Journal of
Biomedical Semantics

RESEARCH

Open Access

# MeSH2Matrix: combining MeSH keywords and machine learning for biomedical relation classification based on PubMed

Houcemeddine Turki[1*†], Bonaventure F. P. Dossou[2,3†], Chris Chinenye Emezue[2,4†], Abraham Toluwase Owodunni[5†], Mohamed Ali Hadj Taieb[1], Mohamed Ben Aouicha[1], Hanen Ben Hassen[6] and Afif Masmoudi[6]

## Abstract

Biomedical relation classification has been significantly improved by the application of advanced machine learning techniques on the raw texts of scholarly publications. Despite this improvement, the reliance on large chunks of raw text makes these algorithms suffer in terms of generalization, precision, and reliability. The use of the distinctive characteristics of bibliographic metadata can prove effective in achieving better performance for this challenging task. In this research paper, we introduce an approach for biomedical relation classification using the qualifiers of co-occurring Medical Subject Headings (MeSH). First of all, we introduce *MeSH2Matrix*, our dataset consisting of 46,469 biomedical relations curated from PubMed publications using our approach. Our dataset includes a matrix that maps associations between the qualifiers of subject MeSH keywords and those of object MeSH keywords. It also specifies the corresponding Wikidata relation type and the superclass of semantic relations for each relation. Using *MeSH2Matrix*, we build and train three machine learning models (Support Vector Machine [*SVM*], a dense model [*D-Model*], and a convolutional neural network [*C-Net*]) to evaluate the efficiency of our approach for biomedical relation classification. Our best model achieves an accuracy of 70.78% for 195 classes and 83.09% for five superclasses. Finally, we provide confusion matrix and extensive feature analyses to better examine the relationship between the MeSH qualifiers and the biomedical relations being classified. Our results will hopefully shed light on developing better algorithms for biomedical ontology classification based on the MeSH keywords of PubMed publications. For reproducibility purposes, *MeSH2Matrix*, as well as all our source codes, are made publicly accessible at https://github.com/SisonkeBiotik-Africa/MeSH2Matrix.

**Keywords** Biomedical relation classification, MeSH keywords, PubMed records, MeSH qualifiers, Machine learning, Integrated gradients, Feature analysis

[†] Houcemeddine Turki, Bonaventure F. P. Dossou, Chris Chinenye Emezue and Abraham Toluwase Owodunni contributed equally to this work.

*Correspondence:
Houcemeddine Turki
turkiabdelwaheb@hotmail.fr
Full list of author information is available at the end of the article

## Introduction

Biomedical relation classification involves identifying and categorizing relationships between biological entities, such as genes, proteins, diseases, and drugs, within biomedical texts. This task is crucial for advancing medical research and has diverse applications [1, 2], as it helps in organizing and extracting valuable insights from the vast and growing body of biomedical literature. Biomedical relation classification has seen significant improvement from the application of advanced machine learning techniques on free-form text from large collections of scholarly publications. Despite this improvement, the reliance on large chunks of raw text makes these algorithms suffer in terms of generalization, precision, and reliability, due to the complexity and variability of natural language [3, 4]. Instead of relying solely on text, our work tackles biomedical relation classification by leveraging an often overlooked aspect of scholarly publications – their bibliographic metadata. It is standard practice for scholarly publications, especially in the biomedical domain, to have a bibliographic metadata section where they provide metadata about the scholarly work. This metadata is machine-readable, and provides simple (as opposed to complex free-form text), concise information about the publication. One such important information provided is the *Medical Subject Headings (MeSH) keywords*, which is represented as a biomedical ontology.

An ontology is made up of two concepts related to each other and is represented using statements in the form of triples: *Subject* (Concept), *Predicate* (Relation Type – encoding the nature of the relationship between the two concepts), and *Object* (Concept) [5]. As a result, biomedical ontologies can be easily enriched, processed, validated, and reused by machines [6]. These ontologies are typically manually curated through human experts (e.g., physicians and ontologists), consortiums (e.g., *Open Biomedical Ontologies Consortium*), and institutions (e.g. *NIH National Center for Biomedical Ontology*) [7].

*Medical Subject Headings* (MeSH) keywords used to annotate PubMed scholarly publications can be a very useful resource for biomedical relation extraction and classification [8]. Based on MeSH taxonomy, it always assigns the same concept to various publications in PubMed using the same term [8]. Consequently, the use of MeSH keywords as input for biomedical ontology engineering can be more efficient than the use of user-generated bibliometric metadata and raw texts of scholarly publications [8]. In this research paper, we propose an approach for biomedical relation classification using the associations of the MeSH keywords in PubMed records. The main aim of the work is to verify whether the associations of the attributes of the subject and object MeSH keywords can contribute to the classification of the analyzed biomedical relations or not. For this, we will use the biomedical relations between MeSH terms as revealed by Wikidata, an open and collaborative knowledge graph available at https://www.wikidata.org [9], to construct the training dataset named MeSH2Matrix for biomedical relation classification using our method. Our dataset provides a matrix of the association between the subject MeSH qualifiers and the object MeSH qualifiers, a relation type, and a superclass for every relation. Then, we will train three machine-learning models on this dataset: a support vector machine (*SVM*), a dense model (*D-Model*), and a convolutional neural network (*C-Net*). This paper builds upon preliminary work on the generation of this large-scale dataset presented at BIR@ECIR 2022 Workshop [10]. The significant additional contribution is the in-depth feature analysis of the C-Net and D-Model models to evaluate how the MeSH qualifier features impact the model's class predictions (using the 5-class prediction case). This feature analysis is elaborately outlined in "Insights from feature analyses" section.

We will begin by giving an overview of MeSH Keywords and how they have contributed so far to enhancing tasks in Biomedical Informatics ("MeSH keywords as a valuable input" section). Then, we will describe Wikidata as a biomedical knowledge resource and explain how it can be used to extract biomedical relations between the MeSH terms ("Wikidata as a biomedical semantic resource" section). We will also provide a literature review of the state-of-the-art of biomedical relation classification ("Biomedical relation classification" section). As well, we will describe the prior works about biomedical relation classification using PubMed to prove the originality of our methods ("Prior works on biomedical relation classification using PubMed" section). After that, we will explain our proposed approach for the development and evaluation of biomedical relation classification based on the MeSH keywords of PubMed scholarly publications ("Methodology" section). Later, we will provide a description of the MeSH2Matrix dataset and assess its quality by comparing its main features to previous research findings ("MeSH2Matrix dataset" section). Subsequently, we will outline our results for the biomedical relation classification using accuracy rates and confusion analysis based on *MeSH2Matrix* ("Experimental results on biomedical relation classification using MeSH2Matrix" section). Afterwards, we will study the significance of the features of the matrices to machine-learning models using Integrated Gradients ("Insights from feature analyses" section). Finally, we will conclude our experiments and provide future directions for our research paper ("Conclusion" section).

**Table 1** Examples of relation types corresponding to the associations of two MeSH Keywords. MeSH Qualifiers contributing the relation type are indicated in bold

| MeSH keyword 1 | MeSH keyword 2 | Relation type |
|---|---|---|
| Sofosbuvir/**therapeutic use** | Hepatitis C/**drug therapy** | Medical Condition Treated |
| Asthma/**complications** | Dyspnea/**etiology** | Symptom |
| Retinopathy/**prevention & control** | Diabetes Mellitus/**complications** | Risk Factor |

## Related works

### MeSH keywords as a valuable input

As a controlled vocabulary, MeSH supports sixteen types of biomedical concepts ranging from anatomical structures to symptoms, diseases, and drugs[1] [11]. This broad coverage of MeSH Terms makes them useful to represent the topics of all scholarly publications [11]. The value of these terms is further increased by the regular revision of MeSH to include new concepts such as COVID-19[2] and SARS-CoV-2[3] and to cover updates in the biomedical nomenclature [12]. That is why it has been used to annotate PubMed records for years by the human curators of the bibliographic database to enable consistent indexing of scholarly papers and intuitive data mining [11, 13]. Beyond its contribution to the enhanced topic granularity in PubMed, MeSH keywords have the ability to represent facets of a given topic through the use of predefined subheadings known as MeSH qualifiers providing more precision to the MeSH keywords of the PubMed records [4, 8]. Currently, there are 89 MeSH qualifiers[4]. representing all the characteristics and features of a biomedical entity as revealed at https://www.nlm.nih.gov/mesh/subhierarchy.html (List at Table S1).

These interesting features of the MeSH keywords encouraged its usage beyond information seeking. MeSH keywords are gaining an increasing popularity in biomedical information retrieval as they allow better accuracy for relation extraction and classification from PubMed [13, 14]. This is due to the restriction of the considered publications to the ones that are most likely to include the required information [14] and the analysis of the co-occurrences of the MeSH Keywords [13]. The better output of clinical knowledge engineering driven by the MeSH terms of the PubMed publications has proven the value of MeSH Keywords, especially when assigned controlled qualifiers, to classify biomedical relations [4, 8]. Essentially, the qualifiers of two co-occurring MeSH keywords can inform us of the nature of the semantic relation between them as shown in Table 1 and Fig. 1. This is particularly motivated by the fact that biomedical publications usually have narrow research scope and do not consequently study multiple and unrelated facets of a given topic unless the publication type is an encyclopedic review [15]. The qualifiers of the MeSH keywords can be easily found as they are simply separated from their corresponding headings using a slash (/). The MeSH keywords of PubMed scholarly publications can be retrieved from the NCBI Entrez API using the Biopython Python Library [16, 17]. The structured format of MeSH keywords and their simple retrieval from the PubMed bibliographic database encourage their usage for the biomedical relation extraction and classification.

### Wikidata as a biomedical semantic resource

Wikidata was created in October 2012 as a knowledge database to support structured data in Wikipedia such as interlanguage links and infoboxes [18]. However, it has grown over the past years to become one of the largest free and open knowledge graphs covering various range of fields, particularly biomedicine [9]. Its collaborative and crowdsourcing-based enrichment, regular updates according to recent advances in the major areas of interest, etc make it the most adequate knowledge base to support ever-changing scholarly evidences, mainly in the context of the COVID-19 pandemic [19]. Currently, Wikidata represents various types of medical entities as items, such as drugs, diseases, genes, proteins, organs, and symptoms [9]. These items are linked to their equivalent entities in external knowledge resources, mainly MeSH [19]. Every entity is assigned a language-independent identifier (so-called *Q-number*) as well as its main names (*labels*), glosses (*descriptions*), and alternative names (*aliases*) in a variety of natural languages, particularly English as shown in Figure D1 [9, 19]. Furthermore, entities are related to other ones using semantic relations (*Subject - Predicate - Object*) where the relation type (*Predicate*) is also a Wikidata entity having its own identifier (*P-number*) and semantic description [9, 20].

[1] https://www.nlm.nih.gov/bsd/disted/meshtutorial/meshtreestructures/index.html.

[2] https://www.ncbi.nlm.nih.gov/mesh/2052179.

[3] https://www.ncbi.nlm.nih.gov/mesh/2052180.

[4] They are only 76 MeSH Qualifiers: *administration & dosage, blood, cerebrospinal fluid, urine, embryology, agonists, antagonists & inhibitors, genetics, immunology, supply & distribution, adverse effects, poisoning*, and *pharmacokinetics* are unintentionally left duplicates as they are included twice in the list of MeSH Qualifiers at https://www.nlm.nih.gov/mesh/subhierarchy.html to highlight multiple contexts of their usage.

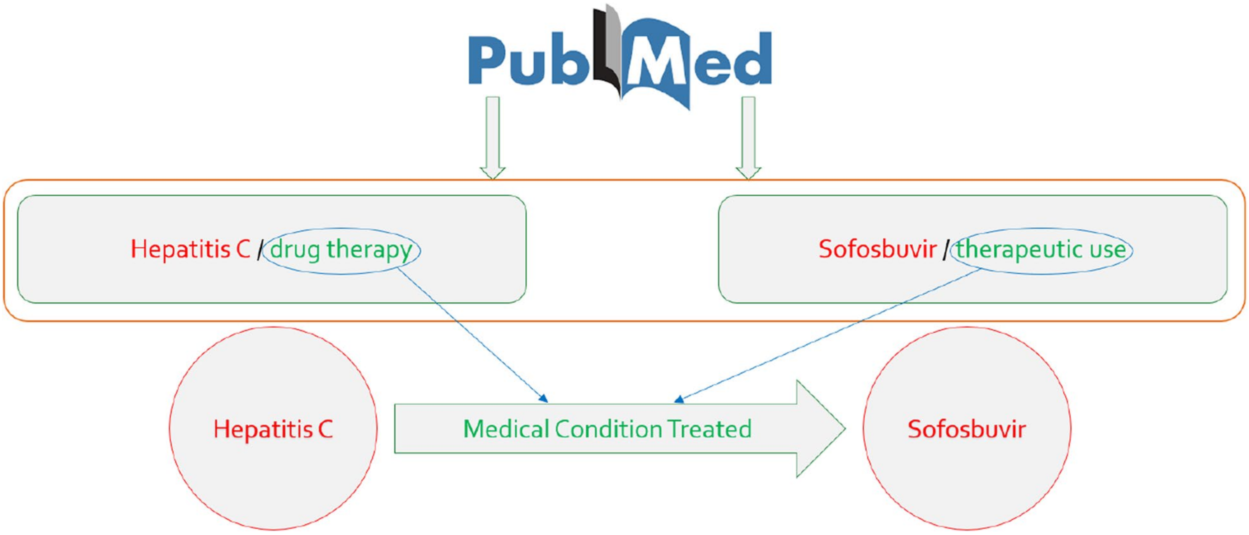

PubMed
Hepatitis C / drug therapy
Sofosbuvir / therapeutic use
Hepatitis C
Medical Condition Treated
Sofosbuvir

**Fig. 1** Principle for converting two complementary MeSH qualifiers into a biomedical relation type

There are two major kinds of relation types: taxonomic or non-taxonomic ones [9, 19]. Taxonomic relations are hierarchical ones that link an entity to its parent class or constituents (e.g., *instance of* [P31] - Figure D2) [9]. Non-taxonomic relation types are non-hierarchical ones that are assigned to define specialized knowledge such as biomedical information (e.g., *signs or symptoms* [P780] - Figure D3) [9]. A non-taxonomic relation can be symmetric: where the subject-object inversion would not affect the meaning of the statement. If this condition is not fulfilled, the relation in non-symmetric [20]. Such a classification of biomedical relation types can be easily inferred from the semantic information about Wikidata relation types, particularly the property constraints providing conditions for the definition of relational statements (Figure D4) [20]. These conditions are not only useful for the validation of semantic information about the semantic relations but also to recognize the distinctive features of relation types. Moreover, Wikidata items are also matched to their equivalents in external resources using non-relational statements in the form of triples where the predicate reveals the aligned resource and the object is the identifier of the concept in the external database [20]. Particularly, *MeSH Descriptor ID* [P486] statements align between MeSH terms and Wikidata items as revealed by Figure D5. As Wikidata knowledge graph is developed using the Resource Description Framework (RDF) format, it can be easily processed to get subsets of semantic information needed to drive knowledge-based applications using a variety of tools, particularly the MediaWiki API[5], the Wikidata SPARQL query service[6], and the Wikibase Integrator Python Library[7].

### Biomedical relation classification

Biomedical relation classification is the assignment of a relation type to a pair of biomedical entities after confirming their semantic relatedness based on natural language processing and information retrieval techniques [3]. Although recent works tried to use pre-trained language models such as *BERT* and *GPT* for biomedical relation classification [21], this gold standard is to create classification algorithms based on two layers; the *Embedding* layer and the *Machine Learning* layer [3]. Embeddings allow converting the environment of the two related concepts into a uni-dimensional (*vector*) [22], bi-dimensional (*matrix*) [23], or tri-dimensional (*box*) [24] array representation [3]. Embeddings provide an abstract representation of the context of the semantic association that can be used to train machine learning models to classify biomedical relations [3]. Machine learning techniques for biomedical relation classification include reinforcement learning [25], neural networks [3], support vector machines [26], and rarely adversarial learning [27] as well as ensemble learning methods, such as *XGBoost* [28]. Most of these works consider a limited number of relation types ($< 10$) for the supervised classification [3, 28]. These relation types mainly cover significant drug and protein interactions, drug adverse effects, drug administration

[5] https://www.wikidata.org/w/api.php.

[6] https://query.wikidata.org.

[7] https://github.com/LeMyst/WikibaseIntegrator.

routes, disease-drug relations, risk factors, and gene expressions [3, 28, 29].

Although several works will try to use *Electronic Health Records* (EHRs) as input for biomedical relation classification [28], most of the previous research has emphasized free online textual resources as inputs for biomedical relation classification algorithms [3, 21]. These resources involve general-purpose databases such as *Wikipedia*, news corpora, and *Common Crawl* [3]. But, the most important resources used in this context are mainly open biomedical bibliographic databases (e.g., *PubMed*) and open-access repositories (e.g., *PubMed Central*) [3, 21]. These resources provide a significant subset of peer-reviewed medical knowledge and can be consequently used to classify biomedical relations with a higher accuracy [30]. In this research work, we continue this trend by developing a matrix-based approach for biomedical relation classification based on the MeSH Keywords of PubMed scholarly publications.

### Prior works on biomedical relation classification using PubMed

In recent years, biomedical relation classification has become a critical aspect of extracting meaningful insights from the growing volume of biomedical literature available in sources such as PubMed. This review explores several key papers that contribute to the field, covering a range of methods and applications. Alimova et al. (2022) [31] address the need for evaluating biomedical relation extraction models on diverse corpora with different entities and relation types. The study compares the performance of BioBERT and LSTM models on four benchmark datasets, including PHAEDRA, i2b2/VA, BC5CDR, and MADE corpora. Notably, their cross-attention LSTM model demonstrates improved cross-domain performance, highlighting the importance of considering diverse domains in biomedical relation classification [31]. Parwez et al. (2023) [32] focus on leveraging neural network-based classification models to extract knowledge from the increasing volume of biomedical texts available in PubMed. The paper introduces an approach to capture both distributional and relational contexts of words to enhance word representation. The proposed method outperforms the GloVe model and exhibits improved accuracy in text classification using learned word representations [32]. Sosa et al. (2023) [33] introduce a novel classification task to associate biological context, such as cell type or tissue information, with protein-protein interactions extracted from text. The study employs syntactic, semantic, and meta-discourse features and introduces the Insider corpora for training classifiers. More specifically, the paper focuses on associating specific biological contexts, such as cell types and tissue information, with protein-protein interactions (PPIs) and gene regulatory interactions. This is achieved by employing a feature-based classifier that analyzes the vast PubMed database to link biomedical context mentioned in the text with extracted PPIs. The paper demonstrates high performance in associating biological contexts with relations. This is a significant advancement, as it allows for a more nuanced understanding of the interactions within biomedical knowledge graphs. By providing context, the study's approach enhances the potential applications of these graphs in various areas of biomedical research and drug discovery. One of the notable aspects of this research is its practical application. The paper includes a case study on dengue fever, showcasing how the methodology can be used to enrich knowledge bases in specific research areas. This case study illustrates the potential impact of the approach on drug discovery and pharmacological inference, highlighting its relevance to current biomedical challenges. In summary, this paper presents a groundbreaking approach to text mining at a scale previously unattainable, providing a method to associate biological contexts with protein-protein interactions in a way that enriches and enhances the utility of biomedical knowledge graphs. This advancement holds significant promise for various applications in biomedical research, drug discovery, and beyond.

As shown, prior works on biomedical relation classification using PubMed have mainly been based on the processing of the abstracts of scholarly publications using machine learning, language models, and semantic annotation. The use of the MeSH keywords in knowledge engineering was mostly restricted to the retrieval of unclassified biomedical relations using the analysis of co-occurrence and citation links in PubMed [34] and the use of weakly-supervised deep learning techniques to adjust or augment MeSH indexing [35].

## Methodology

As shown in Fig. 2, we start with the creation of the MeSH2Matrix dataset using our novel principle of the matrix of correspondence. Then, we explore leveraging the MeSH2Matrix dataset for biomedical relation classification. For this, we focus on machine learning-based methods (particularly neural networks) and train selected models on the MeSH2Matrix dataset. Finally, we perform extensive feature analysis of the trained neural networks to better understand the efficacy of the representations encoded in the MeSH2Matrix in classifying biomedical relation types.

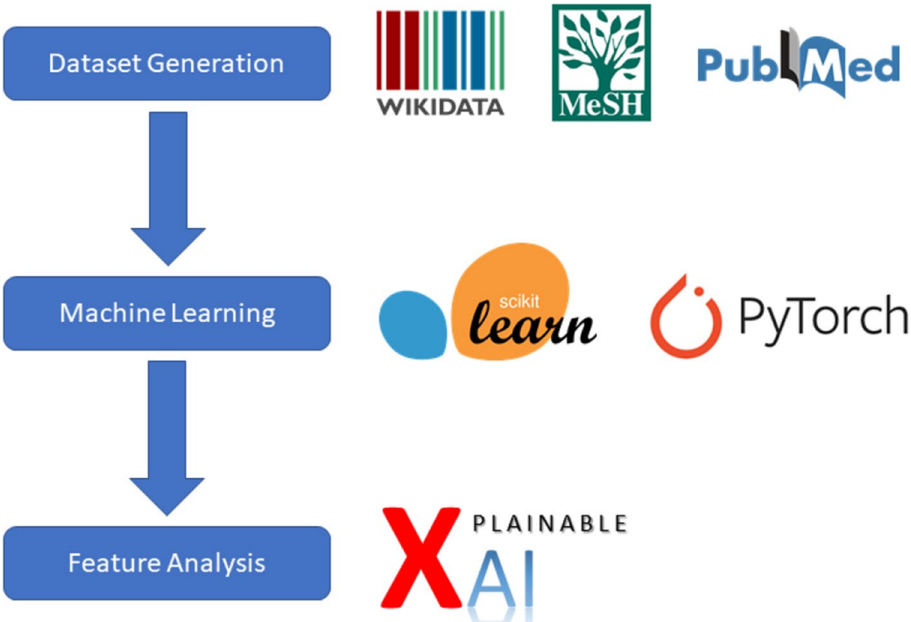


**Fig. 2** Illustration of our methodology. We begin by generating the MeSH2Matrix dataset through our innovative concept of the correspondence matrix. Next we leverage machine learning approaches, particularly neural networks, for the task of biomedical relation classsification using MeSH2Matrix. Lastly, we conduct feature analysis to gain deeper insights into MeSH2Matrix-based classification

## MeSH2Matrix

In this section, we cover the underlying principle of the MeSH2Matrix dataset creation as well as the practical workflow employed.

### *Matrix of correspondence*

We develop our approach upon the assumption that the qualifiers of two co-occurring MeSH terms can outline the type of semantic relation between them, as shown in the examples in Table 1 [4, 8]. Let $t_1$ and $t_2$ be two semantically related MeSH terms that are not assigned a relation type. Our method proposes to first search for the PubMed scholarly publications having both $t_1$ and $t_2$ as MeSH headings. Then for each retrieved record, we extract its qualifiers $q_1$ and $q_2$ (e.g., *therapeutic use* for *Sofosbuvir/therapeutic use*) respectively corresponding to $t_1$ and $t_2$ (e.g., *Sofosbuvir* for *Sofosbuvir/therapeutic use*). This will enable the creation of ($q_1$, $q_2$) pairs as shown in Fig. 3. When a term is assigned two or more qualifiers (e.g., $t_2$ */Z/U* for Paper 3 - Fig. 3), this means that a paper deals with a facet of a characteristic of the considered topic. In such a situation, we consider it as though the qualifiers were independently assigned to the MeSH term for the paper (e.g., $t_2$*/Z* and $t_2$*/U* for Paper 3 - Fig. 3). We restrict the number of considered publications to the 100 most relevant research papers according to PubMed *Best Match* search algorithm [36]. This will prevent matters related to the timeout limit of the NCBI PubMed API (*Error 429*). After the couples of MeSH qualifiers are retrieved, we draw a *matrix of correspondence* ($M(t_1, t_2)$) – this is a square matrix of the qualifiers $(q_1, ..., q_{89})$[8] where each element $m_{i,j}$ is the number of records featuring both $t_1/q_i$ and $t_2/q_j$ as MeSH keywords divided by the total number of records with the two MeSH terms $t_1$ and $t_2$:

$$m_{i,j} = \frac{N(t_1/q_i, t_2/q_j)}{N(t_1, t_2)}, m_{i,j} \in [0, 1]. \quad (1)$$

This *matrix of correspondence* encodes the nature of the semantic relation between $t_1$ and $t_2$. As a practical example, as of March 6, 2022, there are 32 PubMed records where *Hepatitis C* and *Sofosbuvir* are featured together as MeSH headings. From these 32 publications, there are 15 papers where *drug therapy* and *therapeutic use* are the respective qualifiers to *Hepatitis C* and *Sofosbuvir*. In this situation, the value that will be represented for the association between *drug therapy* and *therapeutic use* in the *Hepatitis C-Sofosbuvir* matrix is $15/32 = 0.469$.

### *Building MeSH2Matrix with the matrix of correspondence*

Here we describe the concrete workflow through which we create our MeSH2Matrix dataset. The first step involves extracting biomedical relations from

[8] There are currently 89 pre-defined MeSH qualifiers (76 unique qualifiers and 13 unintentionally repeated ones) as revealed at https://www.nlm.nih.gov/mesh/subhierarchy.html.

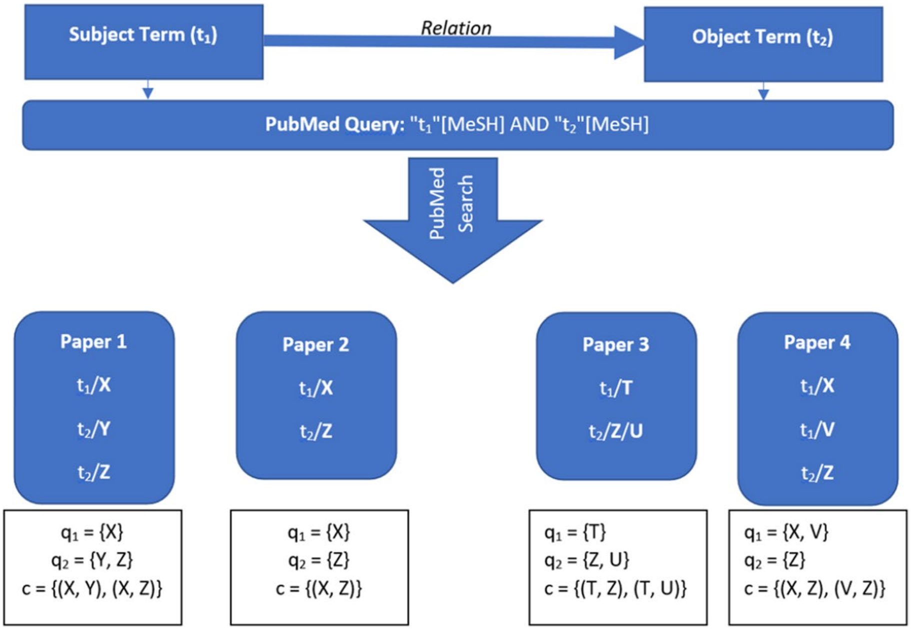


**Fig. 3** Process for the retrieval of the couples of MeSH qualifiers. $t_1$ is the subject MeSH term, $t_2$ is the object MeSH term, $q_1$ are the subject qualifiers, $q_2$ are the object qualifiers, and $c$ is the set of the couples of the extracted MeSH qualifiers

Wikidata with the help of our SPARQL query featured in Figure D6. This is done through identifying all items in Wikidata with a MeSH Descriptor ID (*wdt:P486*), storing these in the *%item* named result set. For each of these items, the query finds any relationships they have with other items. It ensures that these related items also have a MeSH Descriptor ID (*wdt:P486*), storing the identifier value in *?object*. The query then returns the MeSH ID of the original item (*?subject*), the Wikidata property corresponding to the type of relationship (*?reltype*), and the MeSH ID of the related item (*?object*). The result of this extraction is a set $\mathbb{BR} := \{(s_i, r_i, o_i)\}_{i=1}^{N}$ of N tuples $(s_i, r_i, o_i)$ where $s_i$ is a subject term, $o_i$ is an object term and $r_i$ represents their relation type. The output ($\mathbb{BR}$) of the query is saved as a tab-separated values (TSV) file to allow its automatic processing. Recall that the underlying idea of our MeSH2Matrix dataset is the matrix of correspondence $M(s_i, o_i)$ which encodes useful relationships of the subject-object association (using only the qualifiers of $s_i$ and $o_i$). To obtain the matrix $M(s_i, o_i)$ for a given subject term $s_i$ and object term $o_i$ we get their respective qualifiers from PubMed using the NCBI Entrez API. Each term is described in PubMed by a qualifier giving rise to a qualifier couple $(q_a, q_b)$ where $q_a$ is the qualifier assigned to the subject $s_i$ and $q_b$ is the qualifier assigned to the object $o_i$. However, a term association can have many couples of MeSH qualifiers. Based on this, we obtain our matrix of correspondence $M(s_i, o_i)$ through all the extracted qualifier couples respectively for the subject $s_i$ and object term $o_i$. Each matrix is subsequently assigned the relation type $r_i$ corresponding to the subject-object association as a label.

For a better analysis of our proposed approach, we extract the features of the considered Wikidata relation types and verify their names as well as if they are taxonomic, symmetric, or biomedical through the application of SPARQL queries on Wikidata using Wikibase Integrator coupled with human validation.

## Biomedical relation classification with machine learning

This section covers our approach to exploring biomedical relation classification with machine learning. It is important to note that our objective here is to demonstrate the efficacy of machine learning methods, particularly neural networks, and not to obtain the optimal machine learning technique for our task. As a result, this research does not consider many other machine-learning approaches.

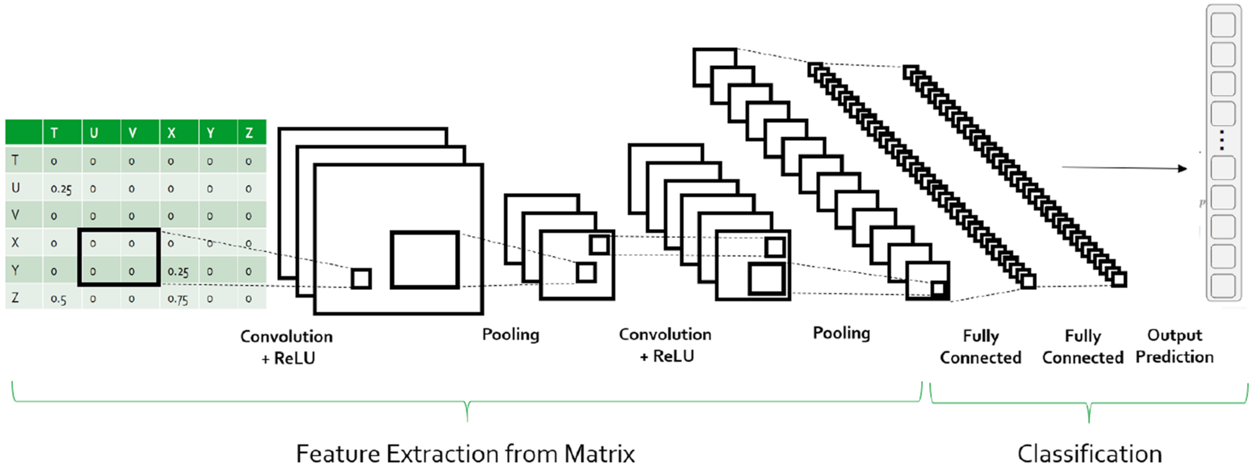
### *Machine learning models explored*

Machine learning-based approaches handle biomedical relation classification as a supervised learning classification task, where labeled data is used to train models. In this paper, we provide benchmark results on our dataset, using three machine learning models:

*SVM:* Support vector machines (SVMs) [37] are best suited for samples with many features because their ability to learn is independent of the features space [38]. They have been used exensively in biomedical classification tasks [39–42] due to their ability to generalize well with data consisting of sparse high-dimensional features. For our baseline, we trained a linear support vector machine. For this, we transformed each 89 × 89 matrix into a single 7921 feature vector.

*D-Model:* Neural networks (NNs) have produced state-of-the-art results in the area of relation classification [27, 42–44]. The major advantage of neural network based approaches lies in thier ability to directly learn the latent feature representation from the labeled training data without requiring experts to carefully craft them [3]. For our experiments with neural networks, we designed *D-Model*, a simple multilayer perceptron with an input layer of output feature size of 3, 960, a hidden layer of 1, 980 and an output layer with an output feature size corresponding to the number of classes [rationale for the choice of the size of neurons: 1). we tested different sizes and this gave the best result, and 2). we followed [45–47] in keeping the hidden layer size between input layer size and output layer size]. ReLU activation function [48] was used between the input and hidden layers to introduce non-linearity. The output layer is connected to a softmax activation function which converts the model's output into a probability over the classes. Although NNs have shown great promise for relation classification, they are highly susceptible to overfitting [49] and require lots of hyperparameter tuning. Therefore, we experimented with regularization techniques (early stopping and dropout) the hyperparameters (learning rate, batch size, etc) in order to produce the best performing *D-Model*.

*C-Net:* Convolutional neural networks (CNNs) are a type of neural networks that can successfully capture the spatial and temporal dependencies in an image through the application of convolution operation and relevant filters. Their potential was first witnessed in computer vision around 2012 [50], and since then have been used extensively even in biomedical relation classification [44, 51, 52]. CNNs perform well on an image dataset better due to the reduction in the number of parameters involved and reusability of their weights - they are therefore best suited for image-type data. Furthermore, with CNNs we can work directly with the 2-dimensional matrix (compared to transforming it for *SVM* and *D-Model*). To explore the impact of CNNs on MeSH2Matrix, we decided to interpret our feature matrix as spatially correlated features and designed *C-Net*, a simple CNN-based architecture made up of four convolution layers (each layer consisting of a 2-dimensional convolution, batch normalization [53], a ReLU activation function [48] and max-pooling) and two fully connected layers (Fig. 4). After passing through the fully connected layers, the final layer uses the softmax activation function which is used to get probabilities of the input matrix being in a particular class. CNN-based models, while being very promising, require practical knowledge to configure the model architecture with regard to the performance [54], and to set the hyperparameters for the best optimization [55]. Similarly, we conducted hyperparameter tuning and optimization in order to explore the reasonable ranges for the sensitive hyperparameters of the classification model.



| | T | U | V | X | Y | Z |
|---|---|---|---|---|---|---|
| T | 0 | 0 | 0 | 0 | 0 | 0 |
| U | 0.25 | 0 | 0 | 0 | 0 | 0 |
| V | 0 | 0 | 0 | 0 | 0 | 0 |
| X | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | 0 | 0 | 0 | 0.25 | 0 | 0 |
| Z | 0.5 | 0 | 0 | 0.75 | 0 | 0 |

**Fig. 4** CNN-based architecture for the classification of the MeSH qualifier-based matrices

### *Experimental setup*

We performed two rounds of classification: one with all relation types (195), and another with 5 categories obtained after grouping the initial 195 relation types (see "MeSH2Matrix dataset" section for more details on grouping). Then, we apply three scenarios of classification to the best-performing model according to the first two rounds to study the effect of various factors on the efficiency of biomedical relation classification based on *MeSH2Matrix*:

- *Scenario 1*: Restriction to the matrices based on 100 scholarly publications or more (Blue in Fig. 5)
- *Scenario 2*: Restriction to the well-represented relation types (10+ matrices)
- *Scenario 3*: Restricted generalization to a unique superclass at once [56]

We split our dataset into training (33, 457 samples), validation (13, 012 samples) - for early stopping, regularization and hyperparameter tuning - and testing (9, 294 samples) - for the final evaluation of the model. For *SVM* training, we merged the training and validation set, making a total of 46, 469 samples for training. For the training of *D-Model* and *C-Net*, we used the Adam optimizer [57]. The code for all our deep learning experiments was written using the PyTorch deep learning framework [58], while for *SVM* we implemented the training using the LinearSVC package.

### *Evaluation metrics*

To assess the efficiency of the three proposed models, we will be based on four basic measures providing insights on the behavior of classification algorithms [59]:

- True Positives (*TP*): the number of items correctly assigned to their respective classes
- True Negatives (*TN*): the number of items correctly not assigned to unrelated classes
- False Positives (*FP*): the number of items mistakenly assigned to unrelated classes
- False Negatives (*FN*): the number of items mistakenly not assigned to their respective classes.

These measures are combined together to provide two main statistical metrics to be used in our study: *Accuracy*, and *F1-Score* [59]. *Accuracy* is defined as the ratio of the number of correct predictions out of the overall number of predictions as clearly revealed in Eq. 2 [59]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

The *F1-Score* combines *Precision* ($\frac{TP}{TP+FP}$) and *Recall* ($\frac{TP}{TP+FN}$) in the following way:

$$F1 = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{3}$$

As sample size per class as well as class imbalance alter the values of the *Accuracy* and the *F1-Score* [60], we additionally consider three metrics that evaluate these two factors for every scenario: *Arithmetic Mean*, *Geometric Mean*, and *GA-Ratio* [61]. Let $x_i$ be the size of the class $i$ and $N$ be the number of classes, the arithmetic mean and geometric mean are defined as:

$$ArithmeticMean = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4}$$

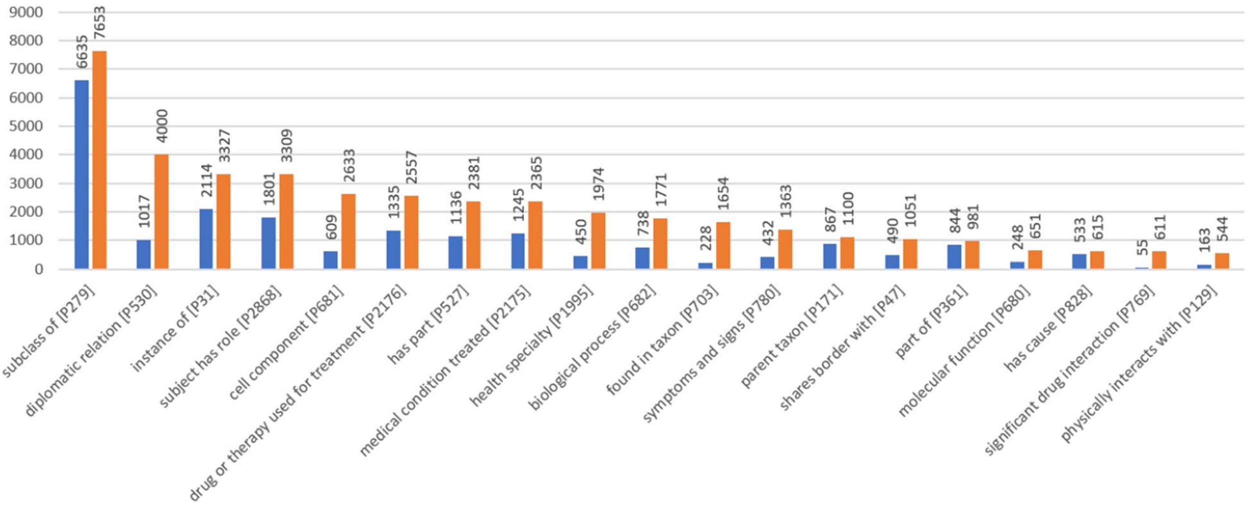


**Fig. 5** Top twenty relation types according to the number of generated matrices: Number of generated matrices (Orange), Number of matrices based on 100+ scholarly publications (Blue)

$$GeometricMean = \sqrt[N]{\prod_{i=1}^{N} x_i} \tag{5}$$

These measures have been shown to efficiently evaluate the sample size per class [61]. The Arithmetic Mean diverges from the Geometric Mean when the class distribution is even [61]. This allows us to consider the complexity of the classification tasks allowing us to judge whether the comparison between different situations of classification is reasonable or not. In this context, the GA-Ratio is the ratio of the arithmetic mean of the size of classes over the geometric means of the size of classes as shown in Eq. 6. It is an efficient metric to evaluate the class distribution of a dataset. As a result of the inequality of arithmetic and geometric means, GA-Ratio always ranges between 0 and 1 where 1 corresponds to an equal class distribution and 0 corresponds to an absolute class imbalance [61]. Here, the Geometric Mean cannot have a value of 0 because every class should at least have one item to exist:

$$GA = \frac{GeometricMean}{ArithmeticMean} \tag{6}$$

To confirm whether the scenario-based evaluation results for the best-performing model apply to the two other models (*SVM* and *C-Net*), we generate confusion matrices for the MeSH2Matrix-driven training of the three machine learning algorithms based on the five superclasses. A confusion matrix is a table that records the associations between the expected classification classes and the predicted ones. Throughout this paper, rows stand for true labels and columns stand for predicted labels. The classes are sorted in the same order in rows and columns for all the generated confusion matrices. As a result, the accurately classified items are featured on the main diagonal line from the top left and bottom right [59].

### Feature analysis

Although machine learning models have gained widespread adoption in recent times, several reservations still exist about how these models could be used and the level of *trust* that should be granted to these models. These reservations do exist because of the black-box nature of these models and this has limited the level of machine-learning model adoption, especially in the medical domain. In their work, [62] demonstrate an opposing non-linear relation that exists between the explainability of a model and the complexity of the model; as a model is trained on a larger amount of data, the complexity of the model increases and the ease of explainability reduces. To make the output of a machine learning model more acceptable in the medical AI domain, more human-based reseasoning needs to be applied [63] in the form of explainability. To enhance the adoption of deep learning models in the medical AI domain, [62] recommended the use of various explainable AI (XAI) techniques hence, the addition of feature analysis and explainability section to this work.

Understanding the predictions of a model using XAI techniques can be broadly categorized into two classes: model-agnostic techniques and model-specific techniques, of which the more popular are the model-agnostic techniques [64]. Due to the nature of our dataset, model-specific feature permutation and ablation have been the simplest strategies for evaluating the feature significance in our models [64]. However, results have shown that the robustness of these techniques is very limited [65]. In addition, although techniques like Shapley Additive Explanations (SHAP) [66] and Local Interpretable Model-agnostic Explanations (LIME) [67] have been seen to be the most widely adopted technique for model explainability, [68] showed that they are less robust and highly prone to adversarial attacks. A more flexible gradient-based technique known as Integrated Gradients (IG) was proposed by [69] and reported to be more robust than earlier mentioned techniques [70], hence our choice for the use of Integrated Gradients as the XAI technique for this work. It is important to note that using the integrated gradients technique for feature importance attribution has its limitations, one of which being that the function learned by the model may be "over-influenced" by just one of the features [71].

Computing the integrated gradients (IG) [69] of neural networks with respect to the input features is a technique that addresses the problem of feature attribution by using a gradient-based approach to satisfy two fundamental axioms: **Sensitivity and Implementation Invariance** that should be satisfied.

**Sensitivity** here implies that a non-zero feature attribution value should be assigned to a feature if it is the only feature differing between a baseline input and input data sample with different predictions. On the other hand, **Implementation Invariance** implies that two neural networks with different architectures are functionally equal if they have equal outputs for each input data sample.

The relevance of integrated gradients, over other techniques, for the explainability of neural networks as highlighted by [69] includes the fact that the use of IG does not require any modification to the original neural network. It can also be used to extract rules from the model and as a tool for debugging deep learning models [69].

Integrated Gradients can be mathematically represented as:

$$IntegratedGradient_i(x) ::= (x_i - x_i') \times \int_0^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

Where $x_i$ and $x_i'$ are the is the input data sample and the baseline input along the $i^{th}$ dimension. In order to understand how each feature in our training dataset contributes to the overall decision-making of our model, we employed integrated gradient as an XAI technique to explain our best two performing models, *D-Model* and *C-Net* at the superclass level. Using IG for this helps us to be able to equally compare a CNN model to an ANN model as opposed to other techniques that would only see a CNN model input to be an image, which is not the case for our dataset. The limitation in the robustness of the permutation explainer in SHAP to not being able to handle datasets with a feature size greater than 225 also informed our choice of integrated gradients.

To compute the integrated gradient for the *D-Model* we set $x_i'$ to be a zero-vector with the same shape as $x_i$ and $\delta F$ was computed at little intervals through moving from $x_i$ (non-zeros) to $x_i'$ (zeros) and then adding the value of $F$ multiplied by the interval size. The computation was done for 8080 test samples and the results are shown throughout "Insights from feature analyses" section (Figs. 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and Supplementary Figures S1 to S10). This same computation was done for the *C-Net* model, however, $x_i'$ was set to a zero-vector of *89 x 89* shape to simulate the image-like input of a CNN model. *89 x 89* represents the dimensions of the *MeSH2Matrix* matrices, as we considered 89 predefined MeSH qualifiers in this work.

## Human subjects research statement

No human subjects were involved in this analysis. The work involved data collected from publicly available datasets. In the next sections, we delve into an extensive examination of the resulting MeSH2Matrix dataset ("MeSH2Matrix dataset" section). In "Experimental results on biomedical relation classification using MeSH2Matrix" section, we describe our models, our related experiments, and corresponding results. Finally, in "Insights from feature analyses" section, we provide an in-depth analysis of our experiments and results.

## MeSH2Matrix dataset

As of December 12, 2021, our SPARQL query (Figure D6) has successfully retrieved 81,000 biomedical relations between MeSH Terms from Wikidata. This is a very significant amount of information as Wikidata only includes

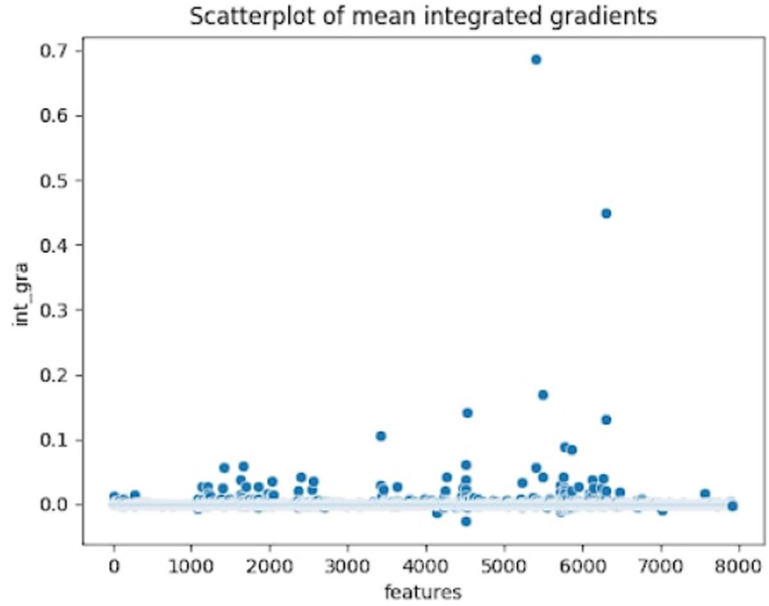


**Fig. 6** Scatterplot of mean integrated gradients for the 7,921 features

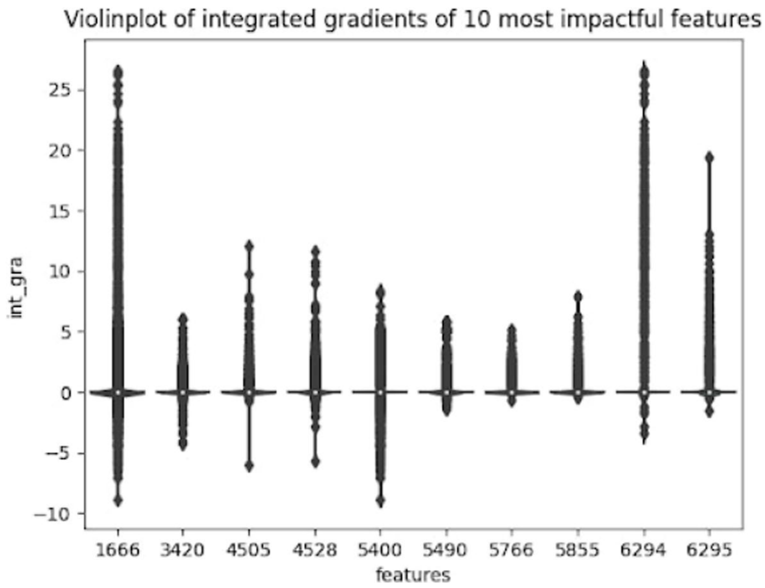


**Fig. 7** Distribution of integrated gradients of 10 most impactful features: 1666 (diagnosis; therapeutic use), 3420 (pharmacology; pharmacology), 4505 (metabolism; enzymology), 4528 (metabolism; drug effects), 5400 (epidemiology; epidemiology), 5490 (ethnology; ethnology), 5766 (therapeutic use; drug therapy), 5855 (administration & dosage; drug therapy), 6294 (drug therapy; therapeutic use), and 6295 (drug therapy; administration & dosage)

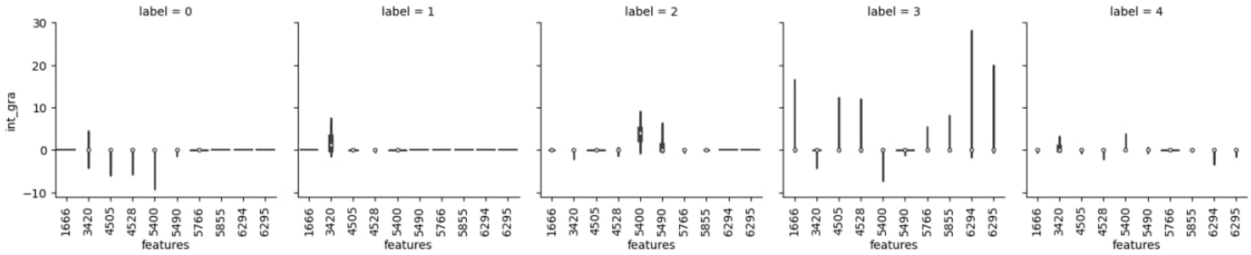


**Fig. 8** Label-based segmentation of the ten most influential features: *Taxonomic* (0), *biomedical symmetric* (1), *non-biomedical symmetric* (2), *biomedical non-symmetric* (3), and *non-biomedical non-symmetric* (4)

99,208 semantic relations between MeSH concepts[9]. We have chosen 81,000 as the number of considered relations in order to simplify the performed computations for the analysis of our proposed approach. When analyzing the biomedical relations extracted for building our dataset, we found out that the supported relation types can be classified into five categories:

- *Non-Biomedical Non-Symmetric* (156 relation types, 17,758 relations),
- *Biomedical Non-Symmetric* (53 relation types, 27,429 relations),
- *Non-Biomedical Symmetric* (12 relation types, 9,000 relations),
- *Biomedical Symmetric* (3 relation types, 1,441 relations), and
- *Taxonomic* (3 relation types, 25,372 relations).

This goes in line with the coverage of various aspects of biomedical knowledge in Wikidata as a multidisciplinary knowledge graph [9, 19]. The extraction of the associations between the subject and object of every semantic relation in the MeSH keywords of PubMed publications has shown that most of the associations are likely to be found in a limited number of publications (Fig. 19A) and that commonly available MeSH associations in the PubMed records are rare (25,227 associations [42.7%] each available in 100 papers or more - Green dot in Fig. 19A). This is evident as scientific productivity follows Lotka's Law, an inverse power law that describes the uneven distribution

[9] For a live update: https://w.wiki/4JN9.

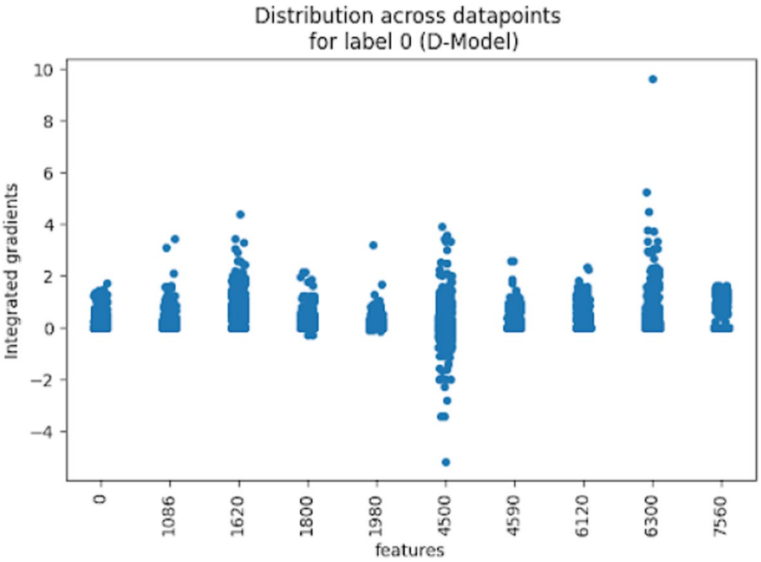


**Fig. 9** Distribution of 10 most impactful features for label 0 [*taxonomic*] (D-model): 0 (analysis; analysis), 1086 (pathology; diagnosis), 1620 (diagnosis; diagnosis), 1800 (etiology; etiology), 1980 (complications; complications), 4500 (metabolism; metabolism), 4590 (biosynthesis; biosynthesis), 6120 (therapy; therapy), 6300 (drug therapy; drug therapy), and 7560 (methods; methods)

of research outputs [72]. When seeing if the extraction of qualifiers describing the MeSH associations has been successful in generating matrices, we found that the probability of the creation of qualifiers' matrices tends to increase with the augmentation of the number of PubMed publications including the MeSH association before reaching a plateau near 1 at twenty publications (Fig. 19B). The existence of biomedical relations in Wikidata that cannot be found in PubMed records and that do not consequently return matrices of correspondence could be explained by the fact that Wikidata is subject to include irrelevant biomedical relations as it is collaboratively edited by human users without any restriction [20]. By contrast, it is important to reveal that the proportion of the generation of the matrices for the MeSH associations available in 100 publications or more is below the plateau with a rate of 73.3% (Red dots in Fig. 19B). To identify the reason behind such an unexpected behavior, we compute the quotient of the

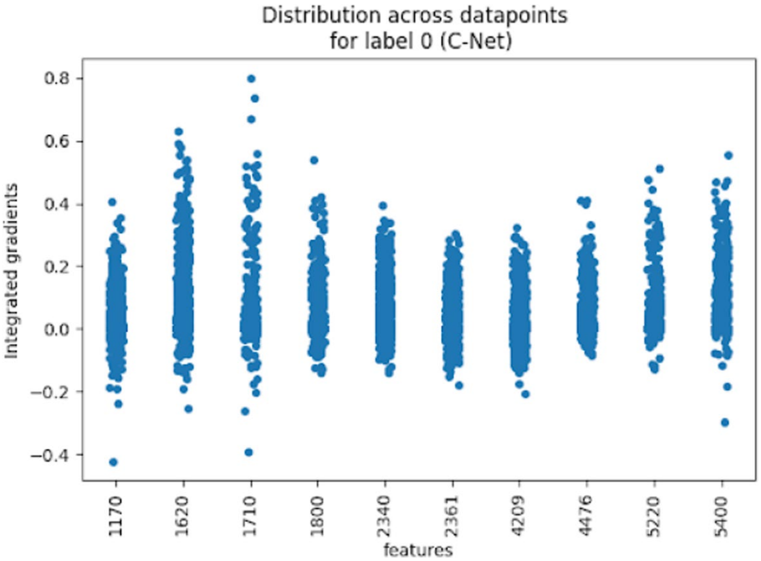


**Fig. 10** Distribution of 10 most impactful features for label 0 [*taxonomic*] (C-Net): 1170 (chemistry; chemistry), 1620 (diagnosis; diagnosis), 1710 (diagnostic imaging; diagnostic imaging), 1800 (etiology; etiology), 2340-2361-4209 (genetics; genetics), 4476 (metabolism; genetics), 5220 (physiopathology; physiopathology), 5400 (epidemiology; epidemiology)

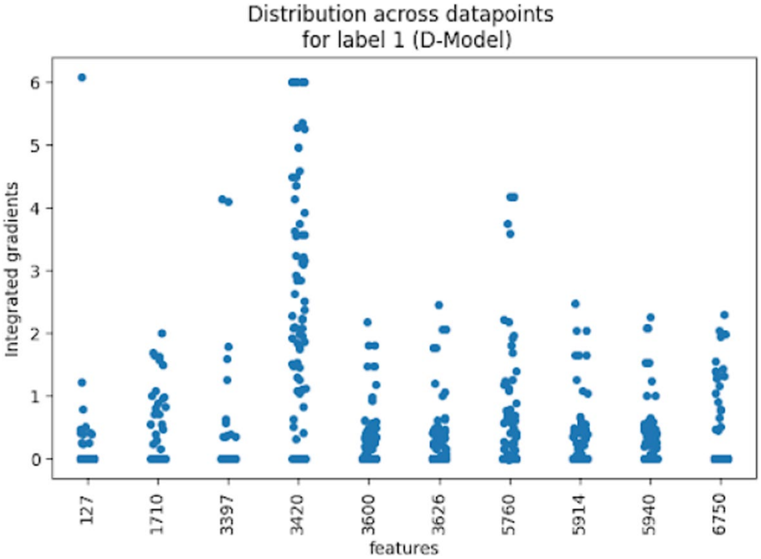


**Fig. 11** Distribution of 10 most impactful features for label 1 [*biomedical symmetric*] (D-model): 127 (blood; pharmacology), 1710 (diagnostic imaging; diagnostic imaging), 3397 (pharmacology; analogs & derivatives), 3420 (pharmacology; pharmacology), 3600-3626-5914-5940 (adverse effects; adverse effects), 5760 (therapeutic use; therapeutic use) and 6750 (surgery; surgery)

MeSH associations not generating matrices and available at least in 100 PubMed papers out of the overall number of the MeSH associations not having qualifiers' matrices for every class of Wikidata relation types. We found out that this rate is significantly higher for *taxonomic* (6,133 out of 13,411, 45.7%) and *non-biomedical non-symmetric* (1,712 out of 8,335, 20.5%) relations than for *biomedical non-symmetric* (1,137 out of 9,498, 12.0%), *non-biomedical symmetric* (189 out of 2,647, 7.1%), and *biomedical symmetric* (7 out of 640, 1.1%). This proves the ability of the MeSH qualifiers to better represent biomedical or symmetric relations than generic and non-symmetric ones.

The obtained dataset of qualifiers' matrices represented 46,469 relations covering the five classes of semantic relation types (195 supported relation types, 54.2% of the matrices based on 100 publications or more): 17,931 biomedical non-symmetric, 11,961 taxonomic, 9,423 non-biomedical non-symmetric, 6,353 non-biomedical symmetric, and

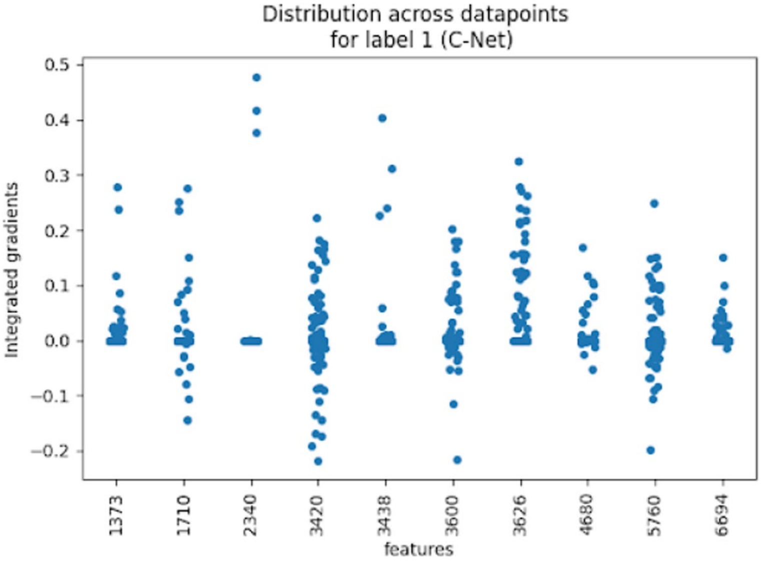


**Fig. 12** Distribution of 10 most impactful features for label 1 [*biomedical symmetric*] (C-Net): 1373 (analogs & derivatives; pharmacology), 1710 (diagnostic imaging; diagnostic imaging), 2340 (genetics; genetics), 3420 (pharmacology; pharmacology), 3438 (pharmacology; pharmacokinetics), 3600-3626 (adverse effects; adverse effects), 4680 (blood; blood), 5760 (therapeutic use; therapeutic use), and 6694 (surgery; diagnostic imaging)

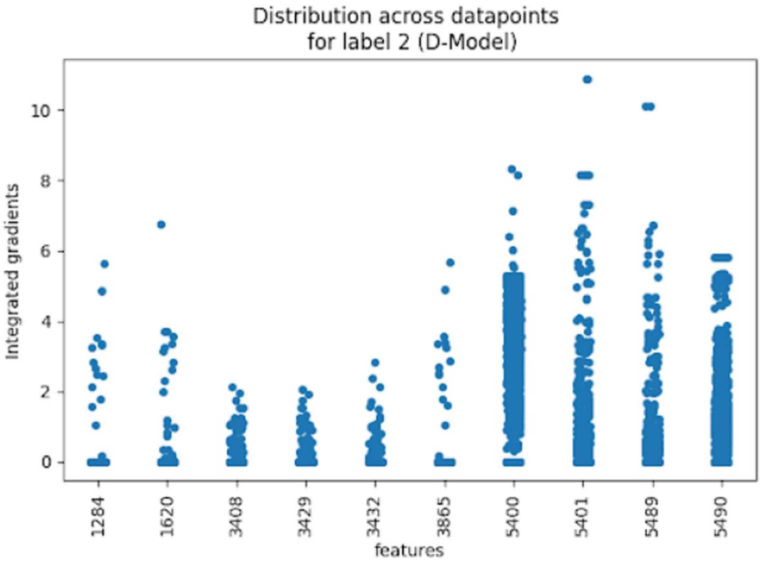


**Fig. 13** Distribution of 10 most impactful features for label 2 [*non-biomedical symmetric*] (D-model): 1284 (agonists; pharmacology), 1620 (diagnosis; diagnosis), 3408-3429 (pharmacology; genetics), 3432 (pharmacology; metabolism), 3865 (agonists; pharmacology), 5400 (epidemiology; epidemiology), 5401 (epidemiology; ethnology), 5489 (ethnology; epidemiology), and 5490 (ethnology; ethnology)

801 biomedical symmetric relations. Unsurprisingly, the most represented relation types in the dataset are mainly taxonomic (e.g., *subclass of* [P279] and *instance of* [P31]) or biomedical non-symmetric ones (e.g., *drug and therapy used for treatment* [P2176] and *symptoms and signs* [P780]) as shown in Fig. 5. A surprising fact is that non-biomedical non-symmetric relations are dominated by *diplomatic relation* [P530] (4,000 out of 9,423 matrices, 42.4%) and *subject has role* [P2868] (3,309 out of 9,423 matrices, 35.1%). The distribution of matrices per relation types follows Lotka's law where a limited number of relation types are attributed a considerable number of matrices. Despite this class imbalance, we have an interesting number of well-represented relation types that can enable biomedical

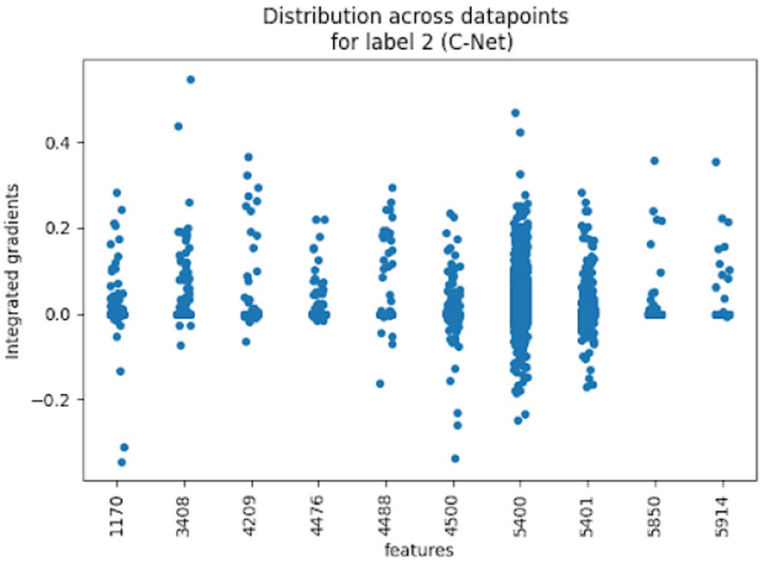


**Fig. 14** Distribution of 10 most impactful features for label 2 [*non-biomedical symmetric*] (C-Net): 1170 (chemistry; chemistry), 3408 (pharmacology; genetics), 4209 (genetics; genetics), 4476 (metabolism; genetics), 4488 (metabolism; pharmacology), 4500 (metabolism; metabolism), 5400 (epidemiology; epidemiology), 5401 (epidemiology; ethnology), 5850 (administration & dosage; administration & dosage), and 5914 (adverse effects; adverse effects)

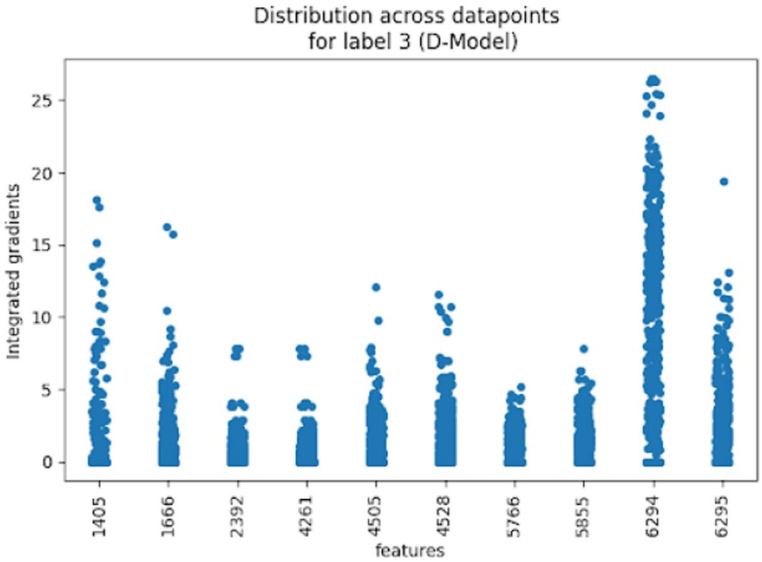


**Fig. 15** Distribution of 10 most impactful features for label 3 [*biomedical non-symmetric*] (D-model): 1405 (analogs & derivatives; drug therapy), 1666 (diagnosis; therapeutic use), 2392-4261 (genetics; drug effects), 4505 (metabolism; enzymology), 4528 (metabolism; drug effects), 5766 (therapeutic use; drug therapy), 5855 (administration & dosage; drug therapy), 6294 (drug therapy; therapeutic use), and 6295 (drug therapy; administration & dosage)

relation classification with high accuracy. This is confirmed even when we only consider the matrices that were based on 100+ scholarly publications as shown in Blue in Fig. 6. This makes our dataset more inclusive than other available corpora for biomedical relation classification only covering a few relation types, particularly *drug interactions*, *drug adverse effects*, and *drug-disease relations* [3].

## Experimental results on biomedical relation classification using MeSH2Matrix

Table 2 shows the results of the three benchmark models on the 195-classification and 5-classification tasks. We refer the reader to "Experimental setup" section for the experimental setup. The metric being used are *accuracy* and *multi-class F1-score* (which is a metric that combines

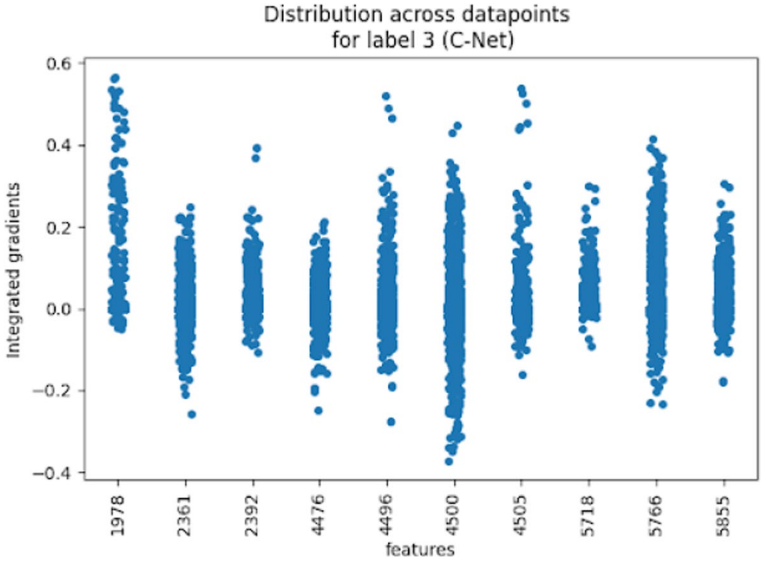


**Fig. 16** Distribution of 10 most impactful features for label 3 [*biomedical non-symmetric*] (C-Net): 1978 (complications; etiology), 2361 (genetics; genetics), 2392 (genetics; drug effects), 4476 (metabolism; genetics), 4496 (metabolism; physiology), 4500 (metabolism; metabolism), 4505 (metabolism; enzymology), 5718 (therapeutic use; complications), 5766 (therapeutic use; drug therapy), and 5855 (administration & dosage; drug therapy)

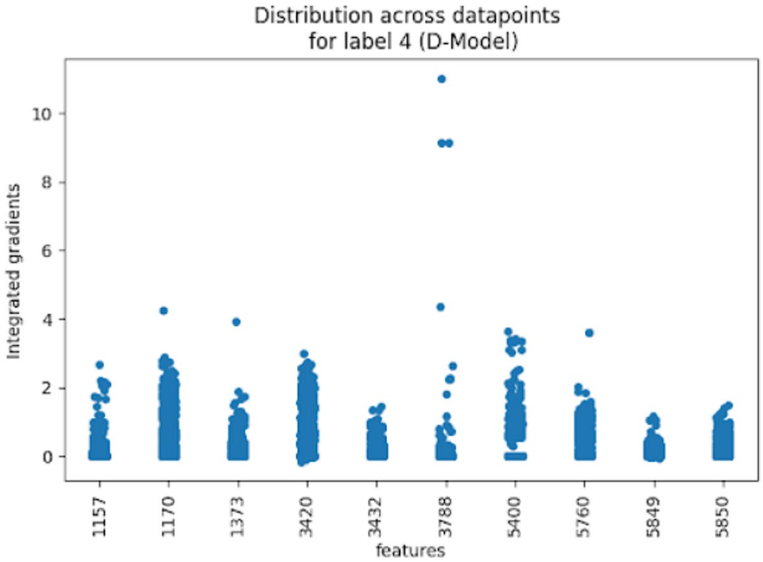


**Fig. 17** Distribution of 10 most impactful features for label 4 [*non-biomedical non-symmetric*] (D-model): 1157 (chemistry; analysis), 1170 (chemistry; chemistry), 1373 (analogs & derivatives; pharmacology), 3420 (pharmacology; pharmacology), 3432 (pharmacology; metabolism), 3788 (toxicity; metabolism), 5400 (epidemiology; epidemiology), 5760 (therapeutic use; therapeutic use), 5849 (administration & dosage; therapeutic use), and 5850 (administration & dosage; administration & dosage)

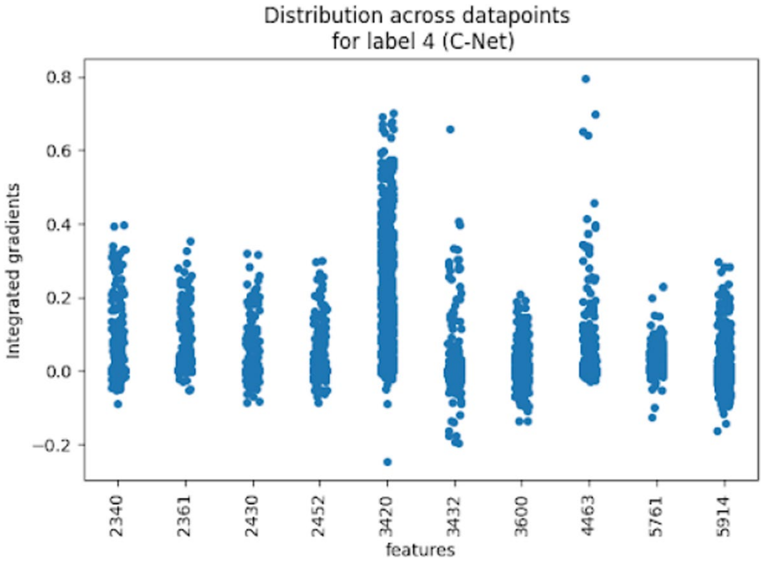


**Fig. 18** Distribution of 10 most impactful features for label 4 [*non-biomedical non-symmetric*] (C-Net): 2340-2361 (genetics; genetics), 2430-2452 (immunology; immunology), 3420 (pharmacology; pharmacology), 3432 (pharmacology; metabolism), 3600-5914 (adverse effects; adverse effects), 4463 (metabolism; chemistry), and 5761 (therapeutic use; administration & dosage)

the precision and recall of the model). It is clear that the three proposed models achieved acceptable accuracy measures that go in line with the recent advances in biomedical relation classification (*F1-Score* between 0.65 and 0.85) [3].

We see a notable improvement in the probabilistic methods (*D-Model* and *C-Net*) over *SVM*. *D-Model* performs the best, although outperforming *C-Net* by a small margin. Another observation is that for all the models, their performance on the 5-class takes was much better than the 195-class ones. This could be because the consideration of five generalized superclasses actually reduces the complexity of the task, making it easier for the model to learn [56]. For example, class generalization

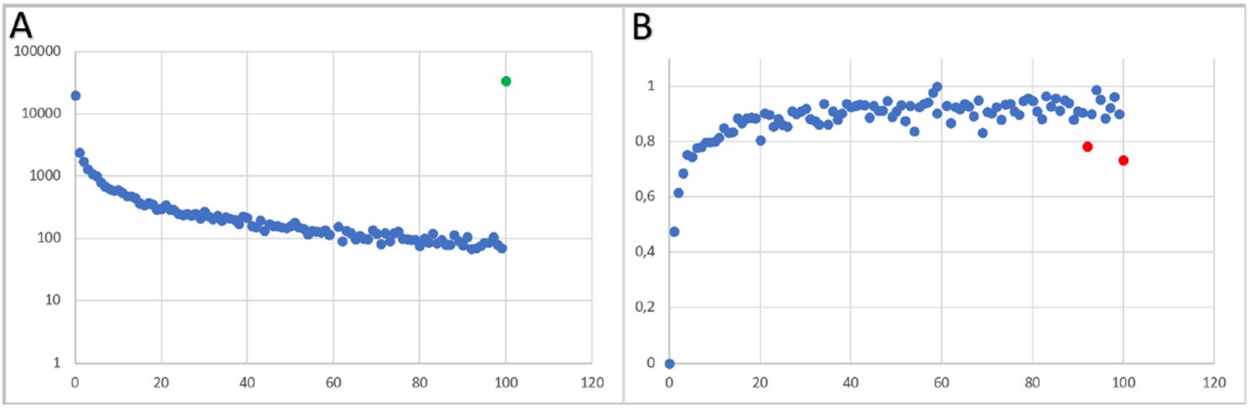


**Fig. 19** Analysis of MeSH Keyword Associations and Qualifier Matrices in PubMed Publications. **19A:** The distribution of the associations between subjects and objects in semantic relations derived from MeSH keywords in PubMed publications. The majority of these associations are concentrated in a limited number of publications. A small fraction (42.7%) of associations are present in 100 or more publications (indicated by the green dot). **19B:** The probability of successful creation of qualifier matrices increases with the number of PubMed publications containing the MeSH association, reaching a plateau near 1 at around twenty publications. However, for associations available in 100 or more publications, the generation rate of these matrices remains below the plateau at 73.3% (indicated by the red dots)

**Table 2** Accuracy [and F1-Score] (in percentage) of the models used in our experiments. In both classes, *D-Model* performs best, followed by *C-Net* and lastly *SVM*. Also, all models performed better on 5 classes compared to 195 classes

| Models | 195 classes | 5 classes |
|---|---|---|
| SVM | 66.43 [61.27] | 78.74 [78.63] |
| D-Model | **70.78** [**66.90**] | **83.09** [**82.94**] |
| C-Net | 70.49 [66.18] | 82.78 [82.61] |

allowed us to be get rid of the closely related taxonomic relation types (i.e., *instance of* [P31], *subclass of* [P279], and *part of* [P361]) and eliminate the effect of the confusion between these three relation types on the accuracy of the models. Another possible reason could be that grouping increased the distribution of some minority classes (classes with a very few samples). On the one hand, due to the enrichment of Wikidata thanks to human efforts, important Wikidata statements can be mistakenly defined for minor relation types [9]. The effect of such deficient relations will become insignificant when the generalization occurs. On the other hand, as of December 12, 2021, 4,522 (4.4%) out of the 99,208 Wikidata relations between MeSH Terms are having the same subject and object as another supported semantic relations between MeSH Concepts[10]. Statements having the same subjects and objects but different relation types are likely to be merged together due to the class generalization, allowing to reduce the confusion between slightly overlapping relation types.

Despite the importance of these results, the adjustment of *MeSH2Matrix* for applying various scenarios of classification ("Experimental results on biomedical relation classification using MeSH2Matrix" section) to *D-Model*, the best performing machine-learning model, can allow a better explanation of the behavior of *MeSH2Matrix*-based biomedical relation classification. The elimination of uncommon semantic relations in *PubMed* (*Scenario 1*) and of under-represented relation types in *MeSH2Matrix* (*Scenario 2*) did not affect the class distribution of training sets as adjusted datasets share the same order of magnitude for *Arithmetic Mean* and *GA-Ratio* as the original *MeSH2Matrix* dataset as shown in Table 3. The relation type-based classification for *Scenario 2* associated with an increase of *Arithmetic Mean* and *GA-Ratio* and the superclass-based classification for *Scenario 1* coupled with a decline of *Arithmetic Mean* and *GA-Ratio* will only matter when accuracy rates significantly lessen for *Scenario 1* and improve for *Scenario 2*. These two scenarios will be important to study the effect of label noise [73] and relation occurrence [8] on the efficiency of the classification and consequently to evaluate the robustness of our approach.

The proposed situations for the restricted generalization of the relation types included in one superclass (*Scenario 3*) seem to be comparable as the arithmetic mean of the number of matrices per relation type ranges between 240.77 and 318.28 and the *GA-Ratio* varies between 0.033 and 0.054 for all the iterations except the merge of non-biomedical non-symmetric relation types (Table 3). Even for the latter, the excess of the rate of items per class (*Matrices per class*: 693.56) and the

[10] Live data: https://w.wiki/4itd.

**Table 3** Descriptive statistics for the situations to be considered for assessing the behavior of the best performing machine-learning model

| Situation | Classes | Geometric mean | Arithmetic mean | GA-Ratio |
|---|---|---|---|---|
| All - Relation Types | 195 | 13.08 | 238.30 | 0.054 |
| All - Superclasses | 5 | 6345.04 | 9293.80 | 0.683 |
| *Scenario 1* - Relation Types | 178 | 10.19 | 141.72 | 0.072 |
| *Scenario 1* - Superclasses | 5 | 2866.20 | 5045.40 | 0.568 |
| *Scenario 2* - Relation Types | 95 | 80.09 | 486.00 | 0.165 |
| *Scenario 2* - Superclasses | 5 | 6308.33 | 9234.00 | 0.683 |
| *Scenario 3* - Taxonomic | 193 | 12.45 | 240.77 | 0.052 |
| *Scenario 3* - Biomedical Non-Symmetric | 146 | 10.57 | 318.28 | 0.033 |
| *Scenario 3* - Biomedical Symmetric | 193 | 12.89 | 240.77 | 0.054 |
| *Scenario 3* - Non-Biomedical Non-Symmetric | 67 | 47.58 | 693.56 | 0.069 |
| *Scenario 3* - Non-Biomedical Symmetric | 186 | 12.24 | 249.83 | 0.049 |

decrease of class imbalance (*GA-Ratio*: 0.069) are not very significant and can be carefully considered as close to the ones of other situations. The comparison between situations where different classes are merged will allow discerning how confusion between relation types occurs for *MeSH2Matrix* [56].

When assessing the effect of *Scenario 1* and *Scenario 2* on the *MeSH2Matrix*-based classification of semantic relations (Fig. 20), it is clear that the removal of minor relation types did not positively affect the classification efficacy as the *Accuracy* and *F1-Score* remained quite the same despite an increase in *Arithmetic mean* and *GA-Ratio*. This proves the robustness of our approach to label noise and endorses that the consideration of a limited number of relation types is not the solution for having a better accuracy of the classification [73]. By contrast, the increase of the *Accuracy* and *F1-Score* of the classification when eliminating uncommon semantic relations co-occurring in less than 100 PubMed records confirms previous findings on the importance of number of associations between terms in scholarly publications for biomedical relation extraction and classification [8]. The lack of such an effect for superclass-based classification shows the limitation of our algorithm to go beyond an accuracy of 84%. Significant improvements for the proposed machine-learning models should be applied to let the classification more reliable.

The evaluation of the effect of restricted generalization of relation types on the accuracy rates of the classification has revealed that this practice was only effective when the generalization concerned non-biomedical relation types and to a lesser extent non-symmetric ones (Fig. 21). The more significant lack of classification of non-biomedical relation types can be explained by the use of a database

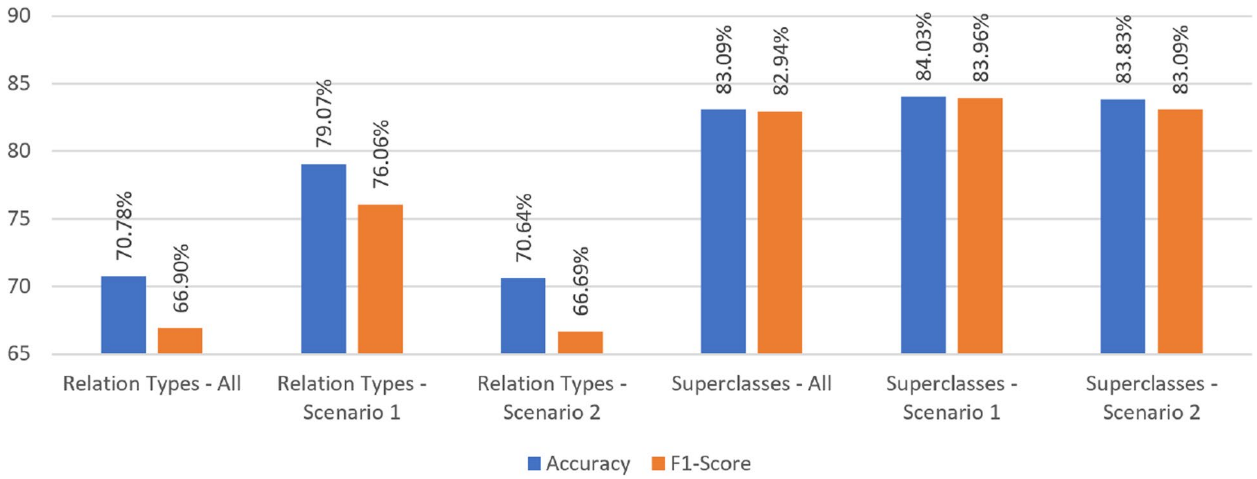


**Fig. 20** *Accuracy* and *F1-Score* for the application of *D-Model* on Scenarios 1 and 2: *Relation Types* corresponds to the MeSH2Matrix classification according to 195 classses, *Superclasses* corresponds to the MeSH2Matrix classification according to 5 categories, and *All* stands for the original *MeSH2Matrix*

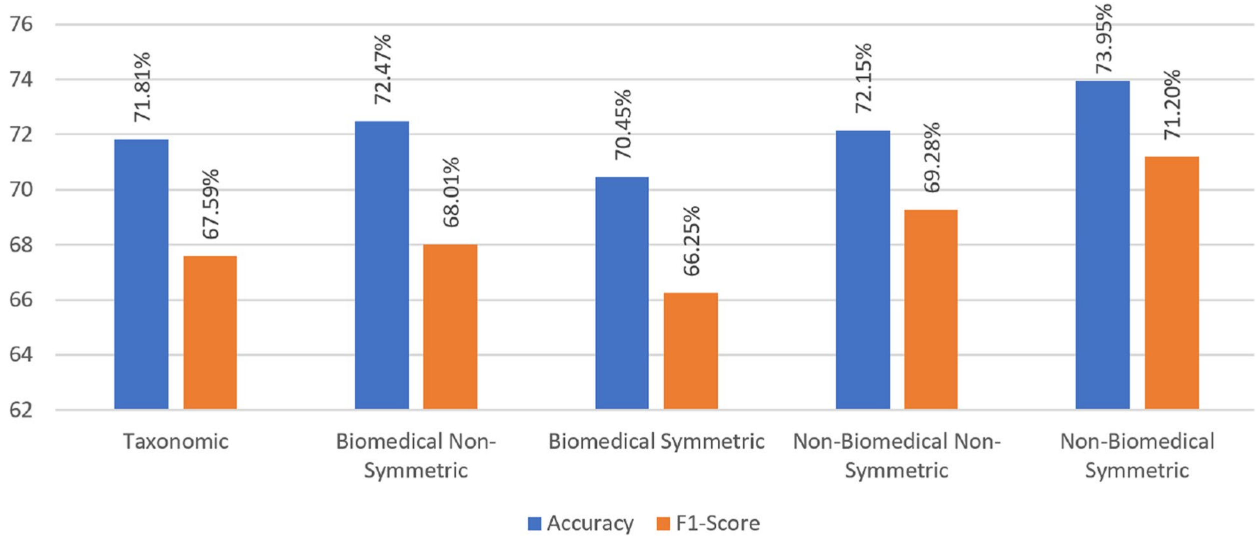


**Fig. 21** *Accuracy* and *F1-Score* for the application of *D-Model* on Scenario 3: Labels reveal the generalized category of relation types for every situation

of biomedical scholarly publications for the creation of *MeSH2Matrix*. In fact, *PubMed* mostly involve biomedical knowledge and consequently provides limited information on *diplomatic relation* [P530] and *shares border with* [P680] statements [13]. Exceptions to this are taxonomic relation types. Such relations constitute the definition of concepts in a given knowledge resource [74] and they can be easily be found in scholarly publications, particularly literature reviews about biomedical entities [75]. As well, when a PubMed publication deals with a concept and its parent class, it is more likely that the same features will be evocated for both entities in that paper, making the two related MeSH terms receive similar qualifiers. This allows the generated matrix to have a symmetric shape that cannot be relevant for other semantic relation types. However, in practice, such an ability to classify taxonomic relations is rarely needed, particularly as *Medical Subject Headings* (MeSH) includes taxonomic relations for all the biomedical concepts it involves [76].

When generating the confusion matrices for the superclass-based classification training of the three proposed models on MeSH2Matrix, we found out that the factors altering the classification accuracy are quite the same across the three algorithms as shown in Fig. 22 although the extent of their influence can significantly differ from a model to another. In fact, we identify that the considerably limited proportion of biomedical symmetic relations considerably altered the efficiency of the classification for this particular superclass (*Circles* in Fig. 22). We also point out that there is a remarkable confusion between taxonomic (expected) and non-symmetric (predicted) relations (*Hexagones* in Fig. 22). This confusion also applies to non-symmetic (expected) and taxonomic (predicted) relations (*Squares* in Fig. 22) as well as to biomedical and non-biomedical non-symmetric relations (*Triangles* in Fig. 22). This kind of confusion is probably motivated by a similarity of patterns between taxonomic and non-symmetric relation types as all these relation types have domains and ranges that are not the same [77]. This finding partly confirm the relative amelioration of accuracy rates when generalizing non-symmetric relations (Fig. 20) and proves that an initial classification of relations as symmetric and non-symmetric can significantly reduce the complexity of our classification. Furthermore, we recognize a significant confusion between non-biomedical symmetric and non-biomedical non-symmetric relations (*Triangles* in Fig. 22). This confirms the large positive effect of the generalization of non-biomedical relations on accuracy rates for *D-Model* (Fig. 20) and can be explained by the fact that the MeSH qualifiers initally used to create the matrices do not capture well non-biomedical semantic relations. This proves in part why MeSH2Matrix-based classification cannot achieve absolute accuracy even when using more efficient machine learning algorithms. Another finding that was not pinpointed through the class generalization experiments is the existence of a bias in the prediction of the considered superclass as taxonomic (*Squares* in Fig. 22). This is mainly due to the very important proportions of taxonomic relations in the *MeSH2Matrix* dataset. This class imbalance seems to more interestingly alter the classification accuracy for classical machine learning models (*SVM*) than for neural networks (*D-Model* and slightly *C-Net*). The slight lack of influence of the other

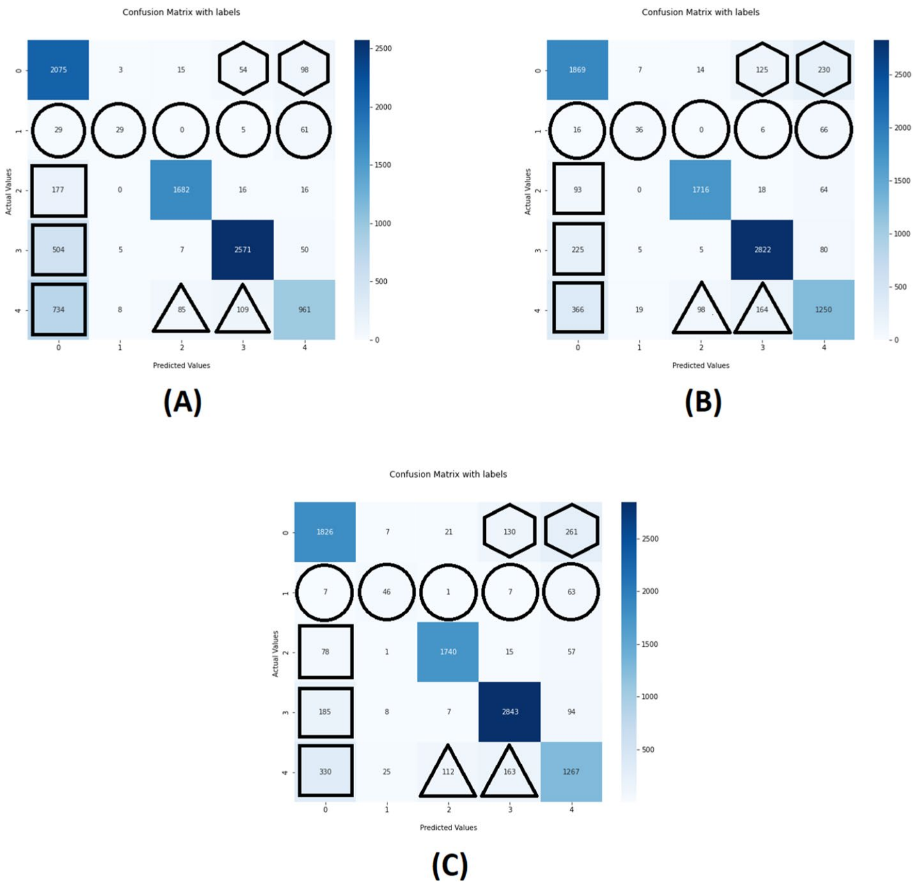


**Fig. 22** Confusion matrices for the superclass-based classification restricted to the MeSH2Matrix training set. **Models**: *SVM* (**A**), *C-Net* (**B**), and *D-Model* (**C**). **Classes**: *Taxonomic* (0), *biomedical symmetric* (1), *non-biomedical symmetric* (2), *biomedical non-symmetric* (3), and *non-biomedical non-symmetric* (4)

confusions for *SVM* (*Haxagones* and *Triangles* in Fig. 22) is not significant and is mainly motivated by the large influence of the bias towards taxonomic relations. This bias significantly explains why *SVM* performed less than neural networks in our classification and prompts the reproduction of our experiments with a more balanced dataset or with a variant of *SVM* that is robust to class imbalance to more objectively compare classical machine learning models to advanced ones [78].

To further study how D-Model and C-Net perform, particularly in the context of biomedical symmetric and taxonomic relations, we provide a comparative analysis of F1-Scores for various relation types using the two models (Table 4). The performance of the two models is generally similar, with minor differences in F1-Scores. However, certain relation types show clear advantages for one model over the other. D-Model outperforms C-Net in relation types such as *diplomatic relation* (P530), *cell component* (P681), *has part* (P527), and *molecular function* (P680). Conversely, C-Net has higher F1-Scores for *instance of* (P31), *drug or therapy used for treatment* (P2176), *medical condition treated* (P2175), and *significant drug interaction* (P769). Both models perform exceptionally well on relations like *medical condition treated* (P2175) and *drug or therapy used for treatment* (P2176), with F1-Scores close to or above 0.9, indicating high accuracy. However, they struggle equally with *shares border with* (P47) and *part of* (P361), where scores are also relatively low. The choice between D-Model and C-Net may thus depend on the specific relation types of interest, particularly as several relation types included in the same superclass may behave differently. Despite this, analyzing superclass-based classification can be useful to discern nuanced differences in model performance across related but distinct relation types.

## Insights from feature analyses

We set out to investigate how the mean integrated gradients were distributed across the 7, 921 features. Figure 6 depicts a scatterplot of the average integrated gradients for each feature. This plot was noisy and not very

**Table 4** Comparison of F1-Scores for the classification of the most common relation types using *D-Model* and *C-Net*. Best model in bold

| Wikidata ID | Relation type | D-Model | C-Net |
|---|---|---|---|
| P279 | Subclass of | **0.655** | 0.654 |
| P530 | Diplomatic relation | **0.852** | 0.850 |
| P31 | Instance of | 0.484 | **0.494** |
| P2868 | Subject has role | 0.803 | **0.805** |
| P681 | Cell component | **0.883** | 0.871 |
| P2176 | Drug or therapy used for treatment | 0.959 | **0.961** |
| P527 | Has part | **0.637** | 0.618 |
| P2175 | Medical condition treated | 0.978 | **0.986** |
| P1995 | Health specialty | **0.881** | 0.856 |
| P682 | Biological process | **0.841** | 0.835 |
| P703 | Found in taxon | **0.903** | 0.881 |
| P780 | Symptoms and signs | 0.779 | **0.780** |
| P171 | Parent taxon | **0.706** | 0.697 |
| P47 | Shares border with | 0.000 | 0.000 |
| P361 | Part of | **0.137** | 0.124 |
| P680 | Molecular function | **0.710** | 0.620 |
| P828 | Has cause | **0.584** | 0.562 |
| P769 | Significant drug interaction | 0.528 | **0.543** |
| P129 | Physically interacts with | 0.868 | **0.870** |

informative. However, it shows that only a limited number of features are positively correlated to the classes. This proves that not all the MeSH qualifiers heavily influence the classification and that some of them can be eliminated without affecting the accuracy of the classification.

To get more informative analyses, we focused on the ten most impactful features (that is the 10 features with the largest positively correlated values). In order to understand how these 10 most impactful features behaved across the dataset, we made a violin plot of the distribution of the 10 most impactful features across the dataset – see Fig. 7. We found that the most represented relation types of the dataset such as *subclass of*, *diplomatic relations*, and *instance of* as well as other common non-biomedical relation types shown at Fig. 5 are not represented by the ten most impactful features. By contrast, common biomedical relation types such as *drug and therapy used for treatment* and *medical condition treated* have several related features among the most impactful ones. An example can be 5766 (therapeutic use; drug therapy) that corresponds to *medical condition treated*.

In order to gain a better understanding of the distribution, we plotted the distribution of the 10 most influential features (features with the most positively correlated gradients) segmented by the dataset's labels – see Fig. 8. Surprisingly, we observe that the most impactful features across the whole dataset (which we found from Fig. 7) seem to be mostly neutral for the classification of relations as *taxonomic* (label 0). This explains the lack of adaptation of the MeSH qualifiers to support non-biomedical relations including taxonomic ones.

Finally, for each of our target labels, we employ integrated gradients ("Feature analysis" section) to examine the influence of the ten most impactful features on the classification performance of our models. This analysis aims to provide deeper insights into the behavior of *D-Model* and *C-Net* for each dataset label. Below, we discuss our findings from feature analysis for each label. For *D-Model*, refer to Figs. 9, 11, 13, 15 and 17, and for *C-Net*, see Figs. 10, 12, 14, 16 and 18.

### Taxonomic (label 0)

Figures 9 and 10 show the 10 most impactful features used by *D-Model* and *C-Net* respectively to classify relations as *taxonomic (label 0)*. It seems that different symmetric features are considered by *D-Model* and *C-Net* for the classification of relations as *taxonomic*. *D-Model* considers features like 1620 (diagnosis; diagnosis) and 6300 (drug therapy; drug therapy) while *C-Net* primarily considers 1710 (chemistry; chemistry), 1620 (diagnosis; diagnosis), 5400 (epidemiology; epidemiology), and 1800 (etiology; etiology) as positive factors for the identification of taxonomic relations. 1710 (chemistry; chemistry) and 5400 (epidemiology; epidemiology) are not even considered by *D-Model* among its ten most impactful features for label 0 (*taxonomic*). Interestingly, we see in Figure S1 that feature 5400 (epidemiology; epidemiology) is rather very negatively correlated to *D-Model*. Features 2340-2361-4209 (genetics; genetics) and 5220 (physiopathology; physiopathology) which are considered among the most significant features of the taxonomic relation classification for *C-Net*, are not among the most impactful features for *D-Model*. The existence of feature 4476 (metabolism; genetics) among the most impactful ones for *C-Net* suggests that several non-symmetric features can stand for taxonomic relations, particularly when the two MeSH qualifiers are used as attributes for the same entity types. In our situation, metabolism[11] and genetics[12] are both subclasses of physiology[13] and they are consequently used as characteristics of organs and species.

The lack of consideration of 5400 (epidemiology; epidemiology) by *D-Model* could partially explain why this model performs better than *C-Net* in the identification

[11] http://id.nlm.nih.gov/mesh/Q000378.

[12] http://id.nlm.nih.gov/mesh/Q000235.

[13] http://id.nlm.nih.gov/mesh/Q000502.

of taxonomic relations as shown in Table 2 and Fig. 22. In fact, this feature stands as a predictive factor for non-biomedical symmetric relations, particularly the ones between countries as revealed in Fig. 8. The attention provided by *C-Net* for 2340-2361-4209 (genetics; genetics) and 1710 (diagnostic imaging; diagnostic imaging) does not mean that *D-Model* does not consider these interesting features. Figure S1 reveals that 1710 (diagnostic imaging; diagnostic imaging) and to a lesser extent 2340-2361-4209 (genetics; genetics) are also positive predictive features for the identification of taxonomic semantic relations for *D-Model*.

#### Biomedical symmetric (label 1)

The analysis of the most positively correlated features for the classification of semantic relations as *biomedical symmetric (label 1)* in *D-Model* (Fig. 11) and *C-Net* (Fig. 12) reveals that many of them are drug-related features involving MeSH subheadings linked to pharmacology like *pharmacology*, *adverse effects*, *therapeutic use*, *pharmacokinetics*, and *analogs* & *derivatives*. This is a direct result of the domination of biomedical non-symmetric relations by significant drug interactions (611 out of 640, 95%) as shown in Fig. 5. The results also show an agreement between *D-Model* and *C-Net* on the usage of symmetric features, particularly 1710 (diagnostic imaging; diagnostic imaging), 3420 (pharmacology; pharmacology), 3600-3626 (adverse effects; adverse effects), and 5760 (therapeutic use; therapeutic use) to identify biomedical symmetric relations.

Furthermore, Figures S2 and S7 show how the 10 most positively correlated features for one model architecture are distributed in another for the classification of label 1. From this we see that while all of the 10 most positively correlated feature for *C-Net* are likewise positively correlated for *D-Model*, features 5914 (adverse effects; adverse effects) and 6750 (surgery; surgery), which are among the 10 most impactful features in *D-Model*, are negatively correlated features for *C-Net*.

#### Non-biomedical symmetric (label 2)

The evaluation of the most impactful features for the classification of relations as label 2 (*non-biomedical symmetric*) by *D-Model* (Fig. 13) and *C-Net* (Fig. 14) has shown 5400 (epidemiology; epidemiology) and 5401 (epidemiology; ethnology) as the most predictive factors for both architectures. *D-Model* also considers 5489 (ethnology; epidemiology) and 5490 (ethnology; ethnology) as main features. These features are ones that have subject and object qualifiers that are equal or corresponding to two MeSH subheadings that are attributed to the same entity types, just as *taxonomic* and *biomedical symmetric* relations, The higher prevalence of *epidemiology* and *ethnology* as impactful MeSH subheadings could be due to the domination of non-biomedical symmetric relations by *diplomatic relations* and *shares border with* statements (Fig. 5). Several other symmetric features related to drugs and diseases such as 1620 (diagnosis; diagnosis) for *D-Model* and 1170 (chemistry; chemistry), 4500 (metabolism; metabolism), 5850 (administration & dosage; administration & dosage), and 5914 (adverse effects; adverse effects) for *C-Net* are also identified as impactful ones. This could be due to the existence of several non-biomedical relation types that can exist between biomedical entities such *physically interacts with* (Fig. 5) that is broader than *significant drug interactions* and involves the interaction of drugs with human genes, proteins, and drugs. This is confirmed by the existence of several non-symmetric features related to associations between drugs and human substances among the most significant features such as 1620 (agonists; pharmacology), 3408-3429 (pharmacology; genetics), 3432 (pharmacology; metabolism), and 4476 (metabolism; genetics). Despite the importance of these non-symmetric features, it is clear that their effect is limited when compared to the ones of symmetric features as shown in Figs. 13 and 14. This confirms again the value of features corresponding to equal subject and object MeSH subheadings to identify symmetric relations.

Figures S3 and S8 show how the 10 most positively correlated features for one model architecture are distributed in another for the classification of label 2. From this we see that all of the 10 most positively correlated features for C-Net are likewise positively correlated for D-Model, except 4476 (metabolism; genet- ics), which seems to be negatively correlated for *D-Model.* Another interesting observation is that the 10 most positively correlated features present in *C-Net* have positive correlation values significantly greater in *D-Model.*

#### Biomedical non-symmetric (label 3)

The analysis of the most significant features for the identification of biomedical non-symmetric relations by *D-Model* (Fig. 15) and *C-Net* (Fig. 16) proves that all the considered features are non-symmetric except 2361 (genetics; genetics) and 4500 (metabolism; metabolism) for *C-Net.* This fact slightly explains the better accuracy of *D-Model* for the classification of relations as biomedical non-symmetric ones as stated in Table 2 and Fig. 22. Even for *C-Net*, these features are both positive and negative predictors and this means that they do not have a conclusive predictive power and are just adding noise to the supervised classification. The consensus between *D-Model* and *C-Net* in identifying several non-symmetric features as positive predictors for biomedical non-symmetric relations – specifically, 2392 (genetics; drug effects), 4505 (metabolism; enzymology), 5766 (therapeutic use; drug therapy), and 5855 (administration & dosage; drug therapy) – demonstrates the effectiveness of these features in easily identifying two significantly represented biomedical non-symmetric relations: *drug or therapy used for treatment* and *medical condition treated* (Fig. 5). A better identification of features for other biomedical non-symmetric relation types would therefore be enabled through an enhanced representation of them in the *MeSH2Matrix* dataset. The fact that most of the features are drug-related involving *analogs & derivatives*, *drug effects*, *therapeutic use*, and *metabolism* is explained by the over-representation of drug-disease semantic relations (Fig. 5).

Figures S4 and S9 show how the 10 most positively correlated features for one model architecture are distributed in another for the classification of label 3. In Figure S9 we see a significantly greater positive correlation from *D-Model* and in Figure S4, we observe that three features – 2361 (genetics; genetics), 4500 (metabolism; metabolism) and 5718 (therapeutic use; complications) – which are among the 10 most impactful for *C-Net*, are negatively correlated features for *D-Model*.

#### Non-biomedical non-symmetric (label 4)

As for non-biomedical non-symmetric relation types, the assessment of the most important features for *D-Model* (Fig. 17) and *C-Net* (Fig. 18) surprisingly finds many symmetric features as impactful: 1170 (chemistry; chemistry) for *D-Model*, 2340-2361 (genetics; genetics) for *C-Net*, 2430-2452 (immunology; immunology) for *C-Net*, 3420 (pharmacology; pharmacology) for *D-Model* and *C-Net*, 5400 (epidemiology; epidemiology) for *D-Model*, and 5850 (administration & dosage; administration & dosage) for *D-Model*. This can be related to the possibility of using non-biomedical relation types such as *has cause* to characterize statements linking between two biomedical relations of the same type. In contrast to biomedical non-symmetric relations, the impact of symmetric features is quite comparable to the one of non-symmetric ones. Even the non-symmetric features that are positively correlated to the classification of relations as non-biomedical non-symmetric ones are linked to MeSH qualifiers that are attributed to the same entity type, particularly 1157 (chemistry; analysis) for *D-Model*, 1373 (analogs & derivatives; pharmacology) for *D-Model*, 3788 (toxicity; metabolism) for *D-Model*, 3432 (pharmacology; metabolism) for *D-Model* and *C-Net*, 4463 (metabolism; chemistry) for *C-Net*, and 5849 (administration & dosage; therapeutic use) for *D-Model*. Such a result is mostly due to the confusion of non-biomedical non-symmetric relations with taxonomic relations as well as with non-biomedical symmetric relations as shown in Fig. 22. This is mainly due to the existence of causal, chemical and biological relation types that are not directly linked to medicine such as *an isotopically modified form of*, *parent cell line* or *does not have cause* and that can be too close to several taxonomic and non-biomedical symmetric relations. We propose a solution to this confusion: splitting the matrices into an array of symmetric features and a matrix of non-symmetric features, separately train the models to classify relations based on symmetric features and non-symmetric ones as shown in Fig. 23, and restrict the use of symmetric features to the classification of taxonomic and symmetric relations.

Figures S5 and S10 show how the 10 most positively correlated features for one model architecture are distributed in another for the classification of label 4. Figure S5 demonstrates that, among the most impactful features for *C-Net*, 3420 (pharmacology; pharmacology) and 3432 (pharmacology; metabolism) are positively correlated with *D-Model*.

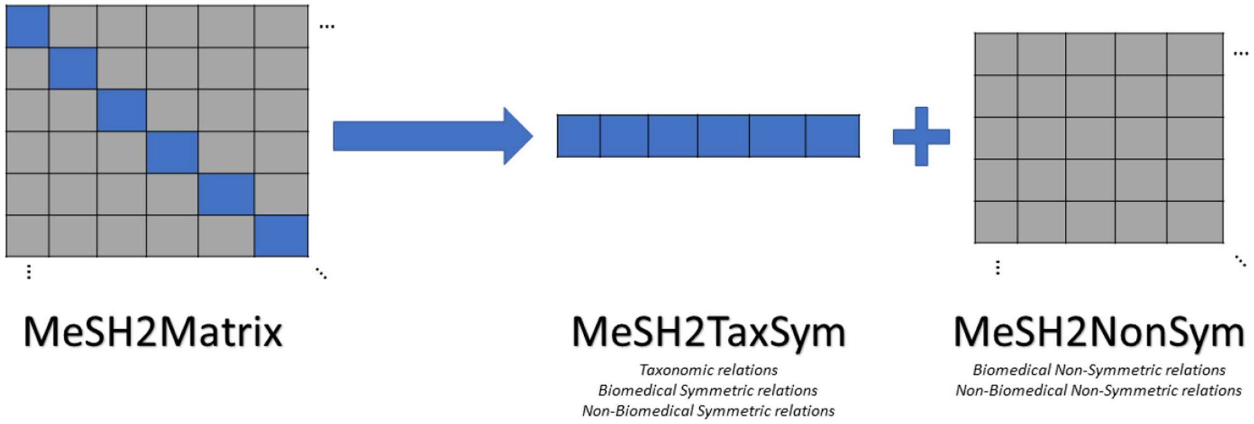

MeSH2Matrix
MeSH2TaxSym
Taxonomic relations
Biomedical Symmetric relations
Non-Biomedical Symmetric relations
MeSH2NonSym
Biomedical Non-Symmetric relations
Non-Biomedical Non-Symmetric relations

**Fig. 23** A diagram about splitting the *MeSH2Matrix* dataset into two specific datasets for classifying specific types of semantic relations

## Conclusion

### Key findings

In this research paper, we proposed a novel approach to the classification of biomedical relations retrieved from Wikidata, an open and collaborative knowledge graph maintained by the Wikimedia Foundation, based on the association between the qualifiers of two semantically related MeSH keywords of PubMed scholarly publications. We generated *MeSH2Matrix* as a large-scale training dataset (covering 195 relation types involved in five superclasses) to enable the MeSH-based biomedical relation classification and we trained three benchmarking machine-learning models (Support Vector Machine [*SVM*], a dense model [*D-Model*], and a convolutional neural network [*C-Net*]) to evaluate the efficiency of our approach to classify various types of biomedical relations. We found an interesting efficiency of our approach in biomedical relation classification proving the promising value of using Bibliometric-Enhanced Information Retrieval towards the improvement of biomedical relation classification. Then, we studied how the behavior of the confusion between the 195 relation types changes as label generalization occurs. Similarly, we created a confusion matrix for the superclass-based classification training. We found that our approach has better behavior in classifying *taxonomic*, *biomedical symmetric*, and *biomedical non-symmetric* relations although several different behaviors can be observed inside the same superclass. Finally, using integrated gradients, we performed comprehensive feature analyses of *C-Net* and *D-Model* to evaluate how the MeSH qualifier features impacted the model's class predictions (using the 5-class prediction case). We identified the top 10 most positively correlated features (which we referred to as impactful features) for each model architecture and class label. We observed that the impactful features vary by class and by model architecture - with some features being positively correlated in one architecture for a given class and negatively correlated in another model for the same class.

### Limitations

Our approach lacks accuracy in classifying non-biomedical relations such as *subject has role* and *shares border with*. This is mainly linked to the lack of MeSH qualifiers that characterize non-biomedical facets of the concepts. As well, a number of MeSH qualifiers do not contribute to the classification of biomedical relations. Getting rid of such MeSH qualifiers will allow rising the speed of the classification training without worrying about losing overall accuracy. From the perspective of using Wikidata as a benchmark for our approach, we found that Wikidata can include inconsistencies due to its open and collaborative editing policy that allows vandalism. This is practically proved by the lack of availability of research articles about a considerable rate of Wikidata relations in PubMed.

### Future directions

As a future direction of this work, we look forward to optimizing our approach for biomedical relation classification by considering the semantic features of the subjects and objects and revising our MeSH2Matrix dataset. Also, we propose to expand our approach into a method for constructing and enriching biomedical knowledge graphs based on PubMed MeSH Keywords. We envision coupling our method with knowledge-based approaches for biomedical entity recognition and relation extraction from the MeSH keywords of scholarly publications as well as to shape-based methods for the validation of biomedical ontologies, particularly *ShEx* and *SHACL*, for such a purpose.

### Abbreviations

| | |
|---|---|
| ANN | Artificial Neural Network |
| API | Application Programming Interface |
| BC5CDR | Biocreative V Chemical Disease Relation |
| BERT | Bidirectional Encoder Representations from Transformers |
| BioBERT | Biomedical Bidirectional Encoder Representations from Transformers |
| CNN | Convolutional Neural Network |
| EHRs | Electronic Health Records |
| FN | False Negatives |
| FP | False Positives |
| GPT | Generative Pre-trained Transformer |
| i2b2/VA | Informatics for Integrating Biology and the Bedside / Veterans Affairs |
| LSTM | Long Short-Term Memory |
| MADE | Medication and Adverse Drug Events from Electronic Health Records |
| MeSH | Medical Subject Headings |
| NCBI | National Center for Biotechnology Information |
| NNs | Neural Networks |
| PPIs | Protein-Protein Interactions |
| PHAEDRA | PHArmacovigilence Entity DRug Annotation |
| RDF | Resource Description Framework |
| SVM | Support Vector Machine |
| TN | True Negatives |
| TSV | Tab-Separated Values |
| TP | True Positives |
| XAI | Explainable Artificial Intelligence |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13326-024-00319-w.

Supplementary Material 1.

**Acknowledgements**

Preliminary results of this research paper have been presented at the 12th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2022 (BIR 2022) held on April 10, 2022 in Stavanger, Norway [10]. We thank the Sisonkebiotik Community, particularly Chris Fourie (University of the Witwatersrand, South Africa), for contributing to the development of the research collaboration that resulted in this research paper. We thank Dr. Ahmed Ben Abdelaziz (University of Sousse, Tunisia) for introducing the Medical Subject Headings to the first author of this research work in March 2016. We also thank Mr. Colin Leong (University of Dayton, United States of America) for his useful comments and discussion.

**Authors' contributions**
Houcemeddine Turki, Bonaventure F. P. Dossou, Chris Chinenye Emezue, and Abraham Owodunni equally contributed to this research paper. Hanen Ben Hassen contributed to the preliminary experiments that led to the establishment of the MeSH2Matrix approach. Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, and Afif Masmoudi supervised the development of the final output.

**Funding**
This work has been supported by the Tunisian Ministry of Higher Education and Scientific Research (MoHESR) within the framework of the Federated Research Project PRFCOV19-D1-P1 and by Craig Newmark Philanthropies, Facebook, and Microsoft through the WikiCred Grant Initiative. This paper is issued as an open-access publication thanks to the "Adapting Wikidata to support clinical practice using Data Science, Semantic Web and Machine Learning" Research Project funded by the Wikimedia Research Fund (G-RS-2204-08368) of the Wikimedia Foundation.

**Availability of data and materials**
For reproducibility purposes, our source code and dataset are currently available at https://github.com/SisonkeBiotik-Africa/MeSH2Matrix. A demo has been made available at https://colab.research.google.com/github/SisonkeBiotik-Africa/MeSH2Matrix/blob/main/MeSH2Matrix_Demo.ipynb to enable users to explore the output of D-Model and C-Net for the relation type-based classification.

**Data availability**
For reproducibility purposes, our source code and dataset are currently available at https://github.com/SisonkeBiotik-Africa/MeSH2Matrix.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha are active contributors to the Wikidata Community.

**Author details**
[1]Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia. [2]Mila Quebec AI Institute, Montreal, Canada. [3]McGill University, Montreal, Canada. [4]Technical University of Munich, Munich, Germany. [5]Faculty of Life Sciences, University of Ilorin, Ilorin, Nigeria. [6]Laboratory of Probability and Statistics, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia.

Received: 14 November 2023 Accepted: 31 August 2024
Published online: 02 October 2024

**References**

1. Shahab E. A Short Survey of Biomedical Relation Extraction Techniques. 2017. arXiv:1707.05850.
2. Huang MS, Han JC, Lin PY, You YT, Tsai RTH, Hsu WL. Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource. Brief Bioinform. 2024;25(3):bbae132. https://doi.org/10.1093/bib/bbae132.
3. Zhang Y, Lin H, Yang Z, Wang J, Sun Y, Xu B, et al. Neural network-based approaches for Biomedical Relation Classification: A Review. J Biomed Inform. 2019;99:103294. https://doi.org/10.1016/j.jbi.2019.103294.
4. Turki H, Hadj Taieb MA, Ben Aouicha M, Fraumann G, Hauschke C, Heller L. Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata. Front Res Metrics Anal. 2021;6:694307. https://doi.org/10.3389/frma.2021.694307.
5. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The ontology for biomedical investigations. PLoS ONE. 2016;11(4):e0154556. https://doi.org/10.1371/journal.pone.0154556.
6. Hoehndorf R, Dumontier M, Gkoutos GV. Evaluation of Research in Biomedical ontologies. Brief Bioinform. 2012;14(6):696–712. https://doi.org/10.1093/bib/bbs053.
7. Konopka BM. Biomedical ontologies–A Review. Biocybernetics Biomed Eng. 2015;35(2):75–86. https://doi.org/10.1016/j.bbe.2014.06.002.
8. Turki H, Hadj Taieb MA, Ben Aouicha M. MeSH qualifiers, publication types and relation occurrence frequency are also useful for a better sentence-level extraction of biomedical relations. J Biomed Inform. 2018;83:217–8. https://doi.org/10.1016/j.jbi.2018.05.011.
9. Turki H, Shafee T, Hadj Taieb MA, Ben Aouicha M, Vrandečić D, Das D, et al. Wikidata: a large-scale collaborative ontological medical database. J Biomed Inform. 2019;99:103292. https://doi.org/10.1016/j.jbi.2019.103292.
10. Turki H, Dossou BFP, Emezue CC, Hadj Taieb MA, Ben Aouicha M, Ben Hassen H, et al. MeSH2Matrix: Machine learning-driven biomedical relation classification based on the MeSH keywords of PubMed scholarly publications. In: Frommholz I, Mayr P, Cabanac G, Verberne S, editors. Proceedings of the 12th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 44th European Conference on Information Retrieval (ECIR 2022). Aachen, Heidelberg: CEUR Workshop Proceedings; 2022. pp. 45–60. http://ceur-ws.org/Vol-3230/paper-07.pdf.
11. Baumann N. How to use the medical subject headings (MeSH). Int J Clin Pract. 2016;70(2):171–4. https://doi.org/10.1111/ijcp.12767.
12. Leydesdorff L, Comins JA, Sorensen AA, Bornmann L, Hellsten I. Cited references and medical subject headings (MeSH) as two different knowledge representations: Clustering and mappings at the paper level. Scientometrics. 2016;109(3):2077–91. https://doi.org/10.1007/s11192-016-2119-7.
13. Lu Y, Figler B, Huang H, Tu YC, Wang J, Cheng F. Characterization of the mechanism of drug-drug interactions from pubmed using MeSH terms. PLoS ONE. 2017;12(4):e0173548. https://doi.org/10.1371/journal.pone.0173548.
14. Tran T, Kavuluru R. Distant supervision for treatment relation extraction by leveraging MeSH subheadings. Artif Intell Med. 2019;98:18–26. https://doi.org/10.1016/j.artmed.2019.06.002.
15. Älgå A, Eriksson O, Nordberg M. Analysis of scientific publications during the early phase of the COVID-19 pandemic: Topic modeling study. J Med Internet Res. 2020;22(11):e21559. https://doi.org/10.2196/21559.
16. Chapman B, Chang J. Biopython: Python tools for computational biology. ACM SIGBIO Newsl. 2000;20(2):15–9. https://doi.org/10.1145/360262.360268.
17. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely available python tools for Computational Molecular Biology and Bioinformatics. Bioinformatics. 2009;25(11):1422–3. https://doi.org/10.1093/bioinformatics/btp163.
18. Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. Commun ACM. 2014;57(10):78–85. https://doi.org/10.1145/2629489.
19. Turki H, Hadj Taieb MA, Shafee T, Lubiana T, Jemielniak D, Ben Aouicha M, et al. Representing COVID-19 information in collaborative knowledge graphs: The case of Wikidata. Semant Web. 2022;13(2):233–64. https://doi.org/10.3233/sw-210444.
20. Turki H, Jemielniak D, Hadj Taieb MA, Labra Gayo JE, Ben Aouicha M, Banat M, et al. Using logical constraints to validate statistical information about COVID-19 in collaborative knowledge graphs: the case of Wikidata. Zenodo. 2021. https://doi.org/10.5281/zenodo.4008358.
21. Lee Y, Son J, Song M. BertSRC: transformer-based semantic relation classification. BMC Med Inform Decis Mak. 2022;22(1). https://doi.org/10.1186/s12911-022-01977-5. https://doi.org/10.1186/s12911-022-01977-5.
22. Camacho-Collados J, Pilehvar MT. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. J Artif Intell Res. 2018;63:743–788. https://doi.org/10.1613/jair.1.11259.
23. Ji G, He S, Xu L, Liu K, Zhao J. Knowledge Graph Embedding via Dynamic Mapping Matrix. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing: Association for Computational Linguistics; 2015. pp. 687–696. https://doi.org/10.3115/v1/P15-1067.

24. Messner J, Abboud R, Ceylan II. Temporal Knowledge Graph Completion Using Box Embeddings. Proceedings of the AAAI Conference on Artificial Intelligence. 2022;36(7):7779–87. https://doi.org/10.1609/aaai.v36i7.20746.
25. Feng J, Huang M, Zhao L, Yang Y, Zhu X. Reinforcement Learning for Relation Classification From Noisy Data. Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1). https://doi.org/10.1609/aaai.v32i1.12063.
26. Zhang Z. Weakly-supervised relation classification for information extraction. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM; 2004. https://doi.org/10.1145/1031171.1031279.
27. Rios A, Kavuluru R, Lu Z. Generalizing biomedical relation classification with neural adversarial domain adaptation. Bioinformatics. 2018;34(17):2973–81. https://doi.org/10.1093/bioinformatics/bty190.
28. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2019;27(1):3–12. https://doi.org/10.1093/jamia/ocz166.
29. Xu R, Li L, Wang Q. dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. BMC Bioinformatics. 2014;15(1). https://doi.org/10.1186/1471-2105-15-105.
30. Huang MS, Lai PT, Lin PY, You YT, Tsai RTH, Hsu WL. Biomedical named entity recognition and linking datasets: survey and our recent development. Brief Bioinform. 2020;21(6):2219–38. https://doi.org/10.1093/bib/bbaa054.
31. Alimova I, Tutubalina E, Nikolenko SI. Cross-Domain Limitations of Neural Models on Biomedical Relation Classification. IEEE Access. 2022;10:1432–9. https://doi.org/10.1109/access.2021.3135381.
32. Parwez MA, Fazil M, Arif M, Nafis MT, Auwul MR. Biomedical Text Classification Using Augmented Word Representation Based on Distributional and Relational Contexts. Comput Intell Neurosci. 2023;2023:1–22. https://doi.org/10.1155/2023/2989791.
33. Sosa DN, Hintzen R, Xiong B, de Giorgio A, Fauqueur J, Davies M, et al. Associating biological context with protein-protein interactions through text mining at PubMed scale. J Biomed Inform. 2023;145: 104474. https://doi.org/10.1016/j.jbi.2023.104474.
34. Shu F, Qiu J, Larivière V. Mapping the biomedical sciences using Medical Subject Headings: a comparison between MeSH co-assignments and MeSH citation pairs. J Med Libr Assoc. 2021;109(3). https://doi.org/10.5195/jmla.2021.1173.
35. Nentidis A, Chatzopoulos T, Krithara A, Tsoumakas G, Paliouras G. Large-scale investigation of weakly-supervised deep learning for the fine-grained semantic indexing of biomedical literature. J Biomed Inform. 2023;146:104499. https://doi.org/10.1016/j.jbi.2023.104499.
36. Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. Best match: New relevance search for Pubmed. PLoS Biol. 2018;16(8):e2005343. https://doi.org/10.1371/journal.pbio.2005343.
37. Cristianini N, Ricci E. In: Kao MY, editor. Support Vector Machines. Boston: Springer US; 2008. pp. 928–932. https://doi.org/10.1007/978-0-387-30162-4_415.
38. Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C, Rouveirol C, editors. Machine Learning: ECML-98. Springer, Berlin Heidelberg: Berlin, Heidelberg; 1998. pp. 137–42.
39. Ben Abacha A, Zweigenbaum P. A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts. In: Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg; 2011. pp. 139–150. https://doi.org/10.1007/978-3-642-19437-5_11.
40. Mavropoulos T, Liparas D, Symeonidis S, Vrochidis S, Kompatsiaris I. A Hybrid Approach for Biomedical Relation Extraction Using Finite State Automata and Random Forest-Weighted Fusion. In: Gelbukh A, editor. Computational Linguistics and Intelligent Text Processing. Cham: Springer International Publishing; 2018. p. 450–62.
41. Muzaffar AW, Azam F, Qamar U. A Relation Extraction Framework for Biomedical Text Using Hybrid Feature Set. Comput Math Methods Med. 2015;2015:1–12. https://doi.org/10.1155/2015/910423.
42. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation Classification via Convolutional Deep Neural Network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin: Dublin City University and Association for Computational Linguistics; 2014. pp. 2335–2344. https://aclanthology.org/C14-1220.
43. dos Santos C, Xiang B, Zhou B. Classifying Relations by Ranking with Convolutional Neural Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing: Association for Computational Linguistics; 2015. pp. 626–634. https://doi.org/10.3115/v1/P15-1061.
44. Peng Y, Lu Z. Deep learning for extracting protein-protein interactions from biomedical literature. In: BioNLP 2017. Vancouver: Association for Computational Linguistics; 2017. pp. 29–38. https://doi.org/10.18653/v1/W17-2304.
45. Heaton J. Introduction to Neural Networks for Java. 2nd ed. Heaton Research: Inc; 2008.
46. Stathakis D. How many hidden layers and nodes? Int J Remote Sens. 2009;30(8):2133–47. https://doi.org/10.1080/01431160802549278.
47. Demuth HB, Beale MH, De Jess O, Hagan MT. Neural Network Design. 2nd ed. Stillwater: Martin Hagan; 2014.
48. Agarap AF. Deep Learning using Rectified Linear Units (ReLU). 2018. arXiv:1803.08375.
49. Mou L, Meng Z, Yan R, Li G, Xu Y, Zhang L, et al. How Transferable are Neural Networks in NLP Applications? In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics; 2016. pp. 479–489. https://doi.org/10.18653/v1/D16-1046.
50. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. https://doi.org/10.1145/3065386.
51. Liu S, Tang B, Chen Q, Wang X. Drug-Drug Interaction Extraction via Convolutional Neural Networks. Comput Math Meth Med. 2016;2016:6918381. https://doi.org/10.1155/2016/6918381.
52. Quan C, Hua L, Sun X, Bai W. Multichannel Convolutional Neural Network for Biological Relation Extraction. BioMed Res Int. 2016;2016:1850404. https://doi.org/10.1155/2016/1850404.
53. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. vol. 37 of Proceedings of Machine Learning Research. Lille: PMLR; 2015. pp. 448–456. https://proceedings.mlr.press/v37/ioffe15.html.
54. Zhang Y, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei: Asian Federation of Natural Language Processing; 2017. pp. 253–263. https://aclanthology.org/I17-1026.
55. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res. 2012;13(null):281–305. https://dl.acm.org/doi/abs/10.5555/2188385.2188395.
56. Turki H, Hadj Taieb MA, Ben Aouicha M. How knowledge-driven class generalization affects classical machine learning algorithms for mono-label supervised classification. In: Proceedings of the 21st International Conference on Intelligent Systems Design and Applications. Online: Springer; 2021. pp. 637–646. https://doi.org/10.1007/978-3-030-96308-8_59.
57. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Bengio Y, LeCun Y, editors. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015. arXiv:1412.6980.
58. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 32. Vancouver: Curran Associates, Inc.; 2019. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
59. Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. 2020. arXiv:2008.05756.
60. Qu W, Balki I, Mendez M, Valen J, Levman J, Tyrrell PN. Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging. Int J Comput Assist Radiol Surg. 2020;15(12):2041–8. https://doi.org/10.1007/s11548-020-02260-6.
61. Jasso G. Measuring Inequality: Using the Geometric Mean/Arithmetic Mean Ratio. Sociol Methods Res. 1982;10(3):303–26. https://doi.org/10.1177/0049124182010003004.

62. Varghese J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. Visceral Med. 2020;36(6):443–9. https://doi.org/10.1159/000511930.
63. Bacciu D, Colombo M, Morelli D, Plans D. ELM Preference Learning for Physiological Data. In: ESANN 2017 - Proceedings, 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. 2017. pp. 99–104. https://www.psy.ox.ac.uk/publications/1049200.
64. Martino FD, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif Intell Rev. 2022. https://doi.org/10.1007/s10462-022-10304-3.
65. Mishra S, Dutta S, Long J, Magazzeni D. A survey on the robustness of feature importance and counterfactual explanations. 2021. arXiv:2111.00358.
66. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Long Beach: Curran Associates Inc.; 2017. pp. 4768–4777. https://dl.acm.org/doi/abs/10.5555/3295222.3295230.
67. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM; 2016. pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.
68. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES '20. New York: Association for Computing Machinery; 2020. pp. 180–186. https://doi.org/10.1145/3375627.3375830.
69. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17. Sydney: JMLR.org; 2017. pp. 3319–3328. https://dl.acm.org/doi/abs/10.5555/3305890.3306024.
70. Kindermans PJ, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, et al. The (Un)reliability of Saliency Methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Berlin: Springer International Publishing; 2019. pp. 267–280. https://doi.org/10.1007/978-3-030-28954-6_14.
71. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. 2017. arXiv:1706.03825.
72. Egghe L, Rousseau R. Theory and practice of the shifted Lotka function. Scientometrics. 2012;91(1):295–301. https://doi.org/10.1007/s11192-011-0539-y.
73. Van Horn G, Mac Aodha O, Song Y, Cui Y, Sun C, Shepard A, et al. The INaturalist Species Classification and Detection Dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 8769–8778. arXiv:1707.06642.
74. Burgun A, Bodenreider O. Aspects of the Taxonomic Relation in the Biomedical Domain. In: Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001. FOIS '01. New York: Association for Computing Machinery; 2001. pp. 222–233. https://doi.org/10.1145/505168.505190.
75. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: Choices, assumptions, and definitions. Ann Rev Stat Appl. 2021;8(1):141–63. https://doi.org/10.1146/annurev-statistics-042720-125902.
76. Newman D, Karimi S, Cavedon L. Using Topic Models to Interpret MEDLINE's Medical Subject Headings. In: Nicholson A, Li X, editors. AI 2009: Advances in Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. pp. 270–279. https://doi.org/10.1007/978-3-642-10439-8_28.
77. Markowitz JA, Terry Nutter J, Evens MW. Beyond is-a and part-whole: More semantic network links. Comput Math Appl. 1992;23(6):377–90. https://doi.org/10.1016/0898-1221(92)90113-V.
78. Abd Elrahman SM, Abraham A. A Review of Class Imbalance Problem. J Netw Innov Comput. 2013;1(2013):332–40. http://ias04.softcomputing.net/jnic2.pdf.

## Publisher's Note