



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Ανάλυση και Επεξεργασία Γεωχωρικών Δεδομένων


Εαρινό Εξάμηνο 2023-2024

Άσκηση 3

Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων

Ονοματεπώνυμο: Κωνσταντίνος Πριμέτης
ID: 03400231

Ιούλιος 2024

	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων	ΙΟΥΛΙΟΣ 2024
	ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	Σελίδα 1 / 12

ΠΕΡΙΕΧΟΜΕΝΑ


ΕΙΣΑΓΩΓΗ	2
ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ.....	3
ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ	5
ΕΚΠΑΙΔΕΥΣΗ - ΑΞΙΟΛΟΓΗΣΗ.....	6
ΣΧΟΛΙΑΣΜΟΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.....	7

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ

<i>Σχήμα 1. Μέση τιμή και τυπική απόκλιση των τιμών των μετρικών που προέκυψαν από τα 5 πειράματα.</i>	8
<i>Σχήμα 2. Τιμή κόστους ανά πείραμα.</i>	9
<i>Σχήμα 3. Πίνακας Σύγχυσης – Πείραμα 3.</i>	10
<i>Σχήμα 4. Πίνακας Σύγχυσης – Πείραμα 5.</i>	11

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

<i>Πίνακας 1. Micro μετρικές.....</i>	7
<i>Πίνακας 2. Weighted μετρικές.....</i>	7


	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 2 / 12

ΕΙΣΑΓΩΓΗ

Η παρούσα εργασία αποτελεί την τρίτη από μια σειρά τριών εργασιών στο πλαίσιο του Μαθήματος «Ανάλυση και Επεξεργασία Γεωχωρικών Δεδομένων» του ΔΠΜΣ «Επιστήμη Δεδομένων και Μηχανική Μάθηση».

Στόχος της συγκεκριμένης εργασίας αποτελεί η εξοικείωση με τηλεπισκοπικά δεδομένα χρονοσειρών μέσα από τον σχεδιασμό και υλοποίηση μιας μεθοδολογίας ταξινόμησης χρονοσειρών Τεχνητών Νευρωνικών Δικτύων. Πιο συγκεκριμένα, πραγματοποιείται λήψη και επεξεργασία γεωχωρικών δεδομένων τύπου .zarr, προετοιμασία των δεδομένων για εκπαίδευση, σχεδιασμός μοντέλου με χρήση αρχιτεκτονικών Transformer για ταξινόμηση αγροτεμαχίων σε κάποια εκ των διαθέσιμων κατηγοριών καλλιέργειας, εκπαίδευση και τέλος αξιολόγηση του μοντέλου.

Η παρούσα έκθεση αποτελεί αναπόσπαστο τμήμα του notebook “Primetis_GBDA2024_Ex3”, το οποίο αποτελεί το κύριο παραδοτέο της εργασίας. Ακολουθώς, διακριτοποιείται και παρουσιάζεται σε τέσσερις (4) ενότητες, βάσει της ροής που ακολουθήθηκε για την ανάπτυξη του μοντέλου. Καθώς δεν τίθενται συγκεκριμένα ζητούμενα από την εκφώνηση της άσκησης, στόχος της συμπληρωματικής έκθεσης αποτελεί η περιγραφή των διαδικασιών που ακολουθήθηκαν κατά την ανάπτυξη, εκπαίδευση και τέλος αξιολόγηση του μοντέλου σε συνδυασμό με τον σχολιασμό της απόδοσής του βάσει των ζητούμενων μετρικών.

	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 3 / 12

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ


Αρχικά, λήφθηκαν τα δεδομένα τα οποία αποτελούνται από ένα σύνολο αρχείων τύπου .zarr συνοδευόμενα από τα μεταδεδομένα τους. Αποτελούν υποσύνολο του συνόλου δεδομένων Timematch, το οποίο περιέχει ετήσιες χρονοσειρές πολυφασματικών δεδομένων Sentinel-2 σε τέσσερις (4) περιοχές της Ευρώπης που λήφθηκαν το έτος 2017. Στη συγκεκριμένη άσκηση έχουν διατηρηθεί δεδομένα που αφορούν μόνο στην περιοχή της Δανίας. Το συγκεκριμένο υποσύνολο δεδομένων πραγματεύεται την ταξινόμηση του εκάστοτε αγροτεμαχίου σε μία εκ των διαθέσιμων κατηγοριών καλλιέργειας (15 στο αρχικό σύνολο).

Η δομή των δεδομένων απαρτίζεται από αγροτεμάχια (η βασική δομή του συνόλου) καθένα από τα οποία αποτελείται από ένα μη σταθερό πλήθος εικονοστοιχείων, ενώ για κάθε εικονοστοιχείο είναι διαθέσιμη μία χρονοσειρά πολυφασματικών δεδομένων (10 κανάλια) με σταθερό πλήθος διαθέσιμων ημερομηνιών, οι οποίες είναι διαθέσιμες στα μεταδεδομένα και κοινές για όλα τα αγροτεμάχια.

Ξεκινώντας την προεπεξεργασία των δεδομένων, εντοπίζονται οι κατηγορίες καλλιέργειας στο δοθέν υποσύνολο, οι οποίες είναι δεκατέσσερις (14). Φιλτράρονται εκτός συνόλου όσες έχουν λιγότερα από διακόσια (200) δειγματικά στοιχεία (σύμφωνα με την υπόδειξη της άσκησης) κι έτσι τελικά απομένουν οχτώ (8) μία εκ των οποίων είναι η κλάση «unknow» (σχολιάζεται εκτενέστερα στο τελευταίο κεφάλαιο της παρούσας έκθεσης). Για τις εναπομείναντες κατηγορίες ορίζεται νέα δεικτοδότηση με ακέραιους αριθμούς, για την καλύτερη μετέπειτα επεξεργασία.


Κατά την προεπεξεργασία των δεδομένων, λαμβάνονται και οι 52 ημερομηνίες που αντιστοιχούν στον χρόνο λήψης. Τελικά αποθηκεύονται ως ακέραιοι (αρχικά σε λίστα) καθένας από τους οποίους αντιστοιχεί στη σειρά της ημερομηνίας λήψης των δειγμάτων μέσα στο έτος 2017 (13^η, 53^η, ..., 363^η ημέρα του 2017). Με αυτόν τον τρόπο (άλλος τύπος δεδομένων – τανυστής) θα εισαχθούν μετέπειτα στο μοντέλο.

Επόμενο στάδιο της επεξεργασίας των δεδομένων αποτελεί η δημιουργία τριών κλάσεων για τον μετασχηματισμό τους. Οι κλάσεις αυτές αφορούν στην τυχαία δειγματοληψία συγκεκριμένου αριθμού εικονοστοιχείων (32) στο σύνολο εκπαίδευσης, κανονικοποίησης και μετατροπής δεδομένων σε τανυστές. Ακόμα, δημιουργείται συνάρτηση η οποία θα εφαρμόσει padding στα

	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 4 / 12

δεδομένα του συνόλου επικύρωσης, έτσι ώστε σε κάθε batch, η διάσταση που αφορά στα εικονοστοιχεία να είναι κοινή (αυτή του μεγαλύτερου δειγματικού στοιχείου στο batch).

Τέλος, δημιουργείται κλάση Dataset (PixelSetData), για τη λήψη δεδομένων και των ετικετών από το δοθέν σύνολο δεδομένων, που χρησιμεύει στη δημιουργία των συνόλων δεδομένων εκπαίδευσης και επικύρωσης.

	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 5 / 12


ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ

Αρχικά σχεδιάζεται μια αρχιτεκτονική Pixel Set Encoder, η οποία έχει ως στόχο την ενσωμάτωση της χωρικής διάστασης των δεδομένων στο μοντέλο. Πιο συγκεκριμένα, αποτελείται από ένα MLP με δύο στρώσεις (Batch Normalization και ReLu συνάρτηση ενεργοποίησης) για την εξαγωγή μιας ενδιάμεσης αναπαράστασης για κάθε εικονοστοιχείο. Στη συνέχεια υπολογίζονται μέση τιμή και τυπική απόκλιση στη διάσταση των εικονοστοιχείων και η ενιαία αναπαράσταση που προκύπτει από τη συνένωσή τους εισάγεται σε ένα δεύτερο MLP για τη περαιτέρω κωδικοποίηση και την τελική ενσωμάτωση για το σύνολο των εικονοστοιχείων.

Στη συνέχεια, με την κλάση PositionalEncoding χρησιμοποιείται ένα στρώμα encoder και υπολογίζεται η κωδικοποίηση της θέσης βάσει των ημερομηνιών της αρχικής λίστας. Σε αυτό το σημείο, η λίστα μετατρέπεται (μέσω κατάλληλης συνάρτησης) σε τανυστή που θα έχει διαστάσεις τέτοιες, ώστε να προσαρμόζεται στον αριθμό των pixels που θα έχουν τελικά τα δειγματικά στοιχεία κάθε batch. Η ενσωμάτωση της χρονικής διάστασης επιτυγχάνεται τελικά μέσω της χρήσης ημι- και συνημιτονοειδών συναρτήσεων.

Περαιτέρω, αναπτύσσεται κλάση που θα πραγματοποιήσει την ταξινόμηση. Αποτελείται από δύο πλήρως συνδεδεμένα στρώματα, μια ReLU συνάρτηση ενεργοποίησης και μία στρώση dropout, εξάγοντας την τελική πρόβλεψη του μοντέλου.

Τέλος, ορίζεται η κλάση του μοντέλου που συνδυάζει τις παραπάνω λειτουργίες σε μια ενιαία αρχιτεκτονική.

	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 6 / 12

ΕΚΠΑΙΔΕΥΣΗ – ΑΞΙΟΛΟΓΗΣΗ

Επόμενο στάδιο αποτελεί η εκπαίδευση του μοντέλου. Σε αυτό το σημείο κρίνεται αναγκαίο να σχολιαστούν κάποιες από τις παραμέτρους που επιλέχθηκαν για την αρχικοποίησή του. Ο αριθμός των εποχών ορίστηκε σε τριάντα (30) για καθένα από τα πειράματα του 5-fold cross validation. Παρατηρήθηκε ότι χρειαζόνταν ένας ικανός αριθμός εποχών (πάνω από 15) για να υπάρξει μείωση στην τιμή της συνάρτησης κόστους. Έτσι, τελικά επιλέχθηκαν οι 30 εποχές. Για την αποφυγή overfitting, τελικά για κάθε πείραμα κρατήθηκαν οι τιμές των μετρικών που αντιστοιχούσαν στην εποχή με τη χαμηλότερη τιμή της val_loss (αντί αυτών της τελευταίας εποχής).

Ως αριθμός των δειγματικών στοιχείων ανά batch επιλέχθηκε το 8. Η πρώτη δοκιμή αφορούσε σε batch size = 32, αλλά τα αποτελέσματα δεν ήταν ικανοποιητικά. Στη συνέχεια, δοκιμάστηκαν 16 και 8 και για το συγκεκριμένο μοντέλο επιλέχθηκε το 8. Η τιμή που αναφέρεται στην αρχική δημοσίευση (128, που βέβαια αφορούσε σε μεγαλύτερο σύνολο δεδομένων) δε δοκιμάστηκε.

Ο ρυθμός εκμάθησης επιλέχθηκε να είναι 10^{-4} . Η αρχική δοκιμή έγινε με 10^{-3} , όμως το μοντέλο παρουσιάζοντας μεγάλη αστάθεια, δε συνέκλινε (τουλάχιστον για τον αριθμό εποχών που δοκιμάστηκε), με τις τιμές της val_loss να πιάνουν πολύ μεγάλες τιμές.

Για την εκπαίδευση του μοντέλου επιλέχθηκε απλή Pytorch (και όχι Lightning module), για καλύτερο «χειροκίνητο» έλεγχο. Πολλές φορές απαιτήθηκε η προσθήκη εκτυπώσεων για τη διόρθωση διαστάσεων δεδομένων που εισάγονταν ή εξάγονταν από τα blocks ή τα στρώματα του μοντέλου και λόγω μεγαλύτερης εξοικείωσης επιλέχθηκε η απλή Pytorch. Ως συνάρτηση κόστους επιλέχθηκε η Cross Entropy και ως optimizer ο Adam.

Τέλος, οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν είναι αυτές που υποδεικνύονται από την εκφώνηση της άσκησης. Πιο συγκεκριμένα, εξήχθησαν πίνακες σύγχυσης για καθένα από τα πέντε (5) πειράματα και αξιολογήθηκαν οι micro και weighted μετρικές F1, Accuracy, Precision και Recall.

ΣΧΟΛΙΑΣΜΟΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Στο παρόν κεφάλαιο επιχειρείται σχολιασμός και ερμηνεία των αποτελεσμάτων, όπως αυτά προέκυψαν από τις μετρικές που αναφέρονται ανωτέρω. Συμπληρωματικά, για την πληρέστερη παρουσίαση τους, παρατίθενται ενδεικτικά διαγράμματα.

Πίνακας 1. Micro μετρικές.

Fold	Accuracy	F1	Precision	Recall
1	0,7841	0,7841	0,7841	0,7841
2	0,8206	0,8206	0,8206	0,8206
3	0,7933	0,7933	0,7933	0,7933
4	0,8846	0,8846	0,8846	0,8846
5	0,8741	0,8741	0,8741	0,8741

Πίνακας 2. Weighted μετρικές.

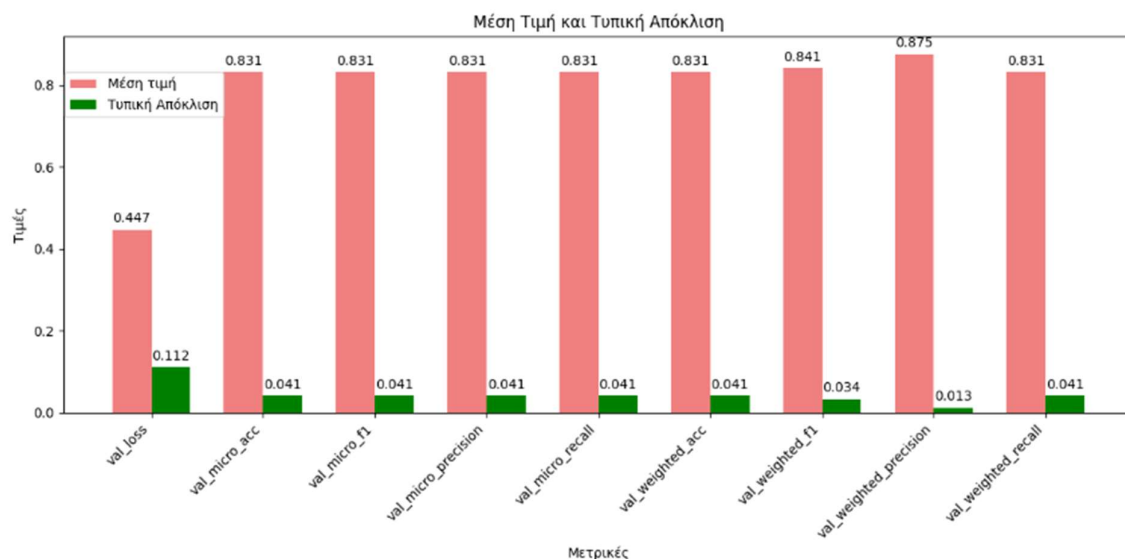
Fold	Accuracy	F1	Precision	Recall
1	0,7841	0,8036	0,8782	0,7841
2	0,8206	0,8282	0,8524	0,8206
3	0,7933	0,8126	0,8732	0,7933
4	0,8846	0,8864	0,8911	0,8846
5	0,8741	0,8755	0,8792	0,8741

Από τον Πίνακα 1 γίνεται αντιληπτό, ότι όλες οι *micro* μετρικές έχουν ίδια τιμή για κάθε πείραμα. Αυτό πραγματικά ισχύει, γιατί σε αυτή την περίπτωση οι υπολογισμοί γίνονται σε επίπεδο δειγματικού στοιχείου όχι σε επίπεδο κατηγορίας. Οι αληθώς και ψευδώς θετικές και οι ψευδώς θετικές και αρνητικές τιμές αθροίζονται για όλες τις κατηγορίες και στη συνέχεια υπολογίζονται οι μετρικές.


Φαίνεται ότι υπάρχει μια σημαντική απόκλιση μεταξύ των ακραίων τιμών, της τάξης περίπου του 0,1. Η απόκλιση αυτή εξηγείται λόγω της παρουσίας της κατηγορίας «unknown», όπως θα γίνει και πιο έντονα φανερό κατά την παρουσίαση των πινάκων σύγχυσης.

Ο Πίνακας 2 παρουσιάζει τα αποτελέσματα των *weighted* μετρικών. Και σε αυτή την περίπτωση, οι τιμές των μετρικών παρουσιάζουν ένα εύρος από 0.78 έως 0.89. Ανάμεσα στις μετρικές, τη μεγαλύτερη τιμή παρουσιάζει εν γένει η μετρική *Precision*, φανερώνοντας καλύτερη πρόβλεψη του δείκτη των αληθώς θετικών προβλέψεων προς το σύνολο των θετικών προβλέψεων (αληθώς και ψευδώς) έναντι των υπόλοιπων μετρικών.

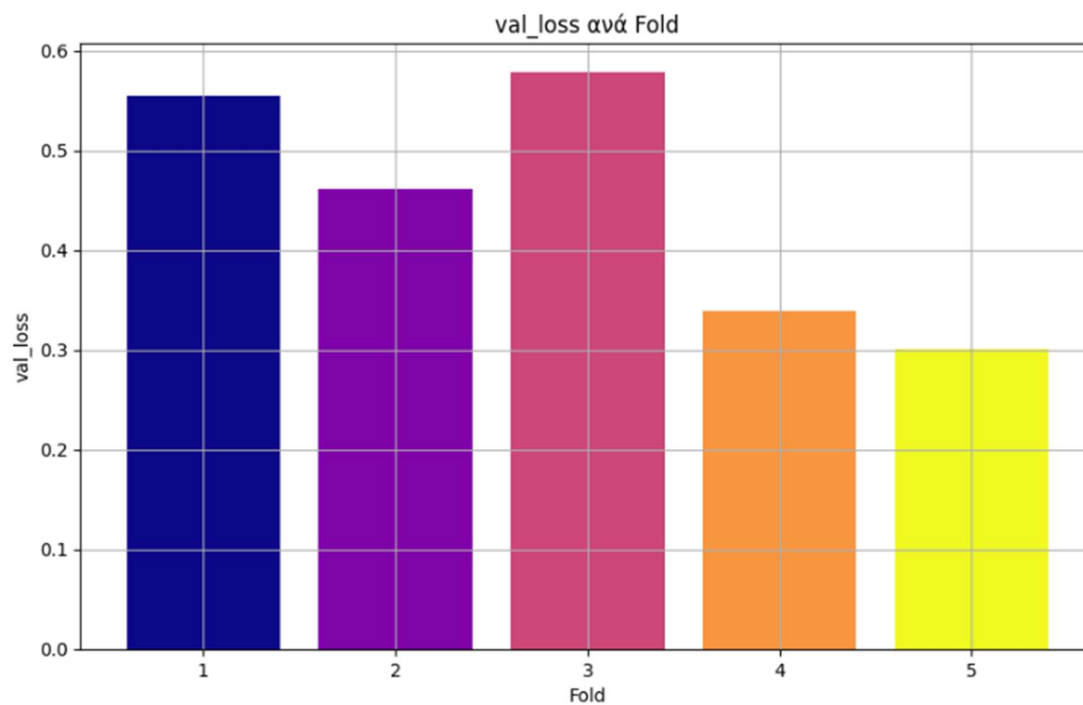
Στο Σχήμα 1 που ακολουθεί παρουσιάζονται η μέση τιμή και η τυπική απόκλιση των τιμών που προέκυψαν για τις μετρικές από τα πέντε (5) πειράματα καθώς και της συνάρτησης κόστους.



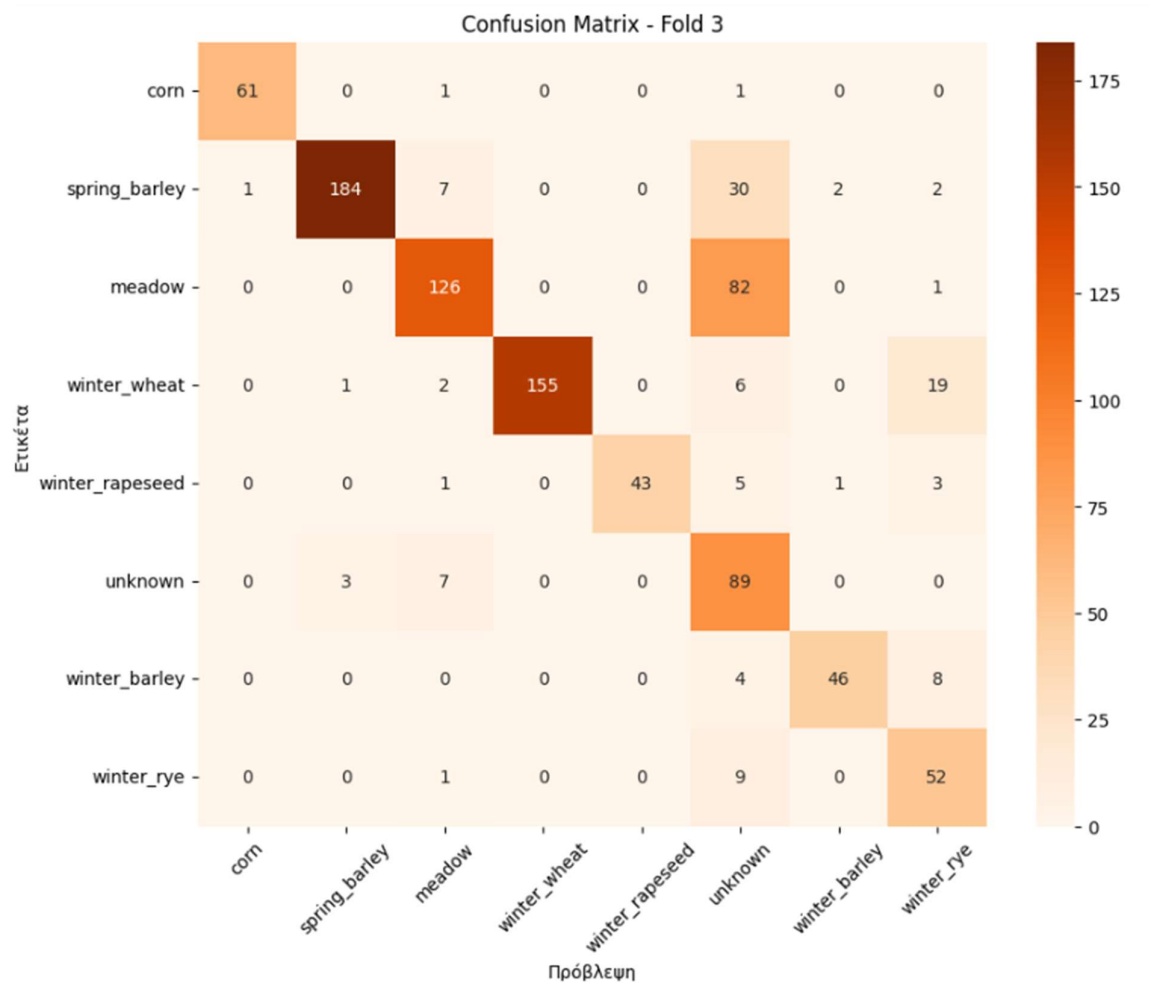
Σχήμα 1. Μέση τιμή και τυπική απόκλιση των τιμών των μετρικών που προέκυψαν από τα 5 πειράματα.

	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 9 / 12

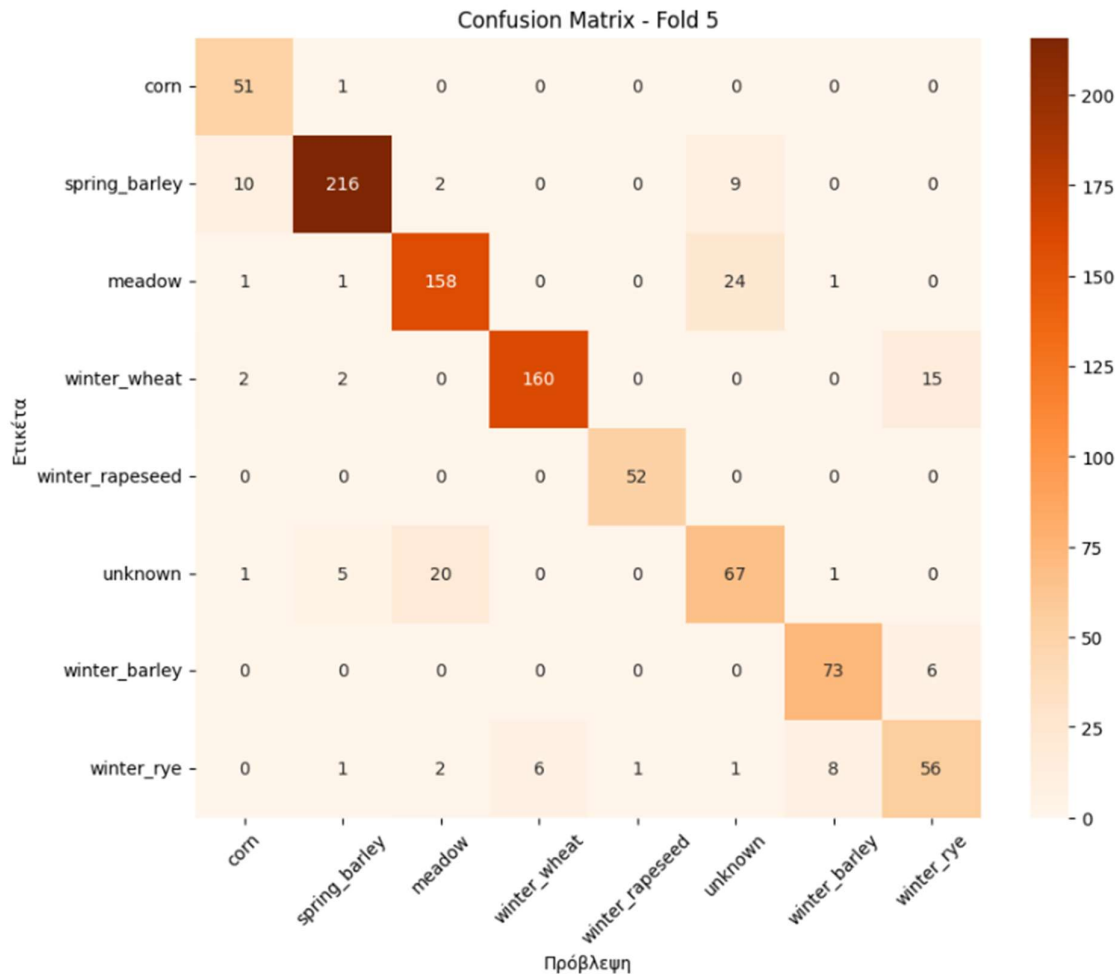
Ακολουθώς, στο Σχήμα 2 που ακολουθεί παρουσιάζονται οι τιμές της συνάρτησης κόστους για καθένα από τα πειράματα που εκτελέστηκαν. Βάσει αυτού, παρουσιάζονται ενδεικτικά οι πίνακες σύγκρισης για τα πειράματα με την υψηλότερη και τη χαμηλότερη τιμή, έτσι ώστε να εξηγηθεί η απόκλιση που υπάρχει στις τιμές των μετρικών μεταξύ τους.



Σχήμα 2. Τιμή κόστους ανά πείραμα.




Σχήμα 3. Πίνακας Σύγχυσης – Πείραμα 3.



Σχήμα 4. Πίνακας Σύγχυσης – Πείραμα 5.

Από τον πίνακα σύγχυσης με την καλύτερη επίδοση (Σχήμα 4) γίνεται αντιληπτό ότι η κύρια διαγώνιος έχει τις μεγαλύτερες τιμές, που επιδεικνύουν σωστή λειτουργία του μοντέλου. Οι αστοχίες που εμφανίζονται, εμπλέκουν κατά κύριο λόγο την κατηγορία «unknown». Ωστόσο, δεν εμφανίζουν υψηλές τιμές.

Αντίθετα, στο Σχήμα 3 οι αντίστοιχες αστοχίες είναι πολύ πιο έντονες. Η κατηγορία «unknown» αποτελεί μια ιδιαίτερη κλάση του μοντέλου, η οποία δύναται να περιέχει δειγματικά στοιχεία που ανήκουν σε οποιαδήποτε από τις υπόλοιπες κλάσεις, όχι μόνο του υποσυνόλου δεδομένων της άσκησης αλλά και του αρχικού συνόλου δεδομένων ή ακόμη και σε νέες κατηγορίες.

	ΑΝΑΛΥΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΓΕΩΧΩΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ Ταξινόμηση χρονοσειρών πολυφασματικών δορυφορικών δεδομένων ΣΥΝΟΔΕΥΤΙΚΗ ΕΚΘΕΣΗ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 12 / 12

Οπότε, στις περιπτώσεις που το σύνολο εκπαίδευσης περιέχει περισσότερα δειγματικά στοιχεία της κλάσης «unknown» λαμβάνονται περισσότερες λανθασμένες προβλέψεις, που αφορούν στην κατηγορία και τιμές μετρικών σημαντικά χαμηλότερες. Μια προσέγγιση που σίγουρα θα βελτίωνε την απόδοση του (διαφορετικού) μοντέλου που θα πρόκυπτε θα ήταν το φιλτράρισμα των δειγματικών στοιχείων της κατηγορίας εκτός του συνόλου δεδομένων της άσκησης. Ωστόσο, η εκφώνηση της άσκησης ήταν σαφής για τις κλάσεις που θα αφαιρεθούν και έτσι η εν λόγω κλάση τελικά διατηρήθηκε.