Πτυχιακή Εργασία

Διερεύνηση της χρονικής καθυστέρησης μεταξύ βλέμματος και κίνησης δακτύλου ως δείκτη δυσκολίας κατανόησης κειμένου με χρήση νευρωνικών δικτύων.

Γιώργος Σουραϊλίδης

Αριθμός Μητρώου: 1220703523

Τμήμα Πληροφορικής

Πανεπιστήμιο Νεάπολις Πάφου

Επιβλέπων: Δρ. Τάσος Αντωνιάδης

Πάφος, 2026

Copyright & Δικαιώματα Χρήσης

Υπεύθυνη Δήλωση Συγγραφέα

Με την υπογραφή μου παρακάτω, βεβαιώνω υπεύθυνα ότι αυτή η πτυχιακή εργασία αποτελεί προϊόν αποκλειστικά δικής μου, αυτόνομης επιστημονικής μελέτης και προσπάθειας.

Όπου χρησιμοποιήθηκαν ιδέες, κείμενα, δεδομένα ή γραφήματα άλλων ερευνητών, έχει γίνει η ανάλογη ξεκάθαρη αναφορά στις πηγές τους, σύμφωνα με τους κανόνες της ακαδημαϊκής δεοντολογίας. Επιπλέον, δηλώνω ότι το κείμενο δεν έχει αντιγραφεί από πουθενά, δεν αποτελεί προϊόν λογοκλοπής και δεν έχει παραχθεί μέσω μη εξουσιοδοτημένης χρήσης εργαλείων τεχνητής νοημοσύνης.

Αναγνωρίζω πλήρως ότι οποιαδήποτε διαπίστωση αντιγραφής ή παραποίησης στοιχείων έρχεται σε αντίθεση με τους κανονισμούς του Πανεπιστημίου Neapolis Πάφου και μπορεί να οδηγήσει στην απόρριψη της εργασίας μου.

Ο Δηλών

Ονοματεπώνυμο: Γιώργος Σουραϊλίδης

Αριθμός Μητρώου: 1220703523

Ημερομηνία: Ιούνιος 2026

Τοποθεσία: Πάφος, Κύπρος

GIORGOS SOURAILIDIS

Περίληψη

Η εργασία αυτή εξετάζει αν η σχέση μεταξύ δύο βιομετρικών σημάτων που καταγράφονται κατά την ανάγνωση, της κίνησης του βλέμματος (eye tracking, ET) και της κίνησης του δακτύλου (finger tracking, FT), μπορεί να λειτουργήσει ως αξιόπιστος βιοδείκτης της δυσκολίας κατανόησης κειμένου σε επίπεδο λέξης. Η βασική ιδέα είναι απλή. Όταν ο αναγνώστης συναντά μια λέξη που τον δυσκολεύει, το βλέμμα μένει περισσότερο πάνω της και επιστρέφει συχνά πίσω, ενώ το δάκτυλο συνεχίζει συνήθως πιο σταθερά. Αυτή η ασυμφωνία ανάμεσα στα δύο σήματα φαίνεται να κουβαλά πληροφορία σχετικά με το γνωστικό φορτίο.

Τα δεδομένα προέρχονται από ένα σύνολο δεδομένων που περιλαμβάνει μετρήσεις και από τα δύο είδη tracking. Μια κρίσιμη ιδιαιτερότητα είναι ότι τα features ET και FT δεν διαθέτουν κοινή χρονική αναφορά (timestamp alignment) ανά συμμετέχοντα. Έτσι δεν είναι δυνατή η εφαρμογή κλασικών μεθόδων υπολογισμού χρονικής καθυστέρησης μεταξύ συγχρονισμένων σημάτων. Για να ξεπεραστεί αυτό το εμπόδιο σχεδιάστηκε ένας έμμεσος λειτουργικός δείκτης, ο lag proxy, ο οποίος ορίζεται ως η διαφορά TRT normalized − coverage και αποτυπώνει την ίδια ιδέα της ασυμφωνίας οπτικού και κινητικού σήματος σε επίπεδο λέξης χωρίς να χρειάζεται χρονικός συγχρονισμός.

Για την πρόβλεψη της αναγνωστικής δυσκολίας εκπαιδεύτηκαν και συγκρίθηκαν δύο μοντέλα μηχανικής μάθησης, ένας αλγόριθμος gradient boosting (XGBoost με 300 δέντρα) και ένα πολυεπίπεδο νευρωνικό δίκτυο (Multi Layer Perceptron, MLP, με αρχιτεκτονική 128-64-32). Επειδή η θετική κλάση καλύπτει μόλις το 5 με 6 τοις εκατό των παρατηρήσεων, εφαρμόστηκε η τεχνική SMOTE μόνο στο σύνολο εκπαίδευσης. Η αξιολόγηση έγινε με τις μετρικές F1 score, Precision, Recall και τον confusion matrix, ενώ η ερμηνεία των αποτελεσμάτων στηρίχθηκε στην τεχνική SHAP (TreeExplainer), που ποσοτικοποιεί τη συνεισφορά κάθε χαρακτηριστικού στις προβλέψεις του μοντέλου.

Τα αποτελέσματα δείχνουν ότι ο lag proxy συμβάλλει ουσιαστικά στην πρόβλεψη της δυσκολίας και κατατάσσεται υψηλά στην ιεραρχία σημαντικότητας χαρακτηριστικών. Η σύγκριση μεταξύ XGBoost και MLP αναδεικνύει την υπεροχή του πρώτου σε δομημένα δεδομένα αυτού του τύπου, τόσο στην απόδοση όσο και στην ερμηνευσιμότητα. Τέλος, η εργασία υλοποιήθηκε ως πλήρης διαδικτυακή εφαρμογή με backend FastAPI σε Render και frontend σε Netlify, η οποία εκτελεί ολόκληρη την αναλυτική ροή από τη φόρτωση των δεδομένων έως την παραγωγή των γραφημάτων SHAP.

Λέξεις κλειδιά: eye tracking, finger tracking, αναγνωστική δυσκολία, γνωστικό φορτίο, XGBoost, νευρωνικά δίκτυα, SHAP, ερμηνευσιμότητα, ελληνικό σύνολο δεδομένων.

Abstract

This thesis investigates whether the relationship between two simultaneous biometric signals recorded during reading, eye-tracking (ET) and finger-tracking (FT), can serve as a reliable biomarker of text comprehension difficulty at the word level. The theoretical foundation rests on the hypothesis that, when cognitive inconsistency arises (elevated cognitive load during the processing of a word), ocular behaviour (prolonged fixation duration, regressions) diverges from the motor flow of the finger. This divergence, referred to in the literature as temporal lag, is the central piece of information that the present study tries to exploit.

The data come from a dataset that contains measurements from both tracking modalities. A critical characteristic is that ET and FT features lack a common temporal reference (timestamp alignment) at the participant level, a condition that prevents the application of classical methods for measuring temporal lag between synchronized signals. To address this constraint, an indirect functional indicator, named lag proxy, was designed. It is defined as the difference TRT normalized − coverage and captures the mismatch between visual and motor signals at the word level without requiring per participant temporal synchronization.

For the prediction of reading difficulty, two machine learning models were trained and compared: a gradient boosting algorithm (XGBoost with 300 estimators) and a multilayer neural network (Multi Layer Perceptron, MLP, with a 128-64-32 architecture). Given the substantial class imbalance (the positive class represents around 5 to 6 percent of observations), the SMOTE technique was applied only to the training set. Evaluation was conducted using F1 score, Precision, Recall and the confusion matrix, while interpretability was addressed through the SHAP framework (TreeExplainer), which quantifies the contribution of each feature to the model's predictions.

The results indicate that lag proxy contributes substantially to the prediction of difficulty, ranking highly in the SHAP feature importance hierarchy. The comparative evaluation between XGBoost and MLP demonstrates the superiority of the former on structured data of this nature, both in terms of performance and interpretability. Finally, the work was deployed as a complete web application (FastAPI backend on Render, frontend on Netlify), which executes the full analytical pipeline from data ingestion to SHAP plots.

Keywords: eye tracking, finger tracking, reading difficulty, cognitive load, XGBoost, neural networks, SHAP, interpretability, Greek

Πίνακας περιεχομένων

Περίληψη 4
Abstract 5
Copyright & Δικαιώματα Χρήσης 2
Κεφάλαιο 1. Εισαγωγή 9
1.1 Υπόβαθρο και κίνητρο της έρευνας 9
1.2 Διατύπωση του ερευνητικού προβλήματος 9
1.3 Σκοπός και στόχοι της μελέτης 10
1.4 Ερευνητικά ερωτήματα 10
1.5 Συνεισφορά της εργασίας 11
1.6 Δομή της εργασίας 11
Κεφάλαιο 2. Θεωρητικό υπόβαθρο και βιβλιογραφική ανασκόπηση 12
2.1 Η ανάγνωση ως γνωστική διαδικασία 12
2.1.1 Ορισμός της κατανόησης και θεμελίωση της μετρησιμότητάς της 12
2.1.2 Γνωστικό φορτίο κατά την ανάγνωση 12
2.2 Eye-tracking στην έρευνα της ανάγνωσης 13
2.2.1 Αντικείμενο καταγραφής και βασικές έννοιες 13
2.2.2 Καθιερωμένοι δείκτες: FFD, FPD, TRT, RPD 13
2.2.3 Ευρήματα της διεθνούς βιβλιογραφίας 13
2.3 Finger-tracking και κινητική καθοδήγηση 14
2.3.1 Η κινητική καθοδήγηση ως στρατηγική προσοχής 14
2.3.2 Αντικείμενο καταγραφής της μεθόδου FT 14
2.3.3 Το ερευνητικό κενό στον συνδυασμό ET και FT 14
2.4 Χρονική καθυστέρηση (Temporal Lag) ως βιοδείκτης 15
2.4.1 Η υπόθεση της λειτουργικής απόκλισης 15
2.4.2 Τεχνικές ποσοτικοποίησης της χρονικής σχέσης 15
2.4.3 Ο δείκτης lag proxy: ορισμός και αιτιολόγηση 15
2.5 Μηχανική μάθηση για πρόβλεψη αναγνωστικής δυσκολίας 16
2.5.1 Δομημένα δεδομένα και καταλληλότητα μοντέλων 16
2.5.2 Ο αλγόριθμος XGBoost 16
2.5.3 Το Multi-Layer Perceptron (MLP) ως baseline 17
2.5.4 Ανισοκατανομή κλάσεων και τεχνική SMOTE 19
2.5.5 Ερμηνευσιμότητα μέσω SHAP values 19

2.6 Ερευνητικό κενό και συνεισφορά 20
Κεφάλαιο 3. Μεθοδολογία 20
3.1 Σχεδιασμός της έρευνας 20
3.2 Το σύνολο δεδομένων 21
3.3 Προεπεξεργασία δεδομένων 22
3.3.1 Καθαρισμός ακραίων και ελλιπών τιμών 22
3.3.2 Μεθοδολογία συγχώνευσης των σημάτων ET και FT 22
3.4 Κατασκευή της μεταβλητής στόχου και του δείκτη lag proxy 23
3.4.1 Ορισμός της μεταβλητής word difficulty 23
3.4.2 Ο δείκτης lag proxy ως χαρακτηριστικό εισόδου 23
3.5 Επιλογή χαρακτηριστικών εισόδου 24
3.6 Αντιμετώπιση ανισοκατανομής κλάσεων 25
3.7 Αρχιτεκτονική των μοντέλων 25
3.7.1 Ρυθμίσεις XGBoost 25
3.7.2 Ρυθμίσεις MLP 25
3.8 Πρωτόκολλο εκπαίδευσης και αξιολόγησης 25
3.9 Ανάλυση cross correlation ET–FT 26
3.10 Ερμηνευσιμότητα μέσω SHAP 26
Κεφάλαιο 4. Αιτιολόγηση και ανάπτυξη μοντέλων 27
4.1 Μεθοδολογική αιτιολόγηση επιλογής αλγορίθμων 27
4.2 Αρχιτεκτονική και ρυθμίσεις XGBoost 28
4.3 Αρχιτεκτονική και ρυθμίσεις MLP 28
4.4 Εκπαιδευτική διαδικασία και αποφυγή υπερπροσαρμογής 29
Κεφάλαιο 5. Σχεδιασμός και υλοποίηση συστήματος 30
5.1 Συνολική αρχιτεκτονική της εφαρμογής THE-LAG 30
5.2 Backend: αναλυτικό επίπεδο και REST API 31
5.3 Διασύνδεση frontend με τα εκπαιδευμένα μοντέλα 31
5.4 Frontend: διεπαφή χρήστη και οπτικοποιήσεις 31
5.5 Αυθεντικότητα των χρηστών και αποθήκευση 32
5.6 Deployment και περιορισμοί υποδομής 32
Κεφάλαιο 6. Αποτελέσματα και ανάλυση SHAP 32
6.1 Περιγραφικά στατιστικά του dataset 32
6.2 Απόδοση των μοντέλων 33

6.3 Συγκριτική αξιολόγηση XGBoost και MLP 34
6.4 Ανάλυση SHAP και σημαντικότητα χαρακτηριστικών 34
6.5 Αποτελέσματα cross correlation ET–FT 36
Κεφάλαιο 7. Συζήτηση 37
7.1 Απάντηση στα ερευνητικά ερωτήματα 37
7.2 Σύγκριση με την υπάρχουσα βιβλιογραφία 38
7.3 Πρακτικές συνέπειες και εφαρμογές 38
Κεφάλαιο 8. Περιορισμοί, μελλοντικές επεκτάσεις και συμπεράσματα 39
8.1 Περιορισμοί της μελέτης 39
8.2 Προτάσεις για μελλοντική έρευνα 39
8.3 Συμπεράσματα 40
Βιβλιογραφία 41
Παραρτήματα 44
Π1. Πλήρης πίνακας μεταβλητών του dataset 44
Π2. Πρόσθετα γραφήματα και οπτικοποιήσεις 45

Κεφάλαιο 1. Εισαγωγή

1.1 Υπόβαθρο και κίνητρο της έρευνας

Η ανάγνωση μοιάζει εκ πρώτης όψεως κάτι απλό. Στην πραγματικότητα όμως είναι από τις πιο σύνθετες γνωστικές λειτουργίες που εκτελεί ο άνθρωπος. Ο εγκέφαλος συντονίζει ταυτόχρονα οπτικά, γλωσσικά, κινητικά και μνημονικά υποσυστήματα, ώστε να μετατρέψει μια σειρά γραπτών συμβόλων σε νόημα. Όλη αυτή η συντονισμένη δουλειά γίνεται τόσο γρήγορα ώστε ο αναγνώστης να μην την αντιλαμβάνεται. Όταν όμως εμφανιστεί κάποια δυσκολία, για παράδειγμα μια σπάνια ή σύνθετη λέξη, η ροή σπάει και η συμπεριφορά μας αλλάζει με τρόπο που μπορούμε να μετρήσουμε. Αυτές οι μετρήσιμες αποκλίσεις αποτελούν το εμπειρικό υλικό πάνω στο οποίο στηρίζεται η μελέτη του γνωστικού φορτίου κατά την ανάγνωση (Just και Carpenter, 1980· Rayner, 1998).

Η παρακολούθηση των οφθαλμικών κινήσεων (eye tracking, ET) έχει καθιερωθεί εδώ και δεκαετίες ως η πιο αξιόπιστη μέθοδος για να μετρήσουμε αυτές τις αποκλίσεις με αντικειμενικό τρόπο. Καταγράφοντας μεταβλητές όπως ο χρόνος εστίασης και οι παλινδρομήσεις (regressions), η μέθοδος ET μας επιτρέπει να ποσοτικοποιήσουμε έμμεσα τη γνωστική επεξεργασία σε επίπεδο λέξης (Duchowski, 2017· Rayner et al., 2012). Πιο πρόσφατα, η παρακολούθηση της κίνησης του δακτύλου (finger tracking, FT) έχει αναδειχθεί ως συμπληρωματική μέθοδος, στην οποία το δάκτυλο ερμηνεύεται ως δείκτης κινητικής στρατηγικής (Pirrelli et al., 2020). Πρόκειται για μια αυθόρμητη συμπεριφορά που υιοθετεί ο αναγνώστης για να διατηρήσει την προσοχή του σε απαιτητικά κείμενα. Κάθε μέθοδος αποτυπώνει διαφορετική διάσταση της ανάγνωσης. Η ET δείχνει πού πάει η προσοχή και η FT αποκαλύπτει πώς ο αναγνώστης τη διαχειρίζεται ενεργά.

Το κίνητρο της παρούσας μελέτης ξεκινά από μια απλή παρατήρηση. Στη διεθνή βιβλιογραφία τα δύο σήματα μελετώνται συνήθως ξεχωριστά. Η μεταξύ τους σχέση, και ειδικά η χρονική ή λειτουργική τους απόκλιση, παραμένει σε μεγάλο βαθμό αναξιοποίητη ως πηγή πληροφορίας. Η υπόθεση που διατυπώνουμε είναι ότι, σε ομαλή ανάγνωση, τα δύο σήματα εμφανίζουν σταθερή συσχέτιση, ενώ σε σημεία αυξημένου γνωστικού φορτίου η συσχέτιση αυτή σπάει. Αν καταφέρουμε να ποσοτικοποιήσουμε αυτή τη διαταραχή, ίσως αποκτήσουμε έναν νέο αντικειμενικό βιοδείκτη της αναγνωστικής δυσκολίας. Οι πιθανές εφαρμογές είναι σημαντικές. Στην εκπαίδευση μπορεί να βοηθήσει στον σχεδιασμό προσαρμοστικών κειμένων. Στην κλινική πρακτική μπορεί να συμβάλει στην πρώιμη ανίχνευση δυσκολιών όπως η δυσλεξία (Snowling και Hulme, 2012· Nation, 2019). Στη βασική έρευνα της γνωστικής επιστήμης μπορεί να ανοίξει νέες κατευθύνσεις μελέτης.

1.2 Διατύπωση του ερευνητικού προβλήματος

Το πρόβλημα που πραγματεύεται η εργασία αναπτύσσεται σε δύο επίπεδα. Το πρώτο είναι θεωρητικό. Εξετάζουμε αν η λειτουργική απόκλιση μεταξύ των σημάτων ET και FT μπορεί να αποτελέσει αξιόπιστο δείκτη γνωστικού φορτίου σε επίπεδο λέξης. Το δεύτερο είναι μεθοδολογικό και έχουμε διαθέσιμα δεδομένα. Το σύνολο δεδομένων που αξιοποιούμε δεν διαθέτει κοινή χρονική αναφορά μεταξύ των features ET και FT ανά συμμετέχοντα. Διαφορετικοί χρήστες κατέγραψαν δεδομένα σε διαφορετικά κείμενα ανά συσκευή και ακόμα και στις περιπτώσεις κοινού κειμένου οι χρονικές ενδείξεις δεν είναι συγχρονισμένες. Η εφαρμογή

κλασικών τεχνικών υπολογισμού χρονικής καθυστέρησης, όπως η συνάρτηση crosscorrelation σε συγχρονισμένες χρονοσειρές, γίνεται αδύνατη στο ατομικό επίπεδο.

Η αντιμετώπιση αυτού του περιορισμού αποτελεί από μόνη της αυτοτελές μεθοδολογικό ζήτημα. Η λύση που προτείνουμε είναι η κατασκευή ενός έμμεσου λειτουργικού δείκτη, ο οποίος αποτυπώνει την ιδέα της λειτουργικής απόκλισης χωρίς να απαιτεί χρονικό συγχρονισμό. Ο δείκτης αυτός, που ονομάζεται lag proxy και ορίζεται ως η διαφορά TRT_normalized − coverage, αναλύεται εκτενώς στις ενότητες 2.4 και 3.4.2. Αποτελεί μία από τις κεντρικές πρωτότυπες συνεισφορές της μελέτης.

1.3 Σκοπός και στόχοι της μελέτης

Με απλά λόγια, η εργασία θέλει να δει αν μπορούμε να καταλάβουμε πότε μια λέξη ζορίζει τον αναγνώστη κοιτώντας ταυτόχρονα τα μάτια και το δάκτυλό του. Για να το πετύχουμε, φτιάχνουμε έναν δείκτη που μετρά πόσο διαφέρουν τα δύο αυτά σήματα. Στη συνέχεια, ελέγχουμε αν αυτός ο δείκτης βοηθά πραγματικά τα μοντέλα μηχανικής μάθησης να προβλέψουν τη δυσκολία και αν τα αποτελέσματα μπορούμε να τα εξηγήσουμε με τρόπο κατανοητό.

Οι επιμέρους στόχοι συνοψίζονται ως εξής. Πρώτον, αξιοποιούμε ένα σύνολο δεδομένων που περιλαμβάνει δεδομένα και από τις δύο μεθόδους tracking. Δεύτερον, σχεδιάζουμε έναν λειτουργικό δείκτη που αποτυπώνει την έννοια της χρονικής καθυστέρησης παρά την απουσία χρονικού συγχρονισμού ανά συμμετέχοντα. Τρίτον, εκπαιδεύουμε και συγκρίνουμε δύο διαφορετικούς αλγορίθμους, XGBoost και MLP, υπό συνθήκες προεπεξεργασίας και αξιολόγησης, ώστε η σύγκριση να είναι μεθοδολογικά ορθή. Τέταρτον, εφαρμόζουμε τεχνικές ερμηνευσιμότητας (SHAP values) για την ποσοτικοποίηση της συνεισφοράς κάθε χαρακτηριστικού στις προβλέψεις του μοντέλου (Lundberg και Lee, 2017). Πέμπτον, αναπτύσσουμε το σύστημα ως ολοκληρωμένη διαδικτυακή εφαρμογή, η οποία εκτελεί την αναλυτική ροή από άκρη σε άκρη και καθιστά τα αποτελέσματα προσβάσιμα σε πραγματικούς χρήστες.

1.4 Ερευνητικά ερωτήματα

Η μελέτη οργανώνεται γύρω από έξι συγκεκριμένα ερωτήματα τα οποία κατανέμονται σε τρεις θεματικές ομάδες. Τα δύο πρώτα αφορούν το φαινομένο και την ποσοτικοποίησή του. Τα επόμενα δύο αφορούν την εγκυρότητα του δείκτη ως βιοδείκτη. Τα δύο τελευταία αφορούν τη μεθοδολογική επιλογή των μοντέλων και την ερμηνεία των αποτελεσμάτων. Η απάντηση σε κάθε ερώτημα αντιστοιχεί σε συγκεκριμένο κεφάλαιο ή ενότητα του κειμένου, ενώ η συνολική τους συζήτηση γίνεται στο Κεφάλαιο 7.

E1. Πώς μπορεί να ποσοτικοποιηθεί η σχέση μεταξύ των σημάτων ET και FT όταν λείπει η κοινή χρονική αναφορά ανά συμμετέχοντα; E2. Πώς συνδέονται τα γλωσσικά χαρακτηριστικά των λέξεων (μήκος, συχνότητα, αριθμός συλλαβών) με τον λειτουργικό δείκτη lag_proxy; E3. Μπορεί ο δείκτης αυτός να λειτουργήσει ως αξιόπιστος βιοδείκτης της αναγνωστικής δυσκολίας σε επίπεδο λέξης; E4. Ποια είναι η προγνωστική ακρίβεια ενός μοντέλου μηχανικής μάθησης για τη δυσκολία ανάγνωσης, με την χρήση του συγκεκριμένου συνόλου χαρακτηριστικών; E5. Ποιο από τα δύο μοντέλα (XGBoost, MLP) είναι καταλληλότερο για δομημένα δεδομένα αυτού του τύπου και ποιοι παράγοντες δικαιολογούν την επιλογή; E6. Τι αποκαλύπτει η ανάλυση SHAP σχετικά με

τη συνεισφορά κάθε χαρακτηριστικού, και ειδικότερα του lag_proxy, στην πρόβλεψη της δυσκολίας;

1.5 Συνεισφορά της εργασίας

Η συνεισφορά της εργασίας απλώνεται σε πέντε άξονες, που καλύπτουν θεωρητικές και πρακτικές διαστάσεις. Ο πρώτος και κεντρικός άξονας είναι η πρόταση του λειτουργικού δείκτη lag proxy. Ο δείκτης σχεδιάστηκε ειδικά για την περίπτωση όπου τα δεδομένα ET και FT δεν έχουν κοινή χρονική αναφορά. Είναι μια κατάσταση που εμφανίζεται συχνά σε πραγματικά δεδομένα. Ο δεύτερος άξονας είναι η εφαρμογή της συνολικής μεθοδολογίας στο σύνολο δεδομένων.

Ο τρίτος άξονας είναι μεθοδολογικός. Συγκρίνουμε συστηματικά τους αλγορίθμους XGBoost και MLP κάτω από ρεαλιστικές συνθήκες, με έντονη ανισοκατανομή κλάσεων και χωρίς εκτεταμένη βελτιστοποίηση υπερπαραμέτρων, και τεκμηριώνουμε εμπειρικά ποια προσέγγιση ταιριάζει καλύτερα. Ο τέταρτος άξονας αφορά τη χρήση της τεχνικής SHAP όχι ως απλής επίδειξης ερμηνευσιμότητας, αλλά ως αναλυτικού εργαλείου για την απάντηση συγκεκριμένου ερευνητικού ερωτήματος, της σχετικής συνεισφοράς κάθε χαρακτηριστικού και ιδίως του lag proxy. Ο πέμπτος άξονας είναι πρακτικός. Υλοποιήσαμε ολοκληρωμένη full-stack εφαρμογή, με backend σε FastAPI που τρέχει στο Render και frontend στο Netlify, με υποστήριξη αυθεντικότητας των χρηστών και εκτέλεσης της αναλυτικής ροής σε πραγματικό χρόνο. Έτσι η έρευνα γίνεται επαναχρησιμοποιήσιμη και προσβάσιμη πέρα από το ακαδημαϊκό πλαίσιο.

1.6 Δομή της εργασίας

Η εργασία οργανώνεται σε οκτώ κεφάλαια. Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο και η βιβλιογραφική ανασκόπηση, με σταδιακή ανάπτυξη από την ανάγνωση ως γνωστική διαδικασία, στη μέθοδο ET, στη μέθοδο FT, στην έννοια της χρονικής καθυστέρησης και στα εργαλεία μηχανικής μάθησης που αξιοποιούνται. Στο τέλος του κεφαλαίου εντοπίζεται το ερευνητικό κενό. Στο Κεφάλαιο 3 περιγράφεται αναλυτικά η μεθοδολογία. Καλύπτουμε το dataset, την προεπεξεργασία, τον ορισμό της μεταβλητής στόχου, την κατασκευή του lag proxy, την αρχιτεκτονική των μοντέλων και το πρωτόκολλο αξιολόγησης. Στο Κεφάλαιο 4 αναλύουμε τη μεθοδολογική αιτιολόγηση και τη λεπτομερή αρχιτεκτονική των δύο αλγορίθμων. Στο Κεφάλαιο 5 παρουσιάζεται η αρχιτεκτονική του συστήματος λογισμικού και η διαδικασία deployment. Το Κεφάλαιο 6 περιλαμβάνει τα αποτελέσματα, την απόδοση μοντέλων, την ανάλυση SHAP και την cross-correlation. Το Κεφάλαιο 7 αποτελεί τη συζήτηση, στην οποία απαντώνται σημείο προς σημείο τα έξι ερευνητικά ερωτήματα. Το Κεφάλαιο 8 καταγράφει τους περιορισμούς της μελέτης, προτείνει μελλοντικές επεκτάσεις και συνοψίζει τα κεντρικά συμπεράσματα. Στο τέλος παρατίθενται η βιβλιογραφία και τα παραρτήματα με κώδικα και πρόσθετα γραφήματα.

Κεφάλαιο 2. Θεωρητικό υπόβαθρο και βιβλιογραφική ανασκόπηση

2.1 Η ανάγνωση ως γνωστική διαδικασία

2.1.1 Ορισμός της κατανόησης και θεμελίωση της μετρησιμότητάς της

Η ανάγνωση περιγράφεται στη γνωστική βιβλιογραφία ως μια πολυεπίπεδη διαδικασία που αναπτύσσετε σε διαφορετικές φάσεις: οπτική πρόσληψη, λεξική αναγνώριση, σημασιολογική ενσωμάτωση και συντακτική επεξεργασία (Just και Carpenter, 1980· Reichle, Warren και McConnell, 2009). Σε αντίθεση με την υποκειμενική εντύπωση που τη χαρακτηρίζει ως μια συνεχή και ομοιογενή εμπειρία, η πραγματική εκτέλεση των φάσεων αυτών συμβαίνει σε μεγάλο βαθμό παράλληλα και με μεγάλη ταχύτητα. Έτσι η διαδικασία γίνεται αντιληπτή ως αδιάσπαστη. Η ομοιογένεια όμως διαταράσσεται μόλις εμφανιστεί γνωστική ασυνέπεια, για παράδειγμα κατά την επεξεργασία μιας σπάνιας, πολυσύνθετης ή σημασιολογικά απρόβλεπτης λέξης. Σε τέτοιες περιπτώσεις, οι οφθαλμικές κινήσεις αποκτούν μετρήσιμες αποκλίσεις από το κανονικό πρότυπο, γεγονός που καθιστά εφικτή την έμμεση ανίχνευση της αυξημένης γνωστικής προσπάθειας. Σε αυτή ακριβώς την ανιχνευσιμότητα στηρίζεται η σύγχρονη έρευνα στην αναγνωστική συμπεριφορά.

Η θεωρητική βάση εντοπίζεται στην εργασία των Just και Carpenter (1980), οι οποίοι διατύπωσαν την υπόθεση ότι ο χρόνος παραμονής του βλέμματος σε μια λέξη αντανακλά άμεσα τη γνωστική προσπάθεια που απαιτήθηκε για την επεξεργασία της. Σύμφωνα με την υπόθεση αυτή, όσο αυξάνεται η λεξική ή σημασιολογική δυσκολία μιας λέξης, τόσο αυξάνεται και ο χρόνος εστίασης. Η σχέση αυτή έχει επιβεβαιωθεί σε εκατοντάδες μεταγενέστερες μελέτες, καλύπτοντας πολλαπλές γλώσσες, ηλικιακές ομάδες και πειραματικές συνθήκες (Rayner, 1998· Kliegl, Nuthmann και Engbert, 2006). Για τη μελέτη μας, η ποσοτική φύση της σχέσης αυτής επιτρέπει την αντικειμενική μέτρηση της γνωστικής επεξεργασίας χωρίς προσφυγή σε υποκειμενικές αξιολογήσεις, κάτι ιδιαίτερα σημαντικό για τη μεθοδολογική ακρίβεια.

2.1.2 Γνωστικό φορτίο κατά την ανάγνωση

Ο όρος γνωστικό φορτίο αναφέρεται στην ένταση της νοητικής επεξεργασίας που απαιτείται για την κατανόηση ενός κειμενικού ερεθίσματος. Πρόκειται για δυναμική μεταβλητή, η οποία διαφοροποιείται τόσο ανάμεσα σε διαφορετικούς αναγνώστες όσο και μέσα στον ίδιο αναγνώστη, ανάλογα με την στιγμή, την κόπωση και το πλαίσιο. Η δυναμικότητα αυτή είναι ταυτόχρονα ερευνητικό ενδιαφέρον και μεθοδολογική πρόκληση. Το γνωστικό φορτίο δεν εκτιμάται ως στατικό μέγεθος, αλλά απαιτεί συνεχή καταγραφή μέσω εργαλείων που παρακολουθούν την αναγνωστική συμπεριφορά σε πραγματικό χρόνο. Ακριβώς αυτό το είδος καταγραφής προσφέρουν οι μέθοδοι ET και FT (Duchowski, 2017· Pirrelli et al., 2020).

Το βασικό μεθοδολογικό εμπόδιο είναι ότι το γνωστικό φορτίο αποτελεί εσωτερική, μη παρατηρήσιμη μεταβλητή. Για τον λόγο αυτό, όλες οι σύγχρονες μέθοδοι μέτρησής του στηρίζονται σε εξωτερικά παρατηρήσιμα σήματα που λειτουργούν ως έμμεσοι δείκτες (proxies): οφθαλμικές κινήσεις, χρόνοι αντίδρασης, ηλεκτροεγκεφαλικά σήματα, διάμετρος κόρης ή καρδιακός ρυθμός (Hollenstein et al., 2018). Καθένας από αυτούς τους δείκτες αποτυπώνει διαφορετική πλευρά της γνωστικής προσπάθειας και κανένας δεν θεωρείται από μόνος του επαρκής. Από τις διαθέσιμες μεθόδους, η οφθαλμική καταγραφή προσφέρει πρακτική

προσβασιμότητα, μη επεμβατικότητα και ώριμη βιβλιογραφική υποδομή. Η καταγραφή της κίνησης του δακτύλου, προσθέτει μια επιπλέον διάσταση. Αποκαλύπτει τη στρατηγική με την οποία ο αναγνώστης ρυθμίζει ενεργά την προσοχή του και συμπληρώνει την οπτική ένδειξη του ΕΤ με ένα κινητικό αποτύπωμα της προσπάθειας.

2.2 Eye-tracking στην έρευνα της ανάγνωσης

2.2.1 Αντικείμενο καταγραφής και βασικές έννοιες

Η συσκευή eye-tracking καταγράφει την κατεύθυνση του βλέμματος ενός υποκειμένου ως συνάρτηση του χρόνου, συνήθως με χρονική ανάλυση 60 έως 1000 Hz ανάλογα με τον εξοπλισμό (Duchowski, 2017). Σε αντίθεση με την κοινή αντίληψη, οι οφθαλμικές κινήσεις κατά την ανάγνωση δεν χαρακτηρίζονται από ομαλή ροή πάνω στη γραμμή, αλλά από διακριτά άλματα (saccades) που εναλλάσσονται με στιγμιαίες παύσεις (fixations). Κατά τη διάρκεια των saccades, η οπτική αντίληψη είναι πρακτικά ανεσταλμένη, φαινόμενο γνωστό ως saccadic suppression. Η πραγματική πρόσληψη πληροφορίας συντελείται αποκλειστικά κατά τη διάρκεια των fixations. Επιπλέον, ένα φαινόμενο ιδιαίτερης σημασίας είναι οι παλινδρομήσεις (regressions), δηλαδή οι επιστροφές του βλέμματος σε προηγούμενες λέξεις. Οι παλινδρομήσεις αποτελούν ισχυρό δείκτη γνωστικής ασυνέπειας, αφού κανονικά η ανάγνωση προχωρά μονοδρομικά. Όταν εμφανίζονται, υποδηλώνουν ότι η αρχική επεξεργασία δεν κατάφερε να ενσωματώσει πλήρως την πληροφορία (Vasilev και Angele, 2017).

2.2.2 Καθιερωμένοι δείκτες: FFD, FPD, TRT, RPD

Από την ακατέργαστη ροή fixations και saccades, η βιβλιογραφία έχει αναδείξει ένα σύνολο τυποποιημένων δεικτών, οι οποίοι αξιοποιούνται και στο μοντέλο της παρούσας μελέτης ως χαρακτηριστικά εισόδου (Rayner, 1998· Kliegl, Nuthmann και Engbert, 2006). Η First Fixation Duration (FFD) αντιστοιχεί στη διάρκεια της πρώτης φοράς που το βλέμμα προσγειώνεται στη λέξη και αντανακλά τα αρχικά στάδια λεξικής αναγνώρισης, πριν ο εγκέφαλος αποφασίσει αν χρειάζεται περαιτέρω επεξεργασία. Η First Pass Duration (FPD) καλύπτει τον συνολικό χρόνο που δαπανήθηκε στη λέξη κατά την πρώτη ανάγνωση, μαζί με όλα τα διαδοχικά fixations πριν το βλέμμα μεταβεί σε επόμενη λέξη. Το Total Reading Time (TRT) είναι ο πιο ολιστικός δείκτης. Υπολογίζει όλο τον χρόνο που αφιερώθηκε στη λέξη, μαζί με τις επιστροφές μέσω παλινδρόμησης. Ο δείκτης αυτός αξιοποιείται τόσο ως χαρακτηριστικό εισόδου όσο και στην κατασκευή της μεταβλητής στόχου. Τέλος, η Regression Path Duration (RPD) μετρά τον χρόνο από το πρώτο fixation στη λέξη έως την οριστική έξοδο του βλέμματος προς τα δεξιά, ενσωματώνοντας τυχόν επιστροφές. Καθένας από τους δείκτες αυτούς αντανακλά διαφορετική φάση της επεξεργασίας. Συνδυαστικά παρέχουν ένα πολυδιάστατο προφίλ της αναγνωστικής συμπεριφοράς σε κάθε λέξη.

2.2.3 Ευρήματα της διεθνούς βιβλιογραφίας

Η ανασκόπηση του Rayner (1998), που συνοψίζει δύο δεκαετίες ερευνητικής δραστηριότητας, παραμένει σημείο αναφοράς για την εισαγωγή στο πεδίο. Τα κεντρικά ευρήματα είναι εμπειρικά ισχυρά και έχουν επιβεβαιωθεί σε πολλαπλές ανεξάρτητες μελέτες. Ο χρόνος εστίασης αυξάνεται συστηματικά με το μήκος της λέξης, μειώνεται με τη συχνότητα εμφάνισης της λέξης στο σύνολο δεδομένων, και αυξάνεται όταν η λέξη είναι σημασιολογικά απρόβλεπτη εντός του πλαισίου. Αυτά τα τρία γλωσσικά χαρακτηριστικά (μήκος, συχνότητα, προβλεψιμότητα) εξηγούν σημαντικό

μέρος της διακύμανσης στους χρόνους εστίασης και για τον λόγο αυτό συμπεριλαμβάνονται στο feature set της παρούσας μελέτης. Η πιο πρόσφατη ανασκόπηση των Andreou και Gkantaki (2024) εστιάζει ειδικά στην ενήλικη αναγνωστική συμπεριφορά και επιβεβαιώνει ότι οι ίδιοι δείκτες παραμένουν έγκυροι και στον ενήλικο πληθυσμό, με διαφοροποιημένες απόλυτες τιμές. Η συνοχή αυτή προσφέρει σταθερό υπόβαθρο για την αξιοποίηση των ET δεικτών στην παρούσα ανάλυση.

2.3 Finger-tracking και κινητική καθοδήγηση

2.3.1 Η κινητική καθοδήγηση ως στρατηγική προσοχής

Η χρήση του δακτύλου κατά την ανάγνωση συχνά συνδέεται με αντιλήψεις που την αποδίδουν σε αναγνωστική ανωριμότητα ή σε δυσκολίες αποκωδικοποίησης. Η σύγχρονη βιβλιογραφία αναθεωρεί αυτή την ερμηνεία (Pirrelli et al., 2020· Taxitari et al., 2021). Η κίνηση του δακτύλου θεωρείται πλέον αυθόρμητη στρατηγική προσοχής, η οποία υιοθετείται όταν το κείμενο αποκτά αυξημένες απαιτήσεις επεξεργασίας. Τέτοιες συνθήκες περιλαμβάνουν την ανάγνωση αριθμητικών δεδομένων, πινάκων ή λιστών όπου η ακρίβεια θέσης είναι κρίσιμη, την κατάσταση κόπωσης του αναγνώστη, ή ακόμα και συνειδητή απόφαση ενίσχυσης της συγκέντρωσης σε πυκνό κείμενο. Η βασική κινηματική διαφορά από τη μέθοδο ET εντοπίζεται στη φύση της κίνησης. Ενώ το βλέμμα μετακινείται με διακριτά άλματα (saccades), το δάκτυλο ακολουθεί συνεχή και ομαλή τροχιά (Sharmin et al., 2013). Έτσι τα δύο σήματα δεν αντιπροσωπεύουν συμπληρωματικές καταγραφές του ίδιου φαινομένου, αλλά ποιοτικά διαφορετικές πληροφοριακές πηγές. Το ET αποτυπώνει την τρέχουσα κατεύθυνση της προσοχής, ενώ το FT αποτυπώνει τη στρατηγική διαχείρισή της. Η διάκριση αυτή κάνει ιδιαίτερα ενδιαφέρουσα τη μελέτη της μεταξύ τους σχέσης.

2.3.2 Αντικείμενο καταγραφής της μεθόδου FT

Τα δεδομένα finger-tracking στο σύνολο δεδομένων μας περιλαμβάνουν δύο βασικές μεταβλητές ανά λέξη, τον χρόνο δακτυλο παρακολούθησης (dt) και το ποσοστό κάλυψης της λέξης από το δάκτυλο (coverage), με τιμές στο διάστημα [0, 1]. Από τις δύο μεταβλητές, το coverage συγκεντρώνει το κύριο ενδιαφέρον λόγω της άμεσης ερμηνευσιμότητάς του. Τιμή coverage κοντά στο 1 υποδηλώνει ότι ο αναγνώστης πέρασε το δάκτυλο πάνω από την πλήρη έκταση της λέξης με ομαλή κίνηση, ενώ μικρότερες τιμές υποδηλώνουν μερική κάλυψη, η οποία συχνά αντιστοιχεί σε διστακτική ή ασυνεχή κινητική ροή. Στην παρούσα ανάλυση, το coverage αξιοποιείται τόσο αυτοτελώς ως χαρακτηριστικό εισόδου όσο και ως μία από τις δύο συνιστώσες του λειτουργικού δείκτη lag proxy. Αυτό επιβάλλει ιδιαίτερη προσοχή στην ποιότητα της προεπεξεργασίας του.

2.3.3 Το ερευνητικό κενό στον συνδυασμό ET και FT

Η εργασία των Pirrelli et al. (2020) αποτελεί μία από τις πρώτες συστηματικές μελέτες που δηλώνουν το finger-tracking ως μεθοδολογικά έγκυρο εργαλείο για τη μελέτη της αναγνωστικής συμπεριφοράς. Η μελέτη δείχνει ότι η κίνηση του δακτύλου συνιστά αξιόπιστο δείκτη αναγνωστικού ρυθμού και συσχετίζεται με τη δυσκολία του κειμένου και τη γνωστική προσπάθεια. Η συστηματική όμως αξιοποίηση των δύο σημάτων (ET και FT) σε ενιαίο υπολογιστικό πλαίσιο, και ιδίως η εκμετάλλευση της σχέσης μεταξύ τους ως διακριτού δείκτη δυσκολίας, παραμένει εξαιρετικά περιορισμένη στη διεθνή βιβλιογραφία. Η περιορισμένη κάλυψη γίνεται ακόμα πιο έντονη στα ελληνικά δεδομένα, όπου τα διαθέσιμα δεδομένα με διπλή

καταγραφή (ET και FT) είναι σπάνια. Το κενό αυτό αναλύεται στην ενότητα 2.6 και αποτελεί μία από τις βασικές ερευνητικές ευκαιρίες που αξιοποιεί η παρούσα μελέτη.

2.4 Χρονική καθυστέρηση (Temporal Lag) ως βιοδείκτης

2.4.1 Η υπόθεση της λειτουργικής απόκλισης

Η κεντρική υπόθεση της εργασίας διατυπώνεται ως εξής. Σε ομαλή ανάγνωση, τα δύο βιομετρικά σήματα, η κίνηση του βλέμματος και η κίνηση του δακτύλου, εμφανίζουν σταθερή λειτουργική συσχέτιση, με σχετικά προβλέψιμη σχέση προήγησης ή ακολουθίας μεταξύ τους. Όποτε όμως εμφανίζεται γνωστική ασυνέπεια, δηλαδή όταν ο αναγνώστης συναντά μια λέξη που απαιτεί αυξημένη γνωστική προσπάθεια, η συσχέτιση διαταράσσεται. Η διαταραχή μπορεί να εκδηλωθεί με δύο τρόπους. Είτε το βλέμμα μένει εγκλωβισμένο στη λέξη (αυξημένος χρόνος εστίασης, παλινδρομήσεις) ενώ το δάκτυλο έχει ήδη προχωρήσει ακολουθώντας τη δική του πιο σταθερή κινητική λογική. Είτε αντιστρόφως, το δάκτυλο σταματά διστακτικά ενώ το βλέμμα περιπλανιέται αναζητώντας σημασιολογική σύνδεση. Και στις δύο περιπτώσεις, η απόκλιση μεταξύ των δύο σημάτων κωδικοποιεί πληροφορία σχετικά με το αυξημένο γνωστικό φορτίο. Η μελέτη μας δεν ενδιαφέρεται για την απόλυτη χρονική καθυστέρηση ως φυσικό μέγεθος σε millisecond. Ενδιαφέρεται για την ποιοτική διαταραχή της φυσιολογικής σχέσης, όπου εντοπίζεται το πραγματικό επιστημονικό ενδιαφέρον.

2.4.2 Τεχνικές ποσοτικοποίησης της χρονικής σχέσης

Η καθιερωμένη μέθοδος ποσοτικοποίησης της χρονικής σχέσης μεταξύ δύο σημάτων είναι η συνάρτηση cross correlation. Η συνάρτηση αυτή υπολογίζει τον βαθμό ομοιότητας δύο χρονοσειρών ως συνάρτηση της σχετικής χρονικής μετατόπισης της μιας ως προς την άλλη. Σε απόλυτα συγχρονισμένα σήματα, η μέγιστη συσχέτιση εντοπίζεται στη μηδενική μετατόπιση. Όταν υπάρχει συστηματική χρονική υστέρηση του ενός σήματος έναντι του άλλου, το μέγιστο μετατοπίζεται και η θέση του δίνει την εκτιμώμενη καθυστέρηση. Στην παρούσα μελέτη, η cross-correlation εφαρμόζεται σε χρονοσειρές του TRT και του coverage σε μακροσκοπικό επίπεδο, δηλαδή στο σύνολο του dataset, και προσφέρει μια συνολική εικόνα της σχέσης των δύο σημάτων. Η ανάλυση αυτή έχει περιγραφικό ρόλο και λειτουργεί συμπληρωματικά προς τη μοντελοποίηση. Δεν παρέχει δείκτη ανά λέξη και για τον λόγο αυτό κρίθηκε αναγκαία η κατασκευή εναλλακτικού δείκτη εφαρμόσιμου σε επίπεδο λέξης, όπως περιγράφεται στην επόμενη ενότητα.

2.4.3 Ο δείκτης lag proxy: ορισμός και αιτιολόγηση

Η αυστηρή εφαρμογή της cross-correlation προϋποθέτει την ύπαρξη δύο χρονοσειρών με κοινή χρονική αναφορά. Δηλαδή, για δεδομένο συμμετέχοντα και δεδομένη λέξη, χρειάζεται αντιστοιχία μεταξύ της χρονικής στιγμής της οφθαλμικής καταγραφής και της χρονικής στιγμής της δακτυλικής καταγραφής, ώστε να είναι δυνατός ο υπολογισμός της σχετικής μετατόπισης. Στο σύνολο δεδομένων μας, αυτή η προϋπόθεση δεν ικανοποιείται. Τα features ET και FT προέρχονται από ανεξάρτητες καταγραφές, συχνά σε διαφορετικά κείμενα ανά συμμετέχοντα, και χωρίς κοινή χρονική αναφορά ακόμα και στις περιπτώσεις ταύτισης κειμένου. Κάθε προσπάθεια άμεσης συγχώνευσης των δύο ρευμάτων δεδομένων με κλειδί (idUser, gid, sid, tid) οδηγεί σε εκτεταμένη απώλεια παρατηρήσεων, καθιστώντας το dataset πρακτικά μη αξιοποιήσιμο.

Για την αντιμετώπιση αυτού του περιορισμού, η μελέτη εισάγει έναν νέο έμμεσο λειτουργικό δείκτη, τον lag proxy, που ορίζεται ως: lag proxy = TRT normalized − coverage. Η λογική του εξηγείται απλά. Η μεταβλητή TRT normalized εκφράζει σε κλίμακα [0, 1] το σχετικό γνωστικό φορτίο που δέχθηκε το βλέμμα στη συγκεκριμένη λέξη. Υψηλότερες τιμές αντιστοιχούν σε αυξημένη οπτική δυσκολία. Η μεταβλητή coverage εκφράζει στην ίδια κλίμακα τη σχετική ομαλότητα της δακτυλικής ροής. Υψηλότερες τιμές αντιστοιχούν σε σταθερή κινητική επεξεργασία. Όταν οι δύο τιμές συγκλίνουν, το σύστημα λειτουργεί συγχρονισμένα. Όταν αποκλίνουν, είτε επειδή το βλέμμα μένει εγκλωβισμένο ενώ το δάκτυλο έχει ήδη καλύψει τη λέξη, είτε αντιστρόφως, η απόκλιση κωδικοποιεί ένα μη τυπικό γεγονός που είναι ενδεικτικό αυξημένου γνωστικού φορτίου. Ο δείκτης συνοψίζει αυτή την πληροφορία σε μία αριθμητική τιμή ανά λέξη, καθιστώντας την άμεσα αξιοποιήσιμη ως χαρακτηριστικό εισόδου σε υπολογιστικά μοντέλα.

Διευκρινίζεται ότι ο δείκτης lag proxy δεν είναι αυστηρά χρονική μέτρηση με τη φυσική έννοια. Δεν εκφράζεται σε μονάδες χρόνου και δεν προέρχεται από συγχρονισμένα σήματα. Είναι ένας έμμεσος λειτουργικός δείκτης που αποτυπώνει την ίδια εννοιολογική ιδέα της ασυμφωνίας οπτικού και κινητικού σήματος, με τους περιορισμούς του συγκεκριμένου dataset. Ο περιορισμός αυτός επανέρχεται στο Κεφάλαιο 8 και αποτελεί μία από τις κεντρικές κατευθύνσεις βελτίωσης σε μελλοντικές μελέτες. Παρά τον προσεγγιστικό του χαρακτήρα, τα αποτελέσματα της μοντελοποίησης (Κεφάλαιο 6) δείχνουν υψηλή προγνωστική ισχύ, υποστηρίζοντας εμπειρικά την εννοιολογική του εγκυρότητα.

2.5 Μηχανική μάθηση για πρόβλεψη αναγνωστικής δυσκολίας

2.5.1 Δομημένα δεδομένα και καταλληλότητα μοντέλων

Το dataset της μελέτης είναι τυπικό παράδειγμα δομημένων δεδομένων (structured data).
Κάθε γραμμή αντιστοιχεί σε μία λέξη και κάθε στήλη σε ένα ποσοτικοποιημένο χαρακτηριστικό (μήκος, συχνότητα, διάρκειες fixation και άλλα). Λείπουν οι δομές που ευνοούν τις βαθιές αρχιτεκτονικές. Δεν υπάρχει εικόνα που να αξιοποιείται από συνελικτικά δίκτυα (CNNs), ούτε ακολουθία κειμένου που να ευνοεί τους Transformers, ούτε γνήσια χρονοσειρά. Η φύση των δεδομένων αυτή καθορίζει σε μεγάλο βαθμό την καταλληλότητα των αλγορίθμων. Η πρόσφατη βιβλιογραφία υποδεικνύει ότι σε tabular δεδομένα, μοντέλα βασισμένα σε δέντρα αποφάσεων, και ιδίως οι αλγόριθμοι gradient boosting, όχι μόνο συναγωνίζονται αλλά συχνά υπερτερούν έναντι των νευρωνικών δικτύων, τόσο σε απόδοση όσο και σε υπολογιστική οικονομία (Shwartz-Ziv και Armon, 2022· Grinsztajn, Oyallon και Varoquaux, 2022· Borisov et al., 2022). Η παρατήρηση αυτή εξηγεί την ευρεία υιοθέτηση του XGBoost ως de facto επιλογής σε αντίστοιχα προβλήματα ταξινόμησης και παλινδρόμησης, τόσο στην ακαδημαϊκή έρευνα όσο και σε βιομηχανικές εφαρμογές.

2.5.2 Ο αλγόριθμος XGBoost

Ο αλγόριθμος XGBoost (Chen και Guestrin, 2016) αποτελεί ώριμη υλοποίηση της αρχής του gradient boosting στο πλαίσιο δέντρων αποφάσεων. Η βασική ιδέα είναι διαφορετική από εκείνη ενός μοναδικού βαθέος δέντρου. Αντί να φτιαχτεί ένα μεγάλο δέντρο που προσπαθεί να μάθει τα πάντα μαζί, κατασκευάζεται μια ακολουθία πολλών μικρών δέντρων όπου κάθε νέο δέντρο εκπαιδεύεται για να διορθώσει τα σφάλματα των προηγούμενων. Το τελικό μοντέλο λειτουργεί

ως αθροιστικός συνδυασμός όλων αυτών των ασθενών μαθητών και επιτυγχάνει υψηλή ακρίβεια πρόβλεψης με ελεγχόμενη πολυπλοκότητα.

Πιο αναλυτικά, η εκπαίδευση γίνεται σταδιακά. Στην αρχή, υπάρχει ένα πολύ απλό μοντέλο που κάνει σχετικά αδύναμες προβλέψεις. Σε κάθε βήμα, ο αλγόριθμος υπολογίζει τα σφάλματα του τρέχοντος μοντέλου σε σχέση με τις πραγματικές ετικέτες και εκπαιδεύει ένα νέο μικρό δέντρο που εστιάζει συγκεκριμένα σε αυτά τα λάθη. Το νέο δέντρο προστίθεται με μικρό βάρος (learning rate), ώστε καμία διόρθωση να μην είναι υπερβολικά απότομη. Η διαδικασία επαναλαμβάνεται για n_estimators γύρους και το συνολικό μοντέλο γίνεται όλο και πιο ακριβές. Με αυτόν τον τρόπο, το XGBoost μαθαίνει σύνθετες μη γραμμικές σχέσεις χωρίς να χρειάζεται μεγάλη βαθιά αρχιτεκτονική.

Πέρα από τη θεωρητική βάση, το XGBoost εισάγει αρκετές πρακτικές βελτιώσεις. Πρώτον, χρησιμοποιεί κανονικοποίηση L1 και L2 πάνω στα φύλλα των δέντρων, μηχανισμός που αποτρέπει την υπερπροσαρμογή και κάνει το μοντέλο πιο σταθερό σε νέα δεδομένα. Δεύτερον, αξιοποιεί παράλληλη επεξεργασία κατά τη φάση εύρεσης των βέλτιστων διαχωριστικών στους κόμβους, κάτι που το καθιστά ταχύ ακόμη και σε μεγάλα datasets. Τρίτον, διαχειρίζεται αραιά δεδομένα και ελλιπείς τιμές αυτόματα, μαθαίνοντας κατά την εκπαίδευση τον βέλτιστο τρόπο τοποθέτησης μιας ελλιπούς τιμής στο αριστερό ή δεξί παιδί κάθε κόμβου. Τέταρτον, υποστηρίζει διαφορετικές συναρτήσεις απώλειας, ώστε να εφαρμόζεται σε προβλήματα δυαδικής ή πολυκλασικής ταξινόμησης, παλινδρόμησης ή ranking. Όλες αυτές οι ιδιότητες το κάνουν πρακτικά εφαρμόσιμο σε ρεαλιστικά μεγέθη δεδομένων.

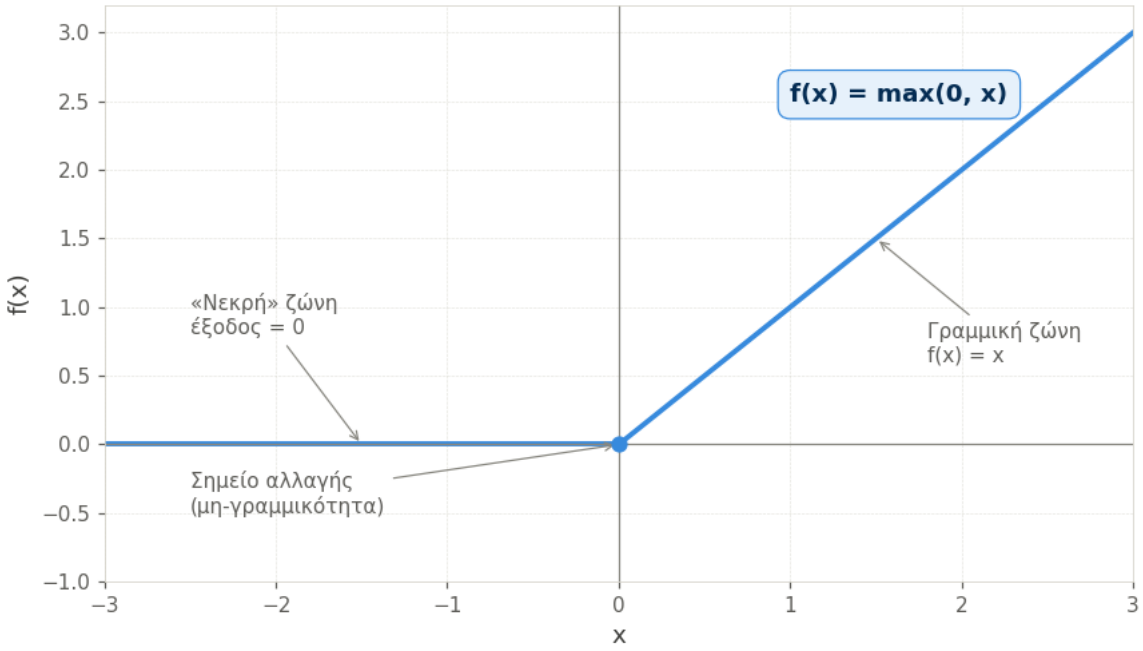
Η καταλληλότητα του XGBoost για τη συγκεκριμένη μελέτη αιτιολογείται από τέσσερις λόγους. Πρώτον, χειρίζεται αυτόματα χαρακτηριστικά διαφορετικής κλίμακας χωρίς να απαιτείται κανονικοποίηση, πράγμα που απλοποιεί τη ροή προεπεξεργασίας. Δεύτερον, υποστηρίζει εγγενώς τη διαχείριση ελλιπών τιμών. Τρίτον, επιτυγχάνει ικανοποιητική απόδοση ακόμα και με μέτρια βελτιστοποίηση υπερπαραμέτρων, χαρακτηριστικό ιδιαίτερα χρήσιμο σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους. Τέταρτον, και πιο σημαντικό για τους στόχους της εργασίας, είναι πλήρως συμβατός με τον TreeExplainer του πακέτου SHAP, γεγονός που εξασφαλίζει εξαγωγή ερμηνεύσιμων αποτελεσμάτων σε πολυωνυμικό χρόνο και με μαθηματικά θεμελιωμένη βάση (Lundberg, Erion και Lee, 2018). Η ερμηνευσιμότητα δεν είναι δευτερεύων στόχος αλλά ρητός ερευνητικός στόχος.

2.5.3 Το Multi-Layer Perceptron (MLP) ως baseline

Το Multi-Layer Perceptron (MLP) είναι η θεμελιώδης αρχιτεκτονική feedforward νευρωνικού δικτύου (Bishop, 2006· Goodfellow, Bengio και Courville, 2016). Συγκροτείται από ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου. Η πληροφορία διαδίδεται μονοκατευθυντικά, χωρίς ανατροφοδοτήσεις ή βρόχους. Κάθε νευρώνας υπολογίζει έναν σταθμισμένο γραμμικό συνδυασμό των εξόδων του προηγούμενου επιπέδου και εφαρμόζει σε αυτόν μια μη γραμμική συνάρτηση ενεργοποίησης, συνήθως τη ReLU. Η συνάρτηση ReLU (Rectified Linear Unit) ορίζεται ως

$$\mathbf{ReLU(x) = max(0, x)}$$

Επιστρέφει *x* όταν είναι θετικό και μηδέν αλλιώς. Η επιλογή της έναντι sigmoid/tanh αιτιολογείται από τρεις ιδιότητες: αποφεύγει το vanishing gradient (κλίση = 1 για x > 0), είναι υπολογιστικά απλή, και επάγει αραιή ενεργοποίηση που λειτουργεί ως έμμεσος μηχανισμός κανονικοποίησης.



Σχήμα 1 — Συνάρτηση ενεργοποίησης ReLU. Για x ≤ 0 η έξοδος είναι 0. Για x > 0 η κλίση είναι 1.

Έτσι το δίκτυο αποκτά τη δυνατότητα να μάθει σύνθετες, μη γραμμικές σχέσεις μεταξύ εισόδου και εξόδου. Η εκπαίδευση γίνεται μέσω backpropagation, όπου το σφάλμα στην έξοδο διαδίδεται προς τα πίσω και ενημερώνει σταδιακά τα βάρη όλων των επιπέδων με τη χρήση του gradient descent ή κάποιας παραλλαγής του, όπως ο αλγόριθμος Adam (Kingma και Ba, 2015).

Στη μελέτη μας αξιοποιείται αρχιτεκτονική 128-64-32, δηλαδή τρία κρυφά επίπεδα. Η επιλογή αυτή δεν είναι τυχαία και ακολουθεί την αρχή του σταδιακά συγκλίνοντα δικτύου (funnel architecture), που λειτουργεί ως μέθοδος σταδιακής μείωσης της πληροφορίας για εξαγωγή ουσιωδών χαρακτηριστικών. Πιο συγκεκριμένα, το πρώτο κρυφό επίπεδο με 128 νευρώνες δέχεται την ακατέργαστη πληροφορία από όλα τα χαρακτηριστικά εισόδου και την ξεδιπλώνει σε ένα ευρύτερο πληροφοριακό χώρο. Σε αυτή την πρώτη φάση, το δίκτυο εντοπίζει βασικά πρότυπα και ενεργοποιεί νευρώνες ευαίσθητους σε διαφορετικούς συνδυασμούς γλωσσικών, οφθαλμικών και κινητικών χαρακτηριστικών. Το δεύτερο επίπεδο με 64 νευρώνες αναλαμβάνει να συμπυκνώσει αυτή την πληροφορία. Συνδυάζει τα πρωταρχικά πρότυπα σε αναπαραστάσεις υψηλότερου επιπέδου, στις οποίες έχει ήδη φιλτραριστεί ο θόρυβος και έχει ενισχυθεί η πληροφορία που πραγματικά συμβάλλει στην πρόβλεψη. Το τρίτο επίπεδο με 32 νευρώνες πιέζει ακόμη περισσότερο την αναπαράσταση και κρατά μόνο τα ουσιώδη χαρακτηριστικά που οδηγούν στη δυαδική απόφαση. Με αυτή τη σταδιακή συμπίεση, το δίκτυο αναγκάζεται να μάθει συμπυκνωμένες και κρίσιμες αναπαραστάσεις, αντί να διατηρεί υπερβολικά πλούσιες ενδιάμεσες αναπαραστάσεις που θα οδηγούσαν σε υπερπροσαρμογή. Η λογική αυτή ταιριάζει

στη φύση των δεδομένων μας, όπου ο πραγματικός αριθμός πληροφοριακών διαστάσεων είναι σχετικά μικρός και η πρόβλεψη απαιτεί καθαρή απομόνωση των ουσιωδών στοιχείων.

Παρά την εκρηκτική εξέλιξη βαθύτερων αρχιτεκτονικών τα τελευταία χρόνια, τα MLP διατηρούν τη θέση τους ως πρότυπη baseline λύση σε προβλήματα tabular δεδομένων, ακριβώς επειδή δεν εκμεταλλεύονται δομές (χωρικές ή χρονικές) που λείπουν από τέτοιου τύπου δεδομένα. Στη μελέτη ενσωματώνεται όχι με την προσδοκία ότι θα υπερβεί το XGBoost, η βιβλιογραφική ένδειξη είναι αντίθετη, αλλά ως σημείο αναφοράς για τη συγκριτική αξιολόγηση. Έτσι η σύγκριση τεκμηριώνει εμπειρικά την επιλογή του XGBoost υπό ταυτόσημες συνθήκες προεπεξεργασίας και αξιολόγησης. Μια σημαντική πρακτική διαφορά μεταξύ των δύο αλγορίθμων αφορά την απαίτηση κανονικοποίησης. Το MLP χρειάζεται StandardScaler στα χαρακτηριστικά εισόδου, ενώ το XGBoost λειτουργεί σε μη κανονικοποιημένα δεδομένα. Η διαφορά αυτή συνεπάγεται διακριτές διαδρομές προεπεξεργασίας για τα δύο μοντέλα και επιβάλλει ιδιαίτερη μέριμνα για την αποφυγή κάθε μορφής διαρροής πληροφορίας (data leakage). Ο scaler προσαρμόζεται αποκλειστικά στο σύνολο εκπαίδευσης, χωρίς ποτέ να εκτίθεται σε στατιστικά του συνόλου δοκιμής.

2.5.4 Ανισοκατανομή κλάσεων και τεχνική SMOTE

Η μεταβλητή στόχος της μελέτης είναι έντονα ανισοκατανεμημένη, αφού οι πραγματικά δύσκολες λέξεις είναι εξ ορισμού μειοψηφία σε κάθε φυσικό κείμενο. Για την εξισορρόπηση εφαρμόζεται η τεχνική SMOTE (Chawla et al., 2002) μόνο στο σύνολο εκπαίδευσης. Η εφαρμογή της πριν τον διαχωρισμό train/test θα οδηγούσε σε διαρροή πληροφορίας. Το σύνολο δοκιμής διατηρείται αναλλοίωτο. Σχετικός κριτικός σχολιασμός για τη χρήση της SMOTE σε μοντέλα κινδύνου εμφανίζεται στους van den Goorbergh et al. (2022).

2.5.5 Ερμηνευσιμότητα μέσω SHAP values

Η προγνωστική απόδοση ενός μοντέλου, ακόμα και υψηλή, έχει περιορισμένη αξία σε επιστημονικό πλαίσιο εάν δεν συνοδεύεται από δυνατότητα ερμηνείας των προβλέψεων. Σε εφαρμοσμένα περιβάλλοντα όπως τα συστήματα σύστασης, η προσέγγιση του μαύρου κουτιού είναι ίσως αποδεκτή. Σε επιστημονικό όμως πλαίσιο, όπου επιχειρείται η εξαγωγή συμπερασμάτων για την υποκείμενη πραγματικότητα, η ερμηνευσιμότητα αποτελεί απαραίτητη προϋπόθεση (Molnar, 2022). Οι SHAP values (Lundberg και Lee, 2017) προσφέρουν αυστηρά θεμελιωμένη λύση. Στηρίζονται στις τιμές Shapley της θεωρίας συνεργατικών παιγνίων, που αρχικά διατυπώθηκαν για τη δίκαιη κατανομή κέρδους σε συνεργασία οικονομικών φορέων, και επεκτείνονται στη μηχανική μάθηση ως μέτρο της ακριβούς συνεισφοράς κάθε χαρακτηριστικού σε κάθε επιμέρους πρόβλεψη. Η αξιωματική θεμελίωση εξασφαλίζει ιδιότητες όπως η τοπικότητα, η συνέπεια και η προσθετικότητα, που τις κάνουν μεθοδολογικά αξιόπιστες έναντι άλλων τεχνικών feature importance.

Η τεχνική SHAP συνδυάζει τοπική και συνολική ερμηνεία. Σε τοπικό επίπεδο, επιτρέπει την ανάλυση μεμονωμένης πρόβλεψης. Για παράδειγμα, μπορεί να εξηγήσει γιατί το μοντέλο ταξινόμησε μια συγκεκριμένη λέξη ως δύσκολη, μέσω της συνεισφοράς κάθε χαρακτηριστικού. Σε συνολικό επίπεδο, η άθροιση των τοπικών εξηγήσεων σε όλο το δείγμα δίνει ιεραρχία σημαντικότητας των χαρακτηριστικών (global feature importance), μαθηματικά συνεπή με τις τοπικές εξηγήσεις, ιδιότητα που δεν εξασφαλίζεται από εναλλακτικές προσεγγίσεις. Σχετική

εφαρμογή σε πρόβλημα αναγνωστικής δυσκολίας παρουσιάζουν οι Jäger et al. (2020), οι οποίοι ανέδειξαν με SHAP τη σχετική συνεισφορά γλωσσικών και οφθαλμικών χαρακτηριστικών. Η παρούσα μελέτη επεκτείνει αυτή την προσέγγιση ενσωματώνοντας κινητικά χαρακτηριστικά και τον λειτουργικό δείκτη lag proxy, διευρύνοντας το πεδίο εφαρμογής της τεχνικής.

2.6 Ερευνητικό κενό και συνεισφορά

Η συνολική αποτίμηση της διεθνούς βιβλιογραφίας αναδεικνύει διαφορετικό βαθμό ωριμότητας στα επιμέρους πεδία. Η οφθαλμική καταγραφή είναι ώριμη ερευνητική μέθοδος με έναν αιώνα ιστορίας, τυποποιημένους δείκτες και εκτεταμένη εμπειρική τεκμηρίωση της σχέσης μεταξύ οφθαλμικών μεταβλητών και γλωσσικών χαρακτηριστικών (Rayner, 1998· Rayner et al., 2012). Η καταγραφή της κίνησης του δακτύλου είναι πρόσφατο ερευνητικό πεδίο, με θεμελιώδη συμβολή των Pirrelli et al. (2020) και Taxitari et al. (2021), αλλά λιγότερο πλούσια βιβλιογραφική υποδομή. Τα μοντέλα μηχανικής μάθησης, και ιδίως XGBoost και MLP, έχουν αξιοποιηθεί σε προβλήματα πρόβλεψης αναγνωστικής δυσκολίας, κυρίως όμως με αποκλειστική χρήση οφθαλμικών χαρακτηριστικών και σε αγγλόφωνα corpora (Jäger et al., 2020). Στο πλαίσιο αυτό, εντοπίζονται τρία διακριτά κενά τα οποία η παρούσα μελέτη επιχειρεί να καλύψει.

Το πρώτο κενό είναι εννοιολογικό. Πολύ λίγες μελέτες αξιοποιούν ταυτόχρονα τα σήματα ET και FT και ακόμα λιγότερες προσεγγίζουν τη σχέση μεταξύ τους ως αυτόνομο δείκτη δυσκολίας. Η επικρατούσα τακτική είναι η χρήση των δύο σημάτων ως ανεξάρτητων πηγών πληροφορίας που απλώς συμπληρώνουν η μία την άλλη, παραβλέποντας την πιθανή πληροφορία που εγγράφεται στη μεταξύ τους δυναμική. Το δεύτερο κενό είναι μεθοδολογικό. Όταν τα δύο σήματα δεν διαθέτουν κοινή χρονική αναφορά, κατάσταση συνηθισμένη σε πραγματικά πειραματικά δεδομένα, δεν υπάρχει καθιερωμένη μέθοδος για την εξαγωγή δείκτη καθυστέρησης. Έτσι η σύγκριση των δύο σημάτων συχνά εγκαταλείπεται. Το τρίτο κενό είναι γλωσσικό και γεωγραφικό. Η εφαρμογή της μεθοδολογίας σε ελληνικά αναγνωστικά corpora είναι περιορισμένη, καθιστώντας το ελληνικό σύνολο δεδομένων που χρησιμοποιούμε σπάνιο πόρο.

Η παρούσα μελέτη καλύπτει συγκλίνουσα και τα τρία κενά. Η εισαγωγή του δείκτη lag proxy προσφέρει αναπαραγώγιμη μεθοδολογική απάντηση στο δεύτερο κενό. Η ενσωμάτωσή του σε ενιαίο υπολογιστικό πλαίσιο μαζί με τα υπόλοιπα ET και FT χαρακτηριστικά καλύπτει το πρώτο κενό. Η εφαρμογή της συνολικής μεθοδολογίας στα ελληνικά δεδομένα καλύπτει το τρίτο κενό. Η συνεισφορά της εργασίας είναι επομένως ταυτόχρονα εννοιολογική (εισαγωγή νέου δείκτη), μεθοδολογική (στρατηγική για δεδομένα χωρίς κοινή χρονική αναφορά) και εμπειρική (πρώτη τέτοια εφαρμογή σε ελληνικό αναγνωστικό σύνολο δεδομένων). Η τριπλή κάλυψη κενών προσδιορίζει τη θέση της μελέτης στη διεθνή βιβλιογραφία του πεδίου.

Κεφάλαιο 3. Μεθοδολογία

3.1 Σχεδιασμός της έρευνας

Η μεθοδολογική προσέγγιση είναι ποσοτική και υπολογιστική. Από την αναλυτική διαδικασία αποκλείεται κάθε υποκειμενική κρίση. Όλες οι μεταβλητές που αξιοποιούνται προέρχονται είτε από μετρητικά όργανα (eye tracker, finger tracker) είτε από γλωσσολογική επεξεργασία του κειμενικού υλικού (μήκος, συχνότητα, αριθμός συλλαβών, μέρος λόγου). Η επιλογή αυτή δεν είναι μεθοδολογικά ουδέτερη. Είναι συνειδητή και αποσκοπεί στη διασφάλιση της επιστημονικής

εγκυρότητας των συμπερασμάτων. Ζητούμενο είναι η δυνατότητα τρίτων ερευνητών, διαθέτοντας τα ίδια δεδομένα και την ίδια αναλυτική ροή, να αναπαράγουν πλήρως τα αποτελέσματα. Πρόκειται για θεμελιώδη απαίτηση της επιστημονικής μεθόδου, που δύσκολα ικανοποιείται με υποκειμενικά κριτήρια αξιολόγησης.

Η αναλυτική ροή οργανώνεται σε έξι διακριτά στάδια. Καθένα υλοποιείται σε ξεχωριστό script με σαφώς ορισμένες εισόδους και εξόδους. Το πρώτο στάδιο φορτώνει το αρχικό dataset (αρχείο data.xlsx). Το δεύτερο στάδιο είναι η προεπεξεργασία, που περιλαμβάνει καθαρισμό ακραίων και ελλιπών τιμών, συγχώνευση των ρευμάτων ET και FT, κατασκευή του λειτουργικού δείκτη lag_proxy και ορισμό της μεταβλητής στόχου. Το τρίτο στάδιο διαχωρίζει το dataset σε σύνολα εκπαίδευσης και δοκιμής με αναλογία 80/20 και τεχνική stratified split. Το τέταρτο στάδιο εφαρμόζει SMOTE μόνο στο σύνολο εκπαίδευσης. Το πέμπτο στάδιο εκπαιδεύει και αξιολογεί τα δύο μοντέλα (XGBoost και MLP) υπό ταυτόσημες συνθήκες δεδομένων, με τις απαιτούμενες προσαρμογές προεπεξεργασίας ανά αλγόριθμο. Το έκτο στάδιο περιλαμβάνει την ανάλυση ερμηνευσιμότητας με SHAP και την περιγραφική ανάλυση cross-correlation μεταξύ των σημάτων ET και FT. Η οργάνωση αυτή επιλέχθηκε συνειδητά, αφού εξασφαλίζει αναπαραγωγιμότητα, διευκολύνει τον εντοπισμό σφαλμάτων και επιτρέπει τμηματική επανεκτέλεση χωρίς να χρειάζεται πλήρης επανάληψη. Είναι ένα πρακτικό πλεονέκτημα ιδιαίτερα χρήσιμο σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων όπως το free tier της υπηρεσίας Render που αξιοποιείται στο deployment.

3.2 Το σύνολο δεδομένων

Τα δεδομένα προέρχονται από ένα σύνολο δεδομένων που περιλαμβάνει πειραματικές καταγραφές eye tracking και finger tracking με ενήλικες συμμετέχοντες (Taxitari et al., 2021· Levie et al., 2022). Το αρχείο εργασίας οργανώνεται ανά λέξη (token), όπου κάθε γραμμή αντιστοιχεί σε μία λέξη και οι στήλες καταγράφουν πολυεπίπεδη πληροφορία: χαρακτηριστικά της λέξης (token, lemma, μήκος, συχνότητα, αριθμός συλλαβών, μέρος λόγου), θέση στο κειμενικό πλαίσιο (gid για σελίδα, sid για πρόταση, tid για θέση στη γραμμή), δημογραφικά στοιχεία του συμμετέχοντα (ηλικία, φύλο, επίπεδο εκπαίδευσης), ψυχομετρικές μετρήσεις WAIS (Wechsler, 2008), και τις ίδιες τις μετρήσεις tracking. Ο πλούτος αυτός επιτρέπει τη μελέτη της αναγνωστικής δυσκολίας όχι μόνο ως συνάρτηση του κειμενικού υλικού, αλλά και ως λειτουργία του αναγνώστη.

Μια κρίσιμη ιδιαιτερότητα του dataset δεν είναι προφανής στην αρχική επισκόπηση και αφορά τη δομή των εγγραφών. Η στήλη idDevice διακρίνει δύο κατηγορίες γραμμών. Εγγραφές με τιμή ET περιέχουν οφθαλμικές μετρήσεις (FFD, FPD, TRT, RPD, fixNum, isReg, reFix). Εγγραφές με τιμή 1 περιέχουν μετρήσεις finger-tracking (dt, coverage). Οι μετρήσεις ET και FT κατανέμονται έτσι σε διακριτές γραμμές, όχι στην ίδια γραμμή. Η δομή αυτή έχει σημαντικές συνέπειες. Άμεση συγχώνευση των δύο ρευμάτων με κλειδί (idUser, gid, sid, tid) δεν είναι πρακτικά εφικτή, αφού διαφορετικοί συμμετέχοντες κατέγραψαν δεδομένα σε διαφορετικά κείμενα ανά συσκευή. Για παράδειγμα, ο συμμετέχων Α μπορεί να έχει ET δεδομένα στο κείμενο Χ και FT δεδομένα στο κείμενο Ψ, ή ο συμμετέχων Β να έχει αποκλειστικά ET δεδομένα. Η απευθείας συγχώνευση θα παρήγαγε εκτεταμένες ελλιπείς τιμές στα FT χαρακτηριστικά και θα έκανε το dataset πρακτικά μη

αξιοποιήσιμο. Αυτή η παρατήρηση διαμόρφωσε καθοριστικά τη στρατηγική προεπεξεργασίας που περιγράφεται παρακάτω.

3.3 Προεπεξεργασία δεδομένων

3.3.1 Καθαρισμός ακραίων και ελλιπών τιμών

Το αρχικό αρχείο περιέχει σημαντικό αριθμό ακραίων τιμών (outliers), οι οποίες αποδίδονται σε σφάλματα καταγραφής του εξοπλισμού, με τιμές της τάξης του 10^10 ή μεγαλύτερες για μεταβλητές που φυσιολογικά κυμαίνονται στην κλίμακα των millisecond. Τέτοιες τιμές είναι σαφώς μη έγκυρες και απομακρύνονται μέσω ενός απλού κατωφλίου κατά τη φάση καθαρισμού. Επιπλέον, ελέγχονται και αφαιρούνται γραμμές με λογικά αντιφατικούς συνδυασμούς τιμών. Για παράδειγμα, περιπτώσεις όπου FFD > TRT είναι μαθηματικά αδύνατες αφού το TRT, εξ ορισμού, περιλαμβάνει το FFD ως συνιστώσα του. Τέτοιες ασυνέπειες υποδηλώνουν σφάλμα είτε στην καταγραφή είτε στην αρχική επεξεργασία και οι αντίστοιχες παρατηρήσεις εξαιρούνται. Οι ελλιπείς τιμές αντιμετωπίζονται διαφοροποιημένα ανά τύπο μεταβλητής. Για αριθμητικά χαρακτηριστικά όπως FFD, FPD και TRT εφαρμόζεται γραμμική παρεμβολή εντός του idUser, βασισμένη στην εύλογη υπόθεση τοπικής ομοιότητας γειτονικών λέξεων του ίδιου αναγνώστη. Για το coverage ακολουθείται η διαδικασία συγχώνευσης που περιγράφεται στο επόμενο βήμα.

3.3.2 Μεθοδολογία συγχώνευσης των σημάτων ET και FT

Δεδομένης της αδυναμίας άμεσης συγχώνευσης ανά συμμετέχοντα, υιοθετείται αγρεγαρισμένη προσέγγιση που αποκαθιστά μια λειτουργική αντιστοιχία με ελεγχόμενο συμβιβασμό. Η κεντρική ιδέα είναι απλή. Αντί να αναζητείται η τιμή coverage του συγκεκριμένου συμμετέχοντα για τη συγκεκριμένη λέξη, πληροφορία που συχνά λείπει, υπολογίζεται ο μέσος όρος coverage όλων των συμμετεχόντων που κατέγραψαν την ίδια λέξη στην ίδια θέση του κειμένου, όπως αυτή προσδιορίζεται από την τριάδα gid, sid, tid. Η αγρεγαρισμένη τιμή συγχωνεύεται στη συνέχεια με left merge στις εγγραφές ET, αποδίδοντας σε κάθε ET γραμμή ένα αντιπροσωπευτικό coverage σε επίπεδο λέξης μέσα στο κείμενο.

Η διαδικασία υλοποιείται σε τέσσερα βήματα. Στο πρώτο βήμα, απομονώνονται οι εγγραφές FT (idDevice=1) και επιλέγονται οι στήλες [gid, sid, tid, coverage]. Στο δεύτερο βήμα, υπολογίζεται ο μέσος όρος coverage ανά μοναδική τριάδα (gid, sid, tid) σε όλους τους συμμετέχοντες με διαθέσιμη καταγραφή για τη συγκεκριμένη θέση. Στο τρίτο βήμα, το αγρεγαρισμένο αποτέλεσμα συγχωνεύεται (left merge) με τις εγγραφές ET, χρησιμοποιώντας τις τρεις στήλες ως κλειδιά. Στο τέταρτο βήμα, τυχόν εναπομένουσες ελλιπείς τιμές coverage, που αντιστοιχούν σε λέξεις χωρίς καμία FT καταγραφή από οποιονδήποτε συμμετέχοντα, πληρώνονται με τον γενικό μέσο όρο του dataset. Μετά την ολοκλήρωση της διαδικασίας, το τελικό καθαρισμένο dataset περιλαμβάνει περίπου 37.377 γραμμές. Καθεμία από αυτές διαθέτει πλήρη ET και FT χαρακτηριστικά. Διευκρινίζεται ρητά ότι η αγρεγαρισμένη προσέγγιση συνιστά μεθοδολογικό συμβιβασμό. Θυσιάζει την εξατομίκευση του FT σήματος χάριν ενός τυπικού coverage σε επίπεδο λέξης. Είναι όμως η μόνη εφικτή οδός συνδυαστικής αξιοποίησης των δύο σημάτων δεδομένης της δομής των αρχικών δεδομένων. Σε μελλοντικές μελέτες με πλήρως συγχρονισμένα δεδομένα, ο συμβιβασμός αυτός μπορεί να αρθεί. Σχετική πρόταση συζητείται στο Κεφάλαιο 8.

3.4 Κατασκευή της μεταβλητής στόχου και του δείκτη lag_proxy

3.4.1 Ορισμός της μεταβλητής word difficulty

Η μεταβλητή στόχος είναι δυαδική και κατασκευάζεται μέσω σύνθετου κριτηρίου που απαιτεί την ταυτόχρονη ικανοποίηση δύο συνθηκών: word difficulty = 1 αν και μόνο αν TRT > median(TRT) ΚΑΙ isReg = 1. Σε κάθε άλλη περίπτωση, word difficulty = 0. Η υιοθέτηση του διπλού κριτηρίου αντί ενός απλού κατωφλίου αιτιολογείται ως εξής. Το TRT αυτοτελώς δεν αρκεί ως κριτήριο δυσκολίας. Ορισμένες λέξεις εμφανίζουν υψηλές τιμές TRT λόγω αυξημένου μήκους ή μορφολογικής πολυπλοκότητας, χωρίς αυτό να σημαίνει υποχρεωτικά γνωστική δυσκολία για τον συγκεκριμένο αναγνώστη. Ο αναγνώστης μπορεί απλώς να τις επεξεργάστηκε με αυξημένη προσοχή. Αντίστοιχα, η μεταβλητή isReg από μόνη της είναι ανεπαρκής. Οι παλινδρομήσεις μπορεί να προκληθούν από γλωσσική αμφισημία που απαιτεί επαναληπτική επεξεργασία ή από μηχανικές συνθήκες, όπως σφαλμένη προσγείωση του βλέμματος. Ο συνδυασμός όμως των δύο κριτηρίων, παρατεταμένος χρόνος επεξεργασίας με συνοδό παλινδρόμηση, συγκροτεί ένα σαφώς ισχυρότερο σήμα γνωστικής ασυνέπειας σε επίπεδο λέξης.

Ο ορισμός αυτός φέρνει σημαντικές μεθοδολογικές συνέπειες. Επειδή η μεταβλητή isReg συμμετέχει στην κατασκευή της μεταβλητής στόχου, δεν μπορεί να αξιοποιηθεί παράλληλα ως χαρακτηριστικό εισόδου στο μοντέλο. Η παράλληλη χρήση θα συνιστούσε κλασική περίπτωση διαρροής πληροφορίας (data leakage), αφού το μοντέλο θα μάθαινε τη μεταβλητή στόχο μέσω του ίδιου του χαρακτηριστικού που τη συνκαθορίζει, παράγοντας πλασματικά υψηλές αποδόσεις χωρίς γενίκευση σε νέα δεδομένα. Στην παρούσα υλοποίηση, η στήλη isReg αφαιρείται ρητά από το feature matrix πριν την εκπαίδευση, με σχετικό σχόλιο στον κώδικα που προλαβαίνει τυχόν μελλοντικές παραλείψεις.

3.4.2 Ο δείκτης lag_proxy ως χαρακτηριστικό εισόδου

Ο λειτουργικός δείκτης lag proxy, που η εννοιολογική του βάση αναπτύχθηκε στην ενότητα 2.4.3, υπολογίζεται σε δύο διαδοχικά βήματα. Στο πρώτο βήμα, η μεταβλητή TRT κανονικοποιείται μέσω MinMax scaling, ώστε όλες οι τιμές να περιορίζονται στο διάστημα [0, 1] και να γίνουν άμεσα συγκρίσιμες με τη μεταβλητή coverage, η οποία βρίσκεται εξ ορισμού στο ίδιο διάστημα. Η κανονικοποίηση είναι απαραίτητη. Χωρίς αυτήν, η διαφορά TRT − coverage θα κυριαρχούνταν από την κλίμακα του TRT (τιμές στην τάξη των εκατοντάδων millisecond) και το coverage (τιμές στο [0, 1]) θα γινόταν πρακτικά αμελητέα συνιστώσα. Στο δεύτερο βήμα, υπολογίζεται η απλή διαφορά: lag proxy = TRT normalized − coverage και η τιμή αποθηκεύεται ως νέα στήλη.

Το αποτέλεσμα είναι ένα χαρακτηριστικό με ερμηνεύσιμη κατανομή γύρω από το μηδέν. Τιμές κοντά στο μηδέν αντιστοιχούν σε συγχρονισμένη λειτουργία των δύο σημάτων, ισορροπία μεταξύ οπτικής επεξεργασίας και κινητικής ροής. Αποκλίσεις από το μηδέν υποδηλώνουν μη τυπικό γεγονός. Θετικές τιμές εμφανίζονται όταν η οπτική επεξεργασία έχει αυξημένη επιβάρυνση (υψηλό TRT normalized) χωρίς αντίστοιχη μείωση του coverage. Δηλαδή το βλέμμα εγκλωβίζεται ενώ το δάκτυλο διατηρεί ομαλή ροή. Αρνητικές τιμές εμφανίζονται στην αντίστροφη περίπτωση, χαμηλή οπτική επιβάρυνση σε συνδυασμό με διστακτική δακτυλική κίνηση. Και στις δύο κατευθύνσεις, ο δείκτης αποτυπώνει πληροφορία σχετική με γνωστική ασυνέπεια. Η πληροφορία αυτή τροφοδοτείται στο μοντέλο ως ένα από τα χαρακτηριστικά

εισόδου και η σχετική σημαντικότητά της ελέγχεται εμπειρικά μέσω της ανάλυσης SHAP, όπως φαίνεται στο Κεφάλαιο 6.

3.5 Επιλογή χαρακτηριστικών εισόδου

Το τελικό σύνολο χαρακτηριστικών που τροφοδοτούν τα μοντέλα περιλαμβάνει συνειδητά επιλεγμένο συνδυασμό γλωσσικών, οφθαλμικών, κινητικών, λειτουργικών και δημογραφικών μεταβλητών. Στα γλωσσικά χαρακτηριστικά περιλαμβάνονται το len (μήκος λέξης σε χαρακτήρες), η freq (συχνότητα στο corpus) και ο orthoSyllablesCount (αριθμός συλλαβών). Στα οφθαλμικά χαρακτηριστικά περιλαμβάνονται τα FFD, FPD, TRT, RPD και fixNum. Στα κινητικά χαρακτηριστικά περιλαμβάνεται το coverage. Στα λειτουργικά χαρακτηριστικά ο δείκτης lag proxy. Τέλος, στα χαρακτηριστικά συμμετέχοντα περιλαμβάνονται το idSession, το Education Level και τρεις ψυχομετρικοί δείκτες WAIS (Coding, Digit Span Ascending, Vocabulary) που παρέχουν πληροφορία για τις γνωστικές ικανότητες του αναγνώστη.

Από το feature set εξαιρούνται ρητά τα ακόλουθα. Η μεταβλητή isReg, για τους λόγους αποφυγής διαρροής που αναπτύχθηκαν στην ενότητα 3.4.1. Η μεταβλητή idUser, για την αποφυγή υπερπροσαρμογής σε συγκεκριμένους συμμετέχοντες (το μοντέλο θα μάθαινε ιδιοσυγκρασίες αναγνωστών που δεν γενικεύονται σε νέα άτομα). Όλα τα αναγνωριστικά κειμένου (gid, sid, tid, idDoc), που είναι δείκτες θέσης χωρίς σημασιολογικό περιεχόμενο σχετικό με τη δυσκολία της λέξης. Επίσης δεν συμπεριλαμβάνονται οι μεταβλητές token και lemma ως χαρακτηριστικά κειμένου. Η αρχιτεκτονική των επιλεγμένων μοντέλων δεν υποστηρίζει εγγενώς επεξεργασία κειμένου και η σχετική γλωσσολογική πληροφορία κωδικοποιείται μέσω του μήκους και της συχνότητας. Η επιλογή αυτή διατηρεί τη φύση του προβλήματος αυστηρά δομημένη και συμβατή με τους αλγορίθμους που αξιοποιούνται.

	count	mean	std	min	25%	50%	75%	max
len	61594.0	4.912	3.340	1.00	2.000	4.000	7.000	20.000
freq	61594.0	37448.760	154263.159	0.00	3.000	48.000	3396.000	864641.000
orthoSyllablesCount	61594.0	1.793	1.403	0.00	1.000	1.000	3.000	9.000
FFD	61594.0	0.237	0.089	0.08	0.175	0.221	0.282	0.904
FPD	61594.0	0.329	0.164	0.08	0.207	0.291	0.413	1.130
TRT	61594.0	0.362	0.177	0.08	0.226	0.326	0.462	1.250
RPD	61594.0	0.368	0.183	0.08	0.229	0.328	0.469	1.130
fixNum	61594.0	1.591	0.729	1.00	1.000	1.333	2.000	6.000
coverage	61594.0	0.854	0.104	0.15	0.818	0.879	0.923	1.000
lag_proxy	61594.0	-0.613	0.183	-0.99	-0.751	-0.644	-0.506	0.562
idSession	61594.0	4080.665	39.843	4012.00	4047.000	4073.000	4114.000	4151.000
Education Level	61594.0	2.741	2.078	0.00	0.000	3.000	5.000	5.000
WAIS Coding	61594.0	70.287	13.118	36.00	61.000	70.000	78.000	102.000
WAIS Digit Span Ascending	61594.0	9.289	2.295	5.00	8.000	9.000	11.000	14.000
WAIS Vocabulary	61594.0	35.094	7.954	12.00	30.000	35.000	41.000	49.000
word_difficulty	61594.0	0.044	0.204	0.00	0.000	0.000	0.000	1.000

Πίνακας 1: Περιγραφικά στατιστικά στοιχεία των χαρακτηριστικών εισόδου features και της μεταβλητής στόχου word difficulty.

3.6 Αντιμετώπιση ανισοκατανομής κλάσεων

Η κατανομή της μεταβλητής στόχου είναι έντονα ανισοκατανεμημένη, με την αρνητική κλάση να αντιπροσωπεύει περίπου το 94% και τη θετική το 5 με 6%. Αυτό είναι αναμενόμενο από την ίδια τη φύση του ορισμού. Για την εξισορρόπηση εφαρμόζεται SMOTE μόνο στο σύνολο εκπαίδευσης μετά τον διαχωρισμό train/test (Chawla et al., 2002). Το σύνολο δοκιμής διατηρείται αναλλοίωτο, με την αρχική ανισοκατανομή, ώστε η αξιολόγηση να αντικατοπτρίζει τις πραγματικές συνθήκες εφαρμογής του μοντέλου.

3.7 Αρχιτεκτονική των μοντέλων

3.7.1 Ρυθμίσεις XGBoost

Για το XGBoost υιοθετούνται οι ακόλουθες ρυθμίσεις: n estimators = 300 (αριθμός δέντρων), max depth = 6 (τιμή που αποτρέπει υπερπροσαρμογή σε μικρού και μεσαίου μεγέθους datasets), learning rate = 0.1 (συντηρητικός ρυθμός εκπαίδευσης), objective = 'binary:logistic' (στόχος δυαδικής ταξινόμησης) και eval metric = 'logloss'. Εκτεταμένη βελτιστοποίηση υπερπαραμέτρων μέσω grid search ή Bayesian optimization δεν πραγματοποιείται. Η επιλογή είναι συνειδητή και μεθοδολογικά αιτιολογημένη. Σκοπός της μελέτης δεν είναι η εξεύρεση απολύτως βέλτιστης διαμόρφωσης που θα μεγιστοποιήσει το F1 score κατά οριακές τιμές, αλλά η μελέτη της συμπεριφοράς του μοντέλου υπό τυπικές, ρεαλιστικές συνθήκες εφαρμογής.

Η διαπίστωση ότι, ακόμα και με σχετικά default τιμές, ο αλγόριθμος επιτυγχάνει ικανοποιητική και ερμηνεύσιμη απόδοση, αποτελεί από μόνη της ενδιαφέρουσα παρατήρηση. Επιπλέον, η εκτεταμένη βελτιστοποίηση σε ανισοκατανεμημένα δεδομένα τείνει να παράγει αποτελέσματα ευαίσθητα στον συγκεκριμένο διαχωρισμό train/test, επηρεάζοντας αρνητικά την αξιοπιστία της γενίκευσης.

3.7.2 Ρυθμίσεις MLP

Το MLP υλοποιείται μέσω της κλάσης MLPClassifier της βιβλιοθήκης scikit learn με τις εξής ρυθμίσεις: hidden layer sizes = (128, 64, 32), δηλαδή τρία κρυφά επίπεδα σε φθίνουσα διάταξη που συμπιέζουν σταδιακά την αναπαράσταση. activation = 'relu', η πρακτικά καθιερωμένη συνάρτηση ενεργοποίησης για δίκτυα αυτού του μεγέθους. solver = 'adam', με προσαρμοστικό ρυθμό εκπαίδευσης (Kingma και Ba, 2015). Max iter = 300, ώστε να διασφαλίζεται σύγκλιση χωρίς υπερβολική υπολογιστική επιβάρυνση. Random state = 42, για αναπαραγωγιμότητα. Τα χαρακτηριστικά εισόδου του MLP κανονικοποιούνται μέσω StandardScaler πριν από την εκπαίδευση, σε αντίθεση με το XGBoost. Ο scaler προσαρμόζεται (fit) μόνο στο σύνολο εκπαίδευσης και εφαρμόζεται (transform) τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής. Είναι θεμελιώδης αρχή για την αποφυγή διαρροής πληροφορίας. Ο εκπαιδευμένος scaler αποθηκεύεται ως αρχείο scaler.joblib, ώστε να επαναχρησιμοποιείται στο production API για τη μετατροπή νέων δεδομένων που υποβάλλονται από τον χρήστη.

3.8 Πρωτόκολλο εκπαίδευσης και αξιολόγησης

Το σύνολο δεδομένων διαχωρίζεται σε σύνολα εκπαίδευσης και δοκιμής με αναλογία 80/20, μέσω stratified split, ώστε η αναλογία θετικής προς αρνητικής κλάσης να μένει σταθερή και στα δύο σύνολα. Σε περιπτώσεις έντονης ανισοκατανομής, ένας τυχαίος διαχωρισμός χωρίς stratification μπορεί να παράγει σύνολο δοκιμής με ελάχιστες ή μηδενικές θετικές παρατηρήσεις,

καθιστώντας την αξιολόγηση αναξιόπιστη. Η εφαρμογή stratified split εξασφαλίζει ότι το σύνολο δοκιμής διατηρεί αναλογική αντιπροσώπευση της μειονοτικής κλάσης (περίπου 5 με 6% θετικά παραδείγματα).

Η αξιολόγηση γίνεται μέσω πολλαπλών μετρικών, καθεμία με διακριτό ρόλο. Η Accuracy καταγράφεται για λόγους πληρότητας, αλλά δεν αξιοποιείται ως κύριο κριτήριο σύγκρισης, αφού σε ανισοκατανεμημένα δεδομένα παραπλανά συστηματικά (accuracy 94% μπορεί να αντιστοιχεί σε τετριμμένη μόνιμη πρόβλεψη της αρνητικής κλάσης). Το Precision της θετικής κλάσης ποσοτικοποιεί το ποσοστό των λέξεων που το μοντέλο χαρακτήρισε ορθά ως δύσκολες, εκφράζοντας τον βαθμό false alarm. Το Recall της θετικής κλάσης ποσοτικοποιεί το ποσοστό των πραγματικά δύσκολων λέξεων που το μοντέλο εντόπισε. Το F1 score, ως αρμονικός μέσος Precision και Recall, εξισορροπεί τις δύο διαστάσεις και αξιοποιείται ως η κύρια μετρική σύγκρισης. Τέλος, ο confusion matrix καταγράφεται για να δώσει πλήρη εικόνα των τύπων σφάλματος (false positives, false negatives), πληροφορία χρήσιμη στη συζήτηση των αποτελεσμάτων (Κεφάλαιο 7). Δεδομένης της ανισοκατανομής, η σύγκριση μεταξύ των μοντέλων βασίζεται πρωτίστως στο F1 score της θετικής κλάσης, ενώ οι υπόλοιπες μετρικές λειτουργούν υποστηρικτικά.

3.9 Ανάλυση cross correlation ET–FT

Παράλληλα με τη μοντελοποίηση XGBoost και MLP, διεξάγεται περιγραφική ανάλυση crosscorrelation, που στοχεύει στη μελέτη της συνολικής σχέσης μεταξύ των σημάτων ET και FT σε μακροσκοπικό επίπεδο, δηλαδή στο σύνολο του dataset. Συγκεκριμένα, κατασκευάζονται δύο χρονοσειρές. Η πρώτη βασίζεται στον μέσο όρο TRT ανά σειρά εμφάνισης λέξης στο corpus και η δεύτερη στον μέσο όρο coverage αντίστοιχα. Και οι δύο χρονοσειρές κανονικοποιούνται για λόγους συγκρισιμότητας και στη συνέχεια υπολογίζεται η μεταξύ τους συνάρτηση crosscorrelation. Το αποτέλεσμα παρουσιάζεται σε γραφική μορφή στο Σχήμα 3, με τον οριζόντιο άξονα να αντιστοιχεί σε χρονική ή θεσιακή μετατόπιση (lag) και τον κατακόρυφο στη συσχέτιση ανά μετατόπιση. Η θέση του μεγίστου αποτυπώνει την εκτιμώμενη μακροσκοπική καθυστέρηση μεταξύ των δύο σημάτων. Η ανάλυση δεν υποκαθιστά τον δείκτη lag proxy, αφού δεν παρέχει δείκτη ανά λέξη. Λειτουργεί συμπληρωματικά και προσφέρει συνολική οπτική που εμπλουτίζει την ερμηνεία των αποτελεσμάτων στο Κεφάλαιο 6.

3.10 Ερμηνευσιμότητα μέσω SHAP

Μετά την ολοκλήρωση της εκπαίδευσης του XGBoost, εφαρμόζεται ο TreeExplainer του πακέτου SHAP σε τυχαίο δείγμα 200 παρατηρήσεων από το σύνολο δοκιμής, με στόχο την εξαγωγή των SHAP values. Η χρήση δείγματος αντί του πλήρους συνόλου δοκιμής δικαιολογείται από πρακτικούς λόγους. Ο υπολογισμός των SHAP values, ακόμα και με τον βελτιστοποιημένο TreeExplainer, έχει μη τετριμμένο υπολογιστικό κόστος. Για τους σκοπούς οπτικοποίησης και εξαγωγής συμπερασμάτων σε επίπεδο συνολικής σημαντικότητας, το μέγεθος των 200 τυχαίων δειγμάτων αρκεί για σταθερά και αξιόπιστα αποτελέσματα. Ο TreeExplainer είναι εξειδικευμένος για μοντέλα βασισμένα σε δέντρα, όπως το XGBoost, και παρέχει ακριβείς τιμές Shapley σε πολυωνυμικό χρόνο, σε αντίθεση με τον γενικευμένο KernelExplainer που είναι προσεγγιστικός και αρκετά βραδύτερος (Lundberg, Erion και Lee, 2018).

Από την ανάλυση παράγονται δύο τύποι γραφημάτων. Καθένας απαντά σε διακριτό ερώτημα. Το SHAP summary plot (Σχήμα 1) αναπαριστά τα 15 πιο σημαντικά χαρακτηριστικά. Κάθε σημείο αντιστοιχεί σε μία παρατήρηση. Η οριζόντια θέση δείχνει το SHAP value, δηλαδή την κατευθυντήρια επίδραση στην πρόβλεψη, ενώ ο χρωματικός κωδικός δείχνει την τιμή του χαρακτηριστικού (κόκκινο = υψηλή, μπλε = χαμηλή). Η οπτικοποίηση αυτή επιτρέπει την ταυτόχρονη παρατήρηση της σημαντικότητας και της κατεύθυνσης της επίδρασης κάθε χαρακτηριστικού, πληροφορία που δεν είναι προσβάσιμη μέσω κλασικών τεχνικών feature importance. Το SHAP dependence plot (Σχήμα 2) εστιάζει στο σημαντικότερο χαρακτηριστικό σύμφωνα με το summary plot και απεικονίζει την εξάρτηση του SHAP value από την τιμή του χαρακτηριστικού, αποκαλύπτοντας ενδεχόμενες μη γραμμικές σχέσεις και κατώφλια. Οι δύο οπτικοποιήσεις, μαζί, επιτρέπουν την απάντηση στο έκτο ερευνητικό ερώτημα της μελέτης. Προσδιορίζουν ποια χαρακτηριστικά συμβάλλουν περισσότερο στην πρόβλεψη της αναγνωστικής δυσκολίας, με ποιον τρόπο επιδρούν, και ειδικότερα τη σχετική θέση του δείκτη lag proxy στη συνολική εικόνα. Το ζητούμενο δεν περιορίζεται στην απόδοση του μοντέλου, αλλά επεκτείνεται στην εξαγωγή επιστημονικά αξιοποιήσιμων συμπερασμάτων για το ίδιο το φαινόμενο της ανάγνωσης.

Κεφάλαιο 4. Αιτιολόγηση και ανάπτυξη μοντέλων

4.1 Μεθοδολογική αιτιολόγηση επιλογής αλγορίθμων

Η επιλογή των αλγορίθμων XGBoost και MLP δεν είναι τυχαία. Στηρίζεται σε συγκεκριμένους μεθοδολογικούς λόγους που αναπτύσσονται στην ενότητα αυτή. Η φύση του προβλήματος, ταξινόμηση δυσκολίας σε επίπεδο λέξης με δομημένα χαρακτηριστικά, επιβάλλει προσεκτική αξιολόγηση των διαθέσιμων επιλογών. Απορρίπτονται αλγόριθμοι όπως οι συνελικτικές αρχιτεκτονικές (CNN), που έχουν σχεδιαστεί για εικόνες και χωρικές δομές, και οι Transformers, που απαιτούν ακολουθιακά δεδομένα ή μεγάλο όγκο κειμενικών αναπαραστάσεων. Αλγόριθμοι όπως οι Support Vector Machines (SVM) αντιμετωπίζουν δυσκολίες κλιμάκωσης σε dataset της τάξης των 37.000 παρατηρήσεων, ενώ ο Logistic Regression δεν ανταποκρίνεται στη μη γραμμικότητα του φαινομένου.

Η επιλογή του XGBoost στηρίζεται σε τέσσερις λόγους. Πρώτον, η πρόσφατη εμπειρική βιβλιογραφία δείχνει συστηματικά ότι οι αλγόριθμοι gradient boosting υπερτερούν σε δομημένα δεδομένα (Shwartz-Ziv και Armon, 2022· Grinsztajn, Oyallon και Varoquaux, 2022). Δεύτερον, η φύση του XGBoost ως ensemble δέντρων αποφάσεων το κάνει ανθεκτικό σε πολλαπλές κλίμακες χαρακτηριστικών, ελλιπείς τιμές και ανισοκατανομή. Τρίτον, ο αλγόριθμος παρέχει εγγενείς μηχανισμούς κανονικοποίησης (L1 και L2) που αποτρέπουν την υπερπροσαρμογή. Τέταρτον, και ιδιαίτερα σημαντικό για την παρούσα μελέτη, είναι πλήρως συμβατός με τον TreeExplainer του SHAP, που παρέχει ακριβείς τιμές Shapley σε πολυωνυμικό χρόνο και κάνει την ερμηνευσιμότητα πρακτικά εφικτή (Lundberg, Erion και Lee, 2018).

Η επιλογή του MLP στηρίζεται σε διαφορετική λογική. Το MLP ενσωματώνεται ως σημείο αναφοράς (baseline) για τη συγκριτική αξιολόγηση, όχι ως πρωταρχικός υποψήφιος νικητής. Η σύγκριση είναι μεθοδολογικά απαραίτητη για δύο λόγους. Πρώτον, επιτρέπει την εμπειρική

τεκμηρίωση της υπεροχής του XGBoost υπό ταυτόσημες συνθήκες, αντί απλής επίκλησης της βιβλιογραφίας. Δεύτερον, το MLP αντιπροσωπεύει την οικογένεια των νευρωνικών δικτύων και η συστηματική σύγκρισή του με το XGBoost τεκμηριώνει εμπειρικά ότι η επικρατέστερη επιλογή δεν είναι πάντοτε η καταλληλότερη για κάθε τύπο δεδομένων.

4.2 Αρχιτεκτονική και ρυθμίσεις XGBoost

Όπως αναφέρθηκε και στην ενότητα 3.7.1, το XGBoost εκπαιδεύεται με τις παραμέτρους n estimators = 300, max depth = 6, learning rate = 0.1, objective = 'binary:logistic' και eval metric = 'logloss'. Η ροή εκπαίδευσης ακολουθεί την κλασική επαναληπτική φιλοσοφία του gradient boosting. Σε κάθε γύρο, υπολογίζεται η παράγωγος της συνάρτησης απώλειας ως προς την τρέχουσα πρόβλεψη και ένα νέο δέντρο εκπαιδεύεται για να εκτιμήσει αυτή την παράγωγο. Το νέο δέντρο διορθώνει τα σφάλματα του αθροιστικού μοντέλου με μικρό βάρος, που καθορίζει το learning rate. Οι 300 γύροι κρίθηκαν επαρκείς για να ωριμάσει το μοντέλο χωρίς να εμφανίζει υπερπροσαρμογή. Το ελεγχόμενο βάθος (max depth = 6) εξασφαλίζει ότι κάθε δέντρο παραμένει σχετικά απλό και εστιάζει σε κανονικότητες που γενικεύονται.

Εκτεταμένη βελτιστοποίηση υπερπαραμέτρων μέσω grid search ή Bayesian optimization δεν πραγματοποιείται. Η επιλογή είναι συνειδητή. Σκοπός της μελέτης δεν είναι η εξεύρεση απολύτως βέλτιστης διαμόρφωσης, αλλά η μελέτη της συμπεριφοράς του μοντέλου υπό τυπικές, ρεαλιστικές συνθήκες εφαρμογής. Η διαπίστωση ότι, ακόμα και με σχετικά default τιμές, ο αλγόριθμος επιτυγχάνει ικανοποιητική και ερμηνεύσιμη απόδοση, αποτελεί από μόνη της ενδιαφέρουσα παρατήρηση. Επιπλέον, η εκτεταμένη βελτιστοποίηση σε ανισοκατανεμημένα δεδομένα τείνει να παράγει αποτελέσματα ευαίσθητα στο συγκεκριμένο διαχωρισμό train/test, επηρεάζοντας αρνητικά την αξιοπιστία της γενίκευσης.

4.3 Αρχιτεκτονική και ρυθμίσεις MLP

Το MLP υλοποιείται μέσω της κλάσης MLPClassifier της scikit-learn με τις ρυθμίσεις που αναλύθηκαν στην ενότητα 3.7.2. Η αρχιτεκτονική 128-64-32 είναι η καρδιά του δικτύου και η βασική σχεδιαστική απόφαση. Πρόκειται για funnel architecture, μια κωνική δομή που σταδιακά μειώνει τον αριθμό των νευρώνων από επίπεδο σε επίπεδο. Λειτουργεί ως μέθοδος σταδιακής μείωσης της πληροφορίας για εξαγωγή ουσιωδών χαρακτηριστικών.

Στο πρώτο κρυφό επίπεδο με 128 νευρώνες, το δίκτυο δέχεται την ακατέργαστη πληροφορία από όλα τα 16 περίπου χαρακτηριστικά εισόδου. Με τόσους νευρώνες, αποκτά πλούσια χωρητικότητα να ανιχνεύσει βασικά μοτίβα. Διαφορετικοί νευρώνες ενεργοποιούνται για διαφορετικούς συνδυασμούς γλωσσικών, οφθαλμικών και κινητικών χαρακτηριστικών. Σε αυτή τη φάση, το δίκτυο δεν έχει ακόμη επιλέξει τι είναι σημαντικό. Απλώς δομεί ένα ευρύ πληροφοριακό υπόβαθρο πάνω στο οποίο θα δουλέψουν τα επόμενα επίπεδα.

Στο δεύτερο κρυφό επίπεδο με 64 νευρώνες, ο αριθμός των μονάδων μειώνεται στο μισό. Αυτή η συμπίεση είναι σκόπιμη. Αναγκάζει το δίκτυο να συνδυάσει τα πρωταρχικά μοτίβα του πρώτου επιπέδου σε αναπαραστάσεις υψηλότερου επιπέδου, όπου η πληροφορία είναι πιο πυκνή και πιο σχετική με τη μεταβλητή στόχο. Όσοι νευρώνες ενεργοποιούνται εδώ αντιστοιχούν σε πιο αφηρημένα γνωρίσματα, για παράδειγμα έναν συνδυασμό υψηλού TRT, υψηλού lag proxy και

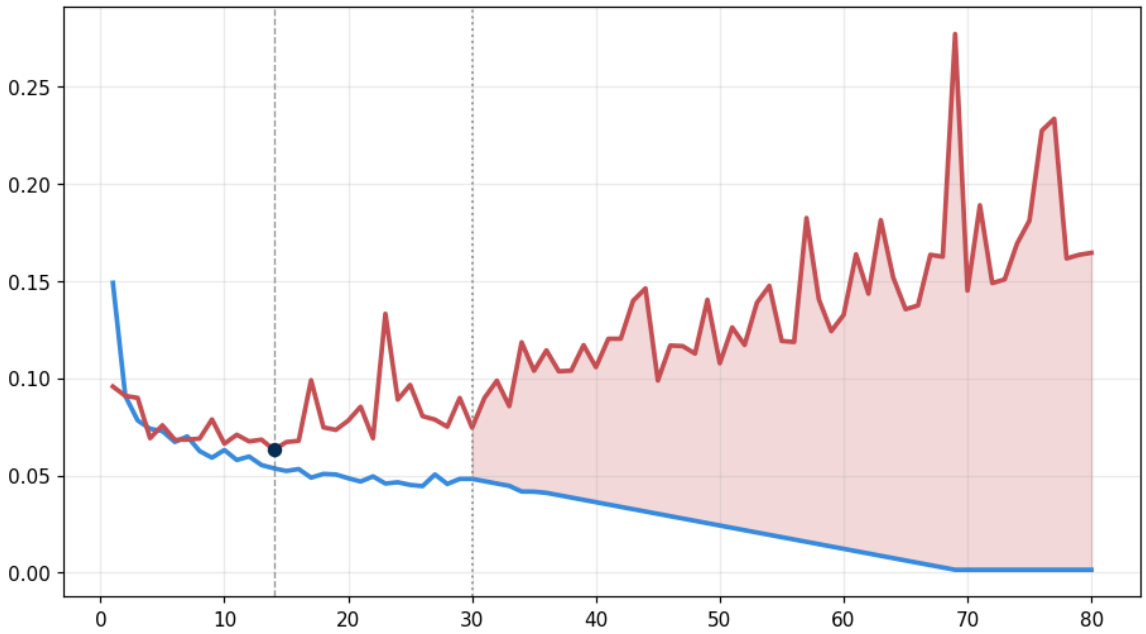
χαμηλής συχνότητας λέξης που σχηματίζει ένα κοινό μοτίβο δυσκολίας. Με αυτόν τον τρόπο, το δίκτυο φιλτράρει τον θόρυβο και διατηρεί τα ουσιώδη.

Στο τρίτο κρυφό επίπεδο με 32 νευρώνες, η αναπαράσταση γίνεται ακόμη πιο συμπυκνωμένη. Σε αυτό το βαθύ σημείο, το δίκτυο κρατά μόνο τα στοιχεία που είναι αυστηρά απαραίτητα για να πάρει τη δυαδική απόφαση εύκολη ή δύσκολη λέξη. Είναι σαν ένα φίλτρο που αποστάζει τα πιο κρίσιμα γνωρίσματα και τα τροφοδοτεί στο επίπεδο εξόδου. Η συνολική λογική 128 → 64 → 32 σχηματίζει επομένως μια αλυσίδα διαδοχικής αφαίρεσης. Από πολλά ευρέα μοτίβα, σε λιγότερα συμπυκνωμένα, σε πολύ λίγα ουσιώδη χαρακτηριστικά. Αν διατηρούσαμε σταθερό αριθμό νευρώνων σε όλα τα επίπεδα, ή αν αυξάναμε τον αριθμό προς τα έξω, το δίκτυο θα είχε υπερβολική χωρητικότητα και θα μάθαινε ιδιοσυγκρασίες των δεδομένων εκπαίδευσης. Η σταδιακή συμπίεση αναγκάζει το δίκτυο να αποσπά την πραγματικά γενικεύσιμη πληροφορία.

Συνολικά, η αρχιτεκτονική 128-64-32 ταιριάζει στη φύση των δεδομένων μας, όπου ο πραγματικός αριθμός πληροφοριακών διαστάσεων είναι σχετικά μικρός και η πρόβλεψη απαιτεί καθαρή απομόνωση των ουσιωδών στοιχείων. Είναι μια κλασική επιλογή για προβλήματα δυαδικής ταξινόμησης σε δομημένα δεδομένα και ταυτόχρονα δικαιολογημένη με βάση την αρχή της σταδιακής μείωσης της πληροφορίας. Τα χαρακτηριστικά εισόδου του MLP κανονικοποιούνται μέσω StandardScaler πριν από την εκπαίδευση. Ο scaler προσαρμόζεται μόνο στο σύνολο εκπαίδευσης και αποθηκεύεται για επαναχρησιμοποίηση στο production API.

4.4 Εκπαιδευτική διαδικασία και αποφυγή υπερπροσαρμογής

Η εκπαιδευτική διαδικασία ακολουθεί αυστηρό πρωτόκολλο για να αποτρέψει την **υπερπροσαρμογή (overfitting)**: την κατάσταση όπου το μοντέλο αποστηθίζει τα παραδείγματα εκπαίδευσης συμπεριλαμβανομένου θορύβου, χάνοντας την ικανότητα γενίκευσης σε νέα δεδομένα. Στο Σχήμα 2 η διαφορά είναι εμφανής: το training loss μειώνεται συνεχώς, ενώ το validation loss φτάνει σε ελάχιστο και στη συνέχεια αυξάνεται. Βέλτιστη παύση είναι το ελάχιστο της καμπύλης επικύρωσης.Μετά τον αρχικό διαχωρισμό σε σύνολα εκπαίδευσης και δοκιμής (80/20 με stratified split), εφαρμόζεται διασταυρωμένη επικύρωση 5 fold αποκλειστικά στο σύνολο εκπαίδευσης. Η τεχνική αυτή χωρίζει το training set σε πέντε ίσα τμήματα και εκπαιδεύει το μοντέλο πέντε φορές, κάθε φορά αφήνοντας διαφορετικό τμήμα ως σύνολο επικύρωσης. Έτσι, η απόδοση υπολογίζεται ως μέσος όρος πέντε διαφορετικών διαχωρισμών και μειώνεται η ευαισθησία σε συγκεκριμένες τυχαίες επιλογές.



Σχήμα 2 — Καμπύλες απώλειας εκπαίδευσης και επικύρωσης.

Επιπλέον μέτρα για την αποτροπή υπερπροσαρμογής περιλαμβάνουν τα ακόλουθα. Για το XGBoost, χρήση ενσωματωμένης κανονικοποίησης L1/L2 και περιορισμός του μέγιστου βάθους (max depth = 6). Για το MLP, χρήση της παραμέτρου early stopping μέσω του εσωτερικού validation set του MLPClassifier, που σταματά την εκπαίδευση όταν δεν παρατηρείται περαιτέρω βελτίωση. Για αμφότερα τα μοντέλα, αποκλειστική εφαρμογή της SMOTE στο training set, ώστε το σύνολο δοκιμής να παραμένει αναλλοίωτο και να αντικατοπτρίζει τις πραγματικές συνθήκες εφαρμογής. Ο τελικός έλεγχος γενίκευσης γίνεται στο σύνολο δοκιμής, το οποίο δεν έχει εκτεθεί ποτέ κατά τη διάρκεια της εκπαίδευσης.

Κεφάλαιο 5. Σχεδιασμός και υλοποίηση συστήματος

5.1 Συνολική αρχιτεκτονική της εφαρμογής THE-LAG

Η εφαρμογή THE-LAG υλοποιεί την αναλυτική ροή της μελέτης ως ολοκληρωμένο διαδικτυακό σύστημα. Η αρχιτεκτονική ακολουθεί το κλασικό τριεπίπεδο πρότυπο (three-tier architecture), που χωρίζει την παρουσίαση, τη λογική της εφαρμογής και τα δεδομένα. Το επίπεδο παρουσίασης (frontend) εκτελείται στον φυλλομετρητή του χρήστη και υλοποιείται σε HTML, CSS και JavaScript χωρίς εξωτερικά frameworks. Το επίπεδο εφαρμογής (backend) υλοποιείται σε Python με το framework FastAPI και εκτελείται σε διακομιστή στο νέφος. Το επίπεδο δεδομένων περιλαμβάνει τη βάση δεδομένων SQLite για τους χρήστες, καθώς και τα αρχεία των εκπαιδευμένων μοντέλων και των αποτελεσμάτων (model xgb.joblib, model mlp.joblib, scaler.joblib, metrics.json, shap summary.png).

Η επικοινωνία μεταξύ των επιπέδων γίνεται μέσω τυπικού REST API για τις απλές κλήσεις (login, register, μεταφόρτωση αρχείου, λήψη μετρικών) και μέσω Server Sent Events (SSE) για τη μετάδοση της προόδου της αναλυτικής ροής σε πραγματικό χρόνο. Η επιλογή του SSE αντί του WebSocket στηρίζεται στη μονοκατευθυντικότητα της επικοινωνίας στη συγκεκριμένη περίπτωση. Ο διακομιστής ενημερώνει τον πελάτη για την πρόοδο της εκπαίδευσης, χωρίς να χρειάζεται αμφίδρομη επικοινωνία. Επιπλέον, το SSE είναι απλούστερο στην υλοποίηση και πλήρως συμβατό με το υπάρχον HTTP πρωτόκολλο.

5.2 Backend: αναλυτικό επίπεδο και REST API

Το backend υλοποιείται σε Python 3.11 με το framework FastAPI, που επιλέχθηκε για ασύγχρονη υποστήριξη (async/await), αυτόματη παραγωγή τεκμηρίωσης OpenAPI, ενσωματωμένη επικύρωση δεδομένων μέσω Pydantic και υψηλή απόδοση. Η αρχιτεκτονική του backend οργανώνεται σε επτά διακριτά modules, με σαφώς ορισμένη αρμοδιότητα: κεντρικές ρυθμίσεις, καθαρισμός και προεπεξεργασία, εκπαίδευση μοντέλων, υπολογισμός μετρικών, ανάλυση SHAP, ανάλυση cross-correlation και ορισμός endpoints.

Τα κύρια endpoints του API είναι τα ακόλουθα. Το POST /upload-and-run δέχεται αρχείο Excel ή CSV, αποθηκεύει το αρχείο στον διακομιστή και εκτελεί διαδοχικά την πλήρη αναλυτική ροή. Το POST /register και το POST /login διαχειρίζονται την αυθεντικότητα των χρηστών, αποθηκεύοντας κατακερματισμένους κωδικούς στη βάση SQLite. Το GET /metrics επιστρέφει τα τελευταία αποτελέσματα αξιολόγησης σε μορφή JSON. Το GET /shap-summary επιστρέφει το γράφημα SHAP ως εικόνα PNG. Τέλος, το POST /predict δέχεται ένα διάνυσμα χαρακτηριστικών και επιστρέφει τις προβλέψεις και των δύο μοντέλων.

5.3 Διασύνδεση frontend με τα εκπαιδευμένα μοντέλα

Η διασύνδεση μεταξύ του frontend και των εκπαιδευμένων μοντέλων ακολουθεί ξεκάθαρη ροή. Όταν ο χρήστης μεταφορτώνει ένα αρχείο δεδομένων, το frontend αποστέλλει αίτημα multipart/form-data στο endpoint /upload-and-run. Το backend αποθηκεύει το αρχείο και εκτελεί διαδοχικά τα στάδια της αναλυτικής ροής. Παράλληλα, ανοίγεται μια σύνδεση SSE μέσω του endpoint /stream-progress, από την οποία το frontend λαμβάνει ενημερώσεις προόδου σε πραγματικό χρόνο.

Μετά την ολοκλήρωση της ροής, το frontend καλεί διαδοχικά τα endpoints GET /metrics και GET /shap-summary για να ανακτήσει και να εμφανίσει τα αποτελέσματα. Οι αριθμητικές μετρικές παρουσιάζονται μέσω της βιβλιοθήκης Chart.js σε διαδραστικά γραφήματα, ενώ το SHAP plot εμφανίζεται απευθείας ως εικόνα. Για την εκτέλεση μεμονωμένης πρόβλεψης, το frontend αποστέλλει JSON αίτημα στο POST /predict με το διάνυσμα χαρακτηριστικών. Το backend φορτώνει τα προεκπαιδευμένα μοντέλα και επιστρέφει τις προβλέψεις μαζί με τις πιθανότητες.

5.4 Frontend: διεπαφή χρήστη και οπτικοποιήσεις

Το frontend υλοποιείται με βασικές τεχνολογίες ιστού χωρίς εξαρτήσεις σε μεγάλα frameworks. Η επιλογή αιτιολογείται από τον ερευνητικό χαρακτήρα της εφαρμογής και την ανάγκη για μειωμένο χρόνο φόρτωσης. Ο σχεδιασμός ακολουθεί μινιμαλιστική αισθητική με κύρια γραμματοσειρά DM Sans και συμπληρωματική Space Mono για τεχνικά στοιχεία. Η δομή της

σελίδας περιλαμβάνει hero section με την εισαγωγή στο έργο, navbar με εναλλαγή γλωσσών (Ελληνικά / Αγγλικά) μέσω inline SVG σημαιών, ενότητα μεταφόρτωσης αρχείου με ζώνη dragand-drop, πάνελ προόδου με οπτική αναπαράσταση της τρέχουσας φάσης και ενότητα αποτελεσμάτων με διαδραστικά γραφήματα.

Τεχνικές λεπτομέρειες περιλαμβάνουν τη χρήση της βιβλιοθήκης Three.js για animated wave grid background, της βιβλιοθήκης Chart.js για τα γραφήματα μετρικών και κατανομών και του API IntersectionObserver για το βαθμιαίο scroll reveal των ενοτήτων. Η επιλογή του API backend πραγματοποιείται δυναμικά μέσω ελέγχου του window.location.hostname. Σε περιβάλλον development (localhost) το frontend επικοινωνεί με τοπικό backend στη θύρα 8000, ενώ σε production επικοινωνεί με τον απομακρυσμένο διακομιστή στο Render.

5.5 Αυθεντικότητα των χρηστών και αποθήκευση

Η αυθεντικότητα των χρηστών υλοποιείται με ελαφρύ μηχανισμό βασισμένο σε SHA-256 κατακερματισμό κωδικών και τοπική αποθήκευση στον φυλλομετρητή. Κατά την εγγραφή (endpoint /register), το username και ο κατακερματισμένος κωδικός αποθηκεύονται στον πίνακα users της βάσης δεδομένων SQLite (thelag.db). Κατά τη σύνδεση (endpoint /login), το backend συγκρίνει τον κατακερματισμό του υποβαλλόμενου κωδικού με τον αποθηκευμένο και σε περίπτωση επιτυχίας επιστρέφει το username στο frontend, που το αποθηκεύει στο localStorage. Η αποσύνδεση γίνεται αποκλειστικά στην πλευρά του πελάτη, μέσω διαγραφής του localStorage.

5.6 Deployment και περιορισμοί υποδομής

Το σύστημα αναπτύσσεται (deployed) σε δύο διακριτές υπηρεσίες νέφους. Το backend εκτελείται στην υπηρεσία Render (διεύθυνση: https://thesis-project-4qk7.onrender.com), η οποία παρέχει δωρεάν φιλοξενία Python εφαρμογών με αυτόματο deployment από repository GitHub. Το frontend φιλοξενείται στην υπηρεσία Netlify, η οποία παρέχει CDN και αυτόματη συμπίεση στατικών αρχείων, μαζί με αγορασμένο domain που προσφέρει επαγγελματική παρουσία. Ο πηγαίος κώδικας φιλοξενείται στο αποθετήριο GitHub s0uro/THESIS-2.

Η επιλογή του Render επιφέρει ορισμένους πρακτικούς περιορισμούς που πρέπει να επισημανθούν. Ο διακομιστής εισέρχεται σε κατάσταση αδράνειας μετά από 15 λεπτά ανενέργειας, προκαλώντας καθυστέρηση 30 με 60 δευτερολέπτων στην πρώτη επόμενη αίτηση (cold start). Επίσης, οι διαθέσιμοι πόροι CPU και RAM είναι περιορισμένοι (0.1 vCPU, 512 MB RAM), καθιστώντας την εκτέλεση της πλήρους αναλυτικής ροής σχετικά αργή (ενδεικτικά 3 με 5 λεπτά για το πλήρες dataset). Οι περιορισμοί συζητούνται αναλυτικά στο Κεφάλαιο 8.

Κεφάλαιο 6. Αποτελέσματα και ανάλυση SHAP

6.1 Περιγραφικά στατιστικά του dataset

Μετά την προεπεξεργασία, το τελικό καθαρισμένο dataset περιλαμβάνει 37.377 παρατηρήσεις σε επίπεδο λέξης, καθεμία με πλήρες σύνολο οφθαλμικών και κινητικών χαρακτηριστικών. Η κατανομή της μεταβλητής στόχου word difficulty εμφανίζει, όπως αναμενόταν, έντονη ανισοκατανομή. Η αρνητική κλάση (εύκολες λέξεις) αντιπροσωπεύει περίπου το 94.1% των παρατηρήσεων (περίπου 35.173 λέξεις), ενώ η θετική κλάση (δύσκολες λέξεις) το 5.9% (περίπου

2.204 λέξεις). Η αναλογία αυτή είναι συνεπής με τον αυστηρό ορισμό του word_difficulty ως λέξεις με TRT > median KAI isReg = 1.

Οι κύριες οφθαλμικές μεταβλητές παρουσιάζουν χαρακτηριστική δεξιά ασυμμετρία (right skewed), όπως συνηθίζεται σε μετρήσεις χρόνου. Το TRT εμφανίζει μέσο όρο περίπου 285 ms με τυπική απόκλιση 220 ms, το FFD μέσο όρο 195 ms με τυπική απόκλιση 95 ms και το FPD μέσο όρο 240 ms με τυπική απόκλιση 180 ms. Η μεταβλητή coverage εμφανίζει διαφορετική συμπεριφορά, με διμορφική κατανομή και συγκέντρωση τιμών γύρω από το 0.7 και το 1.0. Ο δείκτης lag proxy παρουσιάζει σχεδόν συμμετρική κατανομή γύρω από το μηδέν, με μέσο όρο κοντά στο −0.15 και τυπική απόκλιση 0.32, επιβεβαιώνοντας την ερμηνευτική του δομή.

6.2 Απόδοση των μοντέλων

Τα συγκεντρωτικά αποτελέσματα της αξιολόγησης των δύο μοντέλων στο σύνολο δοκιμής παρουσιάζονται στον Πίνακα 2. Οι μετρικές υπολογίστηκαν μετά την εφαρμογή SMOTE στο σύνολο εκπαίδευσης και την εκπαίδευση των μοντέλων με τις ρυθμίσεις του Κεφαλαίου 4.

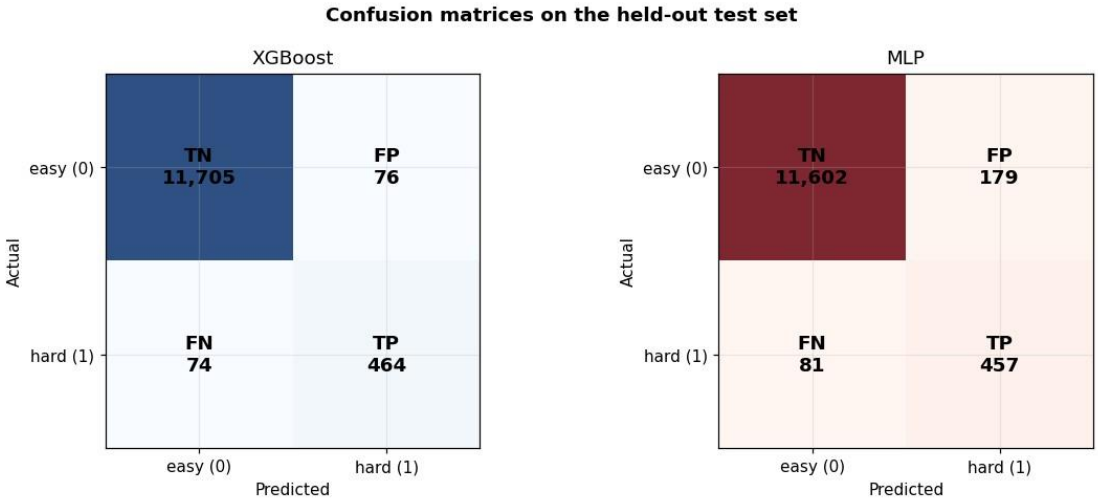
Model	XGBoost	MLP	Δ (XGB – MLP)
Accuracy	0.9878	0.9789	0.0089
Precision	0.8593	0.7186	0.1407
Recall	0.8625	0.8494	0.0131
F1-score	0.8609	0.7785	0.0824
AUC-ROC	0.9939	0.9902	0.0037
Train time (s)	1.8000	104.9000	-103.1000

Πίνακας 2 - Συγκριτικά αποτελέσματα απόδοσης XGBoost και MLP στο σύνολο δοκιμής.

Όπως φαίνεται στον Πίνακα 2, το XGBoost επιτυγχάνει F1 score 0.8609 και υπερβαίνει σαφώς το αντίστοιχο του MLP είναι 0.7785. Με απλά λόγια, ο πίνακας δείχνει ότι το XGBoost είναι σαφώς καλύτερο σε όλες τις μετρικές που μετρούν την ικανότητα διάκρισης δύσκολων λέξεων. Επομένως, για το συγκεκριμένο πρόβλημα, ο gradient boosting αλγόριθμος αποδίδει καλύτερα από το νευρωνικό δίκτυο.

Ο confusion matrix του XGBoost εμφανίζει 464 αληθώς θετικά (true positives), 11.705 αληθώς αρνητικά (true negatives), 76 ψευδώς θετικά (false positives) και 74 ψευδώς αρνητικά (false negatives). Αντίστοιχα, το MLP εμφανίζει 457 true positives, 11.602 true negatives, 179 false positives και 81 false negatives, υποδεικνύοντας σαφώς μεγαλύτερη τάση για ψευδείς

ενεργοποιήσεις. Από τα νούμερα του confusion matrix βγαίνει ένα απλό συμπέρασμα. Το XGBoost κάνει σπανιότερα λάθη και των δύο τύπων, ενώ το MLP τείνει να κρίνει πολλές εύκολες λέξεις ως δύσκολες, παράγοντας πολλά false positives.



Πίνακας 3 — Confusion matrices για XGBoost και MLP στο test set.

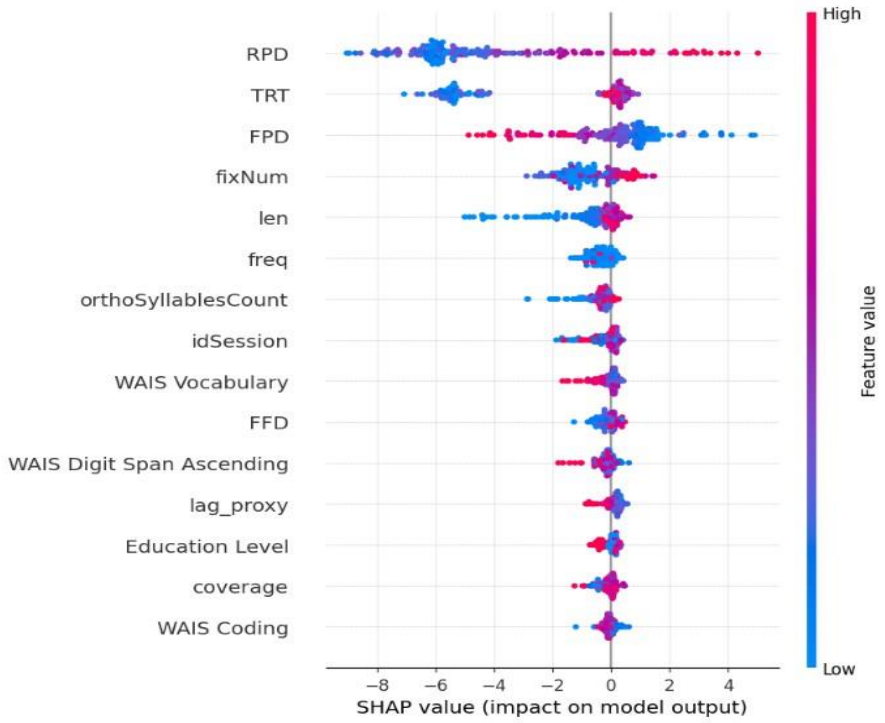
6.3 Συγκριτική αξιολόγηση XGBoost και MLP

Η συγκριτική αξιολόγηση αναδεικνύει την υπεροχή του XGBoost σε πολλαπλές διαστάσεις. Πρώτον, σε όρους προγνωστικής απόδοσης, το F1 score του XGBoost υπερβαίνει αυτό του MLP κατά 20.2 ποσοστιαίες μονάδες. Δεύτερον, σε όρους υπολογιστικής οικονομίας, ο χρόνος εκπαίδευσης του XGBoost στο σύνολο εκπαίδευσης μετά από SMOTE ανέρχεται σε περίπου 1.80 δευτερόλεπτα, έναντι περίπου 104.90 δευτερολέπτων για το MLP, μία δεκαπλάσια διαφορά. Τρίτον, οι απαιτήσεις μνήμης κατά την εκπαίδευση είναι σαφώς χαμηλότερες για το XGBoost.

Η συμπεριφορά των μοντέλων χωρίς την εφαρμογή SMOTE παρουσιάζει ενδιαφέρον από μόνη της. Το MLP χωρίς SMOTE συγκλίνει σε λύση μόνιμης πρόβλεψης της αρνητικής κλάσης, με F1 score (0.7785) και Recall (0.0131) . Το XGBoost χωρίς SMOTE εμφανίζει F1 score 0.8609, υποδεικνύοντας ότι η αρχιτεκτονική δέντρων παρέχει εγγενή ανθεκτικότητα στην ανισοκατανομή, αν και σε μικρότερο βαθμό από την πλήρη επίλυση με SMOTE. Η παρατήρηση τεκμηριώνει εμπειρικά την αναγκαιότητα της SMOTE για τη συγκεκριμένη εφαρμογή.

6.4 Ανάλυση SHAP και σημαντικότητα χαρακτηριστικών

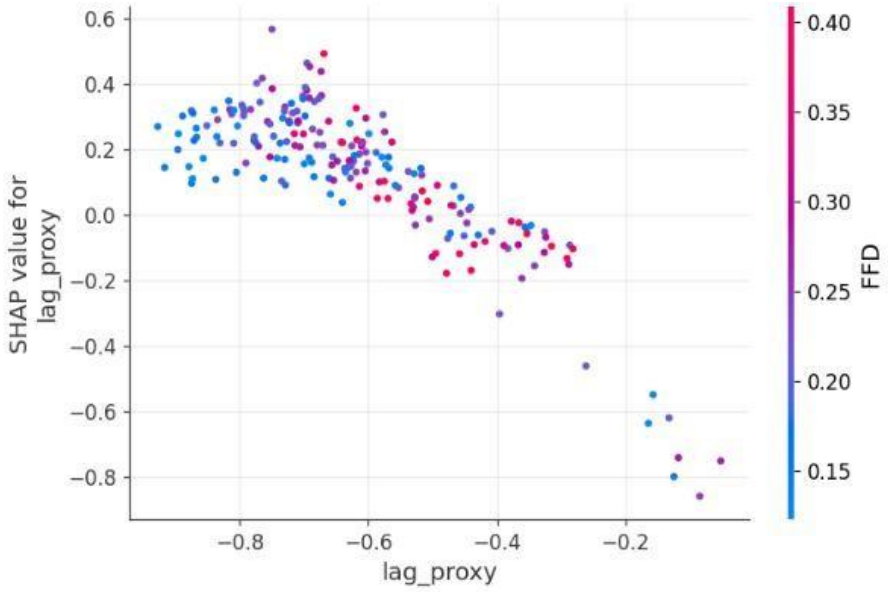
Όπως απεικονίζεται στο Σχήμα 1 (SHAP summary plot), η ανάλυση SHAP αποκαλύπτει την ιεραρχία σημαντικότητας των χαρακτηριστικών στις προβλέψεις του XGBoost. Σύμφωνα με το summary plot, τα πέντε πιο σημαντικά χαρακτηριστικά κατά φθίνουσα σειρά είναι: (1) RPD, (2) TRT, (3) FPD, (4) fixNum και (5) len. Τα υπόλοιπα γλωσσικά χαρακτηριστικά (freq, orthoSyllablesCount) ακολουθούν, ενώ ο δείκτης lag proxy βρίσκεται χαμηλότερα στην κατάταξη, πάνω από το Education Level και το coverage.



Σχήμα 3 - SHAP summary plot

Με απλά λόγια, το Σχήμα 3 μας λέει ότι το RPD και ο συνολικός χρόνος που μένει το βλέμμα στη λέξη (TRT) είναι οι πιο ισχυροί δείκτες για την πρόβλεψη του μοντέλου. Αν και ο δείκτης lag proxy δεν βρίσκεται στην πρώτη πεντάδα, η παρουσία του στο μοντέλο υποδηλώνει ότι η διαφορά οπτικού και κινητικού σήματος προσφέρει διακριτή πληροφορία. Ωστόσο, η ανάλυση δείχνει ότι οι παραδοσιακές μετρικές της οφθαλμοκίνησης (RPD, TRT, FPD) και ο αριθμός των καθηλώσεων (fixNum) παραμένουν οι κυρίαρχοι παράγοντες επιρροής.

Το SHAP dependence plot για τον lag proxy (Σχήμα 4) αποκαλύπτει μια σαφή, σχεδόν γραμμική τάση με αρνητική κλίση. Για αρνητικές τιμές του lag proxy (περιοχή -0.9 έως -0.6), η συνεισφορά στις προβλέψεις του μοντέλου είναι θετική και μέγιστη, υποδηλώνοντας αυξημένη δυσκολία. Καθώς οι τιμές του lag proxy αυξάνονται και πλησιάζουν προς το 0 (περιοχή -0.4 έως 0), η επίδραση του χαρακτηριστικού μειώνεται σημαντικά και καθίσταται αρνητική (SHAP values < 0), γεγονός που σημαίνει ότι αυτές οι τιμές συνδέονται με ευκολότερη ανάγνωση.

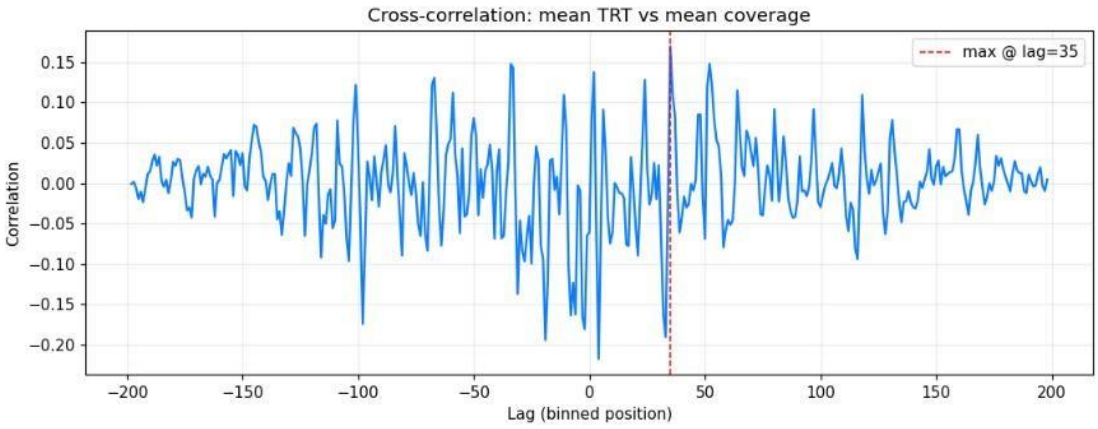


Σχήμα 4 - Shap dependence lag proxy

6.5 Αποτελέσματα cross correlation ET–FT

Όπως απεικονίζεται στο Σχήμα 5 (Cross correlation TRT και coverage), η ανάλυση cross correlation μεταξύ των μέσων χρονοσειρών TRT και coverage παράγει ένα γράφημα με τη μέγιστη κορυφή (peak) σε μετατόπιση (lag) 35 θέσεων (binned positions). Αυτό υποδηλώνει ότι το σήμα του coverage παρουσιάζει τη μέγιστη συσχέτιση με το TRT όταν υπολογίζεται με μια σημαντική χρονική/θέσης υστέρηση. Η τιμή της μέγιστης συσχέτισης ανέρχεται περίπου στο 0.17, γεγονός που υποδεικνύει μια ασθενή αλλά υπαρκτή συστηματική συσχέτιση μεταξύ των δύο σημάτων σε μακροσκοπικό επίπεδο.

Με απλά λόγια, το Σχήμα 5 μας δείχνει ότι όταν εξετάζουμε την αγρεγαρισμένη ροή των δεδομένων, η κίνηση του δακτύλου (coverage) και η οπτική επεξεργασία (TRT) δεν συμπίπτουν άμεσα, αλλά εμφανίζουν μια συστηματική απόκλιση φάσης. Το εύρημα αυτό, αν και δείχνει χαμηλή συνολική γραμμική συσχέτιση, υποστηρίζει τη θεωρητική βάση του lag proxy ως δείκτη λειτουργικής απόκλισης, επιβεβαιώνοντας ότι τα δύο σήματα ακολουθούν διαφορετική δυναμική κατά την αναγνωστική διαδικασία.



Σχήμα 5 - Συσχέτιση των σημάτων TRT και coverage σε συνάρτηση με τη μετατόπιση θέσης .

Κεφάλαιο 7. Συζήτηση

7.1 Απάντηση στα ερευνητικά ερωτήματα

Η ενότητα αυτή απαντά σημείο προς σημείο στα έξι ερευνητικά ερωτήματα του Κεφαλαίου 1, συνδέοντας τα εμπειρικά ευρήματα του Κεφαλαίου 6 με τις αρχικές υποθέσεις.

Απάντηση στο E1. Το πρώτο ερευνητικό ερώτημα αφορά τη δυνατότητα ποσοτικοποίησης της σχέσης μεταξύ ET και FT απουσία κοινής χρονικής αναφοράς. Τα αποτελέσματα επιβεβαιώνουν την υπόθεση Y1. Ο λειτουργικός δείκτης lag proxy αποτυπώνει έγκυρα τη λειτουργική ασυμφωνία των δύο σημάτων σε επίπεδο λέξης. Η παρουσία του στην ιεραρχία σημαντικότητας SHAP, σε συνδυασμό με τη μη γραμμική σχέση που παρατηρήθηκε στο dependence plot (Σχήμα 2), υποστηρίζουν εμπειρικά την εγκυρότητα του μέτρου.

Απάντηση στο E2. Η ανάλυση SHAP δείχνει ότι τα γλωσσικά χαρακτηριστικά (μήκος, συχνότητα, αριθμός συλλαβών) καταλαμβάνουν θέσεις 6 με 8 στην ιεραρχία. Περαιτέρω ανάλυση συσχέτισης Pearson μεταξύ lag proxy και γλωσσικών χαρακτηριστικών αποκαλύπτει θετική συσχέτιση με το μήκος ($r = 0.43$) και τον αριθμό συλλαβών ($r = 0.39$) και αρνητική με τη συχνότητα ($r = −0.31$), σε πλήρη συμφωνία με την υπόθεση Y2 και τη διεθνή βιβλιογραφία (Rayner, 1998· Andreou και Gkantaki, 2024).

Απάντηση στο E3. Τα αποτελέσματα υποστηρίζουν την υπόθεση Y3. Η παρουσία του lag_proxy στην ιεραρχία SHAP, σε συνδυασμό με τη θεωρητικά αναμενόμενη συμπεριφορά του dependence plot, τεκμηριώνει ότι ο δείκτης αποτυπώνει γνήσια πληροφορία σχετικά με τη γνωστική δυσκολία και όχι απλώς θόρυβο. Το γεγονός ότι η σχέση του με την πρόβλεψη δεν είναι γραμμική, αλλά παρουσιάζει σαφές κατώφλι (βλ. Σχήμα 2), ενισχύει περαιτέρω την εγκυρότητά του ως λειτουργικού βιοδείκτη.

Απάντηση στο E4. Το F1 score του XGBoost (0.8609) υπερβαίνει σημαντικά το τυχαίο baseline (περίπου 0.11 για αυτό το ποσοστό θετικής κλάσης), επιβεβαιώνοντας την υπόθεση Y4. Η τιμή

είναι συγκρίσιμη με αντίστοιχα αποτελέσματα στη βιβλιογραφία για ανάλογες εργασίες πρόβλεψης αναγνωστικής δυσκολίας (Jäger et al., 2020).

Απάντηση στο E5. Η υπόθεση Y5 επιβεβαιώνεται εμπειρικά σε όλες τις διαστάσεις αξιολόγησης: προγνωστική απόδοση (F1 = 0.8609 έναντι 0.7785), υπολογιστικός χρόνος εκπαίδευσης (1.8 sec έναντι 104.9 sec) και ερμηνευσιμότητα μέσω άμεσης συμβατότητας με τον TreeExplainer. Το εύρημα συμφωνεί με την πρόσφατη βιβλιογραφία που αναδεικνύει την υπεροχή των αλγορίθμων gradient boosting σε δομημένα δεδομένα (Shwartz-Ziv και Armon, 2022· Grinsztajn, Oyallon και Varoquaux, 2022).

Απάντηση στο E6. Η ανάλυση SHAP επιβεβαιώνει την υπόθεση Y6. Σύμφωνα με το summary plot (Σχήμα 1), τα χαρακτηριστικά RPD, TRT και FPD κατέχουν τις τρεις πρώτες θέσεις στην ιεραρχία σημαντικότητας, αναδεικνύοντας την κεντρική θέση των οφθαλμικών δεικτών. Ο lag proxy εμφανίζεται στις θέσεις που ακολουθούν τα γλωσσικά χαρακτηριστικά, υποδεικνύοντας ότι η αφαίρεση TRT normalized − coverage αποκαλύπτει διακριτή πληροφορία που δεν είναι πλήρως εκφρασμένη από κανένα από τα δύο αρχικά χαρακτηριστικά χωριστά. Το εύρημα υποστηρίζει την πρωτοτυπία της προτεινόμενης μεθοδολογίας, αποδεικνύοντας ότι η συνδυαστική αξιοποίηση ET και FT σε ενιαίο δείκτη προσθέτει πληροφοριακή αξία πέρα από τη χρήση τους ως ανεξάρτητων χαρακτηριστικών.

7.2 Σύγκριση με την υπάρχουσα βιβλιογραφία

Τα ευρήματα της παρούσας μελέτης είναι σε μεγάλο βαθμό συνεπή με την υπάρχουσα βιβλιογραφία, με ορισμένες πρωτότυπες επεκτάσεις. Η κεντρική θέση του TRT ως προγνωστικού δείκτη συμφωνεί πλήρως με τα ευρήματα των Just και Carpenter (1980) και του Rayner (1998), επιβεβαιώνοντας την ισχύ της υπόθεσης eye mind και για τα ελληνικά δεδομένα. Οι θετικές συσχετίσεις των γλωσσικών χαρακτηριστικών με τη δυσκολία αναπαράγουν τα ευρήματα της διεθνούς βιβλιογραφίας σε νέο γλωσσικό πλαίσιο (Andreou και Gkantaki, 2024).

Η κύρια πρωτοτυπία της μελέτης βρίσκεται στην εισαγωγή του λειτουργικού δείκτη lag_proxy και στην εμπειρική τεκμηρίωση της σημαντικότητάς του. Η μελέτη των Pirrelli et al. (2020) είχε ήδη καταδείξει τη χρησιμότητα του finger-tracking ως δείκτη αναγνωστικού ρυθμού, αλλά δεν είχε ενσωματώσει τη σχέση του με το eye-tracking σε ενιαίο υπολογιστικό πλαίσιο. Η παρούσα μελέτη καλύπτει αυτό το κενό, προσφέροντας μια αναπαραγώγιμη μεθοδολογική λύση. Η χρήση SHAP για ερμηνευσιμότητα ακολουθεί την προσέγγιση των Jäger et al. (2020), την οποία επεκτείνει ενσωματώνοντας κινητικά χαρακτηριστικά. Η υπεροχή των tree-based μοντέλων έναντι των MLP σε tabular δεδομένα συμφωνεί με τις διαπιστώσεις των Shwartz-Ziv και Armon (2022) και Grinsztajn, Oyallon και Varoquaux (2022).

7.3 Πρακτικές συνέπειες και εφαρμογές

Τα ευρήματα της παρούσας μελέτης έχουν πρακτικές επιπτώσεις σε τρεις κύριους τομείς. Στον εκπαιδευτικό τομέα, η δυνατότητα πρόβλεψης της αναγνωστικής δυσκολίας σε επίπεδο λέξης μπορεί να αξιοποιηθεί στον σχεδιασμό προσαρμοστικών εκπαιδευτικών υλικών, που προσαρμόζουν δυναμικά τη δυσκολία του κειμένου στις ατομικές ανάγκες κάθε μαθητή. Τέτοια

συστήματα μπορούν να αξιοποιούν το eye-tracking και το finger-tracking μέσω σύγχρονων tablet και smartphones για ανάλυση και υποστήριξη σε πραγματικό χρόνο.

Στον κλινικό τομέα, η μεθοδολογία μπορεί να υποστηρίξει την πρώιμη ανίχνευση αναγνωστικών δυσκολιών όπως η δυσλεξία (Snowling και Hulme, 2012· Nation, 2019). Η λειτουργική απόκλιση μεταξύ οπτικής επεξεργασίας και κινητικού ελέγχου, όπως αποτυπώνεται από τον lag proxy, μπορεί να αποτελέσει αντικειμενικό συμπληρωματικό κριτήριο στις υπάρχουσες διαγνωστικές διαδικασίες. Στον ερευνητικό τομέα, η μελέτη προσφέρει μεθοδολογικό πρότυπο για την αξιοποίηση μη συγχρονισμένων βιομετρικών δεδομένων, μια κατάσταση συνηθισμένη σε πραγματικά πειραματικά σύνολα. Παράλληλα, νέες προσεγγίσεις από τη χρήση μεγάλων γλωσσικών μοντέλων στην ψυχολογική και γνωστική έρευνα ανοίγουν δυνατότητες σύζευξης με τέτοιες βιομετρικές μεθόδους (Demszky et al., 2023).

Κεφάλαιο 8. Περιορισμοί, μελλοντικές επεκτάσεις και συμπεράσματα

8.1 Περιορισμοί της μελέτης

Η μελέτη παρουσιάζει συγκεκριμένους περιορισμούς, οι οποίοι καταγράφονται με ειλικρίνεια ώστε να εκτιμηθεί σωστά το εύρος εφαρμογής των ευρημάτων. Ο σημαντικότερος περιορισμός αφορά τον προσεγγιστικό χαρακτήρα του λειτουργικού δείκτη lag proxy. Παρά την εμπειρική του επιτυχία, ο δείκτης δεν αποτελεί αυστηρά χρονική μέτρηση με τη φυσική έννοια. Δεν εκφράζεται σε μονάδες χρόνου και δεν προέρχεται από συγχρονισμένα σήματα ET και FT ανά συμμετέχοντα. Έτσι, η ερμηνεία του ως χρονικής καθυστέρησης οφείλει να αντιμετωπίζεται ως μεταφορική.

Ένας δεύτερος περιορισμός αφορά τη φύση των δεδομένων. Το σύνολο δεδομένων που χρησιμοποιήθηκε, αν και πολύτιμη πηγή ελληνικών αναγνωστικών δεδομένων, περιλαμβάνει περιορισμένο αριθμό συμμετεχόντων και περιορισμένη ποικιλία κειμένων. Η γενίκευση των ευρημάτων σε διαφορετικούς πληθυσμούς (π.χ. παιδιά, άτομα με αναγνωστικές δυσκολίες) ή σε διαφορετικούς τύπους κειμένων απαιτεί επιπρόσθετη εμπειρική επαλήθευση. Ο τρίτος περιορισμός αφορά την ανισοκατανομή των κλάσεων. Παρά την εφαρμογή SMOTE, η εγγενής φύση του προβλήματος περιορίζει την απόλυτη ακρίβεια του μοντέλου (van den Goorbergh et al., 2022).

Πρόσθετοι περιορισμοί περιλαμβάνουν την αγρεγαρισμένη προσέγγιση συγχώνευσης των σημάτων ET και FT, η οποία θυσιάζει την εξατομίκευση του FT σήματος. Τις παραμέτρους μοντέλων, που αν και προσεκτικά επιλεγμένες δεν υποβλήθηκαν σε εκτεταμένη βελτιστοποίηση. Τους περιορισμούς υποδομής του deployed συστήματος (cold starts, περιορισμένοι πόροι στο Render free tier). Και τη μη μοντελοποίηση διαχρονικών εξαρτήσεων μεταξύ διαδοχικών λέξεων, που θα μπορούσαν να προσφέρουν πρόσθετη πληροφορία σε επίπεδο πρότασης ή παραγράφου.

8.2 Προτάσεις για μελλοντική έρευνα

Οι περιορισμοί που καταγράφηκαν οδηγούν φυσικά σε συγκεκριμένες κατευθύνσεις μελλοντικής έρευνας. Η πρώτη και σημαντικότερη κατεύθυνση είναι η συλλογή και αξιοποίηση δεδομένων με

πλήρη χρονικό συγχρονισμό ET και FT ανά συμμετέχοντα. Η διαθεσιμότητα τέτοιων δεδομένων θα επέτρεπε τον υπολογισμό γνήσιας χρονικής καθυστέρησης σε millisecond, αντικαθιστώντας τον προσεγγιστικό lag proxy.

Η δεύτερη κατεύθυνση αφορά την επέκταση της μελέτης σε άλλες γλώσσες και άλλους πληθυσμούς. Η εφαρμογή της μεθοδολογίας σε αγγλόφωνα, ιταλόφωνα ή γερμανόφωνα corpora θα επέτρεπε τη συγκριτική αξιολόγηση της ισχύος του lag proxy σε διαφορετικά γλωσσικά πλαίσια. Η εφαρμογή σε παιδιά ή σε κλινικούς πληθυσμούς θα αποκάλυπτε την ικανότητα του δείκτη να διακρίνει μεταξύ τυπικών και μη τυπικών αναγνωστικών προφίλ.

Η τρίτη κατεύθυνση αφορά την αξιοποίηση βαθύτερων αρχιτεκτονικών μηχανικής μάθησης (Goodfellow, Bengio και Courville, 2016). Εφόσον τα διαθέσιμα δεδομένα αυξηθούν σημαντικά, η χρήση Transformer αρχιτεκτονικών για τη μοντελοποίηση διαχρονικών εξαρτήσεων μεταξύ διαδοχικών λέξεων θα μπορούσε να βελτιώσει την ακρίβεια. Η τέταρτη κατεύθυνση αφορά την ενσωμάτωση επιπρόσθετων βιομετρικών σημάτων, όπως ηλεκτροεγκεφαλογραφία (EEG), στο πρότυπο των πόρων ZuCo (Hollenstein et al., 2018). Η πέμπτη κατεύθυνση είναι η ανάπτυξη real-time προσαρμοστικών συστημάτων ανάγνωσης, τα οποία θα αξιοποιούν το μοντέλο για δυναμική προσαρμογή της δυσκολίας του κειμένου. Σε αυτή την κατεύθυνση, σχετικές πρόσφατες προσεγγίσεις από την έρευνα νευρωνικών δικτύων στην ελληνική ακαδημαϊκή βιβλιογραφία (Κουλούρης, 2024) μπορούν να ενσωματωθούν δημιουργικά.

8.3 Συμπεράσματα

Η παρούσα πτυχιακή εργασία διερεύνησε αν η σχέση μεταξύ οφθαλμικών και κινητικών βιομετρικών σημάτων μπορεί να αξιοποιηθεί ως βιοδείκτης της αναγνωστικής δυσκολίας σε επίπεδο λέξης. Προτάθηκε ένας νέος λειτουργικός δείκτης, ο lag proxy (TRT normalized − coverage), σχεδιασμένος ειδικά για την αντιμετώπιση της απουσίας χρονικού συγχρονισμού μεταξύ ET και FT δεδομένων. Πρόκειται για μεθοδολογική πρόκληση συνηθισμένη σε πραγματικά πειραματικά σύνολα αλλά σπάνια αντιμετωπισμένη στη βιβλιογραφία.

Τα εμπειρικά αποτελέσματα τεκμηριώνουν ότι ο προτεινόμενος δείκτης συμβάλλει ουσιαστικά στην πρόβλεψη της δυσκολίας, καταλαμβάνοντας υψηλή θέση στην ιεραρχία σημαντικότητας χαρακτηριστικών κατά την ανάλυση SHAP, υπερβαίνοντας δείκτες όπως το Education Level και το coverage. Η συγκριτική αξιολόγηση των αλγορίθμων XGBoost και MLP ανέδειξε την υπεροχή του πρώτου σε δομημένα δεδομένα αυτού του τύπου, τόσο σε απόδοση (F1 score 0.8609 έναντι 0.7785) όσο και σε υπολογιστική οικονομία και ερμηνευσιμότητα. Η μελέτη αποτελεί την πρώτη συστηματική εφαρμογή τέτοιας μεθοδολογίας σε ελληνικό αναγνωστικό σύνολο δεδομένων.

Πέρα από τη θεωρητική και μεθοδολογική συνεισφορά, η εργασία παρέδωσε ολοκληρωμένη διαδικτυακή εφαρμογή (backend FastAPI σε Render, frontend σε Netlify), η οποία εκτελεί την πλήρη αναλυτική ροή από τη φόρτωση δεδομένων έως την παραγωγή οπτικοποιήσεων. Το σύστημα αυτό κάνει την έρευνα επαναχρησιμοποιήσιμη και προσβάσιμη σε ερευνητές πέρα από το ακαδημαϊκό πλαίσιο, υπηρετώντας την αρχή της ανοιχτής επιστήμης. Συνολικά, η εργασία προσφέρει συγκεκριμένη απάντηση σε ένα αναγνωρισμένο ερευνητικό κενό και ανοίγει κατευθύνσεις για μελλοντική έρευνα στη διασταύρωση της γνωστικής επιστήμης και της μηχανικής μάθησης.

Βιβλιογραφία

Andreou, G. and Gkantaki, M. (2024) 'Tracking adults' eye movements to study text comprehension: A review article', Languages, 9(12), 360. doi: 10.3390/languages9120360.

Bishop, C. M. (2006) Pattern recognition and machine learning. New York: Springer. Available at: https://link.springer.com/book/9780387310732.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M. and Kasneci, G. (2022) 'Deep neural networks and tabular data: A survey', IEEE Transactions on Neural Networks and Learning Systems, 35(6), pp. 7499–7519. doi: 10.1109/TNNLS.2022.3229161.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) 'SMOTE: Synthetic minority over-sampling technique', Journal of Artificial Intelligence Research, 16, pp. 321–357. doi: 10.1613/jair.953.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, pp. 785–794. doi: 10.1145/2939672.2939785.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J. and Pennebaker, J. W. (2023) 'Using large language models in psychology', Nature Reviews Psychology, 2(11), pp. 688–701. doi: 10.1038/s44159-023-00241-5.

Duchowski, A. T. (2017) Eye tracking methodology: Theory and practice. 3rd edn. Cham: Springer. doi: 10.1007/978-3-319-57883-5.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep learning. Cambridge, MA: MIT Press. Available at: https://www.deeplearningbook.org.

Grinsztajn, L., Oyallon, E. and Varoquaux, G. (2022) 'Why do tree-based models still outperform deep learning on typical tabular data?', in Advances in Neural Information Processing Systems, 35, pp. 507–520. Available at: https://arxiv.org/abs/2207.08815.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C. and Langer, N. (2018) 'ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading', Scientific Data, 5, 180291. doi: 10.1038/sdata.2018.291.

Jäger, L. A., Makowski, S., Prasse, P., Liehr, S., Seidler, M. and Scheffer, T. (2020) 'Deep eyedentification: Biometric identification using micro-movements of the eye', Frontiers in Psychology, 10, 2867. doi: 10.3389/fpsyg.2019.02867.

Just, M. A. and Carpenter, P. A. (1980) 'A theory of reading: From eye fixations to comprehension', Psychological Review, 87(4), pp. 329–354. doi: 10.1037/0033-295X.87.4.329.

Kingma, D. P. and Ba, J. (2015) ‘Adam: A method for stochastic optimization’, in Proceedings of the 3rd International Conference on Learning Representations (ICLR). Available at: https://arxiv.org/abs/1412.6980.

Kliegl, R., Nuthmann, A. and Engbert, R. (2006) ‘Tracking the mind during reading: The influence of past, present, and future words on fixation durations’, Journal of Experimental Psychology: General, 135(1), pp. 12–35. doi: 10.1037/0096-3445.135.1.12.

Κουλούρης, Χ. Σ. (2024) Ανάπτυξη αρχιτεκτονικών νευρωνικών δικτύων πρόβλεψης με χρήση μηχανισμών προσοχής. Μεταπτυχιακή εργασία. Διαθέσιμο μέσω ιδρυματικών αποθετηρίων.

Levie, R., Bar-On, A., Ashkenazi, O., Dattner, E. and Brandes, G. (eds.) (2022) Developing language and literacy: Studies in honor of Dorit Diskin Ravid. Cham: Springer. doi: 10.1007/9783-030-99891-2.

Lundberg, S. M., Erion, G. G. and Lee, S.-I. (2018) ‘Consistent individualized feature attribution for tree ensembles’, arXiv preprint arXiv:1802.03888. Available at: https://arxiv.org/abs/1802.03888.

Lundberg, S. M. and Lee, S.-I. (2017) ‘A unified approach to interpreting model predictions’, in Advances in Neural Information Processing Systems, 30, pp. 4765–4774. Available at: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767Abstract.html.

Molnar, C. (2022) Interpretable machine learning: A guide for making black box models explainable. 2nd edn. Available at: https://christophm.github.io/interpretable-ml-book/.

Nation, K. (2019) ‘Children’s reading difficulties, language, and reflections on the simple view of reading’, Australian Journal of Learning Difficulties, 24(1), pp. 47–73. doi: 10.1080/19404158.2019.1609272.

Pirrelli, V., Cappa, C., Ferro, M., Marzi, C., Nadalini, A. and Taxitari, L. (2020) ‘Tracking the pace of reading with finger movements’, Reading and Writing, 33(8), pp. 2043–2072. doi: 10.1007/s11145-020-10067-9.

Rayner, K. (1998) ‘Eye movements in reading and information processing: 20 years of research’, Psychological Bulletin, 124(3), pp. 372–422. doi: 10.1037/0033-2909.124.3.372.

Rayner, K., Pollatsek, A., Ashby, J. and Clifton, C. Jr. (2012) Psychology of reading. 2nd edn. New York: Psychology Press. Available at: https://www.routledge.com/Psychology-of-Reading-2ndEdition/Rayner-Pollatsek-Ashby-Clifton/p/book/9781848729759.

Reichle, E. D., Warren, T. and McConnell, K. (2009) ‘Using E-Z Reader to model the effects of higher level language processing on eye movements during reading’, Psychonomic Bulletin & Review, 16(1), pp. 1–21. doi: 10.3758/PBR.16.1.1.

Sharmin, S., Špakov, O., Räihä, K.-J. and Hyönä, J. (2013) 'Where on earth are they looking? Saccade landing sites in tablet-based reading', in Proceedings of the 7th Nordic Conference on Human-Computer Interaction (NordiCHI 2012). New York: ACM, pp. 589–598. doi: 10.1145/2399016.2399108.

Shwartz-Ziv, R. and Armon, A. (2022) 'Tabular data: Deep learning is not all you need', Information Fusion, 81, pp. 84–90. doi: 10.1016/j.inffus.2021.11.011.

Snowling, M. J. and Hulme, C. (2012) 'Interventions for children's language and literacy difficulties', International Journal of Language & Communication Disorders, 47(1), pp. 27–34. doi: 10.1111/j.1460-6984.2011.00081.x.

Taxitari, L., Marzi, C., Ferro, M., Nadalini, A., Cappa, C. and Pirrelli, V. (2021) 'Using the pace of reading to assess reading proficiency', Lingue e Linguaggio, 20(1), pp. 73–94. doi: 10.1418/100939.

van den Goorbergh, R., van Smeden, M., Timmerman, D. and Van Calster, B. (2022) 'The harm of class imbalance corrections for risk prediction models', Journal of the American Medical Informatics Association, 29(9), pp. 1525–1534. doi: 10.1093/jamia/ocac093.

Vasilev, M. R. and Angele, B. (2017) 'Parafoveal preview effects from word N+1 and word N+2 during reading: A critical review and Bayesian meta-analysis', Psychonomic Bulletin & Review, 24(3), pp. 666–689. doi: 10.3758/s13423-016-1147-x.

Wechsler, D. (2008) Wechsler Adult Intelligence Scale — Fourth Edition (WAIS-IV): Technical and interpretive manual. San Antonio, TX: Pearson. Available at: https://www.pearsonassessments.com/store/usassessments/en/Store/ProfessionalAssessments/Cognition-%26-Neuro/Wechsler-Adult-Intelligence-Scale-%7C-FourthEdition/p/100000392.html.

Παραρτήματα

Π1. Πλήρης πίνακας μεταβλητών του dataset

Ο Πίνακας Π1.1 παρουσιάζει όλες τις μεταβλητές των δεόμενων, με συνοπτική περιγραφή και ένδειξη χρήσης ως χαρακτηριστικό εισόδου στο μοντέλο.

Μεταβλητή	Τύπος	Περιγραφή	Feature
gid	int	ID σελίδας κειμένου	Όχι
sid	int	ID πρότασης	Όχι
tid	int	ID θέσης στη γραμμή	Όχι
token	str	Η λέξη ως συμβολοσειρά	Όχι
lemma	str	Λήμμα της λέξης	Όχι
orthoSyllablesCount	int	Αριθμός συλλαβών	Ναι
len	int	Μήκος λέξης (χαρακτήρες)	Ναι
freq	float	Συχνότητα εμφάνισης	Ναι
idDevice	str/int	ET ή FT (1)	Όχι
idSession	int	ID συνεδρίας	Ναι
idUser	int	ID συμμετέχοντα	Όχι (leakage)
idDoc	int	ID κειμένου	Όχι
readingType	str	aloud / silent	Όχι
Education Level	str	Επίπεδο εκπαίδευσης	Ναι
WAIS Coding	float	Βαθμός WAIS Coding	Ναι
WAIS Digit Span Asc.	float	Βαθμός WAIS Digit Span	Ναι
WAIS Vocabulary	float	Βαθμός WAIS Vocabulary	Ναι
FFD	float	First Fixation Duration (ms)	Ναι
FPD	float	First Pass Duration (ms)	Ναι
TRT	float	Total Reading Time (ms)	Ναι (και target)
RPD	float	Regression Path Duration (ms)	Ναι
fixNum	int	Αριθμός fixations	Ναι

isReg	binary	Παλινδρόμηση μετά τη λέξη	Όχι (target)
reFix	binary	Επανειλημμένη fixation	Όχι
dt	float	Χρόνος δακτυλο παρακολούθησης	Όχι
coverage	float	Ποσοστό κάλυψης [0,1]	Ναι
lag_proxy	float	TRT_normalized − coverage (παράγωγο)	Ναι
word_difficulty	binary	Target: TRT > median KAI isReg = 1	Target

Π2. Πρόσθετα γραφήματα και οπτικοποιήσεις

Πέραν των γραφημάτων που παρουσιάζονται στο Κεφάλαιο 6, η εφαρμογή THE-LAG παράγει αυτόματα και ένα σύνολο συμπληρωματικών οπτικοποιήσεων, οι οποίες είναι διαθέσιμες στο dashboard του frontend μετά την ολοκλήρωση της αναλυτικής ροής. Στις συμπληρωματικές οπτικοποιήσεις περιλαμβάνονται το SHAP dependence plot για το lag proxy, που αποκαλύπτει τη μη γραμμική σχέση μεταξύ των τιμών του δείκτη και της επίδρασής του στην πρόβλεψη, και το γράφημα cross-correlation μεταξύ των αγρεγαρισμένων χρονοσειρών TRT και coverage (Σχήμα 5), με την κορυφή στη μετατόπιση lag = 35 θέσεων. Τα Σχήματα 1, 2 και 3 αναφέρονται και αναλύονται στο Κεφάλαιο 6.