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

Biomedical Document Enhancement through Probabilistic Graph Clustering: Indexing and Key Phrase Mining

Jose Mary Golamari[1], D. Haritha[2]

Submitted: 25/08/2023 **Revised**: 16/10/2023 **Accepted**: 25/10/2023

Abstract: As the size of biomedical databases is increasing along with genes and diseases, finding key feature sets is complex due to large data sizes and sparsity problems. The process of extracting relevant biomedical features from documents, ranking them based on their probability, and clustering them is a crucial aspect of biomedical document feature extraction, ranking, and classification. To extract gene/protein features from pre-processed documents, the Abner tagger is used, and the highest probability biomedical features are identified for the graph initialization process. The graph-based clustering is performed based on the relationship between gene/protein terms in the ranked document set. To enhance the quality of the cluster, a novel graph similarity measure is employed, which maximizes the probabilistic entropy measure and prioritizes gene-based ranked document clustering. Experimental results prove that the proposed model has better improvement over the conventional models.

Keywords: *biomedical documents, gene data, feature extraction, document mapping*

1. Introduction

Medical data classification is essential for various applications in information processing and machine learning. Machine learning tools are used to extract valuable information from large datasets, with classification being a commonly used algorithm for mining knowledge rules [1]. The field of biomedical research has seen groundbreaking advancements in the recent past, with the potential to redefine the healthcare paradigm. As we navigate through the 21st century, the convergence of technology with biology, especially the surge in biomedical data, stands poised to transform health outcomes and patient experiences. The crux of this transformation lies in the ability to analyze, interpret, and act upon the huge data produced every day. The term "biomedical data" encompasses a wide spectrum, ranging from molecular and genomic data to clinical records and imaging studies. With the accelerated pace of genomic sequencing, the size of data has expanded exponentially. Concurrently, the digitization of health records has led to the generation of huge amounts of patient-related data. Imaging modalities like MRI, CT scans, and X-rays further augment this data pool. This huge influx of data, often termed 'Big Data,' holds the promise of unlocking intricate human body processes, disease etiologies, and potential therapeutic interventions. Biomedical data analysis stands at the intersection of this revolution, bridging the gap between raw data and actionable insights.

However, with great potential comes significant challenges. The sheer volume, velocity, and variety of biomedical data pose challenges in storage, processing, and interpretation. Traditional data analysis tools need to improve in the face of such complexities. Moreover, the sensitive nature of health-related data necessitates stringent security and privacy measures. As we steer towards a data-driven healthcare model, addressing these challenges becomes imperative [2]. Stemming algorithms are employed to reduce words to their root or stem form, treating morphological variants as equivalent for clustering purposes. Porter's stemming algorithm is a widely used method. Additionally, words occurring in only a few documents are assigned higher weights, often using inverse document frequency (IDF) term weighting to discriminate between documents [3]. Different ranking methods are applied in information retrieval. Content-based ranking methods compute rankings and keywords based on the content of documents. Use-based ranking methods consider the user's past and present navigations to predict and provide keywords for the user. Leading-based rankings rely on link structures within web graphs, aiming to enhance search engine quality [4]. Google and Microsoft employ page ranking algorithms for ranking static ontology structure databases, resulting in a quicker and more efficient retrieval process. These algorithms evaluate link structures in various web graphs to determine rankings. Machine Learning can be implemented through two primary approaches: supervised and unsupervised learning.

Medical data typically comprises sets of biomedical patches and their associations with key terms. Identifying high-risk key terms in medical data can be challenging for doctors. However, patients' knowledge and their clinical histories can assist in the identification of these keyterms [5].

[1] *Research scholar, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram.Guntur, Andhra Pradesh, India.*
[2] *Professor, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram.Guntur, Andhra Pradesh, India.*
* *Corresponding Author Email: golamarijosemary@gmail.com*

Document clustering involves high-dimensional data, often characterized by sparse word-document matrices with positive ordinal values and numerous outliers. Term frequency is commonly used as the term weight, with frequent terms carrying greater importance. The VSM document representation consists of three steps: document indexing, term weight calculation, and similarity identification between documents. Biomedical data classification involves aggregating extensive medical data into useful clusters, each representing a specific subject or context. Traditional classification approaches may group medical data without considering contextual details, leading to inefficient knowledge retrieval. Classification of medical data is crucial for improving data exchange and communication in distributed settings.

Biomedical feature selection techniques typically employ filter or wrapper schemes, assessing the relevance and importance of each feature. Univariate scoring metrics play a significant role in the overall ranking criteria. Managing a large feature space can present challenges in terms of performance and scalability. Researchers have used statistical and mining tools to assist doctors in detecting biomedical keyterms. Computational approaches in text mining can be categorized into top-down and bottom-up approaches. Top-down approaches are user-focused and concentrate on specific criteria of interest, while bottom-up approaches aim to identify all important information [6]. Information Extraction (IE) is a top-down approach where predefined forms are used to extract information. At the same time, Information Retrieval (IR) is a bottom-up approach focused on finding relevant documents from a large set. Both Automatic Text Summarization (ATS) and IE aim to identify relevant information, but they differ in how they present this information to users. Hybrid methods have been explored to improve biomedical keyterm prediction. It's essential to bridge the gap between identifying keyterms and determining the necessary treatments. Classification techniques are employed to diagnose biomedical keyterms, leveraging machine learning tools to analyze extensive datasets. Biomedical feature selection techniques use either filter or wrapper schemes to assess feature relevance and importance. These methods consider relationships between features and class labels. When dealing with a large feature space, predefined numbers of features are often used for classification, and issues related to performance and scalability need to be addressed. In this paper, we propose best suitable method for Biomedical document enhancement through Probabilistic Graph Clustering: Indexing and Key Phrase Mining.

The following are the organization of the paper: We describe the related work of biomedical document enhancement through Probabilistic Graph Clustering in section 2. Section 3 covers proposed work. The section 4 covers the results and discussion. Finally, in Section 5, concludes the paper.

2. Background and Related Work

Historically, the significance of data in medicine traces back centuries. Ancient physicians based their treatments on rudimentary data collection and pattern recognition. However, the dawn of the digital age and the subsequent explosion of biomedical data have mandated a more structured and scientific approach to data analysis [7-10]. Numerous studies have underscored the importance of biomedical data in enhancing healthcare outcomes. [11] highlighted how genomic data could provide insights into disease susceptibility, paving the way for preventive medicine. Similarly, [12] discussed the role of Electronic Health Records (EHR) in predictive analytics, especially in identifying potential disease outbreaks. The potential of imaging data in biomedical analysis has been a focal point of many research endeavors. A study by [13] delved into the role of advanced MRI imaging data in diagnosing neurodegenerative diseases, emphasizing the need for sophisticated data analysis tools. The integration of different data sources, termed multi-modal data integration, has also gained traction. [14] demonstrated how integrating genomic data with clinical records could enhance cancer treatment outcomes. The rise of biomedical data has invariably led to the evolution of data analysis tools tailored for biomedical research. Machine learning and artificial intelligence have emerged as frontrunners in this regard.

A seminal work by [15] showcased the application of deep learning algorithms in analyzing genomic data, demonstrating improved accuracy over traditional methods. Another study by [16] emphasized the role of natural language processing in extracting meaningful insights from unstructured clinical notes. While the potential benefits of biomedical data analysis are manifold, it's essential to address the associated challenges. Data privacy and security have been a recurrent theme in biomedical literature. A study by [17] shed light on the vulnerabilities of EHR systems and proposed a multi-layered security framework. Another challenge is the interoperability of data systems. With multiple sources of biomedical data, ensuring seamless data integration is paramount. In [18], the authors discussed the potential of blockchain technology in ensuring data interoperability without compromising on security. In [19], the authors involved an analysis of various domain representations, proposing a novel embedding method for terms. In a related context, [20] introduced a keyword extraction technique for individual documents. To assess the efficiency of the DIKpE algorithm, a publicly available dataset comprising 215 diverse computer science documents was utilized, irrespective of their content. The evaluation of DIKpE focused on automatically extracting matches between key sentences within the text and the main sentences.

Remarkably, DIKpE outperformed two other algorithms in key phrase extraction without prior training. Another study conducted by X. Mao et al. in 2020 explored unsupervised methods for keyword extraction from voice transcripts. This research delved into the multi-party conference domain, adapting successful algorithms from text transcription. The investigation encompassed aspects such as POS filtering, word clustering, sentence production via the TF-IDF method, and thematic classification for accuracy and quality assessment. The study also employed two unsupervised derogatory word detection techniques for incomplete transcriptions. The authors leveraged Term Frequency-Inverse Text Frequency (TF-IDF) with the Gini Purity Criteria approach to determine transcription topics. Their findings indicated automatic morphological alterations in literacy, particularly evident when comparing Wikipedia pages in four languages: Arabic (383,000), English (50 million), Hungarian (50,000), and Portuguese (211,000). Performance was assessed using standardized cumulative gain and median accuracy, with Arabic outperforming English in certain aspects.

Furthermore, the authors applied concepts from Naïve Bayes and K-Nearest Neighbors (KNN) for precise performance. They also employed relevant pre-processing techniques to enhance the effectiveness of their approach. The authors developed a source code for the Vector Space Model and gathered relevant data for their survey. Their classification techniques, including K-Nearest Neighbors, successfully categorized news items based on content classification when keywords were entered. It's important to note that the methods discussed primarily focused on generating key phrases from document text without the capability to generate label phrases. Most of these text mining approaches rely on representing text documents as bags of words and use vector representations to assign numerical importance values to words within documents. These representations include the Vector Space Model (VSM), probabilistic models, and logical models. The VSM represents documents as vectors in a common vector space, with term weights reflecting the importance of each feature (term) in the document [21-23].

3. Proposed Model

The process of extracting relevant biomedical features from documents, ranking them based on their probability, and clustering them is a crucial aspect of biomedical document feature extraction, ranking, and classification. To extract gene/protein features from pre-processed documents, the Abner tagger is used, and the highest probability biomedical features are identified for the graph initialization process. The graph-based clustering is performed based on the relationship between gene/protein terms in the ranked document set. To enhance the quality of the cluster, a novel graph similarity measure is employed, which maximizes the probabilistic entropy measure and prioritizes gene-based ranked document clustering. Moreover, the pre-processing techniques employed in this investigation, including text normalization, tokenization, and stemming, can be applied to various other text analysis tasks to enhance the precision of the outcomes.

The removal of non-essential words from a stop-word list is an effective strategy to minimize the occurrence of noisy matches and elevate the accuracy of the analysis. In summary, the proposed approach presents a pioneering and effective method to cluster biomedical documents and extract crucial gene-based key-phrases, which can be utilized for in-depth analysis and research in the realm of biomedicine.

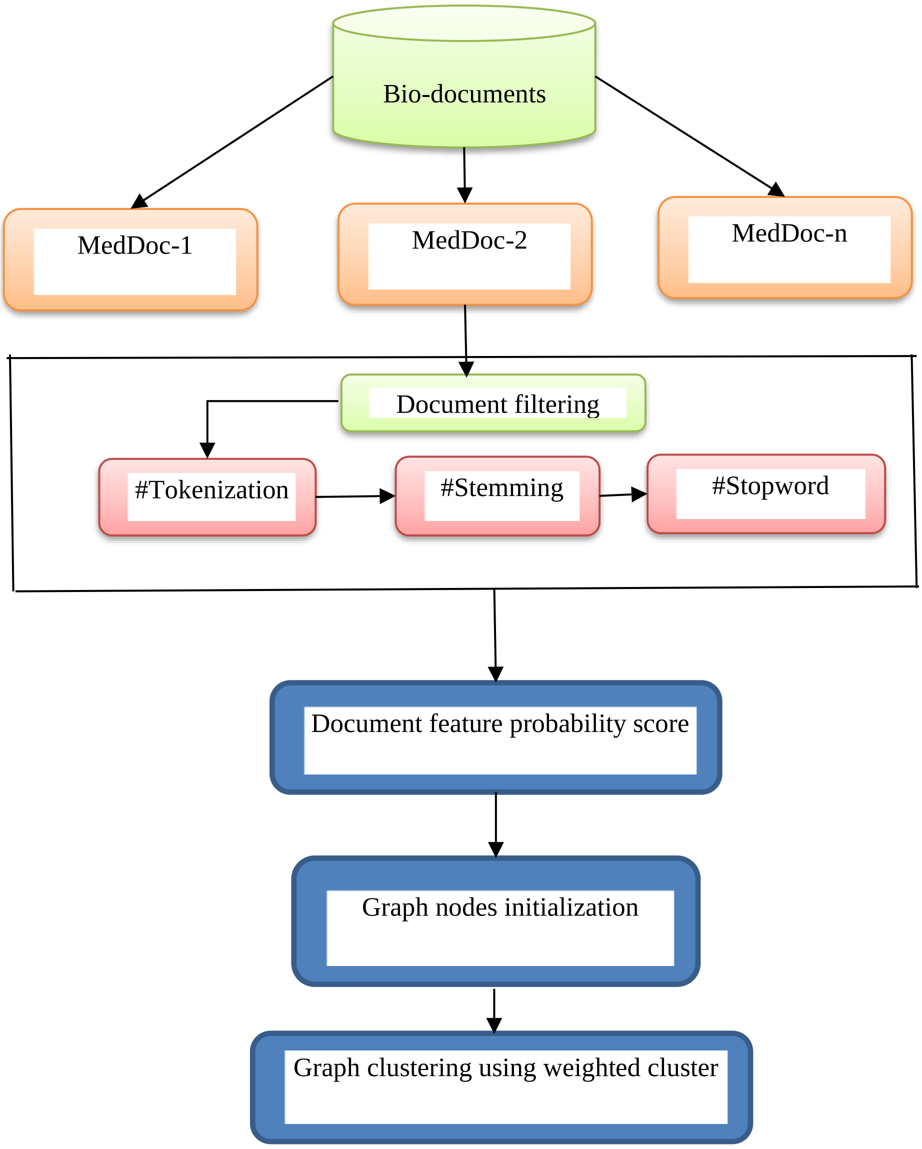

Bio-documents
MedDoc-1
MedDoc-2
MedDoc-n
Document filtering
#Tokenization
#Stemming
#Stopword
Document feature probability score
Graph nodes initialization
Graph clustering using weighted cluster

Fig 1: Proposed framework

• To propose a probabilistic graph clustering approach for indexing and extracting key phrases from large biomedical documents.

• To evaluate the effectiveness of the proposed approach in terms of precision, recall, and F1-score compared to traditional methods.

• To investigate the scalability of the proposed approach for handling large-scale biomedical document sets.

The proposed method utilizes probabilistic graph clustering to index large biomedical documents. This method is designed to effectively cluster similar documents and identify key phrases within each cluster, as shown in Figure 1.

Biomedical Document Preprocessing Algorithm:

D is the set of biomedical XML documents.

λ is the minimum threshold value.

GP is the set of gene protein tags extracted using the Abner library. GPT be the set of gene. Protein tokens in each document D. SD is the synonym dataset.

GPProb(D[i], gpt) represents the gene_protein probability in document *D*[i].

Step 1: Tokenization and Soft-Stemming

For each document d in D, tokenize the document to get a set of tokens Dk. Perform soft-stemming on Dk to reduce tokens to their base forms.

Step 2: Stopword Removal

Remove stopwords from Dk to obtain the processed token set of PBD.

Step 3: Gene and Protein Extraction

Apply the Abner library to extract gene_protein tags from PBD, resulting in the set GP.

Step 4: Tokenization of Gene_Protein Tags

Tokenize the gene_protein tags in GP to get GPT, which contains the gene. Protein tokens.

Step 5: Gene_Protein Probability Calculation

For each gpt in GPT, calculate the gene_protein probability in each document D[i] as follows:

GPProb(D[i], gpt) = Max { Prob(gpt / D[i]) / Prob(gpt)}; for i = 1, 2, ..., *N*

Step 6: Similarity Calculation with Synonym Dataset SD

For each gene protein gn_m in GPT, find the similarity between gene tags and the synonym dataset SD using the formula:

SGP = Sim(gp_m, SD)

Sim(gp_m, SD_j) = Max {Prob(gp_m / SD_j) / (|SD_j| * Prob(gp_m))}; for *m* = 1, 2, ..., |*G*| and *j* = 1, 2, ..., |SD|

Step 7: Document Weight Calculation

Calculate the document weight for each document term Dt_i, gene protein gp_m, and synonym SD_j as follows:

DW(Dt_i, gp_m, SD_j) = Sim(gp_m, SD_j) * ∑ { Prob(Dt_i / SD_j) * Prob((gp_m ∩ Dt_i) / SD_j) }

Step 8: Bio-Key Term Identification

For each biomedical document in D, check the document weights WSGPD[].

If the document weight w_i is greater than λ, mark the corresponding gene protein tags and synonym terms as bio key terms.

Step 9: End

End the algorithm after processing all the biomedical documents in *D*.

Methods used for classification

Fuzzy rough clustering (FRC):

This data clustering method can categorize biological texts. FRC can manage noisy, partial, and overlapping data, making it ideal for biological document classification.

- Preprocess texts by deleting stop words, stemming or lemmatizing words, and decreasing data dimensionality.
- Use fuzzy rough sets to model the documents. This requires fuzzy lower and upper approximations for each category.
- Compute each document's class membership. The document's distance from each category's fuzzy lower and higher estimates determines this.
- Classify each document by the highest membership degree.

Support vector machines: These are ideal for biological document categorization because they can handle high-dimensional data and learn from small datasets. Biomedical document classification using SVM generally involves the following steps:

- Prepare documents.
- Show documents as feature vectors. Bag-of-words, TF-IDF, and word embeddings can achieve this.
- Train the SVM classifier to give the SVM algorithm feature vectors and class labels. SVM will then learn a hyperplane to classify data points.
- New document classification for feature vector extraction and input to the trained SVM classifier. The classifier predicts each new document's class label.

Naïve Bayes: This may classify biomedical texts. The simple and efficient machine learning approach is ideal for text categorization.

Graph based biomedical document clustering

Input: Pre-processed biomedical document PBD[]

Output: Top K clusters.

Procedure:

1. Initialize each node in the graph *G*(*V*, *E*) with document genes/proteins and their synonym frequencies as vertices and document weight as edge weight.

2. Get initial clusters using optimized kmean similarity measure on the graph nodes.

$Pe(v_i.v_j)$= probability that features of v_i, v_j vertices appear in the same document.

$$Pr(v_i.v_j) = \frac{n(v_i.v_j)}{n(v_j)} \qquad (1)$$

$n(v_i.v_j)$ denotes number of documents where both features in v_i, v_j matches.

$$Chisim = \frac{Pr(v_i.v_j)^2}{Pr(v_i),Pr(v_j)} \qquad (2)$$

The proposed method is a probabilistic graph clustering approach to indexing large biomedical documents, with the goal of improving document retrieval efficiency and accuracy. The method involves building a graph representation of the documents, applying a clustering algorithm to group similar documents together, and extracting key phrases from each cluster. By using graph clustering and probabilistic algorithms, this method can identify clusters of documents that are more closely related than those identified by traditional methods and identify clusters that may not be immediately apparent using other approaches. The Gene protein probabilities for each PMID document are extracted by computing the gene-protein probability in each document, finding the similarity between gene tags and the synonym dataset, and calculating the weight of the document based on the similarity between gene/protein term and synonym dataset and the probability of the document text. The resulting WSGPD [] is a weighted synonym gene protein document containing the weight of the document, document text, gene/protein term, and synonym term. Clusters are formed based on the contextual information in the biomedical documents. It is user-specific: *n*=5(default). Here, cluster names are not user-specific. Clusters are named C1.Cn based on the contextual similarity.

Biomedical graph clustering using contextual similarity refers to a technique used to group related entities in large biomedical datasets, such as proteins, genes, and diseases, by considering the context in which they appear. The approach involves representing the biomedical data as a graph, where nodes represent entities and edges represent relationships between them. Contextual similarity is determined based on how often two entities appear together in the same context, such as co-occurrence in scientific literature or shared participation in biological pathways. By leveraging this contextual similarity, biomedical graph clustering aims to identify groups of entities that are functionally related or participate in the same biological processes.

To use an SVM for biomedical document classification, you would first need to collect a set of labeled training data. This data would consist of biomedical documents that have been labeled with their correct classes. Once this have collected the training data, this can use an SVM library to train a classifier.

Prob(gpt) is the probability of the gene/protein term in the entire corpus.

Prob(gpt/*D*[*i*]) is the probability of the gene/protein term in document *D*[*i*].

The max function computes the maximum value of the fraction (*Prob*(gpt/*D*[i])/*Prob*(*gpt*)) over all documents *D*[*i*] containing the gene/protein term.

Sim() function calculates the similarity between gene/protein terms and synonym datasets using the formula given in step 16.

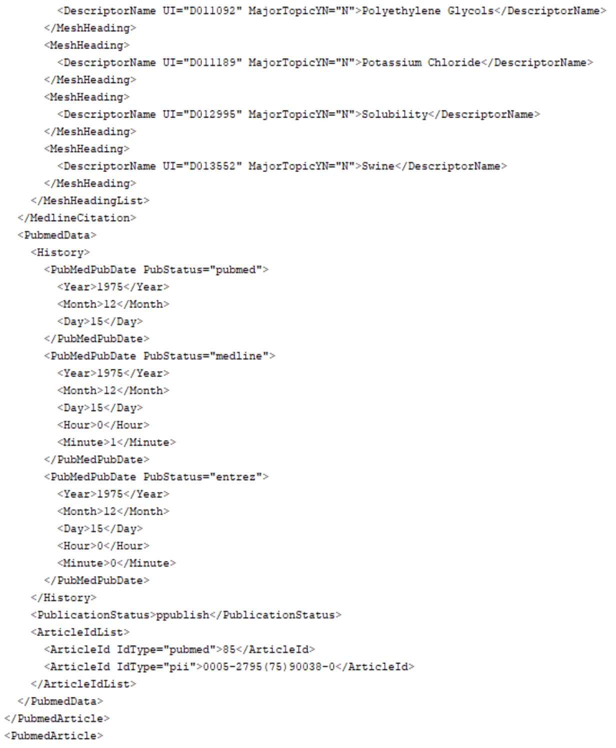
DW() function calculates the weight of the document based on the similarity between gene/protein term and synonym dataset and the probability of the document text.

WSGPD[] is a weighted synonym gene protein document containing the weight of the document, document text, gene/protein term, and synonym term.

A threshold value of 0.5 may be chosen if the algorithm being used outputs scores or probabilities between 0 and 1 and if the goal is to select a subset of documents that have a high degree of relevance to a specific gene-chemical-disease relationship. Here, a threshold value of 0.5 is a reasonable cut-off for selecting the top k documents, as it would include only those documents that are deemed to have a high degree of relevance based on the algorithm's output.

4. Results and Discussion

The experimental results are assessed using an extensive collection of TREC document sets obtained from the repository. The document clustering process involves employing various biomedical datasets, such as Pubmed and Medline XML datasets. To ensure the accuracy of the clustering process, each dataset undergoes a pre-processing phase where uncertain features and noisy content are eliminated. After this pre-processing step, it subject every document to a graph-based clustering algorithm for clustering and classification. Figure 2 shows the sample data in xml format used in this paper for experimentation. Real-time microarray cancer databases are used for the performance of experimental outcomes. The suggested feature selection-based ensemble approaches enhance the efficacy, recall, and accuracy of F-measures on high-dimensional datasets. The suggested model creates decision patterns using the whole training data set, and then its effectiveness is examined using various cross-validations utilizing 10% of the training data as test data. In comparison to traditional methodologies, the proposed ensemble decision-making framework is more effective overall and has a lower false positive rate. The proposed model's ability to reduce error rates on high-dimensional characteristics is its key advantage. Specificity or True negative rate calculates the proportion of people who are correctly identified as not having cancer. True Positive Rate or sensitivity defines the ratio of cancer cases that have been projected to be positive. In contrast, precision calculates the ratio of cancer patients who have been successfully recognized among all those whom the disease has impacted.



```
        <DescriptorName UI="D011092" MajorTopicYN="N">Polyethylene Glycols</DescriptorName>
      </MeshHeading>
      <MeshHeading>
        <DescriptorName UI="D011189" MajorTopicYN="N">Potassium Chloride</DescriptorName>
      </MeshHeading>
      <MeshHeading>
        <DescriptorName UI="D012995" MajorTopicYN="N">Solubility</DescriptorName>
      </MeshHeading>
      <MeshHeading>
        <DescriptorName UI="D013552" MajorTopicYN="N">Swine</DescriptorName>
      </MeshHeading>
    </MeshHeadingList>
  </MedlineCitation>
  <PubmedData>
    <History>
      <PubMedPubDate PubStatus="pubmed">
        <Year>1975</Year>
        <Month>12</Month>
        <Day>15</Day>
      </PubMedPubDate>
      <PubMedPubDate PubStatus="medline">
        <Year>1975</Year>
        <Month>12</Month>
        <Day>15</Day>
        <Hour>0</Hour>
        <Minute>1</Minute>
      </PubMedPubDate>
      <PubMedPubDate PubStatus="entrez">
        <Year>1975</Year>
        <Month>12</Month>
        <Day>15</Day>
        <Hour>0</Hour>
        <Minute>0</Minute>
      </PubMedPubDate>
    </History>
    <PublicationStatus>ppublish</PublicationStatus>
    <ArticleIdList>
      <ArticleId IdType="pubmed">85</ArticleId>
      <ArticleId IdType="pii">0005-2795(75)90038-0</ArticleId>
    </ArticleIdList>
  </PubmedData>
</PubmedArticle>
<PubmedArticle>
```

Fig 2: Sample Data in xml format

Testing data 1:

MZ246.12233 < 0.34

| MZ2.7921478 <0.63 : Cancer (160/0)

| MZ2.7921478 >= 0.63: Normal (2/0)

MZ246.12233 >= 0.34

| MZ17298.153 <0.13 : Cancer (2/0)

| MZ17298.153 >= 0.13 : Normal (25/0)

Testing data 2:

MZ246.41524 < 0.28

| MZ194.41064 <0.13 : Normal (4/0)

| MZ194.41064 >= 0.13

| | MZ0.20108355 <0.03 : Normal (1/0)

| | MZ0.20108355 >= 0.03 : Cancer (123/0)

MZ246.41524 >= 0.28

| MZ280.32307 < 0.22

| | MZ6.3515266 <0.2 : Normal (1/0)

| | MZ6.3515266 >= 0.2 : Cancer (3/0)

| MZ280.32307 >= 0.22 : Normal (58/0)

Figure 3 shows the analysis comparing the proposed method (Graph based biomedical document clustering) is having the highest accuracy of 0.99 as related to conventional methods using accuracy measure.

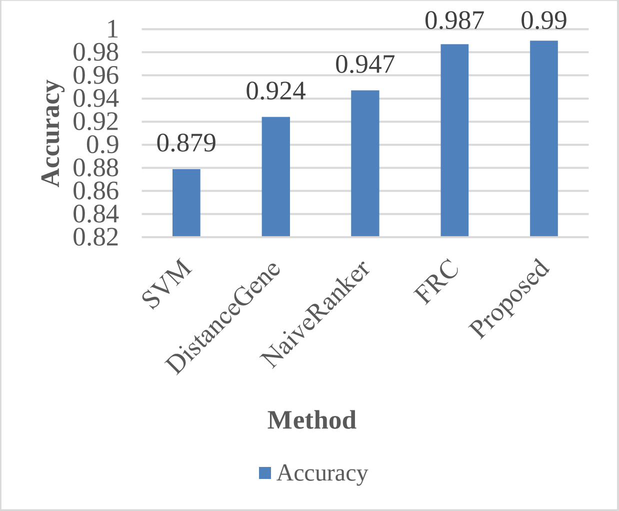


Fig 3: Analysis comparing the proposed model with conventional models using accuracy measure

Figure 4 shows the analysis comparing the proposed method (Graph based biomedical document clustering) is having the highest recall and precision of 0.99 as related to conventional methods using accuracy measure.

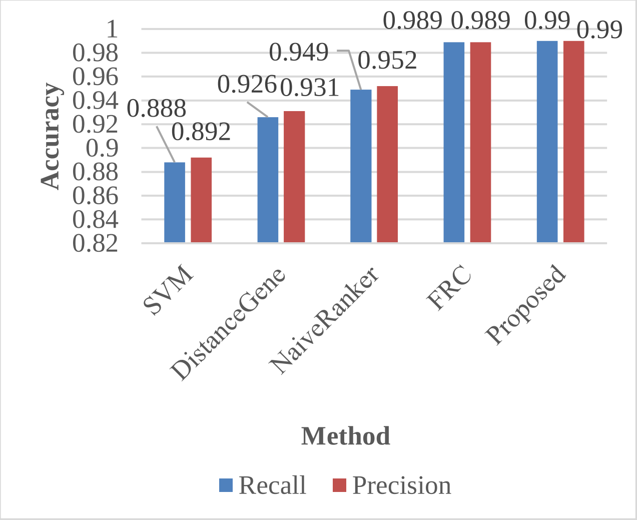


Fig 4: Analysis comparing the proposed method with conventional methods using recall and precision measures

Figure 5 shows the analysis comparing the proposed method (Graph based biomedical document clustering) is having the less computation time of 2987 msec as related to conventional methods using accuracy measure. SVM Ranker is performed poor with high computation time of 4231 msec.

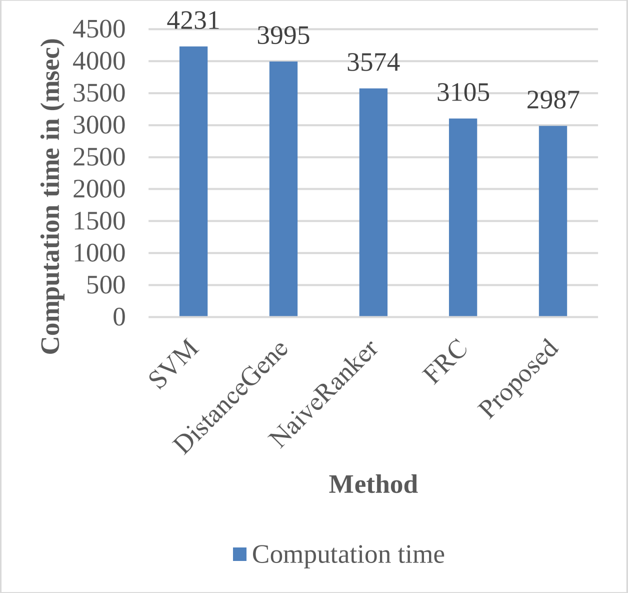


Fig 5: Analysis comparing the proposed method with conventional methods using runtime computation

5. Conclusion

In this paper, an advanced micro-array cancer disease-based biomedical document ranking is implemented on large biomedical document sets. Most of the existing models are independent of biomedical document ranking based on micro-array gene sets. In order to overcome these issues, an advanced feature selection-based classification learning model is proposed to overcome the problem of gene-based biomedical document ranking. A hybrid word embedding method and similarity metric are used to improve the efficiency of the contextual similarity between the document sets. These biomedical key entities are used to map the micro-array data classification patterns for gene the ICD mapping process.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] L. Zhang, W. Lu, H. Chen, Y. Huang, and Q. Cheng, "A comparative evaluation of biomedical similar article recommendation," Journal of Biomedical Informatics, vol. 131, p. 104106, Jul. 2022, doi: 10.1016/j.jbi.2022.104106.

[2] R. Upadhyay, P. K. Padhy, and P. K. Kankar, "A comparative study of feature ranking techniques for epileptic seizure detection using wavelet transform," Computers & Electrical Engineering, vol. 53, pp. 163–176, Jul. 2016, doi: 10.1016/j.compeleceng.2016.05.016.

[3] D. Xiong, Z. Zhang, T. Wang, and X. Wang, "A comparative study of multiple instance learning methods for cancer detection using T-cell receptor sequences," Computational and Structural Biotechnology Journal, vol. 19, pp. 3255–3268, Jan. 2021, doi: 10.1016/j.csbj.2021.05.038.

[4] Y. Wang et al., "A comparison of word embeddings for the biomedical natural language processing," Journal of Biomedical Informatics, vol. 87, pp. 12–20, Nov. 2018, doi: 10.1016/j.jbi.2018.09.008.

[5] Jahiruddin, M. Abulaish, and L. Dey, "A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora," Journal of Biomedical Informatics, vol. 43, no. 6, pp. 1020–1035, Dec. 2010, doi: 10.1016/j.jbi.2010.09.008.

[6] B. G. Patra et al., "A content-based literature recommendation system for datasets to improve data reusability – A case study on Gene Expression Omnibus (GEO) datasets," Journal of Biomedical Informatics, vol. 104, p. 103399, Apr. 2020, doi: 10.1016/j.jbi.2020.103399.

[7] S. Raza, "A COVID-19 Search Engine (CO-SE) with Transformer-based architecture," Healthcare Analytics, vol. 2, p. 100068, Nov. 2022, doi: 10.1016/j.health.2022.100068.

[8] D. Ji, J. Gao, H. Fei, C. Teng, and Y. Ren, "A deep neural network model for speakers coreference resolution in legal texts," Information Processing & Management, vol. 57, no. 6, p. 102365, Nov. 2020, doi: 10.1016/j.ipm.2020.102365.

[9] D. Guo, G. Duan, Y. Yu, Y. Li, F.-X. Wu, and M. Li, "A disease inference method based on symptom extraction and bidirectional Long Short Term Memory networks," Methods, vol. 173, pp. 75–82, Feb. 2020, doi: 10.1016/j.ymeth.2019.07.009.

[10] P. Bota, A. Fred, J. Valente, C. Wang, and H. P. da Silva, "A dissimilarity-based approach to automatic classification of biosignal modalities," Applied Soft Computing, vol. 115, p. 108203, Jan. 2022, doi: 10.1016/j.asoc.2021.108203.

[11] W. Zheng et al., "A graph kernel based on context vectors for extracting drug–drug interactions," Journal of Biomedical Informatics, vol. 61, pp. 34–43, Jun. 2016, doi: 10.1016/j.jbi.2016.03.014.

[12] A. Duque, H. Fabregat, L. Araujo, and J. Martinez-Romo, "A keyphrase-based approach for interpretable ICD-10 code classification of Spanish medical reports," Artificial Intelligence in Medicine, vol. 121, p. 102177, Nov. 2021, doi: 10.1016/j.artmed.2021.102177.

[13] M. Fernández-Pichel, D. E. Losada, and J. C. Pichel, "A multistage retrieval system for health-related

misinformation detection," Engineering Applications of Artificial Intelligence, vol. 115, p. 105211, Oct. 2022, doi: 10.1016/j.engappai.2022.105211.

[14] M. Lentschat, P. Buche, J. Dibie-Barthelemy, and M. Roche, "A new method to extract n-Ary relation instances from scientific documents," Expert Systems with Applications, vol. 209, p. 118332, Dec. 2022, doi: 10.1016/j.eswa.2022.118332.

[15] A. P. Kumar, A. Nayak, M. S. K., S. Goyal, and Chaitanya, "A novel approach to generate distractors for Multiple Choice Questions," Expert Systems with Applications, vol. 225, p. 120022, Sep. 2023, doi: 10.1016/j.eswa.2023.120022.

[16] T. Bikku and R. Paturi, "A novel somatic cancer gene-based biomedical document feature ranking and clustering model," Informatics in Medicine Unlocked, vol. 16, p. 100188, Jan. 2019, doi: 10.1016/j.imu.2019.100188.

[17] M. Sarrouti and S. Ouatik El Alaoui, "A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering," Journal of Biomedical Informatics, vol. 68, pp. 96–103, Apr. 2017, doi: 10.1016/j.jbi.2017.03.001.

[18] L. A. Quintero-Domínguez, C. Morell, and S. Ventura, "A propositionalization method of multi-relational data based on Grammar-Guided Genetic Programming," Expert Systems with Applications, vol. 168, p. 114263, Apr. 2021, doi: 10.1016/j.eswa.2020.114263.

[19] S. Cox et al., "A semantic similarity based methodology for predicting protein-protein interactions: Evaluation with P53-interacting kinases," Journal of Biomedical Informatics, vol. 111, p. 103579, Nov. 2020, doi: 10.1016/j.jbi.2020.103579.

[20] D. L. Rubin, C. F. Thorn, T. E. Klein, and R. B. Altman, "A Statistical Approach to Scanning the Biomedical Literature for Pharmacogenetics Knowledge," Journal of the American Medical Informatics Association, vol. 12, no. 2, pp. 121–129, Mar. 2005, doi: 10.1197/jamia.M1640.

[21] Q. Wang et al., "A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes," Journal of Biomedical Informatics, vol. 105, p. 103418, May 2020, doi: 10.1016/j.jbi.2020.103418.

[22] C. Yan et al., "A survey of automated International Classification of Diseases coding: development, challenges, and applications," Intelligent Medicine, vol. 2, no. 3, pp. 161–173, Aug. 2022, doi: 10.1016/j.imed.2022.03.003.

[23] X. Han et al., "A survey of transformer-based multimodal pre-trained modals," Neurocomputing, vol. 515, pp. 89–106, Jan. 2023, doi: 10.1016/j.neucom.2022.09.136.