International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Novel Approach for Biomedical Text Classification Using Deep Learning and NLP for Disease Prediction

Greeshma G. S.[1], Dr. Bindiya Ahuja[2], Harshita Samota [3], Dr. K. Vanitha[4], Alok Dubey[5], Sheetal Mujoo[6], Saneesh P. S.[7]

Submitted: 17/01/2024 **Revised**: 25/02/2024 **Accepted**: 03/03/2024

Abstract: Biomedical text classification is crucial for automating the analysis of vast biomedical literature to aid in disease prediction, given the exponential growth of biomedical data. Integrating deep learning methods with natural language processing (NLP) has revolutionized this field, offering unprecedented capabilities in understanding and extracting intricate patterns from text data coupled with advanced NLP techniques, enable researchers to uncover hidden associations between biomedical concepts, identify novel biomarkers, and enhance disease prediction accuracy. In this study, we investigate the application of deep learning and NLP for biomedical text classification, presenting a novel framework that harnesses deep neural networks to capture semantic relationships and domain-specific knowledge. Through extensive experimentation on benchmark datasets, we demonstrate the effectiveness of our approach compared to traditional machine learning methods. Our research contributes to advancing biomedical text classification, highlighting the transformative potential of deep learning and NLP in healthcare research and practice.

Keywords*: Biomedical text classification, Deep learning, NLP, Disease prediction*

1. Introduction

As a result of advancements in information and Internet technologies, the total quantity of electronic documents available online has grown exponentially. This massive and unstructured amount of text greatly facilitates automatic text classification [1]. Natural language processing, One of the most crucial areas, categorization aids in the process of assigning text documents to appropriate classifications based on their content. Website categorization, unstructured text categorization, sentiment categorization, spam email filtering, and author identification are the primary uses of the publicly accessible papers, which display several problems and solutions [2]. When using the most often used bag-of-words method, supervised classification algorithms. It may have a sparsity problem due to the short training data and the ease with which some phrases can be ignored. Consequently, increasingly complicated traits have been the focus of contemporary research. Since medical records and literature make up a large portion of medical text, it is naturally the most heavily targeted subfield of text categorization [4]. It also includes the patient's complete medical history and the effects of any prescribed medication. The medical literature contains both old and new accounts of the medical techniques used to diagnose and treat a specific disease [5]. There is a wealth of medical-related data available for data mining because to the proliferation of electronic medical records and related literature made possible by the rise of the internet. Two major problems make text categorization in the medical sector very difficult: first, there are many medical techniques described in the text, and second, there are a few grammatical errors [6] have found widespread usage in image and signal processing, and now medical text categorization is no exception [7].

Some works allow for the classification of medical data on word, phrase, and document levels [8]. You may find a wealth of medical information on diseases, symptoms, treatments, patients' histories, medications, and more on the internet. Sorting them into their appropriate categories will allow them to absorb the most relevant data. This assignment paves the way for subsequent implementation, such the building of an automated medical diagnosing tool and categorization.

1Assistant Professor, Department of Computer science and Engineering
Email: greeshma.s@galgotiasuniversity.edu.in
2Professor, Department CSE, Lingaya's Vidyapeeth
Email: Bindiya.bhatia@gmail.com
3Assistant professor, panipat institute of engineering and technology, Samalkha, Haryana, India
Email: harshitasamota@gmail.com
4Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering,
Karpagam Academy of Higher Education(Deemed to be University) Coimbatore, India
Email: profvanithacse@gmail.com
5Associate professor, Department of Preventive Dental Sciences, College of Dentistry, Jazan University, Jazan, Saudi Arabia.
Email: dentaalok@yahoo.com
6Assistant professor, Department of Maxillofacial Surgery and Diagnostic Sciences, College of Dentistry, Jazan University, Jazan, Saudi Arabia.
Email: sheetalmujoo@yahoo.co.uk
7Former Assistant Professor, Department: Computer Science and Engineering, Sreepathy Institute of Management and Technology, Kerala
Email: saneeshputhurath@gmail.com

Nevertheless, this study discusses a few crucial studies that address medical text categorization. Hughes et al.[10] did a well-known piece of work in medical text classification; practically all researchers in the field acknowledge it. In it, they employed more complicated techniques to express the classification features using CNN. Research trends in clinical text categorization were the exclusive focus of Mujtaba et al.'s extensive analysis of a systematic literature review [11 presented a CNN-based cancer signature text categorization after extensively investigating medical datasets. Some other noteworthy works in medical text classification include the following: integrating neural networks with the technique of attentive rule construction. Some examples of medical text classification applications in the field of general health technology include an instrument based on natural language processing (NLP).

2. Literature Survey

1. In this paper, the authors [1] implemented and evaluated CNN architectures on biomedical text data to automatically categorize articles into relevant topics or classes. The findings probably indicate the effectiveness of CNNs in handling the complexities of biomedical text and their potential to improve the efficiency of indexing tasks in the biomedical domain. The research may have highlighted the advantages of using deep learning methods, particularly CNNs, for processing and organizing large volumes of biomedical literature, contributing to advancements.

2. In this study, the authors [2] likely conducted a performance evaluation of deep learning algorithms for biomedical document classification. They probably assessed the effectiveness of various deep learning techniques in categorizing biomedical documents into relevant classes or topic on a dataset consisting of biomedical documents. The findings likely provided insights into the strengths and weaknesses of different deep learning approaches for handling biomedical text data. This research could have contributed to advancing the field of biomedical informatics by identifying.

3. In this paper, the authors [3] likely developed a methodology where regular expressions are automatically generated to extract relevant features from biomedical text data. These regular expressions might have been used to identify patterns or keywords indicative of specific biomedical concepts or categories. The active learning framework likely involved iteratively selecting the most informative instances for annotation to train the classifier more efficiently. The research probably demonstrated the effectiveness of this approach in improving the performance of biomedical text classification models while reducing the annotation effort required. The findings could have implications for enhancing the scalability and accuracy of text classification systems in biomedical informatics, ultimately facilitating knowledge discovery from large volumes of biomedical literature.

4. In this study, the authors [4] likely focused on constructing. They probably developed and evaluated deep learning architectures tailored to automatically categorize evidence types or levels within biomedical articles. The research might have involved preprocessing the text data, designing neural network architectures, and training the models on a dataset comprising open access biomedical literature. The findings likely demonstrated the feasibility and effectiveness of deep learning approaches in automatically classifying evidence types, potentially contributing to streamlining evidence synthesis processes in biomedical research. This work could have implications for improving the efficiency and accuracy of evidence-based decision-making in healthcare and biomedical sciences by automating the classification of evidence levels from large volumes of literature.

5. In this research, the authors [5] likely focused on elucidating black-box models for biomedical text classification. They probably investigated methods to interpret and explain the predictions made, in the context of biomedical text classification tasks. The study might have explored techniques for generating human-interpretable explanations of model decisions, aiming to enhance transparency. The findings likely provided insights into the inner workings of these models and their decision-making processes, contributing to the interpretability and accountability of machine learning systems in biomedical informatics. This work could have implications for facilitating the adoption of advanced machine learning techniques in healthcare settings by enabling clinicians and researchers to better understand and validate the predictions made by black-box models.

6. In this study, the authors [6] likely developed they probably designed and implemented a neural network-based model capable of automatically classifying sentences into relevant categories or topics within abstracts from both biomedical and computer science domains. The research likely involved preprocessing the text data, training the deep learning classifier on annotated datasets, and evaluating its performance on unseen data. The findings likely demonstrated the effectiveness of the proposed model in accurately categorizing sentences from diverse abstracts, potentially aiding in information retrieval, summarization, or knowledge discovery tasks in biomedical and computer science fields. This work could have implications for improving the efficiency and accuracy of text processing and analysis in interdisciplinary research areas, ultimately facilitating

advancements in both biomedical and computer science domains.

7. In this study, the authors [7] likely conducted an investigation into research methods for techniques. They probably reviewed and analyzed existing literature on the application of topic modeling and deep learning in the context of biomedical document classification. The research might have involved synthesizing insights from various studies to identify trends, challenges, and best practices in this area. This work could have implications for guiding future research directions and informing the development of more effective automated systems for organizing and analyzing biomedical literature.

8. In this paper, the authors [8] likely introduced ML-Net, a deep neural network model designed. They probably developed and evaluated ML-Net to automatically assign multiple labels to biomedical texts, reflecting various relevant concepts or categories present in the documents. The research likely involved training the model on annotated datasets, fine-tuning its parameters, and assessing its performance in accurately predicting multiple labels for biomedical texts. The findings likely demonstrated the effectiveness of ML-Net in handling the complexities of biomedical text data and achieving high accuracy in multi-label classification tasks. This work could have implications for improving information retrieval, document organization, and knowledge extraction processes in biomedical informatics, ultimately facilitating more efficient and comprehensive analysis of biomedical literature.

9. In this study, the authors [9] likely they probably evaluated and compared different active learning strategies to determine their effectiveness in improving the performance of text mining tasks in the biomedical domain. The research may have involved implementing various active learning algorithms, such as uncertainty sampling or query-by-committee, and assessing their impact on the efficiency of text classification, information extraction, or other text mining tasks. The findings likely provided insights into the strengths and limitations of different active learning approaches and their applicability to biomedical text mining scenarios. This work could have implications for optimizing text mining pipelines in biomedical informatics, leading to more accurate and comprehensive analysis of biomedical literature and facilitating knowledge discovery in the field.

10. In this paper, the authors [10] likely probably investigated methods that leverage unlabeled or partially labeled data to train deep learning models for tasks such as document classification, entity recognition, or relation extraction in the biomedical domain. The research may have involved reviewing and analyzing existing literature on unsupervised and self-supervised learning techniques applied to biomedical text mining, as well as proposing novel approaches or improvements. The findings likely provided insights into the potential of these techniques to extract meaningful information from large volumes of unannotated biomedical text data, contributing to advancements in biomedical informatics and facilitating knowledge discovery in biomedical research. This work could have implications for developing more scalable and efficient text mining systems in the biomedical field.

11. In this paper, the authors [11] likely introduced they probably developed a hierarchical classification framework capable of assigning multiple labels to text inputs, where each label represents a specific category or concept. The research may have involved designing a deep learning architecture that can efficiently handle the extreme multi-label classification scenario, where the number of possible labels is very large. The model likely incorporated techniques for hierarchical organization of labels to improve classification accuracy and efficiency. The findings likely demonstrated the effectiveness of the proposed deep neural network in accurately predicting multiple labels for text inputs, particularly in scenarios with a large number of possible labels. This work could have implications for various applications in text classification, recommendation systems, and information retrieval, especially in domains with complex and hierarchical label structures like biomedical informatics or e-commerce.

12. In this study, the authors [12] likely explored the use they probably developed a methodology that leverages thesaurus knowledge to enhance word representations in embedding spaces, aiming to improve the performance of automated classification tasks on biomedical literature. The research may have involved incorporating domain-specific thesauri or ontologies into the word embedding generation process, either through direct integration or through techniques like retrofitting. The findings likely demonstrated the effectiveness of the proposed thesaurus-based word embeddings in capturing semantic relationships and domain-specific knowledge, thereby improving the accuracy of automated classification models for biomedical literature. This work could have implications for advancing text mining and information retrieval techniques in biomedical informatics, ultimately facilitating more efficient and comprehensive analysis of biomedical literature and supporting knowledge discovery in the field.

13. In this study, the authors [13] likely evaluated various active learning strategies to determine their effectiveness in enhancing text mining tasks within the biomedical domain. This research may have involved comparing or relation extraction. The findings likely provided insights into the strengths and limitations of each active learning

approach, facilitating the optimization of text mining pipelines in biomedical informatics. This work contributes to the advancement of text mining techniques in biomedical research, potentially leading to more accurate and comprehensive analyses of biomedical literature and aiding knowledge discovery in the field.

14. In their the authors [14] likely developed GHS-NET to address the challenges of accurately assigning multiple labels to biomedical texts, reflecting various relevant concepts or categories within the documents. This model likely combines shallow neural network architectures with hybridization techniques to improve classification performance. The research likely involved training and evaluating GHS-NET on annotated biomedical text datasets, demonstrating. The findings likely contribute to advancements in biomedical informatics by providing a robust and versatile tool for automatically categorizing biomedical texts, thereby facilitating information retrieval, knowledge organization, and discovery processes in biomedical research and healthcare.

15. In the authors [15] likely investigated various machine learning algorithms and techniques suitable for accurately assigning multiple labels to medical texts, which may cover diverse medical concepts or categories. The research may have involved evaluating the performance of these methods on annotated medical text datasets, considering factors such as classification accuracy, efficiency, and scalability for multi-label medical text classification tasks. This work contributes to advancing the field of medical informatics by identifying effective methods for organizing and analyzing medical text data, ultimately supporting tasks such as clinical decision-making.

NLP

In the realm of biomedical extraction, NLP plays a pivotal role in extracting valuable insights from vast amounts of unstructured text data, including medical literature, clinical notes, and research articles. Through techniques like text classification, NLP aids in categorizing, organizing, and extracting relevant information for various biomedical applications.

In biomedical text classification, NLP algorithms are employed to categorize text documents into predefined classes or categories based on their content, structure, or context. This process facilitates the efficient organization and retrieval of biomedical information, enabling researchers, clinicians, and healthcare professionals to access pertinent literature and insights for decision-making, research, and clinical practice. For instance, NLP can be used to classify medical documents into different categories such as disease types, treatment modalities, patient outcomes, or research methodologies, thereby streamlining information retrieval and knowledge discovery in the biomedical domain. Furthermore, NLP techniques enable the extraction of specific entities and information from biomedical text, including diseases, drugs, genes, proteins, symptoms, and treatment outcomes. By identifying and categorizing these entities, NLP supports tasks such as entity recognition, information extraction, and structured data representation, facilitating deeper analysis and interpretation of biomedical text data. This capability is particularly valuable in biomedical research, drug discovery, clinical decision support, and healthcare analytics, where access to accurate and structured biomedical information is essential for driving innovation and improving patient outcomes.

Moreover, NLP enables the classification of biomedical documents based on various attributes such as study type, medical specialty, research focus, or sentiment. By categorizing documents into relevant classes, NLP facilitates literature mining, evidence synthesis, and knowledge aggregation, empowering researchers and clinicians to access and analyze relevant literature efficiently. Additionally, NLP-powered sentiment analysis can provide insights into patient experiences, healthcare provider attitudes, or drug efficacy, enhancing our understanding of patient-centered care and healthcare delivery.

Overall, NLP plays a crucial role in biomedical extraction by leveraging advanced algorithms and techniques to analyze, categorize, and extract valuable insights from unstructured text data. Through tasks like text classification, entity recognition, and sentiment analysis, NLP enables the efficient processing and interpretation of biomedical information, thereby driving advancements in healthcare, biomedical research, and life sciences. As technology continues to evolve, NLP will continue to play an increasingly important role in unlocking the full potential of biomedical data and transforming healthcare delivery and outcomes for the better.

Word Embedding

Embedding Words. We use GloVe, an unsupervised learning approach, to obtain word vector representations [34]. This is a tool for processing natural language that uses count-based word representations and overall statistics. A vector of real numbers captures the primary semantic features in between the phrases. The cosine similarity or the Euclidean distance may be used to simply calculate the semantic similarity between the two words. This model takes into account both word-level and character-level word segmentation. In order to improve the semantic accuracy, the Word2vec model suggested in [34] makes advantage of the linked properties between words. The dimensionality problem is addressed by

employing a low-dimensional space representation. Word2vec uses two designs for word embedding: CBOW and skip-gram.

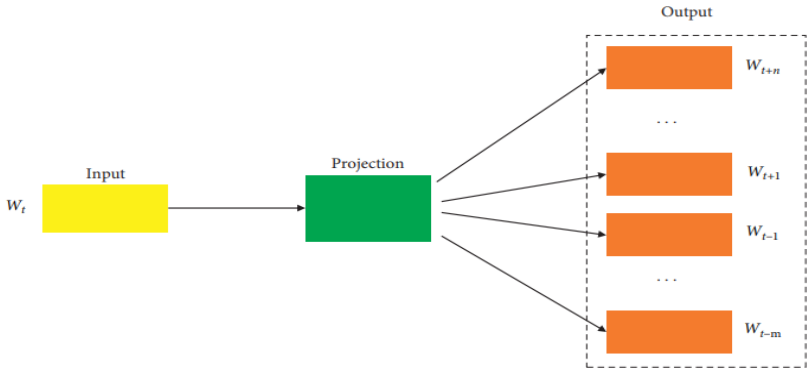

Fig 1: Diagram of word embedding

When it comes to training the word embedding, CBOW is faster than skip-gram. When expressing the semantic information, skip-gram appears to be more accurate. Consequently, this research use the Word2vec model, which is dependent on skip-gram, to train the word embedding problem. You can see the layout of a word embedding module in Figure 1.

CNN

CNN with LSTM Module. Among the many algorithms used in deep learning, CNN stands out. Computer vision and natural language processing have both made good use of this well-known feed-forward neural network's convolutional calculations and deep structure. The authors of this study used a CNN-LSTM hybrid model. Regular Neural Networks (RNNs) are extensively utilized for operations on sequential data. This RNN model takes in both the current input and the previous output and adds them together. In order to take the sequence states into account, the activation function tanh is utilized to regulate it. There will be a multiplication coefficient because the RNN derivative will propagate and communicate to time intervals starting from time t − 1, continuing through time t − 2, and so on up to time l.

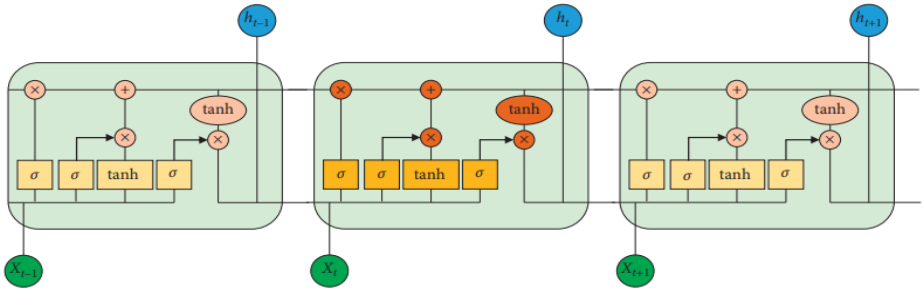

Fig 2: Diagram of a LSTM unit

When multiplication is ongoing, gradient explosion and disappearance happen. Among the primary issues with loss distance dependency is the forward process's insensitivity to start-sequence input, which has little to no impact on subsequent sequences. A number of gates can be used to solve the LSTM problem. ,Long short-term memory (LSTM) gate architectures selectively memorize the input [35]. We commit the most crucial details to memory and let the less significant ones fade into oblivion. hence, the evaluation of the subsequent new data that might be stored in the existing state is accomplished. By feeding the previous state output ht−1 and the current input in a function Xt into a sigmoid function, we may determine the current fresh information that can be readily kept by generating a number between 0 and 1. the last condition The current unit's output is the next-moment Ct, which is derived using the input and forget gates and used to create the hidden layer ht of the next state. ,The output gate determines the output based on the data collected from the cell state.

3. Proposed Approach

This study's technique was derived from previous research and discussions on machine learning, text mining, and connection extraction from medical literature. There were three primary parts to the suggested method: preprocessing, features extraction, and relation extraction. The first part took free-text phrases as input, processed them, and then produced a list of words with annotations. The second part of the system found several sentence-related characteristics that were useful for relation extraction. The architecture of the suggested technique,

DDRel, is depicted in Figure 1. Oddly enough, the output of the preceding component was input into a machine learning component, which completed the detection of connections between medication and illness entities.

Natural Language Processing (NLP) methods formed the basis of the approach's initial phase, preprocessing. It extracted all terms from medical literature pertaining to the biomedical idea (diseases and therapies) and removed any irrelevant or noisy data. It consisted of four main steps: (i) segmentation, (ii) tokenization, (iii) semantic annotation, and (iv) part-of-speech tagging. Sentence Splitting stage, texts are broken into smaller components and given unique identifiers. Sentences are formed by dividing texts using the punctuation marks "," "?," and "!".

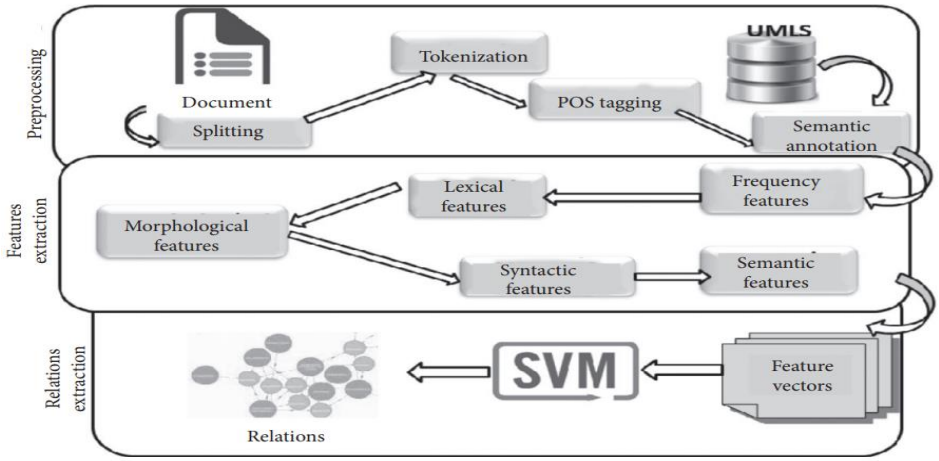

Fig 3: Architecture of Proposed approach

Tokenization stands at subsequent to the splitting of sentences, tokens were extracted from each sentence. A process called tokenization involves breaking down phrases into their component words using non-alphanumeric symbols like space, alien break, or punctuation marks. An XML file containing tokens linked to the following was provided as the end product of tokenization: (i) the sentence identifier, then the token identification, then the length, then the token orthography, then the type, and finally the token itself, which is represented by the string we can see the user-facing XML file display. Tagging Verbal Components. Nouns (NN), verbs (VB), adjectives (JJ), conjunctions (CC), and proper nouns (NNS) are some examples of words that may be Table 1: Extracted Features

"tagged" using the Part-of-Speech (POS) system. The ANNIE POS Tagger algorithm is now live. The final product is an XML file where the grammatical functions of each word are marked shows the user's view of the XML file. "Semantic Annotation," we extract named entities of illnesses and medications. There were a lot of reasons why it was hard to isolate illnesses and medications. There are a number of acronyms and names that are interchangeable for each medical topic. More importantly, in this setting, basic dictionaries are not applicable to novel medications and illnesses. This step involved configuring the Meta-Map system to find the UMLS Meta thesaurus concepts in biomedical texts.

Feature	Example
Frequency features:	
Number of named entities	2
Number of drugs	1
Number of diseases	1
Number of verbs between two NE	1
Number of words between NEs	4
Bag-of-Word	Preliminary = 1; evidence = 1; suggests = 1; that = 1; interferons = 1; beta = 1; may = 1; also = 1; induce = 1; regression = 1; of = 1 metastatic = 1; renal = 1; cell = 1; carcinoma = 1;
Lexical features:	
Sequence of words of the NE	Interferons beta_ metastatic renal cell carcinoma
Sequence of words between every two NEs	May_also_induce_regression
Sequence of 3 words before each NE	Preliminary_evidence_suggests; Also_induce_regression
Sequence of 3 words after each NE	May_also_induce; null
Morphologic features:	
Sequence of lemmas of the words between every two NEs	May_also_induce_regression
Sequence of lemmas of the 3 words before each NE	Preliminary_evidence_suggest; Also_induce_regression
Sequence of lemmas of the 3 words after each NE	May_also_induce; null
Syntactic features:	
Sequence of POS of NE	NNS_NN_JJ_JJ_NN_NN
Sequence of POS of words between every two NEs	MD_RB_VB_NN
Sequence of POS of 3 words before each NE	JJ_NN_VBS; RB_VB_NN
Sequence of POS of 3 words after each NE	MD_RB_VB; NULL
Verbs sequence among every two NEs	Induce;
First verb preceding every NE	Suggest; induce
First verb after every NE	Induce; null
Semantic features:	
Semantic type sequence	TREAT_DIS

These were the characteristics of each word in a sentence: one that in the same group as the words' MeSH mapping, Domain Knowledge, and morphological aspects are semantic kinds as were in the earlier study.

Instead of building characteristics for each token, like in Rosario and Hearst's [30] work, this one did so for each phrase. It was also believed that newly developed characteristics would be better suited for extracting drug-disease relationships, therefore they were included to the dataset.

(i) Features for frequency,

(ii) features for lexical and semantic information,

(iii) features for morphological information,

(iv) features for syntactic information, and

(v) features for drug-disease interactions were suggested in this study.

Features related to frequency. Here are the features that were represented by the frequency:

(i) the total number of NEs (drugs and illnesses are examples of named entities) that appear in the text

(ii)count of drug-related words in the phrase

(iii)the total number of nouns in the phrase that denote illnesses

(iv) the number of verbs that connect each pair of NEs in the sentence

(v) the number of words that connect each pair of NEs in the phrasethe total number of lemmas that connect the two NEs in the given phrase

(vi)Bag-of-Words: a way to organize a sentence's words according to how often each word appears

Here are the lexical features:

(i) The sequence in which words appear in NEs

(ii) The sequence in which words appear in each pair of NEs

(iii) A string of "n" phrases that comes before each NE

(iv) A string of "n" words after each NE

Physical Characteristics. This phase involved the extraction of morphological characteristics, which comprised the following:

(i) The order of the words within each pair of NEs according to the lemma.

(ii) The order of the "n" words before each NE according to the lemma.

(iii) The chain rule for the sequence of "n" words after each NE

These aspects pertain to the point of sale (POS) of every NE and encompass the following:

(i) The word order inside each pair of NEs according to POS

(ii) the word order of "n" words before each NE according to POS

(iii) the word order of "n" words after each NE according to POS

Features Defined by Semantics. The goal of this stage is to identify the phrase's word combinations and extract them. DISEASE and TREAT are the values of these semantic kinds.

Deals with relation extraction. We used a machine learning classifier to extract the drug-disease connections. A categorization procedure that followed relation classes formed the basis of the relation extraction method, which went like this: Relationships with CURE, PREVENT, SIDE EFFECT, and NO CURE all exist. Remedy Fix Dis. Recurrent spontaneous abortions can be treated with intravenous immunoglobulin, for instance. Get ahead of the disease by treating it. For instance, statistics on preventing strokes. The output of the TREAT is DIS, which is a side effect. An example would be a radiation-induced malignant mesodermal mixed tumour in the uterus. The disease cannot be cured with treatment. For instance, it was shown that head lice are resistant to some pesticides. Other relationships include VAGUE, which is extremely ambiguous, ONLY TREAT, which is also ambiguous, and ONLY DIS, which is not specified. Using the features and outputs from the previous stage, this classification helped extract relations between entities. However, classical machine learning classification approaches struggled to handle massive amounts of categorized data. Hence, a support vector machine (SVM) was employed in this method, which performed admirably when applied to data with a high degree of dimensionality [3]. These algorithms take a collection of characteristics observed in a prior stage as input. A machine learning technique finds a hyperplane that maximally divides the feature space into classes using these characteristics. Maximizing the margin is an objective of the support vector machine (SVM) algorithm, which aims to reduce misclassification errors while maximizing class separation.

To categorize the drug-disease connections from biomedical databases, this work utilized a supervised classifier SVM. The purpose of SVM was to differentiate between classes of relations. As a first stage in relation extraction, we supplied the classifier with a training set of feature vectors. We next employed support vector machines (SVMs) with polynomial kernels, as these SVMs have a kernel function and are very appropriate for

our situation. Support vector machines (SVMs) use feature vectors (generated during preprocessing and semantic annotation) to forecast the relation class for every phrase in the data file. Each sentence in this data file is represented by a vector that contains all the characteristics associated with that sentence. Figuring 6 displays the outcomes of the relation extraction. You can see a list of medications and a list of relations when you click on drug-disease relations extraction. On the other hand, you may see which diseases are associated with a medicine when you choose a drug and a connection type (prevent, treat, etc.).

4. Results

The proposed technique was validated through the implementation of a system. We utilized the standard corpus retrieved from MEDLINE 2001 for the studies. No cure, side effect, prevent, cure, and therapy were the four kinds of semantic links that were annotated into this corpus. The MEDLINE 2001 Database of biological literature was used to validate this corpus [30]. The results comparison was validated by using the corpus.

From this matrix, performance metrics were derived for the assessment. For every row in the matrix that represented an actual class, there was a corresponding column that represented the anticipated class. Table 4 displays the confusion matrix for the multiclass classification system that has been put into place.

Table 2: Confusion matrix.

		Predicted class	
		False class	True class
Actual class	False class	TN	FP
	True class	FN	TP

In the CURE class, there are 785 TP classes because those classes actually exist and are predicted to be CURE classes. In the FN class, there are 25 classes (10 + 5 + 10), because those classes are in the CURE class but aren't predicted to be CURE. In the TN class, there are 82 classes (57 + 25 + 0) because they aren't CURE classes but are still predicted to be. Here are the recall, precision, f-score, accuracy, and specificity values obtained from the matrix (Table 4), which were calculated as follows:

Table 3: System confusion matrix.

		Predicted classes			
		CURE	PREVENT	SIDE EFFECT	NO CURE
Actual classes	CURE	785	10	5	10
	PREVENT	2	57	2	2
	SIDE EFFECT	1	2	25	1
	NO CURE	1	1	2	0

Table 4: Comparison of the recall measure.

System	Method	Recall %			
		CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	NLP, polynomial SVM UMLS ontology	96.91	90.48	86.21	0
Abacha and Zweigenbaum [31]	Linear SVM	100	15.15	0	NA
Wang et al. [36]	Pattern, UMLS	89.8	NA	83.3	NA

Table 5: Assessment of accuracy measure.

System	Method	Precision %			
		CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	NLP, polynomial SVM UMLS ontology	99.49	81.43	73.53	0
Abacha and Zweigenbaum [31]	Linear SVM	90.44	15.15	0	NA
Wang et al. [36]	Pattern, UMLS	91.2	NA	92.3	NA

Table 6: Assessment of F-score measure.

System	F-score			
	CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	98.19	85.71	79.37	0
Abacha and Zweigenbaum [31]	95	15.15	0	NA
Frunza et al. [32]	93.6	76.5	50.0	NA
Suchitra and Sudah [34]	90.3			
Muzaffar et al [35]	98.05	93.55	88.89	NA
Wang et al. [36]	90.49	NA	87.56	NA

Table 7: Assessment of the accuracy measure.

System	Accuracy %			
	CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	96.76	97.86	98.52	98.02
Rosario and Hearst [30]	92.6	38.5	20	NA
Suchitra and Sudah [34]			90	
Muzaffar et al [35]	96.1	97.4	96.4	NA

Table 8: Specificity measure.

System	Specificity %			
DDRel approach	CURE	PREVENT	SIDE EFFECT	NO CURE
	95.35	98.42	98.94	98.52

5. Conclusion

In conclusion, the research presented in this paper underscores the critical importance of biomedical text classification using deep learning and NLP techniques for disease prediction. With the exponential growth of biomedical data, the manual analysis of vast amounts of literature has become impractical, necessitating the development of automated text classification methods. The integration of deep learning and NLP not only enables efficient analysis of biomedical text but also holds immense promise in enhancing disease prediction accuracy and improving healthcare outcomes. The main issue addressed in this study is the challenge of effectively analysing and extracting meaningful insights from complex biomedical text data. Traditional machine learning methods often struggle to capture the intricate relationships and nuances present in biomedical literature, leading to suboptimal performance in disease prediction tasks. By leveraging deep learning architectures and advanced NLP techniques, we aim to overcome these limitations and develop more robust and accurate disease prediction models. In summary, the research presented in this paper lays the foundation for advancing biomedical text classification techniques for disease prediction. By harnessing the power of deep learning and NLP, we can unlock new possibilities for improving healthcare decision-making and ultimately enhancing patient outcomes.

References

[1] Rios, A., & Kavuluru, R. (2015, September). Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM conference on bioinformatics, computational biology and health informatics* (pp. 258-267).

[2] Behera, B., Kumaravelan, G., & Kumar, P. (2019, December). Performance evaluation of deep learning algorithms in biomedical document classification. In *2019 11th international conference on advanced computing (ICoAC)* (pp. 220-224). IEEE.

[3] Flores, C. A., Figueroa, R. L., & Pezoa, J. E. (2021). Active learning for biomedical text classification based on automatically generated regular expressions. *IEEE Access*, *9*, 38767-38777.

[4] Burns, G. A., Li, X., & Peng, N. (2019). Building deep learning models for evidence classification from the open access biomedical literature. *Database*, *2019*, baz034.

[5] Moradi, M., & Samwald, M. (2021). Explaining black-box models for biomedical text classification. *IEEE journal of biomedical and health informatics*, *25*(8), 3112-3120.

[6] Gonçalves, S., Cortez, P., & Moro, S. (2020). A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*, *32*, 6793-6807.

[7] Yuk, J., & Song, M. (2018). A study of research on methods of automated biomedical document classification using topic modeling and deep learning. *Journal of the Korean Society for information Management*, *35*(2), 63-88.

[8] Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., & Lu, Z. (2019). ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal*

of the American Medical Informatics Association, *26*(11), 1279-1285.

[9] Naseem, U., Khushi, M., Khan, S. K., Shaukat, K., & Moni, M. A. (2021). A comparative analysis of active learning for biomedical text mining. *Applied System Innovation*, *4*(1), 23.

[10] Nadif, M., & Role, F. (2021). Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings in Bioinformatics*, *22*(2), 1592-1603.

[11] Gargiulo, F., Silvestri, S., Ciampi, M., & De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing*, *79*, 125-138.

[12] Koutsomitropoulos, D. A., & Andriopoulos, A. D. (2022). Thesaurus-based word embeddings for automated biomedical literature classification. *Neural Computing and Applications*, *34*(2), 937-950.

[13] Naseem, U., Khushi, M., Khan, S. K., Shaukat, K., & Moni, M. A. (2021). A comparative analysis of active learning for biomedical text mining. *Applied System Innovation*, *4*(1), 23.

[14] Ibrahim, M. A., Khan, M. U. G., Mehmood, F., Asim, M. N., & Mahmood, W. (2021). GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification. *Journal of biomedical informatics*, *116*, 103699.

[15] Lenivtceva, I., Slasten, E., Kashina, M., & Kopanitsa, G. (2020). Applicability of machine learning methods to multi-label medical text classification. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20* (pp. 509-522). Springer International Publishing.