Γεφυρώνοντας μεγάλα ομικά, γενετικά και ιατρικά δεδομένα για την ευρεία εφαρμογή της Ιατρικής Ακριβείας στην Ελλάδα

ΠΑΡΑΔΟΤΕΟ Π6.3

«Μέθοδοι ανάλυσης κειμένου και μοντέλα επεξεργασίας φυσικής γλώσσας»

Φορέας	Πανεπιστήμιο Θεσσαλίας
Τύπος Παραδοτέου	Έκθεση
Ημερομηνία Υποβολής Παραδοτέου	31 Δεκεμβρίου 2025
Ενότητα Εργασίας	Ενότητα Εργασίας 6 «Ανάπτυξη Μοντέλων Μηχανικής Μάθησης για Εξόρυξη Γνώσης και Κατηγοριοποίησης Ασθενών»

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Πίνακας Εικόνων

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

1 Εισαγωγή

Η ολοένα αυξανόμενη ψηφιοποίηση των ιατρικών δεδομένων και η συστηματική καταγραφή των κλινικών επεισοδίων μέσω ηλεκτρονικών ιατρικών φακέλων (Electronic Health Records – EHRs) δημιουργεί νέες δυνατότητες αλλά και σημαντικές προκλήσεις στην κατανόηση, οργάνωση και αξιοποίησή τους. Στο πλαίσιο της Ενότητας Εργασίας ΕΕ6, το παρόν Παραδοτέο Π6.3 εστιάζει στην ανάπτυξη και αξιολόγηση μεθοδολογιών ανάλυσης ιατρικού κειμένου και μοντέλων επεξεργασίας φυσικής γλώσσας (NLP), με στόχο την αυτόματη εξαγωγή ιατρικών όρων, την παραγωγή ιατρικών λεξικών και τον αξιόπιστο χαρακτηρισμό των ιατρικών γνωματεύσεων. Σύμφωνα με την περιγραφή του φυσικού αντικειμένου της ΕΕ6, η αξιοποίηση των απεικονιστικών δεδομένων και των αντίστοιχων κλινικών αναφορών απαιτεί προηγμένες τεχνικές κατανόησης φυσικής γλώσσας και αυτόματης κωδικοποίησης ιατρικών εννοιών, προκειμένου να υποστηριχθεί η εξατομικευμένη θεραπευτική προσέγγιση και η ουσιαστική γεφύρωση ετερογενών ιατρικών δεδομένων.

Η ανάγκη αυτοματοποίησης είναι ιδιαίτερα έντονη στο πεδίο της ανάθεσης διαγνωστικών κωδικών ICD, διαδικασία που σήμερα εκτελείται χειρωνακτικά και απαιτεί σημαντικό χρόνο και εξειδίκευση. Η ερευνητική κοινότητα έχει αναπτύξει πληθώρα μοντέλων μάθησης για την αντιμετώπιση αυτής της πρόκλησης, από κλασικές μεθόδους ταξινόμησης έως σύγχρονα βαθιά νευρωνικά δίκτυα και attention-based αρχιτεκτονικές. Στο πλαίσιο του παρόντος παραδοτέου, η ομάδα ανέπτυξε και αξιολόγησε νέες προσεγγίσεις πολυετικετικής ταξινόμησης ιατρικού κειμένου, αξιοποιώντας τα δεδομένα των βάσεων MIMIC-III και MIMIC-IV. Συγκεκριμένα, υλοποιήθηκαν μέθοδοι Classifier Chains με τεχνητά νευρωνικά δίκτυα ως βασικούς ταξινομητές, επιτρέποντας την αξιοποίηση των αλληλεξαρτήσεων μεταξύ ICD ετικετών, στοιχείο κρίσιμο για τη βελτίωση της ακρίβειας στην ανάθεση κωδικών.

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Παράλληλα, αναπτύχθηκε μια δεύτερη ερευνητική κατεύθυνση που αφορά τη διεύρυνση και ομογενοποίηση των δεδομένων μέσω: (α) χαρτογράφησης ICD-9 σε ICD-10, προκειμένου να αυξηθεί η συνοχή του συνόλου εκπαίδευσης και να αρθούν περιορισμοί που προκύπτουν από την ταυτόχρονη παρουσία παλαιότερων και νεότερων κωδικοποιήσεων· και (β) δημιουργίας συνθετικών κλινικών κειμένων με χρήση μεγάλων γλωσσικών μοντέλων (LLMs), ώστε να ενισχυθούν σπάνιοι κωδικοί και να αντιμετωπιστούν φαινόμενα ακραίας ανισορροπίας των δεδομένων. Τα συνθετικά κείμενα ενσωματώθηκαν στο σύνολο εκπαίδευσης και αποδεδειγμένα ενίσχυσαν τη γενίκευση και την απόδοση πολλών βαθιών μοντέλων (CNN, MultiResCNN, LAAT, PLM-ICD) στην αυτόματη κωδικοποίηση ICD-10.

Οι μεθοδολογίες που αναπτύχθηκαν δεν περιορίζονται στην απλή ταξινόμηση ιατρικού κειμένου, αλλά συμβάλλουν στην παραγωγή προσαρμοσμένων ιατρικών λεξικών, στη δομημένη εξαγωγή κλινικής πληροφορίας και στην ενίσχυση της διαλειτουργικότητας μεταξύ ετερογενών πηγών ιατρικών δεδομένων. Συνολικά, το Παραδοτέο Π6.3 παρουσιάζει τα θεωρητικά θεμέλια, τις τεχνικές υλοποίησης και τα πειραματικά αποτελέσματα αυτών των προσεγγίσεων, αναδεικνύοντας τη συμβολή τους στην υλοποίηση των στόχων της Πράξης, καθώς και την άμεση αξιοποίησή τους στη μετάβαση προς συστήματα εξατομικευμένης ιατρικής.

2 Δεδομένα και Προεπεξεργασία

Η υλοποίηση των μεθοδολογιών που παρουσιάζονται στο παρόν παραδοτέο βασίστηκε στην αξιοποίηση πραγματικών κλινικών δεδομένων μεγάλης κλίμακας. Ως κύριες πηγές χρησιμοποιήθηκαν οι βάσεις MIMIC-III και MIMIC-IV, οι οποίες περιλαμβάνουν ανώνυμα ιατρικά αρχεία από νοσηλείες στις ΜΕΘ του Beth Israel Deaconess Medical Center. Οι δύο βάσεις αποτελούν διεθνώς αναγνωρισμένα πρότυπα για έρευνα στην κλινική πληροφορική και προσφέρουν έναν πλούτο δεδομένων, τόσο δομημένων όσο και αδόμητων, τα οποία επιτρέπουν την ανάπτυξη,

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

εκπαίδευση και αξιολόγηση αλγορίθμων μηχανικής μάθησης σε πραγματικές συνθήκες.

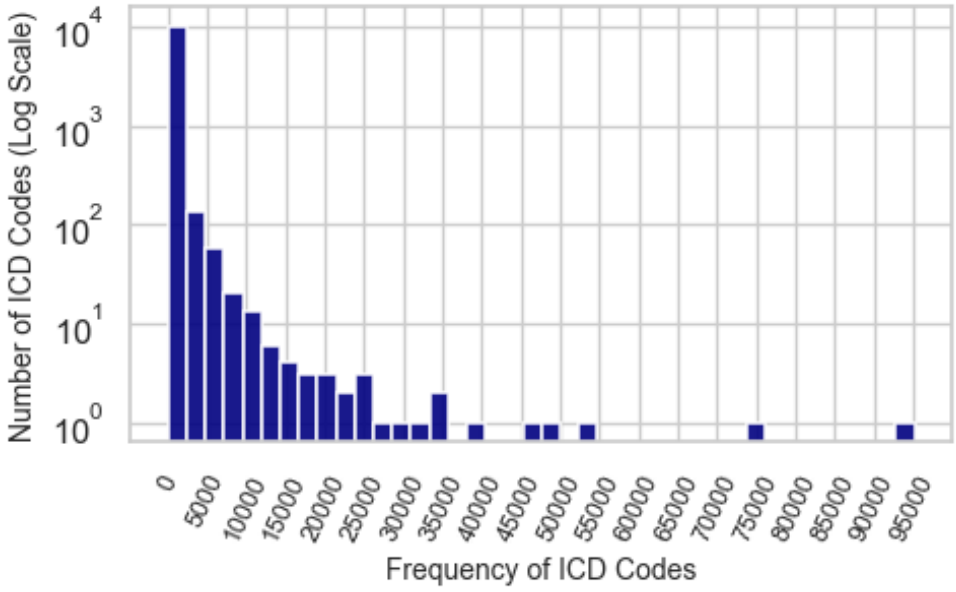
Η βάση MIMIC-III περιλαμβάνει πάνω από 52.000 κλινικές σημειώσεις και χρησιμοποιεί αποκλειστικά το διαγνωστικό σύστημα ICD-9. Αντίθετα, η MIMIC-IV, η οποία αντιπροσωπεύει νεότερη και πιο εκτεταμένη καταγραφή, χρησιμοποιεί συνδυασμό ICD-9 και ICD-10, γεγονός που εισάγει πρόσθετη πολυπλοκότητα όσον αφορά τη συνοχή της κωδικοποίησης. Σε αυτή τη βάση εντοπίζονται περισσότερα από 330.000 κείμενα τύπου discharge summary, καθώς και αναλυτικές πληροφορίες για διαγνώσεις και ιατρικές πράξεις. Το στοιχείο αυτό αποτέλεσε κομβικό σημείο για την ανάπτυξη του pipeline προεπεξεργασίας, καθώς η απουσία ενιαίου συστήματος κωδικοποίησης δυσχεραίνει τόσο την εκπαίδευση όσο και την αξιολόγηση των μοντέλων σε σύγχρονες κλινικές πρακτικές.

Η προεπεξεργασία των αδόμητων κειμένων πραγματοποιήθηκε με ιδιαίτερη προσοχή, καθώς η ποιότητα των χαρακτηριστικών που εξάγονται από το κείμενο καθορίζει σε μεγάλο βαθμό την απόδοση των μοντέλων. Αρχικά, όλα τα κείμενα μετατράπηκαν σε πεζούς χαρακτήρες, ενώ απομακρύνθηκαν αριθμητικά στοιχεία, σημεία στίξης, ειδικοί χαρακτήρες και placeholders που σχετίζονταν με προσωπικά δεδομένα. Η διαδικασία αυτή είναι κρίσιμη για την αποφυγή περιττών διακυμάνσεων στο λεξιλόγιο, καθώς και για τη διασφάλιση της ανωνυμίας των δεδομένων. Στη συνέχεια εφαρμόστηκε tokenization και φίλτρα για τον αποκλεισμό tokens χωρίς σημασιολογική αξία. Ιδιαίτερη προσοχή δόθηκε στη διαχείριση των stopwords· σε αντίθεση με παραδοχές κλασικών μοντέλων, τα stopwords διατηρήθηκαν κατά την εκπαίδευση βαθιών νευρωνικών δικτύων, δεδομένου ότι συμβάλλουν στη διατήρηση της τοπικής και γραμματικής δομής, στοιχείο απαραίτητο για την ορθή λειτουργία attention-based αρχιτεκτονικών.

Παράλληλα με τον καθαρισμό του κειμένου, πραγματοποιήθηκε ενοποίηση των κωδικοποιήσεων ICD. Η συνύπαρξη ICD-9 και ICD-10 στην MIMIC-IV οδηγεί σε ανομοιογένεια και περιορισμό του χρησιμοποιήσιμου όγκου δεδομένων εάν ο

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

ερευνητής επιλέξει μόνο ένα από τα δύο συστήματα. Για τον λόγο αυτό εφαρμόστηκε μια διαδικασία συστηματικής χαρτογράφησης ICD-9 σε ICD-10, αξιοποιώντας δημόσια διαθέσιμα εργαλεία αντιστοίχισης, Python βιβλιοθήκες για την ανάκτηση περιγραφών, καθώς και στοχευμένο χειροκίνητο έλεγχο σε περιπτώσεις που η αντιστοίχιση δεν μπορούσε να πραγματοποιηθεί αυτόματα. Μετά την εφαρμογή του mapping, ο διαθέσιμος όγκος δεδομένων αυξήθηκε σημαντικά: από 122.279 σημειώσεις σε 331.603, κάτι που ισοδυναμεί με αύξηση της τάξης του 171%. Η ενοποίηση του συστήματος κωδικοποίησης υπό το ICD-10 επέτρεψε τη δημιουργία ενός συνεκτικού και πλήρως αξιοποιήσιμου συνόλου εκπαίδευσης, ευθυγραμμισμένου με τη σύγχρονη κλινική πρακτική διεθνώς.



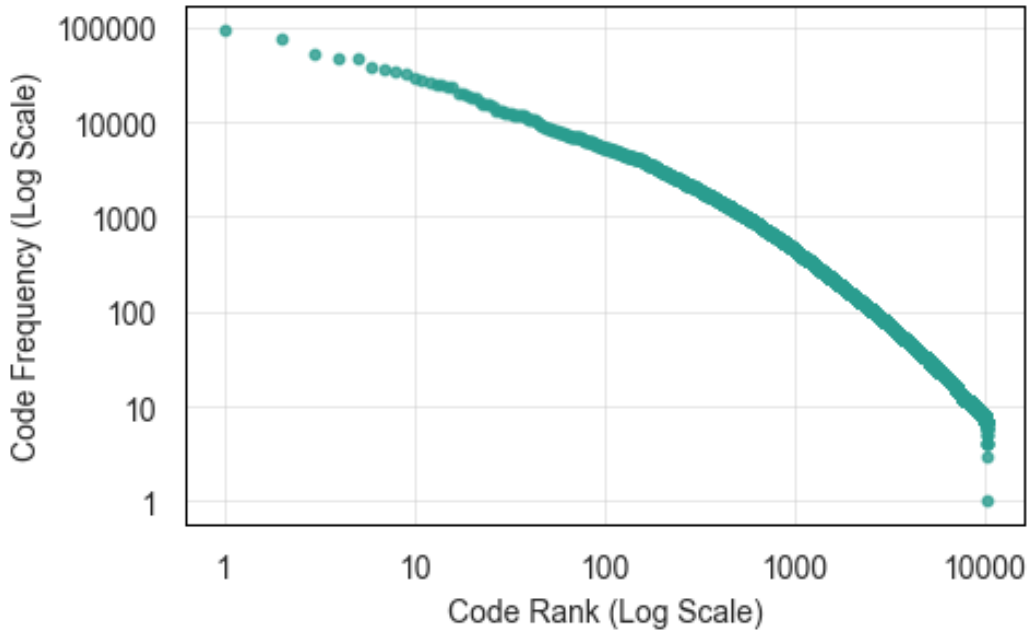
Εικόνα 1: Κατανομή συχνοτήτων ICD-10 μετά το mapping.

Η ανάλυση των συχνοτήτων εμφάνισης των κωδικών ICD αποκάλυψε την ύπαρξη μιας έντονης long-tail κατανομής, χαρακτηριστικής των κλινικών δεδομένων μεγάλης κλίμακας. Μικρός αριθμός κωδικών εμφανίζεται εξαιρετικά συχνά, ενώ χιλιάδες κωδικοί παρουσιάζονται ελάχιστες φορές στο σύνολο εκπαίδευσης. Το φαινόμενο αυτό καταγράφεται χαρακτηριστικά στα αποτελέσματα της μαθηματικής απεικόνισης των συχνοτήτων, η οποία ακολουθεί Zipf-like συμπεριφορά και επιβεβαιώνει ότι η συντριπτική πλειονότητα των ICD-10 κωδικών εμφανίζεται λιγότερο από 30 φορές στο dataset. Η ανισορροπία αυτή μειώνει δραστικά την

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

ικανότητα των μοντέλων να γενικεύσουν επαρκώς και οδηγεί σε υποεκτίμηση των σπάνιων κλινικών φαινοτύπων.

Για την αντιμετώπιση του προβλήματος αυτού αναπτύχθηκε μια στρατηγική ενίσχυσης των δεδομένων (data augmentation) με τη χρήση μεγάλων γλωσσικών μοντέλων. Αρχικά πραγματοποιήθηκε μελέτη συχνοτήτων και κατηγοριοποίηση των κωδικών σε κλάσεις βάσει τεταρτημορίων, ακολουθώντας μια διαδικασία stratified sampling που επιτρέπει την αναλογική εκπροσώπηση διαφορετικών κατηγοριών σπανιότητας. Στη συνέχεια δημιουργήθηκαν ομάδες σπάνιων κωδικών, για τις οποίες παράχθηκαν συνθετικά κείμενα τύπου discharge summary μέσω του μοντέλου LLaMA 3.3. Η παραγωγή συνθετικών κειμένων καθοδηγήθηκε από πραγματικές κλινικές σημειώσεις, ώστε η γλωσσική και δομική μορφή των τεχνητών κειμένων να αντικατοπτρίζει πιστά το ύφος της ιατρικής τεκμηρίωσης. Στο σύνολο παρήχθησαν 16.580 συνθετικές σημειώσεις, οι οποίες ενσωματώθηκαν στο training set, αυξάνοντας σημαντικά την αντιπροσωπευτικότητα των σπάνιων ICD-10 κωδικών.

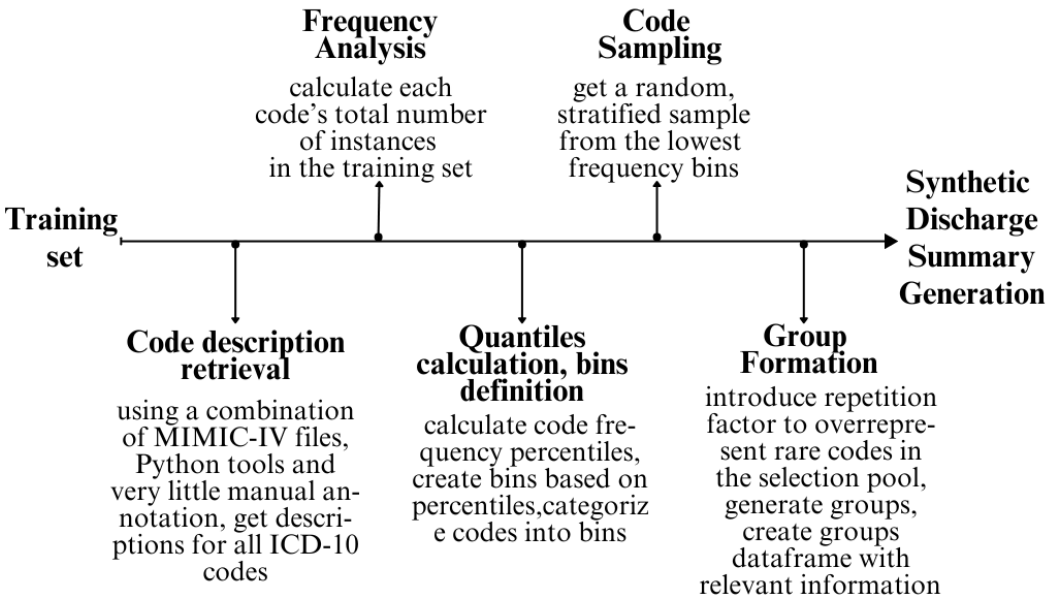


Εικόνα 2: Zipf-like κατανομή συχνοτήτων ICD κωδικών.

Η επίδραση της ενίσχυσης αυτής ήταν ιδιαίτερα εμφανής: κατά μέσο όρο, κάθε σπάνιος κωδικός παρουσίασε αύξηση συχνότητας κατά 256%, γεγονός που βελτίωσε

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

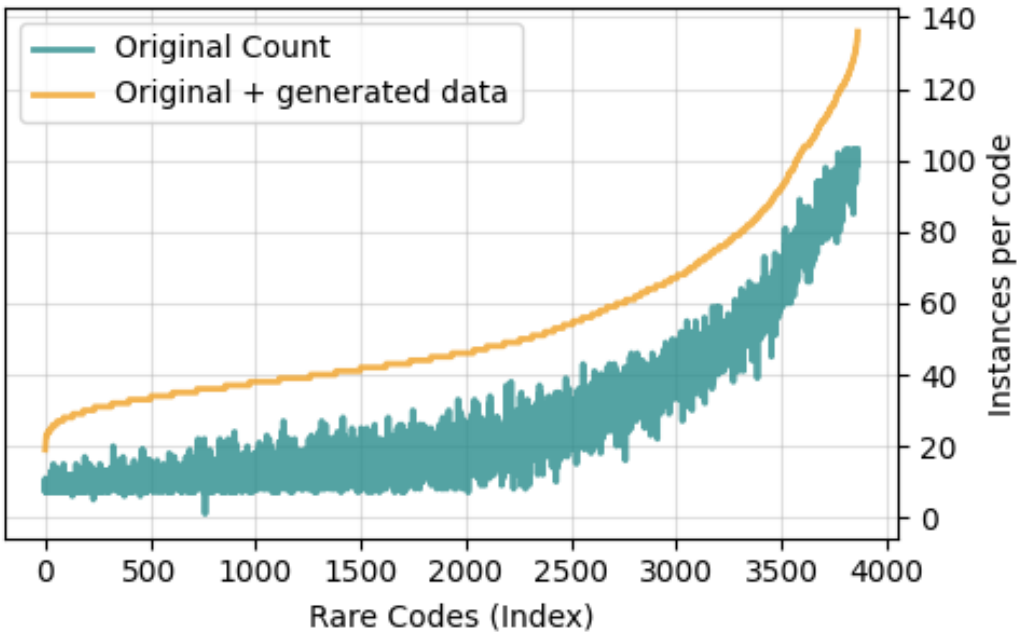
αισθητά τη συμπεριφορά των μοντέλων, ιδιαίτερα ως προς τις μακρο-μετρικές (macro F1, macro precision και macro recall), οι οποίες είναι πιο ευαίσθητες σε σπάνιες κατηγορίες και αποτυπώνουν πληρέστερα την ποιότητα της ταξινόμησης σε ετερογενή σύνολα.



Εικόνα 3: Pipeline παραγωγής συνθετικών δεδομένων.

Η διαδικασία παραγωγής και προεπεξεργασίας των δεδομένων αποτέλεσε απαραίτητο θεμέλιο για την ανάπτυξη των μεθοδολογιών που ακολουθούν. Η προσεκτική διαχείριση των αδόμητων κειμένων, η ενοποίηση της κωδικοποίησης, η αντιμετώπιση της ανισορροπίας και η δημιουργία ενός πλουσιότερου συνόλου εκπαίδευσης συνέβαλαν καθοριστικά στην επίτευξη της ποιότητας και της σταθερότητας των μοντέλων που παρουσιάζονται στις επόμενες ενότητες.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU



Εικόνα 4: Σύγκριση αρχικής και εμπλουτισμένης κατανομής σπάνιων κωδικών.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

3 Μεθοδολογία

Η μεθοδολογική προσέγγιση που ακολουθήθηκε στο πλαίσιο του παρόντος παραδοτέου οργανώνεται γύρω από δύο βασικούς άξονες: αφενός την ανάπτυξη και αξιολόγηση μοντέλων πολυετικετικής ταξινόμησης για την αυτόματη ανάθεση κωδικών ICD σε ιατρικά κείμενα, και αφετέρου την αξιοποίηση σύγχρονων τεχνικών επεξεργασίας φυσικής γλώσσας για την εξαγωγή εννοιών και την παραγωγή δομημένων ιατρικών λεξικών. Οι δύο αυτοί άξονες αλληλοσυμπληρώνονται, καθώς η ποιότητα της ταξινόμησης εξαρτάται σε μεγάλο βαθμό από την ακρίβεια της αναπαράστασης του κειμένου, ενώ η δημιουργία αποδοτικών λεξικών μπορεί να ενισχύσει την ερμηνευσιμότητα και την αξιοποίηση των αποτελεσμάτων.

Το πρώτο μέρος της μεθοδολογίας αφορά την αναπαράσταση του κειμένου σε μορφές κατάλληλες για αλγορίθμους μηχανικής μάθησης. Τα κλινικά κείμενα των βάσεων MIMIC χαρακτηρίζονται από ιδιόμορφη γλώσσα, υψηλή ποικιλότητα όρων και ιδιωματισμών, καθώς και ελλιπή δομή. Για τον λόγο αυτό εξετάστηκαν τόσο κλασικές μεθοδολογίες αναπαράστασης, όπως το TF–IDF, όσο και βαθύτερες σημασιολογικές αναπαραστάσεις μέσω embeddings. Στα παραδοσιακά χαρακτηριστικά TF–IDF, κάθε κείμενο προβάλλεται σε έναν υψηλής διάστασης διανυσματικό χώρο, στον οποίο κάθε διάσταση αντιστοιχεί σε ένα όρο του λεξιλογίου. Η αναπαράσταση αυτή είναι κατάλληλη για γραμμικά μοντέλα, όπως η λογιστική παλινδρόμηση, και αποδείχθηκε ιδιαίτερα αποτελεσματική σε αρχικά πειράματα τόσο στη MIMIC-III όσο και στη MIMIC-IV. Ωστόσο, η έλλειψη κατανόησης σημασιολογικών σχέσεων περιορίζει την απόδοση σε μοντέλα που απαιτούν βαθύτερη γλωσσική κατανόηση.

Για τον λόγο αυτό εξετάστηκαν επίσης μέθοδοι βασισμένες σε word embeddings, όπως το BioWordVec και το Word2Vec, οι οποίες επιτρέπουν τη χαρτογράφηση των λέξεων σε έναν πυκνό, συνεχόμενο χώρο, όπου η γεωμετρική εγγύτητα αντανακλά σημασιολογικές ομοιότητες. Η εκπαίδευση Word2Vec embeddings πάνω στο σύνολο κειμένων της MIMIC-IV προσέφερε προσαρμοσμένες αναπαραστάσεις που

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

αποτυπώνουν ιδιαιτερότητες της ιατρικής ορολογίας, ενώ η χρήση τους σε νευρωνικά δίκτυα βελτίωσε την ικανότητα μάθησης σύνθετων μορφών συνεμφάνισης όρων. Τέλος, για τα πλέον σύγχρονα μοντέλα αξιοποιήθηκαν transformer-based αναπαραστάσεις, όπως εκείνες του RoBERTa (PLM-ICD), οι οποίες ενσωματώνουν contextual πληροφορία και αποδείχθηκαν ικανές να αποδώσουν υψηλότερα ποσοστά ταξινόμησης σε πολύπλοκες κλινικές δομές.

Με βάση τις αναπαραστάσεις αυτές διαμορφώθηκε το μοντέλο πολυετικετικής ταξινόμησης. Η ανάθεση ICD κωδικών αποτελεί τυπικό multi-label πρόβλημα, καθώς κάθε κείμενο μπορεί να αντιστοιχεί σε πολλαπλές διαγνώσεις. Η πιο απλή προσέγγιση, γνωστή ως Binary Relevance, αντιμετωπίζει το πρόβλημα ως σύνολο ανεξάρτητων δυαδικών ταξινομήσεων, όμως η θεμελιώδης υπόθεση ανεξαρτησίας παραβιάζεται πλήρως στα ιατρικά δεδομένα, όπου οι κωδικοί εμφανίζουν ισχυρές αλληλεξαρτήσεις. Για παράδειγμα, η υπέρταση εμφανίζεται συχνά μαζί με δυσλιπιδαιμία, ενώ συγκεκριμένες παθολογίες του αναπνευστικού συνδέονται συστηματικά με συνοδές επιπλοκές. Οι αλληλεξαρτήσεις αυτές αποτελούν κρίσιμη πηγή πληροφορίας, η οποία μπορεί να αξιοποιηθεί για τη βελτίωση της ταξινόμησης.

Σε αυτή τη βάση, η παρούσα εργασία αξιοποίησε και επέκτεινε την τεχνική των Classifier Chains (CC). Η μέθοδος συνδέει μια ακολουθία δυαδικών ταξινομητών, όπου κάθε ταξινομητής προβλέπει έναν ICD κωδικό λαμβάνοντας υπόψη όχι μόνο τα χαρακτηριστικά του κειμένου αλλά και τις προβλέψεις των προηγούμενων ταξινομητών στην αλυσίδα. Με αυτόν τον τρόπο, το μοντέλο καθίσταται ικανό να ενσωματώνει τις εξαρτήσεις μεταξύ των ετικετών, μειώνοντας την αβεβαιότητα και ενισχύοντας την ακρίβεια της ταξινόμησης. Η χρήση ενός νευρωνικού δικτύου ως βασικού ταξινομητή (ANN-CC) προσφέρει ακόμη μεγαλύτερη εκφραστικότητα και δυνατότητα μάθησης μη γραμμικών σχέσεων. Η αρχιτεκτονική που χρησιμοποιήθηκε περιλαμβάνει τέσσερα διαδοχικά πυκνά στρώματα με φθίνουσα διάσταση και ενδιάμεσα Dropout τεχνικές, ώστε να περιοριστεί η υπερεκπαίδευση.

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Ένα κρίσιμο ζήτημα στη μέθοδο CC είναι η σειρά των ετικετών στην αλυσίδα, καθώς διαφορετικές διατάξεις οδηγούν σε διαφορετική διάδοση σφαλμάτων και διαφορετική εκμετάλλευση των εξαρτήσεων. Για τον λόγο αυτό εφαρμόστηκε μια διαδικασία εύρεσης βέλτιστης διάταξης μέσω της τεχνικής Ensemble Classifier Chains (ECC), στην οποία εκπαιδεύθηκαν πολλαπλές αλυσίδες με τυχαίες διατάξεις και η διάταξη που απέδωσε το υψηλότερο ποσοστό ταξινόμησης επιλέχθηκε για την τελική εφαρμογή. Η διαδικασία αυτή βελτίωσε μετρήσιμα την απόδοση τόσο της κλασικής όσο και της νευρωνικής εκδοχής της μεθόδου.

Παράλληλα με τις αλυσίδες ταξινόμησης, εφαρμόστηκαν και πιο σύνθετα μοντέλα βαθιάς μάθησης, τα οποία έχουν καθιερωθεί στη διεθνή βιβλιογραφία για την αυτόματη κωδικοποίηση ICD. Μεταξύ αυτών συγκαταλέγονται το CAML (Convolutional Attention for Multi-Label classification), το MultiResCNN, το LAAT και το PLM-ICD. Τα μοντέλα αυτά ενσωματώνουν label attention μηχανισμούς, οι οποίοι επιτρέπουν στο σύστημα να εντοπίζει τμήματα του κειμένου που συσχετίζονται εντονότερα με κάθε κωδικό, βελτιώνοντας τόσο την ακρίβεια όσο και την ερμηνευσιμότητα της απόφασης. Τα πιο προηγμένα transformer-based μοντέλα, όπως το PLM-ICD, αξιοποιούν προεκπαιδευμένες γλωσσικές αναπαραστάσεις από μεγάλα βιοϊατρικά σώματα κειμένων και επιτυγχάνουν κορυφαίες επιδόσεις σε πληθώρα μετρικών.

Η δεύτερη συνιστώσα της μεθοδολογίας αφορά την εξαγωγή ιατρικών όρων και τη δημιουργία ιατρικών λεξικών. Η διαδικασία αυτή υλοποιήθηκε τόσο μέσω των κλασικών τεχνικών εντοπισμού οντοτήτων (NER) όσο και με τη χρήση LLMs για την αναγνώριση και ομαδοποίηση κλινικών εννοιών. Η ανάλυση των κειμένων επέτρεψε τον εντοπισμό συχνών διαγνωστικών χαρακτηριστικών, όρων απεικονιστικών ευρημάτων και περιγραφών παθολογικών καταστάσεων, τα οποία κανονικοποιήθηκαν και οργανώθηκαν σε θεματικά λεξιλόγια. Τα λεξικά αυτά αποτελούν σημαντικό εργαλείο για τη μελλοντική αξιοποίηση των απεικονιστικών δεδομένων και την ανάπτυξη συστημάτων διαλειτουργικότητας.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Συνολικά, η μεθοδολογία που ακολουθήθηκε ενσωματώνει σύγχρονες τεχνικές NLP, μοντέλα βαθιάς μάθησης, αλγορίθμους πολυετικετικής ταξινόμησης και στρατηγικές εμπλουτισμού δεδομένων. Το συνδυαστικό αυτό πλαίσιο επιτρέπει την αξιόπιστη και αποδοτική ανάλυση μεγάλων συνόλων κλινικών δεδομένων, αποτελώντας ισχυρή βάση για την εφαρμογή των αποτελεσμάτων σε πραγματικά συστήματα υποστήριξης κλινικών αποφάσεων.

4 Πειραματικός Σχεδιασμός

Ο πειραματικός σχεδιασμός που αναπτύχθηκε στο πλαίσιο του παρόντος παραδοτέου αποσκοπεί στη συστηματική αξιολόγηση των μοντέλων πολυετικετικής ταξινόμησης και των μεθόδων επεξεργασίας κλινικού κειμένου που περιγράφηκαν προηγουμένως. Η διαδικασία αυτή διαρθρώθηκε ώστε να επιτρέψει την αποτίμηση της συμβολής κάθε επιμέρους τεχνικής — της προεπεξεργασίας, του εμπλουτισμού δεδομένων, της επιλογής μοντέλου και της ενσωμάτωσης αλληλεξαρτήσεων ετικετών — στη συνολική ποιότητα των αποτελεσμάτων. Ειδική μέριμνα δόθηκε στη δημιουργία ενός πλαισίου σύγκρισης που επιτρέπει την άντληση ασφαλών συμπερασμάτων, τόσο ως προς την αποδοτικότητα όσο και ως προς τη γενικευσιμότητα των μοντέλων.

Η πειραματική διαδικασία διαμορφώθηκε σε δύο βασικές κατευθύνσεις. Η πρώτη αφορά την αξιολόγηση των μοντέλων Classifier Chains και ANN-CC στη βάση MIMIC-III, η οποία χρησιμοποιεί αποκλειστικά ICD-9 κωδικούς. Η επιλογή της MIMIC-III γίνεται λόγω της εκτεταμένης παρουσίας της στη διεθνή βιβλιογραφία, γεγονός που επιτρέπει την άμεση σύγκριση των επιδόσεων των προτεινόμενων μοντέλων με παλαιότερες μεθοδολογίες. Στο πλαίσιο αυτό εξετάστηκαν δύο σενάρια διαφορετικής πολυπλοκότητας: η ταξινόμηση των δέκα συχνότερων ICD-9 κατηγοριών και η ταξινόμηση των είκοσι συχνότερων κατηγοριών. Η διαφοροποίηση των σεναρίων αυτών είναι κρίσιμη, διότι η αύξηση του αριθμού των ετικετών συνεπάγεται μεγαλύτερη ετερογένεια και ισχυρότερες αλληλεξαρτήσεις, οι οποίες ενδέχεται να επηρεάσουν σημαντικά την απόδοση των μοντέλων.

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Η δεύτερη κατεύθυνση αφορά τα πειράματα στη MIMIC-IV, η οποία, μετά τον εμπλουτισμό και την ενοποίηση κωδικοποίησης, προσφέρει ένα εκτενέστερο και ποιοτικά πιο κατάλληλο περιβάλλον για την εκπαίδευση μοντέλων βαθιάς μάθησης. Στην περίπτωση αυτή εξετάστηκαν δύο σενάρια: (α) η εκπαίδευση και αξιολόγηση μοντέλων στη βάση που προέκυψε μετά το mapping ICD-9 → ICD-10 και τις διαδικασίες καθαρισμού, και (β) η εκπαίδευση και αξιολόγηση στη βάση που προέκυψε μετά την ενσωμάτωση των 16.580 συνθετικών κειμένων. Η σύγκριση των δύο αυτών σεναρίων επιτρέπει την ποσοτικοποίηση της συμβολής της τεχνικής data augmentation σε πραγματικές συνθήκες, καθώς τα κλινικά δεδομένα χαρακτηρίζονται από εξαιρετικά μεγάλες κλίμακες ανισορροπίας.

Για την αποφυγή φαινομένων υπερπροσαρμογής (overfitting) εφαρμόστηκε αυστηρή διαδικασία διαχωρισμού των δεδομένων σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Στη MIMIC-IV χρησιμοποιήθηκε multi-label stratified sampling, ώστε να διασφαλιστεί η διατήρηση της κατανομής των ετικετών στα διαφορετικά υποσύνολα, περιορίζοντας την πιθανότητα εμφάνισης ετικετών που υπάρχουν μόνο στο training set ή αποκλειστικά στο test set. Ο διαχωρισμός εφαρμόστηκε σε επίπεδο admission ID, ώστε να αποφευχθεί οποιαδήποτε μορφή διαρροής πληροφορίας (data leakage) μεταξύ των συνόλων. Μετά την ενσωμάτωση των synthetic notes, οι νέες εγγραφές κατανεμήθηκαν αποκλειστικά στο training set, καθώς στόχος τους είναι η ενίσχυση της εκμάθησης και όχι η αλλοίωση των συνθηκών ελέγχου.

Η διαδικασία εκπαίδευσης των μοντέλων ακολούθησε σύγχρονες πρακτικές βελτιστοποίησης. Για τα νευρωνικά δίκτυα χρησιμοποιήθηκε ο βελτιστοποιητής AdamW με γραμμικό warm-up και σταδιακή απομείωση του ρυθμού μάθησης, πρακτική που έχει αποδειχθεί ιδιαίτερα αποτελεσματική σε NLP μοντέλα. Η επιλογή του βέλτιστου μοντέλου έγινε βάσει της απόδοσης στο validation set, με χρήση early stopping όταν δεν παρατηρούνταν βελτίωση επί δέκα συνεχόμενες εποχές. Οι μετρικές που αξιολογήθηκαν περιλαμβάνουν τόσο μικρο- όσο και μακρο-F1, καθώς και συμπληρωματικές μετρικές όπως Jaccard similarity, precision, recall, accuracy,

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

αλλά και πιο εξειδικευμένες μετρικές για multi-label προβλήματα, όπως το precision, το mean average precision (MAP) και το R-precision. Για την εκτίμηση της συνολικής αξιοπιστίας των μοντέλων υπολογίστηκε και το Exact Match Ratio, το οποίο, αν και αυστηρό, επιτρέπει την αποτίμηση του βαθμού πλήρους αντιστοίχισης μεταξύ προβλεπόμενων και πραγματικών ετικετών.

Στο πλαίσιο της αξιολόγησης των Classifier Chains, ιδιαίτερη σημασία δόθηκε στη μελέτη της επίδρασης της διάταξης των ετικετών στην απόδοση του μοντέλου. Πέραν της τυχαίας σειράς, εφαρμόστηκε και η τεχνική εύρεσης βέλτιστης ακολουθίας μέσω επαναλαμβανόμενης εκπαίδευσης ECC μοντέλων με διαφορετικές τυχαίες διατάξεις. Από τα αποτελέσματα επιλέχθηκε εκείνη η διάταξη που μεγιστοποίησε την ακρίβεια, και αυτή χρησιμοποιήθηκε για την εκπαίδευση των CC* και ANN-CC* μοντέλων. Η διαδικασία αυτή επέτρεψε τη διερεύνηση του κατά πόσο η πληροφορία που μεταβιβάζεται κατά μήκος της αλυσίδας βελτιώνεται όταν οι ετικέτες τοποθετούνται σε σειρά που αντανακλά εμπειρικά τα επίπεδα προβλεψιμότητας και συσχέτισής τους.

Για τα μοντέλα βαθιάς μάθησης της βιβλιογραφίας (CAML, MultiResCNN, LAAT, PLM-ICD), ακολουθήθηκαν οι αντίστοιχες αρχιτεκτονικές όπως περιγράφονται στις πρωτότυπες δημοσιεύσεις, με αναπροσαρμογές μόνο στο μέγεθος των attention layers και στην εκπαίδευση word embeddings ώστε να εναρμονίζονται με το προτεινόμενο dataset. Ιδιαίτερα για το PLM-ICD, εφαρμόστηκε η πρακτική εκπαίδευσης σε segments των 128 tokens, ώστε να διατηρηθεί η υπολογιστική αποδοτικότητα χωρίς να χαθεί η συνολική πληροφορία του κειμένου.

Συνολικά, ο πειραματικός σχεδιασμός οργανώθηκε έτσι ώστε να επιτρέπει την αποτίμηση όλων των κρίσιμων παραγόντων που επηρεάζουν την απόδοση ενός πολυετικετικού μοντέλου σε ιατρικά δεδομένα: την ποιότητα και ποσότητα των δεδομένων, τη σημασιολογική αναπαράσταση του κειμένου, την αρχιτεκτονική του μοντέλου, την εκμετάλλευση των εξαρτήσεων μεταξύ ετικετών και την ευαισθησία των μοντέλων σε σπάνιες κατηγορίες. Η ολιστική αυτή προσέγγιση θέτει στέρεες

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

βάσεις για την ανάλυση των αποτελεσμάτων που παρουσιάζονται στην επόμενη ενότητα και εξασφαλίζει την αξιοπιστία των συμπερασμάτων που προκύπτουν.

5 Αποτελέσματα

Η πειραματική διαδικασία που περιγράφηκε στην προηγούμενη ενότητα επέτρεψε τη διεξαγωγή μιας συστηματικής και πολυεπίπεδης αξιολόγησης των μοντέλων πολυετικετικής ταξινόμησης, τόσο σε δεδομένα ICD-9 όσο και σε δεδομένα ICD-10. Τα αποτελέσματα που παρουσιάζονται εδώ αντανακλούν την επίδραση των διαφορετικών μεθοδολογικών επιλογών - προεπεξεργασία, αναπαράσταση κειμένου, αρχιτεκτονική μοντέλου, αξιοποίηση αλληλεξαρτήσεων και εμπλουτισμός δεδομένων - στην τελική απόδοση των συστημάτων. Η ανάλυση οργανώνεται σε δύο μεγάλες κατηγορίες: τα πειράματα στη MIMIC-III, όπου αξιολογήθηκαν οι μέθοδοι Classifier Chains (CC) και οι νευρωνικές τους επεκτάσεις, και τα πειράματα στη MIMIC-IV, όπου εξετάστηκαν τόσο η συμβολή της χαρτογράφησης ICD-9→ICD-10 όσο και η προστιθέμενη αξία των συνθετικών κειμένων σε συνδυασμό με προηγμένα deep learning μοντέλα.

5.1 Αποτελέσματα στη MIMIC-III: Πολυετικετική Ταξινόμηση με Classifier Chains

Η αρχική αξιολόγηση πραγματοποιήθηκε στη βάση MIMIC-III με στόχο να διερευνηθεί η ικανότητα των μοντέλων να εντοπίζουν τις αλληλεξαρτήσεις μεταξύ των ICD-9 κωδικών. Εξετάστηκαν δύο διαφορετικές διαμορφώσεις ετικετών: η ταξινόμηση στις δέκα πιο συχνές ICD-9 κατηγορίες και η ταξινόμηση στις είκοσι πιο συχνές. Η αύξηση του πλήθους των ετικετών συνεπάγεται μεγαλύτερη πολυπλοκότητα, καθώς οι κατηγορίες γίνονται περισσότερο ανισοβαρείς και οι αλληλεξαρτήσεις εντονότερες.

Στο σενάριο των δέκα κατηγοριών, η λογιστική παλινδρόμηση (LR) σημείωσε υψηλές επιδόσεις, ιδιαίτερα σε επίπεδο F1-score, επιβεβαιώνοντας ότι γραμμικά μοντέλα

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

προσαρμόζονται αποτελεσματικά σε προβλήματα με σχετικά μικρό αριθμό ετικετών όταν η αναπαράσταση γίνεται με TF–IDF. Αντίθετα, το μοντέλο LSTM, παρά τη θεωρητική ικανότητά του να συλλαμβάνει χρονικές εξαρτήσεις στο κείμενο, υστέρησε σημαντικά σε macro μετρικές. Το εύρημα αυτό είναι συνεπές με προηγούμενες μελέτες, οι οποίες καταδεικνύουν ότι σε ιατρικά κείμενα μεγάλης έκτασης τα μοντέλα RNN δυσκολεύονται να αποδώσουν χωρίς μηχανισμούς προσοχής.

Η ενσωμάτωση εξαρτήσεων μέσω των Classifier Chains βελτίωσε μετρήσιμα την απόδοση σε σχέση με τα απλά μοντέλα. Η βασική εκδοχή του CC σημείωσε αυξημένη ακρίβεια και καλύτερη ισορροπία μεταξύ precision και recall, υποδηλώνοντας ότι η σειρά των ετικετών προσφέρει χρήσιμη πληροφορία για την ταξινόμηση. Ωστόσο, η μεγαλύτερη βελτίωση παρατηρήθηκε με το ANN-CC, το οποίο εκμεταλλεύεται νευρωνικό δίκτυο ως base classifier. Η χρήση ενός ευέλικτου πολυεπίπεδου perceptron επέτρεψε στο μοντέλο να προσαρμοστεί καλύτερα στη μη γραμμική δομή των κλινικών κειμένων, προσφέροντας σημαντική αύξηση του macro F1-score έναντι όλων των κλασικών μοντέλων.

Όταν ο αριθμός των ετικετών αυξήθηκε σε είκοσι, η δυσκολία του προβλήματος αποτυπώθηκε άμεσα στις μετρικές. Η απόδοση της LSTM υποχώρησε περαιτέρω, ενώ η LR παρέμεινε σταθερή, χωρίς ωστόσο να μπορεί να αντιμετωπίσει τις αυξανόμενες αλληλεπιδράσεις μεταξύ των ετικετών. Τα μοντέλα CC συνέχισαν να υπερέχουν έναντι των ανεξάρτητων ταξινομητών, επιβεβαιώνοντας ότι η εκμετάλλευση των dependencies κρίνεται ζωτική για την πολυετικετική κωδικοποίηση. Το ANN-CC παρέμεινε το καλύτερο από τα μη deep learning μοντέλα, διατηρώντας συνεπή πλεονεκτήματα τόσο στον μικρο- όσο και στον μακρο-μεσαίο χώρο των μετρικών.

Η διερεύνηση της επίδρασης της διάταξης των ετικετών στην αλυσίδα ανέδειξε την αξία της τεχνικής ECC ως heuristic εύρεσης βέλτιστης ακολουθίας. Η εφαρμογή της βέλτιστης αλυσίδας οδήγησε σε περαιτέρω μικρές αλλά στατιστικά συνεπείς βελτιώσεις, ιδίως στο ANN-CC*, όπου το macro F1 ξεπέρασε οριακά τις επιδόσεις της

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

βασικής εκδοχής. Τα αποτελέσματα καταδεικνύουν ότι, μολονότι η σειρά των ετικετών δεν καθορίζει εξ ολοκλήρου την απόδοση, η ορθή επιλογή της μπορεί να περιορίσει τη διάδοση σφαλμάτων κατά μήκος της αλυσίδας και να ενισχύσει τη σταθερότητα του συστήματος.

Συνολικά, τα πειράματα στη MIMIC-III αποδεικνύουν ότι η ενσωμάτωση των εξαρτήσεων μεταξύ των ICD ετικετών αποτελεί κρίσιμο συστατικό για επιτυχημένη ταξινόμηση, και ότι τα ANN-CC μοντέλα προσφέρουν ουσιαστικά πλεονεκτήματα έναντι των παραδοσιακών μεθόδων, ιδιαίτερα όταν ο αριθμός των κωδικών αυξάνεται.

5.2 Αποτελέσματα στη MIMIC-IV: Επίδραση του Mapping ICD-9→ICD-10

Η αξιολόγηση στη MIMIC-IV επικεντρώθηκε αρχικά στη μελέτη της επίδρασης της χαρτογράφησης των ICD-9 κωδικών στις αντίστοιχες ICD-10 τιμές. Η ενοποίηση αυτή αύξησε δραστικά τον αριθμό των διαθέσιμων εγγράφων και μείωσε την ετερογένεια που συνδέεται με την παράλληλη χρήση δύο διαφορετικών συστημάτων κωδικοποίησης. Πριν την εφαρμογή του mapping, τα μοντέλα διέθεταν ένα περιορισμένο σύνολο εκπαίδευσης μετά την εφαρμογή του, ο όγκος των διαθέσιμων κειμένων σχεδόν τριπλασιάστηκε.

Τα deep learning μοντέλα επωφελήθηκαν ιδιαίτερα από τον αυξημένο όγκο δεδομένων. Το Bi-GRU και το CNN κατέγραψαν θετικές μεταβολές σε όλες τις μετρικές, αντανακλώντας την επίδραση της μείωσης του sparsity στον χώρο των λεκτικών αναπαραστάσεων. Το CAML και το MultiResCNN εμφάνισαν ακόμη μεγαλύτερη βελτίωση, καθώς η αύξηση των παραδειγμάτων επέτρεψε στις attention μεθοδολογίες να αποτυπώσουν με μεγαλύτερη ακρίβεια τις σχετιζόμενες εννοιολογικές δομές εντός των κειμένων.

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Το μοντέλο LAAT, το οποίο συνδυάζει LSTM layers με προηγμένο label attention, επωφελήθηκε σε σημαντικό βαθμό από το νέο dataset, παρουσιάζοντας σημαντικές βελτιώσεις τόσο στο micro όσο και στο macro F1. Παρ' όλα αυτά, το PLM-ICD διατήρησε την υπεροχή έναντι όλων των υπόλοιπων μοντέλων, με υψηλότερο AUC-ROC, μεγαλύτερη ακρίβεια στην κατάταξη σπάνιων κωδικών και βελτιωμένες μετρικές R-precision και MAP. Η χρήση contextual transformer embeddings αποδεικνύεται ισχυρά ευεργετική όταν το μοντέλο διαθέτει επαρκή όγκο δεδομένων για fine-tuning.

Ένα ενδιαφέρον εύρημα αποτελεί η μικρή αύξηση των τιμών του Exact Match Ratio μετά το mapping. Αν και η απόλυτη τιμή της μετρικής παραμένει εξαιρετικά χαμηλή - πρακτικά αναμενόμενο για προβλήματα με εκατοντάδες ετικέτες - η αύξηση υποδηλώνει ότι η ομογενοποίηση της κωδικοποίησης βελτιώνει τη συνοχή του label space, καθιστώντας τις προβλέψεις πιο σταθερές.

5.3 Επίδραση των Συνθετικών Κειμένων: Αποτελέσματα Data Augmentation

Μετά την ενσωμάτωση των συνθετικών κειμένων, το σύνολο εκπαίδευσης εμπλουτίστηκε με 16.580 νέες σημειώσεις, οι οποίες αντιστοιχούν αποκλειστικά σε σπάνιους ICD-10 κωδικούς. Η τοπική αύξηση της συχνότητας των underrepresented labels συνέβαλε ουσιαστικά στη βελτίωση των macro-μετρικών, επιτρέποντας στα μοντέλα να μάθουν πληρέστερες αναπαραστάσεις και να αποφύγουν την υπερεστίαση στους συχνότερους κωδικούς.

Η θετική επίδραση του data augmentation ήταν εμφανής για όλα τα μοντέλα, αλλά ιδιαιτέρως για το CAML, το MultiResCNN και το LAAT, τα οποία διαθέτουν μηχανισμούς label attention. Τα μοντέλα αυτά απέκτησαν καλύτερη κάλυψη σπάνιων ετικετών, με το macro F1 να βελτιώνεται σημαντικά σε σχέση με τη βάση μετά το mapping. Επίσης, η αύξηση του R-precision υποδηλώνει ότι οι σχετικοί

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

κωδικοί εντοπίζονται σε υψηλότερες θέσεις κατάταξης, γεγονός κρίσιμο για την κλινική αξιοποίηση των μοντέλων.

Το PLM-ICD διατήρησε την κορυφαία θέση και στο εμπλουτισμένο dataset. Η ικανότητά του να αξιοποιεί contextual πληροφορία το καθιστά ιδιαίτερα ανθεκτικό ακόμη και σε περιπτώσεις σημαντικής ανισορροπίας. Μετά το augmentation, τα κέρδη στο macro F1 και στο MAP ήταν από τα υψηλότερα μεταξύ όλων των μοντέλων, υποδεικνύοντας ότι τα συνθετικά κείμενα δεν προκαλούν θόρυβο αλλά λειτουργούν ενισχυτικά στον χώρο των σημασιολογικών αναπαραστάσεων.

Παρά τη συνολική βελτίωση, το Exact Match Ratio παρέμεινε σε πολύ χαμηλά επίπεδα, γεγονός που επιβεβαιώνει ότι η πλήρης ανάθεση όλων των κωδικών ενός κειμένου εξακολουθεί να αποτελεί εξαιρετικά δύσκολο πρόβλημα. Η παρατήρηση αυτή ευθυγραμμίζεται με τη διεθνή βιβλιογραφία, όπου η μετρική αυτή χρησιμοποιείται περισσότερο ως δείκτης της απόλυτης δυσκολίας του task παρά ως βασικό κριτήριο σύγκρισης.

5.4 Συνολική Συγκριτική Αποτίμηση

Η συνολική ανάλυση των αποτελεσμάτων αναδεικνύει με σαφήνεια ότι:

1. Η εκμετάλλευση αλληλεξαρτήσεων των ετικετών μέσω Classifier Chains βελτιώνει ουσιαστικά την απόδοση έναντι των κλασικών γραμμικών ή recurrent μοντέλων.
2. Η ενοποίηση της κωδικοποίησης ICD-9→ICD-10 αυξάνει κατακόρυφα την αποτελεσματικότητα των μοντέλων βαθιάς μάθησης, ιδιαίτερα εκείνων που βασίζονται σε attention.
3. Το data augmentation με συνθετικά κείμενα αποτελεί μια αποδοτική και ρεαλιστική προσέγγιση για τη βελτίωση της ταξινόμησης σπάνιων κωδικών, χωρίς να θυσιάζεται η συνολική ποιότητα των μοντέλων.

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

4. Τα transformer-based μοντέλα, και ειδικά το PLM-ICD, αποδεικνύονται ανώτερα σε όλα τα σενάρια, επιβεβαιώνοντας τη μετάβαση της NLP έρευνας σε contextual architectures.

Τα ευρήματα αυτά παρέχουν μια ισχυρή βάση για την ενότητα των Συμπερασμάτων και των Μελλοντικών Κατευθύνσεων που ακολουθεί.

6 Συμπεράσματα

Η παρούσα εργασία ανέδειξε τον κρίσιμο ρόλο των σύγχρονων μεθόδων επεξεργασίας φυσικής γλώσσας και της πολυετικετικής μάθησης στην αυτόματη ανάλυση ιατρικών κειμένων και στην κωδικοποίηση κλινικών πληροφοριών. Τα αποτελέσματα που προέκυψαν από τη συστηματική αξιολόγηση των μοντέλων στη MIMIC-III και τη MIMIC-IV αποδεικνύουν ότι η επιτυχής αντιμετώπιση των ιδιαιτεροτήτων των ιατρικών δεδομένων απαιτεί μια ολιστική προσέγγιση, η οποία συνδυάζει τεχνικές προεπεξεργασίας, εμπλουτισμού δεδομένων, αξιοποίησης σημασιολογικών αναπαραστάσεων και κατάλληλων αρχιτεκτονικών μάθησης.

Η ανάλυση στη MIMIC-III κατέδειξε ότι η ενσωμάτωση των αλληλεξαρτήσεων μεταξύ διαγνωστικών κωδικών – μέσω των Classifier Chains και των νευρωνικών επεκτάσεών τους – προσφέρει ουσιαστική βελτίωση στην απόδοση των μοντέλων σε σχέση με απλούστερες γραμμικές ή αναδρομικές προσεγγίσεις. Το γεγονός ότι η ANN-CC εκδοχή υπερείχε σταθερά υποδηλώνει ότι ο συνδυασμός μη γραμμικής μάθησης και αξιοποίησης dependencies αποτελεί ισχυρό εργαλείο για πολυετικετικές ταξινομήσεις στον ιατρικό χώρο. Επιπλέον, η μελέτη της βέλτιστης διάταξης των ετικετών επιβεβαίωσε ότι, αν και δεν μεταβάλλει ριζικά τα αποτελέσματα, μπορεί να περιορίσει τη διάδοση σφαλμάτων και να ενισχύσει τη σταθερότητα των προβλέψεων.

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Η μετάβαση στη MIMIC-IV ανέδειξε τις ανάγκες και τις προκλήσεις της σύγχρονης κλινικής τεκμηρίωσης. Η χαρτογράφηση των ICD-9 κωδικών σε ICD-10 αποδείχθηκε καθοριστικής σημασίας, καθώς επέτρεψε την αξιοποίηση ενός μεγαλύτερου και πιο συνεκτικού όγκου δεδομένων, ο οποίος με τη σειρά του βελτίωσε σημαντικά την απόδοση των deep learning μοντέλων. Η αύξηση της διαθεσιμότητας παραδειγμάτων επηρέασε θετικά τις attention-based αρχιτεκτονικές, ενώ τα transformer μοντέλα, και ειδικότερα το PLM-ICD, αναδείχθηκαν ως οι πιο αποτελεσματικές προσεγγίσεις σε συνθήκες υψηλής πολυπλοκότητας και μεγάλου πλήθους ετικετών.

Η ενσωμάτωση συνθετικών κειμένων με στόχο την αντιμετώπιση της έντονης ανισορροπίας στη συχνότητα των ICD-10 κωδικών αποτέλεσε ένα ακόμη βασικό εύρημα της παρούσας μελέτης. Η παραγωγή 16.580 συνθετικών σημειώσεων συνέβαλε ουσιαστικά στην αντιπροσώπευση των σπάνιων διαγνώσεων, βελτιώνοντας σημαντικά τις macro μετρικές των μοντέλων και επιτρέποντας πληρέστερη μάθηση σε ένα περιβάλλον που κυριαρχείται από long-tail κατανομές. Η προσέγγιση αυτή επιβεβαιώνει ότι τα μεγάλα γλωσσικά μοντέλα μπορούν να αξιοποιηθούν με τρόπο ελεγχόμενο και παραγωγικό, ενισχύοντας τη σταθερότητα και την αποδοτικότητα των συστημάτων αυτόματης κωδικοποίησης.

Πέρα από την πολυετικετική ταξινόμηση, η εργασία συνέβαλε ουσιαστικά και στην εξαγωγή και οργάνωση ιατρικών όρων από τις κλινικές γνωματεύσεις. Η ανάπτυξη προσαρμοσμένων λεξικών, η κανονικοποίηση ορολογίας και η αποτύπωση συχνών εννοιολογικών σχημάτων επιτρέπουν τη μελλοντική αξιοποίηση των δεδομένων σε συστήματα διαλειτουργικότητας και σε εφαρμογές υποστήριξης ιατρικών αποφάσεων. Η δυνατότητα συνδυασμού των παραγόμενων λεξικών με δεδομένα απεικόνισης αποτελεί ένα σημαντικό βήμα προς τη γεφύρωση των κλινικών και απεικονιστικών πληροφοριών, όπως προβλέπεται στους στόχους της Ενότητας Εργασίας ΕΕ6.

Συνολικά, τα ευρήματα του Παραδοτέου Π6.3 επιβεβαιώνουν ότι η ενσωμάτωση προηγμένων τεχνικών NLP στην ανάλυση αδόμητων ιατρικών δεδομένων μπορεί να

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

βελτιώσει σημαντικά την ποιότητα και την αξιοπιστία της αυτόματης διαγνωστικής κωδικοποίησης. Παρά τις ενθαρρυντικές επιδόσεις, ιδιαίτερα των transformer μοντέλων, το Exact Match Ratio παραμένει σε πολύ χαμηλά επίπεδα, αναδεικνύοντας τις εγγενείς δυσκολίες του προβλήματος και την ανάγκη για περαιτέρω μελέτη. Ωστόσο, η συνολική πρόοδος που επιτεύχθηκε καταδεικνύει ξεκάθαρα ότι η σύγχρονη τεχνολογία έχει τη δυνατότητα να υποστηρίξει την αυτοματοποίηση διαδικασιών που έως πρόσφατα θεωρούνταν πρακτικά αδύνατο να ψηφιοποιηθούν με ακρίβεια.

Η συμβολή του παραδοτέου στον συνολικό σκοπό της Πράξης είναι διττή: αφενός παρέχει ισχυρά τεχνικά εργαλεία για την ανάλυση και ταξινόμηση ιατρικών κειμένων, και αφετέρου θέτει τις βάσεις για την ενοποίηση κλινικών, απεικονιστικών και ομικών δεδομένων σε ένα κοινό πληροφοριακό οικοσύστημα. Η δυνατότητα γεφύρωσης αυτών των δεδομένων αποτελεί βασικό βήμα προς την εφαρμογή πρακτικών εξατομικευμένης ιατρικής, όπου η πληροφορία από πολλαπλές πηγές συνδυάζεται ώστε να προσφέρει βελτιωμένη κλινική πρόβλεψη και υποστήριξη.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU