# Bridging the Gap: A Hybrid Approach to Medical Relation Extraction Using Pretrained Language Models and Traditional Machine Learning

Nesma A. Hassan [1], Rania A. Abul Seoud [2], and Dina A. Salem [1,*]

[1] Department of Computer and Software Engineering, Faculty of Engineering, Misr University for Science and Technology, Giza, Egypt
[2] Department of Electrical Engineering, Faculty of Engineering, Fayoum University, Fayoum, Egypt
Email: nesma.abdelaziz@must.edu.eg (N.A.H.); raa00@fayoum.edu.eg (R.A.A.S.); dina.almahdy@must.edu.eg (D.A.S.)
*Corresponding author

***Abstract*—Feature engineering can be time-consuming and challenging, requiring expertise in Natural Language Processing (NLP) techniques and methods. The objective of this study was to explore the use of contextual word embeddings, specifically Bidirectional Encoder Representations from Transformers (BERT)-generated word embeddings, for biomedical relation extraction. The study utilized machine learning models, including Support Vector Machines, Random Forests, and K-nearest neighbor algorithms, to classify relationships between medical entities based on these embeddings. The attention mechanism of a pre-trained BERT model was also used to capture information related to the relationship between medical entities, leading to more advanced biomedical relation extraction. The performance of the machine learning classifiers was evaluated as classification models. The proposed approach outperformed the most recent state-of-the-art model on two publicly available biomedical relation extraction datasets Chemical-Protein Interactions (ChemProt) and Drug-Drug Interactions (DDI), indicating that traditional machine-learning techniques can compete with recent advancements. Experiments on the ChemProt dataset show that the performance of the proposed model's F1-Score is 0.778 and on the DDI dataset, F1-Score is 0.815. This study has demonstrated the potential for using contextual word embeddings and machine learning models for biomedical relation extraction, without the need for extensive manual feature engineering.**

***Keywords*—Bidirectional Encoder Representations from Transformers (BERT), hemical-Protein Interactions (ChemProt), Drug-Drug Interactions (DDI), K-Nearest Neighbor (KNN), Natural Language Processing (NLP), random forest, relation extraction, Support Vector Machine (SVM)**

## I. Introduction

Downstream applications in precision medicine greatly benefit from a comprehensive and accurate knowledge base. However, as the volume of biomedical literature expands rapidly, the disparity between curated information in existing knowledge bases and the available literature widens daily. Manual curation techniques, while still used to ensure data integrity, are expensive and impractical at scale. To expedite the curation process, an automated text mining-based method for relation extraction can extract relation candidates and present them to human curators for verification [1].

In biomedical relation extraction, the manual annotation of chemical and protein/gene entity mentions plays a crucial role in identifying and understanding relationships between these entities in textual data. The process involves human annotators carefully reading through the text and identifying specific mentions of chemicals, proteins, and genes. Annotating chemical entities involves identifying and marking the names of various chemical compounds, molecules, or substances mentioned in the text. This requires annotators to have a good understanding of chemical nomenclature and terminology. They need to recognize not only common chemical names but also synonyms, abbreviations, and variations in spelling or formatting.

On the other hand, annotating protein and gene entities involves identifying and marking the names of proteins and genes mentioned in the text. This requires annotators to be familiar with biological terminology, gene and protein names, and their various aliases. They need to recognize different naming conventions, such as official gene symbols, gene names, or protein names, as well as any variations or synonyms that might be used.

The manual annotation process typically involves annotators highlighting the relevant entity mentioned in the text and associating them with specific entity types, such as chemical, protein, or gene labels. Annotators must follow specific annotation guidelines provided by the project organizers to ensure consistency and accuracy across the annotations. After annotation, these labeled entity mentions serve as the foundation for training machine learning models that can automatically identify and extract chemical-protein/gene relationships from new, unseen text. The annotated data is used to create a labeled dataset for supervised learning, where the models learn

Manuscript received November 19, 2023; revised February 19, 2024; accepted March 14, 2024; published June 13, 2024.

from the annotated examples to generalize and predict relationships in new text.

Natural Language Processing (NLP) primarily focuses on automatic techniques for extracting essential information from unstructured text. Relation Extraction (RE) is a prominent NLP task used to identify and classify relationships between entities of interest. Biomedical relation extraction, a critical task in NLP, involves identifying complex relationships between entities in biomedical texts. For instance, extracting drug-drug interaction relationships or determining the proteins affected by specific chemicals can aid precision medicine, as demonstrated by ChemProt relation extraction. The relationships between entities such as genes, proteins, drugs, and diseases are context-dependent, necessitating a deep understanding of word meanings within specific contexts [2]. These entities exhibit diverse relationships depending on the context in which they appear. For example, the relationship between the gene "BRCA1" and the disease "breast cancer" may differ when discussed in the context of prevention versus treatment. Traditional methods for relation extraction relied on manual feature engineering and rule-based approaches, which were time-consuming, labor-intensive, and required domain-specific knowledge (see Fig. 1).

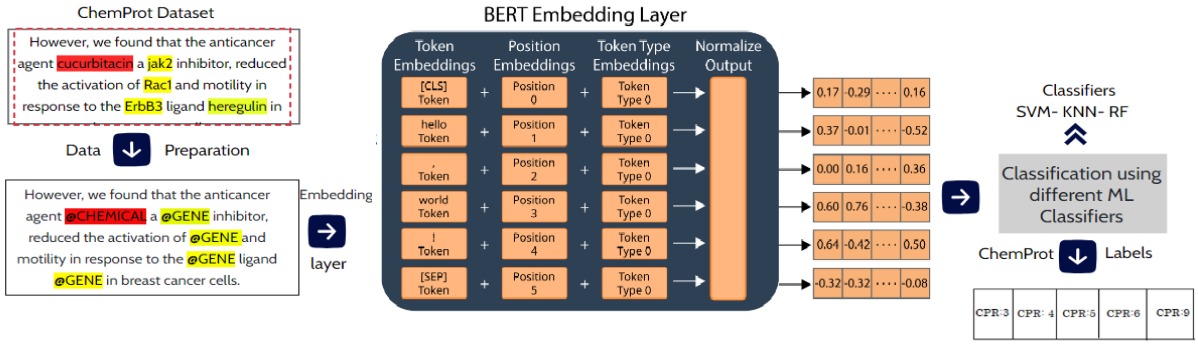


Fig. 1. Graphical abstract.

Most modern neural networks utilize word embeddings as their input layer. Word embeddings automatically capture semantic relationships between words and generate dense, low-dimensional vector representations that serve as inputs for machine learning models. Several word embedding methods, such as Word2Vec [3] and GloVe [4], have been proposed in the past decade. While these methods employ different architectures, they all train word embeddings based on word co-occurrence. Word2Vec predicts the current word given its context (Continuous Bag-Of-Words or Skip-Gram models), while GloVe leverages statistical information from a word-word co-occurrence matrix to consider local and global features of the corpus [5]. However, these word embeddings provide context-independent representations for words, resulting in two limitations [6]. First, word vectors from Word2Vec are static and fail to capture the various meanings of polysemous words accurately. Second, the feature extraction capability of the neural network employed affects the accuracy of relation extraction.

The limitation of context-free language representation lies in its inability to differentiate between words with distinct meanings. For example, a word like "bat" would have the same representation, regardless of whether it refers to a cricket bat or an animal. However, contextual models address this issue by providing a representation for each token based on the surrounding sentence, enabling a more comprehensive understanding of language by machines. The Bidirectional Encoder Representations from Transformers (BERT) model specifically excels in generating contextual representations for individual tokens, and it can even capture the context of entire sentences, sentence pairs, or paragraphs. This contextual awareness enhances the ability of the model to grasp nuanced meanings and relationships within the text, leading to more accurate and nuanced language understanding. While both Word2Vec and BERT aim to represent words as numerical vectors, they differ significantly in their underlying methodologies and resulting embeddings [7].

Word2Vec models produce context-independent embeddings, meaning each word has a single vector representation regardless of its different senses. In contrast, BERT embeddings are context-dependent, allowing multiple vector representations for the same word based on its contextual usage. Additionally, Word2Vec embeddings do not consider word position, whereas BERT explicitly incorporates word position information when calculating embeddings.

Word2Vec pre-trained embeddings can be directly utilized without the need for the model itself. These embeddings are available as a mapping between words and vectors. Conversely, BERT requires the trained model to generate contextual embeddings. The input to BERT is a sentence rather than a single word, as the model needs surrounding context to generate word vectors. The output of BERT is a fixed-length vector representation for the entire input sentence.

Recently, pre-trained language representation models such as Elmo [8] and BERT [9] have demonstrated their effectiveness in improving NLP tasks. Pre-training techniques like the Masked Language Model (MLM) and Next Sentence Prediction (NSP) allow these models to

extract semantic information from large amounts of unlabeled text. Unlike word embeddings, BERT is a contextualized neural embedding model that uses the contextual connections between words or sub-words to train embeddings. The input, a list of words, undergoes embedding into vectors (Input Embeddings), which are then processed by the Transformer Encoder [10]. The output consists of feature embeddings that represent the local context of corresponding words in the document. BERT-based models for biomedical relation extraction have shown excellent performance, achieving state-of-the-art results on various biomedical datasets [11–14]. However, BERT models require substantial computational resources and can be time-consuming. Processing data in parallel with multiple computer resources can handle more text, but accessing powerful GPUs or distributed computing environments is not always feasible. Consequently, there is a need for more precise and lightweight relation extraction models.

By combining machine learning classifiers with contextual embeddings, the accuracy of biomedical relation extraction can significantly improve. This approach enables models to capture the context-specific meanings of words and entities in text, leading to state-of-the-art performance in identifying drug-disease relationships, protein-protein interactions, and other entity relationships in biomedical texts. To address the aforementioned challenges, this paper proposes a novel method that leverages pre-trained transformer architectures specifically for generating input text embeddings. Unlike previous work, which employed the entire architecture, this approach focuses solely on embedding extraction. By doing so, it avoids the computational overhead associated with large parameter sets and time-consuming computations of transformers. Traditional machine learning classifiers, such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF), are utilized for classification purposes. The proposed model's performance is evaluated on two datasets: Chemical-Protein Interactions (ChemProt) and Drug-Drug Interactions (DDI). By combining transformer-based embeddings with traditional machine learning classifiers, the aim is to create a lightweight yet accurate model for biomedical relation extraction. This approach offers a more efficient alternative to full-scale BERT models, enabling faster inference and broader accessibility in resource-limited scenarios. The proposed approach offers a more efficient alternative to the full-scale BERT model, enabling faster inference and broader accessibility in resource-limited scenarios.

Traditional Machine Learning (ML) and Deep Learning (DL) algorithms have found extensive applications across various sectors, leveraging the power of data-driven analysis to make predictions, gain insights, and automate processes.

In the healthcare sector, ML/DL algorithms have been employed for medical image analysis, enabling the detection and diagnosis of diseases such as cancer, cardiovascular conditions, and neurological disorders. They have also been used for clinical decision support systems, helping healthcare professionals with treatment recommendations and patient risk assessment [15–17].

In the finance industry, ML/DL algorithms have been utilized for credit scoring, fraud detection, and algorithmic trading. These algorithms can analyze large volumes of financial data, identify patterns, and make accurate predictions about creditworthiness or detect suspicious transactions that may indicate fraudulent activity [18–20].

ML/DL algorithms have also found applications in the transportation sector, such as in traffic prediction, route optimization, and autonomous vehicles. By analyzing historical traffic data and real-time information, these algorithms can predict traffic patterns, suggest optimal routes, and enable autonomous vehicles to navigate safely [21, 22].

In the energy sector, ML/DL algorithms have been employed for energy demand forecasting, renewable energy optimization, and predictive maintenance of energy infrastructure. These algorithms can analyze historical energy consumption data, weather patterns, and operational parameters to make accurate predictions and optimize energy generation and distribution [23, 24].

Overall, traditional ML/DL algorithms have demonstrated their versatility and effectiveness across various sectors, enabling data-driven decision-making, process automation, and improved efficiency and accuracy in numerous applications.

The proposed research contributions can be described as follows:

- Innovative Use of Pre-trained Transformers: This paper introduces a novel technique that harnesses the capability of pre-trained transformer architectures, specifically to generate text input embeddings. Unlike previous studies, the proposed work focuses solely on embedding extraction, presenting an inventive application of transformer models in natural language processing.
- Computational Efficiency: The proposed method successfully mitigates the computationally intensive procedures associated with transformer models. By not employing the entire architecture, the proposed approach circumvents the massive parameter sets and computationally expensive processes typical of transformers, marking a significant step towards more resource-efficient models.
- Employment of Traditional Machine Learning Classifiers: We integrate established machine learning classifiers for classification needs. This integration of well-known classifiers such as SVM, KNN, and RF for multiclass classification, combined with modern transformer-based techniques, exhibits a unique blend of classic and contemporary methodologies.
- Broad Application and Testing: The performance of the proposed model has been rigorously evaluated on two distinct datasets–ChemProt and DDI. This comprehensive testing further underscores the

versatility and adaptability of the proposed methodology.

- Lightweight and Accurate Model: The primary objective is to amalgamate the benefits of transformer-based embeddings and traditional machine learning classifiers to create a lightweight yet precision-driven model for biomedical relation extraction. The creation of such a model can significantly advance biomedical research and applications.
- Resource-Limited Accessibility: The proposed technique presents a more efficient alternative to full-scale BERT models, enabling quicker inference times and broader accessibility. This factor is particularly impactful in resource-limited scenarios, pushing the boundaries of what's possible in such environments and thereby contributing to the democratization of machine learning applications.

The structure of the remaining sections of this paper is as follows: Section II provides a review of research papers relevant to this study. Section III outlines the design details of the proposed method, including implementation details and experimental settings. Subsequently, Section IV presents the evaluation results and discusses their implications. Finally, the conclusion of the findings is in Section V.

## II. Literature Review

Since pre-trained language models, deep learning techniques, and high-quality word representation (word embedding) have shown performance improvement, extracting various relations in the Biomedical Natural Language Processing (BioNLP) domain has received considerable interest.

Zhu *et al.* [25] employed drug descriptions from Wikipedia and DrugBank to augment the BERT model with semantic knowledge about drug entities. They utilized three types of entity-aware attention to generate sentence representations that incorporated entity information, mutual drug entity information, and drug entity information. The mutual information vector of two drug entities was computed by taking the difference between the BERT embeddings of the two drugs. To obtain drug description information, all drug description documents were processed by a Doc2Vec model to obtain vector representations for each drug entity in the 2013 DDI corpus. The resulting vectors for entity information were then fed into attention layers to retrieve sentence representation vectors that integrated multiple aspects of entity information.

Peng *et al.* [26] developed pre-trained BERT models specifically for biomedical literature, demonstrating their superior performance compared to other pre-trained language models on specific biomedical datasets. Another notable pre-trained BERT model, BioBERT was trained on large-scale biomedical corpora [27]. Li and Ji [28] incorporated GCNN into a BioBERT-based model to integrate dependency structure information. Su and Shanker [29] enhanced BioBERT by adding attention mechanisms in the final layer, leading to the best results on three biomedical extraction datasets. The attention layer outperformed LSTM in capturing important information from the last hidden state vectors.

Asada *et al.* [30] investigated how including diverse drug-related information affects DDI extraction and obtained an F-Score of 85.40, which they regarded as state-of-the-art performance. To construct their approach, they created embedding vectors for a Heterogeneous Knowledge Graph (HKG) of drugs by performing a link prediction task that predicts an entity in the PharmaHKG dataset. When processing the input sentence S, they used the BERT tokenizer to tokenize it into sub-word tokens and then added KG vectors of two drugs to extend it.

Wang *et al.* [31] proposed a BERT-based DDIs detection model and achieved a state-of-the-art result using evidence of supplement-drug interactions from scientific text. Mehryary *et al.* [32] combined a Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) to extract Chemical–Protein Interactions (CPIs) and achieved a high F-Score by a rich set of features. Warikoo *et al.* [33] exploited a set of linguistic features to train a tree kernel classifier to obtain CPIs from biomedical literature.

Generally, the above-mentioned previous methods depend heavily on feature engineering. Deep learning-based methods are another promising way to extract CPIs from biomedical literature. Because of their ability to automatically learn semantic and syntactic information, these methods no longer need to build sophisticated feature engineering or elaborate kernel functions and exhibit more excellent performance. For example, Peng *et al.* [34] proposed an ensemble method to combine three model predictions including Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) [35]. Lu *et al.* [36] proposed an RNN-based model integrating a granular attention mechanism to extract CPIs. Zhang *et al.* [37] proposed an RNN-based model that combines deep contextualized word representations [38] and the multi-head attention mechanism [10] for CPI extraction. Lee *et al.* [27] exploited large-scale biomedical corpora to pre-train Bidirectional Encoder Representations from Transformers (BERT) and classified Chemical-Protein Relation (CPR) types using BioBERT. Overall, deep learning-based methods perform better than traditional machine learning-based methods.

Lung *et al.* [39] extracted a diverse set of features from sentences to build multiple machine-learning models. Similarly, Corbett and Boyle [40] proposed a two-stage approach containing pre-trained LSTM and Bi-LSTM to extract CPIs. They utilized unlabeled data to pre-train word embedding and LSTM in the neural network. Verga *et al.* [41] used an efficient self-attention encoder to form pairwise predictions over entire paper abstracts. Liu *et al.* [42] synthesized GRU and attention pooling. Their experimental results show that the attention mechanism is effective in selecting important words. Peng *et al.* [34] applied Majority vote or stacking to combine the outputs of SVM, CNN, and RNN models.

Furthermore, Zhang *et al.* [37] used contextualized word representations and multi-head attention to learn the presentation for CPIs. They achieved a more competitive result.

During the evaluation of the ChemProt and DDI datasets, we observed a significant data imbalance that affected the performance of the model on the classification task [30, 43, 44]. This issue was present in both the training and test sets and can be attributed to the uneven distribution of classes in these datasets. The majority of previous research in this area did not focus on addressing this problem, with some approaches excluding negative classes from their classification or manually filtering negative samples that were larger than positive samples [45–47]. However, such methods can result in biased models that are not effective in predicting the minority class. First, merely looking at the positive vs. negative classes, only 17.5% of the drug-drug pairs in the DDI dataset and 33.8% of the chem-port pairs are in a positive class, which includes interaction between medical entities. This significant class imbalance can lead to a classifier that is biased towards the majority class(es), resulting in poor performance of the minority class. Therefore, to deal with this, we used a two-stage classification strategy that will be described in Section III.

## III. Materials and Methods

This research paper aims to enhance relation extraction in biomedical texts by combining contextual embedding with machine learning classifiers. By utilizing contextual embedding, the paper proposes a method to generate word embeddings that capture the contextual meaning of words and entities, enabling more accurate identification of complex entity relationships. Integration with machine learning classifiers such as support vector machines, random forest, and K Nearest Neighbor allows these models to achieve state-of-the-art performance in identifying entity relationships in biomedical texts.

This approach eliminates the need for manual feature engineering and provides a flexible and adaptable solution for relation extraction. Language models like BERT have the advantage of automatically learning representations from the text data itself. This allows them to capture both shallow and deep linguistic features, including contextual information, semantics, and syntactic structures. As a result, language models can often outperform traditional machine learning classifiers on certain natural language processing tasks, especially when large amounts of labeled data are available for pretraining and fine-tuning.

The proposed system follows a sequential pipeline, as depicted in Fig. 1, with several key steps. Initially, the input text undergoes pre-processing tasks like sentence segmentation, sentence tokenization, entity extraction, and masking with predefined tags to standardize the text.

In the feature extraction stage, the BERT architecture, specifically BERT, is employed to extract features and generate embeddings. BERT is chosen for its ability to capture contextualized embeddings that are challenging for traditional machine learning classifiers to extract. By leveraging BERT, the classifiers' performance is enhanced by incorporating high-quality contextual features. Notably, the embeddings are computed directly without requiring fine-tuning of the BERT model. The final step in the pipeline is the classification stage, where class imbalance is addressed through binary classification. The samples are classified into positive and negative classes to balance the distribution, and only the positively predicted samples proceed to the next stage. In the subsequent phase, multi-class classification is performed using SVM, KNN, and Random Forest classifiers. This enables evaluation and comparison of the system's performance across different classifiers, facilitating the selection of the most suitable classifier for relation extraction in biomedical texts. An Algorithm for the methodology described in the research paper is shown in Fig. 2.

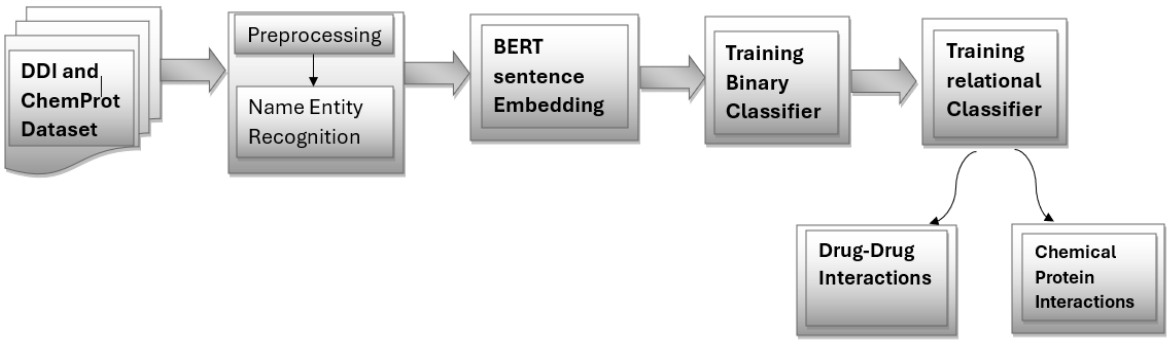


Fig. 2. Relational class extraction framework.

### A. Preprocessing

By following this approach, we achieve a comprehensive and effective methodology for relation extraction, encompassing pre-processing, feature extraction with BERT, and classification using multiple classifiers. This framework enables us to accurately classify relations in biomedical texts and assess the performance of different classifiers, facilitating informed decision-making for relation extraction tasks.

The proposed system's first step is the pre-processing stage, which involves reducing noise in the text and converting it into units that can be used for vectorization. The pre-processing stage includes sentence splitting to

analyze text at the sentence level, tokenization to break sentences into individual words or subwords, and stop word removal to improve text analysis efficiency. To perform Named Entity

Recognition (NER) in the biomedical domain, the scispaCy library [48] was utilized. This library is specifically designed for scientific, biomedical, and clinical text data. To facilitate relation extraction, we replaced the two named entity mentions of interest in the sentence with predefined tags such as @GENE$, @CHEMICAL$, @PROTEIN$, and @DRUG$. This standard pre-processing method enables the identification of chemical-gene relations.

In the example provided, the original sentence "Tomudex treatment increased Cyclin E" has been transformed using this pre-processing method. The chemical entity "Tomudex" and the gene entity "Cyclin E" have been replaced with the tags @CHEMICAL$ and @GENE$, respectively. The resulting sentence is "@CHEMICAL$ treatment resulted in an increase @GENE$". This pre-processed sentence can then be used to identify the chemical-gene relation between Tomudex and Cyclin E, which is that Tomudex treatment increases Cyclin E. By standardizing the pre-processing of text data using predefined tags, relation extraction and other NLP tasks become more efficient and accurate. This is because the tags provide a consistent and predictable format for the input data, making it easier for machine learning algorithms and other NLP tools to work with the data (see Fig. 3).

```
Start

Input: Biomedical text

Pre-processing:
- Sentence segmentation
- Sentence tokenization
- Entity extraction
- Masking with predefined tags

Feature Extraction:
- Utilize BERT architecture (specifically BERT) to extract features and generate embeddings
- Capture contextualized embeddings that are challenging for traditional machine learning cla
ssifiers to extract

Classification:
- Address class imbalance through binary classification
- Classify samples into positive and negative classes to balance the distribution
- Only positively predicted samples proceed to the next stage
- Perform multi-class classification using SVM, KNN, and Random Forest classifiers

Evaluation:
- Evaluate and compare system performance across different classifiers
- Select the most suitable classifier for relation extraction in biomedical texts

Output: Identified entity relationships in biomedical texts

End
```

Fig. 3. Algorithm for the methodology.

### *B. Feature Extraction*

After preprocessing the data, the subsequent step involves converting the tokens into features, resulting in a mathematical vector representation. This vector representation serves as input to the machine learning classifiers. BERT, a pre-trained deep learning model, utilizes a transformer architecture to encode the meaning of words within a sentence. In generating word embeddings, we employed BERT Base uncased [27], which allowed us to calculate the input vectors without fine-tuning directly. The extraction of BERT's embeddings was carried out on Google Colab with TPU (v2–8) acceleration. The process of creating word embeddings with BERT entails tokenizing the input text into individual words or subwords using the BERT tokenizer. Subsequently, the tokenized input is passed through the BERT model, resulting in a sequence of hidden states. These hidden states can be employed to generate word embeddings for each word in the input text by multiplying the hidden states with a learned weight matrix.

To generate word embeddings using BERT, the input text needs to be tokenized using WordPress. Additionally, Special Tokens Like (CLS) and Separator are added to denote the beginning and end of sentences and to differentiate between different segments of text. BERT also requires fixed-length input sequences, which may involve padding or truncation of the input text.

The BERT tokenizer operates in several steps to process the input text. Initially, the text undergoes basic tokenization, where it is divided into tokens based on simple rules. Punctuation marks become individual tokens, while words are split by whitespace. Subsequently, the basic tokens are further broken down into subword units using the WordPiece algorithm. This algorithm constructs a vocabulary of commonly occurring subwords from a large corpus, encompassing complete words and subwords. The tokenizer then searches for each basic token or subword in the vocabulary. Tokens found in the vocabulary remain

unchanged, while those absent are recursively divided into smaller subwords until they match vocabulary entries. For instance, "unhappiness" may be split into "un", "##hap", and "##piness," where "##" denotes subwords that are not standalone words.

Special tokens, such as CLS and SEP, are essential for BERT to mark sentence beginnings, endings, and segment distinctions in multiple sentences. These tokens are added accordingly to the tokenized input. BERT employs segment IDs to differentiate between sentences in the input. For tasks involving a single sentence, all tokens receive the same segment ID. However, when multiple sentences are present, segment IDs are assigned to distinguish between them. To ensure consistent input length, the tokenizer may apply padding or truncation to the tokenized sequence. Padding involves adding special tokens like a token named "PAD" to make all sequences equal in length, while truncation involves removing tokens to satisfy a specified maximum length.

Once these steps are completed, the input text is transformed into a sequence of tokens, encompassing special tokens and potentially subwords. This tokenized sequence can then be fed into the BERT model to generate word embeddings or perform various natural language processing tasks. Finally, the input text needs to be converted into numerical embeddings to be used as input for the classification model. By incorporating these additional preprocessing steps, the input text can be appropriately formatted for BERT, enabling effective feature extraction and embedding generation from the text.

BERT word embeddings offer the advantage of contextual awareness, meaning that the embedding of a word can vary depending on its specific context. In contrast, many other word embedding methods generate a fixed embedding for each word, regardless of context.

### C. Classification

For performing the classification step, In the first stage, we trained a binary SVM classifier to classify samples of both datasets into positive and negative classes. Then, in the second stage, we considered only instances that were classified as positive by the first classifier, and classified them into one of the multi-classes, using a multi-class classifier. A group of machine learning classifiers is used, and the classifiers used to categorize the labels in the dataset were imported from the sci-kit-learn packages [49].

#### 1) Binary classification

Class imbalance is a common issue in machine learning where one class has a significantly larger number of samples than the other. This can lead to biased classifiers that perform better in the majority class and worse in the minority class. It was observed that ChemProt and DDI [50] datasets suffer from class imbalance as shown in Fig. 4. Zhao *et al*. [51] created a more balanced dataset by eliminating surplus negative instances from the SemEval 2013 DDI Extraction dataset through the application of predefined rules. To tackle the significant data imbalance in the proposed study, we opted to employ a binary classifier to categorize samples into positive and negative classes before the implementation of the multiclass classifier. This approach eliminated the necessity of creating specific rules and effectively balanced the sample distribution between the two classes, potentially enhancing the performance of the multiclass classifier. While several classifiers were explored for binary classification, SVM yielded the best results.

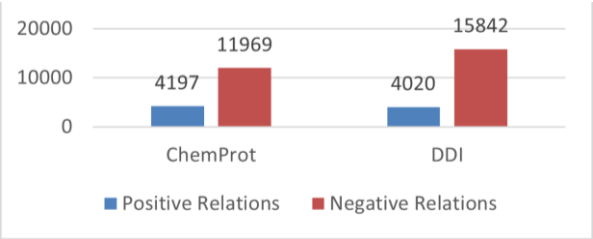


Fig. 4. ChemProt and DDI class imbalance.

#### 2) Multi-class classification

Machine learning classifiers are essential for the second stage of the classification pipeline as they are used to perform multiclass classification. In this stage, the extracted embeddings from BERT are fed into the classifiers to enable the classification of the samples into their respective classes. To achieve this, we evaluated the performance of three popular classifiers: Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Random Forest. SVM is a powerful and widely used classifier that effectively handles high-dimensional data and is suitable for both linear and nonlinear classification problems. KNN is another popular classifier that is useful in identifying patterns in the dataset and classifying samples based on their proximity to other samples. Random Forest is an ensemble learning method that creates multiple decision trees and then aggregates their predictions to make a final decision. By exploring the performance of these three classifiers, we can identify the most suitable and accurate classifier for the given dataset and classification task.

### D. Hyperparameter Tunning

Tables I and II display the parameters that were chosen for the classification process. The classifiers utilized to perform the classification of biomedical entities were imported from sci-kit-learn libraries. After conducting various experiments and testing multiple parameters using GridSearchCV, the optimal parameter values were determined to be the ones mentioned.

TABLE I. PARAMETERS OF CHEMPROT DATASET FOR MULTICLASS CLASSIFICATION

| Model | Parameters | Selected Values |
|---|---|---|
| Random Forest | n_estimators | 20 |
| | Max_depth | 40 |
| | Min_samples_leaf | 1 |
| | Bootstrap | TRUE |
| | Criterion | Gini |
| SVM | C | 10 |
| | Kernel | Poly |
| KNN | N_neighbours | 20 |

TABLE II. PARAMETERS OF DDI DATASET FOR MULTICLASS CLASSIFICATION

| Model | Parameters | Selected Values |
|---|---|---|
| Random Forest | n_estimators | 10 |
| | Max_depth | 30 |
| | Min_samples_leaf | 4 |
| | Bootstrap | FALSE |
| | Criterion | Gini |
| SVM | C | 10 |
| | Kernel | Poly |
| KNN | N_neighbours | 20 |

## IV. RESULTS AND DISCUSSION

This section aims to analyze the performance of the classifiers using test data. We outline the experimental setup and the evaluation metric.

### A. Dataset

The evaluation of the proposed model is performed on two datasets, which are the ChemProt and DDI datasets. These datasets are specifically designed for Relation Extraction (RE) tasks. The ChemProt track corpus in BioCreative VI is used to annotate interactions and explore the recognition of chemical-protein relations from abstracts. The corpus contains directed relations from chemicals/drugs to genes/proteins, indicating how the chemical/drug interacts with the gene/protein. These chemical-protein relations, referred to as "CPR", are classified into ten semantically related classes based on their underlying biological characteristics. For example, interactions that increase the activity or expression of a target gene or protein, such as "activator", "indirect upregulation", and "upregulation", belong to the CPR:3 group, while interactions that decrease the activity or expression of a target gene or protein, such as "downregulation", "indirect downregulation", and "inhibitor", belong to the CPR:4 group. In this task, only relations belonging to CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9 was considered for evaluation purposes, and both chemical and protein/gene entity mentions were manually annotated. Table III provides information on the statistics of the ChemProt corpus.

In the DDIExtraction 2013 shared task, five types of interactions were annotated, and false pairs, which are drug pairs that do not interact. DDI corpus of the 2013 DDI extraction challenge, which is made up of 175 MEDLINE abstracts about DDIs and 730 Drug Bank data. The corpus is divided into two sets: a training set, which includes 572 Drug Bank documents and 142 MEDLINE abstracts, and a test set, which includes 158 Drug Bank documents and 33 MEDLINE abstracts.

TABLE III. STATISTICS OF CHEMPROT CORPUS

| Dataset Classes | Training | Testing |
|---|---|---|
| CPR:3 | 784 | 667 |
| CPR:4 | 2278 | 1667 |
| CPR:5 | 173 | 198 |
| CPR:6 | 235 | 293 |
| CPR:9 | 727 | 644 |
| False | 11969 | 10540 |

Each sentence has annotated lists of all the medications and drug pairs. The following four types of drug-drug interactions (DDIs) are identified among the drug pairs (33508): Advice, Effect, Mechanism, and Int.

- When DDI's pharmacokinetic mechanism is described, a mechanism is assigned.
- When a DDI's effect is described, an effect is assigned.
- When a recommendation or piece of advice regarding a DDI is offered, advice is assigned.
- When a DDI happens but the sentence doesn't give any further details about it, it is allocated.

Table IV provides information on the statistics of the DDI corpus.

TABLE IV. STATISTICS OF DDI CORPUS

| Dataset Classes | Training | Testing |
|---|---|---|
| Advice | 826 | 218 |
| Effect | 1,687 | 356 |
| Mechanism | 1,319 | 302 |
| Int | 188 | 96 |
| false | 15,842 | 4,782 |

### B. Dataset Evaluation

In this section, we analyze the performance of the machine learning classifiers using test data on two benchmark datasets: ChemProt and DDI. The goal of this analysis is to evaluate the effectiveness of the proposed models for label categorization tasks on these datasets. To conduct the evaluation, we first trained the machine learning classifiers on the training data of each dataset. We then tested the classifiers on the test data and recorded their precision, recall, and F1-Score. We repeated this process for each classifier and each dataset. To address the issue of data imbalance in the datasets, we conducted binary classification experiments and compared the performance of several classifiers, Fig. 5. showed that SVM performed better than other tested classifiers. which shows the F1-Score of different classifiers' performance on the two datasets ChemProt and DDI.

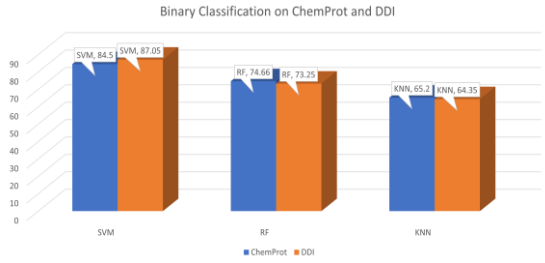


Fig. 5. Binary classification on ChemProt and DDI datasets.

Based on the experimental results presented in Figs. 6 and 7, it is found that the SVM classifier outperformed the other classifiers in terms of precision, recall, and F1-Score. To further evaluate the performance of the SVM classifier, we analyzed its performance on each class of the ChemProt and DDI datasets. The results of this analysis are presented in Tables V and VI, which show

the precision, recall, and F1-Score of the SVM classifier for each class. As expected, the SVM classifier yielded the best results among the classifiers for each class, demonstrating its effectiveness for label categorization tasks.

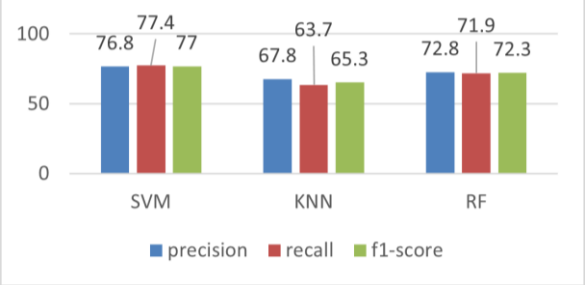


Fig. 6. Chemprot dataset evaluation.

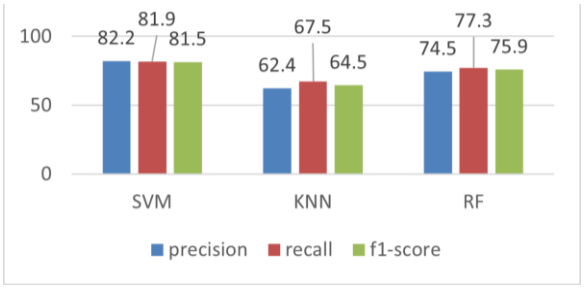


Fig. 7. DDI dataset evaluation.

TABLE V. SVM PERFORMANCE ON EACH CLASS FOR CHEMPROT DATASET

| Classes | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| CPR:3 | 0.749 | 0.747 | 0.748 | 665 |
| CPR:4 | 0.653 | 0.677 | 0.660 | 1661 |
| CPR:5 | 0.733 | 0.806 | 0.768 | 195 |
| CPR:6 | 0.825 | 0.835 | 0.830 | 293 |
| CPR:9 | 0.850 | 0.789 | 0.819 | 644 |
| Weighted Avg | 0.768 | 0.774 | 0.778 | 3458 |

TABLE VI. SVM PERFORMANCE ON EACH CLASS FOR DDI DATASET

| Classes | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Advice | 0.842 | 0.854 | 0.848 | 235 |
| Effect | 0.770 | 0.872 | 0.818 | 360 |
| INT | 0.847 | 0.462 | 0.605 | 95 |
| Mech | 0.862 | 0.843 | 0.853 | 303 |
| Weighted Avg | 0.822 | 0.819 | 0.815 | 993 |

Fig. 8 shows that the confusion matrix of the SVC model is performing very well in the CPR:4 class. However, the model is making more mistakes in the other classes. For example, the model sometimes predicts that samples belong to the class CPR:5 when their true label is CPR:3 or CPR:6. The model makes more mistakes in the CPR:3 and CPR:5 classes than in the other classes. This could be due to several factors, such as the classes being more similar to each other or the model having less training data for these classes.

In Fig. 9 the confusion matrix shows that the SVC model is performing well overall, with an accuracy of 95%. However, there are some areas where the model is making mistakes. For example, the model sometimes predicts that samples belong to the class DDI-advise when their true label is DDI-effect or DDI-int. The model is performing very well in the DDI-effect and DDI-mechanism classes. This could be due to the classes being very distinct from the other classes or the model having a lot of training data for these classes.

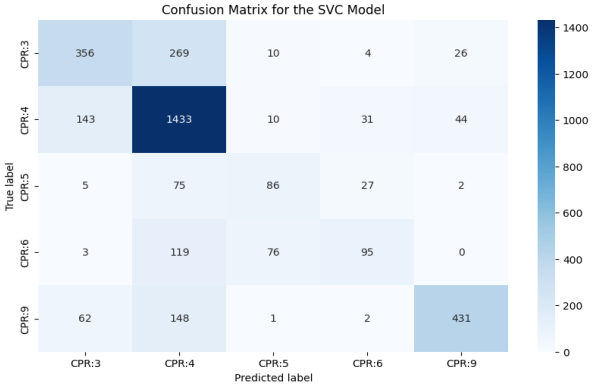


Fig. 8. Confusion matrix of SVM model on ChemProt dataset.

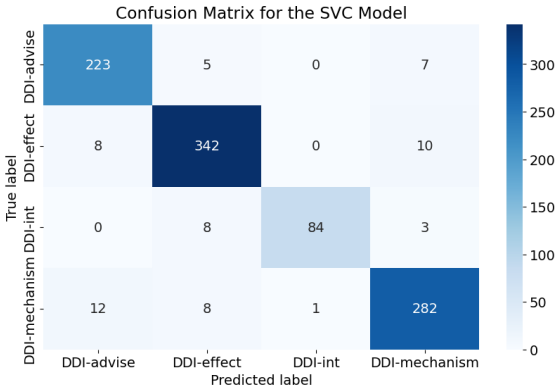


Fig. 9. Confusion matrix of SVM model on DDI dataset.

### C. Comparison with Previous State-of-the-Art Models

Table VII presents a comprehensive comparison of the performance of the proposed model with previous systems regarding Drug-Drug Interaction (DDI) and chemical-protein (ChemProt) relation extraction. The table provides an overview of how the proposed model performs in comparison to the existing systems in this field. It allows us to compare it with different models and systems in terms of various metrics such as precision, recall, F1-Score, and others.

TABLE VII. COMPARISON WITH PREVIOUS SOTA MODELS

| Method | Chemprot F1-Score | Method | DDI F1-Score |
|---|---|---|---|
| Peng *et al.* [34] | 64.10 | Liu *et al.* [52] | 73.7 |
| Corbet and Boyle [40] | 62.58 | Sahu and Anand [53] | 68.6 |
| Verga *et al.* [41] | 50.8 | Zhang *et al.* [54] | 72.9 |
| Biobert [27] | 73.51 | Luo *et al.* [55] | 75.1 |
| Proposed Method | 77.8 | Proposed Method | 81.5 |

## V. CONCLUSION

In this paper, we propose a novel model for text classification in the biomedical domain. The proposed

model combines a pre-trained language model (BERT) with traditional machine learning classifiers. BERT is used to generate embeddings of the input text, which are then passed to the machine learning classifiers for classification. We observed that collecting, aggregating, and enhancing the essential elements of the input text that were mapped to word embeddings led to an increase in the accuracy of the classification task. This is because BERT can learn the semantic relationships between words in a corpus of biomedical text, which helps to improve the representation of the input text. Furthermore, after analyzing the two datasets there is a prevalent issue of class imbalance. Unlike previous approaches that involved manually curating the dataset by removing repeated relations or similar entity names, we opted to grant the classifier complete autonomy in filtering out negative sentences. This decision was based on the observation that allowing the classifier to make such determinations yielded improved performance compared to prior studies and outcomes. In addition to the aforementioned improvements, the proposed model offers several advantages. Firstly, it demonstrates enhanced efficiency compared to previous models that solely rely on BERT. Since machine learning classifiers are used solely for text classification, they do not need to learn the semantic relationships between words, leading to improved computational efficiency. Secondly, it exhibits greater flexibility compared to previous models that solely employ traditional machine learning classifiers. BioBERT's ability to generate embeddings for any text, regardless of the domain, makes it adaptable to various contexts. Lastly, the model showcases increased robustness in comparison to previous models that solely utilize BERT. The machine learning classifiers serve as a means to compensate for any errors made by BERT, enhancing the overall resilience of the system.

In summary, the proposed model introduces a novel approach to text classification in the biomedical domain, combining BERT's language modeling capabilities with traditional machine learning classifiers. The improvements achieved, along with the discussed advantages, contribute to the advancement of text classification methodologies in the biomedical field. It plays a crucial role in biomedical literature analysis by categorizing and organizing scientific articles, enabling efficient navigation, and staying updated with advancements. This improves literature review processes, and decision-making, and facilitates the development of therapies and medical guidelines. In clinical decision support systems, text classification extracts valuable information from medical texts, aiding healthcare professionals in making informed decisions, enhancing patient care, and improving outcomes. It also assists in identifying adverse drug reactions and improving medication management and patient safety.

Text classification is vital in drug discovery and pharmacovigilance, analyzing biomedical literature, clinical trial data, and adverse event reports. It identifies drug targets, predicts efficacy, and detects adverse reactions, accelerating drug discovery, improving safety monitoring, and enhancing pharmacovigilance efforts. In public health, text classification monitors disease outbreaks, detects emerging infectious diseases, and tracks public sentiments, aiding in early detection, response planning, and targeted interventions.

Overall, text classification enhances literature analysis, supports clinical decision-making, accelerates drug discovery, improves pharmacovigilance, and aids in public health monitoring. Its application advances biomedical knowledge and improves patient care through valuable insights, increased efficiency, and informed decision-making.

The proposed model has the potential to be used in a variety of biomedical applications, such as drug discovery, clinical decision support, and natural language processing.

This paper presents a novel model for text classification in the biomedical domain. The model combines a pre-trained language model called BERT with traditional machine learning classifiers. BERT is used to generate word embeddings for the input text, which are then used by the machine learning classifiers for classification. The researchers found that by carefully collecting and enhancing the important elements of the input text, they were able to improve the accuracy of the classification task. BERT's ability to learn semantic relationships between words in biomedical text helps improve the representation of the input text.

Furthermore, after analyzing the two datasets there is a prevalent issue of class imbalance. Unlike previous approaches that involved manually curating the dataset by removing repeated relations or similar entity names, we opted to grant the classifier complete autonomy in filtering out negative sentences. This decision was based on the observation that allowing the classifier to make such determinations yielded improved performance compared to prior studies and outcomes. The proposed model offers several advantages over previous models. Firstly, it is more computationally efficient since the machine learning classifiers only focus on text classification and don't need to learn semantic relationships between words. Secondly, it is more flexible as it can generate embeddings for any text, regardless of the domain. Lastly, the model is more robust as the machine learning classifiers help compensate for any errors made by BERT, improving the overall resilience of the system.

In summary, this novel model combining BERT and machine learning classifiers improves text classification in the biomedical domain. It has various advantages and contributes to advancements in the field. Text classification plays a crucial role in biomedical literature analysis, clinical decision support, drug discovery, pharmacovigilance, and public health monitoring. The proposed model has the potential to be applied in several biomedical applications, such as drug discovery, clinical decision support, and natural language processing.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

Nesma Abdelaziz conducted the research, collected the needed data, and analyzed the collected data; Dina Salem and Nesma Abdel Aziz implemented the proposed modules, interpreted the results, and wrote the paper. Rania Abulseoud provided invaluable insights from her extensive experience in the field. All authors revised the paper, worked on the software enhancement; all authors had approved the final version.

## REFERENCES


[1] H. Lu *et al*., "Extracting chemical–protein interactions from biomedical literature via granular attention-based recurrent neural networks," *Computer Methods Programs Biomed*., vol. 176, pp. 61–68, 2019.

[2] H. Kilicoglu, "Biomedical text mining for research rigor and integrity: Tasks, challenges, directions Briefings," *Bioinformat.*, vol. 19, pp. 1400–414, 2018.

[3] K. W. Church, "word2vec. nat," *Lang. Eng*., vol. 23, pp. 155–162, 2016.

[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. International Conference on Learning Representations*, 2013, pp. 2–4.

[6] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proc. the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 298–307.

[7] C. C. Wang, P. Nulty, and D. Lillis, "A comparative study on word embeddings in deep learning for text classification," in *Proc. 4th International Conference on Natural Language Processing and Information Retrieval (NLPIR '20)*, New York, NY, USA, 2021, pp. 37–46.

[8] M. Peters, "Deep contextualized word representations, for natural language understanding," in *Proc. 2018 EMNLP Workshop BlackboxNLP: An Alyzing and Interpreting Neural Networks for NLP*, 2018, pp. 353–355.

[9] U. Kamath, K. L. Graham, and W. Emara, "Bidirectional encoder representations from transformers (Bert)," *Transformers for Machine Learning*, pp. 43–70, 2022. doi: 10.1201/9781003170082-3

[10] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[11] P. Su and K. V. Shanker, "Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction," *BMC Bioinformatics*, vol. 23, no. 120, 2022. https://doi.org/10.1186/s12859-022-04642-w

[12] S. S. Hiai *et al*., "Relation extraction using multiple pre-training models in biomedical domain," in *Proc. International Conference on Recent Advances in Natural Language Processing*, 2021, pp. 530–537.

[13] H. Fabregat *et al.*, "Negation-based transfer learning for improving biomedical named entity recognition and relation extraction," *Journal of Biomedical Informatics*, 138, 2023.

[14] M. K. Kang and A. Hendawi, "Interaction extraction from biomedical text using relation bioBERT with BLSTM," *Learn. Knowl. Extr*., vol. 5, 2023.

[15] J. Kumari, E. Kumar, and D. Kumar, "A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics," *Arch Computat Methods Eng*., vol. 30, pp. 3673–3701, 2023. https://doi.org/10.1007/s11831-023-09915-y

[16] A. V. L. N. Sujith *et al.*, "Systematic review of smart health monitoring using deep learning and artificial intelligence," *Neuroscience Informatics*, vol. 2, no. 3, 2022.

[17] Y. W. Chen and L. C. Jain, "Deep learning in healthcare," *Paradigms and Applications*, vol. 3, 2020.

[18] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications: A survey," *Applied Soft Computing*, vol. 93, 2020.

[19] J. Huang, J. Chai, and S. Cho, "Deep learning in finance and banking: A literature review and classification," *Frontiers of Business Research in China*, vol. 14, no. 1, pp. 1–24, 2020.

[20] S. Singh *et al*., "Transforming the financial industry through machine and deep learning innovations," in *Proc. International Conference on Intelligent Computing and Optimization*, 2023, pp. 167–176.

[21] H. Nguyen *et al*., "Deep learning methods in transportation domain: A review," *IET Intelligent Transport Systems*, vol. 12, no. 9, pp. 998–1004, 2018.

[22] J. G. Ibañez *et al*., "Deep learning support for intelligent transportation systems," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 3, 2021.

[23] R. O. Ajayi *et al*., "Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques," *Journal of Building Engineering*, vol. 45, 103406, 2020.

[24] M. M. Forootan *et al*., "Machine learning and deep learning in energy systems: A review," *Sustainability*, vol. 14, no. 8, 4832, 2020.

[25] Y. Zhu *et al*., "Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions," *Journal of Biomedical Informatics*, vol. 106, 103451, Jun. 2020. https://doi.org/10.1016/j.jbi.2020.103451

[26] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," *ACLWeb*, vol. 1, 2019.

[27] J. Lee *et al*., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, Sep. 2019. https://doi.org/10.1093/bioinformatics/btz682

[28] D. Li and H. Ji, "Syntax-aware multi-task graph convolutional networks for biomedical relation extraction," in *Proc. Tenth International Workshop on Health Text Mining*, 2019, pp. 28–33. https://doi.org/10.18653/v1/d19-6204

[29] P. Su and K. V. Shanker, "Investigation of BERT model on biomedical relation extraction based on revised fine-tuning mechanism," in *Proc. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 16–19. https://doi.org/10.1109/bibm49941.2020.9313160

[30] M. Asada, M. Miwa and Y. Sasaki, "Integrating heterogeneous knowledge graphs into drug–drug interaction extraction from the literature," *Bioinformatics*, vol. 39, 2022.

[31] L. L. Wang *et al*., "Extracting evidence of supplement-drug interactions from literature," *Computation and Language*, vol. 7, Sep. 2019.

[32] F. Mehryary, J. Bjorne, T. Salakoski, and F. Ginter, "Potent pairing: ensemble of long short-term memory networks and support vector machine for chemical-protein relation extraction," *Database*, vol. 2018, Jan. 2018. https://doi.org/10.1093/database/bay120

[33] N. Warikoo, Y. C. Chang, and W. L. Hsu, "LPTK: A linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task," *Database*, vol. 2, Jan. 2018. https://doi.org/10.1093/database/bay108

[34] Y. Peng, A. Rios, R. Kavuluru, and Z. Lu, "Extracting chemical–protein relations with ensembles of SVM and deep learning models," *Database*, vol. 2, Jan. 2018. https://doi.org/10.1093/database/bay073

[35] J. Zheng and L. Zheng, "A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification," *IEEE Access*, vol. 7, pp. 106673–106685, 2019. https://doi.org/10.1109/access.2019.2932619

[36] H. Lu, L. Li, X. He, Y. Liu, and A. Zhou, "Extracting chemical-protein interactions from biomedical literature via granular attention based recurrent neural networks," *Computer Methods and Programs in Biomedicine*, vol. 176, pp. 61–68, Jul. 2019. https://doi.org/10.1016/j.cmpb.2019.04.020

[37] Y. Zhang, H. Lin, Z. Yang, J. Wang, and Y. Sun, "Chemical–protein interaction extraction via contextualized word representations and multihead attention," *Database*, vol. 2, Jan. 2019. https://doi.org/10.1093/database/baz054

[38] G. Lu, J. Gan, J. Yin, Z. Luo, B. Li, and X. Zhao, "Multi-task learning using a hybrid representation for text classification," *Neural Computing and Applications*, vol. 12, 2018. https://doi.org/10.1007/s00521-018-3934-y

[39] P. Y. Lung, Z. He, T. Zhao, D. Yu, and J. Zhang, "Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering," *Database*, vol. 9, Jan. 2019. https://doi.org/10.1093/database/bay138

[40] P. Corbett and J. Boyle, "Improving the learning of chemical-protein interactions from literature using transfer learning and specialized word embeddings," *Database*, vol. 2, Jan. 2018. https://doi.org/10.1093/database/bay066

[41] P. Verga, E. Strubell, and A. McCallum, "Simultaneously self-attending to all mentions for full-abstract biological relation extraction," arXiv preprint, arXiv:1802.10569, Jan. 2018.

[42] S. Liu *et al.*, "Extracting chemical–protein relations using attention-based neural networks," *Database*, vol. 2, Jan. 2018. https://doi.org/10.1093/database/bay102

[43] M. Faisal and A. Lavelli, "Impact of less skewed distributions on efficiency and effectiveness of biomedical relation extraction," *International Conference on Computational Linguistics*, pp. 205–216, Dec. 2012.

[44] A. Sun, R. Grishman, and S. Sekine, "Semi-supervised relation extraction with large-scale word clustering," in *Proc. the Meeting of the Association for Computational Linguistics*, Jun. 2011, pp. 521–529.

[45] C. Sun *et al*., "A deep learning approach with deep Contextualized word representations for chemical–protein interaction extraction from biomedical literature," *IEEE Access*, vol. 7, pp. 151034–151046, 2019. https://doi.org/10.1109/access.2019.2948155

[46] S. Liu, B. Tang, Q. Chen, and X. Wang, "Drug-drug interaction extraction via convolutional neural networks," *Computational and Mathematical Methods in Medicine*, vol. 2016, pp. 1–8, 2016. https://doi.org/10.1155/2016/6918381

[47] C. Sun *et al*., "Chemical–protein interaction extraction via Gaussian probability distribution and external biomedical knowledge," *Bioinformatics*, vol. 36, no. 15, pp. 4323–4330, May 2020. https://doi.org/10.1093/bioinformatics/btaa491

[48] A. Neumann, "ScispaCy: Fast and robust models for biomedical natural language processing," in *Proc. 18th BioNLP Workshop and Shared Task*, 2019, pp. 319–327.

[49] Scikit-learn: Machine learning in python. [Online]. Available: https://scikit-learn.org/stable/

[50] M. H. Zazo, "The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions," *Journal of Biomedical Informatics*, pp. 914–920, 2013.

[51] Z. Zhao *et al.*, "Drug drug interaction extraction from biomedical literature using syntax convolutional neural network," *Bioinformatics*, vol. 32, pp. 3444–3453, 2016.

[52] X. Liu, "Drug-drug interaction extraction via convolutional Neural networks," *Comput. Math*., pp. 1–9, 2018.

[53] S. Sahu and A. Anand, "Drug-drug interaction extraction from biomedical texts using long short-term memory network," *J. Biomed. Inform.*, vol. 86, pp. 15–24, 2018.

[54] Y. Zhang, "Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths," *Bioinformatics*, pp. 828–835, 2018.

[55] L. Luo, "A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature," *J. Biomed Inform*., vol. 103, 2020.