SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V10I9P108

Volume 10 Issue 9, 72-80, September 2023
© 2023 Seventh Sense Research Group®

Original Article

A Real-Time Cancer-Covid Gene-Set Based Biomedical Document Classification and Ranking Framework for Large Databases

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

[1,2]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India.*

[1]*Corresponding Author : golamarijosemary@gmail.com*

Received: 15 July 2023 Revised: 25 August 2023 Accepted: 13 September 2023 Published: 30 September 2023

Abstract *- Identifying and ranking gene and disease patterns are essential for analyzing and ranking biomedical documents in current biomedical repositories. However, the presence of noise, uncertainty, and missing values in most biomedical databases, coupled with their diverse features and varying levels of gene and disease patterns, makes identifying and ranking high-dimensional patterns across different repositories a complex and challenging task. Data classification algorithms rely on MeSH terms or user-specific keywords to classify documents in conventional biomedical repositories. Nevertheless, these algorithms use static methods to establish relationships among gene sets, which may need to be revised for accurate analysis and ranking of biomedical documents. Locating cancer and COVID genes associated with diseases and their patterns in biomedical repositories is a difficult task. A novel Cancer-Covid gene/disease document classification and ranking approach has been suggested, employing a cross-gene model with machine learning techniques. The proposed method employs an optimized Glove feature extraction technique and an advanced classification model to identify significant features from biomedical documents. Experimental results indicate that this feature extraction method is more effective than other existing techniques in predicting gene-disease relationships in various biomedical documents.*

Keywords *- Cross-domain analysis, Cancer genesets, Covid gene sets.*

1. Introduction

The extraction of information from databases is a critical process for obtaining valuable data and making up-to-date decisions in various industries such as IT, finance, and medicine. As databases continue to grow rapidly, developing efficient machine-learning models for data analysis has become crucial. These models are necessary to extract hidden information from vast amounts of data collected in a distributed and multi-dimensional manner. While conventional models may require significant computational time for pattern discovery, KDD offers a solution for extracting thought-provoking and meaningful relationships and patterns from large databases. Databases are essential in several fields, such as organizational management, product marketing, research and development, disease diagnosis, data security, and MIS decision-making. These outcomes can be adapted to meet diverse objectives and interests. The intricate interrelation of vast amounts of data makes decision-making difficult. The medical industry is witnessing exponential data growth, rendering human analytical skills inadequate. As biotechnology advances rapidly, more biological data is being gathered and made available for analysis. Therefore, it is crucial to devise innovative techniques to extract knowledge from this data [1]. The analysis and interpretation of bio-molecular data have become increasingly important as the amount of data available continues to grow. With this growth comes the potential for discovering new and valuable insights. However, the healthcare industry faces challenges in effectively sharing this wealth of information. Despite the abundance of accessible data, there is a lack of efficient tools capable of extracting the hidden relationships and trends buried within.

To overcome this challenge, data mining and knowledge discovery techniques are being employed to uncover valuable insights in medical and scientific research. These techniques help to identify the most valuable knowledge, unlocking the full potential of bio-molecular data. The purpose of this study is to investigate the potential use of data mining techniques, specifically classification and decision tree algorithms, for analyzing large volumes of medical data obtained from hospitals. The KDD framework provides a variety of methods and strategies that can be employed to extract valuable insights from various data sources. Data mining is a commonly used tool in the field of information technology, which involves discovering hidden knowledge from extensive databases to facilitate informed decision-making and pattern evaluation. The complexity of biomedical documents, which frequently contain multiple versions of MeSH terms and disease types, presents challenges for understanding and analyzing such data.

Therefore, the work addresses these challenges and improves the ability to understand and analyze biomedical data [2]. Exploring extensive databases to identify genes or diseases with comparable characteristics can be a challenging endeavor. Nevertheless, the biomedical industry has introduced several prediction techniques that utilize gene entities and disease prediction processes. The classification and clustering of data in distributed databases can be streamlined by assigning pertinent documents to distinct clusters. An automatic algorithm can further assist in accurately identifying high-quality documents. The objective is to enhance the TREC document's classification, prediction, and clustering methods to simplify the organization of large documents in distributed databases [3]. Microarray data sets play a critical role in comprehending chronic diseases, which can advance from mild symptoms to lethal outcomes over an extended period. Normally, medical records consist of a compilation of documents concerning cancer and its related diseases. Detecting patients at a high risk of developing these ailments can pose a daunting challenge for physicians.

Nonetheless, patients who understand their medical history can aid in the timely detection of the disease. The second method to consider in classifying data involves determining the minimum number of features required to achieve maximum accuracy. These selection methods include filters, wrappers, embedded features, and hybrid methods guided by specific selection criteria. Filter methods are specialized algorithms designed to exclude irrelevant features [4].

In the field of data analysis, there exist five different classes. These classes are used to determine the most optimal feature subset based only on the unique features of the data, such as distance, correlation, and uniformity. Statistical methods are employed to rank genes, and both uniform and multivariate filters are utilized. With the exponential growth of information in various domains, particularly in distributed biomedical repositories, the need for medical data pre-processing has become urgent. This process involves transforming data in centralized databases to generate a summary by selecting only the most relevant information from the source medical data. However, this process often leads to information overload. To address this problem, medical data gene and feature extraction can be implemented [5]. The detection of cancer can be accomplished through the utilization of statistical and mining tools. The analysis of cancer has been extensively studied with the aim of achieving the highest level of accuracy. Implementing mining algorithms is crucial in uncovering the mysteries of this disease. Various hybrid methods have been investigated to enhance the accuracy of cancer prediction. However, there has been a lack of emphasis on identifying the most efficient treatment for individuals suffering from this disease.

To address this issue, a novel approach has been devised to determine the feasibility of accurately detecting cancer and managing suitable treatments. Effective cancer treatment methods can help the detection of cancer-related diseases. The use of classification techniques has become a standard practice among researchers for diagnosing such diseases as the number of fatalities continues to rise and more patient data is made available. However, traditional methods are inadequate in detecting computational variations in complex datasets. Given the complexity of cancer patterns, a cancer classification system must consider multiple types of cancer. To address these challenges, various machine-learning techniques have been proposed. These algorithms are commonly utilized to determine the optimal cancer treatment, but they can also be applied to identify cancer-related diseases [5].

In this paper, we address the subject of gene-set biomedical document classification and present a unique Cancer-Covid gene/disease document classification and ranking strategy that leverages a cross-gene model that employs ML techniques. The proposed approach utilizes an improved Glove feature extraction methodology and a sophisticated classification model to determine important features in biomedical documents.

2. Background and Related Work

In [6], the authors claimed that information retrieval plays a critical role in both engineering and computer science research. This field focuses on retrieving and analyzing knowledge-based data from databases. Their study presented a variety of techniques and models for optimizing information retrieval, such as indexing methods to simplify searches and different approaches to probing for information. Later, they provide a comprehensive overview of the basic concepts that underpin information retrieval systems. Finally, they concluded that information retrieval served as a valuable tool for accessing and exploring large amounts of data. The accumulation of extensive knowledge from a diverse range of documents serves as the foundation for information retrieval.

In [7], the authors revealed that individuals have recognized the importance of identifying and preserving information passed down for millennia. The introduction of personal computers has enabled the storage of vast amounts of data and the extraction of valuable insights. The prevalence of information in the daily lives of individuals has prompted a significant emphasis on research in Information Retrieval (IR) within the field of computer science. IR is a multifaceted process that involves the retrieval and search of extensive databases of knowledge-based data. Their work investigated various techniques and models utilized in IR, such as indexing and searching methods that facilitate the streamlining of the retrieval process.

In [8], the authors state that the importance of information as an essential resource in people's lives underscores the significance of continued research in this area. The internet has become a significant aspect of human life, and a considerable part of this experience involves sifting through the large amount of information available. However, the abundance of data can often lead to confusion as conflicting sources may provide contradictory information. To ensure that users acquire accurate and relevant data efficiently, they need to possess effective strategies and techniques. Their study focuses on the various methods of information retrieval that facilitate the acquisition of necessary data and the optimization of time management.

In [9], the authors conducted a study on query-based document ranking, an advanced information retrieval approach. While numerous technologies have been proposed for document retrieval, their study highlights the most effective techniques. Difficulty in understanding the context of texts or vocabulary used in user queries can result in reduced accuracy and recall rates. To tackle this problem, they developed a document ranking system that employs a hybrid methodology. Additionally, comparable terms are incorporated into the query to improve accuracy. Finally, the K-nearest neighbor technique is utilized to allocate a similarity score and enhance precision.

In [10], the authors analyzed different algorithms for page ranking, comparing their advantages and disadvantages based on various parameters. Later, they proposed a new page ranking system that can predict document rank accurately by combining old term frequency-inverse document weight and new document context weight.

In [11], the authors emphasized that in their research, information retrieval involves searching for documents that match users' queries, indicating their information requirements. They created an innovative framework to simplify this task that integrates an automated text classification system. The Decision Tree technique was recognized as a prevalent and efficient tool in data mining and learning mechanisms. However, it was noted that unreliable or indeterminate data could reduce its output performance. To overcome this issue, feature selection and continuous feature discretization are essential. Although classic algorithms like Naive can be helpful, they are different from the superior capabilities of ANN. Additionally, each input layer is multiplied by a weight that is then transferred to other layers.

In [12], the authors introduced a novel approach to identifying the connections between gene sets. They utilized ontology structures to represent the relationship between genes and their properties. Based on gene clusters and their properties, they proposed a probabilistic method that could predict new gene sets within limited datasets. The method considered the rates of false positives and errors in gene connections and their adjacent clusters.

In [13], the authors presented a novel approach to latent semantic indexing that organizes datasets related to genes. By analyzing 50 biomedical documents, they identified genes and created clusters with a predetermined number. However, the proposed model requires substantial computational resources and time due to the exponential increase in biomedical gene documents. An ant-based feature extraction technique was utilized to address this issue, incorporating a wrapper that predicts classification accuracy in a new framework. A new method for feature selection, called KCALO (Kernel Chaotic Antlion LOpcation), has been developed. This method uses a classification task as its fitness function and measures both the reduction in design criteria and classification accuracy. The feature selection algorithm has been modified using the Ant algorithm and Lawson rings. The LALO (Levy Optimization) algorithm selects the best feature subset, improving classification accuracy within a wrapper. This method has shown a notable 5% increase in performance compared to ALO results. The objective of this program is to explore the use of multi-objective PSO in classifying diseases.

In [14], the authors recognized the challenge posed by uneven class distribution in medical classifications and proposed incremental SVM as a solution [15]. They employed bootstrapping to identify potential candidate support vectors for future iterations to improve the classifier's performance. The sensitivity and specificity achieved were almost comparable to SVM when all available samples at a given incremental level were utilized.

However, there are situations where the class distribution is imbalanced, which can affect the classifier's performance. Additionally, the need for real-time medical data limits the number of cases that can be used for training and testing the classifier, making it a costly process. [16], the abundance of information available on the web presents a significant challenge. The huge amount of data contained within documents pertaining to a particular topic can be overwhelming. However, utilizing ontology as a resource can assist in comprehending this textual data. They proposed a retrieval and annotation model utilizing optimization techniques.

Additionally, the study addresses the challenge of retrieving relevant concepts from corpora. The objective of this study is to utilize the ontology's population to create new instances and achieve semantic document annotation. Practical semantic and ontology annotations are employed to enhance the search process. Their main aim of query paraphrasing is to generate user queries that produce the best possible results. Nevertheless, certain web users face difficulties in making efficient search queries, prompting them to try out different paraphrases to obtain satisfactory outcomes. Arabic, being a language with a plethora of synonyms and hyponyms, requires its information retrieval systems to automate the paraphrasing technique to enhance the synonymizing process. To overcome the time and resource constraints associated with existing query paraphrasing methods, dual query paraphrasing optimization techniques have been proposed as a viable solution [17]. The first method utilizes a genetic algorithm, while the second method uses the ABC-QP algorithm (Artificial Bee Colony Query, to use a different acronym).

In [18], the authors noted that query expansion is an efficient procedure for improving the efficiency of the retrieval process and overcoming search engine obstacles. However, despite its effectiveness, query expansion has limitations in its widespread use as a typical tool in search procedures. To overcome this challenge, a query expansion model is proposed, along with a new metaheuristic known as the Bat Inspired algorithm, to improve the efficiency of the retrieval process. Unlike previous studies that perform these tasks sequentially, this method predicts both the best-related documents and the best keyword growth simultaneously.

In [19], the authors presented a fundamental overview of the information retrieval method using evolutionary computation and presented some crucial models. The process of Drug Discovery and Development is a multifaceted and challenging task that requires constant refinement. Advanced algorithms, such as ant colony optimization and particle swarm optimization, are employed to optimize the process and have been shown to be more effective than traditional computing algorithms. Evolutionary computation is also thoroughly investigated to determine how various techniques can be implemented to improve the optimization process and achieve better outcomes.

The ultimate goal of Drug Discovery and Development is to identify, develop, and commercialize new chemical compounds that can be used to treat specific illnesses. Although this process is complex and intricate, using cutting-edge algorithms and techniques enables medical researchers to streamline the process and provide life-changing treatments to those in need. The development of a dependable and productive medication for patients necessitates a deep comprehension of the disease at a molecular level, as emphasized by [20].

In [21], the authors conducted a study on the frequency of outbreaks of human infectious diseases. This alarming trend can be attributed to the influence of globalization, lifestyle choices, and the distinctive features of microbes. The rise in global diseases can be largely attributed to genetic adaptation to the environment. Determining a cure has become increasingly difficult as the understanding of these diseases grows. However, the co-occurrence-based approach has significant flaws, as pointed out by [22].

One of these errors is the low precision, as not all entities that co-occur in a sentence or document are necessarily related. For example, the sentence "Tamoxifen is" could be followed by various unrelated words. The given example highlights the connection between tamoxifen, cancer, Benicar, and vertigo. Tamoxifen is used to combat cancer, while excessive use of Benicar can lead to vertigo. Although tamoxifen and vertigo are not directly related, there is a correlation between tamoxifen and cancer and benicar and vertigo. Tamoxifen is related to cancer in a "treatment" capacity, whereas Benicar is causally related to vertigo. It is essential to differentiate between various types of relationships, as depending solely on co-occurrence statistics can result in incorrect outcomes.

3. Proposed Modelling

The careful selection of essential features is paramount in analyzing large biomedical document collections utilizing statistical modeling, pattern recognition, and machine learning techniques. The process of selecting features can significantly reduce computational costs, simplify complexity, and mitigate ambiguity. Additionally, implementing optimal feature extraction methods can reduce the dimensions of the features, consequently enhancing accuracy and runtime performance.

The documents will be denoted as Doc-1 to Doc-n to indicate the training biomedical document sets. The initial step involves pre-processing all biomedical data using the Java NLP library. After filtering the input data, the Glove optimization measure is applied to the filtered data to determine the primary and contextual key vectors, as displayed in Figure 1.

The keyword rank is predicted using the main and contextual scores. A probability-based contextual similarity method is implemented to ascertain the score between the gene and disease on the biomedical document sets. Finally, a hybrid classifier is utilized to classify the documents for the ranking process. In order to identify genes and their associated illnesses, MeSH keywords play a vital role as they contain valuable information. To extract meaningful data from biological information, an external classifier is employed that follows specific rules. The primary objective of biomedical document extraction is to analyze unseen data from biomedical databases for pattern recognition. To discover uncommon and interesting rules, an additional measure and minimum support threshold are required. Missing genes and illnesses can be identified by pre-processing each biomedical XML document. Every XML file is carefully analyzed to uncover essential MeSH keywords and terms that aid in identifying genes and their related diseases.

The proposed model delves into the realm of biomedicine and its genetic lexicon, revealing novel associations between genes and diseases via a feature extraction and classification model illustrated in Figure 1. In the document filtering phase, documents are merged with pre-established gene and disease datasets, and input is obtained from the NCBI repository's Medline Database (MDB) and PubMed Database (PDB). Every data instance in the filtering phase provides document particulars and MeSH terms. One million genes are gathered to classify and extract gene tags from the input datasets.

3.1. Multi-Disease Document Filtering and Contextual Similarity

3.1.1. Input

MDB, EDB, PDB, Biomedical Disease list, MGenes, PGenes, EGenes

3.1.2. Procedure

For each biomedical document D[i] in the databases (MDB, EDB, PDB), do the following steps:

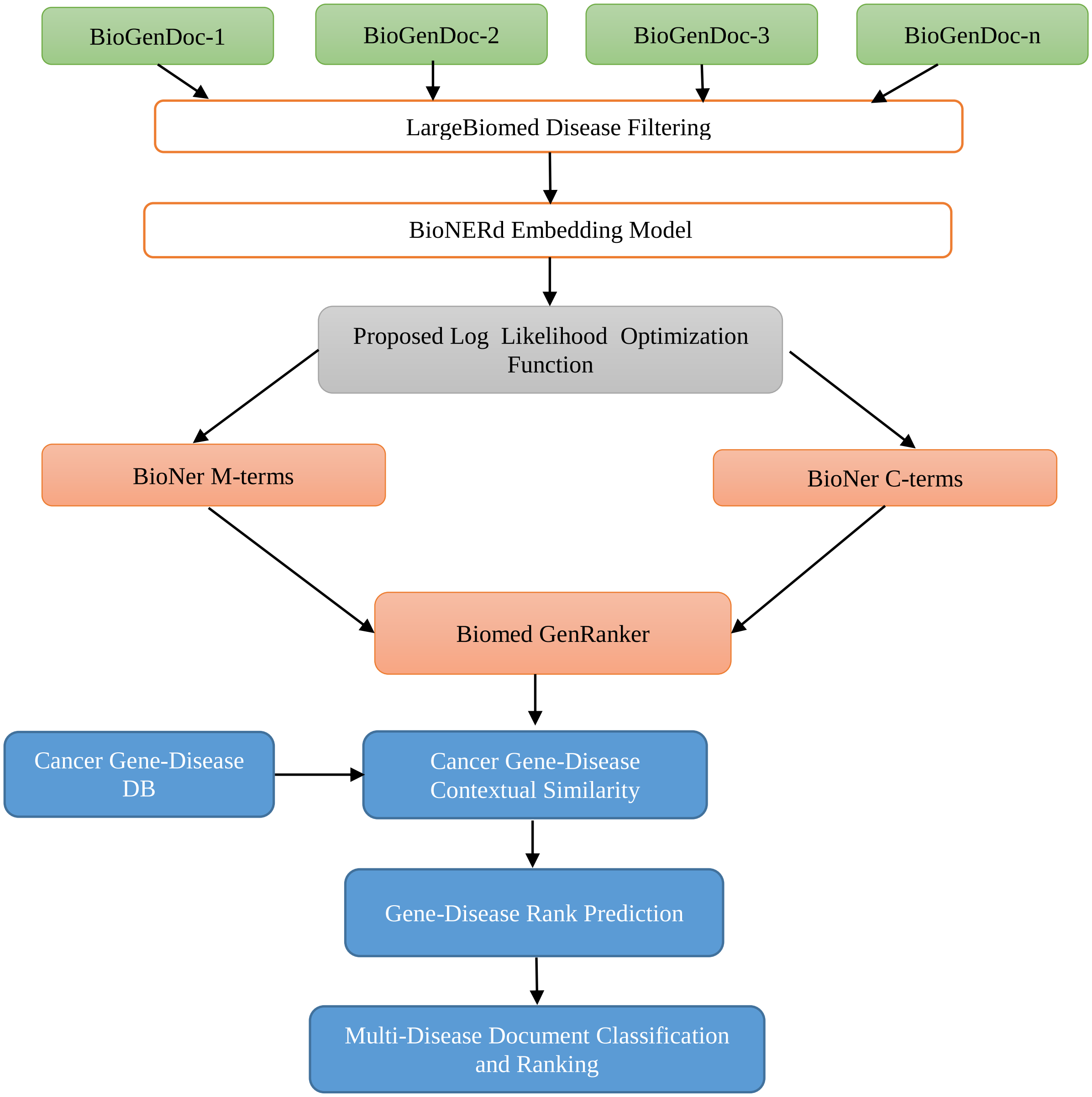
BioGenDoc-1
BioGenDoc-2
BioGenDoc-3
BioGenDoc-n
LargeBiomed Disease Filtering
BioNERd Embedding Model
Proposed Log Likelihood Optimization Function
BioNer M-terms
BioNer C-terms
Biomed GenRanker
Cancer Gene-Disease DB
Cancer Gene-Disease Contextual Similarity
Gene-Disease Rank Prediction
Multi-Disease Document Classification and Ranking

Fig. 1 Proposed framework

Classification of Documents

Depending on the type of document (PubMed or Medline) and if the document belongs to the DiseaseList, add the document to the corresponding set (PDocSet for PubMed, MDocSet for Medline, and EDocSet for others).

Text Processing

For each set of documents (PDocSet, MDocSet, EDocSet), perform the following processing steps:

a) Tokenization

Break down the document text into individual words or tokens (PDTokens for PubMed documents, MDTokens for Medline documents, EDTokens for others).

b) Stemming

Reduce words to their base or root form (SPD[i] for PubMed documents, SMD[i] for Medline documents, SED[i] for others).

c) Stopword Removal

Remove common words that do not contribute much to the meaning of the text (SSPD[i] for PubMed documents, SSMD[i] for Medline documents, SSED[i] for others).

d) Non-Functional Removal

Remove non-functional elements from the processed text (PDFR for PubMed documents, MDFR for Medline documents, EDFR for others).

e) Gene Extraction

Extract gene tags from the processed text (PDGenes for PubMed documents, MDGenes for Medline documents, EDGenes for others).

Similarity Measure: Compute the similarity measure Sim(G_i, C_i, D_i) for each document, where:

G_i, C_i, and D_i are gene, disease, and document instances, respectively.

μ_G (c), μ_C (c), and μ_D (c) are the mean values for each instance type in each biomedical class c.

σ_GeneDB^2 (c), σ_Clist^2 (c), σ_Dlist^2 (c) are the variances for each instance type in each biomedical category c.

P(GeneDB[i]/(CDocs[i])) and P(GeneDB[i]/DGenes[i]) are the probabilities of a gene appearing in the document classes and the extracted gene list, respectively.

The similarity measure is computed as follows:

Sim(G_i,C_i,D_i) = ((max|μ_G (c),μ_C (c),μ_D (c)|)/(min|σ_GeneDB^2 (c),σ_Clist^2 (c),σ_Dlist^2 (c)|)) * (P(GeneDB[i]/(CDocs[i]))/(P(GeneDB[i]/DGenes[i]))

3.1.3. Output

The result would be a similarity measure for each biomedical document, which could be used to find associations between genes and diseases based on how similar they are in terms of their gene, disease, and document features.

3.2. Hybrid Glove Optimization Model

In the hybrid TF-ID-based Word2Vec embedding measure, different aspect sentences and their related feature attributes are taken as input to compute the probability score. Here, the computed probabilistic score is used to select key feature aspect sentences and their related features in the CNN network. Let w_i be the word in the aspect sentence, and its related attributes are named A. The log estimated contextual similarity measure to find the key aspect word in the given dataset is computed as:

$$P_{HTF-ID} = \frac{1}{N} \sum_{i,j} tf_{i.j}.\log p\left(w_i | A_{j+i}\right) \qquad (1)$$

Where, *tf* is the term frequency, and *p* is the conditional probability of occurrence of the word in the given aspect-related features, as shown in eq (1).

The weighted hybrid word2vec of the aspect term is given in eq (2)

$$WE\left(Word2vector_{,i.j}P_{HTF-ID}\right) = maxwordvote\left(\frac{\lfloor n \rfloor_{ij}}{prob\left({}^{n_{ij}}/_{d_j}\right)}\right) \times log_{2\frac{\lfloor D \rfloor}{1+prob\{w_i \in d_j/A_w\}}}, \frac{1}{N} \sum_{i,j} tf_{i.j}.\log p\left(w_i | A_{j+i}\right) \qquad (2)$$

In the glove biomedical feature extraction model, a novel optimization function is proposed to improve the overall efficiency of the feature extraction process for the document classification process.

This work proposes a new optimization function to find and extract the essential main and its correlational features on large biomedical data. The proposed mathematical optimization function is defined as. Define soft constraints for each word pair is computed using the eq (3)

$$C = Costfunction = b_i w_i^T w_j + b_i w_i^T w_j + \theta - \left(log\left(X_{ij}\right)/max\{\|w_i\|\||, \||\|w_j\|\}\right) \qquad (3)$$

$$\eta = weight = f\left(X_{ij}\right) = \begin{cases} \left(\frac{X_{ij}}{x_{max}}\right) \alpha \; if \; x_{ij} < x_{max} \\ 1 \qquad otherwise \end{cases}$$

Where θ, α are the scaling factors of the main and contextual word vectors

The hybrid glove vector model to define the cost function is given in eq (4)

$$J = max\left(\sum_{i=1}^{V}\sum_{j=1}^{V}\eta\left(b_i w_i^T w_j + b_i w_i^T w_j\right) + \theta - log\left(X_{ij}\right) / max\left(\|w_i\|\|, \|\|w_j\|\right)^2\right), WE\left(Hword2vec_{i,j}, P_{HTF-ID}\right) \quad (4)$$

3.3. Feature Ranking Based Cancer-Covid Biomedical Document Classification

In this phase, a hybrid feature extraction measure is proposed to find the various gene-disease-related features for cross-disease data classification. Since most of the traditional feature selection methods are not based on cross-domain features and their relationships for document classification, in this feature extraction measure, a hybrid entropy measure is proposed to classify the features of biomedical terms for the ranking process.

Enhanced Entropy Based Cross-Domain Feature Selection

Step1: Read input filter data.

Step 2: To each feature in the word embedding vector.

Step 3: Compute enhanced entropy using the following formula

$$P[i] = -P(D_i).\log(P(D_i))$$

$$\text{Entropy}(D) = \sum_i P[i]$$

$$P(C_m \mid t_{cw}) > P(C_n \mid t_{cw}) \text{ for } n \neq m$$

u sing eq(1) , $P(C_m / t_{cw})$ is maximized.

According to bayes theory;

$$P(C_m \mid t_{cw}) = \frac{P(t_{cw} \mid C_m)P(C_m)}{P(t_{cw})}$$

Proposed Measure=(Math.cbrt(ent(D)*|D|)* $P(C_m \mid t_{cw})$)/ (chiVal(D));

Step 4: Select the top k gene-disease features in each category from classification patterns.

In the enhanced entropy measure, each cross-domain feature's probability is evaluated to find the contextual top features for the classification problem.

4. Results and Discussions

The proposed model underwent testing in a Java environment, utilizing third-party libraries. Subsequently, the model was applied to extensive biomedical document collections sourced from the NCBI website [https://www.ncbi.nlm.nih.gov/], with results indicating that it demonstrated superior processing speed and accuracy on large datasets.

Our methodology involved developing a machine learning framework based on Java, which was utilized to identify and classify key patterns. We assessed the framework's effectiveness on various Gene datasets, Chemical drug datasets, and biomedical documents, utilizing metrics such as Recall, Precision, F1 value, and Area Under the Curve (AUC) to evaluate gene chemical prediction performance.

Figure 2 describes the cancer feature selection measures and their count on the large biomedical documents. In this table, the proposed model has better essential key features than the conventional measures for gene disease-based pattern evaluation.

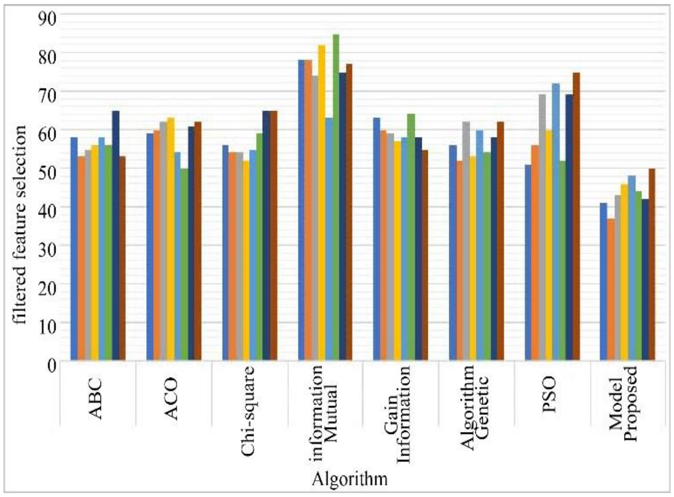

Fig. 2 Performance of proposed model and conventional methods for feature selection measures

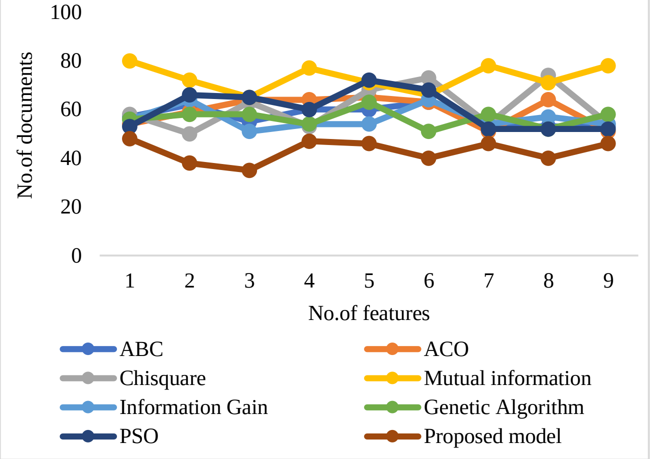

Fig. 3 Covid chemical-disease features extraction using the proposed model and the conventional models

Figure 3 describes the feature selection measures and their count on the large biomedical documents. In this table, the proposed model has better essential key features than the conventional measures for chemical gene-based disease pattern evaluation. Figure 4 describes the feature selection measures and their count on the large biomedical documents. In this table, the proposed model has better essential runtime (ms) than the conventional feature extraction measures for the Cancer-Covid gene, ICD and disease-based pattern evaluation. Figure 5 describes the efficiency of the proposed biomedical document classification model compared to the conventional models for accuracy.

The table shows that the proposed model has a better accuracy value than the traditional approaches on large biomedical Cancer-Covid gene-disease ICD features.

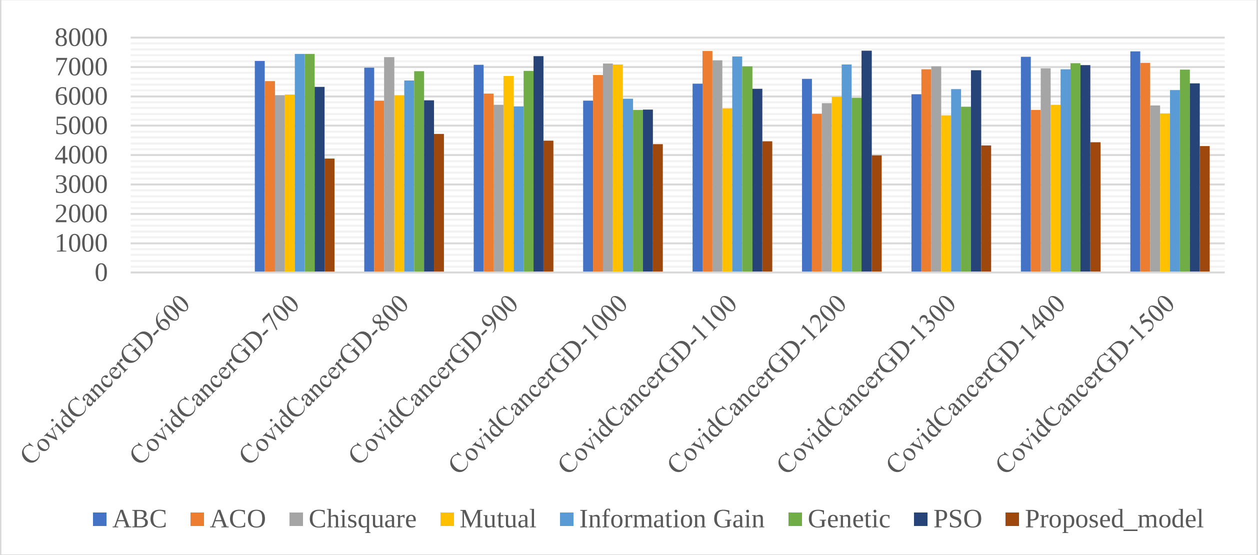

Fig. 4 Performance comparison of runtime (ms) of proposed biomedical document feature extraction method and the traditional methods on biomedical datasets

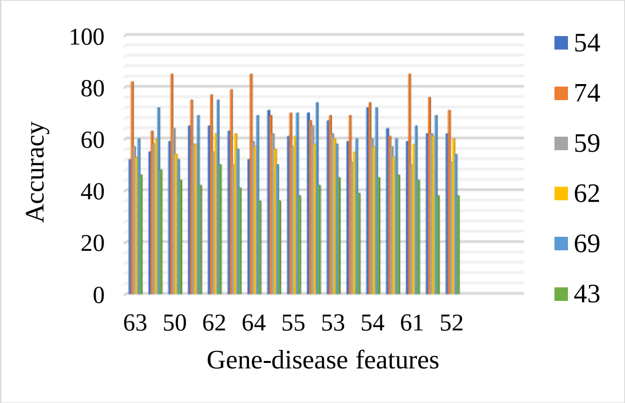

Fig. 5 Performance analysis of accuracy using different traditional frameworks

5. Conclusion

In this paper, we present a novel method for classifying gene-disease patterns and extracting their relational features from large biomedical datasets. The proposed approach has two unique characteristics. First, it demonstrates effectiveness in handling noisy data. Second, we propose a hybrid cross-gene disease document classification model that leverages machine learning frameworks. This paper also introduces an optimized GloVe feature extraction technique and an advanced classification model. These tools are designed to identify key feature sets from biomedical documents. Experimental outcomes indicate that the feature extraction-based gene-disease prediction framework provides better optimization compared to existing state-of-the-art techniques when applied to various biomedical disease documents.

References

[1] Xin Shao et al., “A Clinical Genomics-Guided Prioritizing Strategy Enables Selecting Proper Cancer Cell Lines for Biomedical Research,” *iScience*, vol. 23, no. 11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Yongjing Lin et al., “A Document Clustering and Ranking System for Exploring MEDLINE Citations,” *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 651–661, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[3] Thulasi Bikku, and Radhika Paturi, “A Novel Somatic Cancer Gene-Based Biomedical Document Feature Ranking and Clustering Model,” *Informatics in Medicine Unlocked*, vol. 16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Gaelen P. Adam et al., “A Novel Tool that Allows Interactive Screening of Pubmed Citations Showed Promise for the Semi-Automation of Identification of Biomedical Literature,” *Journal of Clinical Epidemiology*, vol. 150, pp. 63–71, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Mercedes García Carrillo et al., "Academic Dependency: The Influence of the Prevailing International Biomedical Research Agenda on Argentina's CONICET," *Heliyon*, vol. 8, no. 11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] P. Dhanalakshmi, K. Ramani, and B. Eswara Reddy, "An Improved Rank Based Disease Prediction Using Web Navigation Patterns on Bio-Medical Databases," *Future Computing and Informatics Journal*, vol. 2, no. 2, pp. 133–147, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Saeid Balaneshinkordan, and Alexander Kotov, "Bayesian Approach to Incorporating Different Types of Biomedical Knowledge Bases into Information Retrieval Systems for Clinical Decision Support in Precision Medicine," *Journal of Biomedical Informatics*, vol. 98, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Lena Maier-Hein et al., "BIAS: Transparent Reporting of Biomedical Image Analysis Challenges," *Medical Image Analysis*, vol. 66, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Fei Zhu et al., "Biomedical Text Mining and its Applications in Cancer Research," *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 200–211, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[10] Chloé Cabot, Stéfan Darmoni, and Lina F. Soualmia, "Cimind: A Phonetic-Based Tool for Multilingual Named Entity Recognition in Biomedical Texts," *Journal of Biomedical Informatics*, vol. 94, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Maciej Rybinski, Jerry Xu, and Sarvnaz Karimi, "Clinical Trial Search: Using Biomedical Language Understanding Models for Re-Ranking," *Journal of Biomedical Informatics*, vol. 109, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Muhammad Abulaish, Md. Aslam Parwez, and Jahiruddin, "DiseaSE: A Biomedical Text Analytics System for Disease Symptom Extraction and Characterization," *Journal of Biomedical Informatics*, vol. 100, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Nicholas C. Ide et al., "Essie: A Concept-Based Search Engine for Structured Biomedical Text," *Journal of the American Medical Informatics Association*, vol. 14, no. 3, pp. 253–263, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[14] Muhammad Ali Ibrahim et al., "GHS-NET a Generic Hybridized Shallow Neural Network for Multi-Label Biomedical Text Classification," *Journal of Biomedical Informatics*, vol. 116, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Jiho Noh, and Ramakanth Kavuluru, "Improved Biomedical word Embeddings in the Transformer Era," *Journal of Biomedical Informatics*, vol. 120, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Tuan Manh Lai, ChengXiang Zhai, and Heng Ji, "KEBLM: Knowledge-Enhanced Biomedical Language Models," *Journal of Biomedical Informatics*, vol. 143, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Haixia Shang, and Zhi-Ping Liu, "Network-Based Prioritization of Cancer Biomarkers by Phenotype-Driven Module Detection and Ranking," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 206–217, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Satya S. Sahoo et al., "ProvCaRe: Characterizing Scientific Reproducibility of Biomedical Research Studies Using Semantic Provenance Metadata," *International Journal of Medical Informatics*, vol. 121, pp. 10–18, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Sarvnaz Karimi, Justin Zobel, and Falk Scholer, "Quantifying the Impact of Concept Recognition on Biomedical Information Retrieval," *Information Processing & Management*, vol. 48, no. 1, pp. 94–106, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[20] Zan-Xia Jin et al., "Ranking via Partial Ordering for Answer Selection," *Information Sciences*, vol. 538, pp. 358–371, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Grace Wang et al., "Representation of Women in Diagnostic Radiology Residency Programs: Does National Institutes of Health Program Ranking Matter?," *Journal of the American College of Radiology*, vol. 18, no. 1, pp. 185–191, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Shuang Zhu et al., "Research Trends in Biomedical Applications of Two-Dimensional Nanomaterials over the Last Decade – A Bibliometric Analysis," *Advanced Drug Delivery Reviews*, vol. 188, 2022. [CrossRef] [Google Scholar] [Publisher Link]