται σε συγχρονισμό με τη δειγματοληψία της ισχύος, θα συγκροτήσουν το σύνολο δεδομένων που θα χρησιμοποιηθεί για την ανάπτυξη του συστήματος πρόβλεψης.

Το εν λόγω σύνολο δεδομένων είναι δομημένο (tabular) και αποτελείται από 23 στήλες - χαρακτηριστικά και 8387105 γραμμές - δείγματα. Παρακάτω, ακολουθεί μια συνοπτική περιγραφή των 23 χαρακτηριστικών:

Χαρακτηριστικό	Συνοπτική περιγραφή
time	Χρονική σήμανση καταγραφής δεδομένων
gps_h	Υψόμετρο UAV βάσει GPS
gps_easting	Αν. γεωγραφική συντεταγμένη UAV (UTM Easting)
gps_northing	Βόρεια γεωγραφική συντεταγμένη UAV (UTM Northing)
compass	Γωνία προσανατολισμού (heading) UAV
pitch	Γωνία ανόδου/καθόδου UAV
roll	Γωνία κύλισης UAV
received_power_ch1	Ληφθείσα ισχύς στο κανάλι 1
received_power_ch2	Ληφθείσα ισχύς στο κανάλι 2
received_power_ch3	Ληφθείσα ισχύς στο κανάλι 3
received_power_ch4	Ληφθείσα ισχύς στο κανάλι 4
distance	Απόσταση UAV–τερματικού
elevation	Γωνία ανύψωσης UAV ως προς το τερματικό
azimuth	Οριζόντια γωνία (αζιμούθιο) προς UAV
deltaX	Μεταβολή θέσης UAV στον άξονα x
deltaY	Μεταβολή θέσης UAV στον άξονα y
deltaH	Μεταβολή υψομέτρου UAV
received_power_20km_ch1	Ληφθείσα ισχύς ch1 προσομοιωμένης απόστασης 20km
received_power_20km_ch2	Ληφθείσα ισχύς ch2 στα 20km
received_power_20km_ch3	Ληφθείσα ισχύς ch3 στα 20km
received_power_20km_ch4	Ληφθείσα ισχύς ch4 στα 20km
dist_earth	Απόσταση UAV από την επιφάνεια του εδάφους
speed	Ταχύτητα UAV

Πίνακας 2.1: Περιγραφή των χαρακτηριστικών του συνόλου δεδομένων

Κεφάλαιο 3

Πρόβλεψη των απωλειών διάδοσης μεγάλης κλίμακας με μοντέλα μηχανικής μάθησης

3.1 Εισαγωγή

Σε μια SISO (single input single output) ασύρματη ζεύξη, η στιγμιαία ισχύς εκπομπής ($P_T(t)$) συνδέεται με τη στιγμιαία ισχύ λήψης ($P_R(t)$) μέσω της ακόλουθης γενικής σχέσης:

$$P_R(t) = P_T + G_T + G_R - L(t) \tag{3.1}$$

Τα παραπάνω μεγέθη είναι σε μονάδες dB. Με G_T, G_R συμβολίζουμε τα κέρδη των κεραιών εκπομπής και λήψης αντίστοιχα, ενώ με $L(t)$ συμβολίζουμε τις στιγμιαίες απώλειες διάδοσης. Οι στιγμιαίες απώλειες διάδοσης αναλύονται περαιτέρω ως το άθροισμα των απωλειών διάδοσης μικρής ($L_{SS}(t)$) και μεγάλης κλίμακας ($L_{LS}(t)$) τη χρονική στιγμή t:

$$L(t) = L_{LS}(t) + L_{SS}(t) \tag{3.2}$$

Για το κινητό μέρος της ζεύξης (εν προκειμένω το αερόπλοιο), κάθε χρονική στιγμή t αντιστοιχεί και σε μια χωρική θέση r. Οι απώλειες διάδοσης μεγάλης κλίμακας για ένα χωρικό σημείο r ορίζονται (κατ' αρχήν με μη αυστηρό τρόπο) ως η μέση τιμή των στιγμιαίων απωλειών σε μια περιοχή N_r δειγμάτων γύρω από το r, όπου N_r ισούται με "αρκετά" λ. Αντίστοιχα, οι απώλειες διάδοσης μεγάλης κλίμακας για μια χρονική στιγμή t θα ισούνται με τη μέση τιμή των απωλειών σε ένα διάστημα N_t δειγμάτων, με N_t χρονικά δείγματα να αντιστοιχούν σε N_r χωρικά δείγματα. Η σχέση μεταξύ N_r και N_t προκύπτει ως

$$\frac{N_r}{N_t} = \frac{\lambda}{D_s} \tag{3.3}$$

όπου D_s το πλήθος των χρονικών δειγμάτων που αντιστοιχεί σε διανυθείσα απόσταση ίση με λ. Αν v είναι η μέση ταχύτητα του αεροπλοίου και F_s η συχνότητα δειγματοληψίας, τότε:

$$D_s = \frac{F_s}{v} \cdot \lambda \tag{3.4}$$

Το αποτέλεσμα της (3.3) στρογγυλοποιείται στον κοντινότερο ακέραιο. Θέτοντας $N_r = 40 \cdot \lambda$, έχουμε ότι $N_t = 40 \cdot D_s$ και άρα:

$$L_{LS}(t) = \frac{1}{40 \cdot D_s} \sum_{n=-40 \cdot D_s/2}^{(40 \cdot D_s/2)-1} L(t) \tag{3.5}$$

Συνδυάζοντας τις (3.4), (3.1), προκύπτει ότι:

$$L_{LS,i}(t) = P_T + G_T + G_R - \frac{1}{40 \cdot D_s} \sum_{n=-40 \cdot D_s/2}^{40 \cdot D_s/2-1} P_{R,i}(t + n) \tag{3.6}$$

Με δεδομένο, όμως, ότι από τη στιγμιαία ισχύ λήψης που καταγράφεται στο σύνολο δεδομένων, έστω P_R', έχουν αφαιρεθεί τα κέρδη τα κεραιών, τότε θα ισχύει:

$$L_{LS}(t) = P_T - \frac{1}{40 \cdot D_s} \sum_{n=-40 \cdot D_s/2}^{40 \cdot D_s/2-1} P_R'(t + n) \tag{3.7}$$

Εφαρμόζοντας την (3.5) χωριστά για κάθε κανάλι του DP 2x2 MIMO συστήματος, προκύπτουν οι ακόλουθες σχέσεις:

$$L_{LS,RR}(t) = P_T - \frac{1}{40 \cdot D_s} \sum_{n=-40 \cdot D_s/2}^{40 \cdot D_s/2-1} P_{R,RR}'(t + n) \tag{3.8}$$

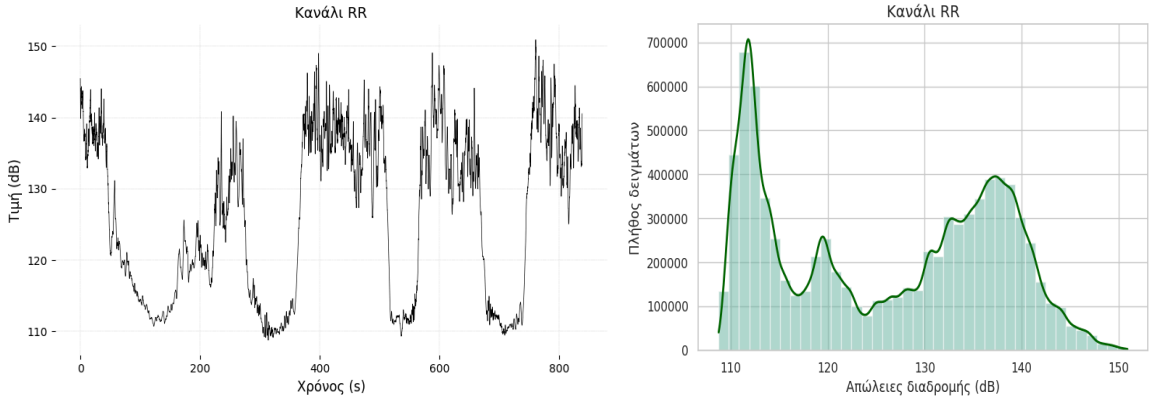
$$L_{LS,LR}(t) = P_T - \frac{1}{40 \cdot D_s} \sum_{n=-40 \cdot D_s/2}^{40 \cdot D_s/2-1} P_{R,LR}'(t + n) \tag{3.9}$$

$$L_{LS,RL}(t) = P_T - \frac{1}{40 \cdot D_s} \sum_{n=-40 \cdot D_s/2}^{40 \cdot D_s/2-1} P_{R,RL}'(t + n) \tag{3.10}$$

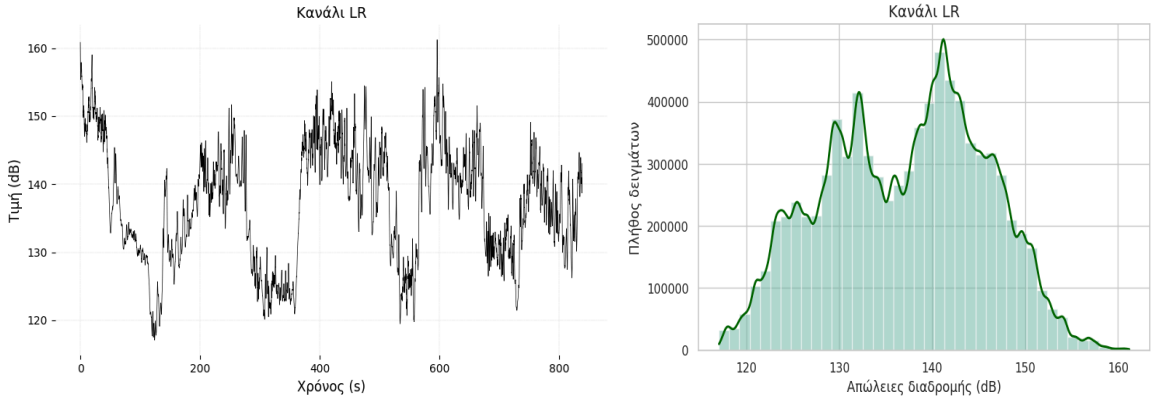
$$L_{LS,LL}(t) = P_T - \frac{1}{40 \cdot D_s} \sum_{n=-40 \cdot D_s/2}^{40 \cdot D_s/2-1} P_{R,LL}'(t + n) \tag{3.11}$$

Οι τέσσερις αυτές μεταβλητές θα αποτελέσουν τις μεταβλητές - στόχους του συστήματος μάθησης. Οι αριθμητικές τους τιμές για το σύνολο εκπαίδευσης προκύπτουν άμεσα από τις μετρήσεις, σε συνδυασμό με το γεγονός ότι η ισχύς εκπομπής είναι σταθερή και ίση με 27dBm.

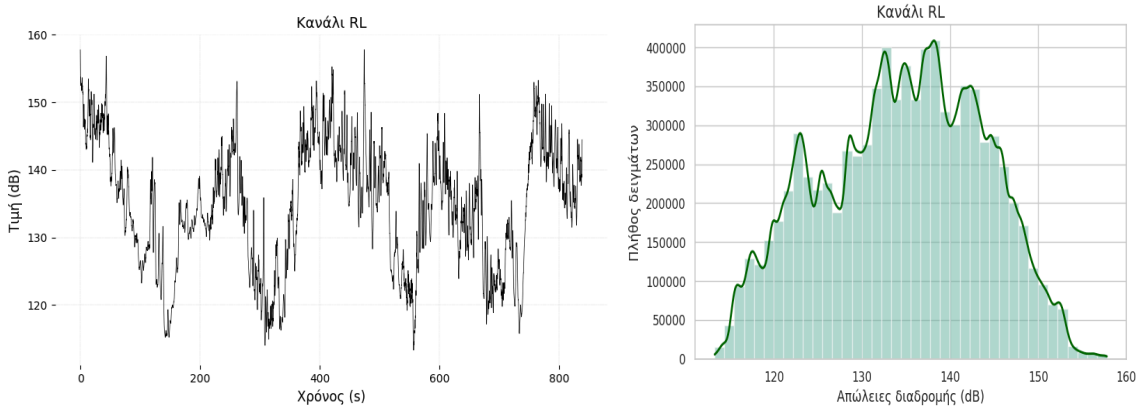
Παρακάτω, απεικονίζονται τα δείγματα εκπαίδευσης των απωλειών διαδρομής για κάθε κανάλι, σε μορφή χρονοσειράς και ιστογράμματος.



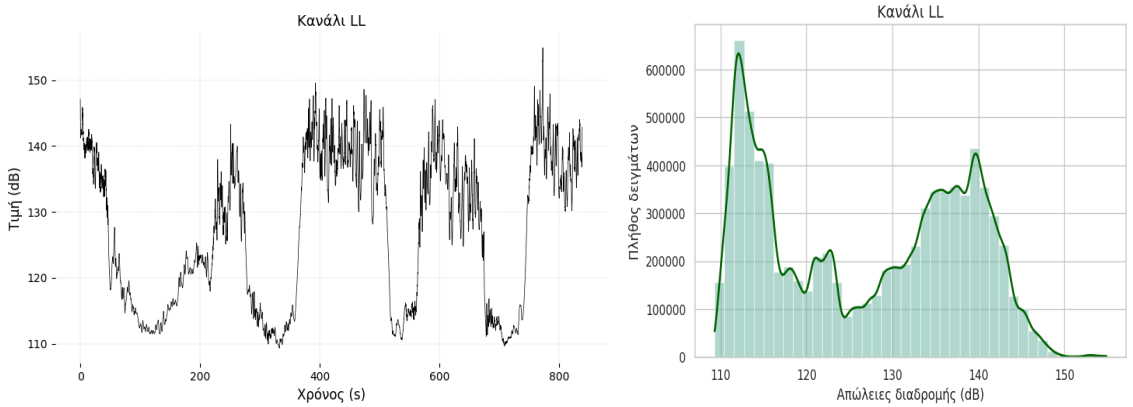
***Εικόνα 3.1**: Χρονοσειρά και ιστόγραμμα απωλειών διαδρομής για το κανάλι RR*



***Εικόνα 3.2**: Χρονοσειρά και ιστόγραμμα απωλειών διαδρομής για το κανάλι LR*



***Εικόνα 3.3**: Χρονοσειρά και ιστόγραμμα απωλειών διαδρομής για το κανάλι RL*



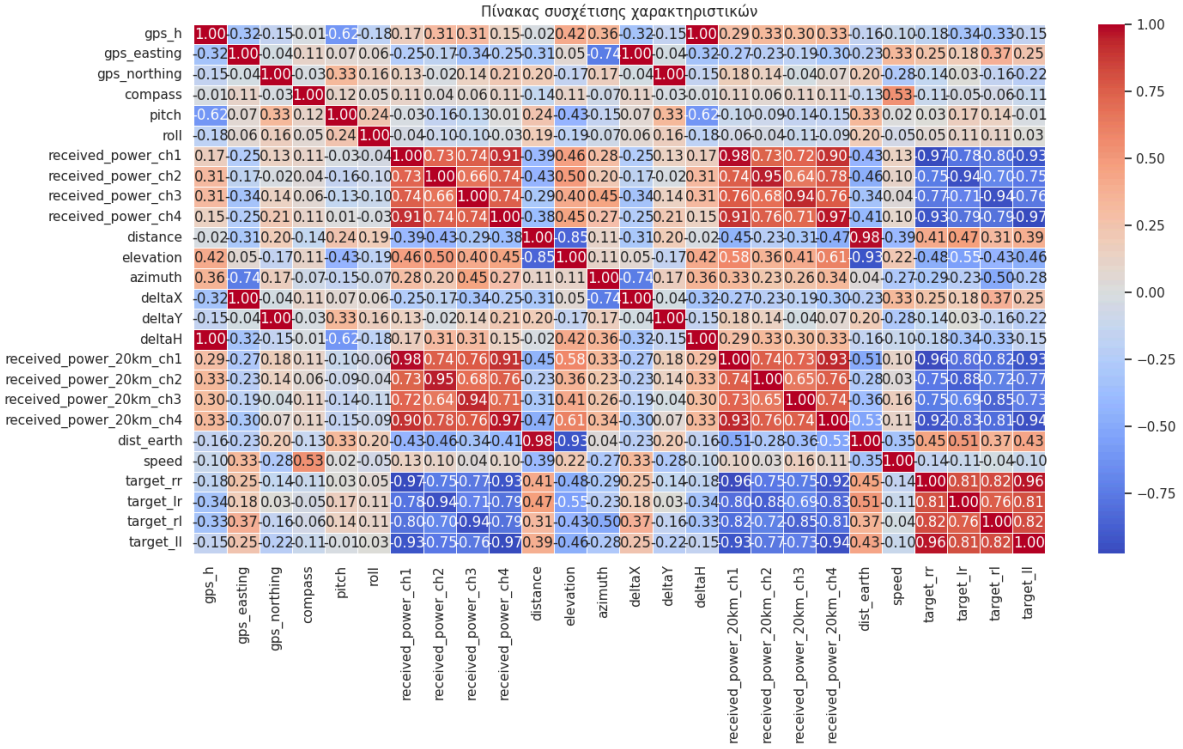
***Εικόνα 3.4**: Χρονοσειρά και ιστόγραμμα απωλειών διαδρομής για το κανάλι LL*

3.2 Στατιστική ανάλυση των δεδομένων με στόχο τη μείωση διάστασης

Η δυνατότητα μείωσης διάστασης των δεδομένων θα διερευνηθεί μέσω των διαδικασιών επιλογής και μετασχηματισμού χαρακτηριστικών.

Ι) Επιλογή χαρακτηριστικών

Όπως αναφέρθηκε και στο πρώτο κεφάλαιο, η επιλογή χαρακτηριστικών θα πραγματοποιηθεί βάσει του συντελεστή συσχέτισης Pearson και της αμοιβαίας πληροφορίας. Παρακάτω, παρουσιάζεται κατ' αρχάς ο πίνακας συσχέτισης μεταξύ των χαρακτηριστικών και των μεταβλητών προς πρόβλεψη:



Εικόνα 3.5: *Πίνακας συσχέτισης μεταξύ χαρακτηριστικών και μεταβλητών - στόχων*

Από τον παραπάνω πίνακα, διαπιστώνουμε σημαντικές συσχετίσεις, τόσο μεταξύ των χαρακτηριστικών, όσο και μεταξύ χαρακτηριστικών και μεταβλητών - στόχων. Μάλιστα, στην πρώτη περίπτωση, υφίστανται ζεύγη μεταβλητών που παρουσιάζουν τέλεια συσχέτιση και συγκεκριμένα οι (deltah, gps_h), (deltaY, gps_northing) και (deltaX, gps_easting). Προφανώς, για καθένα από τα τρία αυτά ζεύγη, η πρώτη μεταβλητή δεν προσφέρει καμία επιπλέον πληροφορία για τις εξόδους σε σχέση με τη δεύτερη, οπότε και αφαιρείται.

Η παραπάνω απόφαση για μείωση της διάστασης ήταν τετριμμένη. Εν γένει, ο προσδιορισμός των χαρακτηριστικών που περιέχουν τη "λιγότερη πληροφορία" για τις εξόδους και άρα είναι δόκιμο να αφαιρεθούν χαρακτηρίζεται από μεγαλύτερη δυσκολία. Κι αυτό, διότι ένα χαρακτηριστικό ενδέχεται να περιέχει λιγοστή πληροφορία είτε επειδή εγγενώς έχει χαμηλή συσχέτιση με τις εξόδους, είτε επειδή είναι υψηλά συσχετισμένο με ένα άλλο χαρακτηριστικό, και άρα η επιπλέον πληροφορία που συνεισφέρει είναι χαμηλή. Οι δύο αυτές περιπτώσεις είναι ετερογενείς και άρα δεν μπορούν να αποτιμηθούν βάσει κάποιας κοινής σχέσης διάταξης. Δεν μπορούμε π.χ. να ισχυριστούμε με ασφάλεια ότι ένα χαρακτηριστικό χαμηλής συσχέτισης με τις εξόδους είναι λιγότερο ή περισσότερο σημαντικό από ένα χαρακτηριστικό με υψηλότερη συσχέτιση με τις εξόδους, αλλά ισχυρά συσχετισμένο με ένα τρίτο χαρακτηριστικό. Τα δύο κριτήρια αυτά, λοιπόν, συνεκτιμώνται, αλλά η τελική απόφαση για το ποια χαρακτηριστικά θα αφαιρεθούν βασίζεται περισσότερο σε σχεδιαστική διαίσθηση παρά σε κάποια αλγοριθμική διαδικασία.

Για τα εναπομείναντα 19 χαρακτηριστικά, λοιπόν, θα εφαρμόσουμε τη διαδικασία επιλογής προοδευτικά σε τέσσερις φάσεις, μειώνοντάς τα από 19 σε 14, 14 σε 10, 10 σε 8 και 8 σε 5 αντίστοιχα. Προτού αναλύσουμε με περισσότερη λεπτομέρεια τις φάσεις αυτές,

θα παρουσιάσουμε, σε μορφή πίνακα, την αμοιβαία πληροφορία μεταξύ των χαρακτηριστικών και των τεσσάρων εξόδων:

Χαρακτηριστικό	**RR**	**LR**	**RL**	**LL**
gps_h	3.517	2.875	2.943	3.473
gps_easting	5.184	4.230	4.282	5.172
gps_northing	4.875	4.003	4.065	4.852
compass	4.697	3.836	3.893	4.671
pitch	2.499	2.130	2.165	2.495
roll	1.215	0.979	0.988	1.218
received_power_ch1	3.679	2.697	2.751	3.496
received_power_ch2	3.411	2.993	2.733	3.391
received_power_ch3	3.092	2.381	2.818	3.074
received_power_ch4	3.266	2.465	2.512	3.446
distance	4.995	4.081	4.155	4.974
elevation	4.874	4.085	4.142	4.859
azimuth	4.745	3.850	3.909	4.724
received_power_20km_ch1	2.224	1.608	1.627	2.050
received_power_20km_ch2	1.578	1.547	1.332	1.589
received_power_20km_ch3	1.408	1.141	1.390	1.399
received_power_20km_ch4	1.934	1.515	1.533	2.050
dist_earth	5.023	4.148	4.215	4.998
speed	0.732	0.577	0.541	0.696

Πίνακας 3.1: Αμοιβαία πληροφορία μεταξύ χαρακτηριστικών και εξόδων

Ακολούθως, παρουσιάζονται τα χαρακτηριστικά που αφαιρούνται σε κάθε φάση, σε συνδυασμό με μια συνοπτική επεξήγηση του συλλογισμού πίσω από την αφαίρεσή τους.

Φάση αφαίρεσης	Χαρακτηριστικά που αφαιρούνται	Επεξήγηση συλλογισμού
19→14	pitch, compass, roll, speed και distance	Τα τέσσερα πρώτα παρουσιάζουν ιδιαίτερα χαμηλές τιμές συσχέτισης κατά Pearson με τις εξόδους. Επιπλέον, με εξαίρεση την compass, παρουσιάζουν επίσης χαμηλή αμοιβαία πληροφορία με τις εξόδους. Η μεταβλητή distance αφαιρείται λόγω ιδιαίτερα υψηλής συσχέτισης Pearson (0.98) με τη μεταβλητή dist_earth. Αφαιρείται αυτή και όχι η dist_earth, καθότι η δεύτερη παρουσιάζει ελαφρώς υψηλότερες τιμές συσχέτισης Pearson και αμοιβαίας πληροφορίας με τις εξόδους.
14→10	received_power_20km για τα 4 κανάλια ισχύος	Παρουσιάζουν υψηλή συσχέτιση με τα αντίστοιχα χαρακτηριστικά της ισχύος λήψης (received_power).
10→8	gps_easting, gps_h	Συνεκτιμώντας τη συσχέτιση Pearson και την αμοιβαία πληροφορία, προκύπτει ότι οι GPS συντεταγμένες του αεροπλοίου παρουσιάζουν λιγότερη πληροφορία συγκριτικά με τα γεωμετρικά χαρακτηριστικά (elevation, dist_earth, azimuth). Η gps_northing διατηρείται, καθότι είναι πιο συσχετισμένη με τις εξόδους σε σχέση με τη gps_h, ενώ η gps_easting παρουσιάζει σχετικά υψηλή απόλυτη συσχέτιση Pearson με το azimuth (0.74).
8→5	gps_northing, azimuth, elevation	Οι μεταβλητές που διατηρούνται είναι τα τέσσερα κανάλια της ισχύος λήψης (received_power) και η dist_earth. Λαμβάνοντας συνδυαστικά υπ' όψιν τα δύο κριτήρια, θεωρούνται αυτές που φέρουν τη μεγαλύτερη ποσότητα πληροφορίας για τις εξόδους.

Πίνακας 3.2: Προοδευτική μείωση διάστασης με επιλογή χαρακτηριστικών

II) Μετασχηματισμός χαρακτηριστικών

Ο μετασχηματισμός χαρακτηριστικών πραγματοποιείται μέσω της μεθόδου PCA. Για λόγους ισότιμης σύγκρισης των διαδικασιών επιλογής και μετασχηματισμού χαρακτηριστικών, οι φάσεις της προοδευτικής μείωσης διάστασης θα είναι οι ίδιες. Σε κάθε φάση, θα αποτιμήσουμε το κατά πόσο η δομή των δεδομένων διατηρείται υπολογίζοντας τη διασπορά των μετασχηματισμένων δεδομένων σε σχέση με τα αρχικά. Αυτό θα μας δώσει μια ιδέα για την απώλεια πληροφορίας που επιφέρει η μείωση διάστασης.

Φάση μείωσης	Διατηρηθείσα διασπορά
19 → 14	0.9983
14 → 10	0.9617
10 → 8	0.9185
8 → 5	0.7998

Πίνακας 3.3: Προοδευτική μείωση διάστασης με PCA

3.3 Επιλογή, εκπαίδευση και ρύθμιση υπερπαραμέτρων μοντέλων

Προσεγγίζοντας το πρόβλημα πρόβλεψης των απωλειών διαδρομής ως ένα στατικό πρόβλημα μάθησης, οι αρχιτεκτονικές που θα χρησιμοποιήσουμε θα είναι αυτές των MLP, SVR και XGB, με τους αντίστοιχους αλγορίθμους μάθησης που περιγράφονται στο 1ο κεφάλαιο. Η ρύθμιση υπερπαραμέτρων πραγματοποιείται μέσω της τεχνικής του randomized search, για τις τυπικές τιμές των υπερπαραμέτρων που περιγράφονται στους Πίνακες (1.1), (1.2) και (1.3).

Αξίζει να σημειωθεί πως το σύνολο επικύρωσης, αλλά και το σύνολο ελέγχου, δεν προκύπτουν με τυχαία δειγματοληψία του συνόλου δεδομένων. Ο λόγος είναι ότι τα δείγματα λαμβάνονται ακολουθιακά και με υψηλή πυκνότητα. Συνεπώς, για ένα δείγμα x του συνόλου επικύρωσης (ή ελέγχου) που λαμβάνεται τυχαία από το σύνολο δεδομένων, είναι εξαιρετικά πιθανό ένα γειτονικό του δείγμα να ανήκει στο σύνολο εκπαίδευσης. Για το γειτονικό αυτό δείγμα, αναμένουμε οι τιμές χαρακτηριστικών, αλλά και εξόδων να είναι παρόμοιες. Η συνάφεια αυτή οδηγεί σε *διαρροή δεδομένων (data leakage)* και άρα σε μη αξιόπιστη αξιολόγηση των μοντέλων. Για να αποφύγουμε αυτό το ενδεχόμενο, θα απομονώσουμε ένα συνεχές τμήμα από το τέλος της χρονοσειράς των απωλειών για το εκάστοτε κανάλι, μήκους ίσου με το 30% της αρχικής. Από τη μορφή των χρονοσειρών στις Εικόνες 3.1-3.4, αναμένουμε το διάστημα αυτό να είναι εν γένει αντιπροσωπευτικό των συνθηκών διάδοσης. Τα σύνολα επικύρωσης και ελέγχου προκύπτουν κατόπιν με δειγματοληψία του εν λόγω διαστήματος.

3.4 Αξιολόγηση μοντέλων

Η αξιολόγηση των μοντέλων πραγματοποιείται για όλες τις φάσεις της μείωσης διάστασης, υπολογίζοντας μετρικές που αποτυπώνουν τόσο το μέσο σφάλμα (RMSE, MAE), όσο και τη διασπορά του σφάλματος (Variance, 99th percentile).

Μπορεί το μέσο σφάλμα να αποτελεί τον δείκτη της γενικής επίδοσης ενός προβλέπτη, η ανάλυση της διασποράς του σφάλματος, όμως, είναι εξίσου σημαντική για τηλεπικοινωνιακά συστήματα: Δεδομένου πως η ζεύξη μεταξύ UAV και τερματικού οφείλει να είναι λειτουργική με μια πολύ μεγάλη πιθανότητα (έστω p) η ρύθμιση των περιθωρίων απωλειών διάδοσης του συστήματος θα πρέπει να πραγματοποιηθεί όχι με βάση το μέσο σφάλμα πρόβλεψης, αλλά με βάση το percentile του σφάλματος που αντιστοιχεί στην p. Προς αποσαφήνιση των παραπάνω, ας υποθέσουμε ότι η ζεύξη θεωρείται σε λειτουργία μόνο αν $L_{LS}(t) \leq L_{LS,max}$ και προδιαγράφεται ώστε η συνθήκη αυτή να είναι αληθής σε ποσοστό του χρόνου τουλάχιστον ίσο με p. Θα πρέπει, λοιπόν, να ισχύει:

$$Pr(L_{LS}(t) \leq L_{LS,max}) \geq p \tag{3.11}$$

Έστω ότι ο προβλέπτης ενσωματώνεται σε ένα σύστημα μετάδοσης με δυνατότητα προσαρμογής στις συνθήκες διαύλου. Τότε, οι μέγιστες επιτρεπτές απώλειες διάδοσης L_{max} θα ρυθμίζονται δυναμικά, βάσει των προβλέψεων του μοντέλου ΜΜ για τις απώλειες διαδρομής ($\hat{L}_{LS}(t)$). Αν, όμως, θέσουμε $L_{LS,max}(t) = \hat{L}_{LS}(t)$, τότε είναι εξαιρετικά πιθανό η συνθήκη $L_{LS}(t) \leq L_{LS,max}(t)$ να παραβιάζεται με πιθανότητα αρκετά χαμηλότερη του $1 - p$, λόγω σφαλμάτων στην πρόβλεψη του $L_{LS}(t)$. Για να ισχύει η (3.11) θα πρέπει να θέσουμε $L_{LS,max}(t) = \hat{L}_{LS}(t) + E_p$, όπου E_p είναι το p-th percentile του απόλυτου σφάλματος πρόβλεψης $|L_{LS}(t) - \hat{L}_{LS}(t)|$. Πράγματι, τότε θα ισχύει

$$Pr(L_{LS}(t) \leq L_{LS,max}(t)) = Pr(L_{LS}(t) \leq \hat{L}_{LS}(t) + E_p) = Pr(L_{LS}(t) - \hat{L}_{LS}(t) \leq E_p) \quad (3.12)$$

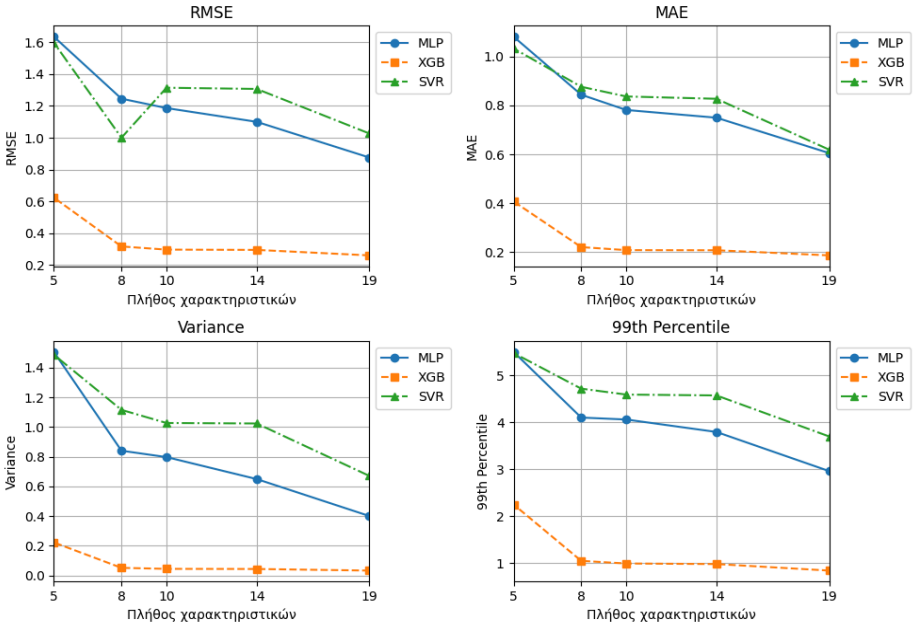
Όμως, εξ ορισμού του E_p θα ισχύει $Pr(L_{LS}(t) - \hat{L}_{LS}(t) \leq E_p) \geq p$ και άρα θα επαληθεύεται η (3.11).

Η απαίτηση για μεγαλύτερες τιμές του $L_{LS,max}(t)$ έρχεται με κόστος την αυξημένη κατανάλωση τηλεπικοινωνιακών πόρων (όπως η ισχύς εκπομπής) ή/και την ανάγκη χρήσης πιο εύρωστων σχημάτων διαμόρφωσης και κωδικοποίησης, με συνέπεια τον μειωμένο ωφέλιμο ρυθμό μετάδοσης. Ως εκ τούτου, για τη βελτιστοποίηση του τηλεπικοινωνιακού συστήματος, είναι κρίσιμο το E_p, και εν γένει η διασπορά του σφάλματος πρόβλεψης, να λαμβάνει χαμηλές τιμές.

Ακολούθως, παρουσιάζονται σε μορφή γραφημάτων οι μετρικές επίδοσης των μοντέλων για κάθε κανάλι, παραμετροποιημένες ως προς το πλήθος των χαρακτηριστικών, τόσο για επιλογή, όσο και για μετασχηματισμό χαρακτηριστικών.

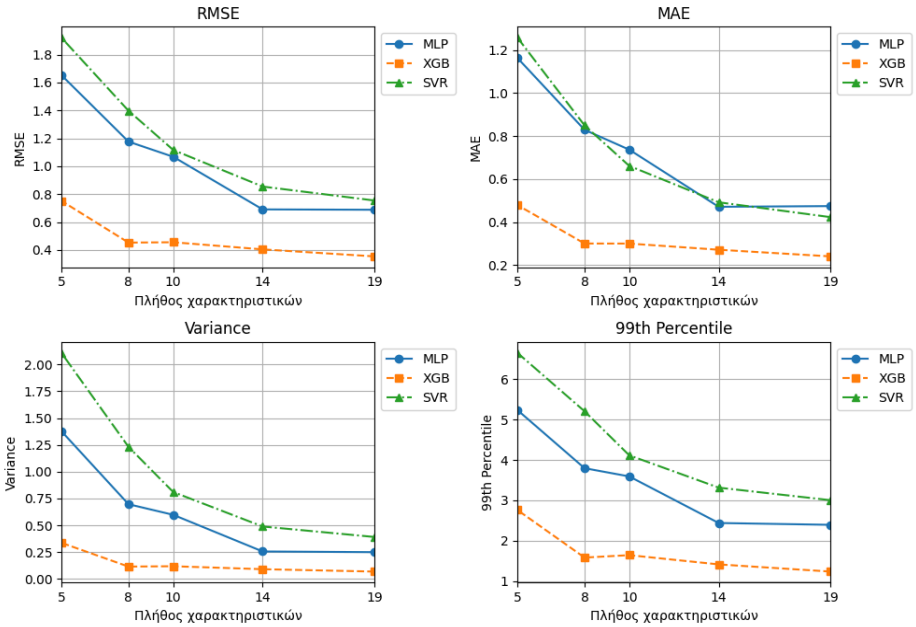
Κανάλι RR

- *Μείωση διάστασης με επιλογή χαρακτηριστικών*



***Εικόνα 3.6**: Γραφική απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (RR - Επιλογή χαρακτηριστικών)*

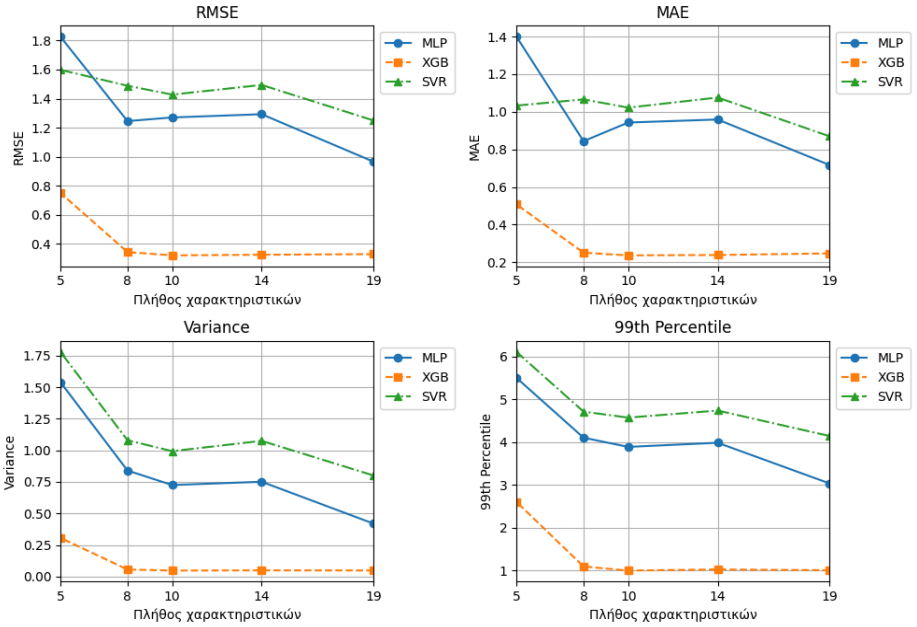
- *Μείωση διάστασης με PCA*



***Εικόνα 3.7**: Απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (RR - PCA)*

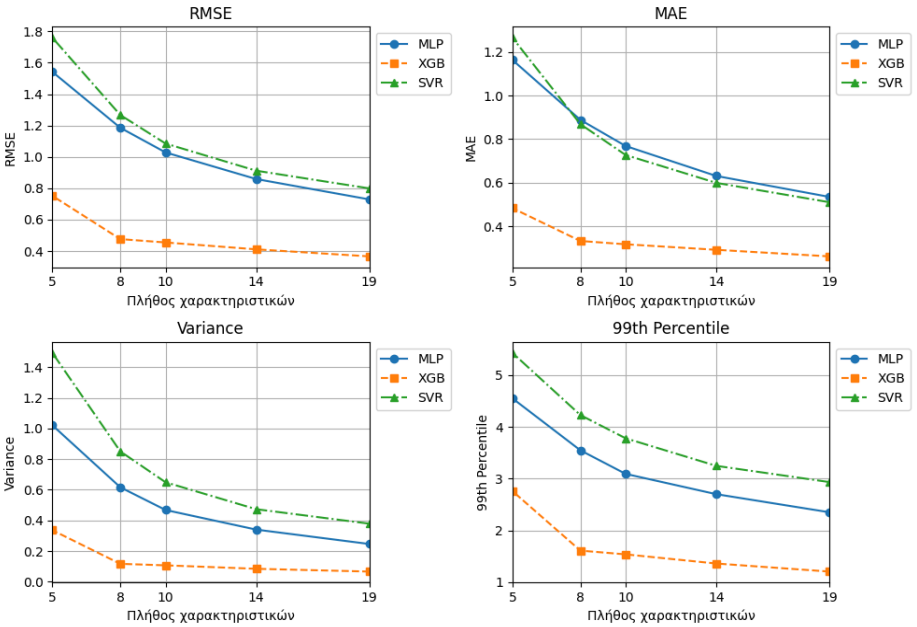
Κανάλι LR

- *Μείωση διάστασης με επιλογή χαρακτηριστικών*



***Εικόνα 3.8**: Γραφική απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (LR - Επιλογή χαρακτηριστικών)*

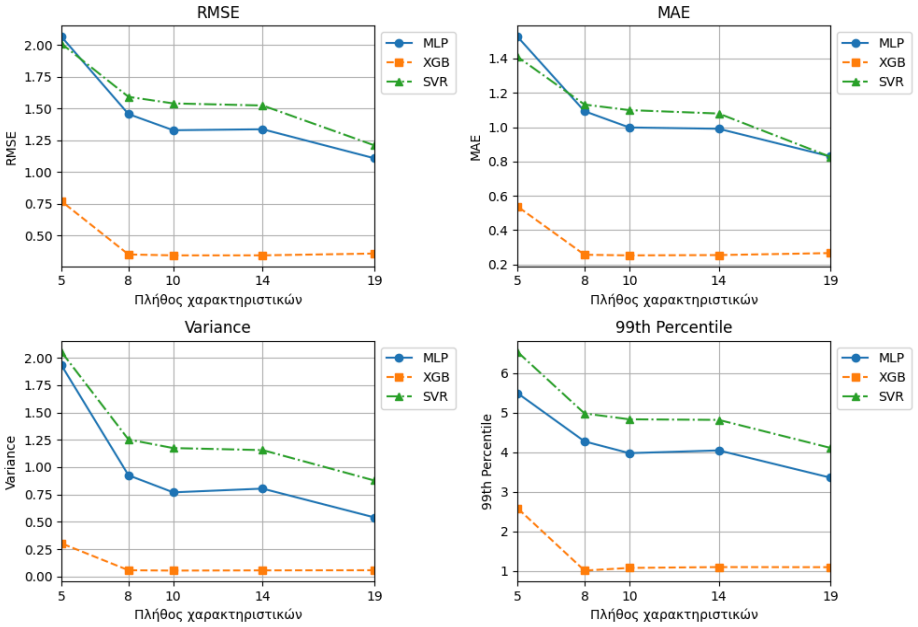
- *Μείωση διάστασης με PCA*



***Εικόνα 3.9**: Απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (LR - PCA)*

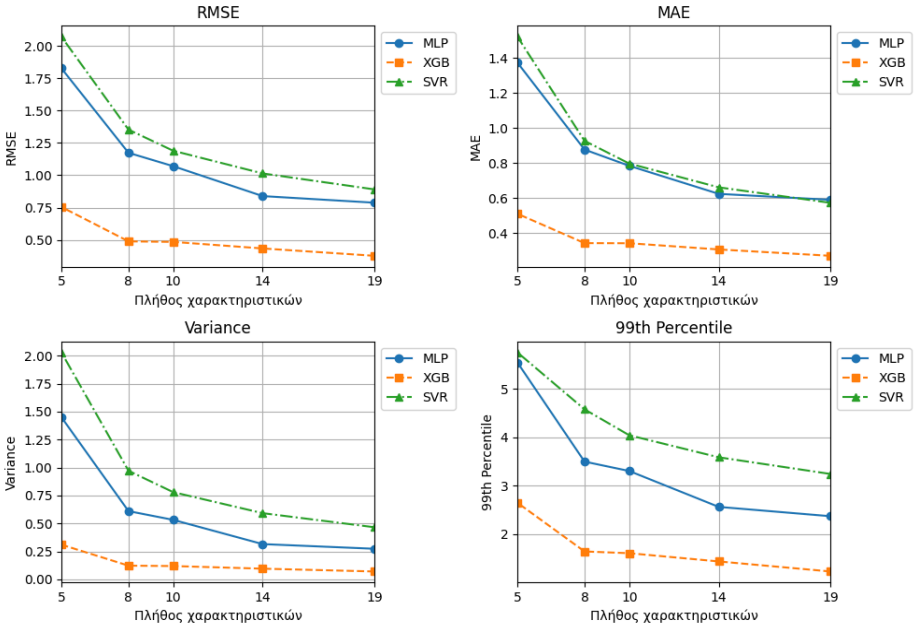
Κανάλι RL

- *Μείωση διάστασης με επιλογή χαρακτηριστικών*



***Εικόνα 3.10**: Γραφική απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (RL - Επιλογή χαρακτηριστικών)*

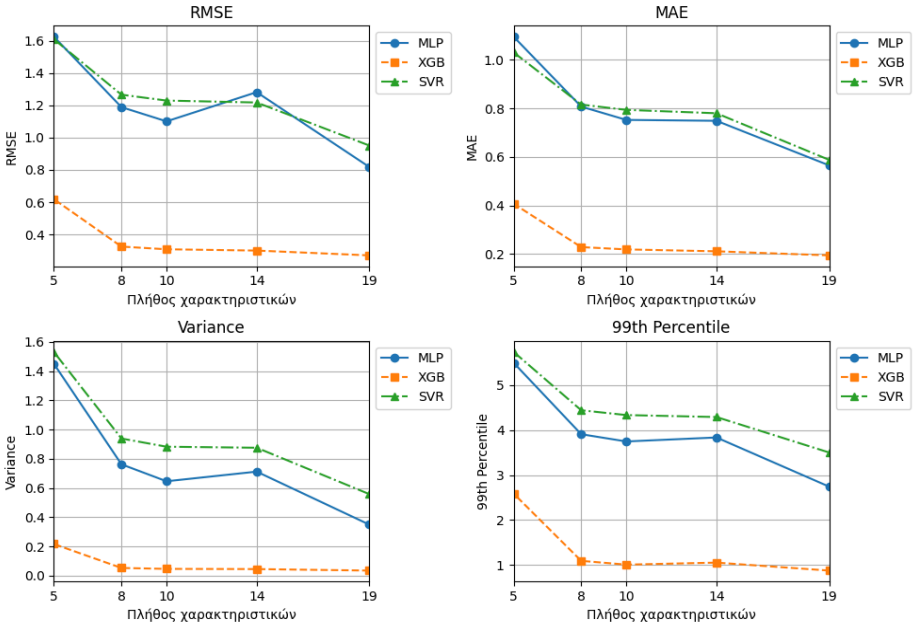
- *Μείωση διάστασης με PCA*



***Εικόνα 3.11**: Απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (RL - PCA)*

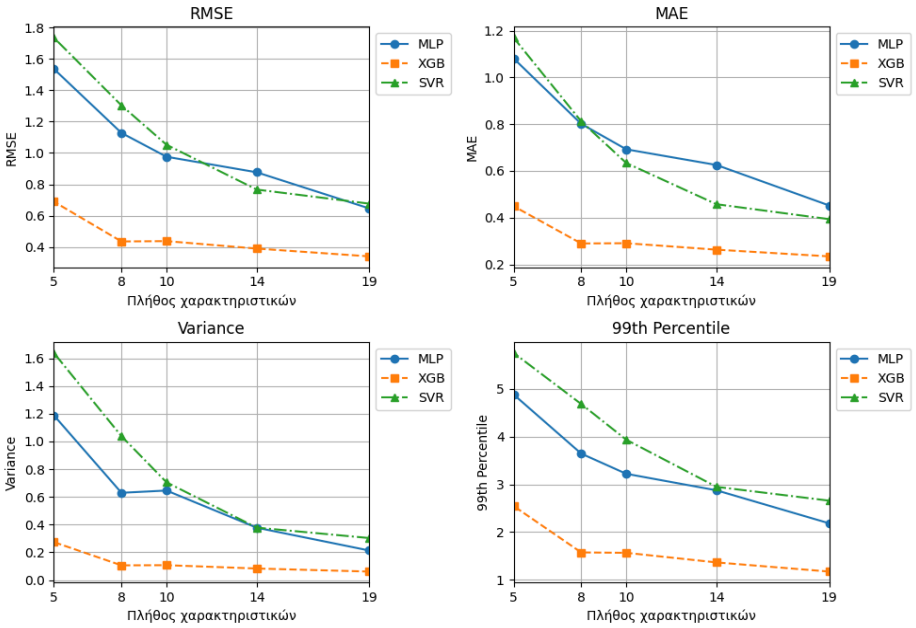
Κανάλι LL

- *Μείωση διάστασης με επιλογή χαρακτηριστικών*



***Εικόνα 3.12**: Γραφική απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (LL - Επιλογή χαρακτηριστικών)*

- *Μείωση διάστασης με PCA*



***Εικόνα 3.13**: Απεικόνιση μετρικών επίδοσης συναρτήσει του πλήθους χαρακτηριστικών (LL - PCA)*

Παρουσιάζονται, επιπλέον, οι δείκτες επίδοσης που επιτυγχάνουν τα μοντέλα στη βέλτιστή τους εκδοχή.

I) Extreme Gradient Boosting (XGB) - 19 χαρακτηριστικά, χωρίς PCA

Μετρική	**RR**	**LR**	**RL**	**LL**
MAE	0.187	0.247	0.267	0.195
RMSE	0.260	0.330	0.359	0.270
Variance	0.0327	0.048	0.058	0.035
99^{th} Percentile	0.838	1.001	1.097	0.879

Πίνακας 3.4: *Μετρικές επίδοσης βέλτιστου XGB*

II) Πολυστρωματικό Perceptron (MLP) - 19 χαρακτηριστικά, με PCA

Μετρική	**RR**	**LR**	**RL**	**LL**
MAE	0.474	0.535	0.589	0.452
RMSE	0.688	0.729	0.788	0.646
Variance	0.249	0.246	0.273	0.213
99^{th} Percentile	2.392	2.350	2.372	2.182

Πίνακας 3.5: *Μετρικές επίδοσης βέλτιστου MLP*

III) Παλινδρομητής Διανυσμάτων Υποστήριξης (SVR) - 19 χαρακτηριστικά, με PCA

Μετρική	**RR**	**LR**	**RL**	**LL**
MAE	0.423	0.51	0.572	0.394
RMSE	0.755	0.7993	0.89	0.677
Variance	0.391	0.3786	0.466	0.303
99^{th} Percentile	3.003	2.933	3.250	2.660

Πίνακας 3.6: *Μετρικές επίδοσης βέλτιστου SVR*

- *Σχολιασμός αποτελεσμάτων:*

Παρατηρούμε, αρχικά, ότι τα μοντέλα, στη βέλτιστή τους εκδοχή, παρουσιάζουν εξαιρετικές επιδόσεις, τόσο σε όρους απόλυτου σφάλματος ακρίβειας, όσο και σε όρους διασποράς, για αμφότερα τα copolarized και crosspolarized κανάλια. Η τιμή του μέσου

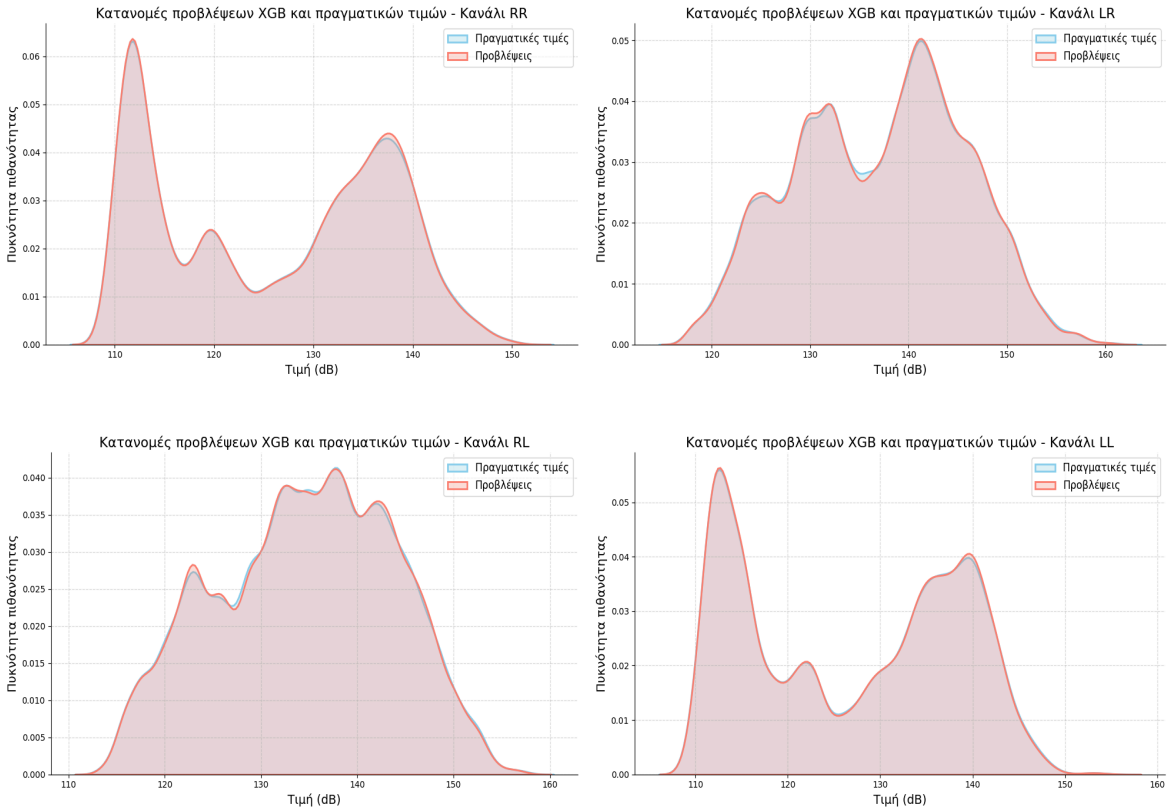
απολύτου σφάλματος (MAE) κυμαίνεται έως και τα 0.6dB, ενώ η τιμή του 99-οστού percentile φτάνει κατά μέγιστο τα 3.1dB. Αυτό σημαίνει ότι, για οποιοδήποτε εκ των τριων αρχιτεκτονικών, μπορούμε να πετύχουμε ένα διάστημα εμπιστοσύνης της τάξης του 99% με κόστος μόλις ένα επιπλέον περιθώριο των 3.1dB. Το περιθώριο αυτό είναι σχεδόν αμελητέο, αν το συγκρίνουμε με την τάξη μεγέθους των απωλειών διαδρομής (βλ. Εικόνες 3.1-3.4). Εστιάζοντας, δε, στο μοντέλο XGB, το οποίο επιτυγχάνει και τη βέλτιστη επίδοση ανάμεσα στα τρία, τότε διαπιστώνουμε ότι το μέσο σφάλμα και το 99-οστό percentile μειώνονται ακόμα περισσότερο, λαμβάνοντας τιμές της τάξης των 0.25dB και 1dB αντίστοιχα. Η ιδιαίτερα εύρωστη, αυτή, συμπεριφορά υποδηλώνει ότι τα μοντέλα μηχανικής μάθησης έχουν κατ' αρχήν τη δυνατότητα να παράγουν ιδιαίτερα αξιόπιστες προβλέψεις των απωλειών διαδρομής, ικανοποιώντας μάλιστα τις δυνητικά αυστηρές προδιαγραφές ενός τηλεπικοινωνιακού συστήματος. Όσον αφορά τα μοντέλα MLP και SVR, συγκρίνοντας μεταξύ των δύο διαπιστώνουμε ότι το MLP επιτυγχάνει εν γένει καλύτερες επιδόσεις, ιδίως στους δείκτες της διασποράς του σφάλματος. Η υπεροχή του μπορεί πιθανώς να αποδοθεί στην από κοινού βέλτιστη σχεδίαση των δικτύων διάκρισης και παλινδρόμησης, η οποία έρχεται σε αντίθεση με την αρχιτεκτονική του SVR, όπου η σχεδίαση των δύο συνιστωσών είναι ανεξάρτητη.

Συσχετίζοντας τώρα τους δείκτες με τη διάσταση του προβλήματος, συνάγουμε ότι η επίδοση των μοντέλων φθίνει σχεδόν ανεπαίσθητα με τη μείωση της διάστασης, ιδίως έως και τα 8 χαρακτηριστικά. Η συμπεριφορά αυτή ήταν αναμενόμενη, καθότι, από τον Πίνακα 3.3, προκύπτει ότι η διασπορά των μετασχηματισμένων δεδομένων διάστασης ίσης με 8 υπερβαίνει το 90% της αρχικής. Ως εκ τούτου, μπορούμε να ισχυριστούμε ότι η εγγενής πολυπλοκότητα της συνάρτησης προς εκτίμηση είναι τέτοια ώστε να απαιτεί πολύ λιγότερα χαρακτηριστικά από τα αρχικά 19. Η πιο εύρωστη συμπεριφορά παρατηρείται εκ νέου για το XGB, όπου για μείωση των χαρακτηριστικών από 19 σε 5, έχουμε αύξηση του MAE και του 99-οστού percentile κατά μόλις 0.5dB και 1.5dB αντίστοιχα.

Συγκρίνοντας τις μεθόδους επιλογής και μετασχηματισμού χαρακτηριστικών, διαπιστώνουμε ότι τα μοντέλα νευρωνικών δικτύων (SVR και MLP) καταγράφουν υψηλότερες επιδόσεις για μείωση διάστασης μέσω PCA σε σχέση με την απλή αφαίρεση χαρακτηριστικών, όπως και αναμενόταν. Αντιθέτως, στην περίπτωση του XGB καταγράφονται ελαφρώς χειρότερες επιδόσεις για μείωση με PCA, οι οποίες όμως είναι αμελητέες (τάξης 0.05dB για το MAE και 0.3dB για το 99-οστό percentile) και άρα δεν χρήζουν περαιτέρω διερεύνησης.

Τέλος, παρατηρούμε εν γένει καλύτερες επιδόσεις στα copolar κανάλια έναντι των crosspolar, γεγονός αναμενόμενο, καθότι, όπως παρατηρούμε και στις Εικόνες 3.1-3.5, οι απώλειες διαδρομής στη δεύτερη περίπτωση παρουσιάζουν μεγαλύτερη στοχαστικότητα

Επιπλέον στα παραπάνω, παρουσιάζεται γραφικά η κατανομή των προβλέψεων του βέλτιστου μοντέλου (XGB για πλήρη διάσταση) σε σχέση με την κατανομή των πραγματικών τιμών, καθότι και η μεταξύ τους απόσταση Kullback - Leibler. Τόσο οπτικά, όσο και αριθμητικά, οι διαφορές που εντοπίζονται είναι ανεπαίσθητες, δείγμα της ποιότητας των παραγόμενων προβλέψεων.



***Εικόνα 3.14**: Κατανομές προβλέψεων βέλτιστου μοντέλου (XGB) και κατανομές πραγματικών τιμών για τα 4 κανάλια*

- *Απόσταση Kullback - Leibler μεταξύ της κατανομής προβλέψεων και της πραγματικής κατανομής για κάθε κανάλι*

Κανάλι	KL Απόσταση
RR	8.094×10^{-5}
RL	1.597×10^{-4}
LR	1.439×10^{-4}
LL	9.229×10^{-5}

***Πίνακας 3.7:** Απόσταση KL μεταξύ της κατανομής προβλέψεων βέλτιστου XGB και της πραγματικής κατανομής για τα 4 κανάλια*

- *Πρόβλεψη απωλειών διαδρομής βάσει γεωμετρικών χαρακτηριστικών*

Εμβαθύνοντας περαιτέρω στο πρόβλημα της πρόβλεψης των απωλειών διαδρομής, έχει ενδιαφέρον να μελετήσουμε τις επιδόσεις των μοντέλων εάν από τα χαρακτηριστικά εισόδου αφαιρέσουμε πλήρως τα κανάλια ισχύος. Στην περίπτωση αυτή, συσχετίζουμε τις

απώλειες αποκλειστικά με τη γεωμετρία της τροχιάς. Οι μετρικές επίδοσης που κατέγραψαν τα μοντέλα παρουσιάζονται παρακάτω:

I) Extreme Gradient Boosting (XGB)

Μετρική	**RR**	**LR**	**RL**	**LL**
MAE	0.161	0.229	0.229	0.174
RMSE	0.226	0.307	0.307	0.246
Variance	0.025	0.042	0.042	0.030
99^{th} Percentile	0.729	0.916	0.919	0.785

Πίνακας 3.8: *Μετρικές επίδοσης XGB για γεωμετρικά χαρακτηριστικά εισόδου*

II) Πολυστρωματικό Perceptron (MLP)

Μετρική	**RR**	**LR**	**RL**	**LL**
MAE	0.832	1.053	1.226	0.853
RMSE	1.275	1.480	1.653	1.280
Variance	0.933	0.985	1.230	0.893
99^{th} Percentile	4.443	4.797	5.157	4.291

Πίνακας 3.9*: Μετρικές επίδοσης MLP για γεωμετρικά χαρακτηριστικά εισόδου*

III) Παλινδρομητής Διανυσμάτων Υποστήριξης (SVR)

Μετρική	**RR**	**LR**	**RL**	**LL**
MAE	1.167	1.489	1.605	1.159
RMSE	1.880	2.1874	2.368	1.863
Variance	2.171	2.569	3.030	2.127
99^{th} Percentile	6.734	7.169	7.630	6.505

Πίνακας 3.10*: Μετρικές επίδοσης SVR για γεωμετρικά χαρακτηριστικά εισόδου*

Παρατηρούμε, ότι, για τα μοντέλα νευρωνικών δικτύων, οι επιδόσεις τους είναι εν γένει χειρότερες σε σχέση με τις αντίστοιχες για δομημένη μείωση διάστασης σε 10. Αυτό είναι αναμενόμενο, καθότι θεωρούμε ότι τα κανάλια ισχύος περιέχουν αρκετή πληροφορία για την έξοδο και επιπλέον, η διαδικασία αφαίρεσης χαρακτηριστικών εδώ δεν χαρακτηρίζεται από κάποια βελτιστότητα. Ωστόσο, για την περίπτωση του XGB, διαπιστώνουμε αναπάντεχα ότι οι επιδόσεις του είναι ανώτερες ακόμα και σε σχέση με την περίπτωση της πλήρους διάστασης. Η (ελαφρά) αυτή βελτίωση μπορεί να αποδοθεί στην έντονη

στοχαστικότητα που παρουσιάζει η εξάρτηση των απωλειών διαδρομής από την ισχύ. Καθότι, λοιπόν, η ισχύς λήψης ενσωματώνει μεταβολές μικρής κλίμακας λόγω πολυδιαδρομικής διάδοσης, είναι πιθανό για παρόμοιες τιμές των απωλειών διαδρομής, οι τιμές της να παρουσιάζουν μεγάλες αποκλίσεις. Το γεγονός αυτό είναι πιθανό να "μπερδεύει" σε κάποιο βαθμό ένα μοντέλο μάθησης, οδηγώντας έτσι σε δυνητικά υποβέλτιστες επιδόσεις.

Συμπερασματικά, από τους παραπάνω δείκτες επίδοσης, συνάγουμε το πολύ χρήσιμο συμπέρασμα ότι οι απώλειες διαδρομής στη ζεύξη μπορούν να προβλεφθούν με μεγάλη ακρίβεια, λαμβάνοντας υπόψη αποκλειστικά τα δεδομένα της τροχιάς του αεροπλοίου σε σχέση με τη θέση του τερματικού.

Κεφάλαιο 4

Πρόβλεψη των απωλειών διάδοσης μικρής κλίμακας με μοντέλα μηχανικής μάθησης

4.1 Εισαγωγή

Οι απώλειες μικρής κλίμακας (ή διαλείψεις μικρής κ υπολογίζονται σύμφωνα με την (3.2) ως εξής:

$$L_{SS}(t) = L(t) - L_{LS}(t) \tag{4.1}$$

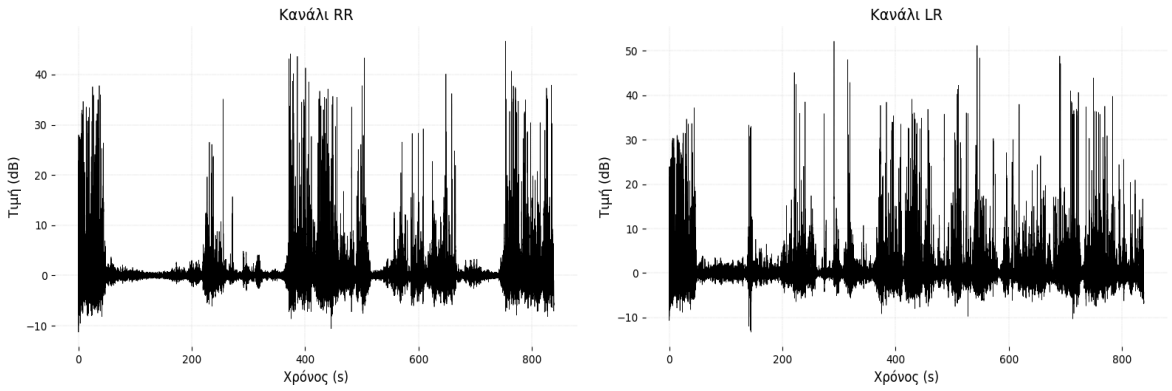
Εφαρμόζοντας τις (3.1), (3.6), παίρνουμε ότι:

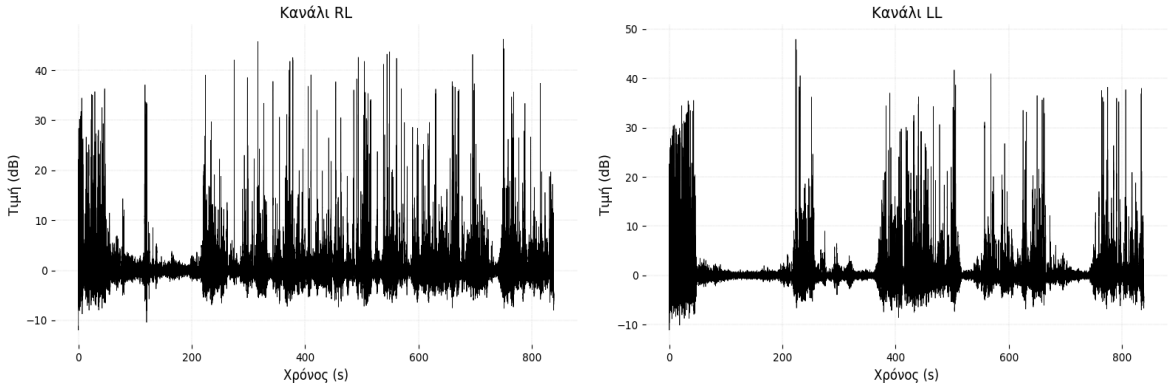
$$L_{SS}(t) = P_T - P_R'(t) - L_{LS}(t) \tag{4.2}$$

Εφαρμόζοντας την (4.2) χωριστά για κάθε κανάλι, προκύπτουν οι ακόλουθες σχέσεις, βάσει των οποίων υπολογίζονται οι αριθμητικές τιμές των δειγμάτων εκπαίδευσης για τις διαλείψεις μικρής κλίμακας:

$$L_{SS,RR}(t) = P_T - P_R'(t) - L_{LS,RR}(t) \tag{4.3}$$
$$L_{SS,LR}(t) = P_T - P_R'(t) - L_{LS,LR}(t) \tag{4.4}$$
$$L_{SS,RL}(t) = P_T - P_R'(t) - L_{LS,RL}(t) \tag{4.5}$$
$$L_{SS,LL}(t) = P_T - P_R'(t) - L_{LS,LL}(t) \tag{4.6}$$

Παρακάτω, παρουσιάζονται γραφικά οι χρονοσειρές των διαλείψεων μικρής κλίμακας για κάθε κανάλι:





***Εικόνα 4.1**: Χρονοσειρές διαλείψεων μικρής κλίμακας για τα 4 κανάλια*

Με απλή οπτική επισκόπηση των παραπάνω γραφικών παραστάσεων, διαπιστώνουμε ότι οι διαλείψεις μικρής κλίμακας μεταβάλλονται στον χρόνο με μεγάλη στοχαστικότητα. Αυτό είναι αναμενόμενο ήδη από τον φυσικό μηχανισμό που τις παράγει, την υπέρθεση, δηλαδή, πολυδιαδρομικών συνιστωσών του ίδιου σήματος, η οποία παρουσιάζει εξαιρετική ευαισθησία στη φάση των συνιστωσών αυτών.

Η στοχαστικότητα των διαλείψεων καθιστά ιδιαίτερα δύσκολη την πρόβλεψη των επόμενων χρονικών τους δειγμάτων, ιδίως για μεγάλα *μήκη πρόβλεψης (prediction lengths)*. Σε τηλεπικοινωνιακές εφαρμογές, ωστόσο, το ενδιαφέρον δεν επικεντρώνεται στην ακριβή πρόβλεψη της ακολουθίας τιμών των διαλείψεων σε ένα επόμενο παράθυρο N δειγμάτων, αλλά στην πρόβλεψη του *βάθους* των διαλείψεων στο παράθυρο αυτό. Πράγματι, μια ασύρματη ζεύξη σχεδιάζεται ώστε να μπορεί να ανεχθεί απώλειες μικρής κλίμακας έως και μια ορισμένη, μέγιστη στάθμη, χωρίς να οδηγηθεί σε διακοπή. Η στάθμη αυτή ονομάζεται *περιθώριο διαλείψεων (Fade Margin - FM)*. Παράλληλα, η ζεύξη προδιαγράφεται ώστε να είναι διαθέσιμη σε ποσοστό του χρόνου κατ' ελάχιστον ίσο με p (*πιθανότητα διαθεσιμότητας*). Θα πρέπει λοιπόν, για μια δεδομένη κατάσταση του διαύλου, το περιθώριο διαλείψεων να καθοριστεί με τρόπο τέτοιο ώστε να ισχύει

$$Pr(L_{SS}(t) < FM(t)) \geq p \qquad (4.7)$$

Θεωρώντας ένα παράθυρο χρονικών δειγμάτων $[t, t+N]$, η παραπάνω σχέση θα ισχύει εξ ορισμού για $FM(t) = L_{ss,p}(t;N)$, όπου $L_{ss,p}(t;N)$ το p-th percentile των διαλείψεων μικρής κλίμακας για το παράθυρο $[t, t+N]$. Θα ποσοτικοποιήσουμε, άρα, το βάθος των διαλείψεων σε ένα χρονικό παράθυρο W μέσω του p-οστό percentile του W.

Θα σχεδιάσουμε, λοιπόν, μοντέλα μηχανικής μάθησης με στόχο την πρόβλεψη του βάθους των διαλείψεων μικρής κλίμακας, διατυπώνοντας το πρόβλημα σε δύο εκδοχές:

1) *Διακριτοποιημένη εκδοχή*: Το μοντέλο εκπαιδεύεται να *ταξινομεί* το βάθος των διαλείψεων ενός μελλοντικού παράθυρου μήκους N σε 3 κλάσεις: Low, Mid και $Deep$. Η κατηγοριοποίηση πραγματοποιείται ως εξής:

   Low εάν $L_{ss,p}(t; N) < 2$
   Mid εάν $2 \leq L_{ss,p}(t; N) < 15$
   $Deep$ εάν $L_{ss,p}(t; N) \geq 15$

   Επιλέγουμε $p = 95$.

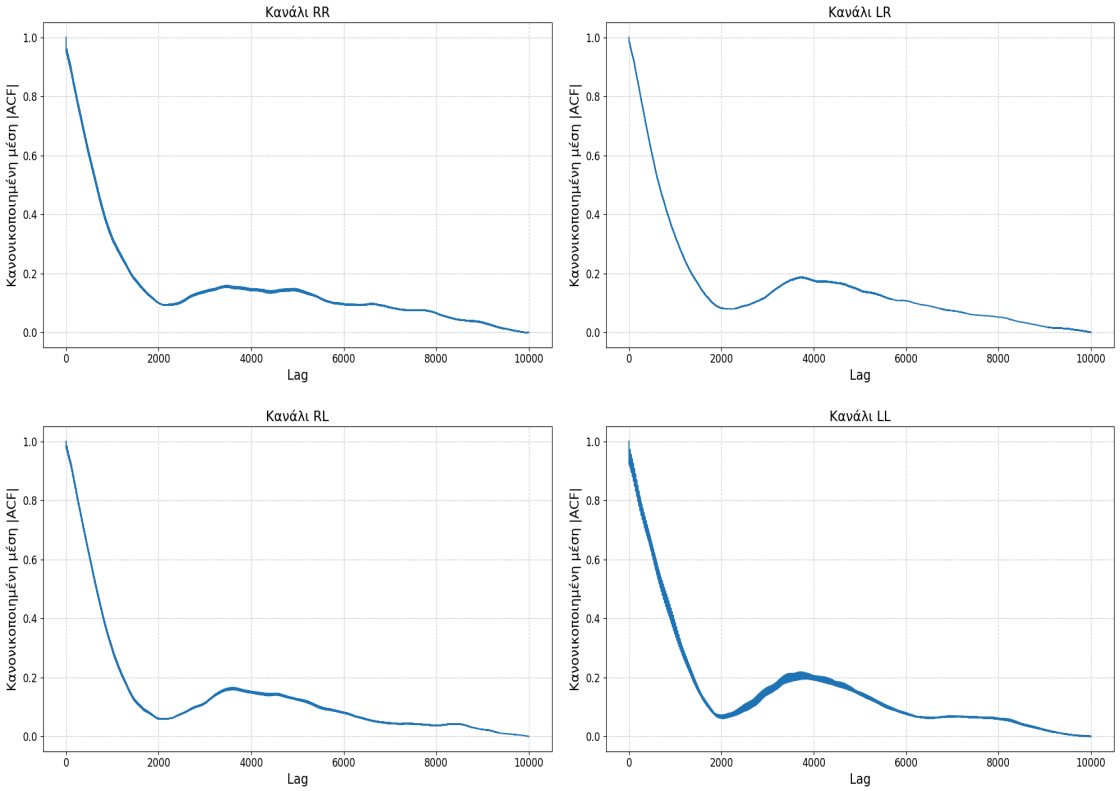
2) *Συνεχής εκδοχή*: Η καταγεγραμμένη χρονοσειρά των διαλείψεων μικρής κλίμακας μετασχηματίζεται σε μια χρονοσειρά από 99-οστά percentiles, τα οποία υπολογίζονται επάνω σε επικαλυπτόμενα παράθυρα 200 δειγμάτων της αρχικής χρονοσειράς. Κατόπιν, τα μοντέλα εκπαιδεύονται ώστε να προβλέπουν τα επόμενα δείγματα της χρονοσειράς των percentiles. Το βήμα επικάλυψης των χρονικών παραθύρων (stride) που αντιστοιχούν σε κάθε percentile είναι τέτοιο ώστε κάθε νέο δείγμα να αντιστοιχεί σε μελλοντικά 5ms (50 χρονικά δείγματα). Αν, δηλαδή, το i-οστό δείγμα percentile αντιστοιχεί σε παράθυρο $[i, i + 200]$, το (i+1)-οστό δείγμα percentile θα αντιστοιχεί σε παράθυρο $[i + 50, i + 250]$.

Ακολούθως, παρουσιάζονται η ροή σχεδίασης και τα αποτελέσματα για τις δύο εκδοχές του προβλήματος:

4.2 *Διακριτοποιημένη εκδοχή*

4.2.1 *Στατιστική ανάλυση των δεδομένων με στόχο τη μείωση διάστασης*

Ο στόχος μας εδώ είναι να προσδιορίσουμε το μήκος συμφραζόμενου N_c, το μήκος δηλαδή της ακολουθίας δειγμάτων εισόδου $(X_{t-N_c}, \ldots, X_t)$ που θα πρέπει να λάβει το σύστημα ώστε να προβλέψει το βάθος των διαλείψεων στο παράθυρο $[t, t + N]$. Όπως επεξηγήθηκε στο πρώτο κεφάλαιο, το N_c καθορίζεται βάσει της συνάρτησης αυτοσυσχέτισης της χρονοσειράς. Καθότι όμως η χρονοσειρά για κάθε κανάλι δεν είναι στάσιμη, θα υπολογίσουμε τη *μέση* απόλυτη συνάρτηση αυτοσυσχέτισης, για καθυστέρηση (lag) έως και 10000.



Εικόνα 4.2: *Κανονικοποιημένη μέση ACF των διαλείψεων μικρής κλίμακας για τα 4 κανάλια*

Βάσει των παραπάνω διαγραμμάτων, θα επιλέξουμε $N_c = 1000$. Για την τιμή αυτή, η ACF ισούται περίπου με το 0.4 της μέγιστης. Υποδειγματοληπτώντας, τέλος, τις χρονοσειρές κατά έναν παράγοντα 10, ώστε να καταστούν πιο διαχειρίσιμες υπολογιστικά, προκύπτει ότι $N_{c,downsampled} = 100$.

4.2.2 *Επιλογή, εκπαίδευση και ρύθμιση υπερπαραμέτρων μοντέλων*

Επιλέγονται μοντέλα για δυναμικά προβλήματα και συγκεκριμένα LSTMs και Stacked RNNs. Η προσαρμογή των αρχιτεκτονικών παλινδρόμησης στη διακριτή φύση του προβλήματος πραγματοποιείται με άμεσο τρόπο, τοποθετώντας ένα *επίπεδο softmax* μετά τα γραμμικά επίπεδα εξόδου (βλ. Σχέσεις 1.26 και 1.36). Η συνάρτηση $softmax(\cdot)$ εφαρμόζεται κατά στοιχείο (element-wise) σε ένα διάνυσμα $y \in \mathbb{R}^C$ και ορίζεται ως

$$softmax(y_i) = \frac{exp(y_i)}{\sum_{j=1}^{C} exp(y_j)} \tag{4.8}$$

Εν προκειμένω, ισχύει $C = 3$, όσες, δηλαδή, και οι κλάσεις. Ουσιαστικά, η $softmax$ κανονικοποιεί τα στοιχεία του y σε τιμές μεταξύ 0 και 1, ώστε να μπορούν να ερμηνευτούν ως *πιθανότητες*. Η κανονικοποίηση αυτή είναι απαραίτητη, καθότι επιθυμούμε το μοντέλο

να αποφασίζει την κλάση στην οποία απεικονίζεται το δείγμα εισόδου x εφαρμόζοντας τον *κανόνα απόφασης Bayes (Bayes' Decision Rule - BDR)*· προβλέποντας, δηλαδή, την a posteriori κατανομή των κλάσεων $Pr(c|x)$ και αποφασίζοντας την κλάση εκείνη για την οποία ισχύει $c^* = argmax_{c\in\{Low,\, Mid, Deep\}} Pr(c|x)$. Αν θεωρήσουμε ότι το διάνυσμα εξόδου του επιπέδου $softmax$ αναπαριστά την $Pr(c|x)$, τότε η εφαρμογή του BDR πραγματοποιείται άμεσα.

Προκειμένου να προσανατολίσουμε το μοντέλο στην εκμάθηση της $Pr(c|x)$, θα ορίσουμε ως συνάρτηση απωλειών τη *Weighted Categorical Cross Entropy (WCCE)*:

$$L_{WCCE}(y, \hat{y}) = -\sum_{i=1}^{C} w_i y_i log(\hat{y}_i) \tag{4.9}$$

Με $\hat{y}$ συμβολίζουμε το διάνυσμα εξόδου του στρώματος softmax, που παράγεται για μια κάποια είσοδο $x \in \mathbb{R}^{N_c}$. Στο x αντιστοιχεί ένα one hot vector y, με τιμή $[1, 0, 0]$ αν x ανήκει στη Low, $[0, 1, 0]$ αν ανήκει στη Mid και $[0, 0, 1]$ αν ανήκει στη $Deep$. Ουσιαστικά, το y αποτελεί το διάνυσμα - στόχο για το $\hat{y}$. Τα βάρη w_i υπολογίζονται ως $w_i = \frac{1}{f_i}$, όπου f_i είναι το ποσοστό των δειγμάτων της i-οστής κλάσης στο σύνολο εκπαίδευσης. Ως αλγόριθμος μάθησης επιλέγεται ο Adam.

Τα βάρη εισάγονται στη συνάρτηση απωλειών με στόχο την αντιστάθμιση της ανισορροπίας των κλάσεων, η οποία παρατηρείται έντονα στο πρόβλημα. Πράγματι, θεωρώντας ενδεικτικά το κανάλι RR, το ποσοστό των δειγμάτων που αντιστοιχούν στις κλάσεις Mid και Deep είναι μόλις 16.5% και 1.17% αντίστοιχα. Σταθμίζοντας, λοιπόν, τους όρους $y_i log(\hat{y}_i)$ με την αντίστροφη συχνότητα των κλάσεων, αυξάνουμε το κόστος των εσφαλμένων ταξινομήσεων των δειγμάτων Mid και Deep. Κατ' αυτόν τον τρόπο, αποτρέπουμε το μοντέλο από το να αναπτύξει προδιάθεση (bias) προς την κυρίαρχη κλάση Low και κατ' επέκταση να υποτιμήσει το βάθος των διαλείψεων.

Επιπρόσθετα στη στάθμισης της συνάρτησης απωλειών, εφαρμόζουμε τεχνητή εξισορρόπηση των κλάσεων στο σύνολο εκπαίδευσης, αφαιρώντας με τυχαίο τρόπο δείγματα της Low, έως ότου να πετύχουμε μια κατανομή της τάξης του 40-35-15. Η αφαίρεση δεδομένων σε ένα πρόβλημα μηχανικής μάθησης προφανώς δεν στοιχειοθετεί βέλτιστη τεχνική, αλλά αποτελεί μια ύστατη λύση που εφαρμόζεται σε περιπτώσεις ισχυρής ανισορροπίας, ιδίως όταν η λιγότερο εκπροσωπούμενη κλάση είναι και η πιο κρίσιμη.

Η ρύθμιση υπερπαραμέτρων πραγματοποιείται μέσω της τεχνικής του randomized search, για τις τυπικές τιμές των υπερπαραμέτρων που περιγράφονται στον Πίνακα 1.4. Για τους λόγους που αναλύθηκαν στην Παράγραφο 3.3, το σύνολο επικύρωσης προκύπτει από ένα απομονωμένο, συνεχές χρονικό διάστημα μήκους 500 χιλιάδων δειγμάτων της αρχικής ακολουθίας, το οποίο επιλέγεται ώστε να είναι κατά το δυνατόν αντιπροσωπευτικό των μέσων συνθηκών διάδοσης.

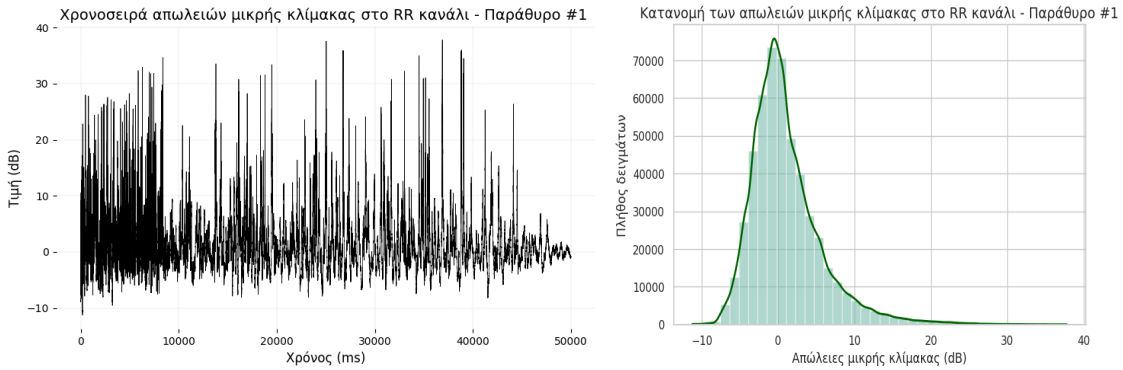
4.2.3 *Αξιολόγηση μοντέλων*

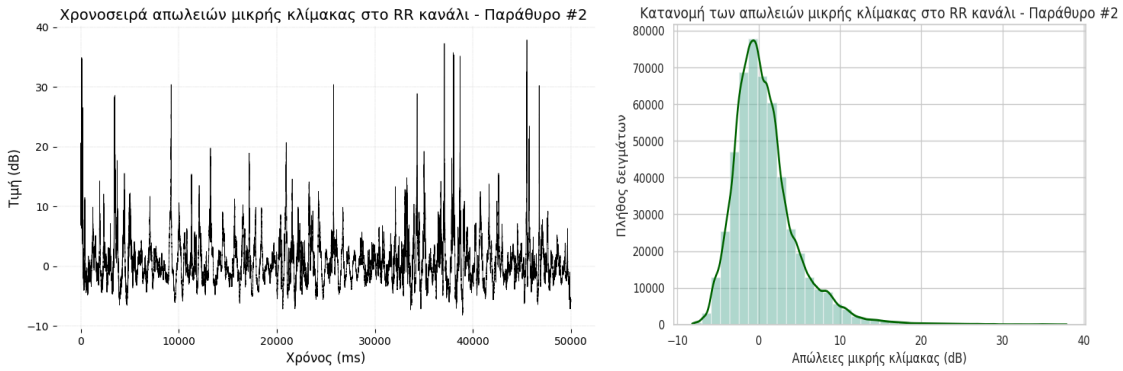
Η αξιολόγηση των μοντέλων πραγματοποιείται βάσει των ακόλουθων μετρικών:

Μετρική	Μαθηματικός Ορισμός	Επεξήγηση
Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	Το ποσοστό των δειγμάτων που ταξινομήθηκαν σωστά στο σύνολο των δειγμάτων, ανεξαρτήτως κλάσης.
Deep Fading Recall	$\text{Recall}_{\text{deep}} = \frac{TP_{\text{deep}}}{TP_{\text{deep}} + FN_{\text{deep}}}$	Η ανάκληση για την κλάση των βαθιών διαλείψεων. Μετρά την ικανότητα του μοντέλου να εντοπίζει σωστά τα δείγματα αυτής της κλάσης, η οποία αποτελεί και την πιο κρίσιμη για τη διαθεσιμότητα του συστήματος
Weighted Avg F1 Score	$\text{F1}_{\text{weighted}} = \sum_{i=1}^{C} \left(\frac{n_i}{\sum_j n_j} \cdot 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \right)$	Σταθμισμένος μέσος όρος του F1 score για όλες τις κλάσεις. Συνδυάζει Precision και Recall με βάση τη σχετική συχνότητα n_i της κάθε κλάσης.

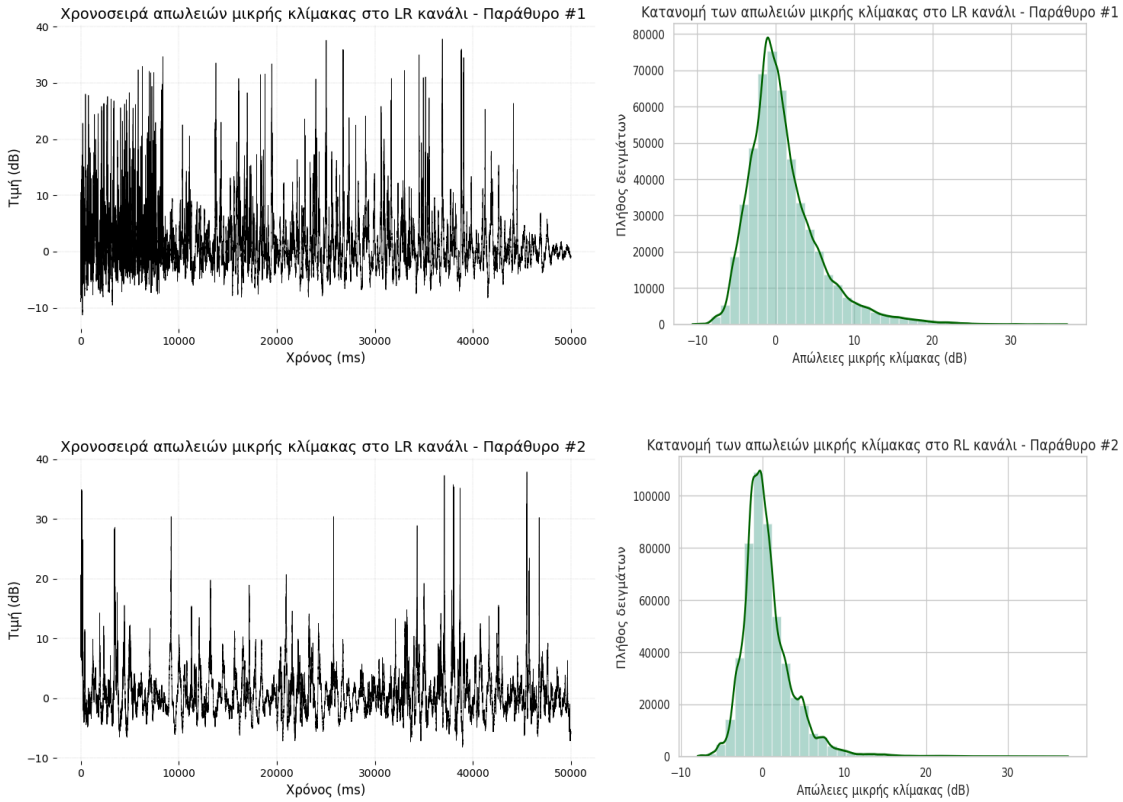
Πίνακας 4.1: Μετρικές επίδοσης μοντέλων ταξινόμησης των διαλείψεων

Παρατηρώντας την Εικόνα 4.1, διαπιστώνουμε οι χρονοσειρές είναι έντονα μη στάσιμες. Προκειμένου, λοιπόν, να αποτιμήσουμε τα μοντέλα με πλήρη και αξιόπιστο τρόπο, εκθέτοντάς τα σε όσο το δυνατόν μεγαλύτερη ποικιλία συνθηκών διάδοσης, θα πραγματοποιήσουμε την αξιολόγησή τους σε δύο παράθυρα μήκους 500 χιλιάδων δειγμάτων, ένα με διαλείψεις ισχυρής έντασης (Παράθυρο #1) και ένα με μέσης έντασης (Παράθυρο #2).

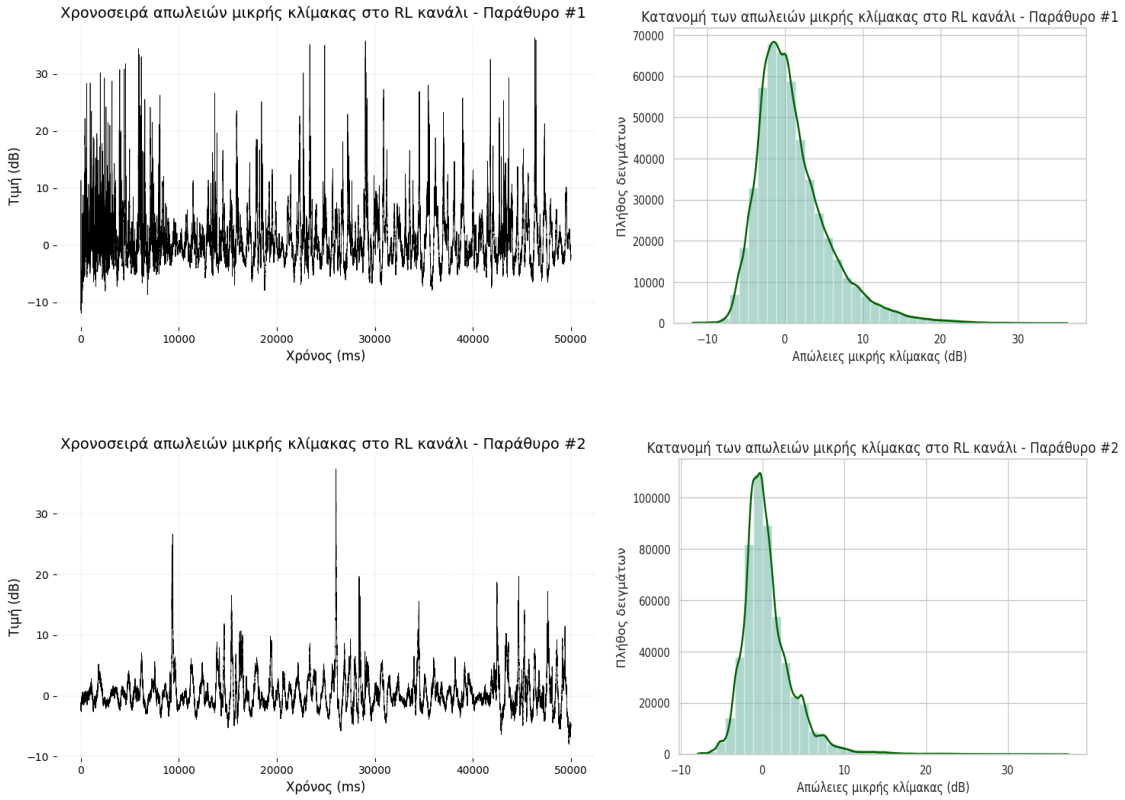




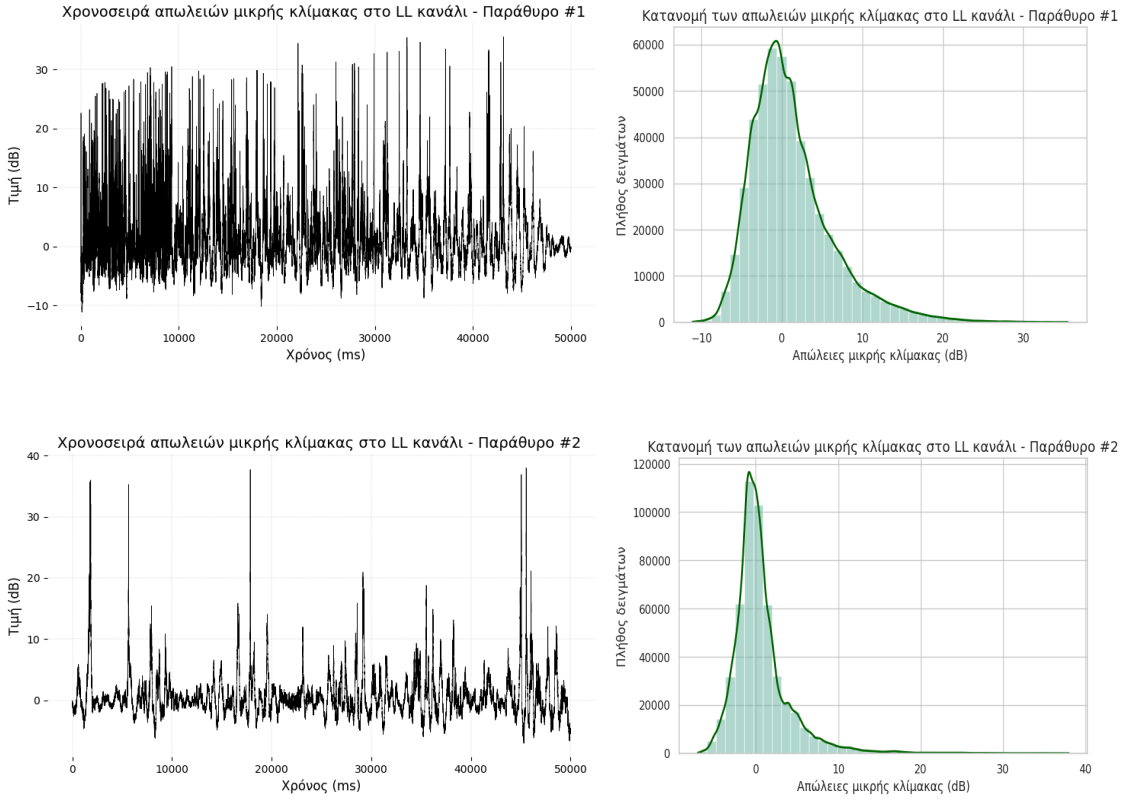
***Εικόνα 4.3**: Χρονοσειρά και ιστόγραμμα απωλειών μικρής κλίμακας για το κανάλι RR*



***Εικόνα 4.4**: Χρονοσειρά και ιστόγραμμα απωλειών μικρής κλίμακας για το κανάλι LR*



***Εικόνα 4.5**: Χρονοσειρά και ιστόγραμμα απωλειών μικρής κλίμακας για το κανάλι RL*

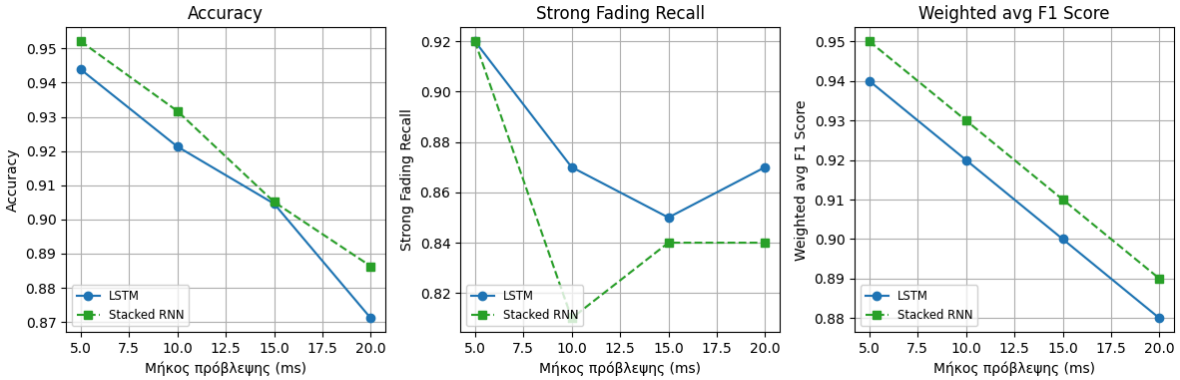


***Εικόνα 4.6**: Χρονοσειρά και ιστόγραμμα απωλειών μικρής κλίμακας για το κανάλι LL*

Ακολούθως, παρουσιάζονται οι γραφικές παραστάσεις των μετρικών επίδοσης που κατέγραψαν τα δύο μοντέλα, παραμετροποιημένες ως προς το μήκος πρόβλεψης.

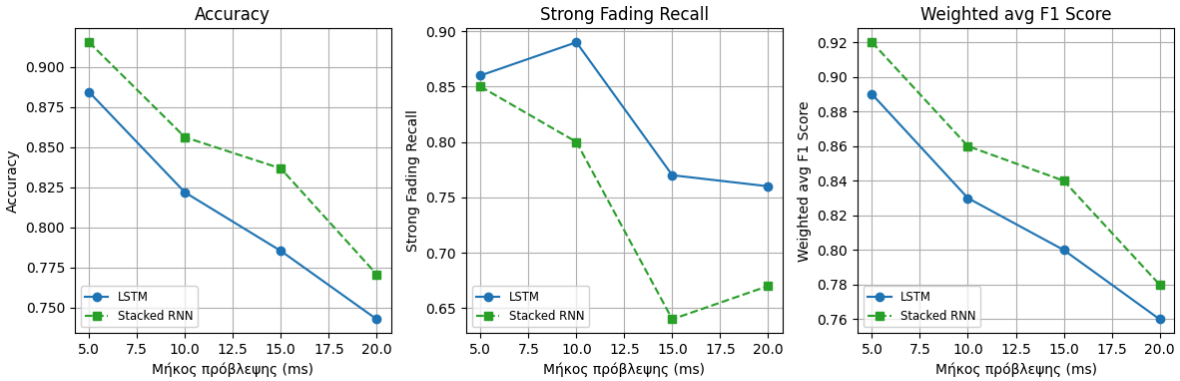
Κανάλι RR

- *Παράθυρο μέσων διαλείψεων*



***Εικόνα 4.7**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (RR - παράθυρο μέσων διαλείψεων)*

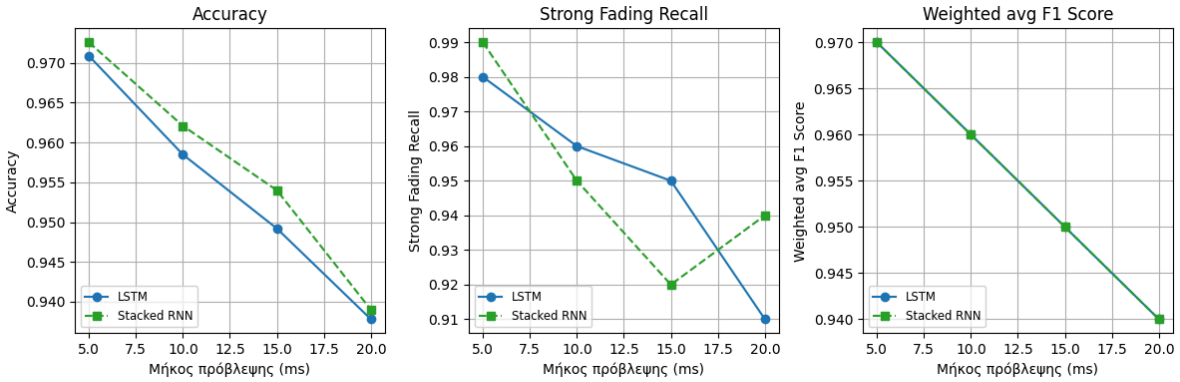
- *Παράθυρο ισχυρών διαλείψεων*



***Εικόνα 4.8**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (RR - παράθυρο ισχυρών διαλείψεων)*

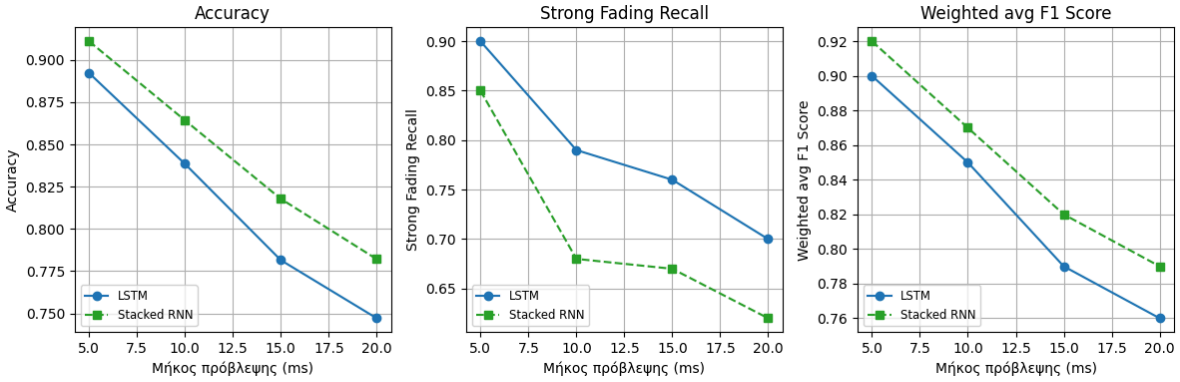
Κανάλι LR

- *Παράθυρο μέσων διαλείψεων*



***Εικόνα 4.9**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (LR - παράθυρο μέσων διαλείψεων)*

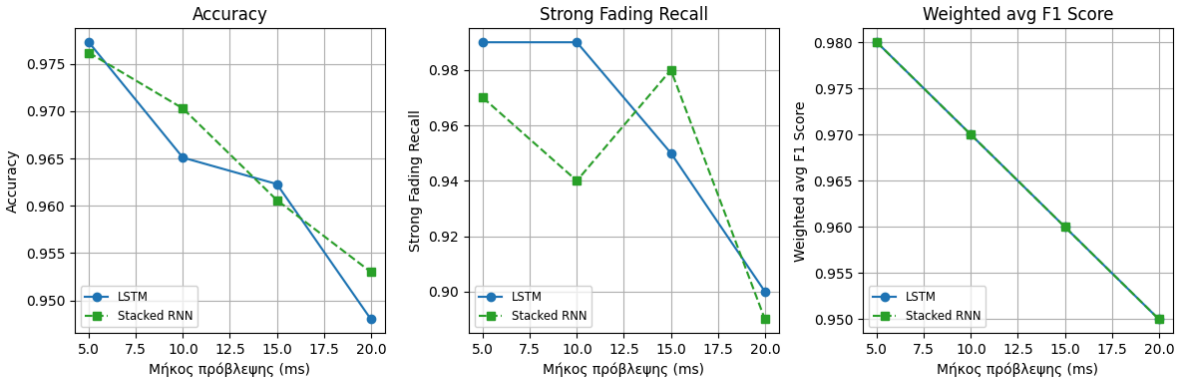
- *Παράθυρο ισχυρών διαλείψεων*



***Εικόνα 4.10**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (LR - παράθυρο ισχυρών διαλείψεων)*

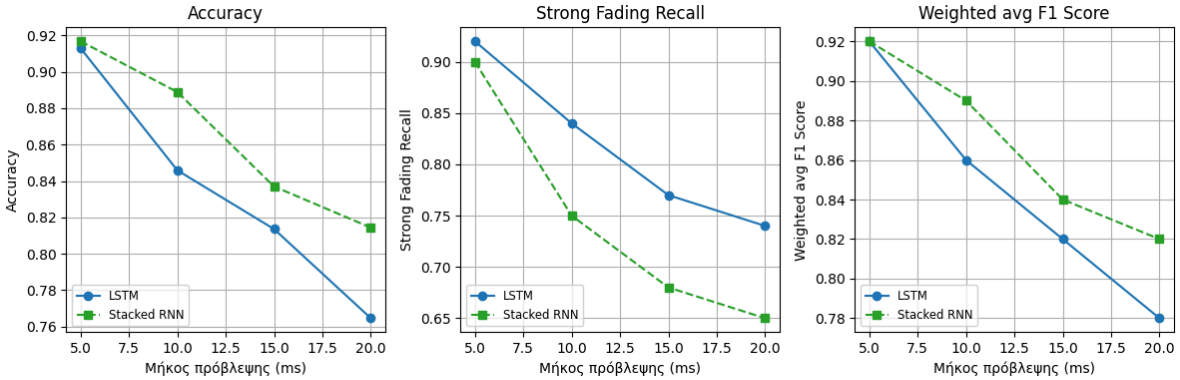
Κανάλι RL

- *Παράθυρο μέσων διαλείψεων*



***Εικόνα 4.11**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (RL - παράθυρο μέσων διαλείψεων)*

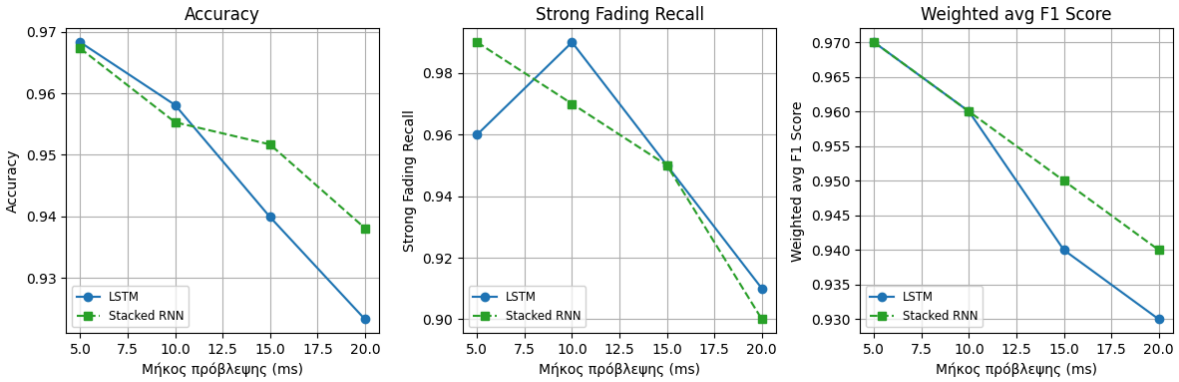
- *Παράθυρο ισχυρών διαλείψεων*



***Εικόνα 4.12**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (RL - παράθυρο ισχυρών διαλείψεων)*

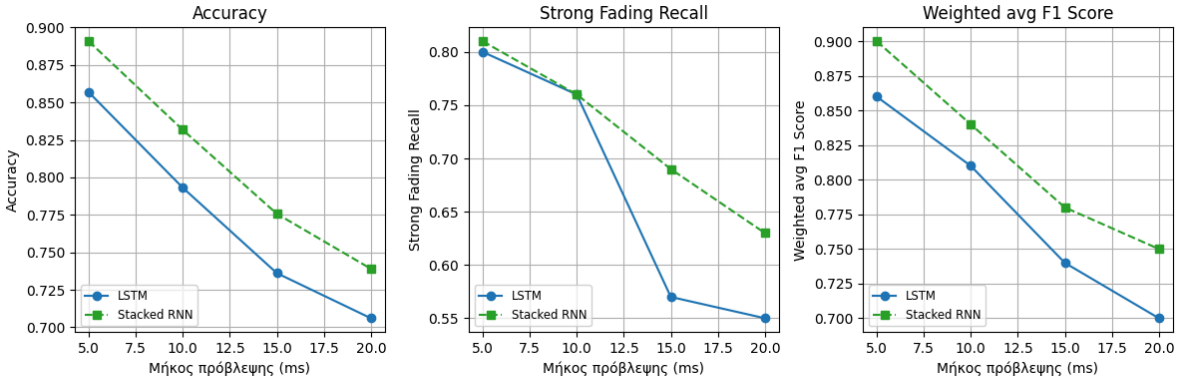
Κανάλι LL

- *Παράθυρο μέσων διαλείψεων*



***Εικόνα 4.13**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (LL - παράθυρο μέσων διαλείψεων)*

- *Παράθυρο ισχυρών διαλείψεων*



***Εικόνα 4.14**: Απεικόνιση μετρικών επίδοσης των μοντέλων ταξινόμησης διαλείψεων συναρτήσει του μήκους πρόβλεψης (LL - παράθυρο ισχυρών διαλείψεων)*

- *Σχολιασμός αποτελεσμάτων:*

Η πρώτη παρατήρηση που εξάγουμε συνίσταται στο ότι όλες οι μετρικές (accuracy, strong fading recall, weighted F1 score) μειώνονται προοδευτικά με την αύξηση του μήκους πρόβλεψης, γεγονός που υποδεικνύει τη δυσκολία της πρόβλεψης των διαλείψεων σε μεγαλύτερους χρονικούς ορίζοντες. Η φθίνουσα τάση, αυτή, είναι πιο έντονη για το παράθυρο των ισχυρών διαλείψεων: Πράγματι, ενώ για μήκος πρόβλεψης ίσο με 5ms οι μετρικές παρουσιάζουν εν γένει υψηλές τιμές (τάξης 86-92% για το accuracy, 80-92% για

το recall, 86-92% για το weighted avg f1 score), για μεγαλύτερους χρονικούς ορίζοντες εμφανίζουν ταχεία επιδείνωση, φτάνοντας τελικά σε επιδόσεις της τάξης <70%. Αντιθέτως, για το παράθυρο των μέσων διαλείψεων, οι επιδόσεις των μοντέλων μειώνονται μεν, παραμένουν όμως σταθερά σε υψηλές στάθμες, υπερβαίνοντας το 90% ακόμη και σε ορίζοντα 20ms.

Εστιάζοντας μεμονωμένα σε καθεμία εκ των μετρικών, έχει ενδιαφέρον να παρατηρήσουμε τις επιδόσεις των μοντέλων στην ανίχνευση της κλάσης των βαθέων διαλείψεων. Πράγματι, αν και η εκπροσώπησή της στο σύνολο δεδομένων (και κατ' επέκταση στις χρονοσειρές ελέγχου) είναι εξαιρετικά χαμηλότερη των άλλων δύο κλάσεων, οι τεχνικές αντιστάθμισης που εφαρμόσαμε (εξισορρόπηση του συνόλου εκπαίδευσης, στάθμιση της συνάρτησης απωλειών) είχαν ως αποτέλεσμα την ανίχνευση των θετικών δειγμάτων της (true positives) με ιδιαίτερα υψηλή ακρίβεια (recall από 85 έως και 99%), ιδίως για μικρότερα μήκη πρόβλεψης. Το γεγονός αυτό είναι ιδιαίτερα σημαντικό, λαμβάνοντας υπ' όψιν ότι οι βαθιές διαλείψεις είναι εκείνες που δοκιμάζουν τα όρια της αξιοπιστίας του συστήματος και, ως εκ τούτου, είναι κρίσιμο να ανιχνεύονται σωστά. Το αντιστάθμισμα βέβαια στην υψηλή ευαισθησία έναντι των βαθύτερων διαλείψεων είναι η εσφαλμένη ταξινόμηση πολλών δειγμάτων της κλάσης Low στη Mid και της Mid στη High. Αυτό επιφέρει μια σχετική μείωση της ακρίβειας (accuracy), η οποία παραταύτα παραμένει υψηλή (τάξης >90%), ιδίως για μέσες διαλείψεις και χαμηλά μήκη πρόβλεψης. Τέλος, ο σταθμισμένος μέσος των F1 scores μας δίνει μια συγκεντρωτική και εξισορροπημένη εκτίμηση της συνολικής επίδοσης του εκάστοτε μοντέλου, λαμβάνοντας υπόψη τόσο την ικανότητά του να εντοπίζει θετικά δείγματα (ανά κλάση), όσο και τη συνέπεια της απόδοσής του μεταξύ των διαφορετικών κλάσεων.

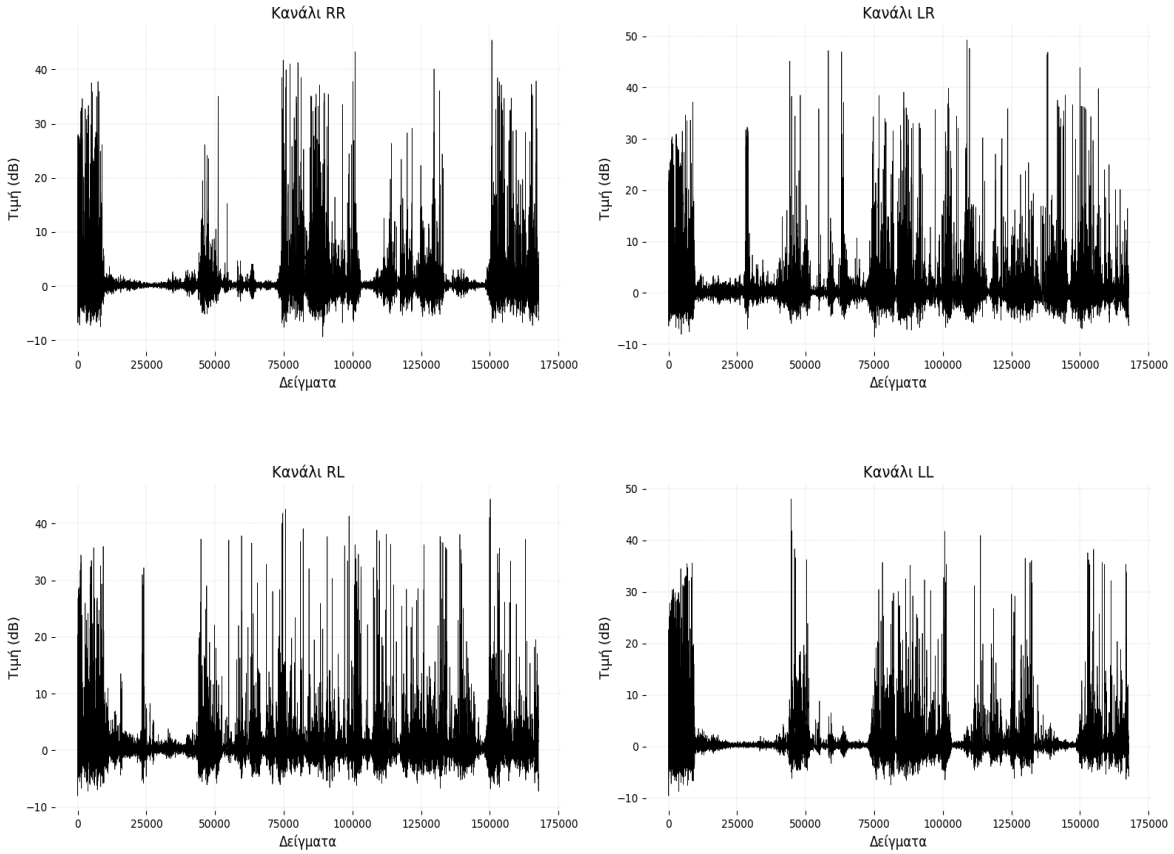
Αποτιμώντας συγκριτικά τα δύο μοντέλα, διαπιστώνουμε ότι το stacked RNN τείνει να εμφανίζει ελαφρώς υψηλότερες επιδόσεις. Πράγματι, κατά τη διάρκεια της εκπαίδευσης διαπιστώθηκε ότι το LSTM τείνει να υπερπροσαρμόζεται στα δεδομένα εκπαίδευσης, με αποτέλεσμα μια σχετική αδυναμία στη γενίκευση έναντι του stacked RNN. Λόγω της υψηλής στοχαστικότητας των διαλείψεων μικρής κλίμακας, τα μοντέλα δεν αναμένεται να αξιοποιούν ιδιαίτερα μακρινές εξαρτήσεις στον χρόνο, οπότε το φαινόμενο των αποσβεννύμενων και εκρηκτικά αυξανόμενων παραγώγων δεν φαίνεται να επηρεάζει τις επιδόσεις του RNN.

Τέλος, έχει ενδιαφέρον να παρατηρήσουμε ότι η επίδοση των μοντέλων είναι σχετικά υψηλότερη στα cross-polarized σε σχέση με τα co-polarized κανάλια. Η διαφορά αυτή μπορεί πιθανώς να αποδοθεί στο γεγονός ότι τα cross-polarized κανάλια εμφανίζουν ελαφρώς βαθύτερες διαλείψεις σε σχέση με τα co-polarized.

4.3 *Συνεχής εκδοχή*

Οι επιδόσεις που καταγράφηκαν στη διακριτοποιημένη εκδοχή του προβλήματος πιστοποιούν την κατ' αρχήν δυνατότητα των μοντέλων να αναπτύξουν ικανότητα πρόβλεψης των διαλείψεων μικρής κλίμακας. Για να μπορέσουμε, όμως, να προσεγγίσουμε το βάθος των διαλείψεων ως μια συνεχή τιμή, αλλά και να εκτιμήσουμε τη διασπορά του

σφάλματος πρόβλεψης, θα πρέπει να μετασχηματίσουμε το πρόβλημα στη συνεχή του εκδοχή, όπως αυτή παρουσιάστηκε στην εισαγωγική παράγραφο του κεφαλαίου. Παρακάτω, απεικονίζονται γραφικά οι χρονοσειρές των 99-οστών percentiles για τα 4 κανάλια.



Εικόνα 4.15: Χρονοσειρές 99οστών percentiles των διαλείψεων μικρής κλίμακας για τα 4 κανάλια

4.3.1 Στατιστική ανάλυση των δεδομένων με στόχο τη μείωση διάστασης

Εν προκειμένω, το μήκος συμφραζομένου τίθεται ως υπερπαράμετρος εκπαίδευσης, οπότε και προσδιορίζεται στο βήμα της ρύθμισης των υπερπαραμέτρων. Ως υποψήφιες τιμές θέτουμε τις 30, 50, 70 και 100.

4.3.2 Επιλογή, εκπαίδευση και ρύθμιση υπερπαραμέτρων μοντέλων

Προσεγγίζοντας το πρόβλημα πρόβλεψης του βάθους των διαλείψεων ως ένα δυναμικό πρόβλημα μάθησης, οι αρχιτεκτονικές που θα χρησιμοποιήσουμε θα είναι αυτές των Stacked RNN, LSTM και GRU. Η ρύθμιση υπερπαραμέτρων πραγματοποιείται μέσω της τεχνικής του randomized search, για τις τυπικές τιμές των υπερπαραμέτρων που

περιγράφονται στον Πίνακα 1.5, ενώ ως αλγόριθμος βελτιστοποίησης χρησιμοποιείται ο Adam. Η επικύρωση και κατόπιν η αξιολόγηση των μοντέλων πραγματοποιείται σε υποακολουθίες ενός συνεχούς παραθύρου μήκους ίσου με το 30% της αρχικής χρονοσειράς των percentiles. Το παράθυρο αυτό επιλέγεται από το τέλος της χρονοσειράς για το εκάστοτε κανάλι και, με επισκόπηση της Εικόνας 4.15, αναμένουμε να είναι εν γένει αντιπροσωπευτικό των μέσων συνθηκών διάδοσης.

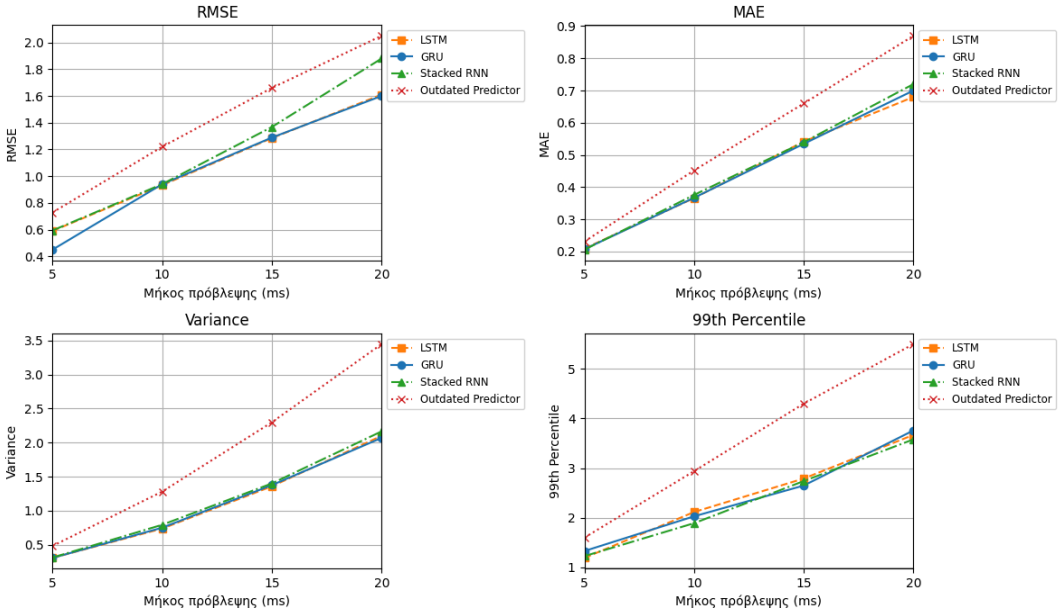
4.3.3 *Αξιολόγηση μοντέλων*

Για την αξιολόγηση των μοντέλων χρησιμοποιούνται, όπως και στο πρόβλημα των απωλειών διαδρομής, μετρικές παλινδρόμησης και συγκεκριμένα αυτές του MAE, RMSE, Variance και 99-οστού percentile του σφάλματος. Ως μοντέλο αναφοράς θα χρησιμοποιήσουμε έναν *ετεροχρονισμένο προβλέπτη (outdated predictor)*, ο οποίος, αξιοποιώντας τη συσχέτιση που παρουσιάζουν οι διαδοχικές τιμές της χρονοσειράς των percentiles, προβλέπει σε ορίζοντα h δειγμάτων βάσει του τελευταίου καταγεγραμμένου δείγματος. Η πρόβλεψη, δηλαδή, που πραγματοποιεί είναι η ακόλουθη:

$$\hat{X}(t + h) = X(t),\ h \in \mathbb{N} \tag{4.10}$$

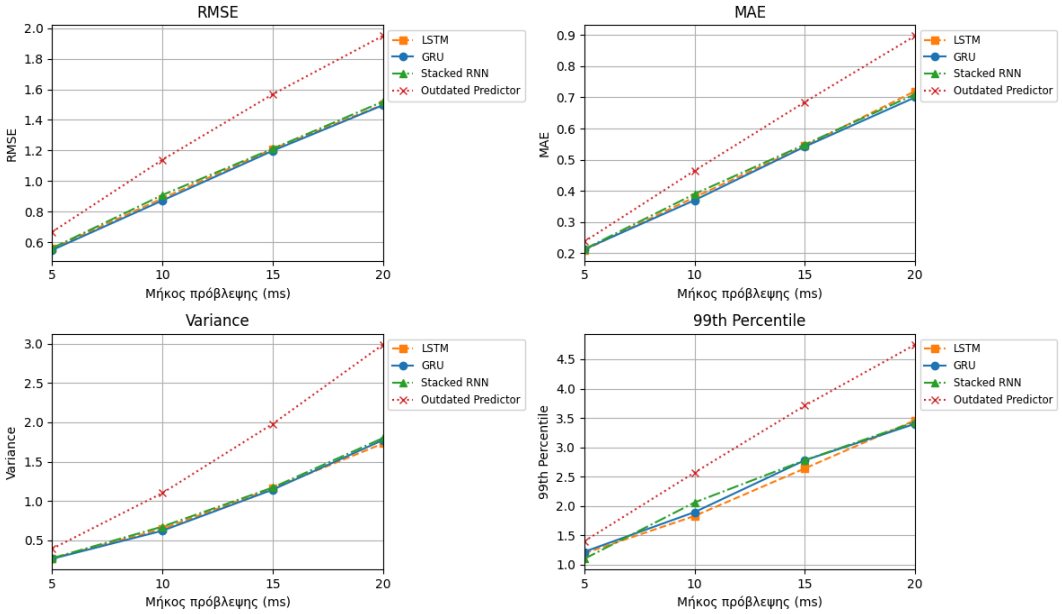
Ουσιαστικά, συγκρίνοντας τις επιδόσεις των μοντέλων μηχανικής μάθησης με τις αντίστοιχες του ετεροχρονισμένου προβλέπτη, επιδιώκουμε να διερευνήσουμε το κατά πόσο τα μοντέλα αναπτύσσουν μια προβλεπτική ικανότητα πέραν της τετριμμένης, η οποία εκφράζεται στη σχέση (4.10). Με δεδομένη την ιδιαίτερα υψηλή στοχαστικότητα που παρουσιάζουν οι διαλείψεις μικρής κλίμακας, κάτι τέτοιο δεν θα πρέπει να θεωρείται αυτονόητο.

Κανάλι RR



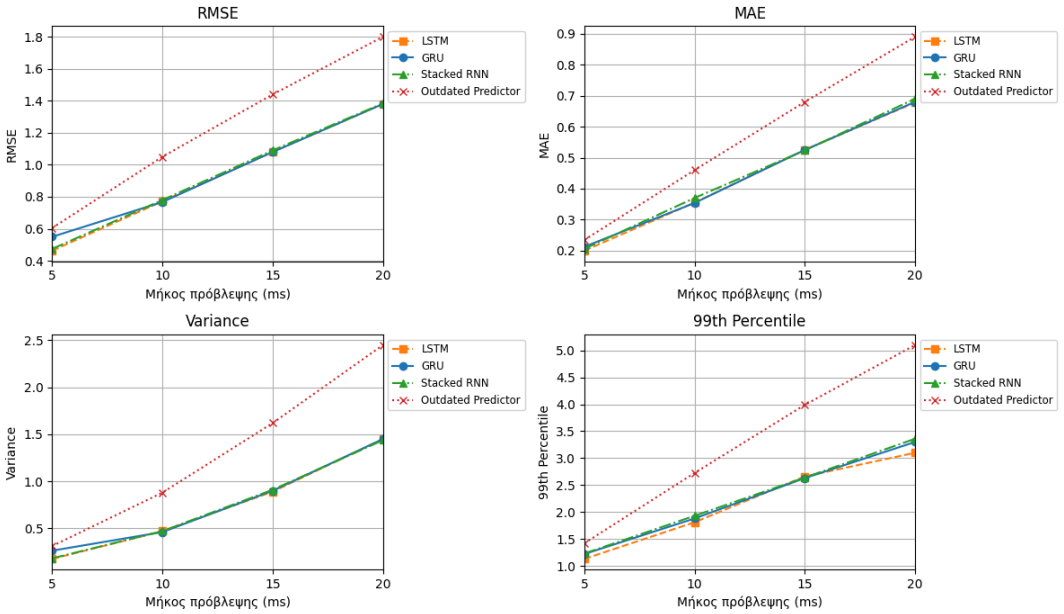
***Εικόνα 4.16**: Απεικόνιση μετρικών επίδοσης των μοντέλων πρόβλεψης του 99οστού percentile των διαλείψεων συναρτήσει του μήκους πρόβλεψης (Κανάλι RR)*

Κανάλι LR



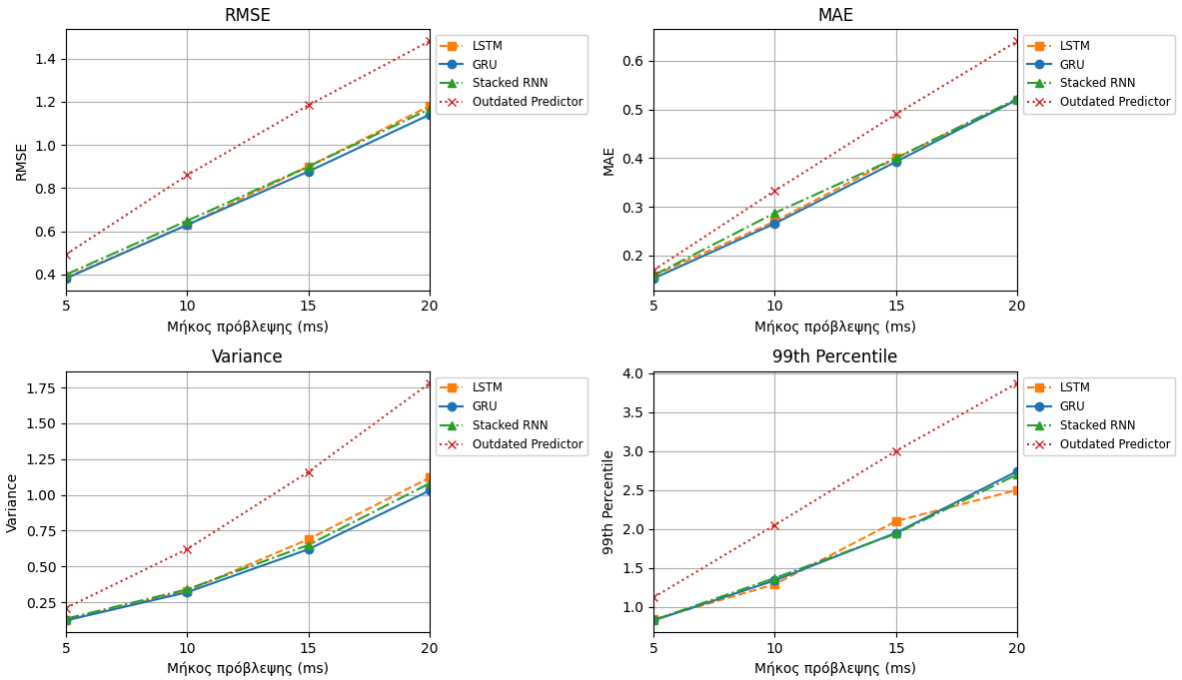
***Εικόνα 4.17**: Απεικόνιση μετρικών επίδοσης των μοντέλων πρόβλεψης του 99οστού percentile των διαλείψεων συναρτήσει του μήκους πρόβλεψης (Κανάλι LR)*

Κανάλι RL



***Εικόνα 4.18**: Απεικόνιση μετρικών επίδοσης των μοντέλων πρόβλεψης του 99οστού percentile των διαλείψεων συναρτήσει του μήκους πρόβλεψης (Κανάλι RL)*

Κανάλι LL



***Εικόνα 4.19**: Απεικόνιση μετρικών επίδοσης των μοντέλων πρόβλεψης του 99οστού percentile των διαλείψεων συναρτήσει του μήκους πρόβλεψης (Κανάλι LL)*

- *Σχολιασμός αποτελεσμάτων:*

Από τα παραπάνω γραφήματα, διαπιστώνουμε κατ' αρχάς την υπεροχή των μοντέλων μηχανικής μάθησης έναντι του ετεροχρονισμένου προβλέπτη. Ανεξαρτήτως καναλιού και μετρικής, τα εκπαιδευμένα μοντέλα πετυχαίνουν σταθερά χαμηλότερα σφάλματα, με την απόκλιση να αυξάνεται όσο αυξάνεται το μήκος πρόβλεψης. Ιδίως στις μετρικές διασποράς, οι διαφορές φτάνουν έως και το 1dB στη Variance και 2dB στο 99-οστό percentile του σφάλματος. Οι διαφορές αυτές δεν είναι αμελητέες, ιδίως αν λάβουμε υπόψη την τάξη μεγέθους των απωλειών μικρής κλίμακας, η οποία αποτυπώνεται στις Εικόνες 4.3-4.6. Σε κάθε περίπτωση, βέβαια, και ανεξαρτήτως της ακριβούς τιμής των διαφορών, η ύπαρξη μιας συστηματικής απόκλισης μεταξύ των δύο ειδών προβλέψεων (μέσω μάθησης και μέσω απλής διατήρησης της μέτρησης) καταδεικνύει το ότι πράγματι, ακόμα και στις δύσκολες συνθήκες του εν λόγω προβλήματος, τα μοντέλα μηχανικής μάθησης κατορθώνουν να εντοπίζουν δομή στα δεδομένα και να την αξιοποιούν, επιτυγχάνοντας ένα συγκριτικό κέρδος επίδοσης.

Τέλος, εστιάζοντας στα μοντέλα μάθησης καθαυτά, διαπιστώνουμε ότι επιτυγχάνουν παρόμοιες επιδόσεις, οι οποίες, αναμενόμενα, φθίνουν με την αύξηση του μήκους πρόβλεψης. Εν γένει, και ιδίως για χαμηλότερα μήκη πρόβλεψης, παρατηρούμε αρκετά χαμηλό μέσο σφάλμα (MAE τάξης 0.25dB και RMSE τάξης 0.6dB), αλλά και διασπορά (Variance τάξης 0.2-0.5dB και 99-οστό percentile τάξης 1-2dB)

Κεφάλαιο 5

Μελλοντικές κατευθύνσεις έρευνας

Το αντικείμενο που πραγματεύτηκε η παρούσα εργασία εντάσσεται στη γενικότερη κατεύθυνση της ενσωμάτωσης τεχνολογιών μηχανικής μάθησης στην τηλεπικοινωνιακή υποδομή [18], [19]. Οι ολοένα αυξανόμενες απαιτήσεις για αξιόπιστη, ταχεία και ασφαλή μετάδοση δεδομένων, σε συνδυασμό με την ανάγκη για βέλτιστη διαχείριση των δικτυακών και τηλεπικοινωνιακών πόρων (ηλεκτρική ισχύς, ηλεκτρομαγνητικό φάσμα, υπολογιστική ισχύς), καθιστούν τη σχεδίαση των δικτύων επικοινωνιών επόμενης γενιάς ένα ιδιαίτερα πολυδιάστατο και πολύπλοκο πρόβλημα. Οι πρωτοφανείς δυνατότητες που παρέχει η μηχανική μάθηση αναφορικά με την επίλυση πολύπλοκων προβλημάτων πρόβλεψης, αναγνώρισης, αλλά και βελτιστοποίησης φαντάζουν ως το πιο ελπιδοφόρο αντιστάθμισμα, με αποτέλεσμα η ερευνητική προσπάθεια για τον συγκερασμό των δύο πεδίων να έχει ενταθεί μαζικά τα τελευταία χρόνια.

Εξειδικεύοντας στο φυσικό στρώμα των επικοινωνιών, η ανάγκη για βέλτιστη διαχείριση πόρων όπως η ισχύς και το φάσμα, καθώς και παροχής αδιάλειπτης διαθεσιμότητας σε πολλαπλούς χρήστες οι οποίοι αντιμετωπίζουν διαφορετικές συνθήκες μετάδοσης, καθιστά επιτακτική ανάγκη την προσαρμογή της μετάδοσης στις δυναμικές συνθήκες διαύλου. Η προσαρμογή αυτή επιτυγχάνεται ρυθμίζοντας παραμέτρους της μετάδοσης (όπως το σχήμα διαμόρφωσης και κωδικοποίησης, η ισχύς εκπομπής, τα κανάλια που ανατίθενται σε κάθε χρήστη) σύμφωνα με μια εκτίμηση στοιχείων της κατάστασης του διαύλου. Δεδομένου όμως πως η κατάσταση του διαύλου αποτελεί μια στοχαστική και χρονικά μεταβαλλόμενη οντότητα, η δυναμική και υψηλής πιστότητας πρόβλεψή της αποτελεί ένα πολύπλοκο πρόβλημα, για το οποίο ενδείκνυται η χρήση τεχνικών μηχανικής μάθησης. Στη βιβλιογραφία ήδη διερευνώνται οι επιδόσεις των τεχνικών αυτών, αξιοποιώντας, κυρίως, δεδομένα προσομοίωσης [20], [21].

Τα αποτελέσματα της παρούσας εργασίας καταδεικνύουν ότι τα μοντέλα μηχανικής μάθησης μπορούν να επιτύχουν ιδιαίτερα υψηλές επιδόσεις στο απαιτητικό πρόβλημα της πρόβλεψης των διαλείψεων μικρής και μεγάλης κλίμακας, κρίσιμων στοιχείων της κατάστασης του διαύλου. Η αξία των αποτελεσμάτων αυτή ενισχύεται αν λάβει κανείς υπόψη πως η εκπαίδευση και αξιολόγηση πραγματοποιείται πάνω σε πραγματικά πειραματικά δεδομένα, τα οποία ενσωματώνουν μια ποικιλία φυσικών φαινομένων που πιθανώς δεν αποτυπώνονται σε προσομοίωση. Μελλοντικά, και προκειμένου να διερευνηθεί σε μεγαλύτερο βάθος η προοπτική χρήσης της μηχανικής μάθησης σε εμπορικά συστήματα επικοινωνιών, θα απαιτηθεί η πραγματοποίηση εκτεταμένων εκστρατειών μετρήσεων, ιδίως για ζεύξεις που δεν υφίσταντο σε δίκτυα παλαιότερης γενιάς, όπως αυτές μεταξύ αεροπλοίων και επίγειων τερματικών.

Επιπρόσθετα στην ανάγκη για περισσότερα δεδομένα, η ενσωμάτωση τεχνολογιών μηχανικής μάθησης σε πραγματικά συστήματα απαιτεί και τη σχεδίαση εξειδικευμένου υλικού, με δυνατότητα υλοποίησης της διαδικασίας παραγωγής προβλέψεων σε πραγματικό

χρόνο. Ιδίως στην περίπτωση της πρόβλεψης των διαλείψεων μικρής κλίμακας, τα χρονικά περιθώρια στα οποία καλείται να αποκριθεί το σύστημα είναι ιδιαίτερα απαιτητικά, τάξης μόνο μερικών ms. Μάλιστα, σε ενσωματωμένα συστήματα, πέραν των περιορισμών στην ταχύτητα επεξεργασίας, υφίστανται επιπλέον περιορισμοί στο μέγεθος της μνήμης, αλλά και στο μήκος λέξης του συστήματος. Ως εκ τούτου, θα πρέπει να διερευνηθεί η δυνατότητα υπολογιστικής υλοποίησης των προβλέψεων σε περιβάλλοντα *Tiny ML* [22],

Τέλος, εστιάζοντας κυρίως στην περίπτωση των δεδομένων χρονοσειρών, αξίζει να διερευνηθεί η χρήση μοντέλων βαθιάς μάθησης τελευταίας τεχνολογίας, όπως οι Time Series Transformers [23]. Η πρόοδος στην επεξεργασία ακολουθιακών δεδομένων, εξαιτίας της ευρείας εφαρμογής σε πεδία της τεχνητής νοημοσύνης όπως η επεξεργασία φωνής και φυσικής γλώσσας, είναι καταιγιστική τα τελευταία χρόνια, συνεισφέροντας μια πληθώρα επιτυχημένων μοντέλων υψηλής χωρητικότητας. Με δεδομένη τη διάχυτη παρουσία των χρονοσειρών σε αρκετούς τομείς της βιομηχανίας, ανάμεσά τους και οι τηλεπικοινωνίες, αναμένουμε η εφαρμογή των συγκεκριμένων μοντέλων να επεκταθεί ιδιαίτερα στα επόμενα χρόνια.

Βιβλιογραφία

[1] Fefferman, C., Mitter, S., & Narayanan, H. (2013). *Testing the manifold hypothesis*.arXiv preprint arXiv:1310.0425. https://doi.org/10.48550/arXiv.1310.0425

[2] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning representations by back-propagating errors*. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, MIT Press.

[3] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). *Support vector regression machines*. In *Advances in Neural Information Processing Systems* (NeurIPS).

[4] Bartlett, P. L., & Mendelson, S. (2002). *Rademacher and Gaussian complexities: Risk bounds and structural results*. *Journal of Machine Learning Research*, 3, 463–482.

[5] Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980

[6] Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition* (4η έκδ.). Academic Press.

[7] Zhang, H., Wang, X., Zhang, C., & Xu, X. (2005). *A fast SMO training algorithm for support vector regression*. In L. Wang, K. Chen, & Y.-S. Ong (Eds.), *Advances in Natural Computation* (ICNC 2005), *Lecture Notes in Computer Science*, Vol. 3610, pp. 221–224. Springer. https://doi.org/10.1007/11539087_26

[8] Haykin, S. S. (2009). *Neural Networks and Learning Machines* (3η έκδ.). Pearson Education, Upper Saddle River, NJ.

[9] Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *The Annals of Statistics*, 29(5), 1189–1232.

[10] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[11] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014).Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1), 1929–1958.

[12] Pascanu, R., Mikolov, T., & Bengio, Y. (2013). *On the difficulty of training recurrent neural networks.* Proceedings of the 30th International Conference on Machine Learning (ICML), PMLR 28(3), 1310–1318. Ανακτήθηκε από https://arxiv.org/abs/1211.5063

[13] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Cho, K. et al. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. Ανακτήθηκε από https://arxiv.org/abs/1406.1078

[15] Κωττής, Π., & Αράπογλου, Π.-Δ. (2011). *Ασύρματες Επικοινωνίες.* Εκδόσεις Τζιόλα.

[16] Psychogios, K., Moraitis, N., & Panagopoulos, A. D. (2024). *Evaluating the performance of a MIMO channel under tree shadowing for low-altitude UAVs.* In *Proceedings of the 5th International Conference on Modern Circuits and Systems Technologies (MOCAST 2024)*,1–4. https://doi.org/10.1109/MOCAST57943.2023.10176758

[17] Nikolaidis, V., Moraitis, N., & Kanatas, A. G. (2017). *Dual-polarized narrowband MIMO LMS channel measurements in urban environments*. *IEEE Transactions on Antennas and Propagation, 65*(2), 900–910. https://doi.org/10.1109/TAP.2016.2637862

[18] Soares Pivoto, D. G., de Figueiredo, F. A. P., Cavdar, C., et al. (2024). *A comprehensive survey of machine learning applied to resource allocation in wireless communications*. *IEEE Communications Surveys & Tutorials*. https://doi.org/10.1109/COMST.2024.3382413

[19] Tran, H. Q., Pham, V. T., & Ngo, S. (2025). *A survey on the applications of machine learning, deep learning, and reinforcement learning in wireless communications*. *Computer and Telecommunication Engineering, 3*(1), 3170. https://doi.org/10.48550/arXiv.2403.18710

[20] Wei, J., & Schotten, H.-D. (2020). *Recurrent Neural Networks with Long Short-Term Memory for Fading Channel Prediction.* In *Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC-Fall)*

[21] Wei, J., & Schotten, H.-D. (2020). *A deep learning method to predict fading channel in multi-antenna systems*. In *Proceedings of the IEEE Vehicular Technology Conference (VTC Spring 2020)*, 1–5. https://doi.org/10.1109/VTC2020-SPRING.2020.9129077

[22] Banbury, C. R., Reddi, V. J., Lam, M., Fu, W., & Holleman, J. (2020). *Benchmarking TinyML Systems: Challenges and Direction*. In Proceedings of the 2020 Conference on Machine Learning and Systems (MLSys).

[23] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). *Transformers in time series: A survey*. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI).