National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Prediction of Drug-to-Drug Interactions through Zero-Shot Learning

Diploma Thesis

of

NIKOLAOS ASTRAS

Supervisor: Panagiotis Tsanakas
Professor

Athens, February 2024

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Prediction of Drug-to-Drug Interactions through Zero-Shot Learning

Diploma Thesis

of

NIKOLAOS ASTRAS

Supervisor: Panagiotis Tsanakas
Professor

Approved by the examination committee on 1st of February 2024.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Panagiotis Tsanakas	Andreas Georgios Stafilopatis	Georgios Matsopoulos
Professor	Professor	Professor

Athens, February 2024

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

. . . . . .
Nikolaos Astras

Περίληψη

Οι αλληλεπιδράσεις φαρμάκων αποτελούν ένα κρίσιμο κομμάτι στη διαχείριση μιας φαρμακευτικής αγωγής. Ορισμένες μελέτες μάλιστα, εκτιμούν ότι αυτές οι αλληλεπιδράσεις μπορεί να είναι υπεύθυνες για έως και 20% των παρενεργειών που απαιτούν νοσοκομειακή νοσηλεία. Η καθιερωμένη μέθοδος για την πρόβλεψη αυτών των αλληλεπιδράσεων είναι μια εκτενής και πολύπλοκη διαδικασία, η οποία βασίζεται στην ανάλυση των φαρμακευτικών ιδιοτήτων των φαρμάκων σε κλινικά αποτελέσματα, βιβλιογραφικές αναφορές κ.α. Βέβαια τα τελευταία χρόνια, ως εναλλακτική λύση, έχουν προκύψει πληθώρα προσεγγίσεων βασισμένες στη μηχανική μάθηση. Αυτές οι μέθοδοι εκμεταλλεύονται το μεγάλο όγκο δεδομένων που είναι πλέον διαθέσιμα στον τομέα της βιοϊατρικής, για τον εντοπισμό σχέσεων μεταξύ φαρμάκων και παρενεργειών, οδηγώντας σε υψηλής ακριβείας προβλέψεις. Σε αυτό που θα προσπαθήσει να συμβάλει η συγκεκριμένη εργασία και τη ξεχωρίζει από άλλες προσεγγίσεις είναι η αξιοποίηση της $Zero-Shot\ Learning$ τεχνικής. Το ZSL είναι μια σύγχρονη τεχνική μηχανικής μάθησης που επιτρέπει στα μοντέλα να γενικεύουν πέρα από τις κατηγορίες που συνάντησαν κατά την εκπαίδευση τους και να κάνουν προβλέψεις για κατηγορίες που δεν έχουν δει ποτέ. Για να το πετύχουμε αυτό, αξιοποιήσαμε ένα ZSL πλαίσιο που βασίζεται στη χαρτογράφηση σχέσεων μεταξύ των χαρακτηριστικών που εξάγουμε από τις κατηγορίες και τα δεδομένα εισόδου. Το πλαίσιο προσπαθεί να αποτυπώσει και να απλοποιήσει τις πολύπλοκες σχέσεις που κρύβονται μεταξύ των ζευγών φαρμάκων και των παρενεργειών. Να σημειωθεί ακόμα ότι ένας συνδυασμός φαρμάκων έχει τη δυναμική να οδηγήσει σε πολλαπλές παρενέργειες και έτσι απαιτούνται κατάλληλες τροποποιήσεις για να ληφθεί υπόψη αυτή η πιθανότητα. Στόχος μας είναι να αναπτύξουμε μια μέθοδο, που με τις απαραίτητες προσαρμογές, θα αποτελέσει ένα πολύτιμο εργαλείο για τον εντοπισμό και τη μείωση των πιθανών αλληλεπιδράσεων φαρμάκων με φαρμάκων που οδηγούν σε παρενέργειες.

Λέξεις Κλειδιά

Αλληλεπιδράσεις φαρμάκων, Μηχανική μάθηση, Διανυσματική αναπαράσταση λέξεων, Ταξινόμηση πολλαπλών ετικετών, Μάθηση χωρίς επαρκή δεδομένα

Abstract

Drug-to-drug interactions (DDIs) are a crucial aspect of medication management. While estimates vary, some studies suggest that DDIs may be responsible for up to 20% of the adverse drug reactions requiring hospitalization. Conventional methods for predicting those interactions rely on analyzing the pharmaceutical properties of drugs, clinical findings, and literature references. In recent years, approaches based on machine learning have emerged as a promising alternative, taking advantage of the vast biomedical data currently available, to identify relations between drugs and side effects, leading to highly accurate predictions. In this thesis, we differentiate by adopting the Zero-shot learning (ZSL) paradigm to tackle the challenge of DDI prediction. ZSL is a modern ML technique, that enables models to generalize beyond the classes encountered during training, and make predictions for unseen classes. To achieve this, we leveraged a ZSL framework that relies on feature vectors extracted from both instances and classes. The framework effectively tries to capture and simplify the complex underlying relationships between different drug pairs and side effects. We should mention that a single drug combination can result in multiple side effects, necessitating appropriate modifications to account for this possibility. Our goal is to develop a DDI prediction pipeline that, with the necessary adjustments, can serve as a valuable resource for identifying and mitigating potential drug-drug interactions.

Keywords

Drug-to-drug interactions, Machine Learning, Zero-Shot Learning, Multi-label classification, NLP, Word Embeddings

to my parents

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντά μου κ. Παναγιώτη Τσανάκα για την πολύτιμη υποστήριξη που μου παρείχε στη διεκπεραίωση της εργασίας μου. Επίσης, οφείλω πολλές ευχαριστίες στον ερευνητή του ΕΚΕΦΕ Δημόκριτος, κ. Δημήτρη Βογιατζή, ο οποίος με την εμπειρία, τις συμβουλές και την υπομονή του, συνέβαλε σημαντικά στην ολοκλήρωση αυτής της διπλωματικής εργασίας.
Ακόμα είμαι ευγνώμων στους γονείς και την αδερφή μου, για την ακλόνητη υποστήριξη και ενθάρρυνση τους. Τέλος, ευχαριστώ όλους όσους συνεργαστήκαμε και όσους με βοήθησαν, τη Νικόλ για τη συνεχή στήριξη της και το φίλο και συμφοιτητή μου Νικόλα με τον οποίο μοιραστήκαμε τόσα πολλά όλα αυτά τα χρόνια.

Table of Contents

List of Figures

List of Tables

Εκτενής Ελληνική Περίληψη

Εισαγωγή

Σε μια φαρμακευτική θεραπεία η συνταγογράφηση περισσότερων του ενός φαρμάκου, είναι συχνά αναγκαία για την αντιμετώπιση μιας ασθένειας. Ωστόσο, αυτοί οι αναγκαίοι συνδυασμοί, μπορεί ανυποψίαστα να οδηγήσουν σε αλληλεπιδράσεις μεταξύ των δραστικών συστατικών των φαρμάκων, οι οποίες κατά συνέπεια μπορεί να προκαλέσουν ανεπιθύμητες παρενέργειες. Είναι σημαντικό να αναφέρουμε ότι αυτές οι παρενέργιες αποτελούν μια από τις κύριες αιτίες των λανθασμένων φαρμακευτικών αγωγών στις ανεπτυγμένες χώρες.

Αυτές οι αλληλεπιδράσεις μπορούν να προκαλέσουν μεταβολές στην αποτελεσματικότητα των φαρμάκων και να οδηγήσουν στην ολική αποτυχία της θεραπείας ή ακόμη σε σοβαρές και εξουθενωτικές παρενέργειες. Εκτιμάται πως το 10–20% των παρενεργειών που καταλήγουν σε νοσηλεία οφείλονται σε αυτές τις αλληλεπιδράσεις. Εξαιρετικά ευάλωτοι είναι οι ηλικιωμένοι ασθενείς, καθώς υπάρχει ισχυρή σχέση μεταξύ της αύξησης της ηλικίας, του αριθμού των συνταγογραφούμενων φαρμάκων και της συχνότητας των πιθανών ανεπιθύμητων αλληλεπιδράσεων [1], [2].

Αναμφίβολα, η ικανότητα να προβλέπουμε αυτές τις αλληλεπιδράσεις εύκολα και με ακρίβεια είναι ένα ανεκτίμητο εργαλείο για τους ιατρούς, καθώς και ένα σημαντικό σημείο ασφάλειας για τους ασθενείς. Κατά κανόνα, η εύρεση των αλληλεπιδράσεων μεταξύ ενός συνδυασμού φαρμάκων, είναι μια εκτενής και πολύπλοκη διαδικασία. Βασίζεται στη μελέτη των φαρμακολογικών ιδιοτήτων των φαρμάκων, σε κλινικά αποτελέσματα, βιβλιογραφικές αναφορές και τις ιδιαιτερότητες του κάθε ασθενή. Είναι δύσκολο λοιπόν, τουλάχιστον με τις μέχρι πρότινος μεθόδους να υπάρξει μια σωστή πρόβλεψη, ειδικά σε μία θεραπεία που υπάρχει μεγάλος αριθμός φαρμάκων [2].

Σε αυτό το εγχείρημα όπως και σε πολλούς άλλους τομείς, τα υπολογιστικά συστήματα έχουν γίνει αναπόσπαστος σύμμαχος. Χρησιμοποιώντας προηγμένους αλγορίθμους και εξειδικευμένο λογισμικό, η δυνατότητα της αυτόματης πρόβλεψης πιθανών αλληλεπιδράσεων γίνεται όλο και πιο επιτεύσιμη. Η έρευνα γύρω από αυτό το θέμα έχει κεντρίσει το ενδιαφέρον πολλών ερευνητών και αποτελεί αντικείμενο μελέτης για πολλούς ακαδημαϊκούς φορείς. Μάλιστα εταιρίες αναπτύσσουν και προσφέρουν τέτοιες λύσεις, είτε ως πρωταρχική τους δραστηριότητα είτε ενσωματωμένες σε μεγαλύτερα συστήματα.

Στο επίκεντρο αυτής της τεχνολογικής εξέλιξης βρίσκεται αναμφίβολα η τεχνητή νοημοσύνη ή ειδικότερα η μηχανική μάθηση [3]. Η αποτελεσματικότητα των μοντέλων μηχανικής μάθησης στη κατανόηση των περίπλοκων βιοχημικών συνδέσεων και στη πρόβλεψη αλληλεπιδράσεων έχει αποτυπωθεί σε εκτενείς μελέτες και παρουσιαστεί σε πολλά καταξιωμένα επιστημονικά περιοδικά. Αυτά τα μοντέλα, συχνά βασίζονται σε νευρωνικά δίκτυα, στη βαθιά μάθηση ή σε

άλλους σύνθετους αλγόριθμους τεχνητής νοημοσύνης. Εκπαιδεύονται σε τεράστιες βάσεις δεδομένων που έχουν δημιουργηθεί από κλινικές δοκιμές, ιατρικά αρχεία και μελέτες.

Σχετική Βιβλιογραφία

Στη μέχρι τώρα βιβλιογραφία διαφορετικοί μέθοδοι προσέγγισης έχουν αξιοποιηθεί από ερευνητές για την παραγωγή προβλέψεων. Από αυτές τις μεθόδους πιο συχνά συναντώνται προσεγγίσεις που βασίζονται στην δομική ομοιότητα των φαρμάκων ή στην εκπαίδευση ενός μοντέλου με προηγμένες ικανότητες ταξινόμησης [4].

Η προσέγγιση που βασίζεται στην ομοιότητα των φαρμάκων ακολουθεί την εξής απλή λογική. Αν το φάρμακο A αλληλεπιδρά με το φάρμακο B και προκαλεί μια συγκεκριμένη παρενέργεια, τότε τα φάρμακα που είναι συγκρίσιμα με το φάρμακο A είναι πιθανό να προκαλέσουν την ίδια παρενέργεια κατά την αλληλεπίδραση τους με το φάρμακο B. Ποικίλοι τύποι νευρωνικών δικτύων χρησιμοποιούνται για αυτό το σκοπό. Είτε για να διευκολύνουν την αναγνώριση των τοπολογικών και δομικών χαρακτηριστικών που σχετίζονται με πιθανές αλληλεπιδράσεις ή για την εξαγωγή των φαρμακοκινητικών προφίλ των φαρμάκων. Οι τεχνικές ημιεπιβλεπόμενης και μη επιβλεπόμενης μάθησης είναι επίσης σε άνοδο, εξερευνώντας μη γνωστές σχέσεις και αναπαραστάσεις στα δεδομένα. Σε ένα άρθρο οι ερευνητές *Vilar et al.* [5], χρησιμοποιούν την μέθοδο αυτή, μέσο ενός διανυσματικού αποκωδικοποιητή για να κωδικοποιήσουν τα μοριακά χαρακτηριστικά των φαρμάκων και να εξάγουν τη μοριακή ομοιότητα τους.

Από την άλλη πλευρά ένας δυαδικός ταξινομητής μπορεί να προσομοιωθεί για τη πρόβλεψη των αλληλεπιδράσεων. Είσοδοι από ζεύγη φαρμάκων που είτε παράγουν παρενέργειες είτε όχι χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου ταξινόμησης. Βεβαίως διάφοροι ταξινομητές όπως ο ταξινομητής Μπέυζ, κ-πλησιέστεροι γείτονες, SVM, κ.τ.λ επιλέγονται για τη δημιουργία του μοντέλου. Χρησιμοποιώντας μια τεχνική πρόβλεψης συνδέσμων, οι *Kastrin et al.* [6] θεώρησαν την πρόβλεψη των αλληλεπιδράσεων ως μια εργασία δυαδικής ταξινόμησης και βασίστηκαν σε πέντε μεγάλες βάσεις δεδομένων DDI

Όλες αυτές οι τεχνολογίες εκμεταλλεύονται των μεγάλο όγκο δεδομένων που είναι διαθέσιμα στη φαρμακολογία, δίνοντας τη δυνατότητα για πιο ακριβείς και γρήγορες προβλέψεις.

Στόχος της παρούσας διπλωματικής εργασίας

Με βάση και τα παραπάνω, η παρούσα διπλωματική διερευνά το πρόβλημα της πρόβλεψης των παρενεργειών που θα δημιουργήσει η αλληλεπίδραση δυο φαρμάκων. Σε αυτό που θα προσπαθήσει να συμβάλει η συγκεκριμένη εργασία και τη ξεχωρίζει από άλλες προσεγγίσεις είναι η αξιοποίηση της $Zero-Shot\ Learning$ τεχνικής. Στοχος μας είναι λοιπόν να κάνουμε προβλέψεις για παρενεργιες που δεν συμπεριλάβαμε στο στάδιο της εκπαίδευσης του μοντέλου. Μέσω μιας μεθόδου *zero shot learning* θα αξιοποιήσουμε τις διανυσματικές αναπαραστάσεις των φαρμάκων και των παρενεργειών, και θα προσπαθήσουμε να δημιουργήσουμε ένα μαθηματικό συσχετισμό μεταξύ των δύο. Θα επιδιώξουμε λοιπόν να εξετάσουμε κατα πόσο μπορούμε να εκμεταλλευτούμε αυτό το συσχετισμό για να κάνουμε προβλέψεις για κλάσεις/παρενέργειες τις οποίες δεν είχαμε συμπεριλάβει στα δεδομένα εκμάθησης του μοντέλου.

Ακόμα στόχος μας είναι να αντιστοιχήσουμε κάθε συνδυασμό φαρμάκων με περισσότερες από μια παρενέργειες, έτσι ο ταξινομητής μας θα χρειαστεί κατάλληλη αναπροσαρμογή.

Επισκόπηση της δομής του μοντέλου

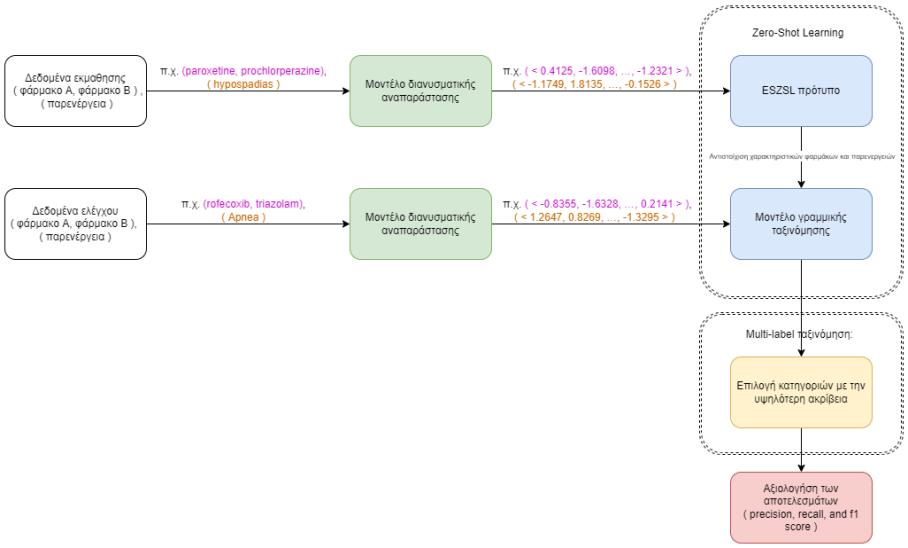


Figure 1. *Διαγραμματική επισκόπηση της δομής του μοντέλου*

Όπως αναφέρθηκε και προηγουμένως η παρούσα διπλωματική ξεχωρίζει λόγω της επιλογής μας να αξιοποιήσουμε το $Zero-Shot\ learning$ (ZSL) πρότυπο. Το ZSL είναι μια τεχνική τεχνητής νοημοσύνης, όπου ένα μοντέλο εκπαιδεύεται ώστε να αναγνωρίζει λέξεις, αντικείμενα ή έννοιες τις οποίες δεν έχει συναντήσει κατά την εκπαίδευσή του [7]. Κατα κανόνα, τα μοντέλα μηχανικής μάθησης απαιτούν μεγάλο όγκο δεδομένων για να μπορέσουν να αναγνωρίσουν ένα συγκεκριμένο αντικείμενο. Αυτό που διακρίνει λοιπόν, τη προσέγγιση του $Zero-Shot\ Learning$, είναι ότι το μοντέλο εκμεταλλεύεται τις γνώσεις που έχει αποκτήσει κατά την εκπαίδευση του και μπορεί να εξάγει γενικεύσεις για νέες κατηγορίες που δεν έχει συναντήσει ποτέ. Αυτό το πετυχαίνει ερμηνεύοντας τη σημασιολογική σημασία των λέξεων ή αντικειμένων μέσα σε έναν πολυδιάστατο χώρο, διευκολύνοντας το μοντέλο να αποτυπώσει και να κατανοήσει τις σχέσεις μεταξύ διαφορετικών εννοιών και να προβλέψει αποτελέσματα για έννοιες που δεν έχει διδαχθεί.

Στη δική μας προσέγγιση επιλέξαμε να χρησιμοποιήσουμε το ZSL, πρότυπο $ESZSL$ που εισαγάγουν ο *Bernardino Romera* και *Philip Torr* στο άρθρο τους "An embarrassingly simple approach to zero-shot learning" [8]. Το $ESZSL$ παρουσιάζει μία καινοτόμο μέθοδο που κάνει χρήση βοηθητικών πληροφοριών για να δημιουργήσει μια σχέση μεταξύ των γνωστών και των άγνωστων κλάσεων. Οι βοηθητικές πληροφορίες στην περίπτωση μας, βασίζονται στις λεκτικές αναπαραστάσεις, γνωστές και ως *word embeddings*, αποτελούν μια μορφή διανυσματικής αναπαράστασης των λέξεων, η οποία επιτρέπει την ποσοτικοποίηση και ανάλυση της σημασιολογικής και συντακτικής πληροφορίας που κρύβουν.

Το $ESZSL$ βασίζεται στην εύρεση μια συνάρτησης η οποία υπολογίζει το ποσοστό συμβατότητας μεταξύ ενός αντικειμένου και μιας κλάσης. Στη περίπτωση μας αυτη η συνάρτηση

έχει ως είσοδο ένα συνδυασμό φαρμάκων και μια παρενέργεια και υπολογίζει τη πιθανότητα του συνδυασμού αυτού να προκαλεί την εξής παρενέργεια. Για να το πετύχει αυτό, χρειάζεται να παρέχουμε στη συνάρτηση ταυτόχρονα τη διανυσματική αναπαράσταση του συνδυασμού των φαρμάκων καθώς επίσης και της παρενέργειας. Τα δεδομένα εκπαίδευσης χρησιμοποιούνται για να βελτιστοποιηθούν οι παραμετροι της συνάρτησης.

Το $ESZSL$ χρησιμοποιεί ένα ευρύ φάσμα μαθηματικών εννοιών και τεχνικών. Στον πυρήνα αυτού του πλαισίου βρίσκονται οι θεμελιώδεις αρχές της γραμμικής άλγεβρας και ειδικότερα οι πράξεις πινάκων. Παρακάτω, παρέχουμε μια σύντομη επισκόπηση της σύνθεσης και των πράξεων μεταξύ των διαφόρων πινάκων που εμπλέκονται στο πλαίσιο.

Με S συμβολίζουμε το πίνακα που περιέχει τις ιδιότητες των κλάσεων, με X το πίνακα που περιέχει τα χαρακτηριστικά των εισόδων και Y εμπεριέχει τη πραγματική κλάση στην οποία ανήκει κάθε είσοδος εκπαίδευσης. Το άρθρο αρχικά παρουσιάζει μια εξίσωση σχεδιασμένη για να διευκολύνει την εκμάθηση ενός γραμμικού ταξινομητή για ένα δεδομένο σύνολο κατηγοριών εκπαίδευσης.

$$\min_{W \in \mathbb{R}^{d \times z}} L(XW, Y) + \Omega(W)$$

Στη παραπάνω εξίσωση, με W αναπαριστώνται οι παράμετροι που πρέπει να βρεθούν, L είναι η επιλεγμένη συνάρτηση απώλειας, και Ω είναι ο κανονικοποιητής. Η συγκεκριμένη επιλογή του L και του Ω μπορεί να οδηγήσει σε πολλαπλές προσεγγίσεις. Στη συγκεκριμένη εξίσωση τα χαρακτηριστικά των κλάσεων δεν χρησιμοποιούνται, με αποτέλεσμα να μην υποστηρίζεται η μεταφορά της γνώσης από το υπάρχον σύνολο κλάσεων σε νέες. Για να ενσωματώσουμε τις πληροφορίες για τα χαρακτηριστικά, η εξίσωση τροποποιείται ως εξής:

$$\min_{V \in \mathbb{R}^{d \times a}} L(XVS, Y) + \Omega(V)$$

Ο πίνακας V προκύπτει από την ισότητα $W = S^T V$. Η χρήση του πίνακα V εισάγει τις ιδιότητες των κλάσεων στην εξίσωση και επιτρέπει τη γενίκευση από τις κλάσεις που χρησιμοποιήθηκαν στο στάδιο της εκπαίδευσης σε νέες. Ο τελικός μας στόχος είναι να διακρίνουμε μεταξύ ενός νέου αδιευκρίνιστου συνόλου κλάσεων z'. Για να γίνει αυτό στην εξίσωση πρέπει να παρέχουμε έναν πίνακα με τις ιδιότητες κάθε κλάσεις που συμβολίζεται από το $S'^{a \times z'}$. Στη συνέχεια, για μια νέα είσοδο x, μια πρόβλεψη μπορεί να δοθεί χρησιμοποιώντας την ακόλουθη δήλωση:

$$\arg\max_i xVS_i$$

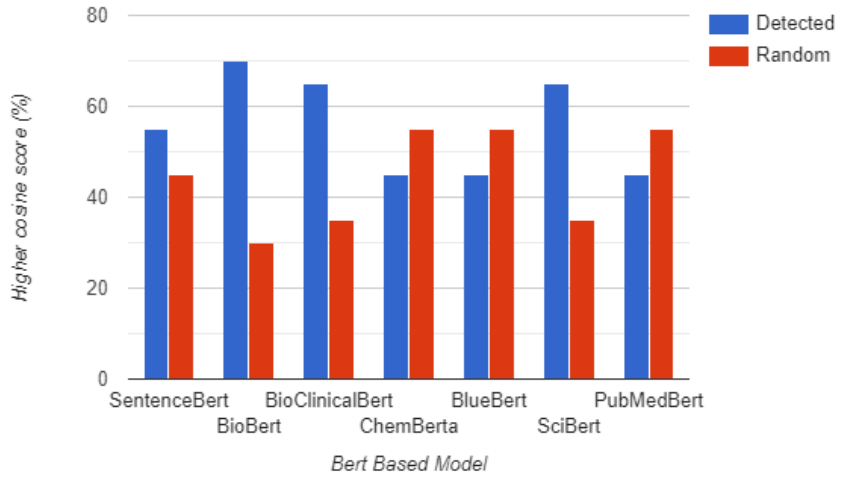
Βάση κάποιων απαραίτητων υποθέσεων για τη μορφή του κανονικοποιητή η εξίσωση για την εύρεση του πίνακα V δίνεται ως εξής:

$$V = (XX^T + \gamma I)^{-1} XYS^T (SS^T + \lambda I)^{-1}$$

Για να μπορέσουμε βέβαια να κατασκευάσουμε το πίνακα V πρέπει να έχουμε στη διάθεση μας τα απαραίτητα δεδομένα και τη δυνατότητα να εξάγουμε τα χαρακτηριστικά/ιδιότητες τους. Για τη συλλογή των δεδομένων χρησιμοποιήθηκε η βάση δεδομένων $TWOSIDES$,

όπου περιέχει λεπτομερής πληροφορίες για 1.318 διαφορετικές παρενέργειες που προκαλούνται από 63.473 συνδυασμούς φαρμάκων. Μετά τη κατάλληλη μορφοποίηση τα δεδομένα ταξινομήθηκαν σε πλειάδες, δυο φαρμάκων και της ανάλογης παρενέργειας που προκαλούν. Σε αυτή τη μορφή τα δεδομένα είναι απλές συμβολοσειρές και δεν μπορούμε μέσα από αυτές να εξάγουμε τη σημασιολογική πληροφορία που χρειαζόμαστε. Για αυτό το λόγο, κάθε κομμάτι των δεδομένων που συλλέξαμε χρειάστηκε να περαστεί από ένα μοντέλο διανυσματικής αναπαράστασης (*word embedding model*), για να μετατραπεί και να αποθηκευτεί στην αντιστοιχη διανυσματική μορφή. Τα μοντέλα διανυσματικής αναπαράστασης λέξεων (*word embedding models*) είναι αλγόριθμοι *NLP* που μετατρέπουν τις λέξεις σε αριθμητικές μορφές, γνωστές ως διανύσματα. Αυτή η μετατροπή διευκολύνει την κατανόηση και την επεξεργασία γλωσσικών δεδομένων σε υπολογιστικό επίπεδο. Χρησιμοποιώντας αυτά τα μοντέλα, οι υπολογιστές είναι ικανοί να ανιχνεύουν ομοιότητες και σημασιολογικές σχέσεις ανάμεσα στις λέξεις.

Στη προσέγγιση μας αξιοποιήσαμε διάφορα μοντέλα που είναι βασισμένα στη τεχνολογία *BERT* της *Google* [9]. Τα μοντέλα αυτά εξάγουν ένα διάνυσμα, συνήθως 768 διαστάσεων για οποιαδήποτε συμβολοσειρά εισαχθεί σε αυτά. Οι διαστάσεις του διανύσματος αποτελούν τα χαρακτηριστικά τα οποία, το μοντέλο χρησιμοποιεί για να αναγνωρίζει μια λέξη ή μια πρόταση. Κάθε διάσταση αποτυπώνει κάποια πτυχή των γλωσσικών, σημασιολογικών ή συμφραζόμενων πληροφοριών του κειμένου. Σχεδιάσαμε ένα μικρό πείραμα ώστε να ελέγξουμε ποια *word embedding model* θα ήταν ιδανικότερο για τη μορφολογία των δεδομένων μας και ταυτόχρονα θα μας έδινε τη καλύτερη αντιστοιχία μεταξύ φαρμάκων και παρενεργειών.

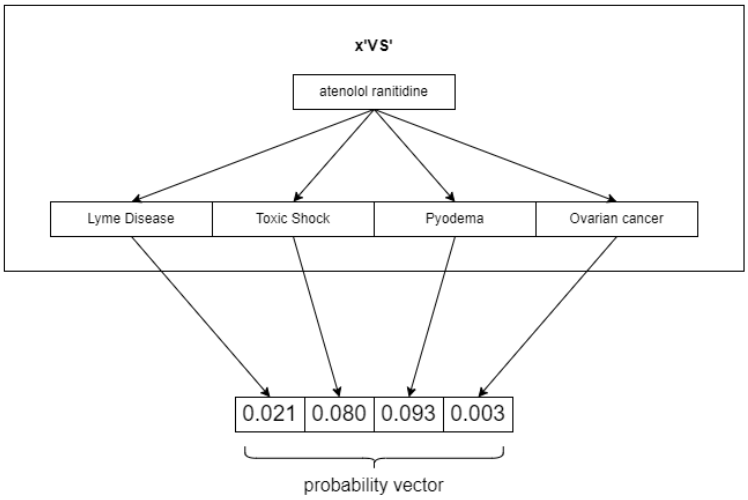


Φιγυρε 2. *Αξιολόγηση διάφορων NLP μοντέλων που έχουν εκπαιδευτεί σε βιοϊατρικά δεδομένα*

Καταλήξαμε στα εξής τρία *BioBert*, *BioClinicalBert*, και *SciBert*. Μέσω αυτών των μοντέλων μπορούμε να εξάγουμε λοιπόν τη μαθηματική αναπαράσταση των χαρακτηριστικών που χρειαζόμαστε για τους συνδυασμούς φαρμάκων αλλα και τις παρενέργειες. Με βάση τα προηγούμενα, μπορούμε με την εξαγωγή των χαρακτηριστικών από τα δεδομένα εκπαίδευσης,

να κατασκευάσουμε το πίνακα V ώστε να το χρησιμοποιήσουμε αργότερα για να κάνουμε προβλέψεις για νέες παρενέργειες.

Όπως είχε αναφερθεί κάθε συνδυασμός φαρμάκων μπορεί να οδηγεί σε περισσότερες από μία παρενέργειες. Ωστόσο η δήλωση $\arg\max_i xVS_i$ που είχαμε παραθέσει προηγουμένως, αντιστοιχεί μοναδικά την είσοδο x με κάποια από της κλάσεις του πίνακα S. Εξετάζοντας την έξοδο της έκφρασης xVS_i, παρατηρούμε ότι το αποτέλεσμα είναι ένα διάνυσμα z' διαστάσεων. Όπου το κάθε στοιχείο του διανύσματος σε μια τυχαία θέση i, αντιστοιχεί στη πιθανότητα της εισόδου x να ανήκει στη κλάση S_i.



Φιγυρε 3. *Παράδειγμα της έξοδου της έκφρασης xVS_i, για ένα νέο συνδυασμό φαρμάκων και τέσσερις νέες παρενέργειες.*

Μπορούμε λοιπόν να εκμεταλλευτούμε κατάλληλα το διάνυσμα αυτό για να κάνουμε περισσότερες από μια προβλέψεις για κάθε είσοδο. Είναι εμφανές ότι θα χρειαστεί να θέσουμε κάποιο όριο το οποίο θα διακρίνει τις προβλεπόμενες πιθανότητες σε ενδεικτικές και μη παρενέργειας. Για να ορίσουμε αυτό το όριο εξετάσαμε διάφορες τεχνικές, αυτή που απέδωσε καλύτερα για το πείραμα μας είναι η εξής: Για κάθε διάνυσμα που περιέχει τις πιθανότητες μιας εισόδου x να ανήκει στις διάφορες κλάσεις, τοποθετούμε σε αύξουσα σειρά τις τιμές των πιθανοτήτων και επιλέγουμε ώς όριο τη τιμή που βρίσκεται στο εκατοστημόριο το οποίο εμείς θέλουμε να δρα ως σημείο αποκοπής. Με αυτή τη τεχνική εξασφαλίζουμε ότι το όριο θα είναι δυναμικό για κάθε αποτέλεσμα μιας τυχαίας εισόδου και εξαρτάται μόνο από το ποσοστό που θα επιλέξουμε εμείς.

Αποτελέσματα

Προτού εξεταστούν τα αποτελέσματα είναι σημαντικό να αναφερθούν οι διάφορες παραδοχές που έγιναν. Κατα τη διάρκεια της αξιολογησης μια πολύ σημαντική υπόθεση που έγινε για τα δεδομένα ελέγχου είναι: ο ορισμός μιας μόνο πραγματικά αληθής ετικέτας για κάθε συνδυασμό φαρμάκων. Έτσι για όλους τους συνδυασμούς για τους οποίους γίνονται προβλέψεις από το μοντέλο, μόνο μια εκ των προβλεπόμενων παρενεργειών θεωρείται ως πραγματικά αληθής. Αναλογίζοντας τη προηγούμενη υπόθεση, είναι εμφανές ότι ο αριθμός των $false\ positives$ θα είναι αυξημένος σε μεγάλο βαθμό, κατά συνέπεια η πιο αξιόπιστη μετρική για τα πειράματα μας είναι η ανάκληση ή $recall$. Σε όλα τα πειράματα καμία από τις παρενέργειες που χρησιμοποιήθηκαν στο στάδιο της εκπαίδευσης δεν συμμετείχε στις παρενέργειες για τις οποίες έγιναν οι προβλέψεις στο κομμάτι του ελέγχου. Στη πρώτη πειραματική διάταξη, 3000 συνδυασμοί φαρμάκων και 150 παρενέργειες, χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου. Ακόμα 1000 νέοι συνδυασμοί και 50 νέες κλάσεις, ήταν διαθέσιμες για την αξιολόγηση του μοντέλου. Αποτελέσματα εξήγαμε από έξι διαφορετικούς συνδυασμούς φαρμάκων και παρενεργειών: (100, 5), (200, 10), (400, 20), (600, 30), (800, 40), (1000, 50). Ακόμα για την εξαγωγή χαρακτηριστικών από τα δεδομένα, εξετάστηκαν 3 διαφορετικά NLP μοντέλα $BioBert$, $BioClinicalBert$, και $SciBert$. Ώστε να εξεταστεί πως η επιλογή των χαρακτηριστικών επηρεάζει την απόδοση. Το ίδιο πείραμα πραγματοποιήθηκε με 5 διαφορετικά πακέτα δεδομένων για το στάδιο εκπαίδευσης άλλα και αξιολόγησης. Στα γραφήματα φαίνεται η διάμεσος της απόδοσης που παρουσίασε το μοντέλο σε αυτά τα πακέτα.

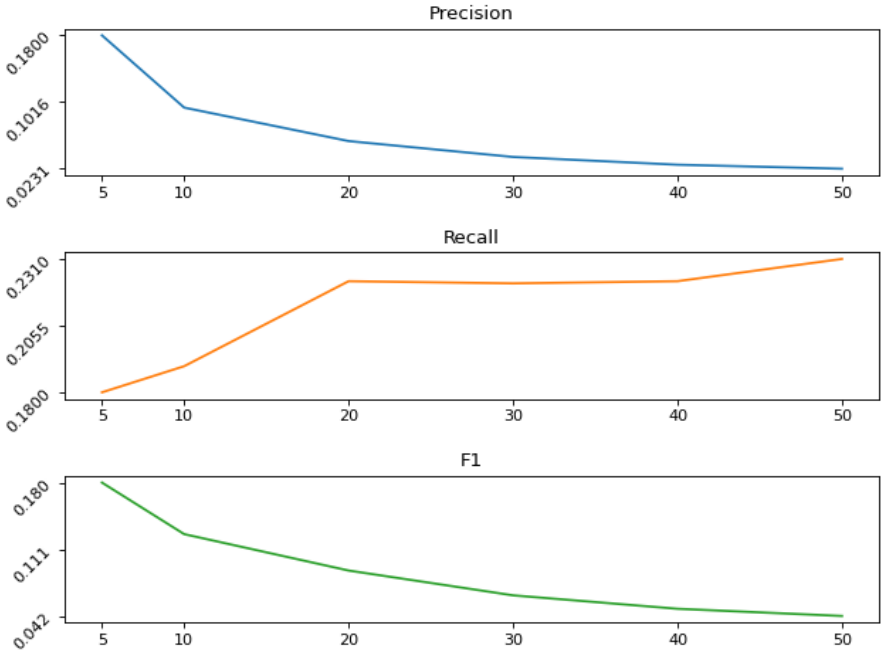


Figure 4. *Διάμεσος των αποτελεσμάτων, με χρήση του BioBert μοντέλου*

$NSideEffects$	$Precision$	$Recall$	$F1$
5	0.18	0.18	0.18
10	0.095	0.19	0.126
20	0.055	0.222	0.089
30	0.036	0.221	0.063
40	0.027	0.222	0.049
50	0.023	0.231	0.042

Table 1. *Διάμεσος των αποτελεσμάτων, με χρήση του BioBert μοντέλου*

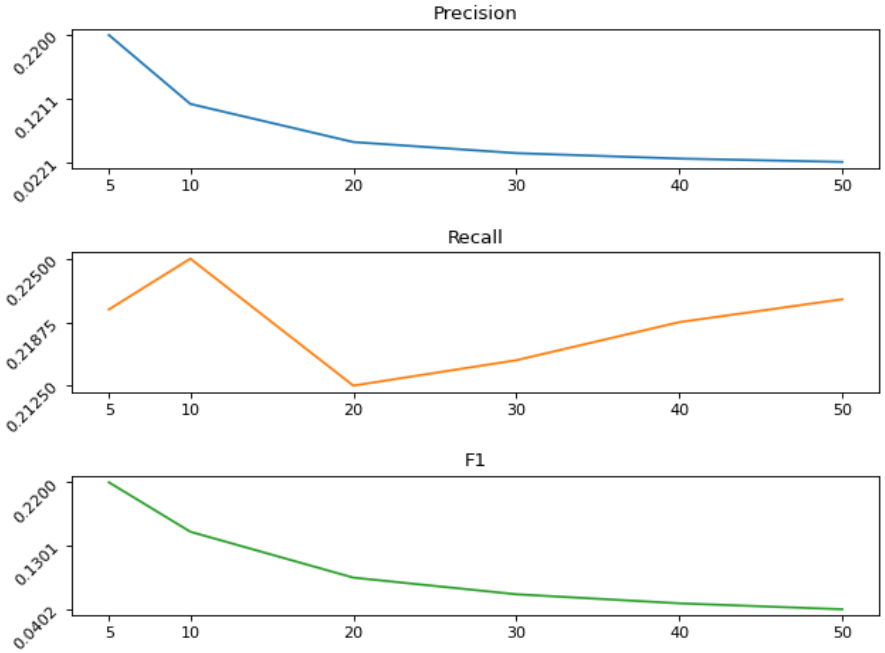


Figure 5. *Διάμεσος των αποτελεσμάτων, με χρήση του BioClinicalBert μοντέλου*

$NSideEffects$	$Precision$	$Recall$	$F1$
5	0.22	0.22	0.22
10	0.112	0.225	0.15
20	0.053	0.212	0.084
30	0.035	0.215	0.061
40	0.027	0.218	0.048
50	0.0221	0.221	0.040

Table 2. *Διάμεσος των αποτελεσμάτων, με χρήση του BioClinicalBert μοντέλου*

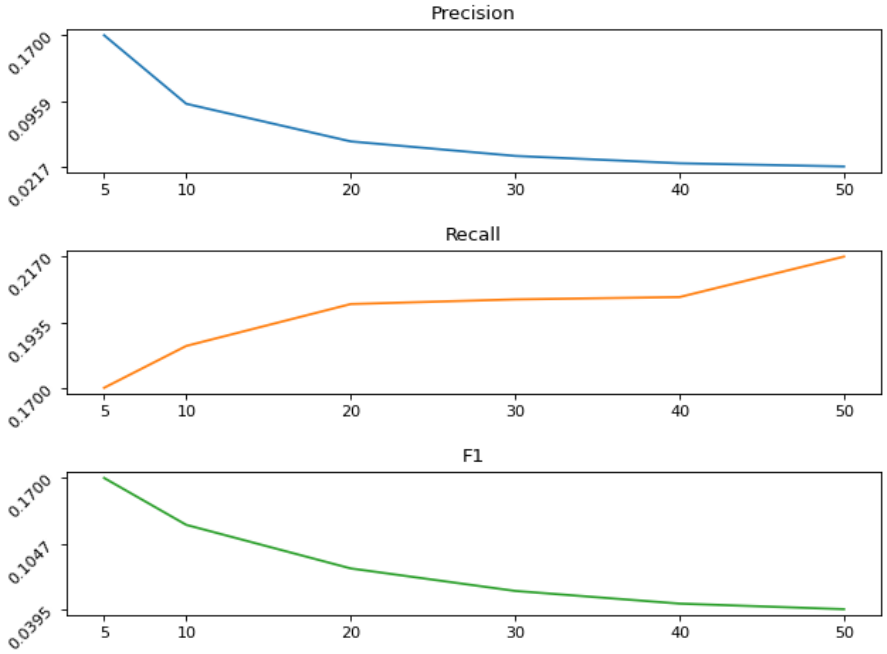


Figure 6. *Διάμεσος των αποτελεσμάτων, με χρήση του SciBert μοντέλου*

$NSideEffects$	$Precision$	$Recall$	$F1$
5	0.17	0.17	0.17
10	0.092	0.185	0.123
20	0.05	0.2	0.08
30	0.036	0.201	0.057
40	0.025	0.202	0.045
50	0.021	0.217	0.039

Table 3. *Διάμεσος των αποτελεσμάτων, με χρήση του SciBert μοντέλου*

Παρότι η ακρίβεια του μοντέλου δεν είναι αυτή που έχουμε συνηθήσει να βλέπουμε με άλλες τεχνικές T.N, αποτελει μια καλή γραμμή αναφοράς για ένα *zero shot learning* πρόβλημα. Όπως αναδεικνύεται και από το ποσοστό ανάκλασης, το μοντέλο έχει την ικανότητα να προβλέπει αλληλεπιδράσεις. Μάλιστα υποδηλώνει ότι η θεμελιώδης προσέγγιση του μοντέλου είναι υποσχόμενη και έχει το δυναμικό να βελτιωθεί περαιτέρω με πρόσθετη ανάπτυξη των χαρακτηριστικών. Μια ακόμα πειραματική διάταξη που εξετάστηκε και έχει αρκετό ενδιαφέρον είναι το κατα πόσο ο αριθμός των κλάσεων που είναι παρούσες κατά τη διάρκεια της εκπαίδευσης του μοντέλου επηρεάζουν την ικανότητα γενίκευσης του. Άρα ένα ακόμα αναγκαίο πείραμα, ήταν να ελεγχθεί πως το σύνολο των διαθέσιμων κλάσεων κατα την εκπαίδευση του μοντέλου επηρεάζει την ακρίβεια του. Ελέγχθηκαν οι περιπτώσεις που το μοντέλο εκπαιδεύτηκε με δεκατρείς, είκοσι πέντε, πενήντα και εκατό κλάσεις, με τις τρεις πρώτες να παρουσιάζουν κάποια αισθητή διάφορα από τα προηγούμενα δεδομένα.

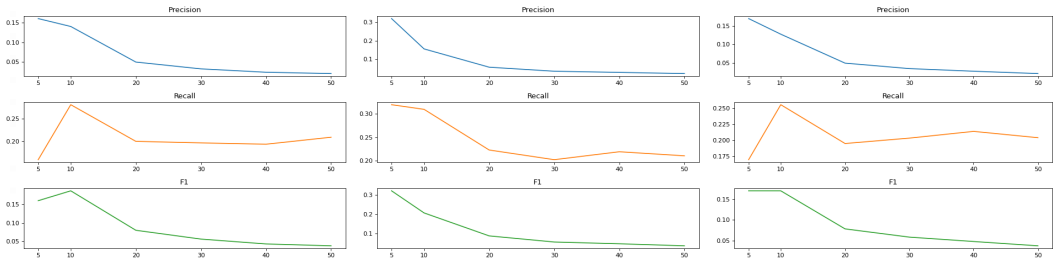


Figure 7. *Χρήση μόνο 25 παρενεργειών για την εκπαίδευση του πίνακα V, Αποτελέσματα κάνοντας χρήση τριών BERT μοντέλων (BioBert, BioClinicalBert, SciBert)*

Αξιοσημείωτη ήταν η αύξηση κατά 15% στο *recall*, που παρατηρήθηκε με τη χρήση του *BioClinicalBert*, όταν το μοντέλο εκπαιδεύτηκε με 25 και 50 κλάσεις, και κλήθηκε να κάνει προβλέψεις μεταξύ 5 και 10 νέων κλάσεων. Τέλος μια ακόμα πειραματική διάταξη που εξετάστηκε, ήταν η αντικατάσταση των κλάσεων εκπαίδευσης (παρενέργειες) από πιο γενικές κατηγορίες στις οποίες μπορεί να ανήκει ένα πλήθος από τις παρενέργειες αυτές. Μια μικρή αύξηση παρατηρήθηκε όταν το μοντέλο τέθηκε να προβλέψει μεταξύ πέντε και δέκα νέων κλάσεων, άλλα στη συνέχεια υπήρξε αισθητή μείωση της απόδοσης.

Συζήτηση και Μελλοντική Δουλειά

Σχολιάζοντας τα προηγούμενα, είναι ευκαιρία για ορισμένες επισημάνσεις στους περιορισμούς της εργασίας, καθώς και σε μελλοντικές βελτιώσεις. Βασικό στοιχείο για τη βέλτιστη επίδοση του μοντέλου, είναι η καλύτερη δυνατή συσχέτιση μεταξύ των χαρακτηριστικών μιας εισόδου με αυτών των διάφορων κλάσεων που είναι διαθέσιμες. Είναι εμφανές ότι όσο πιο αναλυτικά και ουσιώδη είναι τα χαρακτηριστικά που εξάγουμε για τα δεδομένα μας, τόσο αποτελεσματικότερη θα είναι η εκπαίδευση του μοντέλου και η ικανότητα του να κάνει προβλέψεις. Στη παρούσα προσέγγιση τα χαρακτηριστικά εξήχθησαν μέσω προ εκπαιδευμένων μοντέλων. Παρότι τα μοντέλα που επιλέχθηκαν είναι εκπαιδευμένα πάνω σε βιοϊατρικά δεδομένα, το μειονέκτημα τους είναι ότι έχουν κατασκευαστεί για γενική χρήση και σίγουρα όχι για το πολύ εξειδικευμένο πρόβλημα που προσπαθεί να λύσει αυτή η εργασία. Αυτό σημαίνει ότι τα χαρακτηριστικά που εξάγονται δεν είναι τα ακριβώς απαραίτητα για να προκύψει η βελτιστη συσχέτιση μεταξύ των συνδυασμών φαρμάκων και τις διάφορες παρενέργειες, το οποίο επηρεάζει σημαντικά την απόδοση του μοντέλου. Την απόδοση επηρεάζει ακόμα και η δυσκολία της επαλήθευσης των προβλέψεων του μοντέλου, αφού για να θεωρηθεί σωστή οποιαδήποτε πρόβλεψη πρεπει να τεκμηριωθεί από κλινικά δεδομένα. Σε μελλοντική δουλειά, θα ωφελούσε σημαντικά η εκπαίδευση ενός $wordembeddingmodel$, με γνώμονα το συγκεκριμένο πρόβλημα, ώστε τα χαρακτηριστικά που θα εξάγονται να αντανακλούν τις πολύπλοκες βιοχημικές διεργασίες που δρουν στο παρασκήνιο κάθε αλληλεπίδρασης.

Chapter 1

Introduction

1.1 Artificial Intelligence

The focus of Artificial Intelligence (AI), is creating and executing computer systems that can resolve complex problems typically requiring human intelligence [10]. These problems can involve tasks that are either natural or of a high level of complexity. To rephrase it in simpler terms, AI is a computing paradigm, that enhances a machine's capabilities to mimic how humans think and solve complex problems. In the same way as a human, AI can improve and self-correct from the mistakes it makes during the process of solving problems and so it's able to self-improve.

The conception of the idea of Artificial Intelligence isn't as modern as many think, with the concept first appearing as early as 1950 by Alan Turing and his very influential paper on the potential of programming a computer to act intelligently. AI has experienced substantial growth since then, with many advances in areas such as natural language processing, speech recognition, computer vision, and robotics. Due to the exponential growth AI has experienced, many subfields have been constructed with the two most prominent being Machine Learning and Deep Learning.

1.1.1 Machine Learning

Machine Learning (ML) is the subfield of AI, that refers to the ability of machines to resemble intelligent human behavior, using algorithms [11]. ML needs an extensive amount of data, that has to be collected and processed to be used as training data. This is the information under which the ML model will be trained on. Once a model has been selected, it's given the training data and allowed to learn on its own, identifying patterns and making predictions based on that data. Added, the model's parameters can be manually adjusted to help improve its accuracy. Machine learning consists of several learning methods, which include:

- Supervised learning: In supervised learning, the models get trained on information that consists of inputs and their expected output, which are more commonly called labeled data. That data enables the models to learn patterns and improve their accuracy over time. In practice, the models can generalize what they've learned and be able to predict correctly for previously unseen data. Today, supervised learning

is the most commonly used type of machine learning.

- Unsupervised learning: In a different context, unsupervised learning uses unlabeled data and from those tries to extract meaningful patterns not necessarily noticeable by the human eye. The model is left to discover on its own without any external guidance and draw conclusions, previously unthought or unknown.

- Reinforcement learning: Reinforcement learning trains models by experimenting and learning from mistakes, by implementing a feedback-dependent system. The model is trained to analyze input data and generate predictions. If the model's predictions are incorrect, corrective feedback is provided to help the model improve its accuracy. By doing so, the model can gradually learn which steps to take to generate correct predictions [12].

1.1.2 Deep Learning

Deep Learning is a subset of Machine learning that uses algorithms loosely modeled after the structure and function of the brain's neural networks [13]. Therefore, they are referred to as artificial neural networks. Neural networks are constituted of layers of interconnected nodes, with the number of layers oftentimes spanning tens or hundreds. These networks are capable of processing large amounts of data and setting a weight for each connection within the network. Deep learning models learn by consuming an extensive amount of labeled data and have the ability to determine features directly, without requiring manual feature extraction. The above constitutes a sizable contrast with the machine learning workflow, in which features must be manually extracted. Another key difference is the ability of deep learning algorithms to scale with data, as in comparison to ML where the accuracy flattens after a certain amount of training data.

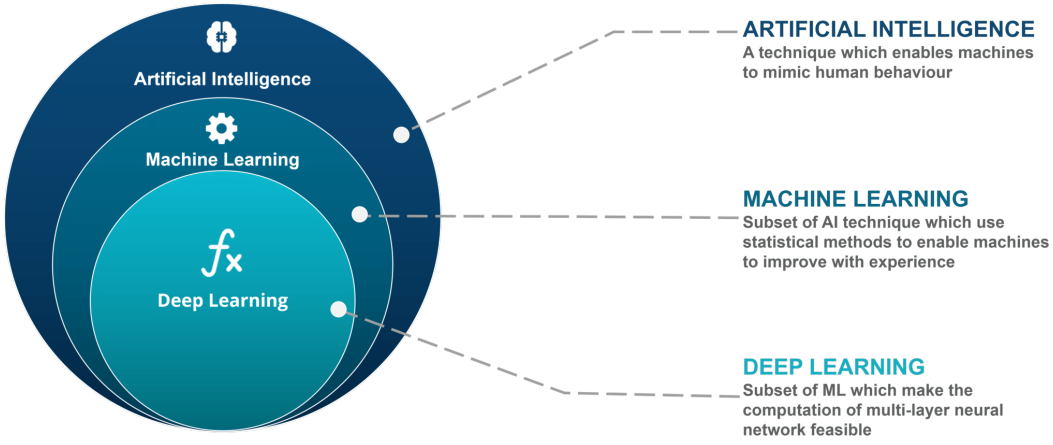


Figure 1.1. *Differences between AI, ML, and Deep Learning*

1.2 Zero-Shot Learning

Zero-Shot learning (ZSL) is a new learning paradigm in machine learning, in which the testing phase includes objects from classes, that were not included in the training, and the model must now classify them correctly [7]. ZSL made its first appearance in papers regarding computer vision. In computer vision, ZSL models were used to match images to new classes, by creating a similarity mapping between the attributes of samples and the classes they belong to.

There has been a rapid increase in the number of zero-shot learning methods proposed yearly. Traditional zero-shot learning methods, work by linking seen and unseen classes through some form of complementary information, which encodes characteristic properties of objects. So even if we haven't encountered a sample before, we can assess the class to which it belongs by observing its properties. These properties could include descriptive features such as color, shape, size, or other attributes that can be used to identify an object or class. That way by providing a high-level description of the new category that relates them to categories previously learned by the machine, we can access semantic information about the category to classify.

Information can be extracted from classes using various tools. We can get the textual definition of each class represented in a vector using modern language processing models. Another, key tool for strengthening the semantic relationship extraction can be manually tagged attributes.

One can see that Zero-shot learning can have a great impact in many applications, as it permits models to categorize new classes without requiring any additional training data.

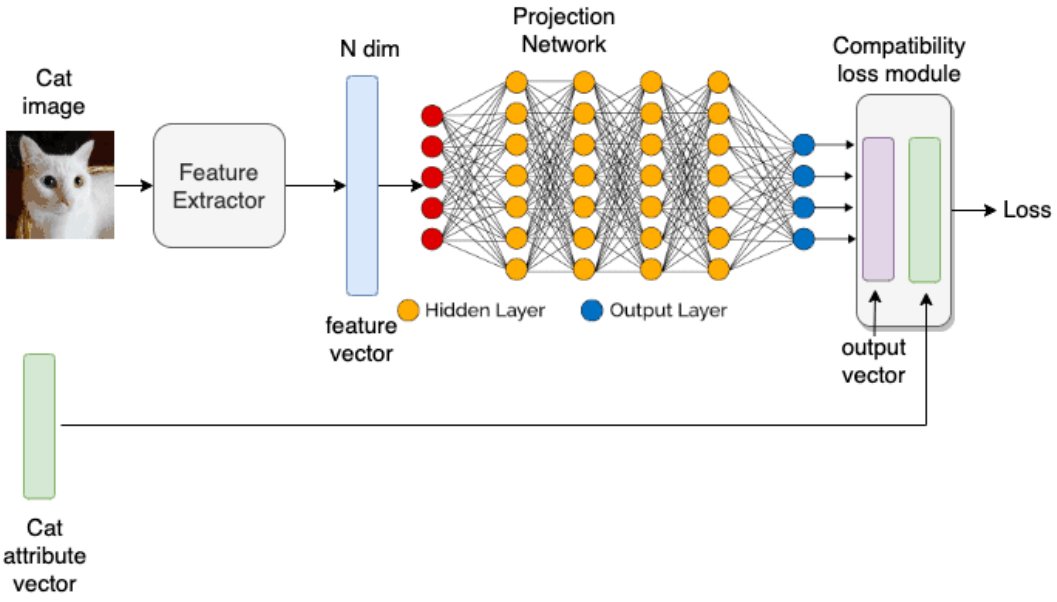


Figure 1.2. *Example of ZSL using embeddings to categorize images*

1.3 Word Embeddings

Word embeddings are a unique process of representing words, that grants words with similar context the ability to have a similar representation. Embeddings not only capture the context of a word but also the syntax and interpretation [14].

Most commonly, words are depicted through word embeddings with a dense numerical vector, usually containing floating point values. Higher dimensional embeddings capture more subtle dependencies between words, but on the other hand, need substantial amounts of data to train.

Word embeddings have been a part of every natural language process model, since their first utilization by researchers at Google in 2013. This signifies their importance and effectiveness, after all, they make it feasible to perform mathematical operations, by numerically representing whole sentences.

Deep learning models are used, to create these multi-layered word representations that capture both their features and their relationships to other words. Although word embedding models get trained in large amounts of generic text data, they can be also tweaked for a specific problem, cutting time and resources from needing to begin again.

Today there are many technologies avai-lable for creating word embeddings. In the past, two of the most influential techniques have been Word2Vec and GloVe. Word2Vec makes use of neural networks and a large dataset, to learn word embeddings. GloVe takes a different approach, adopting a co-occurrence matrix that keeps statistics for each word co-occurrence. More recently, Google's BERT offers state-of-the-art performance in many NLP problems. BERT takes advantage of deep learning architecture used in excessive amounts of text data, named transformers. BERT can process text in all directions and thus can achieve better relations extraction between words.

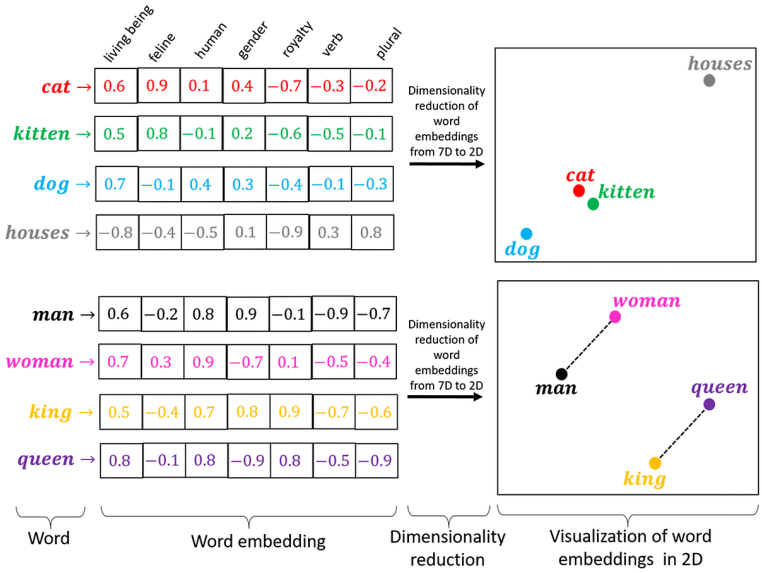


Figure 1.3. *Example of word embeddings vectors and word relationships*

1.4 Multi-label classification

Multi-label classification, unlike conventional classification tasks where a single label is predicted for each sample, involves predicting a combination of multiple labels. It expands upon multi-class classification, where each sample can be unambiguously categorized into one class from a group of several classes [15].

To clarify, in multi-label classification problems, the input is transformed into a binary vector, where one represents the presence of a class associated with the input. There is a wide array of applications for multi-label classification, especially in the area of text and image classification. It has a pivotal role in NLP text categorization tasks and computer vision where multiple labels are needed.

There are different techniques on how to approach these problems, the standard method, known as the binary relevance method, involves training a binary classifier for each label [16]. Then the predictions for each one are joined to accurately output the labels of an unseen input. Another interesting method is the label powerset transformation, where a binary classifier is created for every combination of labels. That way the problem becomes a multiclass classification problem, and can be dealt with accordingly. Furthermore, many classification algorithms have adapted to be able to tackle multi-label tasks as well, for example, the k-nearest neighbors algorithm has been extended to account for multi-label data.

Despite its versatility, multi-label classification comes with a range of unique challenges [17]. The varying frequencies of label occurrences within datasets can lead to imbalanced class distributions, particularly when certain labels rarely appear during training. This can impair the model's ability to generalize to underrepresented labels. Moreover, the computational complexity can significantly increase, as datasets often possess a substantial set of features that the model must train upon. Therefore, more computational resources may be required to handle the increased complexity.

Another challenge lies in evaluating multi-label classification models, as each input can possess multiple correct labels. Several metrics can be employed to assess model performance, including hamming loss, the Jaccard index, precision, recall, and F1-score. The specific metric utilized depends on the objectives and requirements of the task. For instance, in applications where false positives are costly, precision optimization is crucial, while recall prioritization is essential in settings where missing relevant labels pose a greater concern.

1.5 Drug-Drug Interactions

Drug-drug interactions (DDIs) are one of the most common causes of medication errors in developed countries and are responsible for 10–20% of the adverse drug reactions requiring hospitalization [1]. DDIs might occur when multiple drugs are administered conjointly. These interactions may result in either an increase or decrease in efficacy, treatment failure, or may even result in severe and debilitating drug-induced side effects. Extremely vulnerable to these types of interactions are elderly patients, as there is a strong relationship between increasing age, the number of drugs prescribed, and the frequency of potential DDIs [2].

Drug-drug interactions can be categorized into different categories, according to the underlying mechanism by which drugs interact, such as Pharmacokinetic, Pharmacodynamic, and Idiosyncratic [18, 19].

- Pharmacokinetic interactions: These interactions arise when a drug alters the absorption, distribution, and elimination of a coadministered drug. Altering the concentration of a coadministered drug can have severe clinical consequences and can lead to treatment failure or toxicity.

- Pharmacodynamic interactions: Occur when one drug alters the sensitivity or responsiveness of tissues to another drug by having similar (agonistic) or opposing (antagonistic) effects. In other words, combining drugs can have an additive reaction on their effects, leading to an exaggeration or mitigation of them.

- Idiosyncratic interactions: These are adverse drug reactions that are not related to the known pharmacological properties of the drugs and occur in only a small percentage of the population and do not show any apparent relationship.

To effectively mitigate the detrimental consequences of DDIs, healthcare providers need to review patients' medication lists carefully, use computerized systems to alert them to potential problems and educate patients about the importance of reporting all medications they are taking.

Chapter 2

Thesis Outline

As previously mentioned, drug-to-drug interactions can induce adverse effects on patients, which can be dangerous or even life-threatening. Additionally, DDIs can have consequences on the efficacy of drugs, can affect their therapeutic benefit, and make it difficult for doctors to prescribe the appropriate medication regimen. Evidently, being able to predict ahead of time DDIs, can be an essential tool for managing them in a clinical environment and help to ensure that patients receive safe and effective treatment.

One way to approach this issue is by creating a computer system capable of identifying DDIs either by predicting them or identifying known interactions. There has been a significant amount of progress in research using machine learning methods in the past years. Some of the approaches using traditional machine learning methods are discussed shortly.

What is yet to be widely explored, is the application of zero-shot learning to predict DDIs. Zero-shot learning offers the ability to generalize from seen to unseen data, enabling predictions for data not included in its training. Although the intricate nature of drug interactions poses a challenge, the captivating idea of utilizing zero-shot learning keeps the door open for innovative solutions.

2.1 Approaches for Predicting DDIs

Similarity-Based Approach

The concept behind this approach is as follows: if drug A interacts with drug B and induces a specific side effect, then drugs that are comparable to drug A are likely to yield the same side effect when interacting with drug B. Researchers Vilar et al. [5], made use of the similarity-based method by extracting the molecular similarity of drugs, using a bit vector to encode molecular features.

Matrix factorization Approach

This approach reveals concealed relationships in extensive datasets and has become a valuable asset in predicting interactions between drugs. By harnessing latent patterns extracted from existing DDI data, matrix factorization aptly anticipates interactions among drugs sharing similar attributes [20].

Classification-Based Approach

A binary classification problem is simulated to predict DDIs in the traditional classification-based approach. Inputs of DDI and non-DDI pairs are used to train a classification model. Various classifiers like Bayesian, k-nearest neighbor, logistic regression, random forest, and support vector machines (SVM) are employed to build a model. Using a link-prediction technique, Kastrin et al. [6] considered the prediction of unknown drug interactions in five large DDI databases as a binary classification task. They further improved the network topology features by incorporating four semantic characteristics. In the link predictions approach, a graph is constructed using the drugs (or other biomedical entities) as nodes and their connections and interactions as edges. Next, algorithms such as random walk are used to predict missing links between nodes and thus identify missing interactions. This approach has been adopted by many researchers and tweaked according to their study.

2.2 Thesis Description

In simple terms, this thesis aims to predict potential adverse side effects that may arise from the combination of two drugs, where either the side effects or drug pairs have been included at the training stage. To do that, we employ the use of a zero-shot learning framework that has been carefully adjusted to account for the multilabel nature of the problem. Our goal is to first establish a mathematical relationship between a sample of known drug interactions and their side effects. And use this relationship to make predictions for new unseen drug pairs and side effects. Additionally, It's very important to build on this relationship, extracted from the framework, to achieve a one-to-many correlation between drugs and side effects. As with most zero-shot learning problems, the target of this thesis isn't to be a stand-alone mechanism for tackling this very complex problem of identifying DDIs but to be combined with other methods, such as the ones mentioned above. That way we can ensure better results, especially in the case of making predictions for unseen side effects.

Chapter 3

Analysis and Design

3.1 Brief Architecture Review

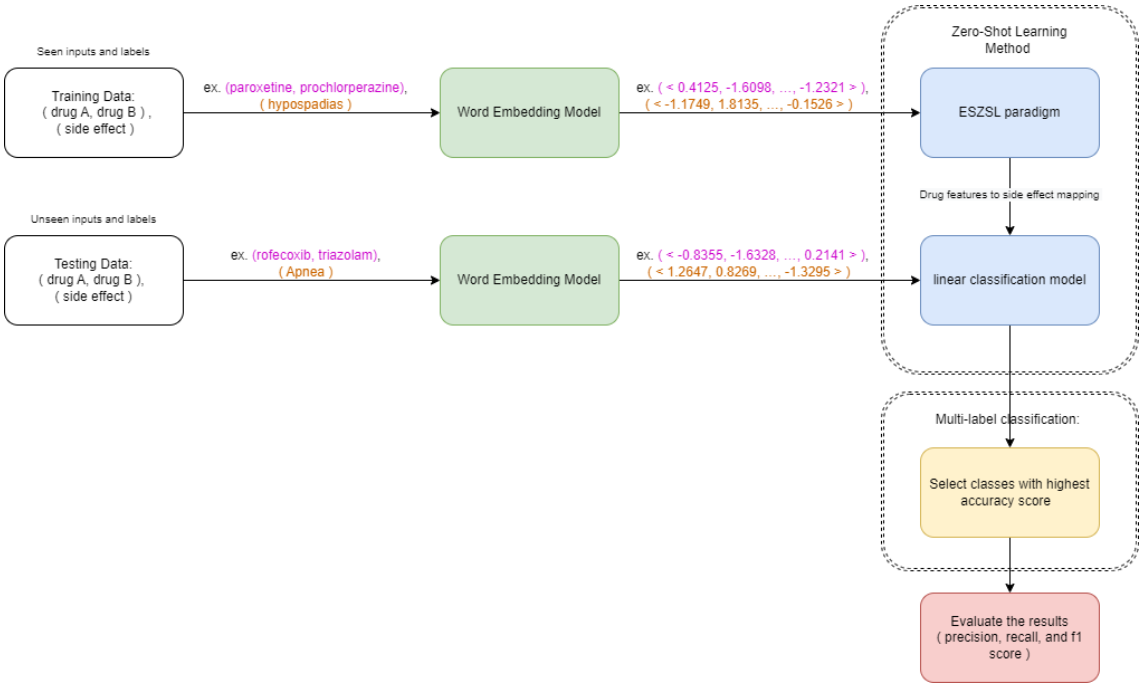


Figure 3.1. *Summary of the thesis architecture*

The following overview provides a high-level account of the steps employed in this thesis architecture, intended as a concise introduction to its different components.

The first step of our extensive process is to gather data from open-source datasets and mold it to align with our requirements. The data format consists of drug pairs and their corresponding known side effects. In this state, the data are nothing but a string of characters and can't provide us with the words and semantic information we very much require.

To solve this problem, each piece of data we collected, is sent through a word embedding model, trained explicitly on biomedical text. The resulting vector, or the yield of the model, is then stored to be used in the subsequent steps.

We have chosen to utilize the ESZSL framework, an acronym for Embarrassingly Simple Zero-Shot Learning. We will examine this framework, in more detail at a later point.

Essentially, this process enables the establishment of a mapping between drug features and their corresponding side effects. Consequently, a linear classification model gets trained, which can be utilized for classes that have not been encountered previously.

Another aspect we do consider in our architecture is that a drug pair might have the potential to produce more than one side effect. Therefore, in the final step of our process, we manipulate the linear classifier to generate predictions for multiple classes and make selections based on the probability scores assigned to each class.

In the final stage, we will employ unseen data to evaluate the effectiveness of our architecture and its ability to make predictions for side effects that were not included in the training stage.

3.2 ESZSL framework

ESZSL is a framework introduced by Bernardino Romera and Philip Torr in their paper "An embarrassingly simple approach to zero-shot learning" [8]. Its inherent goal as a zero-shot learning framework, is to classify objects without previously having any training data for those specific classes. ESZSL takes a novel approach by leveraging auxiliary information and exploiting the relationships between seen and unseen classes.

Two options for auxiliary information in ESZSL are semantic embeddings or attribute vectors linked to the classes. These embeddings capture and represent the semantic information or attributes associated with the classes. By incorporating embeddings into the model, ESZSL has the ability to understand and recognize the intricate relationships that exist among various classes. This enhanced understanding enables the framework to make more informed and accurate classifications, even for classes that have not been previously encountered.

Therefore, the primary goal of ESZSL is to train a function that can accurately calculate the probability of an object belonging to a specific class. This function takes as input an object's features and a class semantic embedding and generates as output the similarity score between the two. To achieve this during the training stage, a labeled dataset consisting of objects with their corresponding class embeddings is utilized to optimize the function's parameters. The training process involves continuously adjusting these parameters to minimize the discrepancy between the predicted similarity scores and the ground truth labels. Below, we'll provide a more in-depth account of the framework's mathematical principles and logic.

ESZSL employs a wide range of mathematical concepts and techniques to tackle the challenge of zero-shot learning. At the core of this framework lies the fundamental principles of linear algebra and specifically specialized matrix operations. Below, we will provide representations and compositions of the various matrices involved in the framework:

- $S^{a \times z}$: Represents a set of classes associated with attributes, where a denotes the number of attributes, and z denotes the number of classes.
- $X^{d \times m}$: Represents a set of instances and their corresponding feature vector, where d denotes the dimensionality of the data, and m denotes the number of instances.

- $Y^{m\times z}$: Represents the ground truth labels of each training instance belonging to any of the z classes.

The paper initially presents an equation designed to facilitate the learning of a linear predictor for a given set of training classes:

$$\min_{W\in\mathbb{R}^{d\times z}} L(XW, Y) + \Omega(W) \qquad (3.1)$$

In the preceding equation, W represents the parameters to be learned, L is the chosen loss function, and Ω is the regularizer. The particular selection of L and Ω can lead to numerous approaches for addressing the problem at hand. In the previous problem, the attributes are not utilized, resulting in a lack of knowledge transfer from the existing set of classes to new classes. To incorporate the given information about attributes, the equation is modified as follows:

$$\min_{V\in\mathbb{R}^{d\times a}} L(XVS, Y) + \Omega(V) \qquad (3.2)$$

The matrix $V^{d\times a}$ is derived by the equality $W = S^T V$. Utilizing matrix V introduces the attributes into the equation and enables the transfer of knowledge from training classes to new ones. Ultimately our objective is to differentiate among a new unseen set of z' classes. The framework needs to be provided with a matrix of their attribute signatures donated by $S'^{a\times z'}$. Then for a new instance x, a prediction can be given using the subsequent statement:

$$\arg\max_i xVS'_i \qquad (3.3)$$

The above formula provides us with a straightforward and accurate method of making predictions, given that we are able to precisely calculate matrix V. In light of this, the authors of the paper introduce two critical assumptions concerning the regularizer. The first assumption focuses on controlling the Euclidean norm of the representations of attribute signatures on the feature space. Controlling the norm promotes fairness and equal consideration of all attribute signatures. The second assumption aims to limit the variance of representations of instances on the attribute space. By ensuring invariance, the model's ability to generalize well to unseen feature distributions is enhanced, making it more reliable and effective in practical scenarios. By incorporating these two properties, the model becomes more balanced, unbiased, and adaptable to diverse data scenarios. A regularizer that accomplishes the previous terms, has a solution for equation (3.2) that can be expressed in closed form in the following way:

$$V = (XX^T + \gamma I)^{-1} XYS^T (SS^T + \lambda I)^{-1} \qquad (3.4)$$

3.3 Multi-label classification

As mentioned before, we are addressing a multi-label classification problem due to the possibility of multiple side effects resulting from each drug interaction. However, a few issues arise when examining the solution provided by the ESZSL framework in equation

(3.3). Using $\arg\max_i$, means that only a single side effect gets selected for each drug pair, based on having the maximum output, obtained from the compatibility function. Hence, some adjustments to the framework are necessary to meet our specific requirements.

If we calculate the compatibility scores for each instance and class embedding, following the standard ESZSL framework. We are presented with multiple scores, each one associated with a distinct instance-class pair. What we need to accomplish is to filter out the pairs that exhibit scores indicative of a relationship. These can be achieved by establishing a threshold for the scores, the threshold acts as a lower bound, indicating the minimum score required for a label to be considered present for an instance. This threshold is essential for making predictions for each instance and its many class pairs.

That threshold can be computed by many different methods [21], for example, the usage of a validation data set or the f1 score to find the optimal trade-off between precision and recall. Our approach uses the top percentile score method, we sort all the compatibility scores in descending order and identify the score at the chosen percentile. This score will be used as the threshold for classifying labels. For each instance, we compare the compatibility scores of the class embeddings to the threshold. If a compatibility score for a particular label is higher than or equal to the threshold, we consider that label as present for the instance. Otherwise, scores below the threshold suggest the absence of the label.

The top percentile score method enables us to adapt the threshold to the specific dataset and distribution of compatibility scores. By sorting the scores and selecting a percentile, we take into account the varying strengths of relationships between drugs and side effects. This approach is particularly suitable for our problem, as we prioritize a higher recall score, aiming to capture as many relevant labels as possible.

3.4 Evaluation

The essence of zero-shot learning lies in its ability to generalize beyond its training data, and make predictions for previously unseen classes. To accurately assess our model's performance in this domain, we employ a separate dataset comprised entirely of novel classes and evaluate how well it predicts the potential drug-to-drug interactions for each instance. We begin by feeding this new dataset through the embedding model, identical to the process for the training data. Next, we use our trained linear classifier in conjunction with our multi-label classification approach to generate predictions. Each instance is associated with a specific set of labels. In a conventional machine learning setting, evaluating the model's accuracy would entail utilizing the true labels present in the testing dataset. However, our challenge arises from the absence of comprehensive clinical data for all possible drug combinations. While we can obtain limited reference labels from known drug-to-drug interactions, we cannot definitively validate or refute the model's predictions for the broader spectrum of interactions. Therefore, we employ a more tailored evaluation method: identifying the overlap between our model's predictions and the true labels we currently possess for each instance. This approach emphasizes the recall of our model, also known as sensitivity, which measures the proportion of correctly predicted positive labels out of the actual positive labels. Recall serves as a crucial indicator of our model's

ability to effectively identify and capture the inherent relationships between drugs and their potential interactions. This assessment, coupled with the limited reference labels from known interactions, provides a comprehensive evaluation of our model's performance in the zero-shot learning domain.

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.5}$$

Chapter 4

Implementation

4.1 Data collection

The first and foremost requirement is to obtain the necessary information to implement the previously described architecture. This information is crucial for both training and testing purposes. Without it, our ability to develop the architecture would be obscured, making it essential to prioritize and acquire the necessary data to proceed with the implementation process. We acquired our data from the publication titled "Modeling polypharmacy side effects with graph convolutional network" [22]. The dataset which interested us the most was the "Drug-drug interaction and side-effect dataset". This information on polypharmacy side effects was extracted from the TWOSIDES database, which provides comprehensive details on 1,318 types of side effects across 63,473 drug combinations. This database was generated using adverse event reporting systems that collect reports from doctors, patients, and drug companies. Below is provided a sample from the dataset.

STITCH 1	STITCH 2	Polypharmacy Side Effect	Side Effect Name
CID000002173	CID000003345	C0151714	hypermagnesemia
CID000002173	CID000003345	C0035344	retinopathy of prematurity
CID000002173	CID000003345	C0004144	atelectasis

Table 4.1. *Sample from the Drug-drug interaction and side-effect dataset*

As mentioned in the preceding sections, a subsequent step requires generating word embeddings for our data. However, the embedding models we tested failed to capture the semantic information of the drugs when given the CID ID of each drug. CID, which stands for Compound Identifier, is a distinct identification number assigned to chemical compounds in the PubChem database. Consequently, we processed the dataset by substituting the CID ID with the most commonly used scientific name for each drug. Additionally, we eliminated the polypharmacy side effect ID as it was not needed in our procedure. Below is provided a sample from the processed dataset.

DRUG 1	DRUG 2	Side Effect Name
ampicillin	fentanyl	hypermagnesemia
ampicillin	fentanyl	retinopathy of prematurity
ampicillin	fentanyl	atelectasis

Table 4.2. *Sample from the processed Drug-drug interaction and side-effect dataset*

The dataset includes 1,048,575 drug-drug-side-effect pairs. Since we will be performing matrix operations later on, we have divided the data into smaller chunks. This approach allows us to maintain smaller matrix dimensions and ensures that our calculations remain manageable. The reduced dataset we have created consists of 4,000 distinct drug pairs and 200 distinct side-effects, some of which will be utilized for training while others will be allocated for testing purposes.

4.2 Selecting an Optimal Word Embedding Model

As aforementioned, we will be passing our data through a word embedding model to extract the features that we will leverage later in the process. Choosing the right word embedding model can depend on many factors, but foremost, it has to do with the task at hand. Since word embeddings, are not the central point of investigation in this thesis but simply a tool, we opted to choose a pre-trained BERT model. Thankfully, there is an array of BERT models fine-tuned on an extensive volume of biomedical text. To evaluate the ability of each model to identify potential drug interactions, we had to design a small-scale test. We selected a pair of drugs and their corresponding side effect from the dataset. Additionally, we selected a side effect that was not yet known to be caused by the chosen drug pair, this way we ensured that the model could recognize both known and unknown interactions. For each model, we passed the drug pair as a single string representing the combined drug names and separately passed the side effect names. This allowed us to embed and compare the drug pair and side effects independently. After embedding, we calculated the cosine similarity scores for the drug pair embedding and its corresponding detected side effect, as well as a score for the drug pair and the side effect not known to be caused by the interaction. In this way, we can evaluate the model's ability to distinguish between drug pairs that have and have not been observed to interact, indicating its suitability for our application. A model that exhibits consistently higher cosine scores for the detected side effect compared to the unknown side effect would be considered more effective in predicting potential interactions. We ran the previous test for several drug pairs and used seven different models (BioBert, ChemBert, BlueBert, BioClinicalBert, SciBert, PubMedBert, Sentence-Bert). The bar graph below illustrates the comparison between cosine scores for clinically detected side effects and randomly selected ones across all candidate models. The blue bar indicates the percentage in which the cosine score was higher for detected side effects, while the red bar represents the corresponding percentage for the randomly selected side effects.

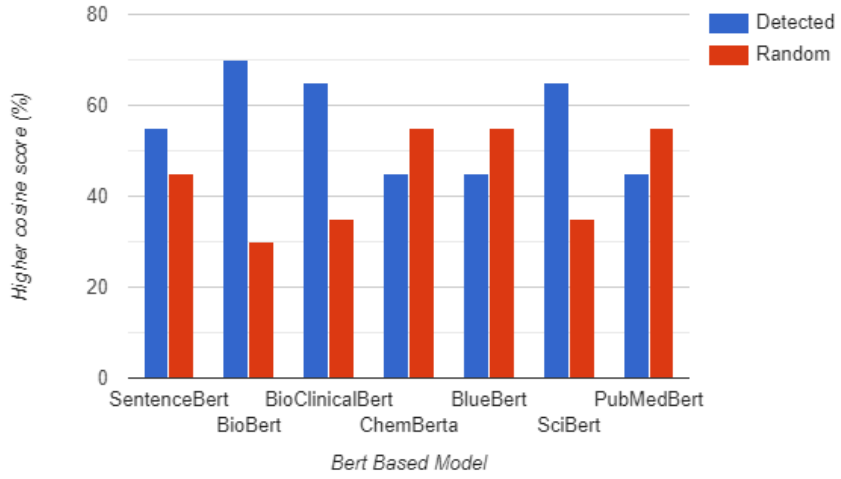


Figure 4.1. *Results of the various models*

As is evident from the bar graph above, the BioBert, BioClinicalBert, and SciBert models emerged as the top performers. While BioBert marginally surpassed the other two, it's not necessarily the optimal choice for our final experiment. This preliminary test serves to narrow down our selection of models, not necessarily identify the absolute best performer.

4.3 Implementing the ESZSL framework

To fully harness the capabilities offered by the ESZSL framework, we must transform the concepts and algorithms outlined in the paper into functional code. For this task, will make use of the Python programming language. Python's intuitive syntax and rich ecosystem of libraries make it a great candidate for our application. Specifically will take advantage of the widely prevalent NumPy library. NumPy's broad range of tools, especially for array operations will significantly assist with translating the process detailed in the paper.

We have divided the code into several core modules, which significantly enhances code readability by breaking it into smaller, self-contained units. Furthermore, this approach improves our capacity to expand and integrate new features seamlessly. For now, will focus on two of them. In the first module, we'll create all the matrices described in the framework, that are integral to construct the main matrix V. In the second module, we'll implement the novel one-line solution proposed in the paper (equation 3.4) and finally produce matrix V.

Previously we discussed that we'll be employing the help of pre-trained word embedding models. The models we choose all have expanded above BERT, a natural language processing model developed by Google, where its generated vector has a dimensionality of 768. By considering that we use these embeddings to extract attributes for both drugs and side effects, it's apparent that both S and X matrices described in Chapter 3, will have

their rows bounded by the dimensionality of 768.

Before proceeding it would be helpful to take a moment and briefly explain how we transform our data from their form in Table 4.2, to accommodate for the framework requirements. To construct matrix S, we take each side effect name and pass it through our chosen embedding model. If there are x number of side effects this will result in a vector with dimensionality of x. Where each element of the vector itself has a dimensionality of 768. To construct the matrix S, will take the aforementioned vector and stack each element as a column, to be able to do this correctly the only prerequisite is that all elements must have the same dimension, something our vector satisfies. The same process is used to construct matrix X, this time by embedding the concatenated string of the two drugs using the chosen embedding model. Finally, matrix Y, containing truth labels, is formed. With x rows where x represents the number of training drug pairs and z columns for the number of training classes. Each row denotes a specific instance and is populated with zeros, except for the class it belongs to, marked with a one. Below is provided a snippet of pseudo-code entailing the steps mentioned above.

Αλγριθμσ 4.1: *Create the input matrices*

```
Let S be an array containing z number of side effects
Let X be an array containing x number of drug pairs
Let Y be an array of size x rows and z columns
for each element in S do
    Transform element into a vector and store it in its original location.
end for
Take all vectors from array S and stack them as columns to make a single 2-D array
for each element in X do
    Transform element into a vector and store it in its original location.
end for
Take all vectors from array X and stack them as columns to make a single 2-D array
for i = 0 to rows - 1 of Y do
    for j = 0 to columns - 1 of Y do
        if the element in the row i belongs in class j then
            Set element in row i and column j to 1
        else
            Set element in row i and column j to 0
        end if
    end for
end for
```

With the matrices constructed, we can focus on solving equation 3.4 and obtaining matrix V. One last thing we should make a note of is the hyperparameters γ and λ which ensure that the model avoids overfitting while maintaining the ability to generalize effectively to unseen data. Selecting optimal hyperparameter values, to ensure the model's highest performance requires a meticulous search process, typically involving trial and error. In our case, we opted for the grid search approach, which involves defining a value range for each hyperparameter and exhaustively evaluating all possible combinations. For our specific application, the most suitable combination of values proved to be γ set to zero

and λ set to three.

Ultimately, we anticipate the dimensions of V to align with the dimensions of the feature vector multiplied by the number of class attributes. We utilize the NumPy library to perform the different matrix operations mentioned earlier. In the pseudocode provided below, we outline the core steps at a higher level.

Αλγοριθμσ 4.2: *Calculating matrix V*

1: Initialize matrix $\mathbf{V}$ with dimensions 768×768 and filled with zeros
2: Set γ to the optimal value determined through experimentation
3: Set λ to the optimal value determined through experimentation
4: Compute matrix multiplication of $\mathbf{X}$ and its transpose
5: Regularize the result of $\mathbf{XX'}$ by adding $10^{\gamma}\mathbf{I}$
6: Calculate the pseudo-inverse of the sum
7: Store the result in **tmp1**
8: Compute matrix multiplication of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{S'}$
9: Store the result in **tmp2**
10: Compute matrix multiplication of $\mathbf{S}$ and its transpose
11: Regularize by adding $10^{\lambda}\mathbf{I}$ to the result of $\mathbf{SS'}$
12: Calculate the pseudo-inverse of the sum
13: Store the result in **tmp3**
14: Compute matrix $\mathbf{V}$ by multiplying **tmp1**, **tmp2**, and **tmp3**

Having completed the above steps, we are now able to utilize the V matrix to make predictions and evaluate the precision of the model.

4.4 Filtering Mechanism for Multi-Label Predictions

As previously mentioned, using equation 3.3 as is, for making predictions is limiting for our type of use. Nevertheless, the equation is still extremely valuable. If we disregard the need to locate the optimal value and remove arg max, the equation will produce a vector. Each element within this vector will contain a probability score. Will delve deeper into what this probability score signifies. Let's indicate as x' the feature vector of a new drug pair instance and S' the matrix containing a new unseen set of side effects and their attributes. The multiplication of $x'VS'$ will result in a vector with its number of components being equal to the number of new side effects in matrix S'. Each component of the vector will hold a score. This score signifies the probability of the instance x' belonging to each class of the S' set. In other words, the probability of drug pair x' inducing a side effect from the set S'. To be more precise, if we were to state that the initial row of matrix S' represents the attributes of side effect y, then the corresponding element in the probability vector's first component would indicate the probability of drug pair x' causing side effect y and so on.

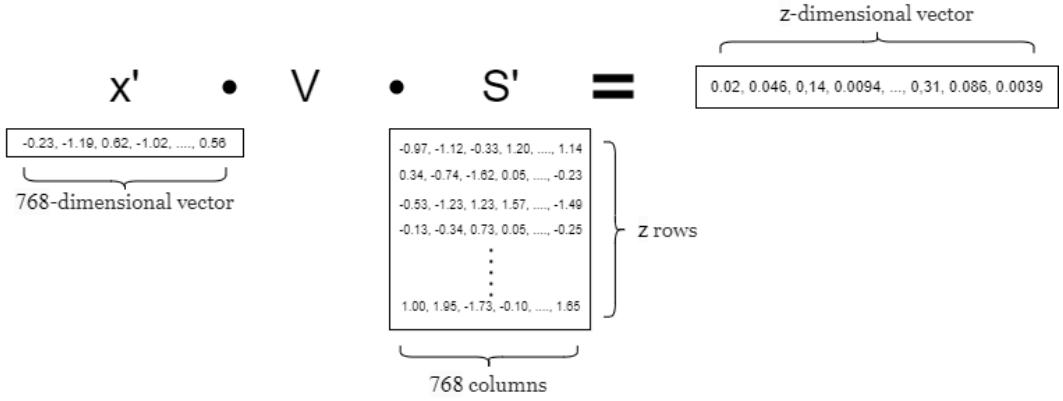


Figure 4.2. *Overview of the $x'VS'$ equation*

As we observe the equation in its current form, it only has the ability to process one instance at a time. Our goal is to enhance this capability by directly feeding the matrix X', which contains the testing instances, to the function. Achieving that is a straightforward task. We will substitute x' with the transpose of the matrix X'. Now, the equation will be represented as $X'^{T}VS'$. The result of the equation will transform into a matrix, where the rows correspond to the number of drug pair instances in X', and the columns are equivalent to the count of side effects in S'. Every row in the resulting matrix will contain the probability scores of distinct instances within X'. The figure below provides a high-level overview of both the process of calculating the probability vector and matrix.

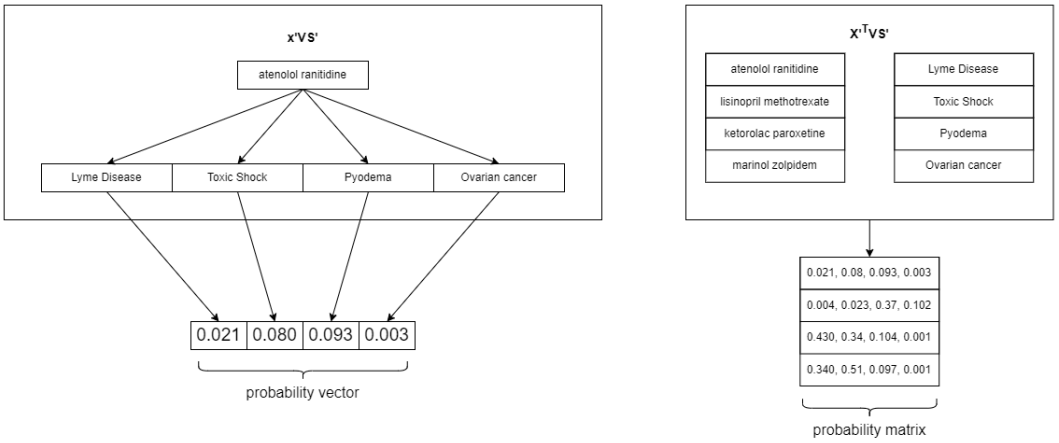


Figure 4.3. *Probability vector and matrix*

In the following phase, we'll process the probability matrix, through a necessary filtering mechanism. Our approach is centered around the establishment of a threshold value. This threshold value serves as a critical criterion for determining which data points should be retained. By applying this threshold to the probability scores, we identify instances that exceed a predefined level of significance and can discard all the rest. Determining the appropriate threshold can be a challenging task. Unlike static thresholds that remain constant across all instances, dynamic thresholds adapt to the characteristics of each row within the probability matrix. A threshold set too low might capture too many irrelevant

relationships, leading to a high rate of false positives. On the other hand, a threshold set too high might miss relevant associations, resulting in false negatives.

To address this, we explored several methods for establishing a threshold. Although we considered alternative options based on statistical methodologies, they often struggled to adapt to the dynamic threshold requirement. As explained in section 3.3, we ultimately opted for a top-percentile approach due to its ability to provide a well-balanced outcome.

The top-percentile approach involves selecting a percentage value, such as 5% or 10%, which represents the upper portion of scores within each row of the probability matrix. This percentage value serves as a filter, identifying the threshold score for each row. The flexibility of this approach shines through as it automatically adjusts to the distribution of scores within each row, capturing the most significant associations while accounting for variation.

Below is a brief pseudo-code snippet illustrating the filtering process. Ultimately, we modify the probability matrix such that it marks a spot with 1 if we identify a relationship between the drug pair and the side effect, otherwise, it is marked as 0.

Αλγοριθμος 4.3: *Transforming the probability matrix*

```
Set P as the probability matrix
for each row in P do
    sorted_vector = sort(row)
    min_index = ⌈(length(sorted_vector) − 1) × percentile⌉
    min = sorted_vector[min_index]
    for each element, index in row do
        if element ≥ min then
            row[index] = 1
        else
            row[index] = 0
        end if
    end for
end for
```

4.5 Evaluation Method

The model's training phase involved the utilization of 3,000 distinct drug pairs and 150 corresponding side effects. For the testing process, 1,000 drug pairs and 50 side effects were reserved. To ensure the robustness and accuracy of the findings, multiple iterations of training and testing were performed, each time incorporating distinct sets of drug pairs and side effects. This systematic approach served to validate the model's performance consistently. By using a separate set of 1,000 drug pairs and 50 side effects, we aimed to measure the model's ability to classify correctly interactions in previously unseen data and test its potential as a zero-shot learning paradigm.

As we have discussed, our approach involves generating multiple predictions for each instance of drug pairs. However by design, in the testing dataset, each instance is associated with only one truth label. Due to the complexity inherent in our data, while we do make

predictions, validating their accuracy or falsehood presents challenges. So we opted to classify as false positive every prognosis except for the ones matching the truth labels. Consequently, each instance can have only one true positive, while the number of false positives is dependent upon the number of predictions we'll permit the model to make for each instance. It's apparent that the high number of false negatives could significantly skew the accuracy of measurements. Hence the recall measurement, which is controlled by true positives and false negatives, is the most accurate evaluation metric for this thesis.

As explored earlier, a probability matrix can be generated by employing the feature matrix of the testing instances denoted as X', the signature matrix S' associated with the testing side effects, along with the V matrix derived from the testing data. Once the probability matrix has been filtered, each instance will have a set of predictions belonging to the interactions it could potentially generate. Before evaluating the predictions, similar to a previous step we must formulate the ground truth label matrix for the testing instances. This procedure enables us to compare the two matrices and calculate the parameters essential for our accuracy metrics.

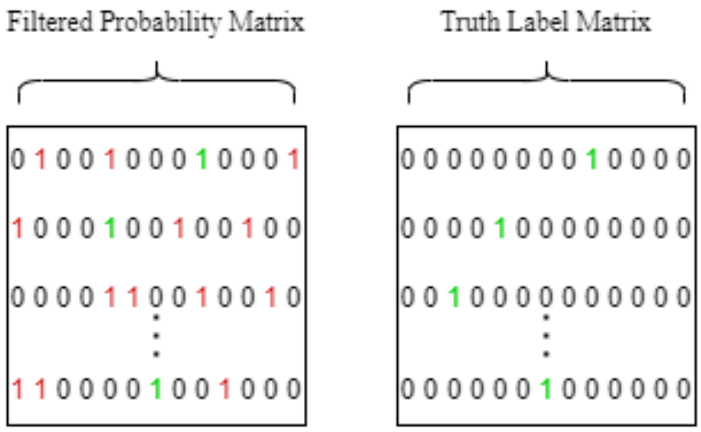


Figure 4.4. *Comparing the Filtered Probability Matrix and the truth label matrix*

Although recall as discussed is the most reliable metric for our model, we still calculated the precision and f1 score. Which were as expected highly distorted by the deliberate large number of false positives. The next section will present the results we extracted from the various datasets we tested the model with. We should mention that we also performed two more experiments. In the first one, we replaced side effects with their corresponding disease classes, for example, "color blindness" will fall under the "nervous system disease" category. This experiment was used to check if the model would be able to generalize better with broader categories. The other experiment we chose to perform, is to use the CID identification for both drugs and side effects, to derive their embeddings. Will discuss both experiments further in the next section which will be accompanied by the results.

Chapter 5

Results

As discussed briefly in previous sections, the fundamental test for our model was composed of the following. For the model's training, 3000 distinct drug pairs and 150 side effects were used. Each side effect in the training set is linked with 20 unique drug pairs. Similarly for the test set 1000 drug pairs and 50 side effects were used. The data comprising every dataset were randomly selected and ensured that no drug combination was duplicated. For all datasets, every metric was calculated for six different number combinations of drug pairs and side effects. The combinations were $(100, 5), (200, 10), (400, 20), (600, 30), (800, 40), (1000, 50)$. Taking into account the effect the embedding model might have on the results, the experiment was repeated using three different models: BioBert, BioClinicalBert, and SciBert. Below are the results of the different datasets corresponding to the different word embedding models.

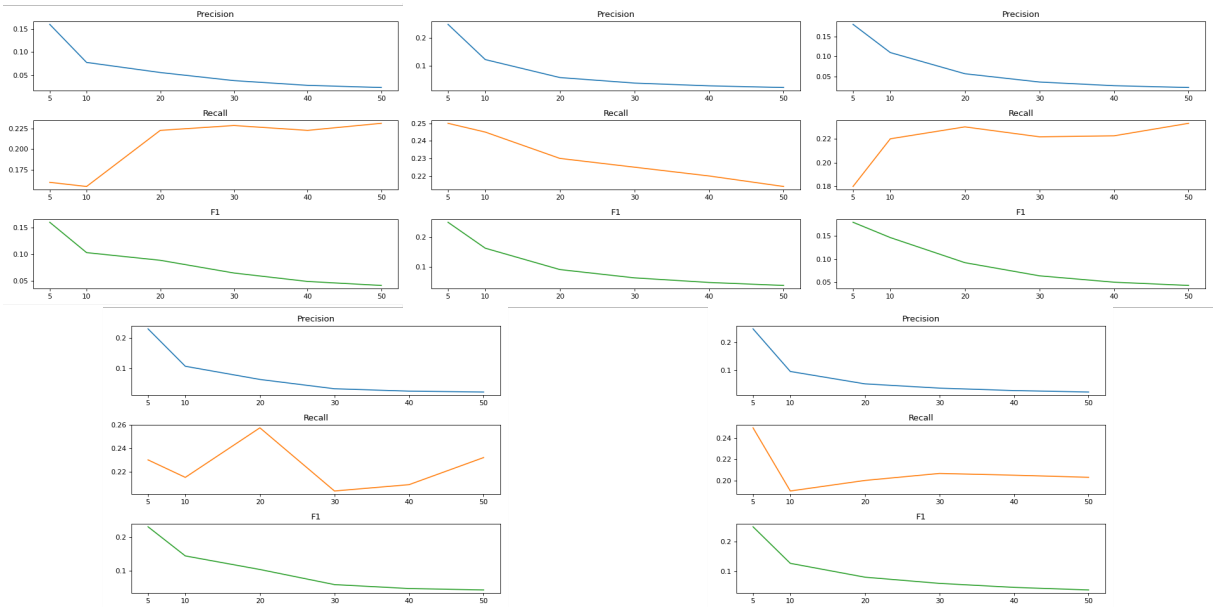


Figure 5.1. *Results from the different datasets using BioBert*

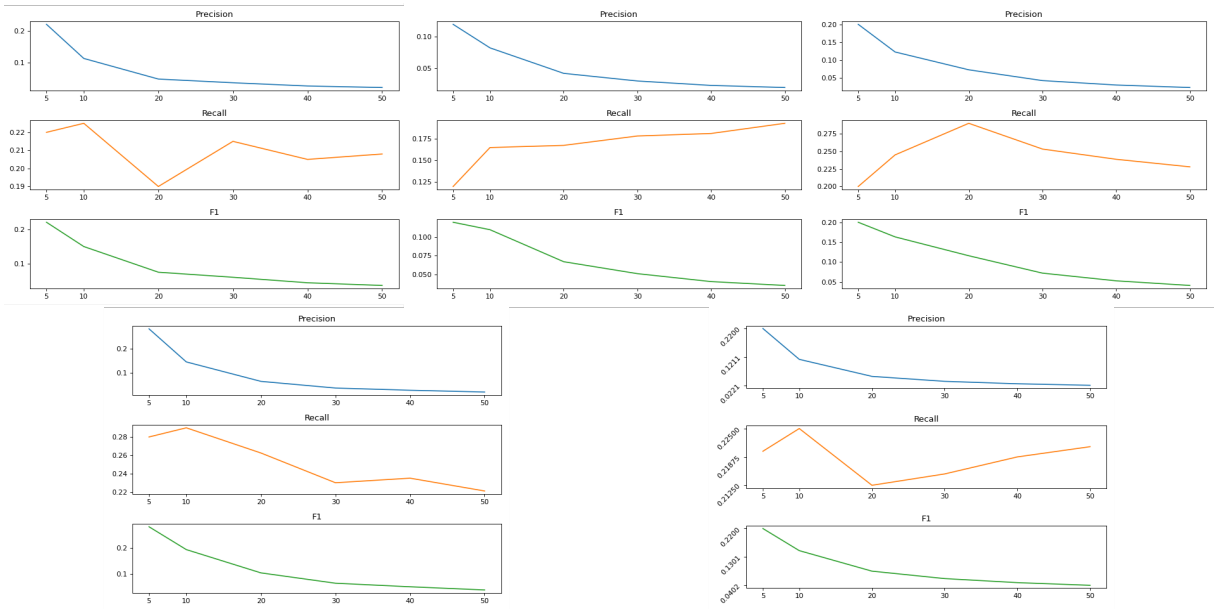


Figure 5.2. *Results from the different datasets using BioClinicalBert*

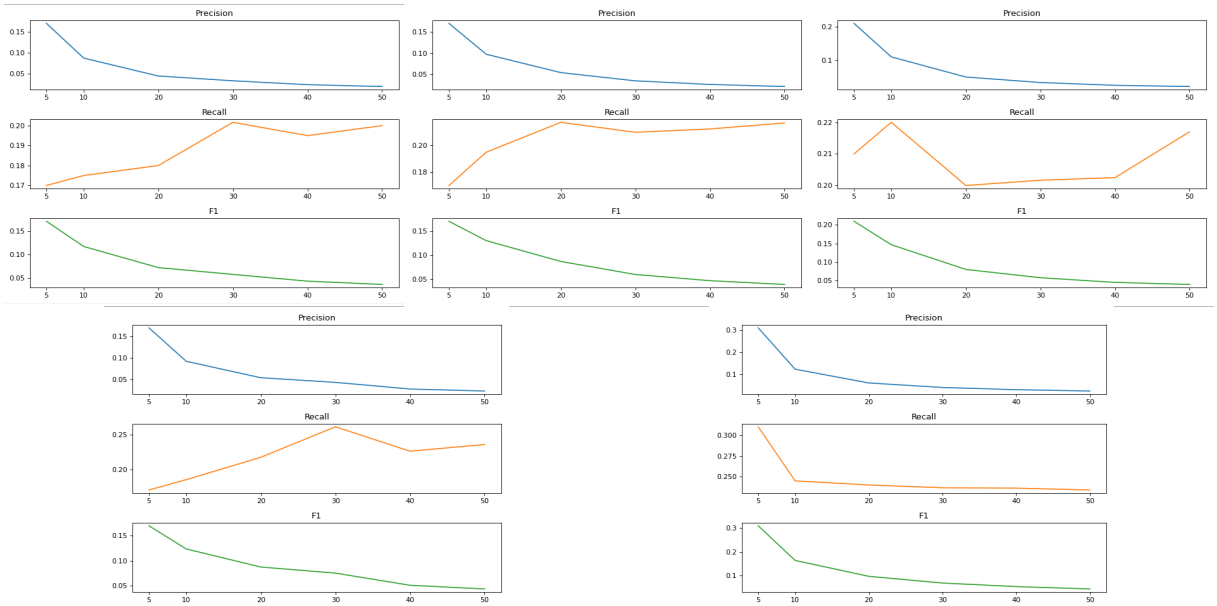


Figure 5.3. *Results from the different datasets using SciBert*

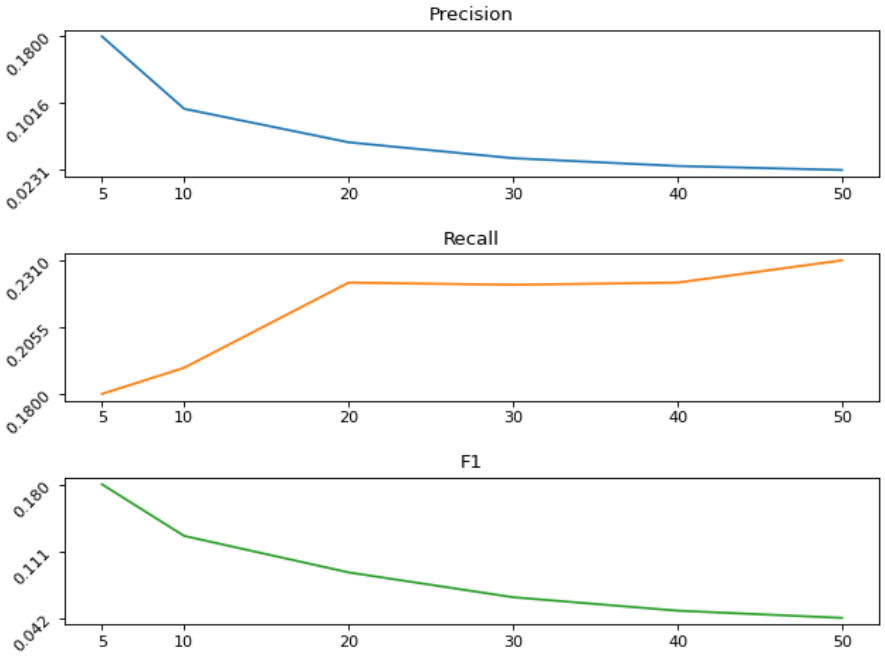


Figure 5.4. *Calculated median from the BioBert results*

Nº Side Effects	Precision	Recall	F1
5	0.18	0.18	0.18
10	0.095	0.19	0.126
20	0.055	0.222	0.089
30	0.036	0.221	0.063
40	0.027	0.222	0.049
50	0.023	0.231	0.042

Table 5.1. *Median accuracy values for the BioBert results*

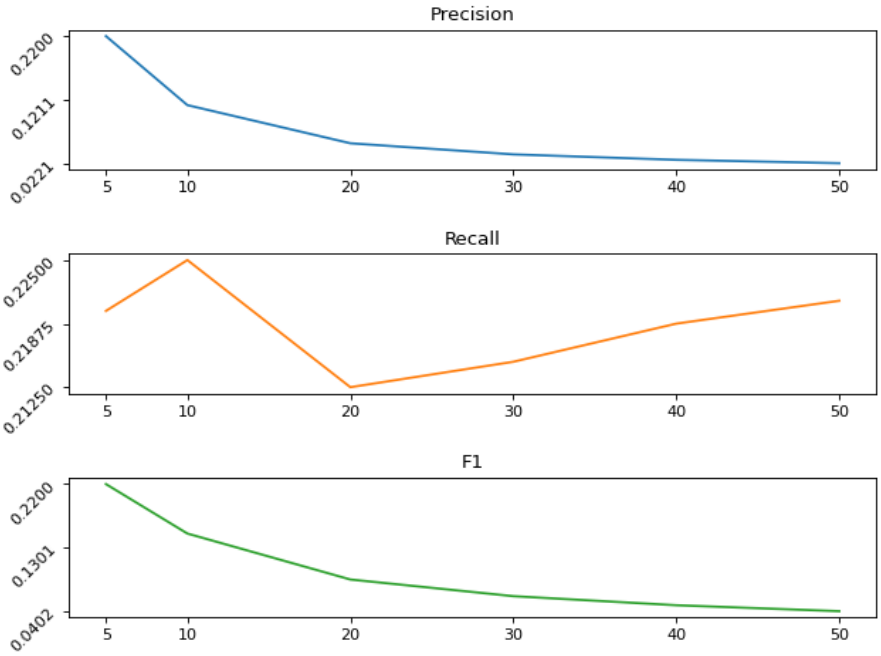


Figure 5.5. *Calculated median from the BioClinicalBert results*

Nº Side Effects	Precision	Recall	F1
5	0.22	0.22	0.22
10	0.112	0.225	0.15
20	0.053	0.212	0.084
30	0.035	0.215	0.061
40	0.027	0.218	0.048
50	0.0221	0.221	0.040

Table 5.2. *Median accuracy values for the BioClinicalBert results*

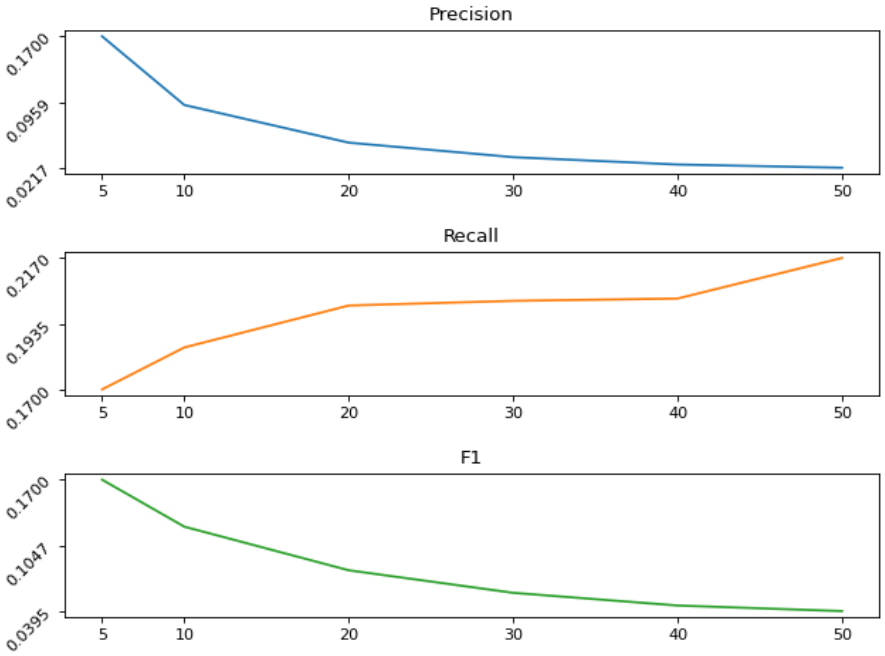


Figure 5.6. *Calculated median from the SciBert results*

Nº Side Effects	Precision	Recall	F1
5	0.17	0.17	0.17
10	0.092	0.185	0.123
20	0.05	0.2	0.08
30	0.036	0.201	0.057
40	0.025	0.202	0.045
50	0.021	0.217	0.039

Table 5.3. *Median accuracy values for the SciBert results*

Let's analyze the results presented above before delving into discussing other configurations. One immediate observation in most of the results is a slight fluctuation in the recall percentage among the first three numbers of available side effects. This is followed by a generally steady but minor increase or decrease in percentage afterward. In general, the recall percentage stays at around 20% but both precision and the f1 score show a consistent decline as the number of false positives grows exponentially. It's worth noting that while BioBert appears to have the best results when all 50 side effects are available, one could argue that the BioClinicalBert results exhibit more consistency, with only minor variation.

It's essential to remember that in this particular setup, the training data significantly outnumbered the testing data. Consequently, the model was trained with a larger number of side effects than it was intended to predict, which can impact its ability to generalize. To gain further insights, it would be beneficial to conduct the next experiment with a reduced number of side effects and assess whether this adjustment improves the results or not. The V matrix was trained using three different combinations of available side effects: 13, 25, and 50. The testing configuration remained unchanged, with the maximum allowable number of side effects still set at 50. Below, the results for the three different configurations on the first dataset are depicted.

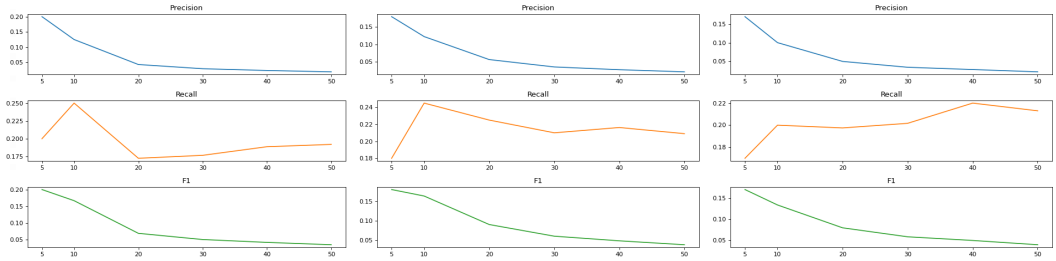


Figure 5.7. *Using 13 Side Effects for Training the V Matrix, Results from Three BERT Models (BioBert, BioClinicalBert, SciBert)*

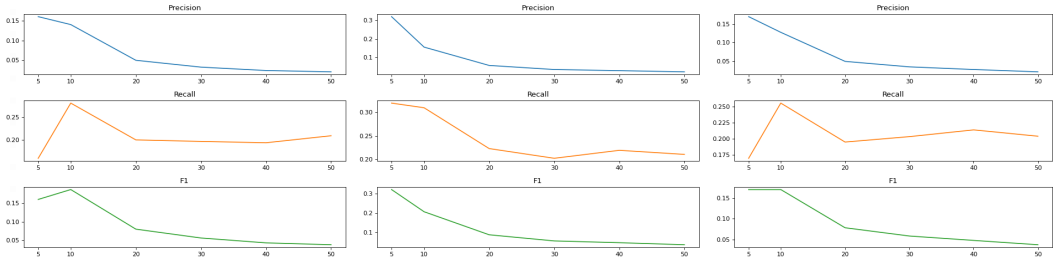


Figure 5.8. *Using 25 Side Effects for Training the V Matrix, Results from Three BERT Models (BioBert, BioClinicalBert, SciBert)*

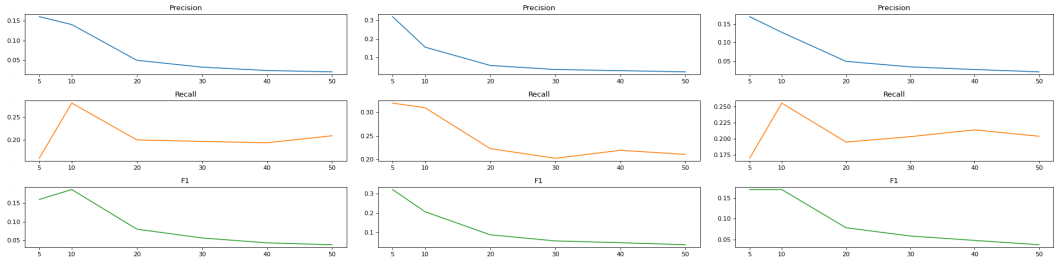


Figure 5.9. *Using 50 Side Effects for Training the V Matrix, Results from Three BERT Models (BioBert, BioClinicalBert, SciBert)*

Our initial observation reveals that the results derived from models utilizing BioBert and SciBert in their processes do not display any significant differentiation from the previous experiments. Nevertheless, the model that incorporated BioClinicalBert into its process yielded some intriguing results. It's helpful to categorize our findings based on the amount of side effects used in the training stage, making two distinctions: one for a small quantity of side effects (13) and another for a larger quantity (25 and 50). In the first case, there was no considerable difference, compared to the results of the initial experiment. However, in the second case, it revealed a significant 15% increase in recall when making predictions between a smaller number of classes (5 and 10). One possible explanation for this increase could be that the reduction in training side effects led to a simplification of feature boundaries within the feature space, thereby aiding the model in more effectively mapping the semantic relationships between drug pairs and side effects. While the exact cause of the model's increased recall remains somewhat unclear, it could also be attributed to having fewer classes to differentiate, allowing it to make clearer distinctions.

One final experiment was conducted to evaluate whether training the model with broader classes would improve its test phase accuracy. By 'boarded classes,' we refer to the disease class that encompasses all associated side effects.

SIDE EFFECT NAME	DISEASE CLASS
color blindness	nervous system disease
refraction disorder	nervous system disease
corneal ulcer	nervous system disease

Table 5.4. *Sample of Side Effects within the Same Disease Class*

For the training dataset, we replaced the side effect names with their corresponding disease classes. The reasoning behind this choice was that the use of more general classes, would provide the model with broader margins and enhance its ability to generalize to unseen classes during the testing phase. This experiment was run on a reduced scale, focusing solely on one of the datasets and utilizing the BioClinicalBert model for embeddings. The results can be seen below in Figure 5.10.

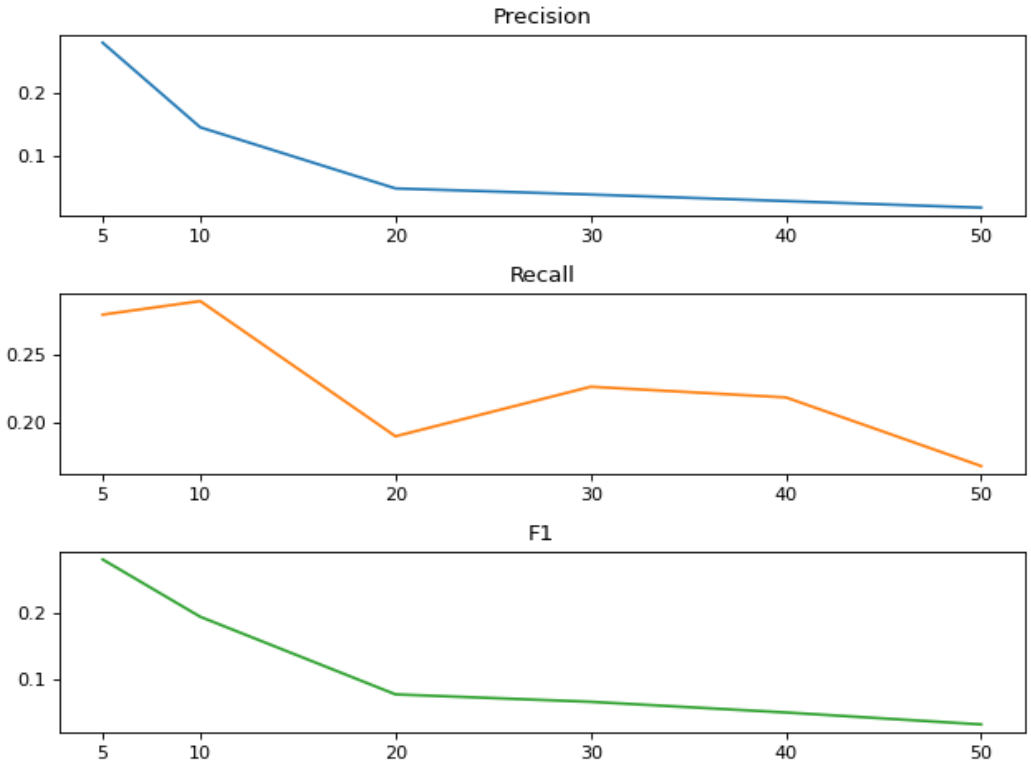


Figure 5.10. *Using broader classes for Training the V Matrix, Results from BioClinicalBert*

The results resemble those of the previous experiment. We observe a notable increase in prediction accuracy with a smaller number of available classes, suggesting that this type of generalization works more effectively when there are fewer classes, making it easier to achieve distinctions. However, as the number of classes increases, it becomes evident that capturing the correct semantic meaning requires more explicit relationships.

Chapter 6

Future Work and Extensions

The experiments above gave us a greater understanding of our model's strengths and weaknesses. From them, we were able to recognize some of our design and process flaws. We need to discuss further our resources, methodology, and, ultimately, our decisions.

6.1 Discussion

Feature Extraction

One pivotal part of our process is the extraction of features from our data. To achieve this we opted to use pre-trained word embedding models. Although the models we chose were trained on a vast amount of medical data, they were trained for general purposes and not the particular task we utilized them. It is clear from the previous in-depth explanation of the zero-shot learning framework that we adopt. The quality of features we derive from our data plays a vital role in the success of our approach. In part, we can attribute some of our design's shortcomings to this. As discussed many times before, the strength of the correlation between drug pairs and side effects greatly depends on those characteristics.

Validating Our Predictions

Inherently the type of problem we choose to take on leads to difficulties when it comes time to validate the guesses made by our model. Even if a prediction isn't present in the recent clinical data, it could be encountered and confirmed at a later point. Leading to an uncertainty of what can be labeled as correct at the time the prediction is made. Simultaneously the fact that we only accept one prediction as truly positive, for every drug combination hinders our ability to use most accuracy metrics. At the same time, the multi-label classification ability that we have added to our structure can not be reliably tested, as a subsequent result of the previous choice.

6.2 Feature Work

Feature Extraction

As underlined above feature extraction and their fineness, are essential for our approach's success. The use of pre-trained models although reasonable, in reality did not yield the best results. For that reason in feature work, the development of a word embedding model solely focused on the problem at hand would be of great benefit [23]. The constructed model could focus on capturing the biochemical properties that are hidden behind every DDI. It could focus on how properties found at the molecular level of each substance, may assist in inducing a specific side effect. An embedding model that could capture and generate a vector, mapping these characteristics, would considerably increase the certainty and accuracy of the predictions made by our architecture.

Validating Our Predictions

Our dataset extraction process could be modified so our testing set, included drug pairs exhibiting more than one confirmed side effect, belonging to the new assortment of classes. This modification would introduce a greater diversity of side effect patterns into our testing set, allowing our model to better generalize and identify DDIs involving multiple side effects. By incorporating drug pairs from the newly defined assortment of classes, we can expand the scope of our model's capabilities, potentially leading to more accurate predictions of complex DDIs.

Bibliography

[1] Ben Snyder, Thomas M Polasek και Matthew P Doogue. *Drug interactions: principles and practice*. *Australian Prescriber*, 35(3):85–88, 2012.

[2] Ke Han, Peigang Cao, Yu Wang, Fang Xie, Jiaqi Ma, Mengyao Yu, Jianchun Wang, Yaoqun Xu, Yu Zhang και Jie Wan. *A Review of Approaches for Predicting Drug–Drug Interactions Based on Machine Learning*. *Frontiers in Pharmacology*, 12:814858, 2022.

[3] Thanh Hoa Vo, Ngan Thi Kim Nguyen και Nguyen Quoc Khanh Le. *Improved prediction of drug-drug interactions using ensemble deep neural networks*. *Medicine in Drug Discovery*, 17:100149, 2023.

[4] Thanh Hoa Vo, Ngan Thi Kim Nguyen, Quang Hien Kha και Nguyen Quoc Khanh Le. *On the road to explainable AI in drug-drug interactions prediction: A systematic review*. *Computational and Structural Biotechnology Journal*, 20:2112–2123, 2022.

[5] Santiago Vilar, Eugenio Uriarte, Lourdes Santana, Nicholas P. Tatonetti και Carol Friedman. *Detection of Drug-Drug Interactions by Modeling Interaction Profile Fingerprints*. *PLoS ONE*, 8(3):e58321, 2013.

[6] Andrej Kastrin, Polonca Ferk και Brane Leskošek. *Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning*. *PLOS ONE*, 13(5):e0196865, 2018.

[7] Yongqin Xian, Bernt Schiele και Zeynep Akata. *Zero-Shot Learning – The Good, the Bad and the Ugly*, 2017.

[8] Bernardino Romera-Paredes και Philip H. S. Torr. *An Embarrassingly Simple Approach to Zero-Shot Learning*. *Visual Attributes*, σελίδες 11–30. Springer International Publishing, 2017.

[9] Jacob Devlin, Ming Wei Chang, Kenton Lee και Kristina N. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018.

[10] Stuart Russell και Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rdη έκδοση, 2009.

[11] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[12] Richard S. Sutton και Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, secondη έκδοση, 2018.

[13] I. Goodfellow, Y. Bengio και A. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016.

[14] *Word embeddings in NLP: A Complete Guide*.

[15] Grigorios Tsoumakas και Ioannis Katakis. *Multi-Label Classification: An Overview*. *International Journal of Data Warehousing and Mining*, 3:1–13, 2009.

[16] Muhammad Atif Tahir, Josef Kittler και Ahmed Bouridane. *Multilabel classification using heterogeneous ensemble of multi-label classifiers*. *Pattern Recognition Letters*, 33(5):513–523, 2012.

[17] Naseer Ahmed Sajid, Atta Rahman, Munir Ahmad, Dhiaa Musleh, Mohammed Imran Basheer Ahmed, Reem Alassaf, Sghaier Chabani, Mohammed Salih Ahmed, Asiya Abdus Salam και Dania AlKhulaifi. *Single vs. Multi-Label: The Issues, Challenges and Insights of Contemporary Classification Schemes*. *Applied Sciences*, 13(11), 2023.

[18] D. Levêque, J. Lemachatti, Y. Nivoix, P. Coliat, R. Santucci, G. Ubeaud-Séquier, L. Beretz και S. Vinzio. *Mécanismes des interactions médicamenteuses d'origine pharmacocinétique*. *La Revue de Médecine Interne*, 31(2):170–179, 2010.

[19] Cara Tannenbaum και Nancy L Sheehan. *Understanding and preventing drug–drug and drug–gene interactions*. *Expert Review of Clinical Pharmacology*, 7(4):533–544, 2014.

[20] Yue Hua Feng, Shao Wu Zhang και Jian Yu Shi. *DPDDI: a deep predictor for drug-drug interactions*. *BMC Bioinformatics*, 21(1), 2020.

[21] Reem Al-Otaibi, Peter Flach και Meelis Kull. *Multi-label Classification: A Comparative Study on Threshold Selection Methods*. 2014.

[22] Marinka Zitnik, Monica Agrawal και Jure Leskovec. *Modeling polypharmacy side effects with graph convolutional networks*. *Bioinformatics*, 34(13):457–466, 2018.

[23] Li Zhang, Tao Xiang και Shaogang Gong. *Learning a Deep Embedding Model for Zero-Shot Learning*. *CoRR*, abs/1611.05088, 2016.

List of Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CID	Compound Identifier
DDI	Drug-to-Drug Interaction
FP	False Positive
ML	Machine Learning
NLP	Natural Language Processing
SVM	Support Vector Machine
TP	True Positive
ZSL	Zero Shot Learning