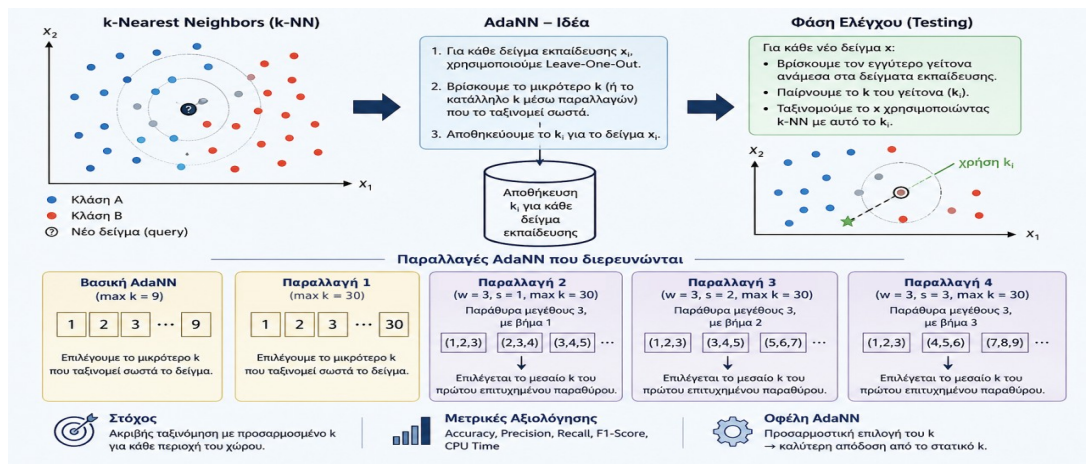


ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΉ ΕΡΓΑΣΙΑ

«Αυτόματος προσδιορισμός της παραμέτρου k στον
κατηγοριοποιητή k εγγύτερων γειτόνων μέσω της τεχνικής
AdaNN και διερεύνηση παραλλαγών της»



Φοιτητής: Μπαντάνης Θεοφάνης
Αρ. Μητρώου:164711.....

Επιβλέπων: Ουγιάρογλου Στέφανος
Βαθμίδα: Καθηγητής

Ημερομηνία: 29/08/2026

Τίτλος Δ.Ε.: Αυτόματος προσδιορισμός της παραμέτρου k στον κατηγοριοποιητή k εγγύτερων γειτόνων μέσω της τεχνικής AdaNN και διερεύνηση παραλλαγών της

Κωδικός Δ.Ε.: 26206

Ονοματεπώνυμο φοιτητή: Μπαντάνης Θεοφάνης

Ονοματεπώνυμο εισηγητή: Ουγιάρογλου Στέφανος

Ημερομηνία ανάληψης Δ.Ε.: 23-03-2026

Ημερομηνία περάτωσης Δ.Ε.: ...29-08-2026.....

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Μπαντάνη Θεοφάνη που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητα και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Στους γονείς μου.

Θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη στους γονείς μου, για την αμέριστη αγάπη, υποστήριξη και ενθάρρυνση που μου προσέφεραν καθ' όλη τη διάρκεια των φοιτητικών μου χρόνων. Η παρουσία τους, η υπομονή τους και η πίστη τους στις δυνατότητές μου αποτέλεσαν σημαντική πηγή δύναμης κάθε δυσκολία και σε κάθε πρόκληση που αντιμετώπισα.

Τους ευχαριστώ για όλες τις θυσίες που έκαναν αλλά και για το γεγονός ότι ήταν πάντα δίπλα μου, δίνοντάς μου τη δύναμη να συνεχίζω και να μην εγκαταλείπω τους στόχους μου. Η ολοκλήρωση των σπουδών μου δεν αποτελεί μόνο προσωπικό επίτευγμα, αλλά και αποτέλεσμα της διαρκούς στήριξης και συμπαράστασής τους.

Η πτυχιακή αυτή εργασία αποτελεί, επομένως, και έναν μικρό φόρο ευγνωμοσύνης προς εκείνους που στάθηκαν δίπλα μου σε όλη αυτή την πορεία. Τους ευχαριστώ από καρδιάς για όλα.

Πρόλογος

Η επιλογή του θέματος της παρούσας διπλωματικής εργασίας προέκυψε από ένα ερώτημα που φαίνεται απλό αλλά αποδεικνύεται επίμονο: πόσους γείτονες πρέπει να κοιτάξει ένας κατηγοριοποιητής προτού αποφασίσει. Ο αλγόριθμος των k εγγύτερων γειτόνων διδάσκεται συνήθως ως το πιο απλό παράδειγμα επιβλεπόμενης μάθησης, όμως η επίδοσή του κρέμεται εξ ολοκλήρου από μία παράμετρο που ο ίδιος ο αλγόριθμος δεν γνωρίζει πώς να επιλέξει. Η ιδέα ότι η παράμετρος αυτή θα μπορούσε να μην είναι ενιαία για ολόκληρο τον χώρο, αλλά να προσαρμόζεται τοπικά σε κάθε περιοχή του, μου φάνηκε εξαιρετικής τόσο κομψή όσο και ελκυστική για πειραματισμό.

Αυτό που με απασχόλησε περισσότερο κατά την εκπόνηση δεν ήταν η υλοποίηση καθαυτή, αλλά η απόσταση ανάμεσα στη διαίσθηση και τη μέτρηση. Παραλλαγές που στο χαρτί έμοιαζαν προφανώς καλύτερες αποδείχθηκαν ουδέτερες ή και ζημιογόνες σε καθαρά δεδομένα, και απέκτησαν νόημα μόνο όταν στα δεδομένα εισήχθη θόρυβος. Η εμπειρία αυτή, ότι δηλαδή μια ιδέα κρίνεται από το πλαίσιο μέσα στο οποίο δοκιμάζεται και όχι από την ευλογοφάνειά της, είναι το σημαντικότερο που αποκόμισα από την εργασία αυτή.

Περίληψη

Ο κατηγοριοποιητής των k εγγύτερων γειτόνων (k -NN) είναι από τους πιο διαδεδομένους αλγορίθμους επιβλεπόμενης μάθησης, ωστόσο η επίδοσή του εξαρτάται κρίσιμα από την τιμή της παραμέτρου k , η οποία παραδοσιακά καθορίζεται ενιαία για ολόκληρο τον χώρο των δεδομένων. Η τεχνική AdaNN αντιμετωπίζει τον περιορισμό αυτόν αναθέτοντας διαφορετική τιμή k σε κάθε σημείο, με βάση τον εγγύτερο γείτονα του στο σύνολο εκπαίδευσης.

Η παρούσα εργασία υλοποιεί τον AdaNN και διερευνά τέσσερις παραλλαγές του. Η πρώτη επεκτείνει το εύρος αναζήτησης του k από $[1..9]$ σε $[1..30]$. Οι άλλες τρεις εισάγουν έναν κανόνα κυλιόμενου παραθύρου: αντί να γίνεται δεκτή η πρώτη τιμή k που κατηγοριοποιεί σωστά ένα σημείο, απαιτείται να την κατηγοριοποιούν σωστά τρεις διαδοχικές τιμές, και ανατίθεται η μεσαία τιμή του παραθύρου. Οι τρεις παραλλαγές διαφέρουν ως προς το βήμα μετατόπισης του παραθύρου (1, 2 και 3). Ως σημείο αναφοράς χρησιμοποιείται και ο προσδιορισμός μίας καθολικής τιμής k μέσω πενταπλής διασταυρούμενης επικύρωσης. Προηγείται συστηματική επισκόπηση των τεχνικών καθολικού και δυναμικού προσδιορισμού της τιμής k που έχουν προταθεί στη βιβλιογραφία, οργανωμένων σε έξι μεθοδολογικές οικογένειες.

Η αξιολόγηση έγινε σε δεκατέσσερα καθιερωμένα σύνολα δεδομένων του αποθετηρίου KEEL, σε δύο εκδοχές το καθένα: στην αρχική τους μορφή και με εισαγωγή θορύβου περίπου 30 % στις ετικέτες του συνόλου εκπαίδευσης. Μετρήθηκαν ορθότητα, ακρίβεια, ανάκληση, F-score και χρόνος CPU. Τα αποτελέσματα δείχνουν ότι σε καθαρά δεδομένα καμία παραλλαγή δεν βελτιώνει τον βασικό AdaNN, ενώ υπό θόρυβο η εικόνα αντιστρέφεται: οι παραλλαγές παραθύρου υπερτερούν του βασικού αλγορίθμου σε εννέα από τα δεκατέσσερα σύνολα, με τη μέση ορθότητα να ανεβαίνει από 80,81 % σε 82,49 %. Ο καθολικός προσδιορισμός του k μέσω διασταυρούμενης επικύρωσης παραμένει ωστόσο ο ισχυρότερος αντίπαλος, κατατασσόμενος πρώτος και στις δεκατέσσερις περιπτώσεις με θόρυβο.

Automatic determination of the parameter k in the k -nearest neighbour classifier via the AdaNN technique and investigation of its variations

Theofanis Bantanis

Abstract

The k -nearest neighbour (k -NN) classifier is among the most widely used supervised learning algorithms, yet its performance depends critically on the value of the parameter k , which is traditionally fixed uniformly across the whole data space. The AdaNN technique addresses this limitation by assigning a different value of k to each query point, based on the optimal k of its nearest neighbour in the training set.

This thesis implements AdaNN and investigates four variations of it. The first extends the search range of k from $[1..9]$ to $[1..30]$. The remaining three introduce a sliding-window rule: instead of accepting the first value of k that classifies a training point correctly, three consecutive values are required to classify it correctly, and the middle value of the window is assigned. The three variations differ in the shift of the window (1, 2 and 3). Determining a single global k through five-fold cross-validation is used as a reference point. A systematic review of the global and dynamic k -determination techniques proposed in the literature, organised into six methodological families, precedes the experimental study.

The evaluation was carried out on fourteen established datasets from the KEEL repository, each in two versions: in its original form and with approximately 30 % label noise injected into the training set. Accuracy, precision, recall, F-score and CPU time were measured. The results show that on clean data no variation improves on baseline AdaNN, whereas under noise the picture is reversed: the window variations outperform the baseline on nine out of fourteen datasets, raising mean accuracy from 80.81 % to 82.49 %. Global determination of k through cross-validation nevertheless remains the strongest competitor, ranking first on all fourteen noisy datasets.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ. Στέφανο Ουγιάρογλου, τόσο για την ανάθεση ενός θέματος που συνδυάζει θεωρητικό ενδιαφέρον με ουσιαστική πειραματική δουλειά, όσο και για την καθοδήγηση σε όλα τα στάδια της εκπόνησης. Οι παρατηρήσεις του για τη διατύπωση των παραλλαγών και για τον σχεδιασμό του πειραματικού πρωτοκόλλου διαμόρφωσαν καθοριστικά την τελική μορφή της εργασίας.

Ευχαριστώ επίσης την οικογένειά μου για την αμέριστη υποστήριξη καθ' όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Πρόλογος	4
Περίληψη	5
Abstract	6
Ευχαριστίες	7
Περιεχόμενα	8
Κατάλογος Σχημάτων	11
Κατάλογος Πινάκων	12
Συντομογραφίες	13
Κεφάλαιο 1ο: Εισαγωγή	14
1.1 Κατηγοριοποίηση	14
1.2 Κατηγοριοποίηση k-εγγύτερων γειτόνων	14
1.3 Η παράμετρος k	16
1.4 Κίνητρο	17
1.5 Συνεισφορά	18
1.6 Οργάνωση της εργασίας	19
Κεφάλαιο 2ο: Απόδοση σταθερής τιμής στην παράμετρο k	20
2.1 Θεωρητική θεμελίωση του κανόνα	20
2.2 Ρύθμιση του k με διασταυρούμενη επικύρωση	20
2.3 Αναλυτικοί και εμπειρικοί κανόνες επιλογής	21
2.4 Άλλες τεχνικές καθολικού προσδιορισμού	21
2.5 Οι περιορισμοί μιας σταθερής τιμής	22
Κεφάλαιο 3ο: Απόδοση δυναμικής τιμής στην παράμετρο k	24
3.1 Επίπεδα προσδιορισμού της τιμής k	24
3.2 Προσεγγίσεις γενετικών αλγορίθμων	24
3.3 Προσεγγίσεις νευρωνικών δικτύων	25
3.4 Προσεγγίσεις προτύπων και συσταδοποίησης	25
3.5 Ευρετικές προσεγγίσεις	26
3.6 Πιθανοτικές προσεγγίσεις	27
3.7 Λοιπές προσεγγίσεις	28
3.8 Συγκριτική επισκόπηση	29
3.9 Τεχνικές μείωσης δεδομένων και θόρυβος στις ετικέτες	30
3.10 Το ερευνητικό κενό	30
Κεφάλαιο 4ο: Ο αλγόριθμος AdaNN	32
4.1 Η κεντρική ιδέα και η υπόθεση εργασίας	32

4.2	Η φάση εκπαίδευσης	33
4.3	Η φάση ελέγχου	34
4.4	Αναλυτικό παράδειγμα	34
4.5	Ιδιότητες και οριακές περιπτώσεις	35
4.6	Υπολογιστική πολυπλοκότητα	36
4.7	Αδυναμίες που κινητοποιούν τις παραλλαγές	36
Κεφάλαιο 5ο: Νέες παραλλαγές του AdaNN		37
5.1	Παραλλαγή 1: επέκταση του εύρους αναζήτησης	37
5.2	Ο κανόνας κυλιόμενου παραθύρου	37
5.3	Παραλλαγές 2, 3 και 4: το βήμα μετατόπισης	38
5.4	Παράδειγμα εφαρμογής των παραλλαγών	38
5.5	Δομικές ιδιότητες των παραλλαγών παραθύρου	39
5.6	Ο αλγόριθμος καθολικού προσδιορισμού ως σημείο αναφοράς	39
5.7	Συγκεντρωτική παρουσίαση των αλγορίθμων	40
Κεφάλαιο 6ο: Υλοποιήσεις σε Python		41
6.1	Βιβλιοθήκες και οργάνωση του κώδικα	41
6.2	Φάση εκπαίδευσης: η συνάρτηση <code>loo_correct_matrix</code>	41
6.3	Ο κανόνας της πρώτης επιτυχίας: η συνάρτηση <code>k_plain</code>	43
6.4	Ο κανόνας κυλιόμενου παραθύρου: η συνάρτηση <code>k_window</code>	43
6.5	Φάση ελέγχου: η συνάρτηση <code>adann_predict</code>	44
6.6	Καθολικός προσδιορισμός: η συνάρτηση <code>find_best_global_k</code>	45
6.7	Πρωτόκολλο μέτρησης χρόνου: η συνάρτηση <code>time_uniformly</code>	46
6.8	Επαλήθευση της ορθότητας	47
Κεφάλαιο 7ο: Πειραματική μελέτη		49
7.1	Σύνολα δεδομένων	49
7.2	Πειραματικό περιβάλλον	51
7.2.1	Εισαγωγή θορύβου στις ετικέτες	51
7.2.2	Μετρικές αξιολόγησης	53
7.2.3	Υλικό και λογισμικό	53
7.2.4	Πρωτόκολλο μέτρησης χρόνου	54
7.3	Πειραματικά αποτελέσματα	54
7.3.1	Καθαρά δεδομένα: ορθότητα	54
7.3.2	Καθαρά δεδομένα: F-score	55
7.3.3	Θορυβώδη δεδομένα: ορθότητα	56
7.3.4	Θορυβώδη δεδομένα: F-score	58
7.3.5	Η μεταβολή που επιφέρει ο θόρυβος	59
7.3.6	Κατανομή των τιμών k	59
7.3.7	Η καθολική τιμή k	62

7.3.8	Υπολογιστικό κόστος	62
7.3.9	Συνολική κατάταξη	63
7.3.10	Ερμηνεία των ευρημάτων	64
7.3.11	Σύγκριση με τη βιβλιογραφία	66
7.3.12	Περιορισμοί της μελέτης	66
Κεφάλαιο 8ο:	Συμπεράσματα και μελλοντικές επεκτάσεις	68
8.1	Συμπεράσματα	68
8.2	Μελλοντικές επεκτάσεις	69
ΒΙΒΛΙΟΓΡΑΦΙΑ		71
ΠΑΡΑΡΤΗΜΑ Α: ΠΗΓΑΙΟΣ ΚΩΔΙΚΑΣ		75
A.1	Φάση εκπαίδευσης: leave-one-out για όλες τις τιμές k	75
A.2	Οι κανόνες επιλογής του k	75
A.3	Φάση ελέγχου	76
A.4	Καθολικός προσδιορισμός του k	76
A.5	Πρωτόκολλο μέτρησης χρόνου	77
ΠΑΡΑΡΤΗΜΑ Β: ΑΝΑΛΥΤΙΚΟΙ ΠΙΝΑΚΕΣ ΜΕΤΡΗΣΕΩΝ		78
B.1	Ακρίβεια και ανάκληση	78
B.2	Χρόνοι CPU	79
B.3	Κατανομή των τιμών k	80

Κατάλογος Σχημάτων

Σχήμα 1.1: Ταξινόμηση ενός άγνωστου σημείου με τον κανόνα k-NN για $k = 1, 3$ και 7 . Η απόφαση μεταβάλλεται ανάλογα με το πλήθος των γειτόνων που λαμβάνονται υπόψη.....	16
Σχήμα 1.2: Όριο απόφασης του k-NN στο σύνολο δεδομένων Banana για $k = 1, 5$ και 25 . Η αύξηση του k εξομαλύνει το όριο και εξαλείφει τους απομονωμένους θύλακες.....	17
Σχήμα 4.1: Οι δύο φάσεις του AdaNN. Η φάση ελέγχου είναι κοινή σε όλες τις παραλλαγές· διαφοροποιείται μόνο ο κανόνας επιλογής στη φάση εκπαίδευσης.....	31
Σχήμα 5.1: Ο κανόνας κυλιόμενου παραθύρου για βήματα $1, 2$ και 3 . Εμφανίζονται τα τέσσερα πρώτα παράθυρα κάθε παραλλαγής και η τιμή k που ανατίθεται σε καθένα.....	36
Σχήμα 6.1: Η συνάρτηση <code>loo_correct_matrix</code> , που υλοποιεί τη φάση εκπαίδευσης με <code>leave-one-out</code> για όλες τις τιμές k	40
Σχήμα 6.2: Η συνάρτηση <code>k_plain</code> , που υλοποιεί τον κανόνα της πρώτης επιτυχίας για τον βασικό AdaNN και την Παραλλαγή 1.....	41
Σχήμα 6.3: Η συνάρτηση <code>k_window</code> , που υλοποιεί τον κανόνα κυλιόμενου παραθύρου των Παραλλαγών $2, 3$ και 4	42
Σχήμα 6.4: Η συνάρτηση <code>adann_predict</code> , που υλοποιεί τη φάση ελέγχου, κοινή σε όλες τις παραλλαγές.	43
Σχήμα 6.5: Η συνάρτηση <code>find_best_global_k</code> , που υλοποιεί τη ρύθμιση του k με πενταπλή στρωματοποιημένη διασταυρούμενη επικύρωση.....	44
Σχήμα 6.6: Η συνάρτηση <code>time_uniformly</code> , που υλοποιεί το πρωτόκολλο μέτρησης χρόνου με κοινό αριθμό επαναλήψεων.....	45
Σχήμα 7.1: Μέγεθος συνόλων εκπαίδευσης και ελέγχου σε λογαριθμική κλίμακα (αριστερά) και πλήθος χαρακτηριστικών και κλάσεων (δεξιά) για κάθε σύνολο δεδομένων.....	48
Σχήμα 7.2: Ποσοστό αλλοιωμένων ετικετών στο σύνολο εκπαίδευσης κάθε συνόλου δεδομένων. Ο θόρυβος δεν αγγίζει καθόλου τα σύνολα ελέγχου.....	49
Σχήμα 7.3: Όριο απόφασης στο Banana με 30% θόρυβο στις ετικέτες εκπαίδευσης. Για $k = 1$ το όριο έχει κατακερματιστεί πλήρως, ενώ για $k = 25$ ανακτάται σε μεγάλο βαθμό.....	50
Σχήμα 7.4: Ορθότητα ανά σύνολο δεδομένων και αλγόριθμο στα καθαρά δεδομένα.....	52
Σχήμα 7.5: F-score ανά σύνολο δεδομένων και αλγόριθμο στα καθαρά δεδομένα.....	53
Σχήμα 7.6: Ορθότητα ανά σύνολο δεδομένων και αλγόριθμο στα θορυβώδη δεδομένα.....	54
Σχήμα 7.7: F-score ανά σύνολο δεδομένων και αλγόριθμο στα θορυβώδη δεδομένα.....	56
Σχήμα 7.8: Μεταβολή της ορθότητας από τα καθαρά στα θορυβώδη δεδομένα. Οι αρνητικές τιμές δηλώνουν πτώση.....	56
Σχήμα 7.9: Μέση τιμή k ανά σημείο εκπαίδευσης, για καθαρά (αριστερά) και θορυβώδη (δεξιά) δεδομένα.....	58
Σχήμα 7.10: Κατανομή των τιμών k που ανατέθηκαν στα σημεία εκπαίδευσης, για τα σύνολα Pen Digits και Landsat Satellite, σε καθαρά και θορυβώδη δεδομένα.....	59
Σχήμα 7.11: Ποσοστό σημείων εκπαίδευσης που κατέληξαν στην εφεδρική τιμή $k = \max k$. Το πλήρες χρώμα αντιστοιχεί σε καθαρά και το ημιδιαφανές σε θορυβώδη δεδομένα.....	59
Σχήμα 7.12: Χρόνος CPU ανά σύνολο δεδομένων και αλγόριθμο σε λογαριθμική κλίμακα (καθαρά δεδομένα).....	61
Σχήμα 7.13: Μέση κατάταξη των έξι αλγορίθμων ως προς την ορθότητα, στις δύο συνθήκες.....	62

Κατάλογος Πινάκων

Πίνακας 2.1: Τεχνικές καθολικού προσδιορισμού της τιμής k	21
Πίνακας 3.1: Τα τέσσερα επίπεδα προσδιορισμού της τιμής k	23
Πίνακας 3.2: Συγκεντρωτική παρουσίαση των μεθόδων δυναμικού προσδιορισμού της τιμής k	28
Πίνακας 4.1: Πίνακας ορθότητας πέντε σημείων εκπαίδευσης για $k = 1..12$	33
Πίνακας 4.2: Τιμές k που αναθέτει ο κανόνας της πρώτης επιτυχίας στα σημεία του Πίνακα 4.1.....	33
Πίνακας 5.1: Τιμές k που αναθέτουν οι έξι κανόνες επιλογής στα σημεία του Πίνακα 4.1.....	37
Πίνακας 5.2: Οι έξι αλγόριθμοι που αξιολογούνται.....	38
Πίνακας 6.1: Οι συναρτήσεις της υλοποίησης και ο ρόλος τους.....	39
Πίνακας 7.1: Τα δεκατέσσερα σύνολα δεδομένων της πειραματικής αξιολόγησης.....	47
Πίνακας 7.2: Σύντομη περιγραφή των συνόλων δεδομένων του αποθετηρίου KEEL.....	47
Πίνακας 7.3: Το περιβάλλον εκτέλεσης των πειραμάτων.....	51
Πίνακας 7.4: Ορθότητα (%) στα καθαρά δεδομένα.....	52
Πίνακας 7.5: Νίκες, ισοπαλίες και ήττες κάθε παραλλαγής έναντι του βασικού AdaNN ($\max k = 9$) ως προς την ορθότητα, στα καθαρά δεδομένα.....	52
Πίνακας 7.6: F-score (%) στα καθαρά δεδομένα.....	53
Πίνακας 7.7: Ορθότητα (%) στα θορυβώδη δεδομένα.....	54
Πίνακας 7.8: Νίκες, ισοπαλίες και ήττες κάθε παραλλαγής έναντι του βασικού AdaNN ($\max k = 9$) ως προς την ορθότητα, στα θορυβώδη δεδομένα.....	55
Πίνακας 7.9: F-score (%) στα θορυβώδη δεδομένα.....	55
Πίνακας 7.10: Μέση τιμή k ανά σημείο εκπαίδευσης, καθαρά δεδομένα.....	57
Πίνακας 7.11: Μέση τιμή k ανά σημείο εκπαίδευσης, θορυβώδη δεδομένα.....	57
Πίνακας 7.12: Η καθολική τιμή k που επιλέχθηκε με πενταπλή διασταυρούμενη επικύρωση, ανά σύνολο δεδομένων.....	60
Πίνακας 7.13: Μέση κατάταξη των αλγορίθμων ως προς την ορθότητα ($1 = \text{καλύτερος}$).....	61
Πίνακας 7.14: Συγκεντρωτικοί μέσοι όροι όλων των μετρικών και των δύο συνθηκών.....	62
Πίνακας B.1: Ακρίβεια (precision, %) στα καθαρά δεδομένα.....	74
Πίνακας B.2: Ανάκληση (recall, %) στα καθαρά δεδομένα.....	74
Πίνακας B.3: Ακρίβεια (precision, %) στα θορυβώδη δεδομένα.....	74
Πίνακας B.4: Ανάκληση (recall, %) στα θορυβώδη δεδομένα.....	75
Πίνακας B.5: Χρόνος CPU (δευτερόλεπτα), καθαρά δεδομένα.....	75
Πίνακας B.6: Χρόνος CPU (δευτερόλεπτα), θορυβώδη δεδομένα.....	76
Πίνακας B.7: Διάμεσος τιμή k , καθαρά δεδομένα.....	76
Πίνακας B.8: Διάμεσος τιμή k , θορυβώδη δεδομένα.....	76
Πίνακας B.9: Ποσοστό (%) σημείων εκπαίδευσης στην εφεδρική τιμή $k = \max k$, καθαρά δεδομένα.....	77
Πίνακας B.10: Ποσοστό (%) σημείων εκπαίδευσης στην εφεδρική τιμή $k = \max k$, θορυβώδη δεδομένα.....	77

Συντομογραφίες

1NC	One Nearest Cluster (μία εγγύτερη συστάδα)
AdaNN	Adaptive k-Nearest Neighbor (προσαρμοστικός κατηγοριοποιητής k εγγύτερων γειτόνων)
Ada-k-NN	Adaptive k-Nearest Neighbor (προσαρμοστικός k-NN βασισμένος σε νευρωνικό δίκτυο)
auROC	Area under the Receiver Operating Characteristic curve
BBO	Biogeography Based Optimization
CNN	Condensed Nearest Neighbor
CPU	Central Processing Unit
CV	Cross Validation (διασταυρούμενη επικύρωση)
Δ.Ε.	Διπλωματική Εργασία
ΔΙ.ΠΑ.Ε.	Διεθνές Πανεπιστήμιο της Ελλάδος
DROP	Decremental Reduction Optimization Procedure
ENN	Edited Nearest Neighbor
G-k-NN	Genetic k-Nearest Neighbor
GS-k-NN	Graph Sparse k-Nearest Neighbor
ICF	Iterative Case Filtering
IKNN	Informative k-Nearest Neighbor
k-NN	k-Nearest Neighbors (k εγγύτεροι γείτονες)
KEEL	Knowledge Extraction based on Evolutionary Learning
LOO	Leave-One-Out
NN	Nearest Neighbor (εγγύτερος γείτονας)
RHC	Reduction through Homogeneous Clusters
shd-k-NN	Subspace Homogeneity based Dynamic k-NN
SMOTE	Synthetic Minority Over-sampling Technique
UCI	University of California, Irvine (Machine Learning Repository)

Κεφάλαιο 1ο: Εισαγωγή

Το κεφάλαιο αυτό εισάγει το πρόβλημα με το οποίο ασχολείται η εργασία. Ορίζεται πρώτα η κατηγοριοποίηση ως τύπος προβλήματος επιβλεπόμενης μάθησης, περιγράφεται στη συνέχεια ο κατηγοριοποιητής των k εγγύτερων γειτόνων και αναλύεται ο ρόλος της παραμέτρου k , η οποία καθορίζει κρίσιμα τη συμπεριφορά του. Ακολουθεί η διατύπωση του κινήτρου της εργασίας και των ερευνητικών ερωτημάτων που τίθενται, η καταγραφή της συνεισφοράς και, τέλος, η παρουσίαση της οργάνωσης των υπόλοιπων κεφαλαίων.

1.1 Κατηγοριοποίηση

Η κατηγοριοποίηση (classification) αποτελεί έναν από τους θεμελιώδεις τύπους προβλημάτων της επιβλεπόμενης μηχανικής μάθησης. Ζητούμενο είναι η ανάθεση μιας ετικέτας κλάσης σε ένα αντικείμενο του οποίου η κλάση είναι άγνωστη, με βάση ένα σύνολο παραδειγμάτων των οποίων η κλάση είναι γνωστή εκ των προτέρων. Τυπικά, έστω σύνολο εκπαίδευσης $TR = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, όπου κάθε x_i ανήκει στον χώρο των χαρακτηριστικών R^d και κάθε y_i σε ένα πεπερασμένο σύνολο ετικετών $C = \{c_1, c_2, \dots, c_m\}$. Ζητούμενο είναι η κατασκευή μιας συνάρτησης $h: R^d \rightarrow C$, η οποία για κάθε νέο, άγνωστο σημείο x επιστρέφει μια πρόβλεψη της κλάσης του. Η ποιότητα της h κρίνεται από την επίδοσή της σε ένα ανεξάρτητο σύνολο ελέγχου TS , το οποίο δεν συμμετείχε καθόλου στην κατασκευή της [5], [6]. Ο διαχωρισμός αυτός είναι απαραίτητος: η επίδοση πάνω στα ίδια τα δεδομένα εκπαίδευσης είναι αισιόδοξα μεροληπτική και δεν αποτελεί ένδειξη γενικευτικής ικανότητας.

Οι αλγόριθμοι κατηγοριοποίησης διακρίνονται σε δύο μεγάλες οικογένειες ανάλογα με το πότε επιτελούν το υπολογιστικό τους έργο. Οι μέθοδοι **πρόθυμης μάθησης** (eager learning), όπως τα δέντρα απόφασης, οι μηχανές διανυσμάτων υποστήριξης ή τα νευρωνικά δίκτυα, επενδύουν σημαντικό χρόνο στη φάση της εκπαίδευσης προκειμένου να συνοψίσουν τα δεδομένα σε ένα συμπαγές μοντέλο, και στη συνέχεια ταξινομούν πολύ γρήγορα. Οι μέθοδοι **οκνηρής μάθησης** (lazy learning) ακολουθούν την αντίστροφη πορεία: αποθηκεύουν αυτούσιο το σύνολο εκπαίδευσης και αναβάλλουν κάθε υπολογισμό μέχρι την υποβολή ενός ερωτήματος [7]. Το πλεονέκτημα της δεύτερης προσέγγισης είναι ότι δεν επιβάλλει καμία παραδοχή για τη μορφή του ορίου απόφασης, το οποίο διαμορφώνεται εξ ολοκλήρου από τα ίδια τα δεδομένα. Το μειονέκτημα είναι ότι τόσο ο χρόνος ταξινόμησης όσο και οι απαιτήσεις μνήμης αυξάνονται γραμμικά με το μέγεθος του συνόλου εκπαίδευσης.

Η αξιολόγηση ενός κατηγοριοποιητή στηρίζεται στον πίνακα σύγχυσης, ο οποίος καταγράφει για κάθε κλάση τα αληθώς θετικά (TP), τα ψευδώς θετικά (FP), τα αληθώς αρνητικά (TN) και τα ψευδώς αρνητικά (FN). Οι μετρικές που προκύπτουν από αυτόν, καθώς και ο τρόπος με τον οποίο συναθροίζονται σε προβλήματα πολλών κλάσεων, ορίζονται αναλυτικά στην ενότητα 7.2, όπου παρουσιάζεται το πειραματικό πλαίσιο της εργασίας.

1.2 Κατηγοριοποίηση k -εγγύτερων γειτόνων

Ο κατηγοριοποιητής των k εγγύτερων γειτόνων (k -Nearest Neighbors, k -NN) κατέχει ιδιαίτερη θέση ανάμεσα στους αλγόριθμους κατηγοριοποίησης. Προτάθηκε ήδη από τη δεκαετία του 1950 [1] και θεμελιώθηκε θεωρητικά τη δεκαετία του 1960 [2], παραμένει όμως έως σήμερα ενεργό αντικείμενο έρευνας. Η μακροβιότητά του δεν οφείλεται σε κάποια εξαιρετική επίδοση, αλλά στον συνδυασμό

απλότητας, απουσίας παραδοχών για την κατανομή των δεδομένων, ευκολίας υλοποίησης, δυνατότητας ενσωμάτωσης νέων παραδειγμάτων χωρίς επανεκπαίδευση και ικανοποιητικής ακρίβειας σε ευρύ φάσμα προβλημάτων [5]. Ο αλγόριθμος δεν κατασκευάζει κανένα ρητό μοντέλο κατά τη φάση της εκπαίδευσης· αντιθέτως, αναβάλλει κάθε υπολογισμό μέχρι τη στιγμή που θα κληθεί να ταξινομήσει, γι' αυτό και κατατάσσεται στις μεθόδους σκληρής μάθησης [7].

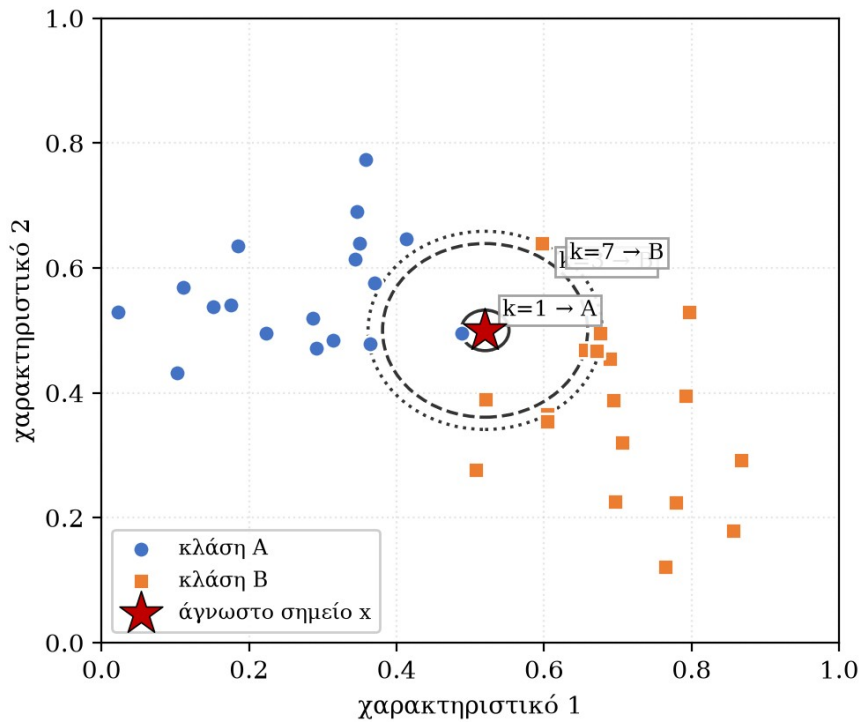
Ο κανόνας ταξινομεί ένα άγνωστο σημείο x με πλειοψηφική ψήφο ανάμεσα στα k σημεία του συνόλου εκπαίδευσης που βρίσκονται πλησιέστερα σε αυτό. Τυπικά, αν $N_k(x)$ συμβολίζει το σύνολο των k εγγύτερων γειτόνων του x , η πρόβλεψη δίνεται από τη σχέση (1.1), όπου $1(\cdot)$ είναι η συνάρτηση δείκτης που λαμβάνει την τιμή 1 όταν η συνθήκη ισχύει και 0 διαφορετικά.

$$h(x) = \arg \max_{c \in C} \sum_{x_j \in N_k(x)} 1(y_j = c) \quad (1.1)$$

Η εγγύτητα ορίζεται μέσω ενός μέτρου απόστασης. Ελλείψει ειδικής γνώσης για το πρόβλημα, χρησιμοποιείται σχεδόν πάντοτε η ευκλείδεια απόσταση, η οποία για δύο σημεία x_i και x_j με d χαρακτηριστικά δίνεται από τη σχέση (1.2) [2], [3]. Επειδή η ευκλείδεια απόσταση αθροίζει τετράγωνα διαφορών, χαρακτηριστικά με μεγαλύτερο εύρος τιμών κυριαρχούν έναντι εκείνων με μικρότερο. Για τον λόγο αυτόν η κανονικοποίηση των χαρακτηριστικών σε κοινό διάστημα, τυπικά στο $[0, 1]$, αποτελεί απαραίτητο προεπεξεργαστικό βήμα και όχι προαιρετική βελτίωση.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^d (a_r(x_i) - a_r(x_j))^2} \quad (1.2)$$

Στο Σχήμα 1.1 απεικονίζεται η λειτουργία του κανόνα σε ένα διδιάστατο παράδειγμα με δύο κλάσεις. Το άγνωστο σημείο βρίσκεται σε περιοχή επικάλυψης, και η απόφαση εξαρτάται από την ακτίνα εντός της οποίας συλλέγονται οι ψήφοι. Για $k = 1$ η απόφαση καθορίζεται από έναν μόνο γείτονα και είναι επομένως εξαιρετικά ευαίσθητη· καθώς το k αυξάνεται, ενσωματώνονται περισσότερες μαρτυρίες και η απόφαση σταθεροποιείται, χωρίς όμως αυτό να σημαίνει ότι γίνεται αναγκαστικά και ορθότερη. Το παράδειγμα καθιστά σαφές ότι η επιλογή του k δεν είναι τεχνική λεπτομέρεια αλλά ουσιαστική απόφαση μοντελοποίησης.



Σχήμα 1.1: Ταξινόμηση ενός άγνωστου σημείου με τον κανόνα k-NN για $k = 1, 3$ και 7 . Η απόφαση μεταβάλλεται ανάλογα με το πλήθος των γειτόνων που λαμβάνονται υπόψη.

Ένα ζήτημα που προκύπτει άμεσα από τη σχέση (1.1) είναι η αντιμετώπιση των ισοπαλιών. Όταν δύο ή περισσότερες κλάσεις συγκεντρώνουν τον ίδιο αριθμό ψήφων, ο κανόνας δεν ορίζει μονοσήμαντα την έξοδο. Οι συνήθεις λύσεις είναι η επιλογή της κλάσης με τον πλησιέστερο εκπρόσωπο, η στάθμιση των ψήφων με το αντίστροφο της απόστασης, ή η ντετερμινιστική επιλογή της κλάσης με τη μικρότερη ετικέτα. Η τελευταία επιλέχθηκε στην παρούσα εργασία, επειδή εφαρμόζεται με τον ίδιο ακριβώς τρόπο σε όλες τις συγκρινόμενες μεθόδους και συνεπώς δεν εισάγει συστηματικό πλεονέκτημα σε καμία από αυτές. Ανάλογο ζήτημα ανακύπτει όταν πολλά σημεία ισαπέχουν από το x : τότε η κατάταξή τους είναι μαθηματικά απροσδιόριστη, γεγονός με πρακτικές συνέπειες που συζητούνται στην ενότητα 6.9.

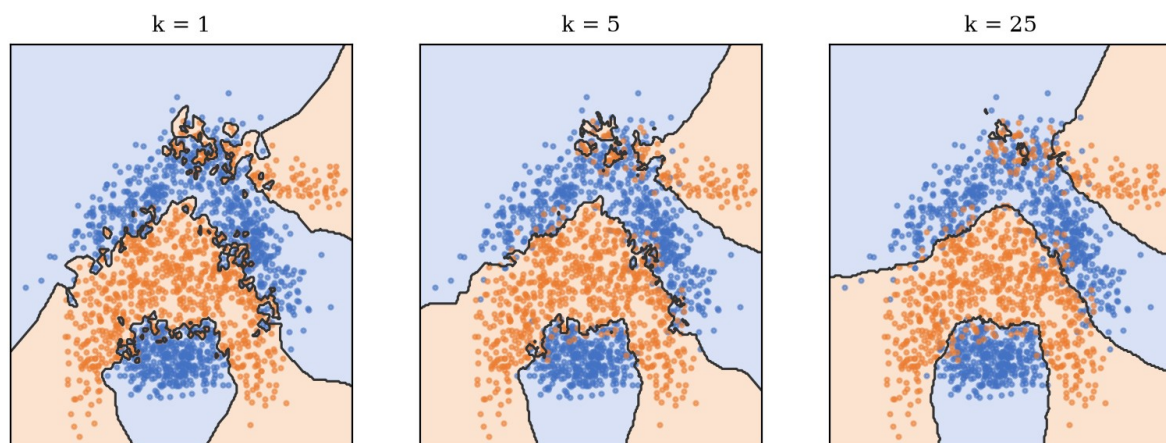
Ένα δεύτερο πρακτικό ζήτημα είναι το κόστος της αναζήτησης. Η εξαντλητική αναζήτηση των k εγγύτερων γειτόνων ενός σημείου απαιτεί τον υπολογισμό n αποστάσεων και έχει επομένως πολυπλοκότητα $O(n \cdot d)$, ενώ για ολόκληρο σύνολο ελέγχου μεγέθους m το κόστος ανέρχεται σε $O(m \cdot n \cdot d)$. Το πρόβλημα αντιμετωπίζεται με χωρικές δομές ευρετηρίασης, με πιο γνωστή το k -d δέντρο [29], το οποίο διαμερίζει αναδρομικά τον χώρο με υπερεπίπεδα κάθετα στους άξονες και επιτρέπει την απόρριψη ολόκληρων υποδέντρων που δεν μπορούν να περιέχουν γείτονα πλησιέστερο από τον ήδη ευρεθέντα. Η αποτελεσματικότητα των δομών αυτών υποβαθμίζεται καθώς αυξάνεται η διαστατικότητα, φαινόμενο γνωστό ως κατάρα της διαστατικότητας.

1.3 Η παράμετρος k

Η απλότητα του k -NN έχει ένα τίμημα: ο αλγόριθμος μεταθέτει ολόκληρο το βάρος της απόφασης σε δύο επιλογές που ο ίδιος δεν καθορίζει, στο μέτρο απόστασης και στο πλήθος των γειτόνων k που λαμβάνονται υπόψη. Το μέτρο απόστασης υπαγορεύεται συνήθως από τη φύση των δεδομένων. Η παράμετρος k , αντιθέτως, δεν έχει προφανή τιμή και δεν προκύπτει από κανέναν αναλυτικό κανόνα.

Η επίδραση της παραμέτρου k στο όριο απόφασης γίνεται καλύτερα κατανοητή μέσω της αντιστάθμισης μεροληψίας-διακύμανσης (bias-variance tradeoff) [5]. Για $k = 1$ ο κατηγοριοποιητής έχει μηδενική μεροληψία ως προς τα δεδομένα εκπαίδευσης, με την έννοια ότι αναπαράγει τέλεια κάθε ετικέτα τους, αλλά μέγιστη διακύμανση: μια μικρή μεταβολή στο σύνολο εκπαίδευσης μπορεί να μεταβάλει δραστικά το όριο. Καθώς το k αυξάνεται, η πρόβλεψη προκύπτει από τη συσσώρευση περισσότερων παρατηρήσεων, η διακύμανση μειώνεται και το όριο εξομαλύνεται. Ταυτόχρονα όμως αυξάνεται η μεροληψία, καθώς η γειτονιά εκτείνεται σε περιοχές του χώρου όπου η υπόθεση της τοπικής ομοιογένειας παύει να ισχύει. Στο όριο, για k ίσο με το πλήθος των σημείων εκπαίδευσης, ο κατηγοριοποιητής εκφυλίζεται και επιστρέφει πάντοτε την πολυπληθέστερη κλάση.

Στο Σχήμα 1.2 απεικονίζεται η επίδραση αυτή πάνω στο σύνολο δεδομένων Banana, ένα διδιάστατο πρόβλημα δύο κλάσεων με μη γραμμικό όριο. Για $k = 1$ το όριο είναι κατακερματισμένο και περιλαμβάνει πολυάριθμους απομονωμένους θύλακες, καθένας από τους οποίους οφείλεται σε ένα και μόνο σημείο εκπαίδευσης. Για $k = 5$ οι περισσότεροι θύλακες έχουν εξαφανιστεί και το όριο ακολουθεί την πραγματική δομή των δεδομένων. Για $k = 25$ το όριο έχει εξομαλυνθεί περαιτέρω και έχει αρχίσει να χάνει τις λεπτές καμπυλώσεις που διαχωρίζουν τις δύο κλάσεις στα άκρα τους. Η βέλτιστη τιμή βρίσκεται προφανώς κάπου ενδιάμεσα, χωρίς όμως να υπάρχει αναλυτικός τρόπος να εντοπιστεί εκ των προτέρων.



Σχήμα 1.2: Όριο απόφασης του k -NN στο σύνολο δεδομένων Banana για $k = 1, 5$ και 25 . Η αύξηση του k εξομαλύνει το όριο και εξαλείφει τους απομονωμένους θύλακες.

Μικρές τιμές του k παράγουν λοιπόν εξαιρετικά ευέλικτα όρια απόφασης που ακολουθούν πιστά τα δεδομένα εκπαίδευσης, αλλά είναι ταυτόχρονα ευάλωτα σε μεμονωμένα εσφαλμένα ή ακραία σημεία. Μεγάλες τιμές εξομαλύνουν το όριο και προσφέρουν ανθεκτικότητα στον θόρυβο, με κόστος όμως την απώλεια λεπτομέρειας σε περιοχές όπου οι κλάσεις εναλλάσσονται πυκνά. Το κρίσιμο σημείο, το οποίο διατρέχει ολόκληρη την εργασία, είναι ότι η ισορροπία ανάμεσα στα δύο άκρα δεν είναι ίδια για όλα τα σύνολα δεδομένων, **ούτε καν για όλες τις περιοχές του ίδιου συνόλου**.

1.4 Κίνητρο

Η καθιερωμένη πρακτική αντιμετώπισης του προβλήματος συνίσταται στον προσδιορισμό μίας και μοναδικής τιμής k , η οποία επιλέγεται μέσω διασταυρούμενης επικύρωσης (cross-validation) πάνω στο σύνολο εκπαίδευσης και στη συνέχεια εφαρμόζεται αναλλοίωτη σε κάθε άγνωστο σημείο [23], [24]. Η προσέγγιση αυτή είναι μεθοδολογικά ορθή και ευρέως αποδεκτή, ενσωματώνει όμως μια σιωπηρή παραδοχή: ότι η βέλτιστη τιμή του k είναι σταθερή σε ολόκληρο τον χώρο των χαρακτηριστικών. Η

παραδοχή αυτή σπανίως ισχύει στην πράξη. Σε ένα ρεαλιστικό σύνολο δεδομένων συνυπάρχουν περιοχές όπου οι κλάσεις διαχωρίζονται καθαρά, και στις οποίες ακόμη και ο εγγύτερος γείτονας αρκεί για σωστή απόφαση, με περιοχές πυκνής επικάλυψης, όπου η απόφαση απαιτεί τη συσσωρευμένη μαρτυρία δεκάδων γειτόνων. Μία καθολική τιμή k αναγκαστικά συμβιβάζεται ανάμεσα σε αυτές τις αντικρουόμενες απαιτήσεις [32].

Η τεχνική AdaNN (Adaptive k-Nearest Neighbor), που προτάθηκε από τους Sun και Huang [3], εγκαταλείπει την παραδοχή αυτή. Αντί για μία καθολική τιμή, ο AdaNN αναθέτει σε κάθε σημείο του συνόλου εκπαίδευσης μια δική του βέλτιστη τιμή k , ορισμένη ως το μικρότερο πλήθος γειτόνων που αρκεί ώστε το σημείο αυτό να κατηγοριοποιηθεί σωστά όταν αντιμετωπιστεί ως άγνωστο. Κατά τη φάση του ελέγχου, κάθε άγνωστο σημείο δανείζεται την τιμή k του εγγύτερου γείτονά του στο σύνολο εκπαίδευσης, με βάση την εύλογη υπόθεση ότι γειτονικά σημεία μοιράζονται παρόμοιες τοπικές συνθήκες. Ο μηχανισμός είναι απλός και δεν απαιτεί καμία επιπλέον παράμετρο πέραν του άνω ορίου αναζήτησης, το οποίο στην αρχική δημοσίευση ορίζεται στο 9.

Το άνω αυτό όριο και ο κανόνας επιλογής της τιμής k είναι ακριβώς τα σημεία στα οποία εστιάζει η παρούσα εργασία. Το όριο 9 είναι μια σύμβαση χωρίς θεωρητική θεμελίωση, και σε δεδομένα με υψηλό επίπεδο θορύβου ενδέχεται να είναι υπερβολικά περιοριστικό: αν καμία τιμή έως το 9 δεν κατηγοριοποιεί σωστά ένα σημείο, ο αλγόριθμος καταφεύγει αναγκαστικά στην τιμή 9, η οποία όμως δεν έχει επιλεγεί επειδή λειτουργεί, αλλά επειδή εξαντλήθηκε το εύρος. Παράλληλα, ο κανόνας «η πρώτη τιμή k που πετυχαίνει» είναι εξ ορισμού εύθραυστος: αρκεί μια τυχαία, οριακή επιτυχία σε μια μικρή τιμή k για να παγιωθεί μια επιλογή που δεν αντανακλά καμία σταθερή τοπική ιδιότητα.

Σκοπός της εργασίας είναι, επομένως, η υλοποίηση του αλγορίθμου AdaNN, ο σχεδιασμός και η υλοποίηση τεσσάρων παραλλαγών του, και η συστηματική πειραματική αξιολόγησή τους σε καθαρά και σε θορυβώδη δεδομένα. Συγκεκριμένα, επιδιώκεται να απαντηθούν τα ακόλουθα πέντε ερευνητικά ερωτήματα:

- **E1.** Βελτιώνει η επέκταση του εύρους αναζήτησης του k από [1..9] σε [1..30] την επίδοση του AdaNN, και υπό ποιες συνθήκες;
- **E2.** Οδηγεί ο κανόνας κυλιόμενου παραθύρου, ο οποίος απαιτεί τρεις διαδοχικές τιμές k να κατηγοριοποιούν σωστά ένα σημείο, σε πιο αξιόπιστη ανάθεση τιμών k σε σχέση με τον κανόνα της πρώτης επιτυχίας;
- **E3.** Πώς επηρεάζει το βήμα μετατόπισης του παραθύρου (1, 2 ή 3) τόσο τις τιμές k που ανατίθενται όσο και την τελική επίδοση;
- **E4.** Πώς συμπεριφέρονται οι παραλλαγές παρουσία θορύβου στις ετικέτες του συνόλου εκπαίδευσης, και επιβεβαιώνεται η υπόθεση ότι ο θόρυβος ευνοεί μεγαλύτερες τιμές k ;
- **E5.** Πώς συγκρίνεται η τοπικά προσαρμοστική επιλογή του k με τον κλασικό καθολικό προσδιορισμό μέσω διασταυρούμενης επικύρωσης, τόσο ως προς την ποιότητα κατηγοριοποίησης όσο και ως προς το υπολογιστικό κόστος;

1.5 Συνεισφορά

Η συνεισφορά της εργασίας συνοψίζεται στα εξής σημεία. **Πρώτον**, παρέχει μια πλήρη, διανυσματοποιημένη υλοποίηση του AdaNN και των τεσσάρων παραλλαγών του σε Python, η ορθότητα της οποίας επαληθεύεται με ανεξάρτητη υλοποίηση εξαντλητικού υπολογισμού. **Δεύτερον**, εισάγει και τυποποιεί τον κανόνα κυλιόμενου παραθύρου ως εναλλακτική στον κανόνα της πρώτης

επιτυχίας, διατυπώνοντάς τον με τρόπο που επιτρέπει τη μελέτη του πλάτους και του βήματος ως ανεξάρτητων παραμέτρων. **Τρίτον**, αξιολογεί όλες τις μεθόδους σε δεκατέσσερα καθιερωμένα σύνολα δεδομένων του αποθετηρίου KEEL υπό δύο συνθήκες θορύβου, παρέχοντας έτσι στοιχεία όχι μόνο για το αν μια παραλλαγή είναι καλύτερη, αλλά και για το **πότε** γίνεται καλύτερη. **Τέταρτον**, ποσοτικοποιεί ρητά την κατανομή των τιμών k που ανατίθενται, στοιχείο που στη βιβλιογραφία συνήθως παραλείπεται και που εδώ αποδεικνύεται καθοριστικό για την ερμηνεία των αποτελεσμάτων.

1.6 Οργάνωση της εργασίας

Το **Κεφάλαιο 2** εξετάζει την απόδοση σταθερής τιμής στην παράμετρο k : τη θεμελίωση του κανόνα, τη ρύθμιση του k με διασταυρούμενη επικύρωση, τους αναλυτικούς και εμπειρικούς κανόνες που έχουν προταθεί, καθώς και τους εγγενείς περιορισμούς κάθε καθολικής επιλογής. Το **Κεφάλαιο 3** παρουσιάζει τις τεχνικές δυναμικού προσδιορισμού της τιμής k , οργανωμένες σε έξι μεθοδολογικές οικογένειες, ακολουθώντας την ταξινόμηση της πρόσφατης βιβλιογραφικής επισκόπησης των Papanikolaou, Evangelidis και Ougiaroglou [32].

Το **Κεφάλαιο 4** αναλύει διεξοδικά τον αλγόριθμο AdaNN, με ψευδοκώδικα, αναλυτικό αριθμητικό παράδειγμα, μελέτη των οριακών περιπτώσεων και ανάλυση της υπολογιστικής πολυπλοκότητας. Το **Κεφάλαιο 5** διατυπώνει τις τέσσερις νέες παραλλαγές που σχεδιάστηκαν στο πλαίσιο της εργασίας. Το **Κεφάλαιο 6** παρουσιάζει τις υλοποιήσεις σε Python, με στιγμιότυπα του κώδικα και επεξήγηση του ρόλου κάθε τμήματος.

Το **Κεφάλαιο 7** περιγράφει την πειραματική μελέτη: τα σύνολα δεδομένων του αποθετηρίου KEEL, το πειραματικό περιβάλλον και τα αποτελέσματα, μαζί με την ερμηνεία τους και τους περιορισμούς τους. Το **Κεφάλαιο 8** συνοψίζει τα συμπεράσματα και προτείνει μελλοντικές επεκτάσεις. Ακολουθούν η βιβλιογραφία και δύο παραρτήματα, με τον πηγαίο κώδικα και τους αναλυτικούς πίνακες μετρήσεων αντίστοιχα.

Κεφάλαιο 2ο: Απόδοση σταθερής τιμής στην παράμετρο k

Το κεφάλαιο αυτό εξετάζει την επικρατέστερη πρακτική αντιμετώπισης του προβλήματος της παραμέτρου k , δηλαδή τον προσδιορισμό μίας σταθερής τιμής που εφαρμόζεται ενιαία σε ολόκληρο τον χώρο των δεδομένων. Παρουσιάζεται πρώτα η θεωρητική θεμελίωση του κανόνα των εγγύτερων γειτόνων, η οποία εξηγεί γιατί η αύξηση του k έχει νόημα. Ακολουθεί η αναλυτική περιγραφή της ρύθμισης του k με διασταυρούμενη επικύρωση, που αποτελεί και το σημείο αναφοράς της πειραματικής μελέτης, και στη συνέχεια παρουσιάζονται οι εναλλακτικές τεχνικές καθολικού προσδιορισμού που έχουν προταθεί στη βιβλιογραφία. Το κεφάλαιο κλείνει με τους εγγενείς περιορισμούς της σταθερής τιμής, οι οποίοι οδηγούν στις δυναμικές προσεγγίσεις του Κεφαλαίου 3.

2.1 Θεωρητική θεμελίωση του κανόνα

Η πρώτη διατύπωση του κανόνα των εγγύτερων γειτόνων αποδίδεται στους Fix και Hodges [1], οι οποίοι σε τεχνική αναφορά του 1951 πρότειναν μια μη παραμετρική μέθοδο διαχωριστικής ανάλυσης βασισμένη στην εγγύτητα. Η εργασία παρέμεινε σχετικά άγνωστη μέχρι τη δημοσίευση των Cover και Hart [2] το 1967, οι οποίοι απέδειξαν το θεμελιώδες αποτέλεσμα ότι, καθώς το μέγεθος του συνόλου εκπαίδευσης τείνει στο άπειρο, το ασυμπτωτικό σφάλμα του κανόνα του ενός εγγύτερου γείτονα φράσσεται άνω από το διπλάσιο του σφάλματος Bayes. Το αποτέλεσμα αυτό είναι αξιοσημείωτο διότι εξασφαλίζει ότι ακόμη και ο απλούστερος δυνατός κανόνας, χωρίς καμία εκπαίδευση και καμία παραδοχή, δεν μπορεί να είναι περισσότερο από δύο φορές χειρότερος από τον θεωρητικά βέλτιστο κατηγοριοποιητή.

Το ίδιο θεώρημα δείχνει επίσης ότι, καθώς το k αυξάνεται με κατάλληλο ρυθμό ως προς το μέγεθος του συνόλου, δηλαδή όταν $k \rightarrow \infty$ και $k/n \rightarrow 0$, το σφάλμα του κανόνα συγκλίνει στο σφάλμα Bayes. Η αύξηση του k έχει επομένως θεωρητική θεμελίωση και όχι μόνο εμπειρική. Το πρόβλημα είναι ότι το αποτέλεσμα αυτό είναι ασυμπτωτικό: δεν λέει τίποτε για το ποια είναι η κατάλληλη τιμή του k όταν το σύνολο εκπαίδευσης έχει πεπερασμένο και συχνά μικρό μέγεθος, ούτε πώς αυτή εξαρτάται από τη διαστατικότητα, την πυκνότητα ή τον θόρυβο των δεδομένων. Στην πράξη, το k παραμένει μια υπερπαράμετρος που πρέπει να καθοριστεί εμπειρικά.

2.2 Ρύθμιση του k με διασταυρούμενη επικύρωση

Η επικρατέστερη τεχνική για τον καθορισμό της «βέλτιστης» τιμής k είναι η διασταυρούμενη επικύρωση, διότι δεν υπάρχει γενικός κανόνας επιλογής της τιμής και η καταλληλότητά της εξαρτάται από το εκάστοτε σύνολο εκπαίδευσης [32]. Η μεθοδολογία θεμελιώθηκε από τον Stone [24] και μελετήθηκε συστηματικά από τον Kohavi [23], ο οποίος συνέκρινε πειραματικά διάφορες παραλλαγές της και κατέληξε υπέρ της στρωματοποιημένης εκδοχής ως καλύτερου συμβιβασμού ανάμεσα σε μεροληψία και διακύμανση.

Στην F -πτυχη εκδοχή της, το σύνολο εκπαίδευσης διαμερίζεται σε F ισομεγέθη υποσύνολα. Κάθε υποσύνολο χρησιμοποιείται μία φορά ως σύνολο επικύρωσης ενώ τα υπόλοιπα λειτουργούν ως σύνολο εκπαίδευσης, και η τελική εκτίμηση προκύπτει ως ο μέσος όρος των F επιμέρους εκτιμήσεων. Η στρωματοποιημένη παραλλαγή (stratified cross-validation) διατηρεί επιπλέον σε κάθε υποσύνολο την αναλογία των κλάσεων του αρχικού συνόλου, μειώνοντας τη διακύμανση της εκτίμησης. Για τη ρύθμιση του k η διαδικασία επαναλαμβάνεται για κάθε υποψήφια τιμή του διαστήματος $[1, K]$ και επιλέγεται η τιμή με τη μεγαλύτερη μέση επίδοση επικύρωσης.

Κρίσιμο σημείο, το οποίο τηρήθηκε αυστηρά στην παρούσα εργασία, είναι ότι η διασταυρούμενη επικύρωση εφαρμόζεται **αποκλειστικά** στο σύνολο εκπαίδευσης. Το σύνολο ελέγχου παραμένει εντελώς αχρησιμοποίητο μέχρι την τελική αξιολόγηση. Η παραβίαση της αρχής αυτής, ακόμη και έμμεση μέσω επαναλαμβανόμενης χρήσης του συνόλου ελέγχου για επιλογή παραμέτρων, οδηγεί σε αισιόδοξα μεροληπτικές εκτιμήσεις που δεν αντανακλούν την πραγματική επίδοση σε νέα δεδομένα.

Το κόστος της προσέγγισης είναι σημαντικό αλλά προβλέψιμο: για σάρωση K υποψήφιων τιμών με F πτυχές απαιτούνται $K \cdot F$ εκπαιδεύσεις και αξιολογήσεις. Στην περίπτωση του k -NN, ωστόσο, το κόστος περιορίζεται δραστικά αν οι γείτονες υπολογιστούν **μία φορά ανά πτυχή** για τη μέγιστη τιμή K και επαναχρησιμοποιηθούν για όλες τις μικρότερες τιμές, τεχνική που υιοθετεί και η υλοποίηση της εργασίας (ενότητα 6.7). Στην ίδια κατεύθυνση, οι Hamerly και Speegle [52] πρότειναν ταχύτερη τεχνική διασταυρούμενης επικύρωσης, η οποία εξετάζει μεγαλύτερο πλήθος τιμών k στον ίδιο χρόνο, μειώνοντας τη χρονική πολυπλοκότητα σε $O(K^*)$, όπου K^* η μέγιστη εξεταζόμενη τιμή.

2.3 Αναλυτικοί και εμπειρικοί κανόνες επιλογής

Παράλληλα με τη διασταυρούμενη επικύρωση, έχουν προταθεί και κλειστοί κανόνες που υπολογίζουν το k απευθείας από χαρακτηριστικά του συνόλου δεδομένων. Ο πιο διαδεδομένος εμπειρικός κανόνας είναι $k \approx \sqrt{n}$, όπου n το πλήθος των σημείων εκπαίδευσης, ο οποίος χρησιμοποιείται ευρέως ως σημείο αναφοράς σε πειραματικές συγκρίσεις. Πρόκειται για πρακτικό κανόνα και όχι για θεωρητικά θεμελιωμένο αποτέλεσμα.

Οι Enas και Choi [53], σε μια από τις παλαιότερες σχετικές μελέτες, υποστήριξαν ότι η βέλτιστη τιμή του k εξαρτάται από τη διαστατικότητα, το μέγεθος και τη δομή του δείγματος, και πρότειναν εξισώσεις υπολογισμού του k συναρτήσει της διαφοράς των αναλογιών των κλάσεων και της διαφοράς των πινάκων συνδιακύμανσης. Οι Hand και Vinciotti [54] εξέτασαν ειδικά την περίπτωση δυαδικών προβλημάτων με ανισόρροπες κλάσεις και παρέθεσαν προσεγγίσεις της μορφής $k \approx n^{1/8}$ ή $k \approx n^{3/8}$. Οι κανόνες αυτοί έχουν το πλεονέκτημα του μηδενικού υπολογιστικού κόστους, αλλά αγνοούν πλήρως την τοπική δομή των δεδομένων και σπανίως υπερτερούν της διασταυρούμενης επικύρωσης στην πράξη.

2.4 Άλλες τεχνικές καθολικού προσδιορισμού

Η βιβλιογραφική επισκόπηση των Papanikolaou, Evangelidis και Ougiaroglou [32] καταγράφει είκοσι οκτώ δημοσιεύσεις που αφορούν τον προσδιορισμό της τιμής k και τις ταξινομεί ανάλογα με το επίπεδο στο οποίο γίνεται η επιλογή. Επτά από αυτές αφορούν καθολικό προσδιορισμό, δηλαδή αναζητούν μία και μόνη τιμή για ολόκληρο το σύνολο δεδομένων, χωρίς όμως να καταφεύγουν αναγκαστικά σε διασταυρούμενη επικύρωση. Οι σημαντικότερες από αυτές συνοψίζονται στη συνέχεια και στον Πίνακα 2.1.

Οι Kardan, Kavian και Esmaceli [33] διατύπωσαν την επιλογή του k ως πρόβλημα καθολικής βελτιστοποίησης και το επέλυσαν με τον εξελικτικό αλγόριθμο Biogeography Based Optimization (BBO). Ο αλγόριθμος χρησιμοποιεί ένα πολυμερές χρωμόσωμα που κωδικοποιεί ταυτόχρονα την επιλογή χαρακτηριστικών, τη στάθμισή τους και την τιμή του k , βελτιστοποιώντας και τα τρία μαζί. Η μέθοδος συγκρίθηκε με έξι εξελικτικούς και δεκατέσσερις μη εξελικτικούς αλγόριθμους σε δέκα σύνολα δεδομένων και υπερτέρησε όλων ως προς την ορθότητα, τον δείκτη Kohen's Kappa και τον ρυθμό μείωσης.

Οι Ferrer-Troyano, Aguilar-Ruiz και Riquelme [40] πρότειναν μια ευρετική διαδικασία με την οποία η τιμή του k επιλέγεται αυτόματα, χωρίς καμία παρέμβαση του χρήστη. Η ιδέα είναι ότι ο αλγόριθμος

αναζητά την τιμή που κατηγοριοποιεί σωστά την πλειονότητα των σημείων εκπαίδευσης. Η προσέγγιση συγκρίθηκε με τον συμβατικό k-NN για περιττές τιμές στο διάστημα [1, 51] σε είκοσι πέντε σύνολα δεδομένων, υπερτερώντας σε δεκατρείς περιπτώσεις, εκ των οποίων σε πέντε η διαφορά ήταν στατιστικά σημαντική.

Οι Nock, Sebban και Bernard [41] δεν προτείνουν διαφορετική διαδικασία επιλογής, αλλά τροποποιούν τον ίδιο τον κανόνα ώστε η επίδοσή του να εξαρτάται λιγότερο από την επιλεγμένη τιμή. Η βασική ιδέα είναι ότι, για δεδομένο k, αν ένα σημείο x ψηφίζει για ένα σημείο y, τότε και το y ψηφίζει για το x, ακόμη και αν δεν ανήκει στους k εγγύτερους γείτονές του. Ο αλγόριθμος συγκρίθηκε με τον 1-NN σε είκοσι εννέα σύνολα δεδομένων και υπερτέρησε, χωρίς ωστόσο στατιστική σημαντικότητα.

Τέλος, στην κατηγορία των πιθανοτικών μεθόδων καθολικού προσδιορισμού, οι Holmes και Adams [50] παρουσίασαν μια αναλυτική πιθανοτική παραλλαγή του k-NN, η οποία αντιμετωπίζει την αβεβαιότητα ως προς την τιμή του k μέσω μιας εκ των προτέρων κατανομής (prior distribution) ορισμένης πάνω σε αυτήν. Η μέθοδος συγκρίθηκε με τον συμβατικό k-NN ως προς τον ρυθμό σφάλματος κατηγοριοποίησης και εμφάνισε ανταγωνιστική επίδοση.

Πίνακας 2.1: Τεχνικές καθολικού προσδιορισμού της τιμής k.

Εργασία	Έτος	Μεθοδολογική οικογένεια	Βασική ιδέα
Enas & Choi [53]	1986	Αναλυτική	Εξισώσεις υπολογισμού του k από τη δομή του δείγματος
Ferrer-Troyano κ.ά. [40]	2001	Ευρετική	Το k που ταξινομεί σωστά την πλειονότητα των σημείων
Holmes & Adams [50]	2002	Πιθανοτική	Εκ των προτέρων κατανομή πάνω στην τιμή του k
Hand & Vinciotti [54]	2003	Αναλυτική	Προσεγγίσεις $k \approx n^{(2/8)}$ και $k \approx n^{(3/8)}$
Nock κ.ά. [41]	2003	Ευρετική	Αμφίδρομη ψηφοφορία, μικρότερη εξάρτηση από το k
Hamerly & Speegle [52]	2012	Άλλη	Ταχύτερη διασταυρούμενη επικύρωση για πολλά k
Kardan κ.ά. [33]	2013	Γενετικός αλγόριθμος	Ταυτόχρονη βελτιστοποίηση χαρακτηριστικών και k (BBO)
Kohavi [23], Stone [24]	1974/ 1995	Διασταυρούμενη επικύρωση	Στρωματοποιημένη F-πτυχη επικύρωση για $k = 1..K$

2.5 Οι περιορισμοί μιας σταθερής τιμής

Ο θεμελιώδης περιορισμός κάθε καθολικής προσέγγισης δεν είναι υπολογιστικός αλλά εννοιολογικός. Η επιλεγμένη τιμή k βελτιστοποιεί τη **μέση** επίδοση σε ολόκληρο τον χώρο των δεδομένων και επομένως εκφράζει έναν συμβιβασμό ανάμεσα σε περιοχές με ριζικά διαφορετικές ανάγκες. Σε ένα σύνολο όπου το 80 % των σημείων βρίσκεται σε καθαρά διαχωρισμένες περιοχές και το 20 % σε ζώνη έντονης επικάλυψης, η βέλτιστη καθολική τιμή θα είναι υπερβολικά μεγάλη για την πρώτη ομάδα και υπερβολικά μικρή για τη δεύτερη.

Ο περιορισμός γίνεται ακόμη πιο έντονος σε σύνολα δεδομένων που συνδυάζουν καλά και κακά διαχωρισμένες κλάσεις, ανισόρροπες κλάσεις, αραιές και πυκνές γειτονιές ή θορυβώδεις υποχώρους. Όπως επισημαίνουν οι Papanikolaou, Evangelidis και Ougiaroglou [32], κάθε υποχώρος ενός συνόλου δεδομένων έχει διαφορετική πυκνότητα και διαφορετικό επίπεδο θορύβου, και επομένως μια καθολικά ορισμένη σταθερή τιμή k δεν είναι κατάλληλη. Αντ' αυτού, θα πρέπει να λαμβάνονται υπόψη τα ιδιαίτερα χαρακτηριστικά του συνόλου δεδομένων και του υποχώρου στον οποίο βρίσκεται κάθε σημείο, ώστε η τιμή k να προσδιορίζεται δυναμικά.

Ένας δεύτερος, λιγότερο προφανής περιορισμός αφορά τις μικρές κλάσεις. Μια μεγάλη καθολική τιμή k εξαφανίζει στην πράξη τις κλάσεις που εκπροσωπούνται από ελάχιστα σημεία: με τριάντα γείτονες, μια κλάση με λίγα παραδείγματα δεν μπορεί ποτέ να συγκεντρώσει πλειοψηφία και επομένως δεν προβλέπεται ποτέ. Το φαινόμενο αυτό δεν αποτυπώνεται στην ορθότητα, καταγράφεται όμως καθαρά στο μακρο-συναθροισμένο F-score, όπως θα φανεί στα πειραματικά αποτελέσματα της ενότητας 7.3. Ακριβώς αυτές οι παρατηρήσεις κινητοποίησαν την ανάπτυξη των τοπικά προσαρμοστικών μεθόδων, οι οποίες εξετάζονται στο επόμενο κεφάλαιο.

Κεφάλαιο 3ο: Απόδοση δυναμικής τιμής στην παράμετρο k

Το κεφάλαιο αυτό παρουσιάζει τις τεχνικές που εγκαταλείπουν την ιδέα μιας ενιαίας τιμής k και την προσδιορίζουν δυναμικά, ανάλογα με τη θέση κάθε σημείου στον χώρο των χαρακτηριστικών. Η οργάνωση του κεφαλαίου ακολουθεί την ταξινόμηση της πρόσφατης βιβλιογραφικής επισκόπησης των Papanikolaou, Evangelidis και Ougiaroglou [32], η οποία εξέτασε είκοσι οκτώ δημοσιεύσεις της περιόδου 1986–2020 και τις ομαδοποίησε σε έξι μεθοδολογικές οικογένειες: γενετικοί αλγόριθμοι, νευρωνικά δίκτυα, πρότυπα και συσταδοποίηση, ευρετικές μέθοδοι, πιθανοτικά πλαίσια και λοιπές προσεγγίσεις. Στο τέλος του κεφαλαίου εντοπίζεται το ερευνητικό κενό που καλείται να καλύψει η παρούσα εργασία.

3.1 Επίπεδα προσδιορισμού της τιμής k

Πριν από την παρουσίαση των επιμέρους μεθόδων, είναι χρήσιμη μια διάκριση ως προς το **επίπεδο ανάλυσης** στο οποίο γίνεται η επιλογή της τιμής k . Η επισκόπηση [32] διακρίνει τέσσερα τέτοια επίπεδα, τα οποία συνοψίζονται στον Πίνακα 3.1. Στο **καθολικό** επίπεδο αναζητείται μία τιμή για ολόκληρο το σύνολο δεδομένων, όπως περιγράφηκε στο Κεφάλαιο 2. Στο επίπεδο της **περιοχής**, ο χώρος διαμερίζεται σε υποχώρους και σε κάθε υποχώρο ανατίθεται δική του τιμή· κάθε άγνωστο σημείο κληρονομεί την τιμή του υποχώρου στον οποίο ανήκει. Στο επίπεδο του **προτύπου**, η διαμέριση γίνεται μέσω αντιπροσωπευτικών σημείων και το άγνωστο σημείο δανείζεται την τιμή του πλησιέστερου αντιπροσώπου. Τέλος, στο επίπεδο του **στιγμιότυπου ελέγχου**, μια ξεχωριστή τιμή k υπολογίζεται για κάθε άγνωστο σημείο χωριστά.

Η διάκριση αυτή δεν είναι τυπολογική λεπτομέρεια αλλά έχει άμεσες πρακτικές συνέπειες. Όσο πιο λεπτομερές είναι το επίπεδο ανάλυσης, τόσο μεγαλύτερη είναι η δυνατότητα προσαρμογής στις τοπικές συνθήκες, αλλά τόσο μικρότερο και το δείγμα πάνω στο οποίο στηρίζεται κάθε εκτίμηση. Στο καθολικό επίπεδο η τιμή k εκτιμάται από χιλιάδες παρατηρήσεις και είναι επομένως στατιστικά σταθερή αλλά αδιάφορη προς την τοπική δομή· στο επίπεδο του στιγμιότυπου η εκτίμηση είναι πλήρως τοπική αλλά στηρίζεται σε ελάχιστη πληροφορία. Η αντιστάθμιση αυτή αποδεικνύεται καθοριστική για την ερμηνεία των πειραματικών αποτελεσμάτων της εργασίας, όπως θα φανεί στην ενότητα 7.3.

Πίνακας 3.1: Τα τέσσερα επίπεδα προσδιορισμού της τιμής k .

Επίπεδο	Πού ορίζεται το k	Πηγή πληροφορίας	Χαρακτηριστικός εκπρόσωπος
Καθολικό	Μία τιμή για όλο το σύνολο	Ολόκληρο το σύνολο εκπαίδευσης	Διασταυρούμενη επικύρωση [23]
Περιοχή	Μία τιμή ανά υποχώρο	Τα σημεία του υποχώρου	G-k-NN [34], shd-k-NN [38]
Πρότυπο	Μία τιμή ανά αντιπρόσωπο	Η γειτονιά του αντιπροσώπου	Garcia-Pedrajas κ.ά. [36]
Στιγμιότυπο ελέγχου	Μία τιμή ανά άγνωστο σημείο	Η τοπική γειτονιά του σημείου	AdaNN [3], GS-k-NN [57]

3.2 Προσεγγίσεις γενετικών αλγορίθμων

Η πρώτη οικογένεια αντιμετωπίζει την επιλογή του k ως πρόβλημα συνδυαστικής βελτιστοποίησης και το επιλύει με εξελικτικές τεχνικές. Πέραν της καθολικής προσέγγισης BBO των Kardan, Kaviani και Esmaeili [33], που παρουσιάστηκε στην ενότητα 2.5, οι Johansson, König και Niklasson [34] πρότειναν μια τοπικά προσαρμοστική εκδοχή, τον αλγόριθμο G-k-NN. Η μέθοδος κατασκευάζει, με γενετικό αλγόριθμο, δέντρα απόφασης που διαμερίζουν τον χώρο αναζήτησης σε διακριτές περιοχές. Σε κάθε

περιοχή ανατίθεται μια σταθερή, βελτιστοποιημένη τιμή k , η οποία στη συνέχεια κληρονομείται από κάθε άγνωστο σημείο ανάλογα με τη θέση του στον χώρο και την περιοχή στην οποία ανήκει.

Η αναζήτηση των k εγγύτερων γειτόνων μπορεί, στη διατύπωση αυτή, να γίνει είτε **τοπικά**, εντός των ορίων της περιοχής, είτε **καθολικά**, σε όλες τις περιοχές, είτε με **μεικτό** τρόπο. Οι τρεις εναλλακτικές συγκρίθηκαν, ως προς την ορθότητα, με τον κλασικό k -NN για $k = 5$ και $k = 11$ σε είκοσι επτά διαφορετικά σύνολα δεδομένων. Ως προς τη στατιστική σημαντικότητα, η καθολική εκδοχή του G - k -NN υπερτερεί σημαντικά και των τριών εκδοχών του συμβατικού k -NN· το ίδιο ωστόσο δεν ισχύει για την τοπική εκδοχή, εύρημα που αποτελεί μια πρώιμη ένδειξη ότι η τοπικότητα δεν αποδίδει αυτομάτως.

3.3 Προσεγγίσεις νευρωνικών δικτύων

Οι Mullick, Datta και Das [35] παρουσίασαν δύο εναλλακτικές παραλλαγές του κλασικού k -NN, τον προσαρμοστικό Ada- k -NN και τον Ada- k -NN2. Και οι δύο αξιοποιούν την πυκνότητα και την κατανομή των δεδομένων προκειμένου να βρουν μια κατάλληλη τιμή k για κάθε άγνωστο σημείο. Ο Ada- k -NN χρησιμοποιεί τεχνητό νευρωνικό δίκτυο για να **μάθει** την κατάλληλη τιμή, ενώ ο Ada- k -NN2 αποτελεί απλοποιημένη εκδοχή που χρησιμοποιεί μια ευρετική ως δείκτη της τοπικής πυκνότητας γύρω από το άγνωστο σημείο. Η ευρετική αυτή βασίζεται στην ιδέα της αναζήτησης εκείνου του k που κατηγοριοποιεί σωστά την πλειονότητα των γειτόνων του σημείου ελέγχου, ιδέα εννοιολογικά συγγενική με εκείνη του AdaNN.

Οι συγγραφείς διεξήγαγαν δύο ομάδες πειραμάτων. Η πρώτη σύγκρινε τους δύο αλγόριθμους με οκτώ άλλους κατηγοριοποιητές σε δεκαεπτά σύνολα δεδομένων, χωρισμένα σε μικρά/μεσαία και μεγάλα. Στα μικρά και μεσαία σύνολα ο Ada- k -NN2 υπερτέρησε όλων ως προς τη μέση ορθότητα και πέτυχε την καλύτερη μέση κατάταξη, ενώ ο Ada- k -NN κατέλαβε την πέμπτη θέση. Στα μεγάλα σύνολα ο Ada- k -NN2 κατέλαβε τη δεύτερη θέση ως προς τη μέση ορθότητα και την πρώτη ως προς τη μέση κατάταξη, ενώ ο Ada- k -NN υποχώρησε στην έκτη και όγδοη θέση αντίστοιχα, γεγονός που υποδεικνύει προβλήματα κλιμάκωσης. Η δεύτερη ομάδα πειραμάτων αφορούσε ανισόρροπες κλάσεις, όπου ο Ada- k -NN2 σε συνδυασμό με ένα απλό σχήμα στάθμισης υπερτέρησε των υπόλοιπων κατηγοριοποιητών.

3.4 Προσεγγίσεις προτύπων και συσταδοποίησης

Η οικογένεια αυτή είναι η πολυπληθέστερη μεταξύ των πρόσφατων δημοσιεύσεων και στηρίζεται στην ιδέα ότι ο χώρος μπορεί να συνοψιστεί σε αντιπροσωπευτικά σημεία ή σε συστάδες, καθεμία από τις οποίες φέρει τη δική της τιμή k .

Οι Garcia-Pedrajas, Romero del Castillo και Cerruela-Garcia [36] διαιρούν τον χώρο αναζήτησης σε υποχώρους, καθέναν από τους οποίους αποτελεί μια γειτονιά που εκπροσωπείται από ένα πρότυπο. Σε κάθε πρότυπο ανατίθεται μια τιμή k , η οποία θεωρείται βέλτιστη εντός των ορίων της γειτονιάς. Η εύρεση των τιμών αυτών γίνεται με άπληστη προσέγγιση: για κάθε πρότυπο δοκιμάζεται η τοπική επίδοση με k να μεταβάλλεται στο διάστημα $[k_{\min}, k_{\max}]$, με συνηθέστερες τιμές $k_{\min} = 1$ και $k_{\max} = 100$, και τελικά ανατίθεται η τιμή με την υψηλότερη επίδοση. Κατά την ταξινόμηση, κάθε άγνωστο σημείο χρησιμοποιεί την τιμή k του εγγύτερου προτύπου του. Δεδομένου ότι η προσέγγιση μπορεί να ενσωματωθεί σχεδόν σε κάθε παραλλαγή του k -NN, οι συγγραφείς πραγματοποίησαν πειράματα με τον τυπικό k -NN, τον προσαρμοστικό k -NN και τον συμμετρικό k -NN, σε ογδόντα σύνολα δεδομένων για κανονικά προβλήματα και εξήντα πέντε για ανισόρροπα, με δεκαπλή διασταυρούμενη επικύρωση. Η προτεινόμενη μέθοδος υπερτέρησε και στις δύο κατηγορίες προβλημάτων.

Οι Bulut και Amasyali [37] παρουσίασαν την προσέγγιση One Nearest Cluster (1NC). Ο χρήστης ορίζει εκ των προτέρων το πλήθος των συστάδων l , ενώ ο αλγόριθμος επιλέγει τα M πλησιέστερα δείγματα

γύρω από το άγνωστο σημείο. Υπολογίζονται όλες οι αποστάσεις ανάμεσα στο σημείο ελέγχου και στα M δείγματα, τα οποία τοποθετούνται σε έναν μονοδιάστατο άξονα από το πλησιέστερο προς το απομακρυσμένο και στη συνέχεια χωρίζονται σε l συστάδες. Η πλειοψηφική ψηφοφορία εφαρμόζεται μόνο εντός της πλησιέστερης συστάδας, οπότε το k προσεγγίζεται ως $k \approx M/l$. Η διαδικασία επαναλαμβάνεται για κάθε σημείο ελέγχου. Σε τριάντα έξι σύνολα δεδομένων ο INC υπερέτρησε εννέα φορές έναντι του 1-NN και του 5-NN, εμφάνισε την υψηλότερη μέση ορθότητα και υπερέτρησε στατιστικά σημαντικά του 1-NN. Οι συγγραφείς υποστηρίζουν ότι τα M και l πρέπει να είναι μικροί ακέραιοι, τόσο για λόγους υπολογιστικού κόστους όσο και επειδή μεγάλες τιμές δίνουν το ίδιο αποτέλεσμα.

Οι Guo και συνεργάτες [11] κατασκευάζουν ένα μοντέλο από το σύνολο εκπαίδευσης, στο οποίο ένα υποσύνολο αντιπροσωπευτικών στιγμιotypών εκπροσωπεί ολόκληρο το σύνολο. Μετά την επιλογή των αντιπροσώπων, η τιμή k επιλέγεται αυτόματα για κάθε αντιπρόσωπο και ισούται με το πλήθος των στιγμιotypών που αυτός καλύπτει· με άλλα λόγια, το πλήθος των γειτόνων καθορίζεται έμμεσα από τη δομή των δεδομένων και όχι από τον χρήστη. Η μέθοδος συγκρίθηκε, ως προς την ορθότητα, με τον C5.0 και τον k -NN σε έξι σύνολα δεδομένων, επιτυγχάνοντας μέση ορθότητα 85,15 % έναντι 81,35 % του C5.0 και 80,90 %, 83,52 % και 83,67 % του k -NN για $k = 1, 3$ και 5 αντίστοιχα.

Η ιδέα της αναπαράστασης του συνόλου εκπαίδευσης με λίγα αντιπροσωπευτικά σημεία αξιοποιείται περαιτέρω από τους Ougiaroglou, Evangelidis και Diamantaras [38] στον αλγόριθμο Subspace Homogeneity based Dynamic k -NN (shd- k -NN). Οι συγγραφείς εφαρμόζουν επαναληπτικά τη διαδικασία συσταδοποίησης k -means μέχρις ότου όλες οι συστάδες γίνουν ομοιογενείς, δηλαδή να περιέχουν σημεία μίας μόνο κλάσης. Η επαναληπτική αυτή διαδικασία παράγει ένα δέντρο συστάδων, στο οποίο κάθε φύλλο αντιστοιχεί σε μια ομοιογενή συστάδα και το βάθος d του φύλλου δίνει πληροφορία για την περιοχή στην οποία βρίσκεται το άγνωστο σημείο. Όταν πρέπει να ταξινομηθεί ένα άγνωστο σημείο, ο αλγόριθμος εντοπίζει το κεντροειδές της πλησιέστερης ομοιογενούς συστάδας και το αντίστοιχο d , και εφαρμόζει μία από πέντε ευρετικές που υπολογίζουν το k ως συνάρτηση του d . Η μέθοδος συγκρίθηκε με τον «βέλτιστο» k -NN, όπως προκύπτει από πενταπλή διασταυρούμενη επικύρωση, καθώς και με τους 1-NN, 5-NN, 10-NN, $\sqrt{N}$ -NN και $\sqrt{(N/2)}$ -NN, σε δεκατέσσερα αρχικά σύνολα δεδομένων και δεκαεννέα παραλλαγές τους με προσθήκη τυχαίου θορύβου. Όλες οι ευρετικές εμφάνισαν ανταγωνιστική και, σε ορισμένες περιπτώσεις, καλύτερη επίδοση από τον «βέλτιστο» k -NN.

3.5 Ευρετικές προσεγγίσεις

Η αξιοποίηση ευρετικών αποτελεί συνηθισμένη τεχνική βελτίωσης του συμβατικού k -NN, και στην κατηγορία αυτή ανήκει και ο αλγόριθμος AdaNN που αποτελεί το αντικείμενο της παρούσας εργασίας.

Οι Ougiaroglou και συνεργάτες [39] εφαρμόζουν σταδιακό υπολογισμό των εγγύτερων γειτόνων σε δομές R-δέντρων, με την αναζήτηση να διακόπτεται μόλις ικανοποιηθούν ορισμένα ευρετικά κριτήρια. Έτσι, κάθε άγνωστο σημείο ταξινομείται με μη σταθερό k . Η σημαντικότερη ιδιότητα της σταδιακής αναζήτησης είναι ότι οι γείτονες ανακαλύπτονται με τη σειρά της απόστασής τους από το σημείο ερωτήματος, γεγονός που επιτρέπει την ανακάλυψη του $(k+1)$ -οστού γείτονα εφόσον έχουν ήδη ανακαλυφθεί οι προηγούμενοι k . Οι τρεις προτεινόμενες ευρετικές πρόωρης διακοπής είναι: (α) η ψηφοφορία διακόπτεται όταν ικανοποιηθεί ένα προκαθορισμένο κατώφλι για την κυρίαρχη κλάση· (β) η ψηφοφορία διακόπτεται όταν οι εναπομείνουσες ψήφοι είναι λιγότερες από τη διαφορά $k -$ (τρέχουσες ψήφοι της κυρίαρχης κλάσης), δηλαδή όταν το αποτέλεσμα δεν μπορεί πλέον να αντιστραφεί· και (γ) η ψηφοφορία διακόπτεται όταν ένα προκαθορισμένο πλήθος **διαδοχικών** γειτόνων ψηφίζει όλο υπέρ της κυρίαρχης κλάσης. Η τρίτη αυτή ευρετική είναι ιδιαίτερα σημαντική για την παρούσα εργασία, καθώς

εισάγει την ιδέα ότι μια απόφαση θεωρείται αξιόπιστη όταν επιβεβαιώνεται από διαδοχικές παρατηρήσεις — ιδέα την οποία ο κανόνας κυλιόμενου παραθύρου του Κεφαλαίου 5 μεταφέρει από τη φάση ελέγχου στη φάση επιλογής της τιμής k . Οι ευρετικές δοκιμάστηκαν σε δύο σύνολα δεδομένων ως προς την ορθότητα και το υπολογιστικό κόστος, με τη δεύτερη να υπερτερεί σε ορθότητα και την πρώτη σε κόστος.

Ο αλγόριθμος **AdaNN** των Sun και Huang [3] ανήκει επίσης στην οικογένεια αυτή και ταξινομείται στο επίπεδο του στιγμιότυπου ελέγχου. Σύμφωνα με αυτόν, σε κάθε στιγμιότυπο του συνόλου εκπαίδευσης ανατίθεται διαφορετική τιμή k , ίση με το ελάχιστο πλήθος γειτόνων που πρέπει να συμπεριληφθούν στην πλειοψηφική ψηφοφορία ώστε το στιγμιότυπο να κατηγοριοποιηθεί σωστά. Στη συνέχεια, για κάθε άγνωστο σημείο, ο αλγόριθμος εντοπίζει τον εγγύτερο γείτονά του και την τιμή k που του έχει ανατεθεί, και εφαρμόζει τον k -NN με αυτήν την τιμή. Η μέθοδος συγκρίθηκε, ως προς την ορθότητα, με τον συμβατικό k -NN για $k \in [1, 9]$ και με τον Ada- k -NN [35] σε δεκαπέντε σύνολα δεδομένων· κατατάχθηκε δεύτερη σε έξι σύνολα και τρίτη σε ένα, με τον Ada- k -NN να υπερτερεί συνολικά. Ο αλγόριθμος αναλύεται εξαντλητικά στο Κεφάλαιο 4.

Τέλος, οι Baoli, Qin και Shiwen [42] ανέπτυξαν μια παραλλαγή προσανατολισμένη στην κατηγοριοποίηση κειμένου. Ο προτεινόμενος αλγόριθμος δεν εισάγει νέα διαδικασία επιλογής, αλλά κατασκευάζει μια παραλλαγή του k -NN λιγότερο εξαρτημένη από την τιμή του k : το k θα πρέπει να είναι ανάλογο του πλήθους των στιγμιότυπων εκπαίδευσης που ανήκουν στην κατηγορία στην οποία πρόκειται να ταξινομηθεί το σημείο ελέγχου. Έτσι, για την ταξινόμηση σε μια κατηγορία με πολλά παραδείγματα χρησιμοποιείται μεγαλύτερο k σε σύγκριση με μια κατηγορία με λίγα. Τα πειράματα σε δύο σύνολα δεδομένων κειμένου έδειξαν ότι η παραλλαγή είναι λιγότερο ευαίσθητη στην επιλογή του k και, ως συνέπεια, αντιμετωπίζει αποτελεσματικά την ανισορροπία κλάσεων.

3.6 Πιθανοτικές προσεγγίσεις

Οι πιθανοτικές μέθοδοι αποτελούν την πολυπληθέστερη οικογένεια στο σύνολο της βιβλιογραφίας, είναι όμως και η παλαιότερη: η διάμεση χρονολογία δημοσίευσης των εργασιών της είναι το 2007 [32].

Οι Johansson, Boström και König [43] εισήγαγαν την έννοια των «σφαιρών εμπιστοσύνης», αξιοποιώντας τον εκτιμητή Laplace $P_{\text{ClassA}} = (k+1)/(N+C)$, όπου k είναι το πλήθος των στιγμιότυπων εκπαίδευσης που ανήκουν στην κλάση A , N το συνολικό πλήθος των στιγμιότυπων εντός της σφαίρας και C το πλήθος των κλάσεων. Κάθε στιγμιότυπο εκπαίδευσης περιβάλλεται από μια σφαίρα εμπιστοσύνης που περιλαμβάνει άλλα στιγμιότυπα. Για την κατασκευή της, ο αλγόριθμος εξετάζει τους εγγύτερους γείτονες έναν προς έναν, ξεκινώντας από τον πλησιέστερο, και διακόπτει είτε όταν η τιμή του εκτιμητή αρχίσει να μειώνεται (κατασκευή «πρόθυμου» τύπου) είτε όταν εξεταστούν όλα τα στιγμιότυπα, οπότε επιλέγεται η σφαίρα με την υψηλότερη τιμή του εκτιμητή («ολική» κατασκευή). Η μέθοδος συγκρίθηκε ως προς την ορθότητα και το auROC με τον συμβατικό k -NN για σταθερές τιμές $k = 5, 11$ και 17 σε δεκαοκτώ σύνολα δεδομένων, υπερτερώντας σημαντικά σε δεκατρία από αυτά.

Οι Bhattacharya, Ghosh και Chowdhury [44], [45] αξιοποιούν την τοπική κατανομή των δεδομένων κατασκευάζοντας μια υπερσφαίρα γύρω από κάθε άγνωστο σημείο, ώστε να αποτυπωθεί η κατανομή των στιγμιότυπων εκπαίδευσης γύρω του. Επιπλέον, εισάγουν την έννοια του βάρους κομβικότητας (hubness weight), δηλαδή της πιθανότητας ένα σημείο να ανήκει στη γειτονιά ενός συγκεκριμένου σημείου. Ο συνδυασμός των δύο εννοιών οδηγεί στην ανάθεση μοναδικής τιμής k σε κάθε άγνωστο σημείο. Η μέθοδος συγκρίθηκε με τον συμβατικό k -NN για σταθερές τιμές $k = 1, 3, 5, 7$ και $\lceil \sqrt{N} \rceil$, καθώς και με άλλες παραλλαγές [43], [58], σε δεκαπέντε σύνολα δεδομένων, υπερτερώντας στις περισσότερες περιπτώσεις.

Οι Zhang και Li [46] υποστηρίζουν ότι το μέγεθος της γειτονιάς πρέπει να είναι ευέλικτο προκειμένου να αντιμετωπιστεί το πρόβλημα της ανισόρροπης κατηγοριοποίησης. Έτσι, για δεδομένο k , προτείνουν τη διεύρυνση της εξεταζόμενης γειτονιάς έως ότου συμπεριληφθούν $k/2$ στιγμιότυπα της σπάνιας κλάσης. Η προσέγγιση συγκρίθηκε, ως προς το $auROC$ και με ανάλυση κυρτού περιβλήματος, με παραλλαγές του k -NN (ENN και CCW- k -NN) και με καθιερωμένες τεχνικές αντιμετώπισης της ανισορροπίας (SMOTE, MetaCost) σε δώδεκα σύνολα δεδομένων, εμφανίζοντας ανταγωνιστική επίδοση.

Οι Song και συνεργάτες [47] υποστηρίζουν ότι η παράμετρος k δεν επαρκεί από μόνη της για να καθορίσει το μέγεθος της γειτονιάς και εισάγουν μια νέα παράμετρο I , η οποία μετρά το πλήθος των **πληροφοριακών** στιγμιότυπων εντός της επιλεγμένης γειτονιάς. Ένα σημείο εκπαίδευσης θεωρείται πληροφοριακό αν βρίσκεται κοντά στο σημείο ελέγχου και ταυτόχρονα μακριά από στιγμιότυπα άλλων κλάσεων. Η επιλογή των k και I γίνεται με διασταυρούμενη επικύρωση. Η μέθοδος αποδείχθηκε λιγότερο ευαίσθητη στην επιλογή του k σε σύγκριση με τον συμβατικό k -NN, εμφανίζοντας ταυτόχρονα ανταγωνιστική επίδοση έναντι ευρέως χρησιμοποιούμενων κατηγοριοποιητών, όπως οι μηχανές διανυσμάτων υποστήριξης.

Ο Ghosh [49] επιλέγει μοναδική τιμή k για κάθε άγνωστο σημείο με βάση την κατανομή των κλάσεων γύρω του, εμφανίζοντας αξιοσημείωτη επίδοση ως προς το τυπικό σφάλμα, ενώ σε προγενέστερη εργασία του [48] είχε προτείνει μπεϋζιανή προσέγγιση για τη βελτιστοποίηση της επιλογής του k με βάση την κατανομή των δεδομένων, η οποία υπερέτερσε των συνήθων τεχνικών διασταυρούμενης επικύρωσης. Οι Wang, Neskonvic και Cooper [10] αξιοποιούν την τοπική στατιστική εμπιστοσύνη: αντί για σταθερό k , η τιμή ρυθμίζεται με βάση την τοπική κατανομή των δεδομένων μέχρις ότου επιτευχθεί ένα προκαθορισμένο επίπεδο εμπιστοσύνης. Ο αλγόριθμος δοκιμάστηκε σε πέντε σύνολα δεδομένων, με τον συμβατικό k -NN να εμφανίζει ελαφρώς χαμηλότερα ποσοστά σφάλματος. Τέλος, οι Dhurandhar και Dobra [51] διατύπωσαν εκφράσεις που εξηγούν στατιστικά τη συμπεριφορά του k -NN και προτείνουν την κατασκευή πολλαπλών κατανομών για τα άγνωστα σημεία, ώστε να επιλέγεται η τιμή k που μειώνει τον ρυθμό σφάλματος στο μεγαλύτερο δυνατό πλήθος κατανομών.

3.7 Λοιπές προσεγγίσεις

Στην τελευταία κατηγορία εντάσσονται μέθοδοι που δεν ταιριάζουν στις προηγούμενες οικογένειες. Οι Zhong και συνεργάτες [55] παρουσίασαν μια επαναληπτική διαδικασία που παράγει ένα διάστημα $[k_{min}, k_{max}]$ εντός του οποίου αναζητείται δυναμικά η τιμή του k . Στη συνέχεια, ο αλγόριθμος κατασκευάζει καμπύλες τάσης μεταβολής, με βάση την αναλογία των ορθά κατηγοριοποιημένων στιγμιότυπων για κάθε k εντός του διαστήματος, και η τελική τιμή επιλέγεται με βάση τρία κριτήρια που εξαρτώνται από το σχήμα των καμπυλών. Η μέθοδος υπερέτερσε του συμβατικού k -NN με σταθερές τιμές $k = 1, 5, 7$ και $\sqrt{N}$ ως προς την ανάκληση και το F-score.

Οι Anava και Levy [56] υιοθετούν σχήμα στάθμισης και επιχειρούν να ελαχιστοποιήσουν την απόσταση ανάμεσα στην παραγόμενη πρόβλεψη και την πραγματική τιμή, βελτιστοποιώντας ταυτόχρονα το πλήθος των επιλεγμένων εγγύτερων γειτόνων για κάθε άγνωστο σημείο. Η προσέγγιση συγκρίθηκε, ως προς το τυπικό σφάλμα, με τον συμβατικό k -NN και τον εκτιμητή Nadaraya-Watson σε οκτώ σύνολα δεδομένων, υπερτερώντας σε επτά από αυτά, με τη διαφορά να είναι στατιστικά σημαντική σε τρία.

Οι Zhang και συνεργάτες [57] πρότειναν το πλαίσιο αναλυτικής βελτιστοποίησης Graph Sparse k -NN (GS- k -NN), με στόχο την εκμάθηση διαφορετικής, βέλτιστης τιμής k για κάθε άγνωστο σημείο, αξιοποιώντας την ήδη γνωστή κατανομή των δεδομένων. Η μέθοδος συγκρίθηκε, ως προς την ορθότητα και το ριζικό μέσο τετραγωνικό σφάλμα, με τον συμβατικό k -NN για $k = 5$ και με την παραλλαγή L- k -

NN σε δώδεκα σύνολα δεδομένων, σε προβλήματα κατηγοριοποίησης, παλινδρόμησης και συμπλήρωσης ελλειπουσών τιμών, εμφανίζοντας υψηλότερη ορθότητα και χαμηλότερο σφάλμα. Στην ίδια κατεύθυνση κινούνται και οι εργασίες των Zhang και συνεργατών [8], [9], οι οποίες διατυπώνουν την επιλογή του k ως πρόβλημα βελτιστοποίησης με αραιή αναπαράσταση και προτείνουν τη δομή k *Tree, που αποθηκεύει στα φύλλα ενός δέντρου απόφασης τόσο τη βέλτιστη τιμή k όσο και το υποσύνολο των σημείων εκπαίδευσης που απαιτούνται, επιτυγχάνοντας ταυτόχρονα τοπική προσαρμογή και επιτάχυνση της φάσης ελέγχου.

3.8 Συγκριτική επισκόπηση

Ο Πίνακας 3.2 συνοψίζει τις κυριότερες μεθόδους δυναμικού προσδιορισμού της τιμής k , με τη μεθοδολογική τους οικογένεια, το επίπεδο στο οποίο γίνεται η επιλογή και το πλήθος των συνόλων δεδομένων που χρησιμοποιήθηκαν στην αξιολόγησή τους. Ο πίνακας βασίζεται στον συγκεντρωτικό πίνακα της επισκόπησης [32], περιορισμένο στις μεθόδους δυναμικού προσδιορισμού.

Πίνακας 3.2: Συγκεντρωτική παρουσίαση των μεθόδων δυναμικού προσδιορισμού της τιμής k .

Εργασία	Έτος	Οικογένεια	Επίπεδο επιλογής	Σύνολο α
Guo κ.ά. [11]	2003	Πρότυπα & συσταδοποίηση	Πρότυπο	6
Baoli κ.ά. [42]	2004	Ευρετική	Στιγμιότυπο ελέγχου	2
Ghosh [48]	2006	Πιθανοτική	Περιοχή	11
Ghosh [49]	2007	Πιθανοτική	Περιοχή	14
Wang κ.ά. [10]	2007	Πιθανοτική	Περιοχή	5
Song κ.ά. [47]	2007	Πιθανοτική	Στιγμιότυπο ελέγχου	12
Ougiaroglou κ.ά. [39]	2007	Ευρετική	Στιγμιότυπο ελέγχου	2
Johansson κ.ά. [43]	2008	Πιθανοτική	Στιγμιότυπο ελέγχου	18
Johansson κ.ά. [34]	2008	Γενετικός αλγόριθμος	Περιοχή	27
Sun & Huang [3] (AdaNN)	2010	Ευρετική	Στιγμιότυπο ελέγχου	15
Zhang & Li [46]	2013	Πιθανοτική	Στιγμιότυπο ελέγχου	12
Bhattacharya κ.ά. [44]	2014	Πιθανοτική	Στιγμιότυπο ελέγχου	15
Zhang κ.ά. [57]	2014	Άλλη	Στιγμιότυπο ελέγχου	12
Bhattacharya κ.ά. [45]	2015	Πιθανοτική	Στιγμιότυπο ελέγχου	15
Anava & Levy [56]	2016	Άλλη	Στιγμιότυπο ελέγχου	8
Garcia-Pedrajas κ.ά. [36]	2017	Πρότυπα & συσταδοποίηση	Πρότυπο	145
Bulut & Amasyali [37]	2017	Πρότυπα & συσταδοποίηση	Πρότυπο	36
Zhong κ.ά. [55]	2017	Άλλη	Περιοχή	1
Mullick κ.ά. [35]	2018	Νευρωνικά δίκτυα	Περιοχή	17
Ougiaroglou κ.ά. [38]	2020	Πρότυπα & συσταδοποίηση	Πρότυπο	33

Από τον πίνακα προκύπτουν ορισμένες παρατηρήσεις που καταγράφει και η επισκόπηση [32]. Πρώτον, οι πιθανοτικές προσεγγίσεις κυριαρχούν αριθμητικά αλλά είναι κατά κανόνα οι παλαιότερες, ενώ οι πιο πρόσφατες εργασίες στρέφονται προς τα πρότυπα και τη συσταδοποίηση. Δεύτερον, το διάμεσο πλήθος συνόλων δεδομένων που χρησιμοποιούνται για τον έλεγχο των μεθόδων είναι μόλις δώδεκα, μέγεθος που θεωρείται ανεπαρκές για αξιόπιστη αξιολόγηση, ιδίως όταν δεν συνοδεύεται από ελέγχους στατιστικής σημαντικότητας: μόνο σε δώδεκα από τις είκοσι οκτώ εργασίες διεξήχθη τουλάχιστον ένας στατιστικός έλεγχος. Τρίτον, οι περισσότερες μέθοδοι συγκρίνονται με τον k -NN για μερικές σταθερές τιμές k και όχι με τη «βέλτιστη» τιμή που θα προέκυπτε από διασταυρούμενη επικύρωση, γεγονός που καθιστά τη σύγκριση ευνοϊκή για τις προσαρμοστικές μεθόδους.

3.9 Τεχνικές μείωσης δεδομένων και θόρυβος στις ετικέτες

Παράλληλα με την επιλογή του k , σημαντικό τμήμα της βιβλιογραφίας ασχολείται με τη μείωση του μεγέθους του συνόλου εκπαίδευσης, τόσο για λόγους αποδοτικότητας όσο και για την απομάκρυνση θορυβωδών σημείων. Ο κανόνας CNN του Hart [15] ήταν η πρώτη τέτοια προσπάθεια, επιλέγοντας ένα υποσύνολο ικανό να ταξινομεί σωστά όλα τα υπόλοιπα σημεία. Ο κανόνας ENN του Wilson [16] ακολούθησε αντίστροφη λογική, αφαιρώντας τα σημεία που δεν κατηγοριοποιούνται σωστά από τους γείτονές τους. Η διάκριση αυτή είναι ουσιώδης: ο CNN αποσκοπεί στη συμπύκνωση, ο ENN στον καθαρισμό, ενώ ο Tomek [17] πρότεινε τροποποιήσεις που συνδυάζουν τα δύο. Μεταγενέστερες εργασίες, όπως η οικογένεια DROP των Wilson και Martinez [18] και ο ICF των Brighton και Mellish [19], εκλέπτουν τις ιδέες αυτές, ενώ οι García και συνεργάτες [20] παρείχαν συστηματική ταξινόμηση και εκτενή πειραματική σύγκριση του πεδίου. Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι μέθοδοι μείωσης βασισμένες σε ομοιογενή συσταδοποίηση, όπως ο RHC των Ougiaroglou και Evangelidis [12], [13] και οι ιεραρχικές επεκτάσεις του [14], οι οποίες αναγνωρίζουν ακριβώς τις περιοχές όπου η τοπική πυκνότητα και η καθαρότητα των κλάσεων διαφέρουν.

Η δεύτερη γραμμή έρευνας που συνδέεται άμεσα με την παρούσα εργασία αφορά τον θόρυβο στις ετικέτες. Ως θόρυβος ετικετών (label noise ή class noise) ορίζεται η κατάσταση κατά την οποία ένα μέρος των παραδειγμάτων εκπαίδευσης φέρει λανθασμένη ετικέτα κλάσης, φαινόμενο εξαιρετικά συνηθισμένο σε πραγματικά δεδομένα. Η επισκόπηση των Frénay και Verleysen [21] αποτελεί τη συστηματικότερη καταγραφή του πεδίου: οι συγγραφείς διακρίνουν τρία μοντέλα θορύβου, ανάλογα με το αν η πιθανότητα αλλοίωσης μιας ετικέτας είναι ανεξάρτητη από τα δεδομένα, εξαρτάται μόνο από την πραγματική κλάση, ή εξαρτάται και από τα χαρακτηριστικά του σημείου. Το απλούστερο από αυτά, ο ομοιόμορφος θόρυβος ανεξάρτητος από τα δεδομένα, είναι και το πλέον διαδεδομένο στην πειραματική βιβλιογραφία, καθώς επιτρέπει τον ακριβή έλεγχο του επιπέδου αλλοίωσης. Οι Zhu και Wu [22] έδειξαν πειραματικά ότι ο θόρυβος ετικετών υποβαθμίζει την επίδοση σημαντικά περισσότερο από τον θόρυβο χαρακτηριστικών ισοδύναμης έντασης.

Για τον k -NN ειδικότερα, ο βασικός μηχανισμός άμυνας έναντι του θορύβου είναι η **αύξηση του k** , καθώς ένα εσφαλμένα επισημειωμένο σημείο υπερψηφίζεται από τους ορθούς γείτονές του. Η εναλλακτική οδός είναι ο καθαρισμός του συνόλου εκπαίδευσης πριν την ταξινόμηση, με τεχνικές τύπου ENN [16]. Οι δύο προσεγγίσεις δεν είναι ισοδύναμες: ο καθαρισμός αφαιρεί οριστικά σημεία και ενέχει τον κίνδυνο απώλειας γνήσιων αλλά ασυνήθιστων παραδειγμάτων, ενώ η αύξηση του k διατηρεί όλη την πληροφορία και μεταθέτει το βάρος στη διαδικασία ψηφοφορίας. Η παρούσα εργασία κινείται στη δεύτερη κατεύθυνση, εξετάζοντας αν οι μηχανισμοί επιλογής του k μπορούν να ανακαλύψουν **αυτόματα** ότι απαιτούνται μεγαλύτερες τιμές, χωρίς να τους δοθεί καμία πληροφορία για την παρουσία θορύβου.

3.10 Το ερευνητικό κενό

Από την ανασκόπηση των Κεφαλαίων 2 και 3 προκύπτουν τέσσερις παρατηρήσεις που οριοθετούν τη συνεισφορά της εργασίας. **Πρώτον**, το άνω όριο 9 στην αρχική διατύπωση του AdaNN [3] δεν συνοδεύεται από θεωρητική ή εμπειρική δικαιολόγηση, ούτε εξετάζεται η ευαισθησία της μεθόδου ως προς αυτό. **Δεύτερον**, ο κανόνας επιλογής «η πρώτη τιμή k που πετυχαίνει» υιοθετείται χωρίς να συγκριθεί με εναλλακτικούς κανόνες, παρότι είναι εξ ορισμού ευαίσθητος σε μεμονωμένες τυχαίες επιτυχίες: η ιδέα των διαδοχικών επιβεβαιώσεων, που εμφανίζεται ως ευρετική πρόωρης διακοπής στη [39], δεν έχει μεταφερθεί στη φάση επιλογής της τιμής k .

Τρίτον, η αξιολόγηση του AdaNN στην αρχική δημοσίευση γίνεται αποκλειστικά σε δεδομένα χωρίς τεχνητά εισαγόμενο θόρυβο, με αποτέλεσμα να μην ελέγχεται η συμπεριφορά του ακριβώς στις συνθήκες όπου η επιλογή μεγαλύτερων τιμών k έχει τη μεγαλύτερη σημασία. **Τέταρτον**, όπως επισημαίνει και η επισκόπηση [32], οι προσαρμοστικές μέθοδοι συγκρίνονται συνήθως με τον k -NN για λίγες σταθερές τιμές k και όχι με τον καθολικό προσδιορισμό μέσω διασταυρούμενης επικύρωσης, ο οποίος αποτελεί το πραγματικό σημείο αναφοράς της πράξης.

Η παρούσα εργασία απαντά και στις τέσσερις παρατηρήσεις. Επεκτείνει το άνω όριο και μετρά την επίπτωση, εισάγει και αξιολογεί έναν εναλλακτικό κανόνα επιλογής βασισμένο σε κυλιόμενο παράθυρο διαδοχικών επιτυχιών, διεξάγει ολόκληρη την πειραματική αξιολόγηση παράλληλα σε καθαρά και σε θορυβώδη δεδομένα, και χρησιμοποιεί ως σημείο αναφοράς τον καθολικό προσδιορισμό του k με πενταπλή στρωματοποιημένη διασταυρούμενη επικύρωση. Το επόμενο κεφάλαιο διατυπώνει με ακρίβεια τον βασικό αλγόριθμο.

Κεφάλαιο 4ο: Ο αλγόριθμος AdaNN

Το κεφάλαιο αυτό αναλύει διεξοδικά τον αλγόριθμο AdaNN (Adaptive k-Nearest Neighbor) των Sun και Huang [3], ο οποίος αποτελεί τη βάση της εργασίας. Παρουσιάζονται η κεντρική ιδέα και η υπόθεση εργασίας στην οποία στηρίζεται, οι δύο φάσεις λειτουργίας του με τον αντίστοιχο ψευδοκώδικα, ένα αναλυτικό αριθμητικό παράδειγμα που καθιστά τον μηχανισμό απολύτως συγκεκριμένο, οι ιδιότητες και οι οριακές περιπτώσεις του, η υπολογιστική του πολυπλοκότητα και, τέλος, οι αδυναμίες που κινητοποιήσαν τον σχεδιασμό των παραλλαγών του Κεφαλαίου 5.

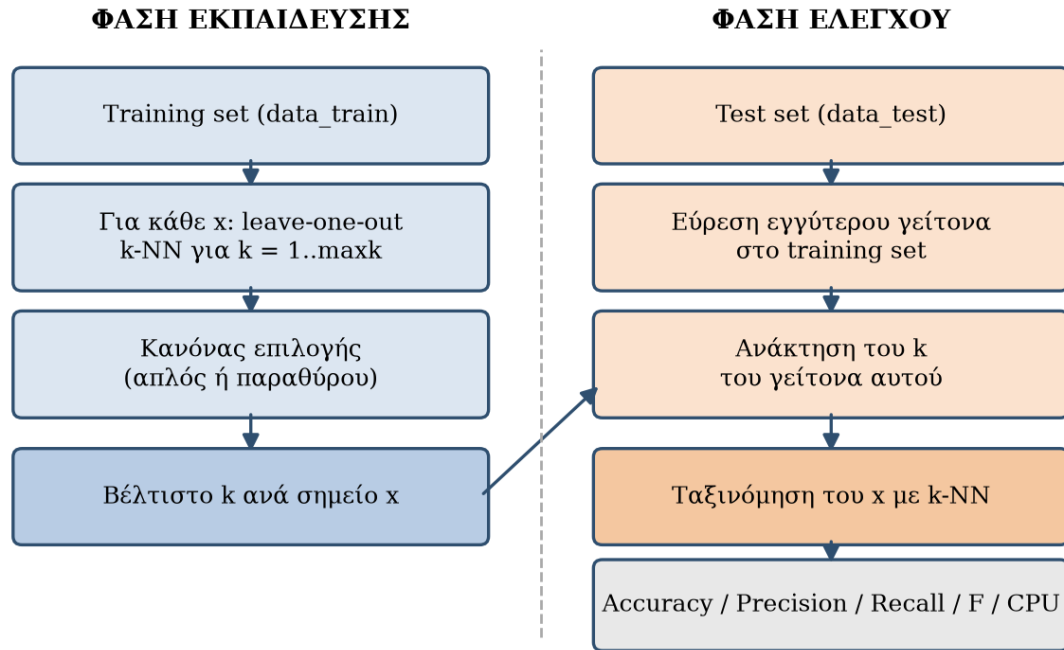
4.1 Η κεντρική ιδέα και η υπόθεση εργασίας

Ο AdaNN ξεκινά από την παρατήρηση που διατυπώθηκε στην ενότητα 2.6: μια καθολική τιμή k είναι αναγκαστικά συμβιβαστική, διότι ο χώρος των δεδομένων δεν είναι ομοιογενής. Αντί να αναζητήσει την καλύτερη μέση τιμή, ο αλγόριθμος αναθέτει σε **κάθε σημείο του συνόλου εκπαίδευσης** μια δική του τιμή k και, κατά την ταξινόμηση, μεταφέρει την τιμή αυτή στα άγνωστα σημεία μέσω του εγγύτερου γείτονα.

Η υπόθεση εργασίας, όπως τη διατυπώνουν οι Sun και Huang [3], είναι διπλή. Πρώτον, ότι η τιμή k που κατηγοριοποιεί σωστά ένα σημείο εκπαίδευσης όταν αυτό αντιμετωπιστεί ως άγνωστο αποτελεί καλή ένδειξη για το πόσο ευρεία πρέπει να είναι η γειτονιά στην περιοχή αυτή. Δεύτερον, ότι ένα άγνωστο σημείο μοιράζεται τις τοπικές συνθήκες του εγγύτερου γείτονα του στο σύνολο εκπαίδευσης, και επομένως η τιμή k που λειτούργησε για εκείνον έχει υψηλή πιθανότητα να λειτουργήσει και για αυτό.

Η υπόθεση αυτή είναι ισχυρότερη από όσο φαίνεται με την πρώτη ματιά, καθώς η επίδοση του συστήματος εξαρτάται πλέον από την ποιότητα δύο **διαδοχικών** εκτιμήσεων: ότι η τιμή k του γείτονα είναι όντως κατάλληλη για την περιοχή, και ότι το άγνωστο σημείο ανήκει πράγματι στην ίδια τοπική περιοχή με τον γείτονα του. Αν κάποια από τις δύο αποτύχει, το σφάλμα δεν περιορίζεται σε ένα σημείο αλλά μεταδίδεται σε ολόκληρη την περιοχή για την οποία ο συγκεκριμένος γείτονας αποτελεί πλησιέστερο σημείο. Το γεγονός αυτό αποδεικνύεται καθοριστικό για την ερμηνεία των αποτελεσμάτων υπό θόρυβο (ενότητα 7.3).

Το σημαντικότερο πλεονέκτημα του AdaNN έναντι των υπόλοιπων προσαρμοστικών μεθόδων του Κεφαλαίου 3 είναι η **οικονομία του μηχανισμού**. Δεν εισάγει καμία νέα υπερπαράμετρο πέραν του άνω ορίου αναζήτησης $\max k$, δεν απαιτεί κατασκευή βοηθητικού μοντέλου, δεν χρειάζεται συσταδοποίηση ή εκπαίδευση νευρωνικού δικτύου και δεν μεταβάλλει το μέτρο απόστασης. Το μόνο που προσθέτει είναι ένα προεπεξεργαστικό βήμα *leave-one-out* πάνω στο σύνολο εκπαίδευσης και ένας κανόνας μεταφοράς. Ακριβώς αυτή η απλότητα τον καθιστά κατάλληλο υπόβαθρο για τη μελέτη παραλλαγών: κάθε μεταβολή στον κανόνα επιλογής μπορεί να απομονωθεί και να αξιολογηθεί καθαρά, χωρίς να συγχέεται με την επίδραση άλλων μηχανισμών.



Σχήμα 4.1: Οι δύο φάσεις του AdaNN. Η φάση ελέγχου είναι κοινή σε όλες τις παραλλαγές· διαφοροποιείται μόνο ο κανόνας επιλογής στη φάση εκπαίδευσης.

4.2 Η φάση εκπαίδευσης

Η φάση εκπαίδευσης επεξεργάζεται αποκλειστικά το σύνολο εκπαίδευσης και παράγει, για κάθε σημείο του, μια τιμή k . Για κάθε σημείο x του συνόλου εκπαίδευσης εφαρμόζεται η μέθοδος **leave-one-out**: το σημείο εξαιρείται προσωρινά από το σύνολο και αντιμετωπίζεται ως άγνωστο, ταξινομούμενο από τα υπόλοιπα σημεία. Η εξαίρεση αυτή είναι απαραίτητη, διότι διαφορετικά κάθε σημείο θα ήταν πάντοτε ο πλησιέστερος γείτονας του εαυτού του και θα ταξινομούσαν σωστά για $k = 1$, καθιστώντας ολόκληρη τη διαδικασία άσκοπη.

Στη συνέχεια δοκιμάζονται διαδοχικά οι τιμές $k = 1, 2, \dots, \max k$ και για κάθε μία καταγράφεται αν το x κατηγοριοποιείται σωστά. Το αποτέλεσμα της διαδικασίας είναι, για κάθε σημείο, ένα διάνυσμα $\max k$ δυαδικών τιμών, δηλαδή συνολικά ένας πίνακας ορθότητας διαστάσεων $n \times \max k$. Ο πίνακας αυτός αποτελεί τη μοναδική είσοδο του κανόνα επιλογής και είναι ο λόγος για τον οποίο όλες οι παραλλαγές του Κεφαλαίου 5 μπορούν να συγκριθούν καθαρά μεταξύ τους: υπολογίζεται μία φορά και τροφοδοτεί όλους τους κανόνες.

Ο κανόνας επιλογής του αρχικού AdaNN ορίζει ως βέλτιστο k το **μικρότερο** πλήθος γειτόνων που αρκεί για τη σωστή κατηγοριοποίηση του σημείου. Αν καμία τιμή στο διάστημα $[1, \max k]$ δεν το επιτυγχάνει, ανατίθεται η εφεδρική τιμή $\max k$. Στην αρχική δημοσίευση το $\max k$ ορίζεται στο 9. Η προτίμηση προς τη μικρότερη τιμή έχει διαισθητική βάση, καθώς μια μικρή τιμή υποδηλώνει ότι το σημείο βρίσκεται σε καθαρή περιοχή όπου η απόφαση λαμβάνεται εύκολα· ταυτόχρονα όμως καθιστά τον κανόνα ευαίσθητο, αφού μία και μόνη επιτυχία σε μικρό k παγιώνει την επιλογή.

Αλγόριθμος 4.1: Φάση εκπαίδευσης του AdaNN.

Είσοδος : σύνολο εκπαίδευσης $TR = \{(x_1, y_1), \dots, (x_n, y_n)\}$, άνω όριο $\max k$
 Έξοδος : πίνακας $k_opt[1..n]$ με μία τιμή k ανά σημείο εκπαίδευσης

1 για κάθε $x_i \in TR$:

```

2      N ← οι maxk εγγύτεροι γείτονες του x_i στο TR \ {x_i}
3      votes ← μηδενικό διάνυσμα πλήθους κλάσεων
4      για k = 1 έως maxk:
5          votes[ y(N[k]) ] ← votes[ y(N[k]) ] + 1
6          correct[i][k] ← ( argmax(votes) = y_i )
7      # κανόνας επιλογής: πρώτη επιτυχία
8      αν υπάρχει k με correct[i][k] = αληθές:
9          k_opt[i] ← min { k : correct[i][k] = αληθές }
10     αλλιώς:
11         k_opt[i] ← maxk          # εφεδρική τιμή
12     επίστρεψε k_opt

```

Δύο σημεία του Αλγορίθμου 4.1 αξίζουν προσοχής. Το πρώτο είναι η γραμμή 2: η αναζήτηση των γειτόνων γίνεται **μία φορά** για τη μέγιστη τιμή maxk και όχι ξεχωριστά για κάθε k. Το δεύτερο είναι η γραμμή 5: οι ψήφοι ανά κλάση συσσωρεύονται προοδευτικά, προσθέτοντας έναν γείτονα τη φορά, ώστε η μετάβαση από την τιμή k στην τιμή k+1 να κοστίζει μία μόνο πράξη. Ο συνδυασμός των δύο καθιστά τη σάρωση ολόκληρου του διαστήματος [1, maxk] ουσιαστικά δωρεάν, όπως αναλύεται στην ενότητα 4.6.

4.3 Η φάση ελέγχου

Η φάση ελέγχου χρησιμοποιεί τις τιμές k της φάσης εκπαίδευσης για να ταξινομήσει τα άγνωστα σημεία. Κάθε άγνωστο σημείο εντοπίζει τον εγγύτερο γείτονα του στο σύνολο εκπαίδευσης, ανακτά την τιμή k που έχει ανατεθεί σε αυτόν, και ταξινομείται με πλειοψηφική ψήφο ανάμεσα στους αντίστοιχους k εγγύτερους γείτονές του. Το κρίσιμο σημείο, και η βασική μεθοδολογική επιλογή της εργασίας, είναι ότι η φάση αυτή είναι **πανομοιότυπη σε όλες τις παραλλαγές**. Επομένως, κάθε διαφορά που παρατηρείται στα αποτελέσματα οφείλεται αποκλειστικά στον κανόνα με τον οποίο ανατέθηκαν οι τιμές k κατά την πρώτη φάση.

Αλγόριθμος 4.2: Φάση ελέγχου του AdaNN.

Είσοδος : TR, k_opt[1..n], σύνολο ελέγχου TS
 Έξοδος : προβλέψεις $\hat{y}$ για κάθε σημείο του TS

```

1  για κάθε x ∈ TS:
2      nn1 ← ο εγγύτερος γείτονας του x στο TR
3      k ← k_opt[ nn1 ]
4      N ← οι k εγγύτεροι γείτονες του x στο TR
5       $\hat{y}(x)$  ← η κλάση με τις περισσότερες ψήφους στο N
6      (ισοπαλία: η κλάση με τη μικρότερη ετικέτα)
7  επίστρεψε  $\hat{y}$ 

```

Σημειώνεται ότι ο εγγύτερος γείτονας nn₁ της γραμμής 2 συμπεριλαμβάνεται πάντοτε στη γειτονιά N της γραμμής 4, καθώς αποτελεί τον πρώτο από τους k εγγύτερους γείτονες. Το σημείο που δίνει την τιμή k συμμετέχει, δηλαδή, και στην ψηφοφορία.

4.4 Αναλυτικό παράδειγμα

Για να γίνει απολύτως συγκεκριμένος ο μηχανισμός, εξετάζεται ένα παράδειγμα με πέντε σημεία εκπαίδευσης. Ο Πίνακας 4.1 δίνει τον πίνακα ορθότητας που παράγει η φάση εκπαίδευσης για τιμές k από 1 έως 12: το σύμβολο $\checkmark$ δηλώνει ότι το σημείο κατηγοριοποιείται σωστά με τη συγκεκριμένη τιμή k, ενώ το $\times$ ότι κατηγοριοποιείται εσφαλμένα. Τα πέντε σημεία έχουν επιλεγεί ώστε να καλύπτουν τις χαρακτηριστικές περιπτώσεις που συναντώνται στην πράξη.

Πίνακας 4.1: Πίνακας ορθότητας πέντε σημείων εκπαίδευσης για $k = 1..12$.

Σημείο	1	2	3	4	5	6	7	8	9	10	11	12
x_1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
x_2	×	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓
x_3	×	×	×	×	×	✓	✓	✓	✓	✓	✓	✓
x_4	✓	×	×	×	×	×	×	×	×	×	×	×
x_5	×	×	×	×	×	×	×	×	×	×	×	×

Το σημείο x_1 βρίσκεται σε καθαρή περιοχή και κατηγοριοποιείται σωστά για κάθε τιμή k . Το x_2 παρουσιάζει μια **μεμονωμένη, οριακή** επιτυχία στο $k = 2$, η οποία αμέσως ανατρέπεται στο $k = 3$ και σταθεροποιείται μόνο από το $k = 4$ και έπειτα· πρόκειται για την περίπτωση που καθιστά τον κανόνα της πρώτης επιτυχίας εύθραυστο. Το x_3 βρίσκεται σε δύσκολη περιοχή και απαιτεί έξι γείτονες προτού η απόφαση γίνει σωστή. Το x_4 κατηγοριοποιείται σωστά μόνο για $k = 1$, συμπεριφορά τυπική ενός απομονωμένου σημείου. Τέλος, το x_5 δεν κατηγοριοποιείται σωστά για καμία τιμή k , γεγονός που υποδηλώνει είτε ακραίο σημείο είτε εσφαλμένη ετικέτα.

Ο Πίνακας 4.2 δίνει τις τιμές που αναθέτει ο κανόνας της πρώτης επιτυχίας για τα δύο άνω όρια που εξετάζονται στην εργασία. Η διαφορά τους εντοπίζεται αποκλειστικά στο σημείο x_5 , το οποίο καταφεύγει στην εφεδρική τιμή: με $\max k = 9$ λαμβάνει 9, ενώ με $\max k = 30$ λαμβάνει 30. Για όλα τα υπόλοιπα σημεία η επέκταση του εύρους είναι εντελώς ουδέτερη — παρατήρηση που εξηγεί, όπως θα φανεί στην ενότητα 7.3, γιατί η Παραλλαγή 1 μεταβάλλει ελάχιστα την επίδοση σε καθαρά δεδομένα και σημαντικά σε θορυβώδη.

Πίνακας 4.2: Τιμές k που αναθέτει ο κανόνας της πρώτης επιτυχίας στα σημεία του Πίνακα 4.1.

Σημείο	Πρώτη επιτυχία	AdaNN ($\max k = 9$)	Παραλλαγή 1 ($\max k = 30$)
x_1	$k = 1$	1	1
x_2	$k = 2$	2	2
x_3	$k = 6$	6	6
x_4	$k = 1$	1	1
x_5	καμία	9 (εφεδρική)	30 (εφεδρική)

4.5 Ιδιότητες και οριακές περιπτώσεις

Από τη διατύπωση του αλγορίθμου προκύπτουν ορισμένες ιδιότητες που καθορίζουν τη συμπεριφορά του. Η **πρώτη** αφορά την εκφύλιση σε 1-NN. Στη χειρότερη περίπτωση, όπου σε όλα τα σημεία εκπαίδευσης έχει ανατεθεί $k = 1$, ο AdaNN συμπίπτει πλήρως με τον κανόνα του ενός εγγύτερου γείτονα. Η περίπτωση αυτή δεν είναι θεωρητική: σε καθαρά, καλά διαχωρισμένα σύνολα δεδομένων η συντριπτική πλειονότητα των σημείων λαμβάνει πράγματι $k = 1$, όπως επιβεβαιώνουν οι μετρήσεις της ενότητας 7.3. Η ιδιότητα αυτή εξηγεί γιατί ο AdaNN συμπεριφέρεται σχεδόν σαν 1-NN σε εύκολα προβλήματα.

Η **δεύτερη** ιδιότητα αφορά την εφεδρική τιμή. Τα σημεία που δεν κατηγοριοποιούνται σωστά για καμία τιμή k λαμβάνουν την τιμή $\max k$. Η τιμή αυτή δεν έχει επιλεγεί επειδή λειτουργεί, αλλά επειδή εξαντλήθηκε το εύρος αναζήτησης· πρόκειται δηλαδή για ομολογία αποτυχίας και όχι για επιλογή. Είναι ενδιαφέρον ότι τα σημεία αυτά ταυτίζονται εννοιολογικά με εκείνα που θα αφαιρούσε ένας κανόνας καθαρισμού τύπου ENN [16], παρατήρηση στην οποία επανέρχεται η ενότητα 7.3.

Η **τρίτη** ιδιότητα αφορά την ευαισθησία στο $\max k$. Το άνω όριο δεν είναι απλώς ένα τεχνικό όριο αναζήτησης, αλλά καθορίζει και την τιμή που λαμβάνουν όλα τα «δύσκολα» σημεία. Η μεταβολή του από 9 σε 30 επηρεάζει επομένως δύο διαφορετικές ομάδες σημείων: εκείνα που θα μπορούσαν να βρουν

μια τιμή k που πράγματι λειτουργεί κάπου μεταξύ 10 και 30, και εκείνα που εξακολουθούν να αποτυγχάνουν και λαμβάνουν πλέον ισχυρότερη εξομάλυνση.

Η **τέταρτη** ιδιότητα αφορά την ασυμμετρία της μεταφοράς. Η τιμή k ενός σημείου εκπαίδευσης μεταφέρεται σε όλα τα άγνωστα σημεία για τα οποία αυτό αποτελεί τον εγγύτερο γείτονα, δηλαδή σε ολόκληρο το κελί Voronoi του. Η επιρροή κάθε σημείου εκπαίδευσης είναι επομένως ανάλογη του μεγέθους του κελιού του: σημεία σε αραιές περιοχές επηρεάζουν πολύ μεγαλύτερο τμήμα του χώρου από σημεία σε πυκνές περιοχές. Πρόκειται για μια σιωπηρή στάθμιση που ο αλγόριθμος δεν ελέγχει.

4.6 Υπολογιστική πολυπλοκότητα

Το κόστος της φάσης εκπαίδευσης καθορίζεται από την εύρεση των $\max k$ εγγύτερων γειτόνων για καθένα από τα n σημεία εκπαίδευσης. Με χρήση χωρικής δομής ευρετηρίασης το κόστος αυτό είναι $O(n \cdot \log n \cdot \max k)$ κατά μέσο όρο, ενώ με εξαντλητική αναζήτηση ανέρχεται σε $O(n^2 \cdot d)$. Κρίσιμη παρατήρηση, την οποία αξιοποιεί η υλοποίηση, είναι ότι η σάρωση όλων των τιμών k από 1 έως $\max k$ **δεν** απαιτεί επανάληψη της αναζήτησης. Οι γείτονες υπολογίζονται μία φορά και ταξινομούνται κατά απόσταση· στη συνέχεια οι ψήφοι ανά κλάση συσσωρεύονται προοδευτικά. Το επιπλέον κόστος της σάρωσης περιορίζεται έτσι σε $O(n \cdot \max k \cdot m)$, όπου m το πλήθος των κλάσεων, ποσότητα αμελητέα σε σχέση με την αναζήτηση.

Η φάση ελέγχου απαιτεί την εύρεση έως $\max(k_opt)$ γειτόνων για καθένα από τα t σημεία ελέγχου, με κόστος $O(t \cdot \log n \cdot \max(k_opt))$. Ως προς τη μνήμη, ο αλγόριθμος απαιτεί την αποθήκευση του πίνακα γειτόνων μεγέθους $n \times \max k$, δηλαδή $O(n \cdot \max k)$, καθώς και του διανύσματος k_opt μεγέθους n . Συγκριτικά, ο αλγόριθμος καθολικού προσδιορισμού με F -πτυχη διασταυρούμενη επικύρωση εκτελεί F αναζητήσεις γειτόνων, μία ανά πτυχή, πάνω σε υποσύνολα μεγέθους $(F-1)n/F$ το καθένα, και επιπλέον μία τελική αναζήτηση για το σύνολο ελέγχου. Αναμένεται επομένως να είναι συστηματικά ακριβότερος από κάθε παραλλαγή του AdaNN, πρόβλεψη που επιβεβαιώνεται από τις μετρήσεις της ενότητας 7.3.

4.7 Αδυναμίες που κινητοποιούν τις παραλλαγές

Η ανάλυση των προηγούμενων ενοτήτων αναδεικνύει δύο συγκεκριμένες αδυναμίες του βασικού αλγορίθμου, οι οποίες αποτελούν και το αντικείμενο των παραλλαγών που παρουσιάζονται στο επόμενο κεφάλαιο.

Η **πρώτη** είναι το αυθαίρετο άνω όριο. Η τιμή 9 της αρχικής δημοσίευσης δεν συνοδεύεται από θεωρητική ή εμπειρική τεκμηρίωση. Σε δεδομένα με υψηλό επίπεδο θορύβου το όριο αυτό ενδέχεται να είναι υπερβολικά περιοριστικό: όσο περισσότερα σημεία καταφεύγουν στην εφεδρική τιμή, τόσο μεγαλύτερο μέρος του χώρου ταξινομείται με μια τιμή που δεν επιλέχθηκε επειδή λειτουργεί.

Η **δεύτερη**, και ουσιαστικότερη, είναι η ευθραυστότητα του κανόνα της πρώτης επιτυχίας. Όπως δείχνει το σημείο x_2 του Πίνακα 4.1, μια μεμονωμένη επιτυχία σε κάποια μικρή τιμή k μπορεί να είναι προϊόν μιας οριακής ψηφοφορίας που ανατρέπεται με την προσθήκη ενός και μόνο γείτονα. Ο κανόνας δεν διαθέτει κανέναν μηχανισμό να διακρίνει μια τέτοια τυχαία επιτυχία από μια σταθερή τοπική ιδιότητα, και η επιλογή που παγιώνει μεταδίδεται σε ολόκληρο το κελί Voronoi του σημείου. Το επόμενο κεφάλαιο εισάγει έναν κανόνα που απαιτεί **επιβεβαίωση** από περισσότερες της μίας διαδοχικές τιμές k , ακριβώς για να αντιμετωπίσει το ενδεχόμενο αυτό.

Κεφάλαιο 5ο: Νέες παραλλαγές του AdaNN

Το κεφάλαιο αυτό διατυπώνει με ακρίβεια τις τέσσερις παραλλαγές του AdaNN που σχεδιάστηκαν και υλοποιήθηκαν στο πλαίσιο της εργασίας. Η πρώτη μεταβάλλει μόνο το άνω όριο αναζήτησης, ενώ οι υπόλοιπες τρεις αντικαθιστούν τον κανόνα της πρώτης επιτυχίας με έναν κανόνα κυλιόμενου παραθύρου που απαιτεί διαδοχικές επιβεβαιώσεις. Ακολουθεί ο αλγόριθμος καθολικού προσδιορισμού που χρησιμοποιείται ως σημείο αναφοράς, καθώς και συγκεντρωτική παρουσίαση των έξι αλγορίθμων που αξιολογούνται πειραματικά. Υπενθυμίζεται ότι η φάση ελέγχου παραμένει κοινή σε όλες τις παραλλαγές· διαφοροποιείται αποκλειστικά ο κανόνας επιλογής της φάσης εκπαίδευσης.

5.1 Παραλλαγή 1: επέκταση του εύρους αναζήτησης

Η **Παραλλαγή 1** διατηρεί ακέραιο τον κανόνα της πρώτης επιτυχίας και μεταβάλλει μόνο το άνω όριο, από 9 σε 30. Η αλλαγή είναι ελάχιστη ως προς την υλοποίηση αλλά ουσιώδης ως προς τη συμπεριφορά, για δύο λόγους. Πρώτον, τα σημεία που προηγουμένως κατέληγαν στην εφεδρική τιμή 9 έχουν πλέον την ευκαιρία να βρουν μια τιμή k που πράγματι λειτουργεί, κάπου μεταξύ 10 και 30. Δεύτερον, τα σημεία που εξακολουθούν να μην κατηγοριοποιούνται σωστά για καμία τιμή λαμβάνουν πλέον την τιμή 30 αντί της 9, η οποία προσφέρει σημαντικά ισχυρότερη εξομάλυνση.

Και οι δύο επιδράσεις αναμένεται να είναι εντονότερες σε θορυβώδη δεδομένα, όπου το ποσοστό των σημείων που καταφεύγουν στην εφεδρική τιμή είναι πολλαπλάσιο. Αντιστρόφως, σε καθαρά δεδομένα, όπου η συντριπτική πλειονότητα των σημείων βρίσκει επιτυχία ήδη στο $k = 1$ ή $k = 2$, η παραλλαγή αναμένεται πρακτικά ουδέτερη. Η πρόβλεψη αυτή επιβεβαιώνεται πλήρως από τα πειραματικά αποτελέσματα.

5.2 Ο κανόνας κυλιόμενου παραθύρου

Οι **Παραλλαγές 2, 3 και 4** αντικαθιστούν τον κανόνα της πρώτης επιτυχίας με έναν κανόνα που απαιτεί επιβεβαίωση. Αντί να γίνεται δεκτή η πρώτη τιμή k που κατηγοριοποιεί σωστά το σημείο, εξετάζονται **τριάδες διαδοχικών τιμών**. Μια τριάδα θεωρείται επιτυχής μόνο αν **και οι τρεις** τιμές που περιέχει κατηγοριοποιούν σωστά το σημείο. Η πρώτη επιτυχής τριάδα κερδίζει, και στο σημείο ανατίθεται η **μεσαία** τιμή της. Αν καμία τριάδα δεν επιτύχει, ανατίθεται η εφεδρική τιμή 30.

Το σκεπτικό πίσω από τον κανόνα αυτόν είναι η αναζήτηση **σταθερότητας** αντί για την αναζήτηση της πρώτης επιτυχίας. Μια μεμονωμένη επιτυχία σε κάποια τιμή k μπορεί να είναι τυχαία, προϊόν μιας οριακής ψηφοφορίας που θα αντιστρεφόταν με την προσθήκη ενός ακόμη γείτονα. Αντιθέτως, τρεις διαδοχικές επιτυχίες υποδηλώνουν ότι η σωστή απόφαση διατηρείται καθώς η γειτονιά διευρύνεται, δηλαδή ότι η επιλογή αντανακλά μια σταθερή τοπική ιδιότητα και όχι μια σύμπτωση. Η ιδέα συγγενεύει άμεσα με την τρίτη ευρετική πρόωρης διακοπής των Ougiaroglou και συνεργατών [39], η οποία διακόπτει την ψηφοφορία όταν ένα προκαθορισμένο πλήθος διαδοχικών γειτόνων ψηφίζει υπέρ της ίδιας κλάσης· εδώ όμως η απαίτηση διαδοχικών επιβεβαιώσεων μεταφέρεται από τη φάση ελέγχου στη φάση **επιλογής της τιμής k** .

Η ανάθεση της μεσαίας τιμής, αντί της πρώτης ή της τελευταίας του παραθύρου, τοποθετεί την επιλογή στο κέντρο της περιοχής σταθερότητας, μεγιστοποιώντας το περιθώριο και προς τις δύο κατευθύνσεις: όποια και αν είναι η κατεύθυνση της απόκλισης του άγνωστου σημείου από τον γείτονά του, η επιλεγμένη τιμή απέχει τουλάχιστον μία θέση από το όριο της περιοχής όπου η απόφαση παραμένει σωστή.

Αλγόριθμος 5.1: Ο κανόνας κυλιόμενου παραθύρου.

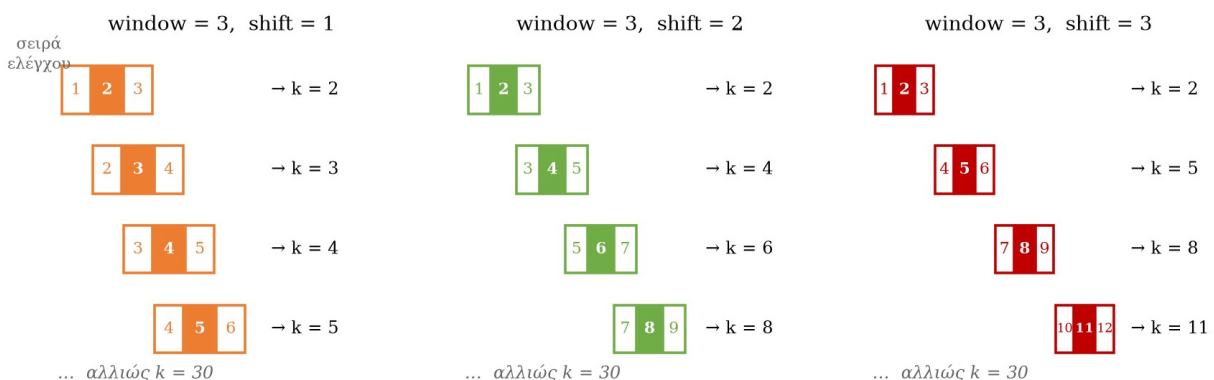
Είσοδος : πίνακας ορθότητας `correct[1..n][1..maxk]`, πλάτος w , βήμα s
Έξοδος : πίνακας `k_opt[1..n]`

```
1  για κάθε σημείο  $i$ :  
2       $k\_opt[i] \leftarrow \max k$                                 # εφεδρική τιμή  
3       $start \leftarrow 1$   
4      όσο  $start + w - 1 \leq \max k$ :  
5          αν correct[i][start .. start+w-1] είναι όλα αληθή:  
6               $k\_opt[i] \leftarrow start + \text{floor}(w/2)$         # μεσαία τιμή του παραθύρου  
7              διάκοψε  
8           $start \leftarrow start + s$                             # μετατόπιση του παραθύρου  
9  επέστρεψε  $k\_opt$ 
```

5.3 Παραλλαγές 2, 3 και 4: το βήμα μετατόπισης

Οι τρεις παραλλαγές διαφέρουν αποκλειστικά ως προς το **βήμα μετατόπισης** s του παραθύρου, όπως απεικονίζεται στο Σχήμα 5.1. Στην **Παραλλαγή 2** το βήμα είναι 1 και τα παράθυρα επικαλύπτονται σχεδόν πλήρως, εξετάζοντας κάθε δυνατή τριάδα: $[1,2,3]$, $[2,3,4]$, $[3,4,5]$ και ούτω καθεξής. Στην **Παραλλαγή 3** το βήμα είναι 2 και η επικάλυψη περιορίζεται σε ένα στοιχείο: $[1,2,3]$, $[3,4,5]$, $[5,6,7]$. Στην **Παραλλαγή 4** το βήμα είναι 3 και τα παράθυρα είναι ξένα μεταξύ τους, διαμερίζοντας το διάστημα $[1, 30]$ σε δέκα διαδοχικές τριάδες: $[1,2,3]$, $[4,5,6]$, $[7,8,9]$.

Η αύξηση του βήματος έχει δύο αντίρροπες συνέπειες. Αφενός μειώνει το πλήθος των παραθύρων που εξετάζονται και επομένως αυξάνει το ενδεχόμενο να προσπεραστεί μια περιοχή σταθερότητας που δεν είναι ευθυγραμμισμένη με το πλέγμα των παραθύρων· ένα σημείο του οποίου η μοναδική σταθερή περιοχή είναι η $[2,3,4]$ θα εντοπιστεί από την Παραλλαγή 2 αλλά όχι από τις άλλες δύο. Αφετέρου, οδηγεί σε συστηματικά μεγαλύτερες τιμές k , καθώς η αναζήτηση προχωρά ταχύτερα προς τα δεξιά. Το βήμα λειτουργεί δηλαδή ως ρυθμιστής της **επιθετικότητας** με την οποία η αναζήτηση μετακινείται προς μεγαλύτερες γειτονιές.



Σχήμα 5.1: Ο κανόνας κυλιόμενου παραθύρου για βήματα 1, 2 και 3. Εμφανίζονται τα τέσσερα πρώτα παράθυρα κάθε παραλλαγής και η τιμή k που ανατίθεται σε καθένα.

5.4 Παράδειγμα εφαρμογής των παραλλαγών

Ο Πίνακας 5.1 εφαρμόζει και τους τέσσερις κανόνες στα πέντε σημεία του παραδείγματος του Πίνακα 4.1, κάνοντας ορατές τις διαφορές τους. Το σημείο x_i , που πετυχαίνει παντού, λαμβάνει από όλες τις παραλλαγές παραθύρου την τιμή 2, δηλαδή τη μεσαία τιμή του πρώτου παραθύρου $[1,2,3]$, αντί για την

τιμή 1 του βασικού κανόνα. Το x_2 , που εμφανίζει την οριακή επιτυχία στο $k = 2$, δεν λαμβάνει πλέον την τιμή 2 αλλά τιμές 5 ή 6, ανάλογα με το βήμα, καθώς η πρώτη πραγματικά σταθερή περιοχή του είναι η $[4,5,6]$. Το x_3 λαμβάνει 7 ή 8 αντί για 6. Το x_4 , που πετυχαίνει μόνο στο $k = 1$, δεν διαθέτει καμία σταθερή τριάδα και καταφεύγει στην εφεδρική τιμή 30, παρότι ο βασικός κανόνας του είχε αναθέσει την τιμή 1.

Η τελευταία αυτή περίπτωση είναι διαφωτιστική. Ο κανόνας παραθύρου δεν είναι απλώς αυστηρότερος· είναι αυστηρότερος με τρόπο που μπορεί να **αντιστρέψει** πλήρως την επιλογή, μετατρέποντας μια πολύ μικρή τιμή σε πολύ μεγάλη. Αν το x_4 είναι γνήσιο απομονωμένο σημείο μιας μικρής κλάσης, η μεταβολή αυτή είναι επιζήμια· αν είναι σημείο με αλλοιωμένη ετικέτα, είναι ακριβώς η επιθυμητή συμπεριφορά. Ποια από τις δύο περιπτώσεις κυριαρχεί εξαρτάται από τα δεδομένα, και αυτό ακριβώς μετρά η πειραματική μελέτη του Κεφαλαίου 7.

Πίνακας 5.1: Τιμές k που αναθέτουν οι έξι κανόνες επιλογής στα σημεία του Πίνακα 4.1.

Σημείο	AdaNN $k \leq 9$	AdaNN $k \leq 30$	$w=3, s=1$	$w=3, s=2$	$w=3, s=3$
x_1	1	1	2	2	2
x_2	2	2	5	6	5
x_3	6	6	7	8	8
x_4	1	1	30	30	30
x_5	9	30	30	30	30

5.5 Δομικές ιδιότητες των παραλλαγών παραθύρου

Ένα σημείο που αξίζει να επισημανθεί ιδιαίτερος είναι ότι ο κανόνας παραθύρου **δεν μπορεί ποτέ να αναθέσει την τιμή $k = 1$** , καθώς η μικρότερη δυνατή μεσαία τιμή είναι το 2. Το γεγονός αυτό αποκλείει εξ ορισμού τη συμπεριφορά του ενός εγγύτερου γείτονα, η οποία στα καθαρά δεδομένα είναι συχνά η βέλτιστη. Πρόκειται για δομικό χαρακτηριστικό του κανόνα και όχι για επιλογή υλοποίησης: γενικότερα, για πλάτος παραθύρου w , η μικρότερη δυνατή τιμή που μπορεί να ανατεθεί είναι $1 + \text{floor}(w/2)$.

Μια δεύτερη δομική ιδιότητα είναι ότι ο κανόνας παραθύρου αναθέτει πάντοτε τιμή **μεγαλύτερη ή ίση** εκείνης του κανόνα της πρώτης επιτυχίας. Πράγματι, αν το πρώτο παράθυρο που ικανοποιείται ξεκινά στη θέση start , τότε όλες οι τιμές $\text{start}, \text{start}+1, \dots, \text{start}+w-1$ είναι επιτυχείς, οπότε η πρώτη επιτυχία εμφανίζεται το αργότερο στη θέση start , ενώ ο κανόνας παραθύρου αναθέτει $\text{start} + \text{floor}(w/2)$. Η μονότονη αυτή σχέση εξηγεί γιατί οι μέσες τιμές k των παραλλαγών παραθύρου είναι συστηματικά υψηλότερες, τόσο σε καθαρά όσο και σε θορυβώδη δεδομένα.

Η τρίτη ιδιότητα αφορά το ποσοστό των σημείων που καταφεύγουν στην εφεδρική τιμή. Επειδή η συνθήκη επιτυχίας είναι αυστηρότερη, τα σημεία που δεν την ικανοποιούν είναι περισσότερα από εκείνα του βασικού κανόνα. Το ποσοστό της εφεδρικής τιμής αυξάνεται επομένως σε κάθε παραλλαγή παραθύρου, και μάλιστα μονότονα ως προς το βήμα, όπως επιβεβαιώνουν οι μετρήσεις του Παραρτήματος Β.

5.6 Ο αλγόριθμος καθολικού προσδιορισμού ως σημείο αναφοράς

Ως σημείο αναφοράς χρησιμοποιείται ο κλασικός καθολικός προσδιορισμός του k που περιγράφηκε στην ενότητα 2.3. Ο αλγόριθμος εκτελεί στρωματοποιημένη πενταπλή διασταυρούμενη επικύρωση πάνω στο σύνολο εκπαίδευσης για κάθε τιμή k από 1 έως 30, δηλαδή τριάντα ανεξάρτητες επικυρώσεις συνολικά, και επιλέγει την τιμή που πέτυχε τη μεγαλύτερη μέση ορθότητα. Με την τιμή αυτή ταξινομείται στη συνέχεια ολόκληρο το σύνολο ελέγχου με απλό k -NN.

Η σύγκριση με τον αλγόριθμο αυτόν είναι ουσιώδης για δύο λόγους. Αφενός απαντά στο πρακτικό ερώτημα αν η πολυπλοκότητα της τοπικής προσαρμογής ανταποδίδεται σε επίδοση έναντι της απλούστερης και ευρέως χρησιμοποιούμενης εναλλακτικής. Αφετέρου καλύπτει το κενό που εντοπίστηκε στην ενότητα 3.11: οι περισσότερες προσαρμοστικές μέθοδοι της βιβλιογραφίας συγκρίνονται με τον k-NN για λίγες αυθαίρετες σταθερές τιμές k και όχι με τη βέλτιστη τιμή που θα επέλεγε ένας προσεκτικός χρήστης.

5.7 Συγκεντρωτική παρουσίαση των αλγορίθμων

Στον Πίνακα 5.2 συνοψίζονται οι έξι αλγόριθμοι που αξιολογούνται πειραματικά, μαζί με τις παραμέτρους τους και τον κανόνα επιλογής που εφαρμόζει καθένας. Οι συντομογραφίες της τελευταίας στήλης χρησιμοποιούνται σε όλους τους πίνακες και τα γραφήματα του Κεφαλαίου 7.

Πίνακας 5.2: Οι έξι αλγόριθμοι που αξιολογούνται.

Αλγόριθμος	maxk	Κανόνας επιλογής	Συντομογραφία
best global k	30	5-fold CV για $k = 1..30$, μία καθολική τιμή	best global k
AdaNN (βάση)	9	μικρότερο k που πετυχαίνει, αλλιώς 9	AdaNN $k \leq 9$
Παραλλαγή 1	30	μικρότερο k που πετυχαίνει, αλλιώς 30	AdaNN $k \leq 30$
Παραλλαγή 2	30	παράθυρο 3, βήμα 1, μεσαία τιμή	w=3, s=1
Παραλλαγή 3	30	παράθυρο 3, βήμα 2, μεσαία τιμή	w=3, s=2
Παραλλαγή 4	30	παράθυρο 3, βήμα 3, μεσαία τιμή	w=3, s=3

Το υπολογιστικό κόστος των παραλλαγών διαφέρει ελάχιστα από εκείνο του βασικού αλγορίθμου. Ο κανόνας επιλογής, είτε πρόκειται για την πρώτη επιτυχία είτε για το κυλιόμενο παράθυρο, εφαρμόζεται πάνω στον ήδη υπολογισμένο πίνακα ορθότητας και κοστίζει $O(n \cdot \max k)$ πράξεις, ποσότητα αμελητέα σε σχέση με την αναζήτηση γειτόνων. Η μόνη ουσιαστική διαφορά κόστους προέρχεται από την επέκταση του maxk από 9 σε 30, η οποία απαιτεί την ανάκτηση περισσότερων γειτόνων ανά σημείο. Όπως θα φανεί στην ενότητα 7.3, η τριπλάσια αύξηση του άνω ορίου δεν τριπλασιάζει τον χρόνο αλλά τον αυξάνει κατά περίπου 30 %, διότι το κυρίαρχο κόστος είναι η κατασκευή της δομής ευρετηρίασης και η ίδια η αναζήτηση.

Κεφάλαιο 6ο: Υλοποιήσεις σε Python

Το κεφάλαιο αυτό παρουσιάζει την υλοποίηση των αλγορίθμων σε γλώσσα Python. Για κάθε συνάρτηση παρατίθεται στιγμιότυπο του πηγαίου κώδικα και εξηγείται αναλυτικά ποιο τμήμα του επιτελεί ποια λειτουργία, ώστε η αντιστοίχιση ανάμεσα στη διατύπωση των Κεφαλαίων 4 και 5 και στον εκτελέσιμο κώδικα να είναι απολύτως σαφής. Το κεφάλαιο κλείνει με τη διαδικασία επαλήθευσης της ορθότητας της υλοποίησης, η οποία σε μια πειραματική εργασία αποτελεί προϋπόθεση για την αξιοπιστία κάθε συμπεράσματος.

6.1 Βιβλιοθήκες και οργάνωση του κώδικα

Η υλοποίηση έγινε σε Python 3.13 με χρήση δύο μόνο εξωτερικών βιβλιοθηκών: της NumPy [31] για τους διανυσματικούς υπολογισμούς και της scikit-learn [30] για την αναζήτηση εγγύτερων γειτόνων, τη στρωματοποιημένη διαμέριση των πτυχών και τον υπολογισμό των μετρικών. Η αναζήτηση γειτόνων γίνεται μέσω της κλάσης NearestNeighbors με παράμετρο `algorithm="auto"`, η οποία επιλέγει αυτόματα μεταξύ k-d δέντρου, ball tree και εξαντλητικής αναζήτησης ανάλογα με τη διαστατικότητα και το μέγεθος του συνόλου. Στα πειράματα η διαστατικότητα κυμαίνεται από 2 έως 40, εύρος εντός του οποίου η αυτόματη επιλογή λειτουργεί ικανοποιητικά.

Η βασική σχεδιαστική επιλογή είναι ότι η υλοποίηση **διαχωρίζει πλήρως** τον υπολογισμό του πίνακα ορθότητας από τον κανόνα επιλογής. Ο πίνακας ορθότητας υπολογίζεται μία φορά ανά σύνολο δεδομένων και τροφοδοτεί όλους τους κανόνες, οι οποίοι υλοποιούνται ως ανεξάρτητες, καθαρές συναρτήσεις πάνω σε αυτόν. Ο διαχωρισμός αυτός δεν είναι μόνο θέμα αποδοτικότητας: εγγυάται ότι κάθε διαφορά στα αποτελέσματα οφείλεται αποκλειστικά στον κανόνα και όχι σε τυχόν διαφορές του υποκείμενου υπολογισμού. Ο Πίνακας 6.1 συνοψίζει τις συναρτήσεις της υλοποίησης και τον ρόλο τους.

Πίνακας 6.1: Οι συναρτήσεις της υλοποίησης και ο ρόλος τους.

Συνάρτηση	Φάση	Ρόλος	Αλγόριθμοι που εξυπηρετεί
<code>loo_correct_matrix</code>	Εκπαίδευση	Παράγει τον πίνακα ορθότητας $n \times \max k$ με <code>leave-one-out</code>	Όλες οι παραλλαγές AdaNN
<code>k_plain</code>	Εκπαίδευση	Κανόνας πρώτης επιτυχίας, εφεδρική τιμή <code>maxk</code>	AdaNN $k \leq 9$, Παραλλαγή 1
<code>k_window</code>	Εκπαίδευση	Κανόνας κυλιόμενου παραθύρου, μεσαία τιμή	Παραλλαγές 2, 3 και 4
<code>adann_predict</code>	Έλεγχος	Μεταφορά της τιμής k μέσω του εγγύτερου γείτονα	Όλες οι παραλλαγές AdaNN
<code>find_best_global_k</code>	Εκπαίδευση	Στρωματοποιημένη 5-fold CV για $k = 1..30$	best global k
<code>batch_majority_vote</code>	Κοινή	Διανυσματική πλειοψηφική ψηφοφορία	Όλοι
<code>time_uniformly</code>	Μέτρηση	Κοινός αριθμός επαναλήψεων ανά σύνολο δεδομένων	Όλοι

6.2 Φάση εκπαίδευσης: η συνάρτηση `loo_correct_matrix`

Η συνάρτηση του Σχήματος 6.1 αποτελεί τον πυρήνα της υλοποίησης. Υλοποιεί τον Αλγόριθμο 4.1 και παράγει, με ένα και μόνο ερώτημα γειτόνων, τον πίνακα ορθότητας για κάθε σημείο εκπαίδευσης και κάθε τιμή k .

```

adann.py — loo_correct_matrix()

1 def loo_correct_matrix(X_train, y_train, max_k):
2     """
3     Leave-one-out kNN πάνω στο σύνολο εκπαίδευσης.
4     Επιστρέφει πίνακα correct διαστάσεων (n, max_k) όπου correct[i, k-1]
5     είναι True όταν ο k-NN με τη συγκεκριμένη τιμή k κατηγοριοποιεί σωστά
6     το σημείο i (με το ίδιο το σημείο εξαιρεμένο από τη γειτονιά του).
7     """
8     n = len(X_train)
9     n_classes = int(y_train.max()) + 1
10    rows = np.arange(n)
11
12    nbrs = NearestNeighbors(n_neighbors=max_k + 1, algorithm="auto").fit(X_train)
13    _, idx = nbrs.kneighbors(X_train)
14
15    # Αφαίρεση του ίδιου του σημείου. Με διπλότυπα σημεία δεν βρίσκεται
16    # πάντα στη στήλη 0, οπότε εντοπίζεται ρητά.
17    is_self = idx == rows[:, None]
18    self_pos = np.where(is_self.any(axis=1), is_self.argmax(axis=1), max_k)
19    keep = np.arange(max_k + 1)[None, :] != self_pos[:, None]
20    nn = idx[keep].reshape(n, max_k)
21
22    counts = np.zeros((n, n_classes), dtype=np.int32)
23    correct = np.zeros((n, max_k), dtype=bool)
24    for k in range(1, max_k + 1):
25        counts[rows, y_train[nn[:, k - 1]]] += 1
26        correct[:, k - 1] = counts.argmax(axis=1) == y_train
27    return correct

```

Σχήμα 6.1: Η συνάρτηση loo_correct_matrix, που υλοποιεί τη φάση εκπαίδευσης με leave-one-out για όλες τις τιμές k.

Η ανάλυση του κώδικα ανά τμήμα έχει ως εξής. Οι **γραμμές 12–13** εκτελούν τη μοναδική αναζήτηση γειτόνων, ζητώντας maxk + 1 γείτονες αντί για maxk. Ο επιπλέον γείτονας είναι απαραίτητος επειδή, στην αναζήτηση του συνόλου εκπαίδευσης πάνω στον εαυτό του, ένας από τους επιστρεφόμενους γείτονες είναι το ίδιο το σημείο.

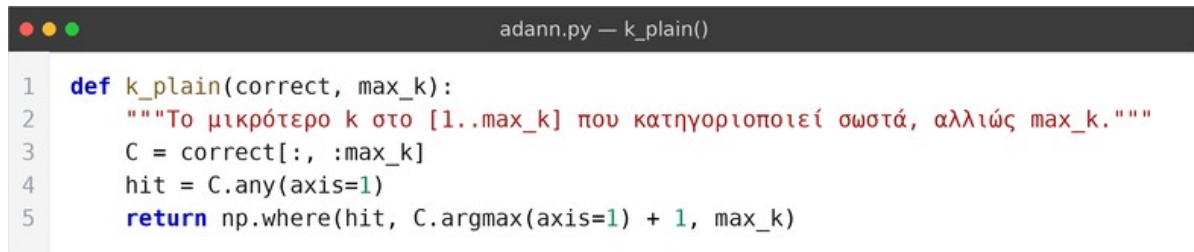
Οι **γραμμές 17–20** υλοποιούν τη ρητή αφαίρεση του σημείου από τη γειτονιά του. Η προφανής υλοποίηση, που απλώς παραλείπει την πρώτη στήλη του πίνακα γειτόνων, είναι **λανθασμένη** όταν υπάρχουν διπλότυπα σημεία: αν δύο σημεία έχουν ταυτόσημα διανύσματα χαρακτηριστικών, το ίδιο το σημείο δεν βρίσκεται αναγκαστικά στη θέση 0. Η γραμμή 17 εντοπίζει σε ποια στήλη εμφανίζεται ο δείκτης του ίδιου του σημείου, η γραμμή 18 αντιμετωπίζει και την περίπτωση όπου το σημείο απουσιάζει εντελώς από τη λίστα, και οι γραμμές 19–20 κατασκευάζουν τον τελικό πίνακα γειτόνων διαστάσεων $n \times \text{maxk}$. Το σφάλμα αυτό είναι δύσκολο να εντοπιστεί, διότι επηρεάζει μόνο τα σύνολα δεδομένων που περιέχουν διπλότυπα και εκεί μόνο έναν μικρό αριθμό σημείων.

Οι **γραμμές 22–26** αποτελούν το σημείο στο οποίο η υλοποίηση κερδίζει την αποδοτικότητά της. Ο πίνακας counts διαστάσεων $n \times \text{πλήθος_κλάσεων}$ συγκρατεί τις τρέχουσες ψήφους κάθε σημείου. Σε κάθε επανάληψη του βρόχου προστίθεται **ένας** γείτονας (γραμμή 25) και ελέγχεται αν η πλειοψηφούσα κλάση συμπίπτει με την πραγματική ετικέτα (γραμμή 26). Έτσι, ολόκληρη η σάρωση $k = 1..\text{maxk}$ κοστίζει maxk φθηνές διανυσματικές ενημερώσεις αντί για maxk ανεξάρτητες ταξινομήσεις. Σημειώνεται ότι η συνάρτηση argmax της NumPy επιστρέφει τον **μικρότερο** δείκτη σε περίπτωση

ισοπαλίας, γεγονός που υλοποιεί ακριβώς τον κανόνα επίλυσης ισοπαλιών που περιγράφηκε στην ενότητα 1.2 και εφαρμόζεται με τον ίδιο τρόπο σε όλες τις παραλλαγές.

6.3 Ο κανόνας της πρώτης επιτυχίας: η συνάρτηση `k_plain`

Η συνάρτηση του Σχήματος 6.2 υλοποιεί τον κανόνα επιλογής του βασικού AdaNN. Είναι η ίδια συνάρτηση που εξυπηρετεί και την Παραλλαγή 1: η μόνη διαφορά ανάμεσα στους δύο αλγορίθμους είναι η τιμή της παραμέτρου `max_k`, που καλείται με 9 και 30 αντίστοιχα.



```
adann.py — k_plain()
1 def k_plain(correct, max_k):
2     """Το μικρότερο k στο [1..max_k] που κατηγοριοποιεί σωστά, αλλιώς max_k."""
3     C = correct[:, :max_k]
4     hit = C.any(axis=1)
5     return np.where(hit, C.argmax(axis=1) + 1, max_k)
```

Σχήμα 6.2: Η συνάρτηση `k_plain`, που υλοποιεί τον κανόνα της πρώτης επιτυχίας για τον βασικό AdaNN και την Παραλλαγή 1.

Η **γραμμή 3** περικόπτει τον πίνακα ορθότητας στις πρώτες `max_k` στήλες· έτσι, ο ίδιος πίνακας που υπολογίστηκε για `maxk = 30` μπορεί να χρησιμοποιηθεί και για τον βασικό AdaNN με `maxk = 9`, χωρίς επανάληψη της αναζήτησης γειτόνων. Η **γραμμή 4** εντοπίζει, για κάθε σημείο, αν υπάρχει έστω μία επιτυχής τιμή. Η **γραμμή 5** συνδυάζει τα δύο: η `argmax` επί ενός δυαδικού πίνακα επιστρέφει τη θέση της **πρώτης** αληθούς τιμής, οπότε η πρόσθεση της μονάδας δίνει την πρώτη επιτυχή τιμή `k`· όπου δεν υπάρχει καμία επιτυχία, η `np.where` αναθέτει την εφεδρική τιμή `max_k`. Ολόκληρος ο κανόνας υλοποιείται δηλαδή σε τρεις διανυσματικές πράξεις, χωρίς κανέναν βρόχο πάνω στα σημεία.

6.4 Ο κανόνας κυλιόμενου παραθύρου: η συνάρτηση `k_window`

Η συνάρτηση του Σχήματος 6.3 υλοποιεί τον Αλγόριθμο 5.1 και εξυπηρετεί και τις τρεις παραλλαγές παραθύρου· η διαφοροποίησή τους γίνεται αποκλειστικά μέσω της παραμέτρου `shift`, η οποία λαμβάνει τις τιμές 1, 2 και 3 για τις Παραλλαγές 2, 3 και 4 αντίστοιχα.

```

1  def k_window(correct, max_k, window=3, shift=1):
2      """
3      Κερδίζει το πρώτο παράθυρο `window` διαδοχικών τιμών k που ΟΛΕΣ
4      κατηγοριοποιούν σωστά το σημείο· ανατίθεται η μεσαία τιμή του.
5      Τα παράθυρα ξεκινούν στο 1 και προχωρούν κατά `shift`. Εφεδρική: max_k.
6      """
7      n = correct.shape[0]
8      k_opt = np.full(n, max_k, dtype=int)
9      undecided = np.ones(n, dtype=bool)
10
11     start = 1
12     while start + window - 1 <= max_k:
13         block = correct[:, start - 1:start - 1 + window].all(axis=1)
14         newly = undecided & block
15         k_opt[newly] = start + window // 2      # μεσαία τιμή του παραθύρου
16         undecided &= ~newly
17         if not undecided.any():
18             break
19         start += shift
20     return k_opt

```

Σχήμα 6.3: Η συνάρτηση `k_window`, που υλοποιεί τον κανόνα κυλιόμενου παραθύρου των Παραλλαγών 2, 3 και 4.

Οι **γραμμές 8–9** αρχικοποιούν τη διαδικασία: όλα τα σημεία λαμβάνουν εξ αρχής την εφεδρική τιμή `max_k` και σημειώνονται ως «μη αποφασισμένα». Η στρατηγική αυτή απλοποιεί τον κώδικα, διότι η εφεδρική τιμή δεν χρειάζεται να ανατεθεί σε κάποιο ξεχωριστό βήμα στο τέλος: όποιο σημείο δεν βρει επιτυχές παράθυρο διατηρεί απλώς την αρχική του τιμή.

Ο βρόχος των **γραμμών 12–19** διατρέχει τα παράθυρα και όχι τα σημεία. Η **γραμμή 13** είναι το κεντρικό σημείο του κανόνα: η έκφραση `correct[:, start-1 : start-1+window].all(axis=1)` ελέγχει **ταυτόχρονα για όλα τα σημεία** αν και οι τρεις διαδοχικές τιμές του τρέχοντος παραθύρου είναι επιτυχείς, επιστρέφοντας ένα δυαδικό διάνυσμα μήκους `n`. Η **γραμμή 14** απομονώνει εκείνα τα σημεία που ικανοποιούν τη συνθήκη **και** δεν έχουν αποφασιστεί σε προηγούμενο παράθυρο· η σύζευξη αυτή είναι που υλοποιεί την απαίτηση να κερδίζει το **πρώτο** επιτυχές παράθυρο. Η **γραμμή 15** αναθέτει τη μεσαία τιμή του παραθύρου, `start + window // 2`, και η **γραμμή 16** αφαιρεί τα σημεία αυτά από το σύνολο των μη αποφασισμένων.

Οι **γραμμές 17–18** αποτελούν βελτιστοποίηση πρόωρου τερματισμού: μόλις αποφασιστούν όλα τα σημεία, ο βρόχος διακόπτεται χωρίς να εξεταστούν τα υπόλοιπα παράθυρα. Τέλος, η **γραμμή 19** μετατοπίζει το παράθυρο κατά `shift` θέσεις, υλοποιώντας έτσι τη μόνη διαφορά ανάμεσα στις τρεις παραλλαγές. Αξίζει να σημειωθεί ότι ολόκληρος ο κανόνας απαιτεί το πολύ $\text{ceil}(\text{max_k} / \text{shift})$ επαναλήψεις, δηλαδή το πολύ 28 για `shift = 1`, ανεξάρτητα από το μέγεθος του συνόλου δεδομένων.

6.5 Φάση ελέγχου: η συνάρτηση `adann_predict`

Η συνάρτηση του Σχήματος 6.4 υλοποιεί τον Αλγόριθμο 4.2 και είναι κοινή σε όλες τις παραλλαγές: δέχεται το διάνυσμα `k_opt` που παρήγαγε ο εκάστοτε κανόνας και δεν γνωρίζει ποιος κανόνας το παρήγαγε.

```

adann.py — adann_predict()

1  def adann_predict(X_train, y_train, k_opt, X_test):
2      """0 εγγύτερος γείτονας κάθε σημείου ελέγχου δίνει την τιμή k."""
3      n_classes = int(y_train.max()) + 1
4      max_k = int(k_opt.max())
5
6      nbrs = NearestNeighbors(n_neighbors=max_k, algorithm="auto").fit(X_train)
7      _, nn = nbrs.kneighbors(X_test)
8
9      ks = k_opt[nn[:, 0]]
10     y_pred = np.empty(len(X_test), dtype=int)
11     for kv in np.unique(ks):
12         sel = ks == kv
13         y_pred[sel] = batch_majority_vote(y_train[nn[sel, :kv]], n_classes)
14     return y_pred

```

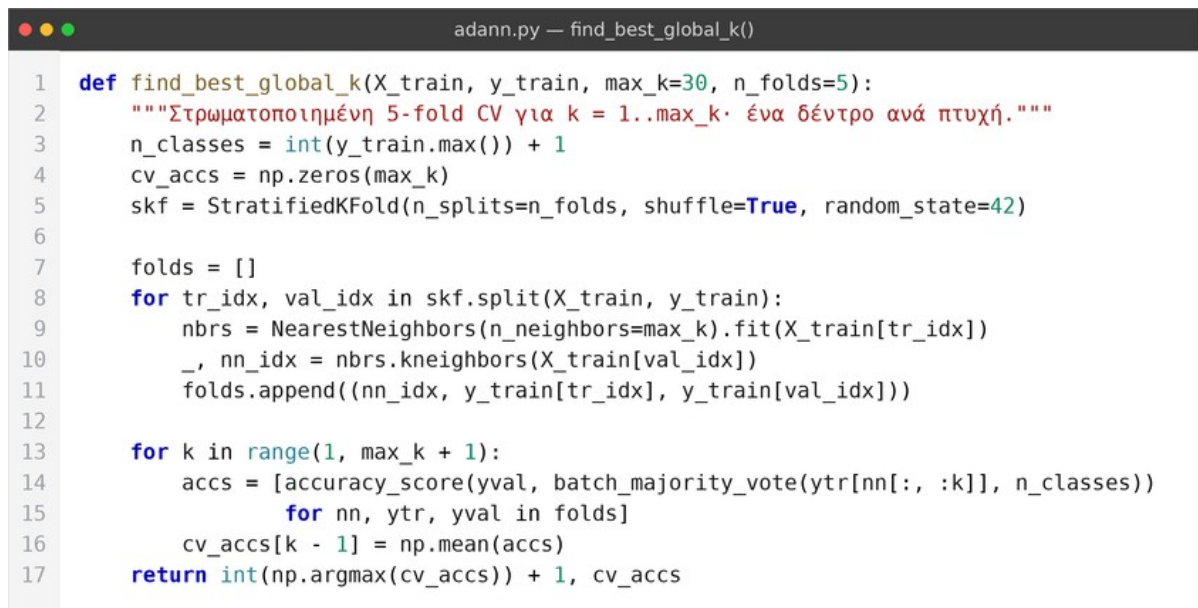
Σχήμα 6.4: Η συνάρτηση `adann_predict`, που υλοποιεί τη φάση ελέγχου, κοινή σε όλες τις παραλλαγές.

Η **γραμμή 6** ζητά `max_k` γείτονες για κάθε σημείο ελέγχου, όπου `max_k` είναι η μεγαλύτερη τιμή που έχει ανατεθεί σε οποιοδήποτε σημείο εκπαίδευσης· η αναζήτηση γίνεται επομένως μία φορά και καλύπτει κάθε ενδεχόμενη ανάγκη. Η **γραμμή 9** υλοποιεί τον πυρήνα της ιδέας του AdaNN σε μία μόνο έκφραση: το `nn[:, 0]` είναι ο εγγύτερος γείτονας κάθε σημείου ελέγχου και το `k_opt[nn[:, 0]]` η τιμή `k` που του έχει ανατεθεί.

Ο βρόχος των **γραμμών 11–13** ομαδοποιεί τα σημεία ελέγχου ανά τιμή `k`. Η ομαδοποίηση αυτή είναι απαραίτητη για την αποδοτικότητα: επειδή διαφορετικά σημεία ελέγχου χρησιμοποιούν διαφορετικό πλήθος γειτόνων, δεν είναι δυνατή μία ενιαία διανυσματική ψηφοφορία για όλα. Αντί όμως για έναν βρόχο ανά σημείο ελέγχου, ο κώδικας εκτελεί έναν βρόχο ανά **διακριτή τιμή** `k`, δηλαδή το πολύ 30 επαναλήψεις, μέσα στις οποίες η ψηφοφορία γίνεται διανυσματικά για όλα τα σημεία της ομάδας ταυτόχρονα.

6.6 Καθολικός προσδιορισμός: η συνάρτηση `find_best_global_k`

Η συνάρτηση του Σχήματος 6.5 υλοποιεί τον αλγόριθμο αναφοράς, δηλαδή τη ρύθμιση του `k` με στρωματοποιημένη πενταπλή διασταυρούμενη επικύρωση.



```

1 def find_best_global_k(X_train, y_train, max_k=30, n_folds=5):
2     """Στρωματοποιημένη 5-fold CV για k = 1..max_k· ένα δέντρο ανά πτυχή."""
3     n_classes = int(y_train.max()) + 1
4     cv_accs = np.zeros(max_k)
5     skf = StratifiedKFold(n_splits=n_folds, shuffle=True, random_state=42)
6
7     folds = []
8     for tr_idx, val_idx in skf.split(X_train, y_train):
9         nbrs = NearestNeighbors(n_neighbors=max_k).fit(X_train[tr_idx])
10        _, nn_idx = nbrs.kneighbors(X_train[val_idx])
11        folds.append((nn_idx, y_train[tr_idx], y_train[val_idx]))
12
13    for k in range(1, max_k + 1):
14        accs = [accuracy_score(yval, batch_majority_vote(ytr[nn[:, :k]], n_classes))
15                for nn, ytr, yval in folds]
16        cv_accs[k - 1] = np.mean(accs)
17    return int(np.argmax(cv_accs)) + 1, cv_accs

```

Σχήμα 6.5: Η συνάρτηση `find_best_global_k`, που υλοποιεί τη ρύθμιση του `k` με πενταπλή στρωματοποιημένη διασταυρούμενη επικύρωση.

Ο βρόχος των **γραμμών 8–11** εκτελείται μία φορά ανά πτυχή και αποθηκεύει, για κάθε πτυχή, τους `max_k` εγγύτερους γείτονες κάθε σημείου επικύρωσης μαζί με τις αντίστοιχες ετικέτες. Το ουσιώδες σημείο είναι ότι η δαπανηρή αναζήτηση γειτόνων εκτελείται **πέντε φορές συνολικά** και όχι πέντε φορές για καθεμία από τις τριάντα υποψήφιες τιμές: κατασκευάζονται δηλαδή πέντε δομές αναζήτησης αντί για εκατόν πενήντα.

Ο βρόχος των **γραμμών 13–16** αξιοποιεί τα αποθηκευμένα αποτελέσματα. Για κάθε τιμή `k`, η έκφραση `nn[:, :k]` περικόπτει απλώς τον ήδη υπολογισμένο πίνακα γειτόνων στις `k` πρώτες στήλες, και η ψηφοφορία εκτελείται διανυσματικά. Η **γραμμή 17** επιστρέφει την τιμή με τη μεγαλύτερη μέση ορθότητα επικύρωσης, μαζί με ολόκληρη την καμπύλη `cv_accs`, η οποία χρησιμοποιείται για διαγνωστικούς σκοπούς. Παρά τη βελτιστοποίηση αυτή, ο αλγόριθμος παραμένει ο ακριβότερος από τους έξι, όπως τεκμηριώνεται στην ενότητα 7.3.

6.7 Πρωτόκολλο μέτρησης χρόνου: η συνάρτηση `time_uniformly`

Η μέτρηση των χρόνων εκτέλεσης παρουσίασε ένα πρακτικό πρόβλημα που απαιτούσε ειδικό χειρισμό. Στο λειτουργικό σύστημα των πειραμάτων η συνάρτηση `time.process_time()` έχει διακριτική ικανότητα περίπου 15,6 χιλιοστών του δευτερολέπτου, με αποτέλεσμα ολόκληρος ο αγωγός για μικρά σύνολα, όπως το Iris με 120 σημεία εκπαίδευσης, να μετράται ως μηδέν. Η λύση που υιοθετήθηκε υλοποιείται από τη συνάρτηση του Σχήματος 6.6.

```

experiment.py — time_uniformly()

1  def time_uniformly(pipelines):
2      """
3      Εκτελεί κάθε αγωγή ενός συνόλου δεδομένων και επιστρέφει
4      {όνομα: (αποτέλεσμα, δευτερόλεπτα CPU)} μαζί με τον αριθμό επαναλήψεων.
5      Ο κοινός αριθμός επαναλήψεων κρατά τους χρόνους συγκρίσιμους.
6      """
7      raw = {}
8      for name, fn in pipelines.items():
9          t0 = time.process_time()
10         raw[name] = (fn(), time.process_time() - t0)
11
12     times = [t for _, t in raw.values()]
13     t_min, t_max = min(times), max(times)
14     reps = MAX_REPEATS if t_min <= 0 else min(MAX_REPEATS,
15                                             int(np.ceil(MIN_TIMED / t_min)))
16     if t_max > 0:
17         # φραγή του συνολικού χρόνου
18         reps = min(reps, max(1, int(TIME_BUDGET / t_max)))
19     if reps <= 1:
20         return raw, 1
21
22     out = {}
23     for name, fn in pipelines.items():
24         t0 = time.process_time()
25         for _ in range(reps):
26             fn()
27         out[name] = (raw[name][0], (time.process_time() - t0) / reps)
28     return out, reps

```

Σχήμα 6.6: Η συνάρτηση `time_uniformly`, που υλοποιεί το πρωτόκολλο μέτρησης χρόνου με κοινό αριθμό επαναλήψεων.

Ο βρόχος των **γραμμών 8–10** αποτελεί το πρώτο πέρασμα: κάθε αγωγός εκτελείται μία φορά και καταγράφεται ο χρόνος του. Οι **γραμμές 13–15** υπολογίζουν τον αριθμό επαναλήψεων με βάση τον **ταχύτερο** αγωγό: αν αυτός μετρηθεί κάτω από το κατώφλι `MIN_TIMED`, η μέτρηση βρίσκεται στην ανάλυση του ρολογιού και απαιτείται επανάληψη. Οι **γραμμές 16–17** φράσσουν τον συνολικό χρόνο με βάση τον βραδύτερο αγωγό, ώστε τα μεγάλα σύνολα δεδομένων να μην οδηγήσουν σε υπερβολικό χρόνο εκτέλεσης.

Η λεπτομέρεια ότι ο αριθμός επαναλήψεων είναι **κοινός** για όλους τους αλγόριθμους του ίδιου συνόλου δεδομένων (γραμμές 21–26) δεν είναι διακοσμητική, αλλά αποτελεί προϋπόθεση εγκυρότητας. Αν κάθε αλγόριθμος επαναλαμβανόταν ανεξάρτητα μέχρι να ξεπεράσει το κατώφλι, τότε ένας ταχύς αλγόριθμος θα μετριοταν ως μέσος όρος πολλών εκτελέσεων, με θερμές κρυφές μνήμες, ενώ ένας βραδύτερος θα μετριοταν ως μία μεμονωμένη ψυχρή εκτέλεση. Η σύγκριση θα ήταν συστηματικά μεροληπτική υπέρ του δεύτερου. Ο αριθμός επαναλήψεων που χρησιμοποιήθηκε καταγράφεται για κάθε σύνολο δεδομένων στο αρχείο καταγραφής και κυμαίνεται από 1, για τα μεγάλα σύνολα, έως 16 για το Iris.

6.8 Επαλήθευση της ορθότητας

Η υλοποίηση είναι πλήρως διανυσματοποιημένη για λόγους απόδοσης, γεγονός που την καθιστά ταχύτερη αλλά και δυσκολότερη στην επιθεώρηση. Για τον λόγο αυτόν αναπτύχθηκε **ανεξάρτητη υλοποίηση αναφοράς**, η οποία υπολογίζει τα ίδια μεγέθη με τον προφανή και αργό τρόπο: ρητός

υπολογισμός όλων των αποστάσεων, ταξινόμηση, και ρητή καταμέτρηση ψήφων ανά σημείο και ανά τιμή k , με χρήση δομών της τυπικής βιβλιοθήκης της Python αντί για πράξεις πινάκων. Η σύγκριση των δύο υλοποιήσεων έγινε σε τέσσερα σύνολα δεδομένων και στις δύο συνθήκες θορύβου, για όλες τις παραλλαγές.

Η επαλήθευση ανέδειξε ένα εύρημα που αξίζει να καταγραφεί. Στα σύνολα Pima, Ecoli και Iris η συμφωνία των δύο υλοποιήσεων είναι απόλυτη, ενώ στο Balance Scale εμφανίζονται αποκλίσεις. Η αιτία δεν είναι σφάλμα αλλά **ισοπαλία αποστάσεων**: τα χαρακτηριστικά του Balance Scale είναι κβαντισμένα σε πέντε μόνο διακριτές τιμές, με αποτέλεσμα το 100 % των σημείων να έχει τουλάχιστον μία ισοπαλία απόστασης μέσα στους τριάντα πρώτους γείτονές του. Όταν πολλά σημεία ισαπέχουν, η κατάταξή τους είναι μαθηματικά απροσδιόριστη και οι δύο υλοποιήσεις μπορούν νόμιμα να επιλέξουν διαφορετική σειρά.

Για να επιβεβαιωθεί η ερμηνεία αυτή, η σύγκριση επαναλήφθηκε περιοριζόμενη στα σημεία των οποίων οι τριάντα πρώτες αποστάσεις δεν περιέχουν καμία ισοπαλία. Σε αυτά τα σημεία η συμφωνία των δύο υλοποιήσεων είναι **100 % σε όλες τις παραλλαγές και σε όλα τα σύνολα**. Το Balance Scale δεν διαθέτει κανένα τέτοιο σημείο, γεγονός που εξηγεί πλήρως τη συμπεριφορά του. Το συμπέρασμα είναι διπλό: αφενός η υλοποίηση είναι ορθή, αφετέρου τα αποτελέσματα για το Balance Scale, και σε μικρότερο βαθμό για το Iris, όπου το 68 % των σημείων εμφανίζει ισοπαλίες, εξαρτώνται εν μέρει από τη σειρά κατάταξης ισαπεχόντων γειτόνων και όχι μόνο από τον αλγόριθμο.

Πέραν της σύγκρισης με την υλοποίηση αναφοράς, ελέγχθηκε και ο κανόνας παραθύρου έναντι συνθετικών περιπτώσεων με γνωστό αναμενόμενο αποτέλεσμα, κατασκευασμένων ώστε να καλύπτουν τα ακραία σενάρια: σημείο που πετυχαίνει μόνο στην πρώτη τριάδα, σημείο που πετυχαίνει μόνο στην τελευταία, σημείο του οποίου η περιοχή επιτυχίας δεν ευθυγραμμίζεται με το πλέγμα των παραθύρων για συγκεκριμένο βήμα, και σημείο που δεν πετυχαίνει ποτέ. Οι περιπτώσεις αυτές αντιστοιχούν ακριβώς στα σημεία x_1 έως x_5 του παραδείγματος του Πίνακα 4.1 και όλες επαληθεύτηκαν. Ο πλήρης κώδικας των συναρτήσεων που παρουσιάστηκαν παρατίθεται στο Παράρτημα Α.

Κεφάλαιο 7ο: Πειραματική μελέτη

Το κεφάλαιο αυτό παρουσιάζει την πειραματική μελέτη της εργασίας. Περιγράφονται πρώτα τα σύνολα δεδομένων που χρησιμοποιήθηκαν και οι ιδιαιτερότητές τους, στη συνέχεια το πειραματικό περιβάλλον, δηλαδή ο τρόπος εισαγωγής του θορύβου, οι μετρικές αξιολόγησης, το υλικό και το λογισμικό και το πρωτόκολλο μέτρησης του χρόνου, και τέλος τα πειραματικά αποτελέσματα μαζί με την ερμηνεία τους και τους περιορισμούς τους. Συνολικά εκτελέστηκαν 168 πειράματα, δηλαδή έξι αλγόριθμοι επί δεκατέσσερα σύνολα δεδομένων επί δύο συνθήκες θορύβου.

7.1 Σύνολα δεδομένων

Η αξιολόγηση πραγματοποιήθηκε σε δεκατέσσερα σύνολα δεδομένων, τα οποία προέρχονται από το αποθετήριο **KEEL** (Knowledge Extraction based on Evolutionary Learning) [27] και, στην πλειονότητά τους, έλκουν την καταγωγή τους από το αποθετήριο UCI [28]. Το KEEL επιλέχθηκε διότι διαθέτει τα σύνολα ήδη κανονικοποιημένα και σε έτοιμες διαμερίσεις διασταυρούμενης επικύρωσης, στοιχείο που διευκολύνει την αναπαραγωγιμότητα και καθιστά τα αποτελέσματα άμεσα συγκρίσιμα με τη βιβλιογραφία των μεθόδων εγγύτερων γειτόνων, η οποία χρησιμοποιεί ευρέως τα ίδια σύνολα.

Τα χαρακτηριστικά των συνόλων συνοψίζονται στον Πίνακα 7.1. Η επιλογή τους δεν είναι τυχαία: καλύπτουν εύρος μεγέθους από 150 έως σχεδόν 58.000 σημεία, διαστατικότητα από 2 έως 40 χαρακτηριστικά και πλήθος κλάσεων από 2 έως 26. Περιλαμβάνουν τεχνητά σύνολα με ελεγχόμενη γεωμετρία (Banana, Twonorm), σύνολα αναγνώρισης προτύπων (Letter, Pen Digits, Texture), σύνολα από τις φυσικές και βιολογικές επιστήμες (MAGIC, Ecoli, Yeast), ιατρικά δεδομένα (Pima) και δεδομένα τηλεπισκόπησης (Landsat). Η ποικιλομορφία αυτή είναι απαραίτητη ώστε τα συμπεράσματα να μην εξαρτώνται από τα ιδιαίτερα χαρακτηριστικά ενός συγκεκριμένου τύπου προβλήματος.

Πίνακας 7.1: Τα δεκατέσσερα σύνολα δεδομένων της πειραματικής αξιολόγησης.

Κωδ.	Σύνολο δεδομένων	Όνομα στο KEEL	Σύνολο	Train	Test	Χαρακτ.	Κλάσεις
bl	Balance Scale	balance	625	500	125	4	3
bn	Banana	banana	5.300	4.240	1.060	2	2
ecl	Ecoli	ecoli	336	269	67	7	8
iris	Iris	iris	150	120	30	4	3
lir	Letter Image Recognition	letter	20.000	16.000	4.000	16	26
ls	Landsat Satellite	satimage	6.435	5.148	1.287	36	6
mgt	MAGIC Gamma Telescope	magic	19.020	15.216	3.804	10	2
pd	Pen Digits	penbased	10.992	8.794	2.198	16	10
ph	Phoneme	phoneme	5.404	4.324	1.080	5	2
pm	Pima Indians Diabetes	pima	768	615	153	8	2
sh	Shuttle	shuttle	57.999	46.400	11.599	9	7
tn	Twonorm	twonorm	7.400	5.920	1.480	20	2
txr	Texture	texture	5.500	4.400	1.100	40	11
ys	Yeast	yeast	1.484	1.188	296	8	10

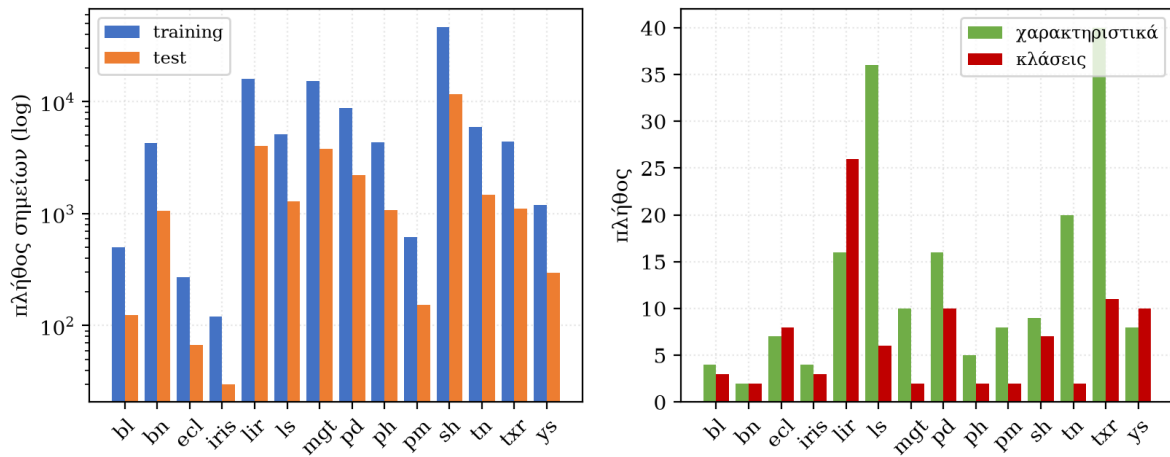
Ο Πίνακας 7.2 περιγράφει συνοπτικά το πρόβλημα που πραγματεύεται κάθε σύνολο, καθώς και τις ιδιαιτερότητες που έχουν σημασία για την ερμηνεία των αποτελεσμάτων.

Πίνακας 7.2: Σύνομη περιγραφή των συνόλων δεδομένων του αποθετηρίου KEEL.

Σύνολο δεδομένων	Περιγραφή και ιδιαιτερότητες
Balance Scale	Προσομοίωση ψυχολογικών πειραμάτων: η κατάσταση μιας ζυγαριάς (κλίση αριστερά, κλίση δεξιά, ισορροπία) προβλέπεται από το βάρος και την απόσταση εκατέρωθεν. Τα

Σύνολο δεδομένων	Περιγραφή και ιδιαιτερότητες
	χαρακτηριστικά είναι κβαντισμένα σε πέντε μόνο διακριτές τιμές, γεγονός που προκαλεί εκτεταμένες ισοπαλίες απόστασης.
Banana	Τεχνητό διδιάστατο σύνολο δύο κλάσεων σε σχήμα μπανάνας, με έντονα μη γραμμικό όριο και επικάλυψη. Χρησιμοποιείται ευρέως για την οπτικοποίηση ορίων απόφασης, όπως στα Σχήματα 1.2 και 7.3.
Ecoli	Πρόβλεψη της θέσης εντοπισμού πρωτεϊνών στο βακτήριο E. coli από βιοχημικά χαρακτηριστικά της αλληλουχίας. Έντονη ανισορροπία: η μικρότερη κλάση περιλαμβάνει μόλις δύο σημεία εκπαίδευσης.
Iris	Το κλασικό σύνολο του Fisher: τρία είδη ίριδας διακρίνονται από μήκος και πλάτος σεπάλων και πετάλων. Πολύ μικρό μέγεθος, με μόλις τριάντα σημεία ελέγχου.
Letter Image Recognition	Αναγνώριση των είκοσι έξι κεφαλαίων λατινικών γραμμάτων από δεκαέξι στατιστικά χαρακτηριστικά εικόνας. Το σύνολο με το μεγαλύτερο πλήθος κλάσεων.
Landsat Satellite	Ταξινόμηση τύπων εδαφοκάλυψης από φασματικές τιμές πολυφασματικών δορυφορικών εικόνων Landsat, με τριάντα έξι χαρακτηριστικά ανά εικονοστοιχείο.
MAGIC Gamma Telescope	Διάκριση καταιγισμών ακτίνων γάμμα από αδρονικό υπόβαθρο σε τηλεσκόπιο Cherenkov. Δυσχερές πρόβλημα με σημαντική επικάλυψη κλάσεων.
Pen Digits	Αναγνώριση χειρόγραφων ψηφίων από δειγματοληπτημένες συντεταγμένες της κίνησης της γραφίδας. Καλά διαχωρισμένες κλάσεις και πολύ υψηλή ορθότητα.
Phoneme	Διάκριση ρινικών από στοματικά φωνήεντα με βάση πέντε αρμονικά χαρακτηριστικά του φάσματος της φωνής.
Pima Indians Diabetes	Πρόβλεψη διαβήτη σε γυναίκες του πληθυσμού Pima από οκτώ ιατρικές μετρήσεις. Μικρό σύνολο με έντονη επικάλυψη κλάσεων και χαμηλή ορθότητα για όλους τους κατηγοριοποιητές.
Shuttle	Δεδομένα τηλεμετρίας διαστημικού λεωφορείου. Το μεγαλύτερο σύνολο της μελέτης, με ακραία ανισορροπία: η κυρίαρχη κλάση καλύπτει περίπου το 80 % των σημείων.
Twonorm	Τεχνητό σύνολο είκοσι διαστάσεων: δύο κλάσεις παράγονται από πολυμεταβλητές κανονικές κατανομές με διαφορετικό μέσο και ίδια διασπορά.
Texture	Αναγνώριση έντεκα τύπων υφής από σαράντα χαρακτηριστικά που προκύπτουν από ροπές τέταρτης τάξης. Το σύνολο με τη μεγαλύτερη διαστατικότητα.
Yeast	Πρόβλεψη της θέσης εντοπισμού πρωτεϊνών σε κύτταρα ζύμης. Δέκα κλάσεις με έντονη ανισορροπία και εκτεταμένη επικάλυψη· το δυσκολότερο σύνολο της μελέτης.

Όλα τα σύνολα διατίθενται ήδη κανονικοποιημένα, με κάθε χαρακτηριστικό κλιμακωμένο στο διάστημα $[0, 1]$, και με τις ετικέτες κλάσης κωδικοποιημένες ως διαδοχικοί ακέραιοι ξεκινώντας από το μηδέν. Η κανονικοποίηση είναι, όπως αναλύθηκε στην ενότητα 1.2, προϋπόθεση για τη σωστή λειτουργία της ευκλείδειας απόστασης. Ο διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο ελέγχου δίνεται εκ των προτέρων και αντιστοιχεί σε αναλογία 80 % προς 20 %, δηλαδή στην πρώτη πτυχή μιας πενταπλής διαμέρισης. Ο διαχωρισμός αυτός διατηρήθηκε αυτούσιος για όλους τους αλγόριθμους και για τις δύο συνθήκες θορύβου, ώστε κάθε σύγκριση να αφορά ακριβώς τα ίδια δεδομένα. Στο Σχήμα 7.1 απεικονίζονται γραφικά τα μεγέθη, η διαστατικότητα και το πλήθος κλάσεων των συνόλων.



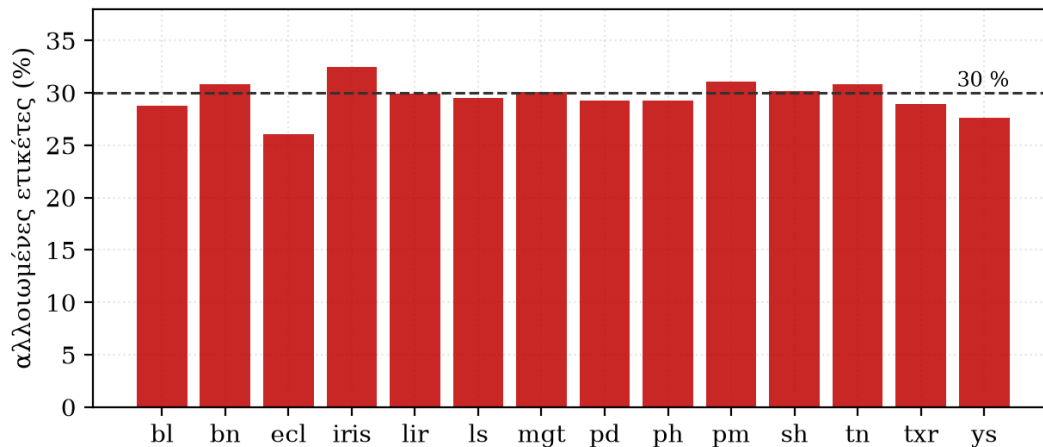
Σχήμα 7.1: Μέγεθος συνόλων εκπαίδευσης και ελέγχου σε λογαριθμική κλίμακα (αριστερά) και πλήθος χαρακτηριστικών και κλάσεων (δεξιά) για κάθε σύνολο δεδομένων.

Δύο από τα σύνολα παρουσιάζουν έντονη ανισορροπία κλάσεων που αξίζει να επισημανθεί, καθώς επηρεάζει την ερμηνεία των μετρικών. Στο Ecoli η μικρότερη κλάση περιλαμβάνει μόλις δύο σημεία εκπαίδευσης και στο Yeast τέσσερα, αριθμοί μικρότεροι από το πλήθος των πτυχών της διασταυρούμενης επικύρωσης. Στα δύο αυτά σύνολα η στρωματοποίηση δεν είναι εφικτή για τις συγκεκριμένες κλάσεις, γεγονός που καθιστά την εκτίμηση του καθολικού k λιγότερο αξιόπιστη. Ο περιορισμός αφορά αποκλειστικά τον αλγόριθμο καθολικού προσδιορισμού· οι παραλλαγές του AdaNN δεν χρησιμοποιούν διασταυρούμενη επικύρωση και επομένως δεν επηρεάζονται.

7.2 Πειραματικό περιβάλλον

7.2.1 Εισαγωγή θορύβου στις ετικέτες

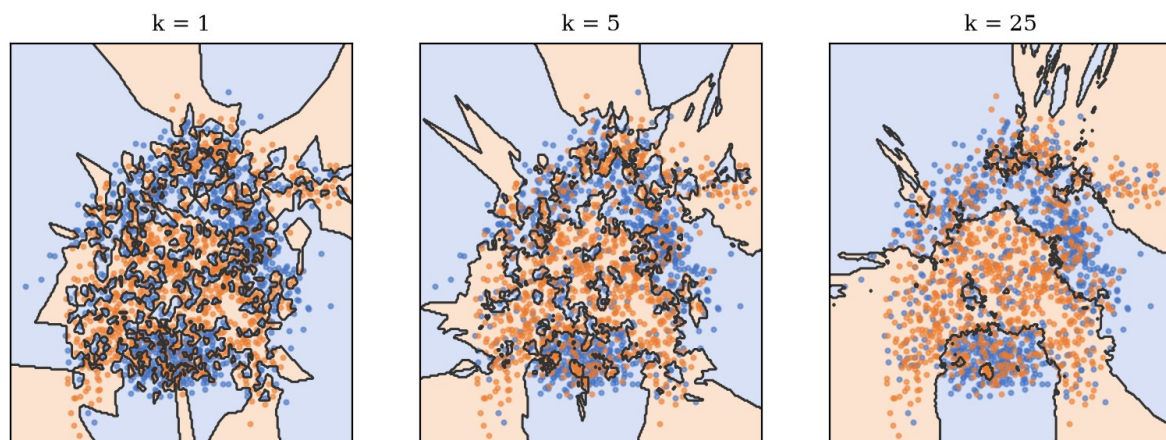
Κάθε σύνολο δεδομένων χρησιμοποιήθηκε σε δύο εκδοχές. Η πρώτη, που αναφέρεται ως **καθαρή** (clean), είναι η αρχική μορφή του συνόλου. Η δεύτερη, που αναφέρεται ως **θορυβώδης** (noisy), προκύπτει με τυχαία αλλοίωση περίπου του 30 % των ετικετών του συνόλου εκπαίδευσης. Ο θόρυβος ανήκει στην κατηγορία του ομοιόμορφου θορύβου ανεξάρτητου από τα δεδομένα, το απλούστερο και πλέον διαδεδομένο μοντέλο στην πειραματική βιβλιογραφία [21]. Το ακριβές ποσοστό ανά σύνολο κυμαίνεται από 26,0 % έως 32,5 % και απεικονίζεται στο Σχήμα 7.2.



Σχήμα 7.2: Ποσοστό αλλοιωμένων ετικετών στο σύνολο εκπαίδευσης κάθε συνόλου δεδομένων. Ο θόρυβος δεν αγγίζει καθόλου τα σύνολα ελέγχου.

Καθοριστικό στοιχείο του σχεδιασμού, το οποίο επαληθεύτηκε ρητά, είναι ότι **ο θόρυβος υπάρχει μόνο στο σύνολο εκπαίδευσης**. Τα σύνολα ελέγχου των δύο εκδοχών είναι απολύτως ταυτόσημα, με μηδενική διαφορά ετικετών, ενώ και τα διανύσματα χαρακτηριστικών παραμένουν αναλλοίωτα και στις δύο εκδοχές. Ο σχεδιασμός αυτός δεν είναι τυχαίος: επιτρέπει τη μέτρηση της ανθεκτικότητας κάθε αλγορίθμου έναντι διαφθαρμένης πληροφορίας εκπαίδευσης, διατηρώντας ταυτόχρονα ένα καθαρό και σταθερό κριτήριο αξιολόγησης. Αν ο θόρυβος υπήρχε και στο σύνολο ελέγχου, η πτώση της ορθότητας θα ανακατευόταν με την αδυναμία επίτευξης υψηλής επίδοσης έναντι εσφαλμένων ετικετών αναφοράς, και η ερμηνεία των αποτελεσμάτων θα ήταν αδύνατη.

Ο λόγος για τον οποίο ο θόρυβος αποτελεί κεντρικό στοιχείο του πειραματικού σχεδιασμού είναι ότι μετατοπίζει τη βέλτιστη τιμή του k προς τα πάνω. Ένα μεμονωμένο σημείο με εσφαλμένη ετικέτα δημιουργεί, για $k = 1$, έναν ολόκληρο θύλακα λανθασμένης κατηγοριοποίησης γύρω του, το μέγεθος του οποίου καθορίζεται από την πυκνότητα των γειτονικών σημείων. Αντιθέτως, για μεγαλύτερες τιμές k το εσφαλμένο σημείο συμμετέχει ως μία μόνο ψήφος ανάμεσα σε πολλές και υπερψηφίζεται από τους ορθά επισημειωμένους γείτονές του. Το Σχήμα 7.3 δείχνει το σύνολο Banana μετά την εισαγωγή θορύβου 30 % στις ετικέτες εκπαίδευσης και πρέπει να συγκριθεί απευθείας με το Σχήμα 1.2.



Σχήμα 7.3: Όριο απόφασης στο Banana με 30 % θόρυβο στις ετικέτες εκπαίδευσης. Για $k = 1$ το όριο έχει κατακερματιστεί πλήρως, ενώ για $k = 25$ ανακτάται σε μεγάλο βαθμό.

Η σύγκριση των δύο σχημάτων είναι αποκαλυπτική. Στα καθαρά δεδομένα η μετάβαση από $k = 1$ σε $k = 25$ βελτιώνει το όριο μόνο οριακά, καθώς ήδη για $k = 5$ η δομή έχει αποτυπωθεί ικανοποιητικά. Στα θορυβώδη δεδομένα, αντιθέτως, η μετάβαση αυτή είναι καθοριστική: για $k = 1$ το όριο είναι πρακτικά άχρηστο, ενώ για $k = 25$ ανακτάται σε μεγάλο βαθμό η αρχική δομή των δύο κλάσεων. Το γεγονός αυτό υποδεικνύει ότι κάθε μηχανισμός αυτόματης επιλογής του k οφείλει να είναι σε θέση να φτάνει σε αρκετά μεγάλες τιμές όταν οι συνθήκες το απαιτούν, παρατήρηση που θεμελιώνει άμεσα την Παραλλαγή 1.

7.2.2 Μετρικές αξιολόγησης

Η ορθότητα (accuracy) ορίζεται ως το ποσοστό των σωστά κατηγοριοποιημένων σημείων επί του συνόλου και δίνεται από τη σχέση (7.1). Παρότι είναι η πιο διαδεδομένη μετρική, είναι παραπλανητική σε προβλήματα με έντονη ανισορροπία κλάσεων: σε ένα σύνολο όπου το 99 % των σημείων ανήκει σε μία κλάση, ένας κατηγοριοποιητής που επιστρέφει πάντοτε αυτή την κλάση επιτυγχάνει ορθότητα 99 % χωρίς να έχει μάθει τίποτε [25].

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (7.1)$$

Για τον λόγο αυτόν συνυπολογίζονται η ακρίβεια (precision) και η ανάκληση (recall), οι οποίες δίνονται από τις σχέσεις (7.2) και (7.3). Η ακρίβεια εκφράζει το ποσοστό των προβλέψεων μιας κλάσης που ήταν όντως ορθές, ενώ η ανάκληση το ποσοστό των σημείων μιας κλάσης που εντοπίστηκαν. Οι δύο μετρικές βρίσκονται σε αντιστάθμιση μεταξύ τους και συνοψίζονται στο F-score, τον αρμονικό μέσο τους, που δίνεται από τη σχέση (7.4). Ο αρμονικός μέσος επιλέγεται αντί του αριθμητικού επειδή τιμωρεί αυστηρότερα τις περιπτώσεις όπου η μία από τις δύο μετρικές είναι πολύ χαμηλή.

$$Precision = TP / (TP + FP) \quad (7.2)$$

$$Recall = TP / (TP + FN) \quad (7.3)$$

$$F = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (7.4)$$

Σε προβλήματα με περισσότερες από δύο κλάσεις οι τρεις τελευταίες μετρικές υπολογίζονται ανά κλάση και στη συνέχεια συναθροίζονται. Η παρούσα εργασία χρησιμοποιεί τη **μακρο-συνάθροιση** (macro averaging), δηλαδή τον απλό μέσο όρο των τιμών όλων των κλάσεων, χωρίς στάθμιση με το πλήθος των σημείων κάθε κλάσης. Η επιλογή αυτή αποδίδει σε κάθε κλάση ίση βαρύτητα ανεξάρτητα από το μέγεθός της και είναι επομένως καταλληλότερη για σύνολα με έντονη ανισορροπία, όπως το Shuttle ή το Yeast [25]. Για τα δυαδικά προβλήματα χρησιμοποιείται ο συμβατικός δυαδικός ορισμός ως προς τη θετική κλάση. Πέραν των μετρικών ποιότητας μετρήθηκε και ο χρόνος CPU ολόκληρου του αγωγού κάθε αλγορίθμου.

7.2.3 Υλικό και λογισμικό

Όλα τα πειράματα εκτελέστηκαν στο ίδιο μηχάνημα, χωρίς παράλληλη εκτέλεση άλλων υπολογιστικά απαιτητικών εργασιών, ώστε οι μετρήσεις χρόνου να είναι συγκρίσιμες μεταξύ τους. Οι σχετικές λεπτομέρειες παρατίθενται στον Πίνακα 7.3. Οι απόλυτες τιμές χρόνου εξαρτώνται προφανώς από το υλικό και δεν έχουν αυτοτελή σημασία· εκείνο που ενδιαφέρει είναι οι **σχετικές** διαφορές μεταξύ των αλγορίθμων, οι οποίες παραμένουν σταθερές ανεξάρτητα από την πλατφόρμα.

Πίνακας 7.3: Το περιβάλλον εκτέλεσης των πειραμάτων.

Στοιχείο	Τιμή
Λειτουργικό σύστημα	Microsoft Windows 11

Στοιχείο	Τιμή
Γλώσσα προγραμματισμού	Python 3.13
Βιβλιοθήκη μηχανικής μάθησης	scikit-learn
Βιβλιοθήκη αριθμητικών υπολογισμών	NumPy
Μέτρο χρόνου	time.process_time() (χρόνος CPU διεργασίας)
Δομή αναζήτησης γειτόνων	αυτόματη επιλογή (k-d tree / ball tree / brute)
Πλήθος πτυχών CV	5, στρωματοποιημένες
Σπόρος τυχαιότητας	42 (σταθερός)

7.2.4 Πρωτόκολλο μέτρησης χρόνου

Ο χρόνος CPU μετράται με τη συνάρτηση `time.process_time()`, η οποία καταγράφει τον χρόνο επεξεργαστή που καταναλώνει η διεργασία και όχι τον πραγματικό χρόνο ρολογιού, αποκλείοντας έτσι την επίδραση άλλων διεργασιών του συστήματος. Η μέτρηση καλύπτει **ολόκληρο** τον αγωγό κάθε αλγορίθμου, δηλαδή τόσο τη φάση εκπαίδευσης όσο και τη φάση ελέγχου. Κάθε αλγόριθμος εκτελείται **ανεξάρτητα**, χωρίς να επαναχρησιμοποιεί ενδιάμεσα αποτελέσματα άλλων, ακόμη και όταν αυτό θα ήταν υπολογιστικά συμφέρον· διαφορετικά, οι χρόνοι δεν θα ήταν συγκρίσιμοι μεταξύ τους. Η αντιμετώπιση του προβλήματος της ανάλυσης του μετρητή, μέσω κοινού αριθμού επαναλήψεων ανά σύνολο δεδομένων, περιγράφηκε αναλυτικά στην ενότητα 6.8.

7.3 Πειραματικά αποτελέσματα

Η παρουσίαση των αποτελεσμάτων ακολουθεί τη σειρά: ορθότητα και F-score στα καθαρά δεδομένα, οι ίδιες μετρικές στα θορυβώδη δεδομένα, η μεταβολή που επιφέρει ο θόρυβος, η κατανομή των τιμών k που ανατίθενται, η καθολική τιμή k , το υπολογιστικό κόστος και, τέλος, η συνολική κατάταξη των αλγορίθμων. Ακολουθούν η ερμηνεία των ευρημάτων, η σύγκριση με τη βιβλιογραφία και οι περιορισμοί της μελέτης. Σε κάθε πίνακα η τελευταία γραμμή δίνει τον μέσο όρο των δεκατεσσάρων συνόλων.

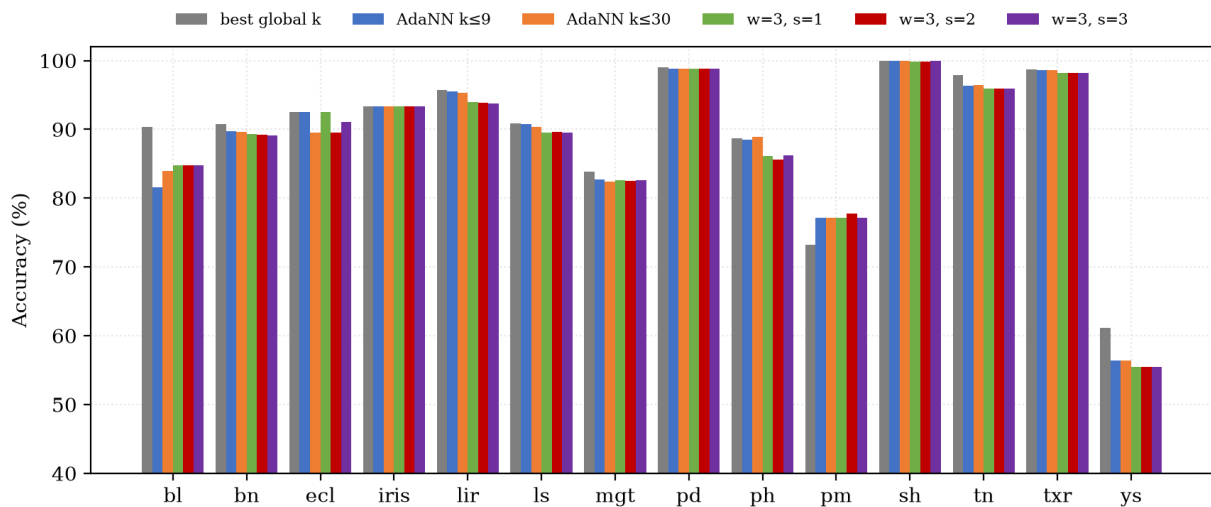
7.3.1 Καθαρά δεδομένα: ορθότητα

Ο Πίνακας 7.4 και το Σχήμα 7.4 παρουσιάζουν την ορθότητα στα καθαρά δεδομένα. Η πρώτη παρατήρηση είναι ότι οι διαφορές μεταξύ των έξι αλγορίθμων είναι μικρές: οι μέσοι όροι κυμαίνονται από 88,19 % έως 89,73 %, δηλαδή σε εύρος μικρότερο της μιάμισης ποσοστιαίας μονάδας. Ο καθολικός προσδιορισμός του k επιτυγχάνει τον υψηλότερο μέσο όρο, ενώ ανάμεσα στις παραλλαγές του AdaNN την καλύτερη επίδοση έχει ο **βασικός** αλγόριθμος με $\max k = 9$, με 88,71 %.

Πίνακας 7.4: Ορθότητα (%) στα καθαρά δεδομένα.

ds	best global k	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	90,40	81,60	84,00	84,80	84,80	84,80
bn	90,75	89,72	89,62	89,34	89,25	89,15
ecl	92,54	92,54	89,55	92,54	89,55	91,04
iris	93,33	93,33	93,33	93,33	93,33	93,33
lir	95,75	95,53	95,35	93,92	93,85	93,77
ls	90,91	90,75	90,37	89,51	89,67	89,51
mgt	83,89	82,70	82,39	82,65	82,54	82,60
pd	99,04	98,82	98,82	98,82	98,82	98,82
ph	88,70	88,52	88,89	86,11	85,65	86,20
pm	73,20	77,12	77,12	77,12	77,78	77,12
sh	99,96	99,91	99,91	99,90	99,90	99,91
tn	97,84	96,35	96,42	95,88	95,88	95,88

ds	best global k	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
txr	98,73	98,64	98,64	98,18	98,18	98,18
ys	61,15	56,42	56,42	55,41	55,41	55,41
M.O.	89,73	88,71	88,63	88,39	88,19	88,27



Σχήμα 7.4: Ορθότητα ανά σύνολο δεδομένων και αλγόριθμο στα καθαρά δεδομένα.

Καμία από τις τέσσερις παραλλαγές δεν βελτιώνει τον βασικό αλγόριθμο κατά μέσο όρο. Η Παραλλαγή 1, που απλώς επεκτείνει το εύρος αναζήτησης, καταγράφει 88,63 %, οριακά χαμηλότερα από τη βάση. Οι τρεις παραλλαγές παραθύρου κινούνται επίσης ελαφρώς χαμηλότερα, με τιμές 88,39 %, 88,19 % και 88,27 % αντίστοιχα. Η εικόνα επιβεβαιώνεται και από τη σύγκριση ανά σύνολο δεδομένων, που παρουσιάζεται στον Πίνακα 7.5: κάθε παραλλαγή χάνει από τη βάση σε περισσότερα σύνολα από όσα κερδίζει.

Πίνακας 7.5: Νίκες, ισοπαλίες και ήττες κάθε παραλλαγής έναντι του βασικού AdaNN ($\max k = 9$) ως προς την ορθότητα, στα καθαρά δεδομένα.

Παραλλαγή	Νίκες	Ισοπαλίες	Ήττες
AdaNN k≤30	3	6	5
w=3, s=1	1	4	9
w=3, s=2	2	2	10
w=3, s=3	1	4	9

Υπάρχουν ωστόσο και εξαιρέσεις που αξίζει να σημειωθούν. Στο Balance Scale όλες οι παραλλαγές με $\max k = 30$ υπερτερούν αισθητά της βάσης, με την ορθότητα να ανεβαίνει από 81,60 % σε 84,80 %. Στο Ecoli η Παραλλαγή 2 φτάνει το 92,54 %, εξισούμενη με τον καθολικό προσδιορισμό. Στο Pima όλες οι παραλλαγές ξεπερνούν τόσο τη βάση όσο και τον καθολικό προσδιορισμό, ο οποίος καταγράφει εδώ τη χειρότερη σχετική του επίδοση με 73,20 % έναντι 77,12 % της βάσης. Πρόκειται για σύνολα μικρού έως μεσαίου μεγέθους με σημαντική επικάλυψη κλάσεων, δηλαδή ακριβώς τις συνθήκες υπό τις οποίες η τοπική προσαρμογή του k έχει νόημα.

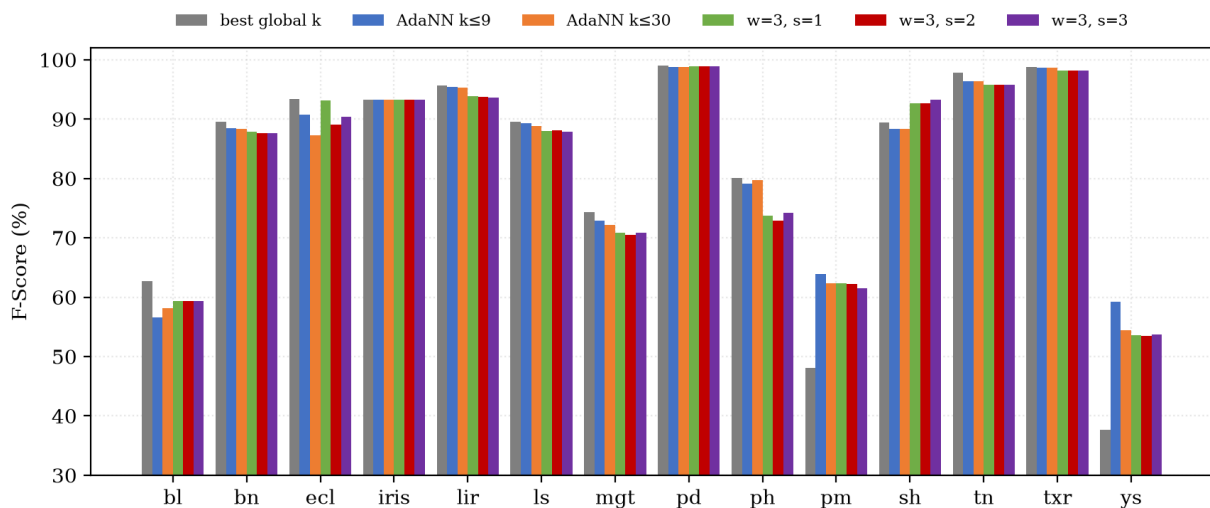
7.3.2 Καθαρά δεδομένα: F-score

Ο Πίνακας 7.6 και το Σχήμα 7.5 παρουσιάζουν το F-score, το οποίο δίνει διαφορετική εικόνα από την ορθότητα. Εδώ ο **βασικός AdaNN** έχει τον υψηλότερο μέσο όρο με 83,66 %, ξεπερνώντας τον καθολικό προσδιορισμό που περιορίζεται σε 82,10 %. Η αντιστροφή αυτή είναι ενδεικτική και οφείλεται στα σύνολα με έντονη ανισορροπία κλάσεων. Στο Yeast, για παράδειγμα, ο καθολικός προσδιορισμός

επιλέγει $k = 30$ και επιτυγχάνει ορθότητα 61,15 %, υψηλότερη από κάθε παραλλαγή, αλλά F-score μόλις 37,66 % έναντι 59,24 % της βάσης.

Πίνακας 7.6: F-score (%) στα καθαρά δεδομένα.

ds	best global k	AdaNN $k \leq 9$	AdaNN $k \leq 30$	w=3, s=1	w=3, s=2	w=3, s=3
bl	62,74	56,60	58,18	59,34	59,34	59,34
bn	89,53	88,42	88,32	87,84	87,64	87,65
ecl	93,43	90,79	87,22	93,16	89,04	90,37
iris	93,27	93,27	93,27	93,27	93,27	93,27
lir	95,66	95,44	95,25	93,83	93,76	93,68
ls	89,57	89,30	88,84	87,98	88,15	87,90
mgt	74,34	72,92	72,13	70,90	70,52	70,89
pd	99,06	98,83	98,84	98,85	98,85	98,85
ph	80,07	79,12	79,73	73,78	72,85	74,18
pm	48,10	63,92	62,37	62,37	62,22	61,54
sh	89,45	88,40	88,40	92,66	92,66	93,24
tn	97,83	96,33	96,40	95,76	95,76	95,76
txr	98,75	98,68	98,68	98,22	98,22	98,22
ys	37,66	59,24	54,38	53,59	53,47	53,67
M.O.	82,10	83,66	83,00	82,97	82,55	82,75



Σχήμα 7.5: F-score ανά σύνολο δεδομένων και αλγόριθμο στα καθαρά δεδομένα.

Η εξήγηση είναι εκείνη που προαναγγέλθηκε στην ενότητα 2.6: μια μεγάλη καθολική τιμή k εξαφανίζει τις μικρές κλάσεις. Με τριάντα γείτονες, μια κλάση που εκπροσωπείται από ελάχιστα σημεία δεν μπορεί ποτέ να συγκεντρώσει πλειοψηφία και επομένως δεν προβλέπεται ποτέ. Η ορθότητα δεν το αντιλαμβάνεται, καθώς τα σημεία αυτά είναι ελάχιστα, αλλά το μακρο-συναθροισμένο F-score, που αποδίδει ίση βαρύτητα σε κάθε κλάση, καταγράφει την απώλεια. Το εύρημα αυτό αναδεικνύει ένα πραγματικό πλεονέκτημα της τοπικής προσαρμογής: η ανάθεση μικρών τιμών k στις περιοχές των μικρών κλάσεων τούς επιτρέπει να επιβιώσουν.

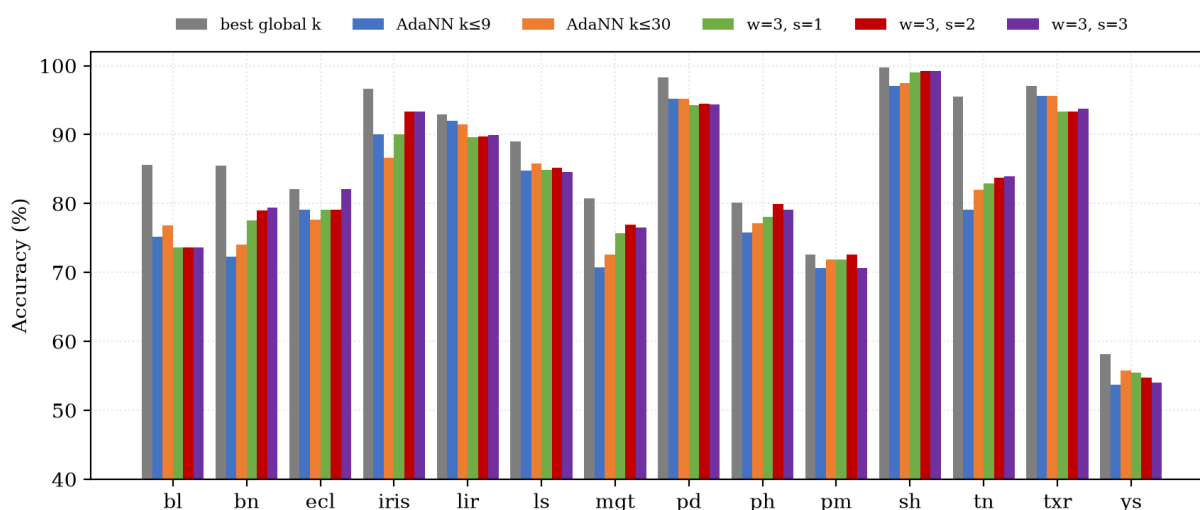
7.3.3 Θορυβώδη δεδομένα: ορθότητα

Ο Πίνακας 7.7 και το Σχήμα 7.6 παρουσιάζουν την ορθότητα στα θορυβώδη δεδομένα, όπου η εικόνα **αντιστρέφεται πλήρως**. Ο βασικός AdaNN με $\max k = 9$ πέφτει στο 80,81 % και γίνεται ο **χειρότερος** από τους έξι αλγορίθμους. Όλες οι παραλλαγές τον ξεπερνούν: η Παραλλαγή 1 φτάνει το 81,44 %, η

Παραλλαγή 2 το 81,81 %, η Παραλλαγή 3 το 82,49 % και η Παραλλαγή 4 το 82,46 %. Η βελτίωση της καλύτερης παραλλαγής έναντι της βάσης ανέρχεται σε 1,68 ποσοστιαίες μονάδες.

Πίνακας 7.7: Ορθότητα (%) στα θορυβώδη δεδομένα.

ds	best global k	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	85,60	75,20	76,80	73,60	73,60	73,60
bn	85,47	72,26	74,06	77,55	78,96	79,43
ecl	82,09	79,10	77,61	79,10	79,10	82,09
iris	96,67	90,00	86,67	90,00	93,33	93,33
lir	92,92	92,00	91,53	89,62	89,70	89,90
ls	88,97	84,77	85,78	84,85	85,24	84,62
mgt	80,73	70,77	72,63	75,71	76,95	76,55
pd	98,27	95,22	95,22	94,27	94,45	94,36
ph	80,09	75,83	77,13	78,06	79,91	79,07
pm	72,55	70,59	71,90	71,90	72,55	70,59
sh	99,72	97,09	97,47	99,04	99,19	99,19
tn	95,47	79,12	81,96	82,91	83,78	83,99
txr	97,09	95,64	95,64	93,36	93,36	93,73
ys	58,11	53,72	55,74	55,41	54,73	54,05
M.O.	86,70	80,81	81,44	81,81	82,49	82,46



Σχήμα 7.6: Ορθότητα ανά σύνολο δεδομένων και αλγόριθμο στα θορυβώδη δεδομένα.

Η υπεροχή δεν είναι αποτέλεσμα λίγων ακραίων τιμών αλλά συστηματική, όπως δείχνει ο Πίνακας 7.8. Η Παραλλαγή 1 και η Παραλλαγή 3 κερδίζουν τη βάση σε εννέα από τα δεκατέσσερα σύνολα, ενώ οι Παραλλαγές 2 και 4 σε οκτώ. Η μεγαλύτερη μεμονωμένη βελτίωση καταγράφεται στο Banana, όπου η ορθότητα ανεβαίνει από 72,26 % στη βάση σε 79,43 % με την Παραλλαγή 4, δηλαδή κατά 7,17 ποσοστιαίες μονάδες. Σημαντικές βελτιώσεις παρατηρούνται επίσης στο Iris, από 90,00 % σε 93,33 %, και στο Ecoli, όπου η Παραλλαγή 4 φτάνει το 82,09 % και εξισώνεται με τον καθολικό προσδιορισμό.

Πίνακας 7.8: Νίκες, ισοπαλίες και ήττες κάθε παραλλαγής έναντι του βασικού AdaNN (maxk = 9) ως προς την ορθότητα, στα θορυβώδη δεδομένα.

Παραλλαγή	Νίκες	Ισοπαλίες	Ήττες
AdaNN k≤30	9	2	3
w=3, s=1	8	2	4
w=3, s=2	9	1	4
w=3, s=3	8	1	5

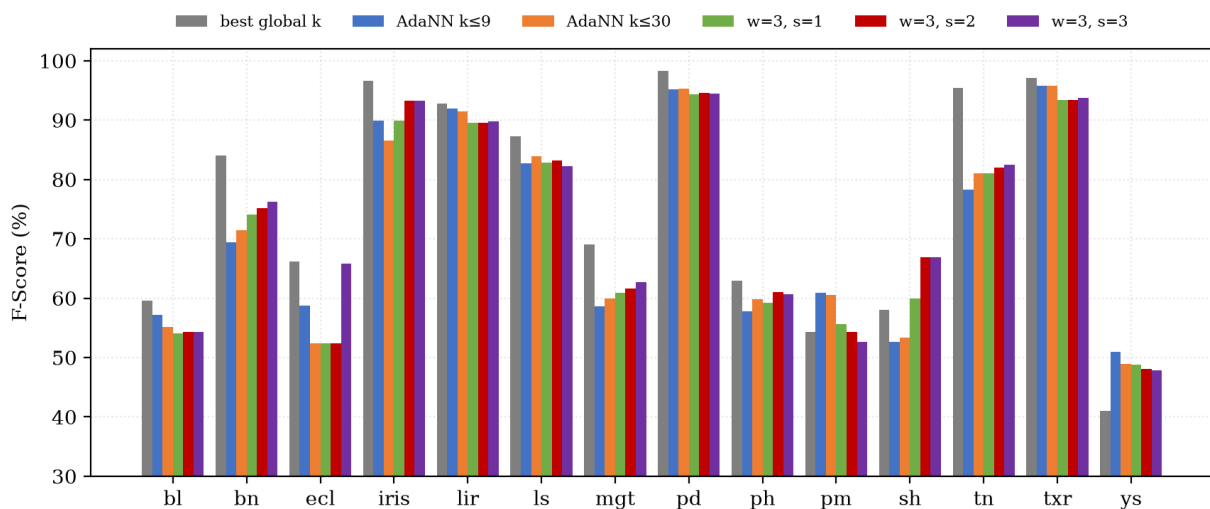
Παρά τη βελτίωση αυτή, ο καθολικός προσδιορισμός του k κυριαρχεί στα θορυβώδη δεδομένα με μέση ορθότητα 86,70 %, υπεροχή άνω των τεσσάρων ποσοστιαίων μονάδων έναντι της καλύτερης παραλλαγής. Όπως θα φανεί στη συνολική κατάταξη, κατατάσσεται πρώτος και στα δεκατέσσερα σύνολα δεδομένων χωρίς εξαίρεση. Το εύρημα αυτό είναι από τα σημαντικότερα της εργασίας και ερμηνεύεται αναλυτικά παρακάτω.

7.3.4 Θορυβώδη δεδομένα: F-score

Ο Πίνακας 7.9 και το Σχήμα 7.7 παρουσιάζουν το F-score στα θορυβώδη δεδομένα. Η κατάταξη των παραλλαγών ακολουθεί εκείνη της ορθότητας, με τη βάση στο 71,44 % και την Παραλλαγή 4 να επιτυγχάνει το υψηλότερο ποσοστό μεταξύ των AdaNN, 73,08 %. Αξιοσημείωτο είναι ότι εδώ ο καθολικός προσδιορισμός διατηρεί την υπεροχή του και στο F-score, με 75,92 %, σε αντίθεση με ό,τι συνέβαινε στα καθαρά δεδομένα. Ο θόρυβος εξαφανίζει το πλεονέκτημα που είχε η τοπική προσαρμογή ως προς τις μικρές κλάσεις, διότι οι μικρές κλάσεις είναι ακριβώς εκείνες που διαφθείρονται αναλογικά περισσότερο.

Πίνακας 7.9: F-score (%) στα θορυβώδη δεδομένα.

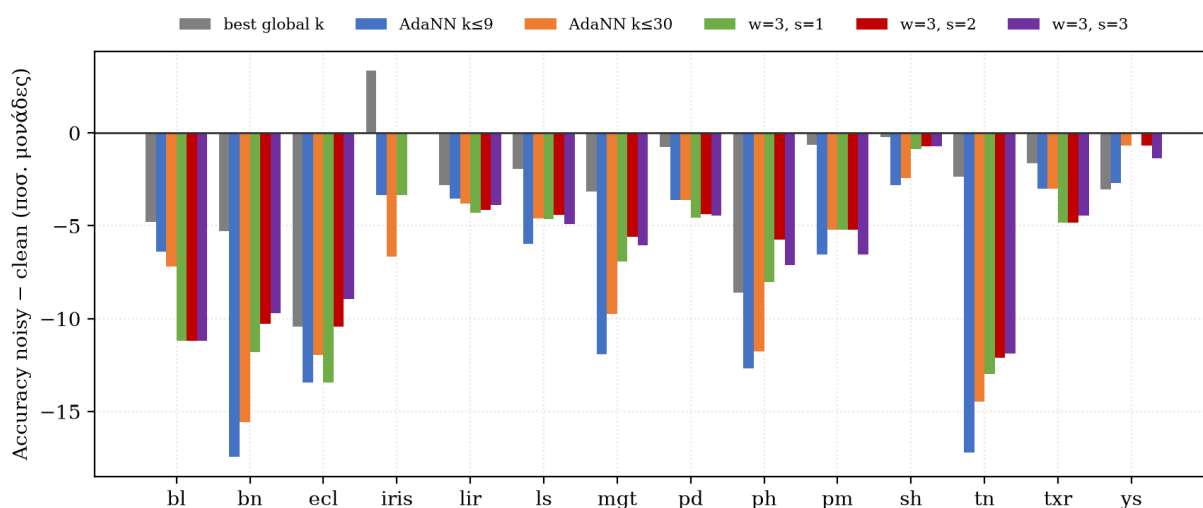
ds	best global k	AdaNN $k \leq 9$	AdaNN $k \leq 30$	w=3, s=1	w=3, s=2	w=3, s=3
bl	59,64	57,16	55,12	54,05	54,32	54,29
bn	84,02	69,38	71,50	74,13	75,14	76,25
ecl	66,22	58,72	52,45	52,44	52,44	65,84
iris	96,66	89,95	86,60	89,97	93,27	93,27
lir	92,83	91,92	91,45	89,54	89,61	89,81
ls	87,31	82,78	83,88	82,83	83,21	82,30
mgt	69,01	58,63	59,92	60,88	61,65	62,65
pd	98,32	95,20	95,26	94,34	94,54	94,44
ph	62,99	57,84	59,84	59,21	61,04	60,63
pm	54,35	60,87	60,55	55,67	54,35	52,63
sh	58,01	52,61	53,38	59,99	66,85	66,90
tn	95,40	78,32	81,05	81,02	81,98	82,54
txr	97,14	95,75	95,74	93,41	93,41	93,77
ys	41,01	50,98	48,87	48,83	48,07	47,84
M.O.	75,92	71,44	71,11	71,17	72,13	73,08



Σχήμα 7.7: F-score ανά σύνολο δεδομένων και αλγόριθμο στα θορυβώδη δεδομένα.

7.3.5 Η μεταβολή που επιφέρει ο θόρυβος

Το Σχήμα 7.8 απεικονίζει άμεσα τη διαφορά ορθότητας μεταξύ των δύο συνθηκών, ανά σύνολο δεδομένων και αλγόριθμο. Οι αρνητικές τιμές αντιστοιχούν σε πτώση λόγω θορύβου. Η εικόνα επιβεβαιώνει ότι ο βασικός AdaNN υφίσταται τη μεγαλύτερη ζημία σχεδόν παντού, ενώ οι παραλλαγές με $\max k = 30$ απορροφούν καλύτερα την επίδραση. Ιδιαίτερα διαφωτιστικό είναι το Banana, όπου η βάση χάνει σχεδόν 17,5 ποσοστιαίες μονάδες ενώ η Παραλλαγή 4 περιορίζει την απώλεια σε λιγότερες από 10.



Σχήμα 7.8: Μεταβολή της ορθότητας από τα καθαρά στα θορυβώδη δεδομένα. Οι αρνητικές τιμές δηλώνουν πτώση.

7.3.6 Κατανομή των τιμών k

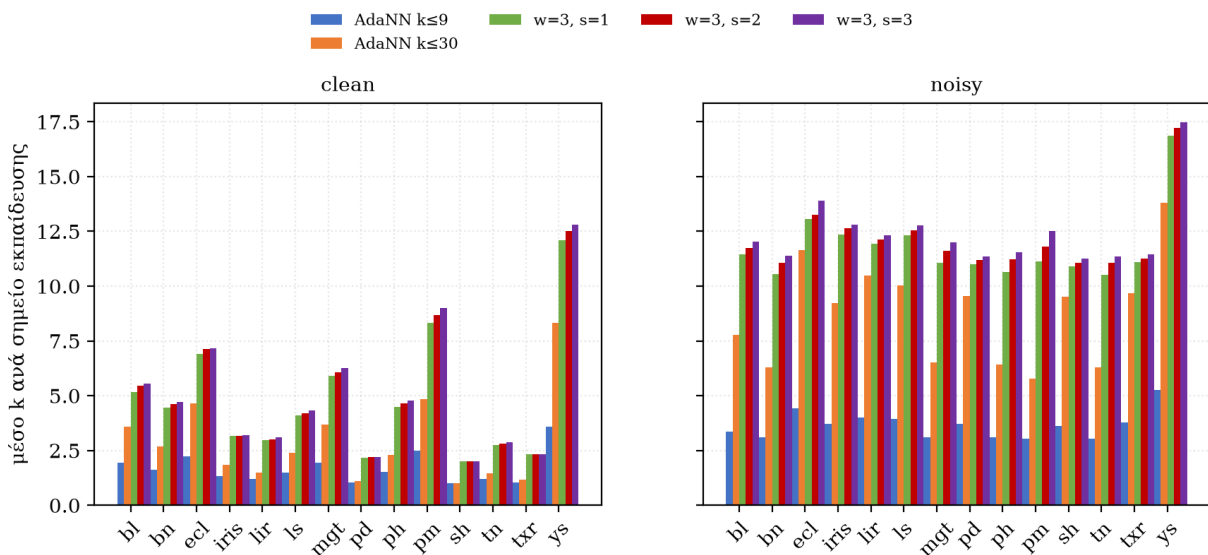
Η ερμηνεία των παραπάνω αποτελεσμάτων προϋποθέτει γνώση των τιμών k που πράγματι ανατέθηκαν. Οι Πίνακες 7.10 και 7.11 δίνουν τη μέση τιμή k ανά σύνολο δεδομένων και παραλλαγή, για τις δύο συνθήκες αντίστοιχα, ενώ το Σχήμα 7.9 τις απεικονίζει συγκριτικά. Το εύρημα είναι σαφές και συστηματικό: ο θόρυβος αυξάνει δραματικά τις τιμές k σε κάθε παραλλαγή και σε κάθε σύνολο δεδομένων χωρίς εξαίρεση.

Πίνακας 7.10: Μέση τιμή k ανά σημείο εκπαίδευσης, καθαρά δεδομένα.

ds	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	1,96	3,58	5,17	5,45	5,54
bn	1,63	2,69	4,47	4,62	4,70
ecl	2,25	4,65	6,91	7,14	7,16
iris	1,32	1,85	3,18	3,18	3,19
lir	1,20	1,48	2,96	3,02	3,09
ls	1,50	2,38	4,09	4,21	4,32
mgt	1,93	3,69	5,92	6,08	6,25
pd	1,03	1,10	2,18	2,19	2,19
ph	1,51	2,31	4,50	4,65	4,77
pm	2,50	4,86	8,32	8,68	9,01
sh	1,00	1,01	2,02	2,02	2,02
tn	1,20	1,47	2,74	2,81	2,87
txr	1,05	1,16	2,32	2,33	2,33
ys	3,58	8,33	12,10	12,51	12,80
M.O.	1,69	2,90	4,78	4,92	5,02

Πίνακας 7.11: Μέση τιμή k ανά σημείο εκπαίδευσης, θορυβώδη δεδομένα.

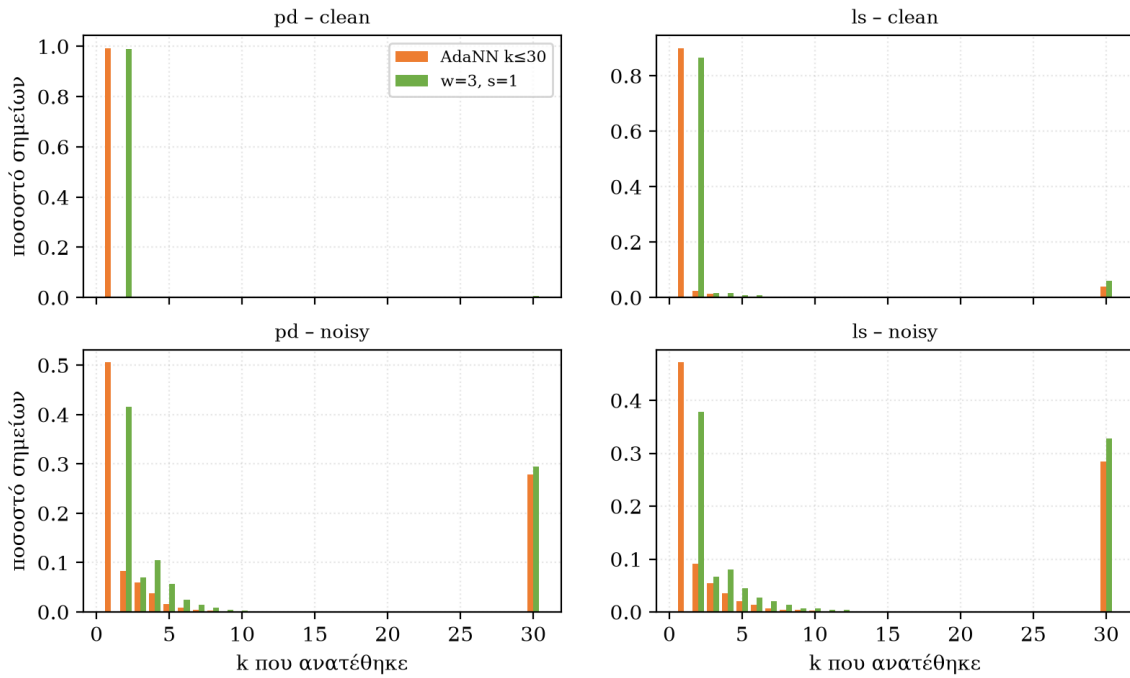
ds	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	3,36	7,77	11,45	11,73	12,02
bn	3,09	6,31	10,55	11,08	11,39
ecl	4,41	11,64	13,06	13,25	13,90
iris	3,71	9,24	12,35	12,65	12,82
lir	4,01	10,48	11,93	12,12	12,33
ls	3,95	10,04	12,32	12,55	12,78
mgt	3,12	6,51	11,05	11,61	11,99
pd	3,72	9,56	11,00	11,19	11,37
ph	3,10	6,41	10,65	11,23	11,56
pm	3,03	5,78	11,12	11,81	12,52
sh	3,63	9,51	10,90	11,08	11,27
tn	3,05	6,31	10,53	11,07	11,34
txr	3,78	9,69	11,09	11,26	11,45
ys	5,25	13,81	16,88	17,21	17,49
M.O.	3,66	8,79	11,78	12,13	12,44



Σχήμα 7.9: Μέση τιμή k ανά σημείο εκπαίδευσης, για καθαρά (αριστερά) και θορυβώδη (δεξιά) δεδομένα.

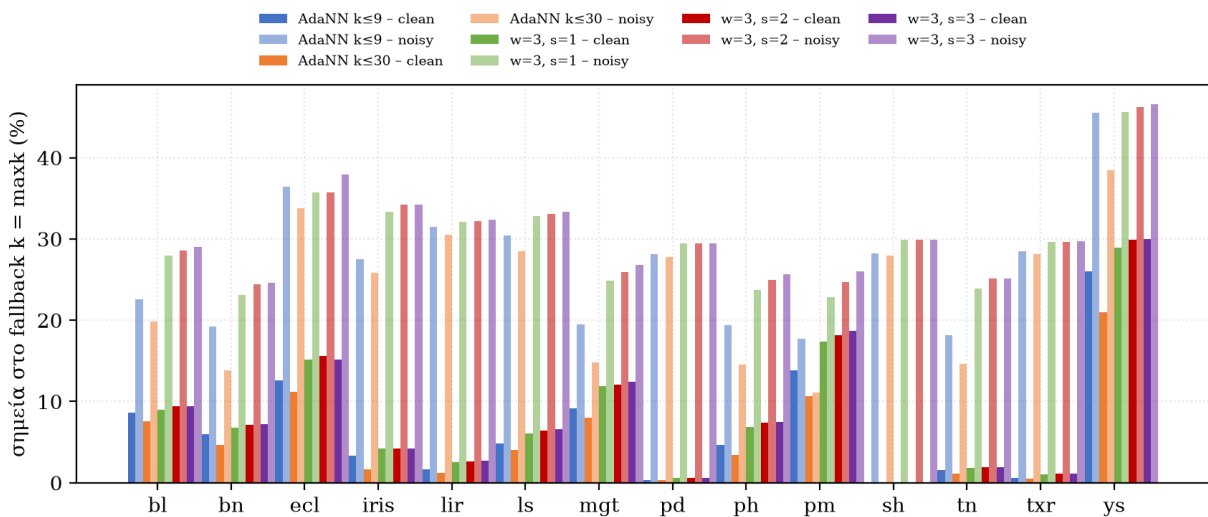
Σε αριθμούς: στα καθαρά δεδομένα ο βασικός AdaNN αναθέτει μέση τιμή 1,69, η Παραλλαγή 1 2,90 και οι παραλλαγές παραθύρου γύρω στο 5,02. Στα θορυβώδη δεδομένα οι αντίστοιχες τιμές εκτοξεύονται σε 3,66, 8,79 και 12,44. Η αύξηση αυτή δεν προγραμματίστηκε από κανένα σημείο του αλγορίθμου: κανένας μηχανισμός δεν πληροφορήθηκε ότι τα δεδομένα περιέχουν θόρυβο. Προέκυψε αποκλειστικά επειδή, υπό θόρυβο, οι μικρές τιμές k παύουν να κατηγοριοποιούν σωστά τα σημεία εκπαίδευσης και η αναζήτηση αναγκάζεται να προχωρήσει προς τα δεξιά. Πρόκειται για την ακριβή επιβεβαίωση της υπόθεσης που διατυπώθηκε στο ερευνητικό ερώτημα Ε4.

Το Σχήμα 7.10 δείχνει την πλήρη κατανομή, και όχι μόνο τον μέσο όρο, για δύο αντιπροσωπευτικά σύνολα. Στα καθαρά δεδομένα η κατανομή είναι έντονα συγκεντρωμένη στις μικρές τιμές, με τη συντριπτική πλειονότητα των σημείων να λαμβάνει $k = 1$ ή $k = 2$. Στα θορυβώδη δεδομένα η κατανομή απλώνεται σε ολόκληρο το διάστημα και εμφανίζει έντονη κορυφή στην τιμή 30, δηλαδή στην εφεδρική τιμή. Η διαφορά ανάμεσα στον απλό κανόνα και τον κανόνα παραθύρου είναι επίσης ορατή: ο δεύτερος μετατοπίζει συστηματικά τη μάζα της κατανομής προς τα δεξιά, ακριβώς όπως προβλέπει η δεύτερη δομική ιδιότητα της ενότητας 5.6.



Σχήμα 7.10: Κατανομή των τιμών k που ανατέθηκαν στα σημεία εκπαίδευσης, για τα σύνολα Pen Digits και Landsat Satellite, σε καθαρά και θορυβώδη δεδομένα.

Το Σχήμα 7.11 παρουσιάζει το ποσοστό των σημείων που κατέληξαν στην εφεδρική τιμή, δηλαδή δεν κατηγοριοποιήθηκαν σωστά για καμία τιμή k ή τριάδα. Στα καθαρά δεδομένα ο μέσος όρος κυμαίνεται από 5,4 % έως 8,4 %, ενώ στα θορυβώδη εκτινάσσεται σε 23,5 % έως 30,8 %. Το ποσοστό αυτό πλησιάζει το ίδιο το επίπεδο θορύβου, γεγονός αναμενόμενο: ένα σημείο του οποίου η ετικέτα έχει αλλοιωθεί δύσκολα θα κατηγοριοποιηθεί σωστά από τους ορθά επισημειωμένους γείτονές του, για οποιαδήποτε τιμή k .



Σχήμα 7.11: Ποσοστό σημείων εκπαίδευσης που κατέληξαν στην εφεδρική τιμή $k = \max k$. Το πλήρες χρώμα αντιστοιχεί σε καθαρά και το ημιδιαφανές σε θορυβώδη δεδομένα.

7.3.7 Η καθολική τιμή k

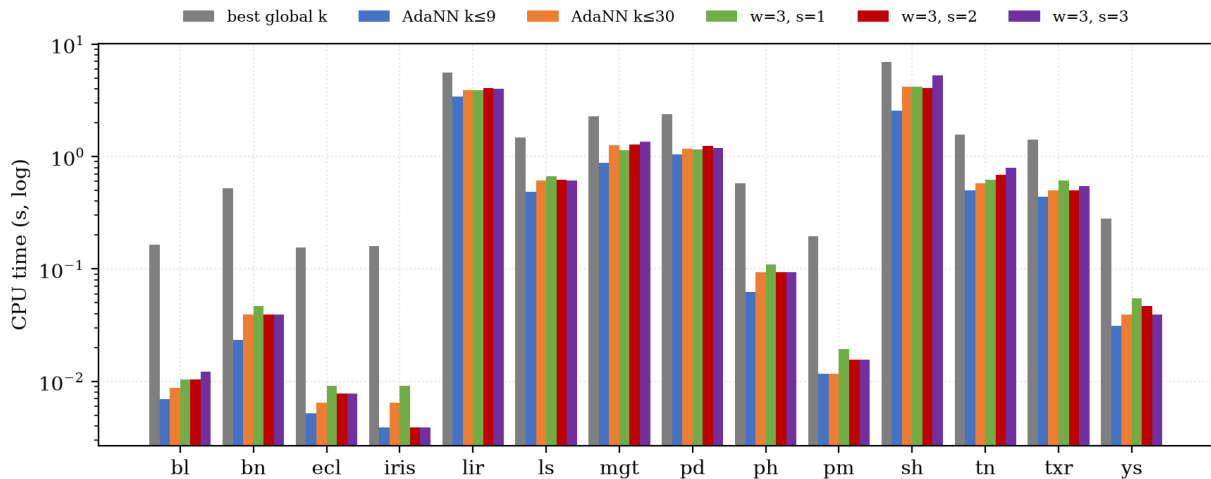
Ο Πίνακας 7.12 παραθέτει την τιμή k που επέλεξε η διασταυρούμενη επικύρωση για κάθε σύνολο δεδομένων και στις δύο συνθήκες. Η αντιπαράβολή είναι εξαιρετικά διαφωτιστική. Στα καθαρά δεδομένα η επιλογή είναι $k = 1$ για πέντε από τα δεκατέσσερα σύνολα, δηλαδή η διασταυρούμενη επικύρωση διαπιστώνει ότι ο απλός κανόνας του ενός εγγύτερου γείτονα είναι ο βέλτιστος. Στα θορυβώδη δεδομένα καμία τιμή δεν πέφτει κάτω από το 9, ενώ σε πέντε σύνολα επιλέγεται η μέγιστη δυνατή τιμή 30, γεγονός που υποδηλώνει ότι η βέλτιστη τιμή ενδέχεται να βρίσκεται ακόμη υψηλότερα και να περιορίζεται από το άνω όριο της σάρωσης.

Πίνακας 7.12: Η καθολική τιμή k που επιλέχθηκε με πενταπλή διασταυρούμενη επικύρωση, ανά σύνολο δεδομένων.

ds	best k (clean)	best k (noisy)
bl	28	23
bn	26	27
ecl	5	10
iris	11	17
lir	1	9
ls	3	13
mgt	11	30
pd	1	12
ph	1	30
pm	24	30
sh	1	18
tn	30	30
txr	1	13
ys	30	18

7.3.8 Υπολογιστικό κόστος

Το Σχήμα 7.12 παρουσιάζει τους χρόνους CPU σε λογαριθμική κλίμακα. Η ιεραρχία είναι σταθερή σε όλα τα σύνολα δεδομένων και συμφωνεί απόλυτα με την ανάλυση πολυπλοκότητας της ενότητας 4.6. Ο καθολικός προσδιορισμός είναι ο ακριβότερος, με μέσο χρόνο 1,70 δευτερόλεπτα, καθώς εκτελεί πέντε αναζητήσεις γειτόνων για τη διασταυρούμενη επικύρωση συν μία τελική. Ο βασικός AdaNN είναι ο ταχύτερος με 0,68 δευτερόλεπτα, αφού χρειάζεται μόλις δέκα γείτονες ανά σημείο. Οι παραλλαγές με $\max k = 30$ κινούνται ενδιάμεσα, γύρω στα 0,88 έως 1,00 δευτερόλεπτα.



Σχήμα 7.12: Χρόνος CPU ανά σύνολο δεδομένων και αλγόριθμο σε λογαριθμική κλίμακα (καθαρά δεδομένα).

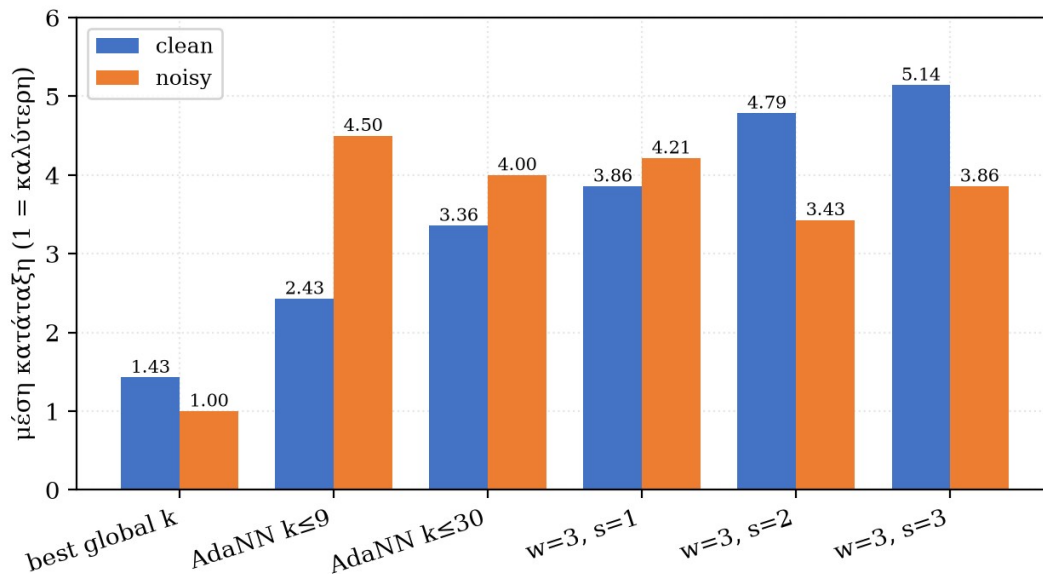
Αξίζει να επισημανθεί ότι η τριπλάσια αύξηση του άνω ορίου, από 9 σε 30, δεν τριπλασιάζει τον χρόνο αλλά τον αυξάνει κατά περίπου 30 %. Η αιτία είναι ότι το κυρίαρχο κόστος είναι η κατασκευή της δομής ευρετηρίασης και η ίδια η αναζήτηση, ενώ η επιπλέον σάρωση των τιμών k υλοποιείται με προοδευτική συσσώρευση ψήφων και είναι σχεδόν δωρεάν. Το εύρημα αυτό έχει πρακτική σημασία: η επέκταση του εύρους αναζήτησης, που όπως φάνηκε βελτιώνει σημαντικά την επίδοση υπό θόρυβο, κοστίζει ελάχιστα. Ακόμη και στην ακριβότερη περίπτωση, οι παραλλαγές του AdaNN παραμένουν περίπου διπλάσια ταχύτερες από τον καθολικό προσδιορισμό.

7.3.9 Συνολική κατάταξη

Για συνολική αποτίμηση χρησιμοποιείται η μέση κατάταξη, μετρική που προτείνεται από τον Demšar [26] για τη σύγκριση κατηγοριοποιητών σε πολλά σύνολα δεδομένων. Σε κάθε σύνολο οι έξι αλγόριθμοι κατατάσσονται από 1 (καλύτερος) έως 6 (χειρότερος) με βάση την ορθότητα, και υπολογίζεται ο μέσος όρος των κατατάξεων. Το πλεονέκτημα της προσέγγισης είναι ότι δεν επηρεάζεται από τις απόλυτες διαφορές, οι οποίες ποικίλλουν σημαντικά μεταξύ συνόλων. Τα αποτελέσματα δίνονται στον Πίνακα 7.13 και απεικονίζονται στο Σχήμα 7.13.

Πίνακας 7.13: Μέση κατάταξη των αλγορίθμων ως προς την ορθότητα (1 = καλύτερος).

Αλγόριθμος	Μέση κατάταξη (clean)	Μέση κατάταξη (noisy)
best global k	1,43	1,00
AdaNN $k \leq 9$	2,43	4,50
AdaNN $k \leq 30$	3,36	4,00
w=3, s=1	3,86	4,21
w=3, s=2	4,79	3,43
w=3, s=3	5,14	3,86



Σχήμα 7.13: Μέση κατάταξη των έξι αλγορίθμων ως προς την ορθότητα, στις δύο συνθήκες.

Η αντιστροφή που περιγράφηκε στις προηγούμενες ενότητες αποτυπώνεται εδώ με τη μεγαλύτερη σαφήνεια. Στα καθαρά δεδομένα η σειρά των παραλλαγών του AdaNN είναι ακριβώς αντίστροφη της πολυπλοκότητάς τους: η βάση κατατάσσεται 2,43, η Παραλλαγή 1 3,36 και οι παραλλαγές παραθύρου ακολουθούν με 3,86, 4,79 και 5,14. Στα θορυβώδη δεδομένα η σειρά αναστρέφεται: η βάση πέφτει στην τελευταία θέση με 4,50, ενώ η Παραλλαγή 3 ανεβαίνει στην καλύτερη θέση μεταξύ των AdaNN με 3,43. Ο καθολικός προσδιορισμός κατατάσσεται 1,00, δηλαδή πρώτος σε **κάθε ένα** από τα δεκατέσσερα σύνολα δεδομένων.

Ο Πίνακας 7.14 συνοψίζει όλα τα παραπάνω σε μία εικόνα, παραθέτοντας τους μέσους όρους ορθότητας, F-score και χρόνου CPU για κάθε αλγόριθμο και για τις δύο συνθήκες.

Πίνακας 7.14: Συγκεντρωτικοί μέσοι όροι όλων των μετρικών και των δύο συνθηκών.

Αλγόριθμος	Acc clean	Acc noisy	F clean	F noisy	CPU clean	CPU noisy
best global k	89,73	86,70	82,10	75,92	1,696	1,728
AdaNN $k \leq 9$	88,71	80,81	83,66	71,44	0,678	0,734
AdaNN $k \leq 30$	88,63	81,44	83,00	71,11	0,883	0,958
$w=3, s=1$	88,39	81,81	82,97	71,17	0,895	1,210
$w=3, s=2$	88,19	82,49	82,55	72,13	0,908	0,910
$w=3, s=3$	88,27	82,46	82,75	73,08	0,997	0,894

7.3.10 Ερμηνεία των ευρημάτων

Γιατί οι παραλλαγές δεν βοηθούν σε καθαρά δεδομένα. Η αποτυχία των παραλλαγών παραθύρου στα καθαρά δεδομένα δεν είναι τυχαία ούτε οφείλεται σε ατέλεια της υλοποίησης· είναι δομική συνέπεια του ίδιου του κανόνα. Όπως επισημάνθηκε στην ενότητα 5.6, το παράθυρο πλάτους τρία **δεν μπορεί ποτέ να αναθέσει την τιμή $k = 1$** . Στα καθαρά δεδομένα, όμως, η τιμή $k = 1$ είναι ακριβώς η βέλτιστη για τη συντριπτική πλειονότητα των σημείων: τα ιστογράμματα του Σχήματος 7.10 δείχνουν ότι η κατανομή των τιμών k είναι έντονα συγκεντρωμένη στη μονάδα, ενώ ο Πίνακας 7.12 δείχνει ότι ακόμη και η διασταυρούμενη επικύρωση επιλέγει $k = 1$ σε πέντε σύνολα. Ο κανόνας παραθύρου αποκλείει εξ ορισμού τη βέλτιστη επιλογή και αναγκάζεται να αναθέσει μεγαλύτερες τιμές, οι οποίες εξομαλύνουν χωρίς λόγο ένα όριο απόφασης που δεν χρειάζεται εξομάλυνση.

Η ποσοτική επιβεβαίωση δίνεται από τους μέσους όρους των τιμών k . Στα καθαρά δεδομένα ο βασικός AdaNN αναθέτει μέση τιμή 1,69, ενώ οι παραλλαγές παραθύρου κινούνται γύρω στο 5,02, δηλαδή τριπλάσια. Πρόκειται για συστηματική υπερεκτίμηση σε συνθήκες όπου δεν υπάρχει τίποτε να αντισταθμιστεί. Η ίδια λογική εξηγεί και γιατί η Παραλλαγή 1, που διατηρεί τον κανόνα της πρώτης επιτυχίας και απλώς επεκτείνει το εύρος, ζημιώνεται πολύ λιγότερο: τα σημεία που πετυχαίνουν με $k = 1$ συνεχίζουν να λαμβάνουν $k = 1$, και μόνο τα λίγα σημεία που κατέφευγαν στην εφεδρική τιμή επηρεάζονται. Το συμπέρασμα υπερβαίνει τη συγκεκριμένη εργασία: ένας μηχανισμός που προσθέτει ανθεκτικότητα δεν είναι ουδέτερος όταν η ανθεκτικότητα δεν χρειάζεται· έχει κόστος, το οποίο πληρώνεται σε ακρίβεια.

Γιατί οι παραλλαγές βοηθούν υπό θόρυβο. Υπό θόρυβο η κατάσταση αντιστρέφεται διότι αντιστρέφεται η ίδια η ανάγκη. Ένα σημείο εκπαίδευσης του οποίου η ετικέτα έχει αλλοιωθεί βρίσκεται πλέον περιτριγυρισμένο από γείτονες διαφορετικής, αλλά ορθής, κλάσης. Ο κανόνας της πρώτης επιτυχίας είναι ιδιαίτερα ευάλωτος σε αυτή τη συνθήκη: ένα αλλοιωμένο σημείο μπορεί να κατηγοριοποιηθεί «σωστά», δηλαδή σύμφωνα με την εσφαλμένη του ετικέτα, για κάποια μικρή τιμή k , αν τύχει να έχει κοντά του ένα ακόμη αλλοιωμένο σημείο. Ο κανόνας θα παγιώσει τότε μια πολύ μικρή τιμή k , η οποία θα μεταδοθεί σε όλα τα σημεία ελέγχου που έχουν αυτό το σημείο ως εγγύτερο γείτονα.

Ο κανόνας παραθύρου προστατεύει ακριβώς από αυτό το σενάριο. Η απαίτηση τριών διαδοχικών επιτυχιών καθιστά την τυχαία σύμπτωση πολύ λιγότερο πιθανή, διότι απαιτεί η εσφαλμένη ετικέτα να επιβεβαιωθεί τρεις φορές καθώς η γειτονιά διευρύνεται. Το σημαντικότερο ίσως εύρημα της εργασίας είναι ότι η προσαρμογή αυτή συμβαίνει **αυτόματα**: κανένας αλγόριθμος δεν πληροφορήθηκε ποτέ ότι τα δεδομένα περιέχουν θόρυβο, ούτε του δόθηκε παράμετρος που να ρυθμίζει την ανθεκτικότητα. Ο μηχανισμός λειτουργεί δηλαδή ως αυτόματος ανιχνευτής τοπικής δυσκολίας, χωρίς να χρειάζεται να γνωρίζει την αιτία της.

Γιατί υπερτερεί ο καθολικός προσδιορισμός υπό θόρυβο. Το πλέον απροσδόκητο εύρημα είναι ότι ο καθολικός προσδιορισμός κατατάσσεται πρώτος σε **κάθε ένα** από τα δεκατέσσερα θορυβώδη σύνολα, με μέση κατάταξη 1,00 και μέση ορθότητα 86,70 %. Η εξήγηση βρίσκεται στην **πηγή της πληροφορίας** που χρησιμοποιεί κάθε μέθοδος. Ο AdaNN εκτιμά την τιμή k κάθε σημείου από **μία και μοναδική** παρατήρηση: το αν το ίδιο το σημείο κατηγοριοποιείται σωστά. Όταν η ετικέτα του σημείου αυτού είναι αλλοιωμένη, η εκτίμηση δεν είναι απλώς θορυβώδης αλλά συστηματικά λανθασμένη, και το σφάλμα μεταδίδεται σε ολόκληρο το κελί Voronoi του. Με 30 % αλλοιωμένες ετικέτες, περίπου το ένα τρίτο των τιμών k που ανατίθενται προέρχεται από διαφθαρμένη πηγή. Ο καθολικός προσδιορισμός, αντιθέτως, εκτιμά μία μόνο τιμή από τον μέσο όρο πέντε πτυχών πάνω σε ολόκληρο το σύνολο εκπαίδευσης· ο θόρυβος επηρεάζει και αυτή την εκτίμηση, αλλά εξομαλύνεται από τη συνάθροιση χιλιάδων παρατηρήσεων.

Με άλλα λόγια, ο θόρυβος πλήττει την προσαρμοστική μέθοδο σε **δύο επίπεδα ταυτόχρονα**: διαφθείρει τα δεδομένα πάνω στα οποία εφαρμόζεται η ψηφοφορία, όπως ακριβώς και στην καθολική μέθοδο, αλλά διαφθείρει επιπλέον και τον μηχανισμό επιλογής της παραμέτρου. Η καθολική μέθοδος υφίσταται μόνο την πρώτη επίπτωση. Το εύρημα αυτό συνδέεται άμεσα με τη διάκριση των επιπέδων προσδιορισμού της ενότητας 3.2: η τοπική προσαρμογή έχει νόημα όταν οι τοπικές εκτιμήσεις είναι αξιόπιστες, και υπό υψηλό θόρυβο η ίδια η τοπικότητα γίνεται αδυναμία.

Η επίδραση του βήματος μετατόπισης. Οι τρεις παραλλαγές παραθύρου διαφέρουν μόνο ως προς το βήμα, και οι διαφορές τους είναι μικρές αλλά όχι τυχαίες. Στα καθαρά δεδομένα η Παραλλαγή 2 με βήμα 1 κατατάσσεται καλύτερα (3,86) από τις Παραλλαγές 3 και 4 (4,79 και 5,14), γεγονός αναμενόμενο: το μικρότερο βήμα εξετάζει κάθε δυνατή τριάδα και εντοπίζει την πρώτη περιοχή

σταθερότητας όπου κι αν βρίσκεται, αναθέτοντας έτσι τις μικρότερες δυνατές τιμές k . Στα θορυβώδη δεδομένα η σειρά αλλάζει και η Παραλλαγή 3 με βήμα 2 προηγείται (3,43 έναντι 4,21). Το βήμα λειτουργεί δηλαδή ως ρυθμιστής της επιθετικότητας με την οποία η αναζήτηση προχωρά προς μεγαλύτερες τιμές, όπως επιβεβαιώνει η μονότονη αύξηση της μέσης τιμής k από 11,78 για βήμα 1 σε 12,44 για βήμα 3. Το γεγονός ότι το βέλτιστο βήμα είναι το **ενδιάμεσο**, και όχι το ακραίο, υποδηλώνει ισορροπία ανάμεσα στην πυκνότητα ελέγχου και την ταχύτητα προόδου.

7.3.11 Σύγκριση με τη βιβλιογραφία

Τα ευρήματα της εργασίας συμφωνούν με τη γενική εικόνα της βιβλιογραφίας σε δύο σημεία και την εμπλουτίζουν σε ένα τρίτο. Πρώτον, επιβεβαιώνεται η θεμελιώδης παρατήρηση των Cover και Hart [2] ότι μεγαλύτερες τιμές k οδηγούν σε καλύτερη προσέγγιση του σφάλματος Bayes όταν τα δεδομένα το επιτρέπουν· εδώ η ανάγκη αυτή προκύπτει τεχνητά μέσω του θορύβου. Δεύτερον, επιβεβαιώνεται η διαπίστωση των Zhu και Wu [22] ότι ο θόρυβος ετικετών είναι ιδιαίτερα επιζήμιος, καθώς η μέση ορθότητα του βασικού AdaNN πέφτει από 88,71 % σε 80,81 %, δηλαδή σχεδόν οκτώ ποσοστιαίες μονάδες.

Το σημείο στο οποίο η εργασία προσθέτει κάτι νέο αφορά τη σχέση ανάμεσα στην τοπική προσαρμογή και τον θόρυβο. Όπως καταγράφει η επισκόπηση [32] και επιβεβαιώνει ο Πίνακας 3.2, η βιβλιογραφία των προσαρμοστικών μεθόδων παρουσιάζει συνήθως αποτελέσματα σε καθαρά δεδομένα και τεκμηριώνει βελτίωση έναντι του k -NN με λίγες σταθερές τιμές k . Τα παρόντα αποτελέσματα δείχνουν ότι η σύγκριση αυτή μπορεί να είναι παραπλανητική αν το σημείο αναφοράς δεν περιλαμβάνει καθολικό προσδιορισμό μέσω διασταυρούμενης επικύρωσης, και ότι υπό θόρυβο η ιεραρχία ενδέχεται να αντιστραφεί. Η παρατήρηση συνδέεται με τη διαπίστωση των Fréney και Verleysen [21] ότι η αξιολόγηση αλγορίθμων υπό θόρυβο απαιτεί ξεχωριστό πειραματικό σχεδιασμό και δεν προκύπτει ως επέκταση της αξιολόγησης σε καθαρά δεδομένα.

Αξίζει τέλος να σημειωθεί η σύνδεση με τις μεθόδους καθαρισμού δεδομένων τύπου ENN [16]. Ο κανόνας ENN αφαιρεί από το σύνολο εκπαίδευσης τα σημεία που δεν κατηγοριοποιούνται σωστά από τους γείτονές τους, δηλαδή ακριβώς τα σημεία που στην παρούσα εργασία καταλήγουν στην εφεδρική τιμή $k = \max k$. Το ποσοστό αυτό ανέρχεται σε περίπου 23,5 % στα θορυβώδη δεδομένα, τιμή που πλησιάζει το επίπεδο του εισαχθέντος θορύβου. Ο AdaNN, δηλαδή, εντοπίζει ουσιαστικά τα ίδια σημεία που θα αφαιρούσε ένας καθαριστής, αλλά αντί να τα διαγράψει τους αναθέτει μεγάλη τιμή k .

7.3.12 Περιορισμοί της μελέτης

Η εργασία υπόκειται σε περιορισμούς που πρέπει να ληφθούν υπόψη κατά την ερμηνεία των αποτελεσμάτων. Ο πρώτος και σοβαρότερος αφορά τον **διαχωρισμό των δεδομένων**: χρησιμοποιήθηκε ένας μόνο διαχωρισμός 80/20 ανά σύνολο δεδομένων, και όχι πλήρης διασταυρούμενη επικύρωση για την τελική αξιολόγηση. Οι διαφορές που καταγράφονται, ιδίως στα μικρά σύνολα όπως το Iris με μόλις τριάντα σημεία ελέγχου, εμπίπτουν σε μεγάλο βαθμό στη διακύμανση του διαχωρισμού: μία μόνο διαφορετική πρόβλεψη μεταβάλλει εκεί την ορθότητα κατά περισσότερες από τρεις ποσοστιαίες μονάδες. Οι μέσοι όροι και οι μέσες κατατάξεις σε δεκατέσσερα σύνολα είναι σαφώς πιο αξιόπιστοι από τις επιμέρους τιμές.

Ο δεύτερος περιορισμός αφορά την **απουσία ελέγχου στατιστικής σημαντικότητας**. Οι διαφορές αναφέρονται ως έχουν, χωρίς έλεγχο Friedman ή μετέπειτα κατά ζεύγη συγκρίσεις, όπως προτείνει ο Demšar [26]. Δεδομένου ότι πολλές από τις διαφορές στα καθαρά δεδομένα είναι της τάξης του μισού της ποσοστιαίας μονάδας, είναι πιθανό ένα μέρος τους να μην είναι στατιστικά σημαντικά. Αντιθέτως,

η αντιστροφή που παρατηρείται στα θορυβώδη δεδομένα είναι αρκετά μεγάλη και αρκετά συστηματική, με οκτώ έως εννέα νίκες στα δεκατέσσερα σύνολα, ώστε να θεωρείται ασφαλής. Σημειώνεται ότι το πρόβλημα αυτό δεν είναι ιδιαίτερο της παρούσας εργασίας: όπως καταγράφει η επισκόπηση [32], μόνο σε δώδεκα από τις είκοσι οκτώ σχετικές δημοσιεύσεις διεξήχθη έστω ένας στατιστικός έλεγχος.

Ο τρίτος περιορισμός αφορά τις **ισοπαλίες απόστασης**, όπως αναλύθηκε στην ενότητα 6.9. Στο Balance Scale το 100 % των σημείων εμφανίζει ισοπαλία μέσα στους τριάντα πρώτους γείτονες και στο Iris το 68 %. Στα σύνολα αυτά η κατάταξη των ισαπεχόντων γειτόνων είναι μαθηματικά απροσδιόριστη, και επομένως τα αποτελέσματα εξαρτώνται εν μέρει από την εσωτερική σειρά που επιστρέφει η βιβλιοθήκη αναζήτησης. Τα συμπεράσματα για τα δύο αυτά σύνολα πρέπει να αντιμετωπίζονται με επιφύλαξη· τα υπόλοιπα δώδεκα σύνολα δεν επηρεάζονται ουσιαστικά.

Ο τέταρτος περιορισμός αφορά το **εύρος του πειραματικού πλέγματος**. Εξετάστηκε ένα μόνο επίπεδο θορύβου, περίπου 30 %, ενώ η συμπεριφορά σε ενδιάμεσα επίπεδα, όπως 5 %, 10 % ή 20 %, παραμένει άγνωστη και είναι πιθανό να αποκαλύπτει το σημείο καμπής όπου η ιεραρχία των μεθόδων αντιστρέφεται. Ομοίως, το πλάτος του παραθύρου διατηρήθηκε σταθερό στο τρία, χωρίς να εξεταστούν πλάτη δύο, τέσσερα ή πέντε, και το άνω όριο περιορίστηκε στο 30, τιμή που όπως φάνηκε στον Πίνακα 7.12 είναι δεσμευτική για ορισμένα θορυβώδη σύνολα. Τέλος, ο θόρυβος που εξετάστηκε είναι ομοιόμορφος και ανεξάρτητος από τα δεδομένα, το απλούστερο από τα τρία μοντέλα που διακρίνουν οι Fréney και Verleysen [21].

Κεφάλαιο 8ο: Συμπεράσματα και μελλοντικές επεκτάσεις

Το κεφάλαιο αυτό συνοψίζει τα ευρήματα της εργασίας, απαντά ρητά στα πέντε ερευνητικά ερωτήματα που τέθηκαν στην ενότητα 1.4, καταγράφει τη συνεισφορά και προτείνει κατευθύνσεις για περαιτέρω έρευνα.

8.1 Συμπεράσματα

Η εργασία υλοποίησε τον αλγόριθμο AdaNN και τέσσερις παραλλαγές του, και τους αξιολόγησε σε δεκατέσσερα σύνολα δεδομένων του αποθετηρίου KEEL υπό δύο συνθήκες θορύβου, με σημείο αναφοράς τον καθολικό προσδιορισμό του k μέσω πενταπλής στρωματοποιημένης διασταυρούμενης επικύρωσης. Συνολικά εκτελέστηκαν 168 πειράματα, καθένα από τα οποία μέτρησε ορθότητα, ακρίβεια, ανάκληση, F-score και χρόνο CPU. Η ορθότητα της υλοποίησης επαληθεύτηκε έναντι ανεξάρτητης υλοποίησης αναφοράς, με απόλυτη συμφωνία σε κάθε σημείο του οποίου η κατάταξη γειτόνων δεν είναι απροσδιόριστη λόγω ισοπαλίας.

E1: Βελτιώνει η επέκταση του εύρους από [1..9] σε [1..30]; Η απάντηση εξαρτάται από τη συνθήκη. Στα καθαρά δεδομένα η επέκταση είναι ουδέτερη έως ελαφρώς επιζήμια, με τη μέση ορθότητα να μεταβάλλεται από 88,71 % σε 88,63 %. Στα θορυβώδη δεδομένα η επέκταση είναι σαφώς ωφέλιμη, ανεβάζοντας τη μέση ορθότητα από 80,81 % σε 81,44 % και κερδίζοντας τη βάση σε εννέα από τα δεκατέσσερα σύνολα. Δεδομένου ότι το επιπλέον υπολογιστικό κόστος είναι μόλις περίπου 30 %, η επέκταση συνιστάται ως προεπιλογή.

E2: Οδηγεί ο κανόνας παραθύρου σε πιο αξιόπιστη ανάθεση; Ναι, αλλά μόνο όταν η αξιοπιστία απειλείται. Σε καθαρά δεδομένα ο κανόνας είναι επιζήμιος, διότι αποκλείει δομικά την τιμή $k = 1$ που είναι εκεί η βέλτιστη. Υπό θόρυβο ο κανόνας αποδίδει, υπερτερώντας της βάσης σε οκτώ έως εννέα σύνολα και ανεβάζοντας τη μέση ορθότητα κατά έως 1,68 ποσοστιαίες μονάδες. Η αξία του δηλαδή δεν είναι απόλυτη αλλά εξαρτημένη από τις συνθήκες.

E3: Πώς επηρεάζει το βήμα μετατόπισης; Το βήμα λειτουργεί ως ρυθμιστής της ταχύτητας με την οποία η αναζήτηση προχωρά προς μεγαλύτερες τιμές k . Μεγαλύτερο βήμα παράγει συστηματικά μεγαλύτερες τιμές, όπως επιβεβαιώνει η μονότονη αύξηση της μέσης τιμής k από 11,78 σε 12,44 στα θορυβώδη δεδομένα. Το βέλτιστο βήμα δεν είναι ακραίο: υπό θόρυβο υπερτερεί το ενδιάμεσο βήμα 2, ενώ σε καθαρά δεδομένα προηγείται το βήμα 1. Οι διαφορές μεταξύ των τριών παραλλαγών είναι πάντως μικρότερες από τη διαφορά τους έναντι της βάσης.

E4: Επιβεβαιώνεται ότι ο θόρυβος ευνοεί μεγαλύτερα k ; Απολύτως και χωρίς εξαίρεση. Σε κάθε σύνολο δεδομένων και σε κάθε παραλλαγή η μέση τιμή k αυξήθηκε υπό θόρυβο, με τον βασικό AdaNN να μεταβαίνει από 1,69 σε 3,66 και την Παραλλαγή 1 από 2,90 σε 8,79. Η προσαρμογή συνέβη αυτόματα, χωρίς κανέναν αλγόριθμο να πληροφορηθεί για την παρουσία θορύβου. Το ίδιο επιβεβαιώνεται και από τον καθολικό προσδιορισμό, όπου η επιλεγμένη τιμή δεν πέφτει ποτέ κάτω από το 9 στα θορυβώδη δεδομένα, ενώ στα καθαρά είναι $k = 1$ για πέντε σύνολα.

E5: Πώς συγκρίνεται με τον καθολικό προσδιορισμό; Ο καθολικός προσδιορισμός υπερτερεί ως προς την ορθότητα και στις δύο συνθήκες, και στα θορυβώδη δεδομένα κατατάσσεται πρώτος σε όλα τα σύνολα (μέση κατάταξη 1,00). Έχει όμως δύο μειονεκτήματα. Πρώτον, είναι περίπου διπλάσια ακριβότερος υπολογιστικά. Δεύτερον, και σημαντικότερο, στα καθαρά δεδομένα υστερεί σαφώς ως προς το F-score, με 82,10 % έναντι 83,66 % του AdaNN, διότι μια μεγάλη καθολική τιμή k εξαφανίζει

τις μικρές κλάσεις. Η επιλογή μεθόδου εξαρτάται επομένως από το αν η εφαρμογή ενδιαφέρεται για τη συνολική ορθότητα ή για την ισότιμη μεταχείριση όλων των κλάσεων.

Η κύρια συνεισφορά της εργασίας δεν είναι η ανακάλυψη μιας παραλλαγής που υπερτερεί καθολικά, αλλά η τεκμηρίωση ότι **η αξία των παραλλαγών αυτών είναι εξαρτημένη από τις συνθήκες**, μαζί με την εξήγηση του μηχανισμού που παράγει την εξάρτηση. Η παράλληλη αξιολόγηση σε καθαρά και θορυβώδη δεδομένα αποκάλυψε μια πλήρη αντιστροφή της ιεραρχίας, η οποία θα είχε μείνει αόρατη σε οποιαδήποτε από τις δύο συνθήκες μεμονωμένα. Δευτερευόντως, η εργασία παρέχει μια επαληθευμένη και αποδοτική υλοποίηση, στην οποία η σάρωση όλων των τιμών k κοστίζει όσο μία αναζήτηση γειτόνων, καθώς και ποσοτικά στοιχεία για την κατανομή των τιμών k , τα οποία σπανίως αναφέρονται στη βιβλιογραφία και αποδείχθηκαν καθοριστικά για την ερμηνεία.

Το γενικότερο συμπέρασμα υπερβαίνει τη συγκεκριμένη οικογένεια αλγορίθμων. Η εργασία ξεκίνησε από ένα απλό ερώτημα: αν ένας αυστηρότερος κανόνας επιλογής του k βελτιώνει τον AdaNN. Η απάντηση αποδείχθηκε ότι εξαρτάται εξ ολοκλήρου από το αν τα δεδομένα το χρειάζονται. Σε καθαρά δεδομένα η αυστηρότητα κοστίζει· υπό θόρυβο αποδίδει. Ένας μηχανισμός ανθεκτικότητας δεν είναι ποτέ δωρεάν και οφείλει να αξιολογείται στις συνθήκες για τις οποίες σχεδιάστηκε.

8.2 Μελλοντικές επεκτάσεις

Από τα ευρήματα και τους περιορισμούς της εργασίας προκύπτουν οι ακόλουθες κατευθύνσεις, ταξινομημένες κατά σειρά προτεραιότητας:

- **Κλιμάκωση του επιπέδου θορύβου.** Επανάληψη των πειραμάτων για θόρυβο 5 %, 10 %, 20 % και 40 % ώστε να εντοπιστεί το σημείο καμπής στο οποίο οι παραλλαγές αρχίζουν να υπερτερούν της βάσης. Η γνώση του σημείου αυτού θα μετέτρεπε το ποιοτικό συμπέρασμα της εργασίας σε πρακτικό κριτήριο επιλογής.
- **Πλήρης διασταυρούμενη επικύρωση και έλεγχοι σημαντικότητας.** Χρήση και των πέντε πτυχών κάθε συνόλου αντί μίας, με αναφορά μέσου όρου και τυπικής απόκλισης, και εφαρμογή του ελέγχου Friedman με μετέπειτα κατά ζεύγη συγκρίσεις κατά Demšar [26].
- **Διερεύνηση του πλάτους παραθύρου.** Το πλάτος διατηρήθηκε σταθερό στο τρία. Πλάτη δύο, τέσσερα και πέντε θα επέτρεπαν τη μελέτη της αντιστάθμισης ανάμεσα στην αυστηρότητα του κριτηρίου σταθερότητας και στην ευελιξία των τιμών k που παράγονται.
- **Άρση του άνω ορίου.** Σε πέντε θορυβώδη σύνολα η διασταυρούμενη επικύρωση επέλεξε την οριακή τιμή 30, ένδειξη ότι το όριο είναι δεσμευτικό. Επέκταση σε 50 ή 100 θα έδειχνε πού πραγματικά σταματά το όφελος.
- **Συνδυασμός με καθαρισμό δεδομένων.** Τα σημεία που καταλήγουν στην εφεδρική τιμή είναι ουσιαστικά εκείνα που θα αφαιρούσε ένας κανόνας τύπου ENN [16]. Η προεφαρμογή καθαρισμού πριν τον AdaNN, ή η αντιμετώπιση των σημείων αυτών ως μη έγκυρων πηγών τιμής k κατά τη φάση ελέγχου, αποτελούν φυσικές επεκτάσεις.
- **Σταθμισμένη ψηφοφορία.** Η αντικατάσταση της απλής πλειοψηφίας με ψηφοφορία σταθμισμένη κατά το αντίστροφο της απόστασης θα μετρίαζε την επίδραση των απομακρυσμένων γειτόνων στις μεγάλες τιμές k , όπου η υπόθεση της τοπικής ομοιογένειας εξασθενεί.
- **Υβριδική προσέγγιση.** Δεδομένου ότι ο καθολικός προσδιορισμός υπερτερεί στην ορθότητα και ο AdaNN στο F-score σε καθαρά δεδομένα, αξίζει να διερευνηθεί σχήμα που χρησιμοποιεί

την καθολική τιμή ως κάτω όριο και επιτρέπει τοπική απόκλιση μόνο όπου η τοπική εκτίμηση κρίνεται αξιόπιστη.

- **Σύγκριση με άλλες δυναμικές μεθόδους.** Η εργασία σύγκρινε τον AdaNN μόνο με τον καθολικό προσδιορισμό. Η σύγκριση με μεθόδους άλλων οικογενειών του Κεφαλαίου 3, όπως ο shd-k-NN [38] ή ο Ada-k-NN2 [35], υπό τις ίδιες συνθήκες θορύβου, θα τοποθετούσε τα ευρήματα σε ευρύτερο πλαίσιο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] E. Fix and J. L. Hodges, "Discriminatory analysis: Nonparametric discrimination, consistency properties," USAF School of Aviation Medicine, Randolph Field, TX, Tech. Rep. 4, 1951.
- [2] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, Jan. 1967.
- [3] S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," in Proc. 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Yantai, China, 2010, pp. 91-94.
- [4] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York, NY: Wiley-Interscience, 2001.
- [5] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 2009.
- [6] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY: Springer, 2006.
- [7] D. Wettschereck, D. W. Aha, and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," Artificial Intelligence Review, vol. 11, no. 1-5, pp. 273-314, 1997.
- [8] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 5, pp. 1774-1785, May 2018.
- [9] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for kNN classification," ACM Transactions on Intelligent Systems and Technology, vol. 8, no. 3, art. 43, 2017.
- [10] J. Wang, P. Neskovic, and L. N. Cooper, "Improving nearest neighbor rule with a simple adaptive distance measure," Pattern Recognition Letters, vol. 28, no. 2, pp. 207-213, 2007.
- [11] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in On The Move to Meaningful Internet Systems 2003, Lecture Notes in Computer Science, vol. 2888. Berlin, Germany: Springer, 2003, pp. 986-996.
- [12] S. Ougiaroglou and G. Evangelidis, "RHC: a non-parametric cluster-based data reduction for efficient k-NN classification," Pattern Analysis and Applications, vol. 19, no. 1, pp. 93-109, 2016.
- [13] S. Ougiaroglou and G. Evangelidis, "Efficient dataset size reduction by finding homogeneous clusters," in Proc. 5th Balkan Conference in Informatics (BCI), Novi Sad, Serbia, 2012, pp. 168-173.
- [14] S. Ougiaroglou, G. Evangelidis, and D. A. Dervos, "FHC: an adaptive fast hybrid method for k-NN classification," Logic Journal of the IGPL, vol. 23, no. 3, pp. 431-450, 2015.
- [15] P. E. Hart, "The condensed nearest neighbor rule," IEEE Transactions on Information Theory, vol. 14, no. 3, pp. 515-516, May 1968.
- [16] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-2, no. 3, pp. 408-421, Jul. 1972.
- [17] I. Tomek, "Two modifications of CNN," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-6, no. 11, pp. 769-772, 1976.

- [18] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 257-286, 2000.
- [19] H. Brighton and C. Mellish, "Advances in instance selection for instance-based learning algorithms," *Data Mining and Knowledge Discovery*, vol. 6, no. 2, pp. 153-172, 2002.
- [20] S. García, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: taxonomy and empirical study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 417-435, Mar. 2012.
- [21] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845-869, May 2014.
- [22] X. Zhu and X. Wu, "Class noise vs. attribute noise: a quantitative study," *Artificial Intelligence Review*, vol. 22, no. 3, pp. 177-210, 2004.
- [23] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Canada, 1995, pp. 1137-1143.
- [24] M. Stone, "Cross-validated choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B*, vol. 36, no. 2, pp. 111-133, 1974.
- [25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427-437, 2009.
- [26] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.
- [27] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255-287, 2011.
- [28] M. Kelly, R. Longjohn, and K. Nottingham, "The UCI Machine Learning Repository," University of California, Irvine. [Online]. Available: <https://archive.ics.uci.edu>
- [29] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509-517, Sep. 1975.
- [30] F. Pedregosa et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [31] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357-362, Sep. 2020.
- [32] M. Papanikolaou, G. Evangelidis, and S. Ougiaroglou, "Dynamic k determination in k-NN classifier: a literature review," in *Proc. 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2021, pp. 1-8.
- [33] A. A. Kardan, A. Kavian, and A. Esmaili, "Simultaneous feature selection and feature weighting with K selection for KNN classification using BBO algorithm," in *Proc. 5th Conference on Information and Knowledge Technology (IKT)*, 2013, pp. 349-354.
- [34] U. Johansson, R. König, and L. Niklasson, "Evolving a locally optimized instance based learner," in *Proc. International Conference on Data Mining (DMIN)*, 2008, pp. 124-129.

- [35] S. S. Mullick, S. Datta, and S. Das, "Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5713-5725, 2018.
- [36] N. Garcia-Pedrajas, J. A. Romero del Castillo, and G. Cerruela-Garcia, "A proposal for local k values for k-nearest neighbor rule," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 2, pp. 470-475, Feb. 2017.
- [37] F. Bulut and M. F. Amasyali, "Locally adaptive k parameter selection for nearest neighbor classifier: one nearest cluster," *Pattern Analysis and Applications*, vol. 20, no. 2, pp. 415-425, May 2017.
- [38] S. Ougiaroglou, G. Evangelidis, and K. I. Diamantaras, "Dynamic k-NN classification based on region homogeneity," in *Communications in Computer and Information Science*, vol. 1259. Cham, Switzerland: Springer, 2020, pp. 27-37.
- [39] S. Ougiaroglou, A. Nanopoulos, A. N. Papadopoulos, Y. Manolopoulos, and T. Welzer-Druzovec, "Adaptive k-nearest-neighbor classification using a dynamic number of nearest neighbors," in *Advances in Databases and Information Systems (ADBIS)*, Berlin, Germany: Springer, 2007, pp. 66-82.
- [40] F. J. Ferrer-Troyano, J. S. Aguilar-Ruiz, and J. C. Riquelme, "Non-parametric nearest neighbor with local adaptation," in *Progress in Artificial Intelligence (EPIA 2001)*, *Lecture Notes in Computer Science*, vol. 2258. Berlin, Germany: Springer, 2001, pp. 22-29.
- [41] R. Nock, M. Sebban, and D. Bernard, "A simple locally adaptive nearest neighbor rule with application to pollution forecasting," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, no. 8, pp. 1369-1382, Dec. 2003.
- [42] L. Baoli, L. Qin, and Y. Shiwen, "An adaptive k-nearest neighbor text categorization strategy," *ACM Transactions on Asian Language Information Processing*, vol. 3, no. 4, pp. 215-226, Dec. 2004.
- [43] U. Johansson, H. Boström, and R. König, "Extending nearest neighbor classification with spheres of confidence," in *Proc. 21st International Florida Artificial Intelligence Research Society Conference (FLAIRS-21)*, 2008, pp. 282-287.
- [44] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, "Test point specific k estimation for kNN classifier," in *Proc. 22nd International Conference on Pattern Recognition (ICPR)*, 2014, pp. 1478-1483.
- [45] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, "A probabilistic framework for dynamic k estimation in kNN classifiers with certainty factor," in *Proc. 8th International Conference on Advances in Pattern Recognition (ICAPR)*, 2015, pp. 1-5.
- [46] X. Zhang and Y. Li, "A positive-biased nearest neighbour algorithm for imbalanced classification," in *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, *Lecture Notes in Artificial Intelligence*, vol. 7819. Berlin, Germany: Springer, 2013, pp. 293-304.
- [47] Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, "IKNN: informative k-nearest neighbor pattern classification," in *Knowledge Discovery in Databases (PKDD 2007)*, Berlin, Germany: Springer, 2007, pp. 248-264.
- [48] A. K. Ghosh, "On optimum choice of k in nearest neighbor classification," *Computational Statistics and Data Analysis*, vol. 50, no. 11, pp. 3113-3123, Jul. 2006.

- [49] A. K. Ghosh, "On nearest neighbor classification using adaptive choice of k ," *Journal of Computational and Graphical Statistics*, vol. 16, no. 2, pp. 482-502, Jun. 2007.
- [50] C. C. Holmes and N. M. Adams, "A probabilistic nearest neighbour method for statistical pattern recognition," *Journal of the Royal Statistical Society: Series B*, vol. 64, no. 2, pp. 295-306, May 2002.
- [51] A. Dhurandhar and A. Dobra, "Probabilistic characterization of nearest neighbor classifier," *International Journal of Machine Learning and Cybernetics*, vol. 4, no. 4, pp. 259-272, Aug. 2013.
- [52] G. Hamerly and G. Speegle, "Efficient model selection for large-scale nearest-neighbor data mining," in *Data Security and Security Data (BNCOD 2010)*, *Lecture Notes in Computer Science*, vol. 6121. Berlin, Germany: Springer, 2012, pp. 37-54.
- [53] G. G. Enas and S. C. Choi, "Choice of the smoothing parameter and efficiency of k -nearest neighbor classification," in *Statistical Methods of Discrimination and Classification*, vol. 12. Elsevier, 1986, pp. 235-244.
- [54] D. J. Hand and V. Vinciotti, "Choosing k for two-class nearest neighbour classifiers with unbalanced classes," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1555-1562, Jun. 2003.
- [55] X.-F. Zhong, S.-Z. Guo, L. Gao, H. Shan, and J.-H. Zheng, "An improved k -NN classification with dynamic k ," in *Proc. 9th International Conference on Machine Learning and Computing (ICMLC)*, New York, NY, USA, 2017, pp. 211-216.
- [56] O. Anava and K. Y. Levy, " k -Nearest neighbors: from global to local," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 4923-4931.
- [57] S. Zhang, M. Zong, K. Sun, Y. Liu, and D. Cheng, "Efficient k NN algorithm based on graph sparse reconstruction," in *Advanced Data Mining and Applications (ADMA 2014)*, *Lecture Notes in Computer Science*, vol. 8933. Cham, Switzerland: Springer, 2014, pp. 356-369.
- [58] N. Tomašev, M. Radovanović, D. Mladenović, and M. Ivanović, "Hubness-based fuzzy measures for high-dimensional k -nearest neighbor classification," *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 3, pp. 445-458, Jun. 2014.

ΠΑΡΑΡΤΗΜΑ Α: ΠΗΓΑΙΟΣ ΚΩΔΙΚΑΣ

Στο παράρτημα αυτό παρατίθενται τα βασικά τμήματα της υλοποίησης, τα οποία αναλύθηκαν στο Κεφάλαιο 6. Ο πλήρης κώδικας, μαζί με το σενάριο επαλήθευσης, το σενάριο παραγωγής των σχημάτων και τα αρχεία δεδομένων, συνοδεύει την εργασία σε ηλεκτρονική μορφή. Ο κώδικας είναι γραμμένος σε Python και χρησιμοποιεί τις βιβλιοθήκες NumPy και scikit-learn.

A.1 Φάση εκπαίδευσης: leave-one-out για όλες τις τιμές k

Η συνάρτηση που ακολουθεί αποτελεί τον πυρήνα της υλοποίησης (Σχήμα 6.1). Παράγει, με ένα και μόνο ερώτημα γειτόνων, τον πίνακα ορθότητας για κάθε σημείο εκπαίδευσης και κάθε τιμή k.

```
def loo_correct_matrix(X_train, y_train, max_k):
    """
    Leave-one-out kNN πάνω στο σύνολο εκπαίδευσης.
    Επιστρέφει πίνακα correct διαστάσεων (n, max_k) όπου correct[i, k-1]
    είναι True όταν ο k-NN με τη συγκεκριμένη τιμή k κατηγοριοποιεί σωστά
    το σημείο i (με το ίδιο το σημείο εξαιρεμένο από τη γειτονιά του).
    """
    n = len(X_train)
    n_classes = int(y_train.max()) + 1
    rows = np.arange(n)

    nbrs = NearestNeighbors(n_neighbors=max_k + 1, algorithm="auto").fit(X_train)
    _, idx = nbrs.kneighbors(X_train)

    # Αφαίρεση του ίδιου του σημείου. Με διπλότυπα σημεία δεν βρίσκεται
    # πάντα στη στήλη 0, οπότε εντοπίζεται ρητά.
    is_self = idx == rows[:, None]
    self_pos = np.where(is_self.any(axis=1), is_self.argmax(axis=1), max_k)
    keep = np.arange(max_k + 1)[None, :] != self_pos[:, None]
    nn = idx[keep].reshape(n, max_k)

    counts = np.zeros((n, n_classes), dtype=np.int32)
    correct = np.zeros((n, max_k), dtype=bool)
    for k in range(1, max_k + 1):
        counts[rows, y_train[nn[:, k - 1]]] += 1
        correct[:, k - 1] = counts.argmax(axis=1) == y_train
    return correct
```

A.2 Οι κανόνες επιλογής του k

Οι δύο συναρτήσεις που ακολουθούν υλοποιούν τους δύο κανόνες επιλογής (Σχήματα 6.2 και 6.3). Η πρώτη αντιστοιχεί στον βασικό AdaNN και στην Παραλλαγή 1, με διαφορετική τιμή της παραμέτρου max_k. Η δεύτερη υλοποιεί τον κανόνα κυλιόμενου παραθύρου των Παραλλαγών 2, 3 και 4, με διαφορετική τιμή της παραμέτρου shift.

```
def k_plain(correct, max_k):
    """Το μικρότερο k στο [1..max_k] που κατηγοριοποιεί σωστά, αλλιώς max_k."""
    C = correct[:, :max_k]
    hit = C.any(axis=1)
    return np.where(hit, C.argmax(axis=1) + 1, max_k)
```

```
def k_window(correct, max_k, window=3, shift=1):
    """
    Κερδίζει το πρώτο παράθυρο `window` διαδοχικών τιμών k που ΟΛΕΣ
    κατηγοριοποιούν σωστά το σημείο· ανατίθεται η μεσαία τιμή του.
    Τα παράθυρα ξεκινούν στο 1 και προχωρούν κατά `shift`. Εφεδρική: max_k.
    """
    n = correct.shape[0]
    k_opt = np.full(n, max_k, dtype=int)
    undecided = np.ones(n, dtype=bool)

    start = 1
    while start + window - 1 <= max_k:
        block = correct[:, start - 1:start - 1 + window].all(axis=1)
        newly = undecided & block
        k_opt[newly] = start + window // 2      # μεσαία τιμή του παραθύρου
        undecided &= ~newly
        if not undecided.any():
            break
        start += shift
    return k_opt
```

A.3 Φάση ελέγχου

Η φάση ελέγχου είναι κοινή σε όλες τις παραλλαγές (Σχήμα 6.4). Τα σημεία ελέγχου ομαδοποιούνται ανά τιμή k ώστε η ψηφοφορία να εκτελείται διανυσματικά για κάθε ομάδα, αντί για έναν βρόχο ανά σημείο.

```
def adann_predict(X_train, y_train, k_opt, X_test):
    """Ο εγγύτερος γείτονας κάθε σημείου ελέγχου δίνει την τιμή k."""
    n_classes = int(y_train.max()) + 1
    max_k = int(k_opt.max())

    nbrs = NearestNeighbors(n_neighbors=max_k, algorithm="auto").fit(X_train)
    _, nn = nbrs.kneighbors(X_test)

    ks = k_opt[nn[:, 0]]
    y_pred = np.empty(len(X_test), dtype=int)
    for kv in np.unique(ks):
        sel = ks == kv
        y_pred[sel] = batch_majority_vote(y_train[nn[sel, :kv]], n_classes)
    return y_pred
```

A.4 Καθολικός προσδιορισμός του k

Ο αλγόριθμος αναφοράς (Σχήμα 6.5). Οι γείτονες υπολογίζονται μία φορά ανά πτυχή και επαναχρησιμοποιούνται για όλες τις τριάντα τιμές του k, ώστε να κατασκευάζονται πέντε δομές αναζήτησης αντί για εκατόν πενήντα.

```
def find_best_global_k(X_train, y_train, max_k=30, n_folds=5):
    """Στρωματοποιημένη 5-fold CV για k = 1..max_k· ένα δέντρο ανά πτυχή."""
    n_classes = int(y_train.max()) + 1
    cv_accs = np.zeros(max_k)
    skf = StratifiedKFold(n_splits=n_folds, shuffle=True, random_state=42)
```

```

folds = []
for tr_idx, val_idx in skf.split(X_train, y_train):
    nbrs = NearestNeighbors(n_neighbors=max_k).fit(X_train[tr_idx])
    _, nn_idx = nbrs.kneighbors(X_train[val_idx])
    folds.append((nn_idx, y_train[tr_idx], y_train[val_idx]))

for k in range(1, max_k + 1):
    accs = [accuracy_score(yval, batch_majority_vote(ytr[nn[:, :k]], n_classes))
             for nn, ytr, yval in folds]
    cv_accs[k - 1] = np.mean(accs)
return int(np.argmax(cv_accs)) + 1, cv_accs

```

A.5 Πρωτόκολλο μέτρησης χρόνου

Η συνάρτηση που υλοποιεί το πρωτόκολλο της ενότητας 7.2 (Σχήμα 6.6). Το ουσιώδες σημείο είναι ότι ο αριθμός επαναλήψεων υπολογίζεται μία φορά ανά σύνολο δεδομένων και εφαρμόζεται σε όλους τους αλγορίθμους του.

```

def time_uniformly(pipelines):
    """
    Εκτελεί κάθε αγωγό ενός συνόλου δεδομένων και επιστρέφει
    {όνομα: (αποτέλεσμα, δευτερόλεπτα CPU)} μαζί με τον αριθμό επαναλήψεων.
    Ο κοινός αριθμός επαναλήψεων κρατά τους χρόνους συγκρίσιμους.
    """
    raw = {}
    for name, fn in pipelines.items():
        t0 = time.process_time()
        raw[name] = (fn(), time.process_time() - t0)

    times = [t for _, t in raw.values()]
    t_min, t_max = min(times), max(times)
    reps = MAX_REPEATS if t_min <= 0 else min(MAX_REPEATS,
                                              int(np.ceil(MIN_TIMED / t_min)))
    if t_max > 0:
        # φραγή του συνολικού χρόνου
        reps = min(reps, max(1, int((TIME_BUDGET / t_max))))
    if reps <= 1:
        return raw, 1

    out = {}
    for name, fn in pipelines.items():
        t0 = time.process_time()
        for _ in range(reps):
            fn()
        out[name] = (raw[name][0], (time.process_time() - t0) / reps)
    return out, reps

```

ΠΑΡΑΡΤΗΜΑ Β: ΑΝΑΛΥΤΙΚΟΙ ΠΙΝΑΚΕΣ ΜΕΤΡΗΣΕΩΝ

Στο παράρτημα αυτό παρατίθενται οι μετρήσεις που δεν συμπεριλήφθηκαν στο Κεφάλαιο 7 για λόγους οικονομίας: η ακρίβεια και η ανάκληση, οι χρόνοι CPU και για τις δύο συνθήκες, καθώς και η διάμεσος και το ποσοστό εφεδρικής τιμής των κατανομών k . Σε όλους τους πίνακες η τελευταία γραμμή δίνει τον μέσο όρο των δεκατεσσάρων συνόλων δεδομένων.

B.1 Ακρίβεια και ανάκληση

Πίνακας B.1: Ακρίβεια (precision, %) στα καθαρά δεδομένα.

ds	best global k	AdaNN $k \leq 9$	AdaNN $k \leq 30$	w=3, s=1	w=3, s=2	w=3, s=3
bl	60,22	54,30	55,85	57,57	57,82	57,57
bn	92,49	90,83	90,63	91,48	92,03	91,07
ecl	94,58	94,31	92,86	95,40	92,74	94,49
iris	94,44	94,44	94,44	94,44	94,44	94,44
lir	95,66	95,44	95,26	93,91	93,83	93,75
ls	89,94	89,55	89,04	88,25	88,42	88,18
mgt	85,63	82,19	82,41	87,77	88,22	87,42
pd	99,06	98,84	98,85	98,87	98,87	98,87
ph	84,19	86,08	87,08	84,06	83,20	83,59
pm	73,08	70,45	72,50	72,50	75,68	73,68
sh	91,31	91,73	91,73	98,31	98,31	99,41
tn	98,10	96,73	96,73	98,57	98,57	98,43
txr	98,74	98,66	98,66	98,21	98,21	98,21
ys	39,56	62,34	56,70	56,46	56,10	56,20
M.O.	85,50	86,14	85,91	86,84	86,89	86,81

Πίνακας B.2: Ανάκληση (recall, %) στα καθαρά δεδομένα.

ds	best global k	AdaNN $k \leq 9$	AdaNN $k \leq 30$	w=3, s=1	w=3, s=2	w=3, s=3
bl	65,50	59,12	60,82	61,68	61,68	61,68
bn	86,75	86,13	86,13	84,47	83,64	84,47
ecl	92,73	88,28	83,90	91,40	87,03	88,07
iris	93,33	93,33	93,33	93,33	93,33	93,33
lir	95,69	95,49	95,30	93,90	93,82	93,74
ls	89,30	89,17	88,77	87,96	88,12	87,87
mgt	65,68	65,53	64,13	59,47	58,73	59,62
pd	99,06	98,83	98,83	98,84	98,84	98,84
ph	76,32	73,21	73,52	65,73	64,80	66,67
pm	35,85	58,49	54,72	54,72	52,83	52,83
sh	92,41	90,17	90,17	89,71	89,71	89,71
tn	97,56	95,94	96,08	93,10	93,10	93,23
txr	98,78	98,70	98,70	98,24	98,24	98,24
ys	37,94	58,39	54,24	53,56	53,41	53,59
M.O.	80,49	82,20	81,33	80,44	79,81	80,13

Πίνακας B.3: Ακρίβεια (precision, %) στα θορυβώδη δεδομένα.

ds	best global k	AdaNN $k \leq 9$	AdaNN $k \leq 30$	w=3, s=1	w=3, s=2	w=3, s=3
bl	57,53	57,22	54,95	55,63	56,58	56,28
bn	84,20	69,81	71,58	78,03	81,40	80,46
ecl	73,44	70,48	53,53	61,99	61,99	73,37
iris	96,97	90,74	87,78	90,24	94,44	94,44
lir	92,91	92,01	91,60	89,81	89,88	90,05
ls	88,13	83,07	84,19	83,16	83,61	82,76
mgt	80,55	58,98	62,49	71,19	75,40	72,20

ds	best global k	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
pd	98,33	95,30	95,35	94,62	94,79	94,69
ph	70,38	60,07	62,59	66,15	72,03	68,77
pm	64,10	56,45	58,93	61,36	64,10	59,52
sh	64,93	49,02	49,79	55,28	61,67	61,65
tn	96,93	81,34	85,22	90,91	92,07	90,61
txr	97,14	95,82	95,81	93,78	93,77	94,11
ys	45,78	55,40	54,49	54,23	52,55	52,07
M.O.	79,38	72,55	72,02	74,74	76,73	76,50

Πίνακας B.4: Ανάκληση (recall, %) στα θορυβώδη δεδομένα.

ds	best global k	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	62,21	57,46	55,83	53,83	53,83	53,83
bn	83,85	68,94	71,43	70,60	69,77	72,46
ecl	63,61	56,63	52,26	50,77	50,77	63,09
iris	96,67	90,00	86,67	90,00	93,33	93,33
lir	92,89	91,99	91,50	89,63	89,70	89,89
ls	86,81	82,82	83,83	82,89	83,22	82,27
mgt	60,36	58,28	57,54	53,18	52,14	55,33
pd	98,32	95,16	95,22	94,25	94,46	94,35
ph	57,01	55,76	57,32	53,58	52,96	54,21
pm	47,17	66,04	62,26	50,94	47,17	47,17
sh	54,61	85,93	84,59	78,03	78,11	78,12
tn	93,91	75,51	77,27	73,07	73,88	75,78
txr	97,16	95,74	95,72	93,32	93,33	93,69
ys	38,44	49,42	46,24	46,23	46,02	45,97
M.O.	73,79	73,55	72,69	70,02	69,91	71,39

B.2 Χρόνοι CPU

Οι χρόνοι αναφέρονται σε δευτερόλεπτα χρόνου CPU διεργασίας και καλύπτουν ολόκληρο τον αγωγό, δηλαδή φάση εκπαίδευσης και φάση ελέγχου. Για τα μικρά σύνολα δεδομένων πρόκειται για μέσο όρο πολλαπλών εκτελέσεων, σύμφωνα με το πρωτόκολλο της ενότητας 7.2.

Πίνακας B.5: Χρόνος CPU (δευτερόλεπτα), καθαρά δεδομένα.

ds	best global k	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	0,16	0,01	0,01	0,01	0,01	0,01
bn	0,52	0,02	0,04	0,05	0,04	0,04
ecl	0,16	0,01	0,01	0,01	0,01	0,01
iris	0,16	0,00	0,01	0,01	0,00	0,00
lir	5,61	3,42	3,88	3,91	4,08	4,00
ls	1,48	0,48	0,61	0,67	0,62	0,61
mgt	2,30	0,88	1,27	1,14	1,28	1,36
pd	2,38	1,05	1,17	1,16	1,25	1,19
ph	0,58	0,06	0,09	0,11	0,09	0,09
pm	0,20	0,01	0,01	0,02	0,02	0,02
sh	6,95	2,58	4,16	4,16	4,08	5,25
tn	1,56	0,50	0,58	0,62	0,69	0,80
txr	1,41	0,44	0,50	0,61	0,50	0,55
ys	0,28	0,03	0,04	0,05	0,05	0,04
M.O.	1,70	0,68	0,88	0,89	0,91	1,00

Πίνακας Β.6: Χρόνος CPU (δευτερόλεπτα), θορυβώδη δεδομένα.

ds	best global k	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	0,18	0,01	0,01	0,01	0,02	0,02
bn	0,64	0,03	0,05	0,05	0,05	0,06
ecl	0,15	0,01	0,01	0,01	0,01	0,01
iris	0,14	0,00	0,00	0,00	0,00	0,00
lir	5,05	3,44	3,80	3,75	3,62	3,75
ls	1,66	0,64	0,78	0,75	0,75	0,70
mgt	2,52	1,28	1,38	1,27	1,16	1,20
pd	2,45	1,05	1,36	1,36	1,36	1,30
ph	0,61	0,06	0,11	0,11	0,09	0,14
pm	0,20	0,01	0,02	0,02	0,02	0,02
sh	7,30	2,61	4,38	8,08	4,30	3,81
tn	1,62	0,56	0,80	0,88	0,70	0,84
txr	1,41	0,55	0,69	0,62	0,62	0,62
ys	0,27	0,03	0,04	0,04	0,04	0,04
M.O.	1,73	0,73	0,96	1,21	0,91	0,89

Β.3 Κατανομή των τιμών k

Η διάμεσος τιμή k συμπληρώνει τη μέση τιμή που δόθηκε στο Κεφάλαιο 7 και είναι ανθεκτικότερη στην επίδραση των σημείων που καταλήγουν στην εφεδρική τιμή. Η αντιπαραβολή των δύο μεγεθών είναι διαφωτιστική: σε αρκετά σύνολα η διάμεσος παραμένει πολύ χαμηλή ενώ ο μέσος όρος είναι υψηλός, ένδειξη ότι η κατανομή είναι έντονα διμερής, με μια μεγάλη μάζα σημείων σε πολύ μικρές τιμές και μια δεύτερη συγκέντρωση στην εφεδρική τιμή.

Πίνακας Β.7: Διάμεσος τιμή k, καθαρά δεδομένα.

ds	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	1,0	1,0	2,0	2,0	2,0
bn	1,0	1,0	2,0	2,0	2,0
ecl	1,0	1,0	2,0	2,0	2,0
iris	1,0	1,0	2,0	2,0	2,0
lir	1,0	1,0	2,0	2,0	2,0
ls	1,0	1,0	2,0	2,0	2,0
mgt	1,0	1,0	2,0	2,0	2,0
pd	1,0	1,0	2,0	2,0	2,0
ph	1,0	1,0	2,0	2,0	2,0
pm	1,0	1,0	2,0	2,0	2,0
sh	1,0	1,0	2,0	2,0	2,0
tn	1,0	1,0	2,0	2,0	2,0
txr	1,0	1,0	2,0	2,0	2,0
ys	1,0	1,0	4,0	4,0	5,0
M.O.	1,0	1,0	2,1	2,1	2,2

Πίνακας Β.8: Διάμεσος τιμή k, θορυβώδη δεδομένα.

ds	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	1,0	1,0	4,0	4,0	5,0
bn	1,0	1,0	4,0	4,0	5,0
ecl	2,0	2,0	4,0	4,0	5,0
iris	1,0	1,0	4,0	4,0	5,0
lir	2,0	2,0	4,0	4,0	5,0
ls	2,0	2,0	4,0	4,0	5,0
mgt	1,0	1,0	4,0	4,0	5,0
pd	1,0	1,0	4,0	4,0	5,0

ds	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
ph	1,0	1,0	4,0	4,0	5,0
pm	1,0	1,0	4,0	6,0	5,0
sh	1,0	1,0	3,0	4,0	5,0
tn	1,0	1,0	4,0	4,0	5,0
txr	2,0	2,0	4,0	4,0	5,0
ys	5,0	5,0	16,0	16,0	17,0
M.O.	1,6	1,6	4,8	5,0	5,9

Πίνακας Β.9: Ποσοστό (%) σημείων εκπαίδευσης στην εφεδρική τιμή $k = \max_k$, καθαρά δεδομένα.

ds	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	8,6	7,6	9,0	9,4	9,4
bn	6,0	4,7	6,8	7,1	7,2
ecl	12,6	11,2	15,2	15,6	15,2
iris	3,3	1,7	4,2	4,2	4,2
lir	1,7	1,2	2,5	2,6	2,7
ls	4,8	4,0	6,1	6,4	6,6
mgt	9,2	8,0	11,9	12,1	12,4
pd	0,3	0,3	0,6	0,6	0,6
ph	4,7	3,4	6,9	7,4	7,5
pm	13,8	10,7	17,4	18,2	18,7
sh	0,0	0,0	0,1	0,1	0,1
tn	1,6	1,1	1,8	1,9	1,9
txr	0,6	0,5	1,0	1,1	1,1
ys	26,0	21,0	28,9	29,9	30,0
M.O.	6,7	5,4	8,0	8,3	8,4

Πίνακας Β.10: Ποσοστό (%) σημείων εκπαίδευσης στην εφεδρική τιμή $k = \max_k$, θορυβώδη δεδομένα.

ds	AdaNN k≤9	AdaNN k≤30	w=3, s=1	w=3, s=2	w=3, s=3
bl	22,6	19,8	28,0	28,6	29,0
bn	19,2	13,8	23,1	24,4	24,6
ecl	36,4	33,8	35,7	35,7	37,9
iris	27,5	25,8	33,3	34,2	34,2
lir	31,5	30,5	32,1	32,2	32,4
ls	30,4	28,5	32,8	33,1	33,3
mgt	19,5	14,8	24,9	25,9	26,8
pd	28,1	27,8	29,5	29,5	29,5
ph	19,4	14,5	23,7	25,0	25,7
pm	17,7	11,1	22,8	24,7	26,0
sh	28,2	28,0	29,9	29,9	29,9
tn	18,2	14,6	23,9	25,1	25,1
txr	28,5	28,1	29,6	29,6	29,7
ys	45,5	38,5	45,6	46,2	46,6
M.O.	26,6	23,5	29,6	30,3	30,8