How to Cite:
Reddy, V. S. N., & Midhunchakkaravarthy, D. (2022). An Efficient multi-class SVM and Bayesian network based biomedical document ranking and classification framework using Gene-disease and ICD drug discovery databases. *International Journal of Health Sciences*, *6*(S3), 1291–1308. https://doi.org/10.53730/ijhs.v6nS3.5551

An Efficient multi-class SVM and Bayesian network based biomedical document ranking and classification framework using Gene-disease and ICD drug discovery databases

V. Shiva Narayana Reddy
Research Scholar, Lincoln University College, Malaysia

Divya Midhunchakkaravarthy
Professor, Lincoln University College, Malaysia

***Abstract*---**Biomedical document feature extraction and ranking play an essential role in the real-time document key phrase extraction and ranking. International classification of disease (ICD-10) is a list of medical related terms such as disease symptoms, abnormal discovery and disease signs. In most of the conventional methods, finding, extraction and ranking of biomedical disease patterns with the gene terms help to rank the phrase or document. However, the contextual disease patterns of these methods are independent of gene entities, disease entities and drug discovery codes for document ranking and summarization. Conventional word embedding models such as gain ratio, entropy, Glove, chi-square and probabilistic measures are used to find the essential key terms and its relationships using static gene disease databases. The main objective of the proposed work is to optimize the word embedding model along with the key-phrase ranking and classification. Most of the biomedical applications use pre-trained gene-disease database with limited number of gene names for key phrase ranking and classification process. In this work, an integrated gene-disease database and ICD drug database codes are used to train the model using the optimized SVM classification model and Bayesian estimation model. In this work, an ensemble learning model is integrated to CNN framework to classify the document sets using the gene disease database and ICD codes. Experimental results are implemented on biomedical documents using ICD codes and gene disease databases with different statistical metrics. Results show that the present framework has better true negative rate, error rate and precision than the stat of art algorithms.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
***Corresponding author**: Reddy, V. S. N.; Email: Shiva_reddy5017@yahoo.com*
Manuscript submitted: 18 Nov 2021, Manuscript revised: 09 Feb 2022, Accepted for publication: 27 March 2022

Keywords---gene data, machine learning, classification, keyphrase ranking, drug discovery.

Introduction

The Machine-learning framework used for the discovery of medications and bioactive substances prioritization with pharmacological impacts and their streamlining [1]. In Bioinformatics, the main goal is to make more and more medicines in a short amount of time that are usually safe. There is now a new field called computer-aided drug design that is separate from the rest of the field. Bioinformatics is the field that deals with the huge growth in natural information. This has led to the development of important and optional databases of nucleic sequences, patterns of proteins, and structures. . Many applications use the WordNet lexical database in conjunction with clustering algorithms. Clustering algorithms using ontologies such as WordNet are presented by [2]. They acquaint lexical chains with a select subset of semantic features (core semantics) that define the document's theme and are useful for clustering in order to fully utilise the semantic data from Word-Net. The lexical chain is a technique for extracting a set of semantically linked terms from a page and capturing the main subject. However, while the proposed approach eliminated many of the problems associated with traditional clustering, it still has certain drawbacks, such as essential words not found in the WordNet lexicon not being reflected as concepts for similarity calculations. It can be improved by studying a wider database of knowledge, such as Wikipedia. [3] provide a novel multi-document summarizer that generates the summary using the Yago ontological knowledge base. They use Yago ontology to identify the essential notion using entity identification and disambiguation.

This stage can be improved even further by employing entropy-based sentence selection. The proposed method can be extended to include multilingual summary. Text Mining is an interdisciplinary field that employs both data mining and other disciplines like information extraction and retrieval, classification / classification information and text summary techniques. Typical text mining techniques include grading, clustering, extraction of information, recovery of information, categorization, ranking, extraction of subject subjects, etc. Documents are first prepared during text clustering and features are extracted using different techniques. The features retrieved are then grouped together. For feature extraction and text clustering, various methods are available. The key terms are generally extracted in words or phrases during the clustering process. The characteristics are first extracted using various formulas in the form of phrases. Then, the phrases are weighed and each document extracts a certain ratio of phrases with the highest weight. The summary phrases are then clustered with the WordNet seminal. The traditional class imbalance approach is used to solve the problem. A number of advanced methods are developed to achieve increased precision in standard classification approaches to solve the class-imbalanced problem. The cost-sensitive method of learning generally includes a cost matrix for all error or case categories. It aims mainly to facilitate learning from imbalanced data sets. In the process of sampling the minority class, this mechanism has an equivalent effect. It can end by overfitting training with

specific rules or regulations. ANN is a characterized statistical learning algorithm, based on the Neural Network. The neural network is a neuronal network to detect cases when activated. Approximation and feature assessment can be performed by looking for network sizes that take inputs. The interconnected neurons of ANNs may also, besides pattern recognition, be defined as input/output measurement in line with the machinery learnings (pattern recognition). K- The nearest neighbor (KNN) is an instance-based classification system to support kernel learning in anomaly repositories. The kernel design classification function KNN optimizes the anomaly pattern removal function [4].

This model thus has a phrase cluster through the relationship and classification process at document level. A number of relevant documents that would have been classed as low on a best match search will be grouped with other relevant documents (via inter-documental linkages), improving the efficiency of the IR system. When the cluster hypothesis is applied to a collection of papers, the relevant documents are well segregated from the non-related documents (i.e. grouped separately). The efficacy of hierarchical clustering can be measured using cluster searches that discover the cluster that best fits the query [5]. The number of clusters, documents, and the number, type, and dimension of characteristics accessible for the clustering technique are referred to as the document representation. The process of identifying the most effective subset of features to employ for clustering is known as feature selection. The application of one or more input alterations to generate additional important features is one of the extraction features [6]. These two strategies can be used to provide a set of characteristics suitable for clustering applications. A similarity function in pair form is commonly used to determine the document's similarity. There are numerous techniques to execute the text-clustering grouping phase. The clustering algorithms begin with as many clusters as there are records, with each cluster containing only one record.

The cluster pairs are then mixed one by one until the number of clusters is reduced to k. The pairs of merged clusters are the stages that are closest to each other. If the combination continues, it will result in a hierarchy of clusters with a single cluster having all of the records at the top. This traditional Hierarchy is in the form of a tree (a dendrogram). Individual elements are at one end of this hierarchy, whereas each element at the other end is a single cluster. Agglomerative hierarchical clustering algorithms can be described as gullible in an algorithmic sense. A wide range of agglomerative hierarchical clustering techniques has been proposed. Such hierarchical algorithms can be split into two method categories with relative ease. The first group consists of connection methods, single, complete, weighted and uneven average connection methods. These methods may be used to display a graph. The second hierarchical group of clustering methods are methods for specifying the cluster centers (as an average or a weighted average of the cluster member vectors). The central, median, and minimum variance methods [7] include these methods. Clustering is the responsibility of grouping together similar objects [8] Clustering was initially a method used in an information retrieval system to improve precision or to recall[9]. This is the best way to browse documents or organize search engine results in reply to user query or to help users rapidly identify and focus on the appropriate result set. Clustering has become the most efficient way to navigate

documents. The document clustering is an uncontrolled learning from unstructured textual data and it helps to improve efficiency in the retrieval of different information (IR) tasks. The hypothesis indicates that if a cluster document is relevant to a search request, other documents within the same cluster will also probably be relevant. This is because the clustering of documents is very important, because the clustering of good data results in the creation of good clusters, whereas the clustering of raw data produces poor clusters. Preprocessing involves removing undesirable text, removing stopwords, stemming, reducing functions, etc.

All of these steps lead to reduced dimensions and thus the clustering process is simplified. The representation in a common vector space of a set of documents is called the vector space model. Also known as a 'word bag,' the space model is supposed to appear in a separate mode with an immaterial order. Documents are shown as vectors for the characteristics representing the terms of the collection. A meaningful feature unit is called an index term in each word. Clustering documents of similar groups to create a consistent cluster. For accurate clustering it is necessary to define the closeness between a couple of objects in terms of a pair of wise similarities or distance. Similarity is accurately reflected in similarity measures between two data objects. Similarity measure is the main input to a clustering algorithm. Various similarity or distance measures lead to various cluster formations. Similarity or distance measurements are therefore very important for a grouping algorithm's outcome and success. The depictions of words do not deal with the lexical variation, semantic variation, syntactical variation and morphological variation. However, there is scope to explore these different similarity measures more closely with different clustering algorithms.

Text documents are treated as a sequence of words to overcome these problems. A suffix tree data structure maintains the sequential relationship between words and documents. A suffix tree is a data structure used to match and query strings efficiently. A large amount of suffixes were studied and used. It has been used in fundamental string issues such as strings, [10] text compression and the most time-consuming substring. Suffix trees are used because the search time is separate from the string length. A string suffix tree is just a compact selection of all its suffixes. The documents can be considered as word sequences instead of characters. Text document is considered to be a string of words in the STD model of a suffix tree document. Classification can take many forms, from fully automated human intervention systems[11] to semi-automatic systems that use a hybrid human machine approach. The structural modifications of airways may create remodeling of the wall and this may lead to decrease in luminal diameter. An effective gene selections algorithm (NMI), Correlation-based Selection Feature (CFS), and Particle Swarm Optimization (PSO) are proposed to be integrated into a set technique, and SVM with leave-one-out cross validation is used as a classification. Fix and Hodges, which is one of the simplest and most popular classification, was first introduced to the k-nearest neighbor method (k-NNN). As a learner, the k-NN has a simple strategy.

The k-NN classification has two phases, the first phase consists of the determination of the k-neighbors and the second phase of class definition using the neighbors. It maintains all training instances instead of generating an explicit

model. This takes a test example feature in a vector form and finds the Euclidean distance from each example to the vector representation. The closest sample to the test sample is called the closest neighbor. As the trained sample is in some sense the most similar in terms of our trial samples, it makes sense to allocate its class marking to the test sample. The removal of such negligible features by selecting adequate and fewer information elements helps to overcome the 'dimensionality curse' problem and helps to make learning more efficient. Feature selection techniques do not alter, but simply select a subset, the original representation of the variables. So, the original semantics of the variables are preserved. The methods of selection are described as filters, wrappers, embedded feature and hybrid methods based on the selection criteria. The filter methods are separate algorithms of categories [12]. They select the best feature subset only based on the distinctive data characteristics, such as distance, correlation, and consistency. They are using gene ranking statistical methods. Either uniform or multivariate filters are used. For each feature, univariate methods provide a value and multivariate measures take groups of functions or characteristic interaction into account.

Background

The traditional model based on a graph only takes into account the question of the sentence node and the sentence node relation. In this system, the prediction and the measurement sensitive to query is taken into consideration from sentence to sentence edge. Another model that uses TextRank with certain differences has been presented. This procedure uses a shortest way to generate the TextRank summary. The majority of work in multi-document classification is based on content coverage, centrality, or maximum marginal based approaches [13]. They conducted a thorough literature study on the biomedical text to discover distinct approaches, areas of application, research trends, and research gaps. They discovered that research has changed from single document ranking to multi-document ranking, and that interest in knowledge rich approaches is expanding faster than interest in knowledge poor methods. Future study on cognitive aspects of classification is needed to increase knowledge in this topic. Recent studies have focused on hybrid methodologies, which incorporate statistics, linguistics, machine learning, and language processing techniques. They proposed a method for automatically creating abstracts for biomedical research papers.

They analysed the appropriateness of text classification methodologies in terms of content coverage and human perspective using two approaches: extractive and abstractive. Despite advanced research in the biomedical domain, some research gaps must be filled for future studies, such as a heavy reliance on English literature, a scarcity of extrinsic evaluation studies, and a lack of standard corpora and reference standards to support the development of ranking in the biomedical domain. Another biomedical rankingapproach, [14] increased performance by incorporating domain knowledge into the Bayesian ranking process using the Unified Medical Language System (UMLS) paradigm. They used six features to extract relevant themes from biomedical text in their technique. In the biomedical field, the Unified Medical Language System (UMLS) has proven to be a helpful information source for ranking. The language of the document being

condensed must be mapped onto the ideas it contains when the UMLS is used. The usage of e-materials in education has grown dramatically in recent years. Many colleges and universities record their lectures and notes and make them available on the internet. One of the challenges with these materials is that they must be both basic and comprehensive. [15] introduced a classification application that solves the learner's problems and creates a simple summary. They came up with a collection of learner-dependent readability-related features for extracting essential sentences to help with this.

The majority of present text classification research is focused on short text documents. There are just a few studies in the literature that can summarise a long text document, such as a novel. To summarise the Novel documents, [16] developed a topic modelling based extraction approach. The underlying concept behind topic modelling is that the text is viewed as a collection of separate themes. Local and worldwide distribution of features or topics can be found in a long text novel. As a result, topic modelling is a good choice for novel text summary. The LDA topic modelling approach was utilised. Following the topic modelling process, all of the candidate phrases that are too large to fit in the compression ratio are extracted. Finally, the most important candidate sentences are chosen based on the repetition and topic diversity of each candidate sentence. By extracting semantic entities, the proposed method can be enhanced. [17] Developed a word-sentence co-ranking based technique to increase sentence scoring to address this issue. They argue that words and sentences have mutual impact in the sense that words appearing in high-scoring sentences gain high scores, and sentences containing high-scoring words should be regarded high-scoring sentences.

They integrate word-sentence relationships into a graph-based ranking algorithm in the proposed work so that mutual impact can more precisely supply the intrinsic status of words and sentences. For single document classification, the suggested co-ranking technique is used. It is possible to extend co-rank for multi document ranking. For text segment extraction, [18] presented a hybrid approach. There are two parts to how it works. In the first stage, a context-based features matrix is constructed, and segments are represented as nodes in the second stage. For rating text parts, they devised the Hopfied Network algorithm. Because semantic information is considerably closer to human perception, it can provide more valuable indications for sentence rating in ranking systems, thanks to recent advances in natural language processing and shallow semantic parsing techniques. Finally, a greedy algorithm is used to choose highly scored sentences for inclusion in the summary. Deep semantic information, rather than shallow semantic information, can be used to improve the current technique. They also presented a graph ranking model-based text classification technique. Their method consists of two steps. The first operation searches the document for cohesive chunks, while the second operation ranks the individual chunks and selects the phrases. [19] presented single document ranking based on graph.

It can be improved by incorporating several rule-based algorithms that can help resolve anaphora in sentences between nouns and pronouns. Graph-based approaches for multi-document extractive ranking, in particular, treated sentences as Bags of words and ignored the semantic structure of sentences and

semantic links between phrases. These solutions use a content closeness metric to determine sentence similarity, which may not be able to separate several sentences that are semantically equivalent. As a result, the final summary may include redundant information. [20]proposed a clustered genetic semantic graph-based approach that automatically groups comparable sentences and generates abstractive summaries via language generation. [21] provide an excellent overview of graph-based approaches in recent technical papers. They looked at the functional components and maturity of graph-based methods in the field of natural language processing and comprehension. They employed feature extraction criteria to train Maximum Entropy (ME), Naive-bayes, and support vector machines in their supervised model (SVM). Because text characteristics are language agnostic, it is possible to utilize different or additional features depending on the language and kind of text to improve the summary.

Proposed Model

In the proposed approach, a large set of multi-domain document sets are taken as input for keyphrase ranking and classification process. In this work, initially, biomedical document sets are filtered using the Stanford NLP library for non-special characters removal, tokenization, stemming etc. Gene-disease database is used to find the contextual similarity between the gene-disease to the contextual keyphrase word embedding vectors. Similarly, ICD drug codes are used to find the contextual similarity between the gene-disease, ICD code and keyphrases. Gene-disease, disease chemical contextual similarities are used to evaluate the document classification and ranking process for large biomedical document sets. The overview of the proposed framework is represented in figure 1. Here, an efficient particle swarm optimization(PSO) approach is used to select the key ranked features to the Bayesian graph based deep learning framework for document ranking process. The selected ranked documents are classified using the multi-class SVM classification framework on large document candidate sets.

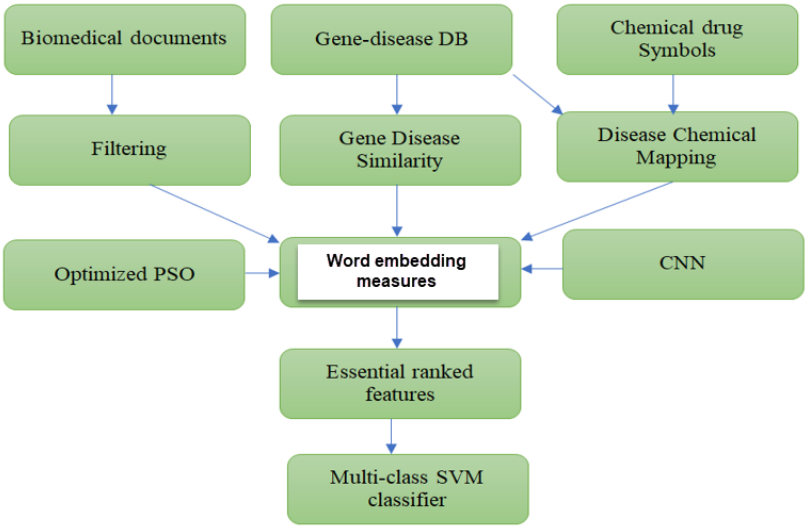

Figure1. Gene-disease and ICD based biomedical document ranking and classification

Proposed model is implemented in three phases:

- Data collection and contextual similarity computation
- Word embedding measures and contextual similarities for keyphrase extraction and ranking
- Phrase level and document level classification framework.

Phase#1: Data collection and contextual similarity computation

In this phase, biomedical documents are taken as input from the multiple domains such as pubmed and medline for document feature extraction, ranking and classification process. Initially, all the document sets are filtered using the Stanford NLP library for tokenization, stemming and contextual similarity between gene-disease and ICD codes.
The contextual dependency between the

$$\text{DependencyMeasure}(s1,s2) = \max\{\sum_{p=1} \text{sim}(\text{GV}(w_p), \text{PV})/|s1|, \sum_{p=1} \text{sim}(\text{ICDV}(w_p), \text{PV})/|s1|\}$$

$$\text{sim}(w, \text{PV}[]) = \sum_{i=1}^{|\text{PV}|} \text{prob}(w / \text{PV}[i]).\text{cosine}(w, \text{PV}[i])/|\text{PV}|$$

Phase 2: Word embedding and features extraction

In this phase, each biomedical document is processed to find the essential key terms using the glove optimization model and the contextual similarity measures.

- Step1: Find the word embedding vectors using proposed Glove optimization mdodel.
- Step 2: To each feature in the embedding vector , gene-disease and ICD codes, compute the contextual phrase similarity using the following formula. The following are the phrase wise and sentence wise contextual similarity ranking computations for the document ranking and classification process.

$$\text{WordVectorRank(WVR)}$$
$$\Pr(WV[]) = -\Pr ob(F_i).\log(Prob(F_i))$$
$$Ent(F, WV[]) = \sum_i \Pr(WV[i]/F)\log(\Pr(WV[i]))$$
$$WVR(WV[]) = (\sqrt[3]{Ent(F)} * |T| * CEntropy(w_a / F)) / Chisq(F)$$
$$\text{PhraseVectorRank(P VR)}$$
$$\Pr(PV[]) = -\Pr ob(p_a / F_i).\log(Prob(p_a / F_i))$$
$$PVR(PV[]) = \Pr(PV[]) * CEntropy(w_a / PV[]) / N;$$
$$\text{SentenceVectorRank(S VR)}$$
$$\Pr(SVR[]) = -\Pr ob(s_a / F_i).\log(Prob(s_a / F_i))$$
$$PVR(SVR[]) = \Pr(PV[]) * CEntropy(w_a / SV[]) * CEntropy(w_a / PV[]) / \max\{w_a / (PV \cap SV)\};$$

Contextual relationship between Glove vector and Gene sets

Let v1 denotes the count of gene sets in D.
Let v2 denotes the count of glove vector entities in D.
The contextual relationship between the glove vectors and gene-sets are given by

$$Sim(vc1[], vc2[]) = Prob(vc1/vc2)/|vc1|.|vc2|$$

Contextual relationship between Glove vector and ICD codes

Let w_G and w_{ICD} be the glove vectors and ICD codes then the contextual similarity is is given as

$$Sim(w_G, w_{ICD}) = \frac{|w_G \cap w_{ICD}|}{\max(w_G, w_{ICD})}$$

Phase-3: Phrase level and document level classification framework

Initially, document dependency graph DDG →(V, E) is constructed using the bayes network based edge set and vertex set as E and V. Here ,V is initialized with document features, disease and chemical drug vector sets and E is initialized with contextual similarity between the node vertices. The sample graph based disease relationships are extracted in a hierarchical manner as shown in figure 2.

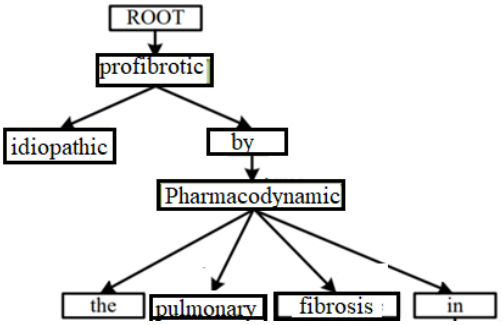

Figure 2. Sample graph based disease pattern extraction

Figure 2, describes the dependency parse tree of the sample biomedical gene disease pattern. In this example, the relationship between the disease and its chemical symbols are identified for feature extraction process.All the features in the PSO approach and graph based weighted tokens are given to word embedding process of CNN framework as shown in figure 3. The word embeddings for these words can be initialized to Glove model as pre-trained vectors.

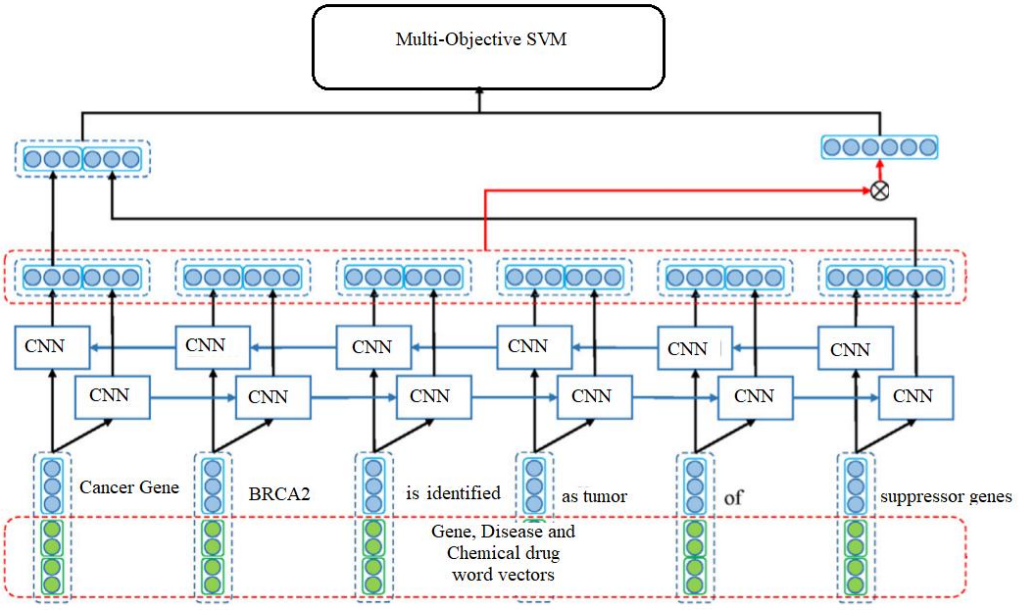

Figure 3. Proposed Graph based CNN framework for disease2drug prediction

Different filters with varying window sizes (here, only the height is different) slide over the full E rows in the convolution layer, i.e. the filter width is usually the same as the E width. Each filter conducts E-convolution, generating various function maps. A max-over-time pooling operation is applied to the elements located in the same feature map to extract the most important feature. Max-over-time pooling operation is performed on the word embedding features to find the essential key relationships among the gene disease based chemical drug terms as shown in Figure 3. In the proposed CNN framework, a multi-objective SVM is proposed to the classify the input tokens for chemical drug prediction based on the gene disease patterns.

Input the CNN word embedding features for data classification

For each feature set do

for each disease in GDP.

do

Apply SVM multi-class optimization models as

$$WV[] = P(c_i / t_j) = \frac{p(c_i)\prod_{\forall j} P(t_i / c_i) + |t_i \cap WV[]|}{p(t_j)}$$

$$PV[] = P(c_i / PV_j) = \frac{p(c_i)\prod_{\forall j} P(t_i / c_i) + |t_i \cap PV[]|}{p(t_j \cap PV_j)}$$

$$SV[] = P(c_i / SV_j) = \frac{p(c_i)\prod_{\forall j} P(t_i / c_i) + |t_i \cap SV[]|}{p(t_j \cap SV_j)}$$

Here, word level, phrase level and sentence level contextual conditional probabilities are computed in order to find the best optimization function in multi-class SVM model.

$$\min_{W_k, a_k} \frac{1}{2}\|W_k\|_1^2 + \tau_m + \sum_{i=1}^{l} a_i \left(y_i \left[ker < x, y > \cdot w + b \right] - 1 + \xi_i \right) - \sum_{i=1}^{l} \gamma_i \xi_i$$

$$s.t\ ker < x, y > \cdot w + b \geq 1 - \xi_i^n - \tau_m,$$

$$\xi_i^n > 0$$

$$\tau_m > 0; m = 1 \ldots classes$$

$$\xi_i = \max\{vc_i, pv_i, cv_i\}$$

Here kernel function ker(x,y) represents the kernel functions defined from gene disease vector space to chemical symbol vector space.

$$Ker < x, y > = e^{-\xi_i^n \log\left(\sum \| x - y \|^2\right)} \quad \text{if x==y}$$

$$=. e^{-\xi_i^n \log\left(\sum \| x - y \|^{1/2}\right)} \quad \text{if x<y}$$

$$= e^{-\xi_i^n \log\left(\sum \| y \|^2\right)} \quad \text{if x > y}$$

Step 4: Test data is predicted to the class y based on the largest decision values as

$$arg\ max\{ W_k^T D_i + b_k \}$$

Experimental setup and results

The proposed model is simulated in Amazon AWS cloud server with 32GB of RAM. In the AWS server, different word embedding model are used to select to find the best processing speed and accuracy on large datasets. In the proposed work, java based deep learning framework is implemented in the large AWS instance to filter the essential key patterns and classification models. In order to improve the classification speed, proposed model is simulated with multi-class classification model on arge features space. We evaluate the performance of our frameworks on different Gene dataset, ICD and biomedical documents. Here, different metrics such asaccuracy, Precision, and Area under the Curve (AUC) are used to evaluate the performance of gene disease and drug prediction.

Table1
Performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for accuracy computation

TestSample	Glove+RF	MI+SVM	PSO+SVM	ProposedModel
#1	0.9	0.94	0.92	0.96
#2	0.91	0.93	0.92	0.96
#3	0.92	0.92	0.9	0.96
#4	0.89	0.92	0.92	0.96
#5	0.89	0.94	0.92	0.97
#6	0.89	0.94	0.91	0.98
#7	0.9	0.9	0.92	0.96
#8	0.92	0.91	0.9	0.98
#9	0.9	0.91	0.91	0.96
#10	0.92	0.94	0.9	0.98
#11	0.93	0.93	0.9	0.96
#12	0.91	0.91	0.91	0.97
#13	0.9	0.91	0.91	0.98
#14	0.91	0.92	0.9	0.97
#15	0.93	0.92	0.91	0.96
#16	0.9	0.9	0.91	0.97
#17	0.91	0.92	0.91	0.97
#18	0.9	0.95	0.9	0.97
#19	0.93	0.9	0.92	0.98
#20	0.91	0.91	0.93	0.97
#21	0.9	0.94	0.92	0.97
#22	0.92	0.91	0.92	0.97
#23	0.91	0.92	0.92	0.96
#24	0.93	0.9	0.91	0.96
#25	0.92	0.92	0.92	0.97

Table 1, describes the performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for accuracy computation. As given in table, proposed framework has better accuracy than the conventional approaches for large biomedical document sets.

Table2
Performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for recall computation

TestSample	Glove+RF	MI+SVM	PSO+SVM	ProposedModel
#1	0.89	0.94	0.9	0.98
#2	0.9	0.94	0.91	0.97
#3	0.92	0.91	0.91	0.97
#4	0.91	0.9	0.92	0.98
#5	0.92	0.91	0.93	0.97
#6	0.92	0.95	0.91	0.97
#7	0.9	0.94	0.91	0.98
#8	0.91	0.94	0.9	0.96
#9	0.9	0.94	0.91	0.98
#10	0.9	0.91	0.91	0.97
#11	0.9	0.92	0.91	0.97
#12	0.92	0.92	0.92	0.97
#13	0.89	0.94	0.91	0.96
#14	0.91	0.91	0.91	0.98
#15	0.9	0.94	0.9	0.98
#16	0.9	0.93	0.93	0.96
#17	0.92	0.9	0.92	0.97
#18	0.91	0.91	0.92	0.98
#19	0.92	0.93	0.91	0.97
#20	0.93	0.91	0.93	0.98
#21	0.91	0.92	0.9	0.96
#22	0.93	0.9	0.9	0.97
#23	0.92	0.93	0.91	0.97
#24	0.91	0.94	0.9	0.98
#25	0.9	0.91	0.91	0.96

Table 2, describes the performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for recallcomputation. As given in table, proposed framework has better recall than the conventional approaches for large biomedical document sets.

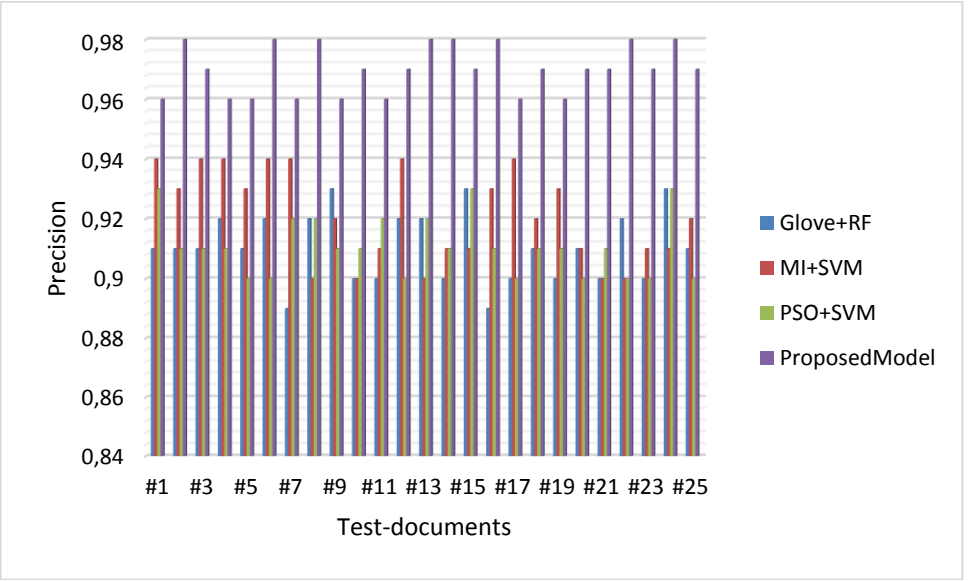

Figure 4. Performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for precision computation

Figure 4, describes the performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for recall computation. As given in figure, proposed framework has better precision than the conventional approaches for large biomedical document sets.

Figure 5. Performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for ROC computation.

Figure 5, describes the performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for recall computation. As given in figure, proposed framework has better ROC than the conventional approaches for large biomedical document sets.

Table 3
Performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for F-measure computation

TestSample	Glove+RF	MI+SVM	PSO+SVM	ProposedModel
#1	0.91	0.92	0.93	0.97
#2	0.9	0.91	0.91	0.98
#3	0.9	0.92	0.9	0.97
#4	0.9	0.91	0.93	0.96
#5	0.93	0.91	0.91	0.97
#6	0.93	0.92	0.93	0.97
#7	0.9	0.91	0.92	0.98
#8	0.9	0.92	0.93	0.97
#9	0.9	0.92	0.93	0.98
#10	0.92	0.94	0.92	0.97
#11	0.91	0.94	0.9	0.96
#12	0.91	0.92	0.9	0.96
#13	0.89	0.94	0.93	0.97
#14	0.92	0.93	0.9	0.96
#15	0.92	0.93	0.93	0.97
#16	0.91	0.9	0.91	0.97
#17	0.91	0.92	0.93	0.98
#18	0.91	0.94	0.9	0.98
#19	0.91	0.95	0.92	0.97
#20	0.91	0.94	0.91	0.98
#21	0.9	0.92	0.9	0.97
#22	0.92	0.9	0.93	0.96
#23	0.89	0.94	0.93	0.96
#24	0.92	0.9	0.9	0.96
#25	0.9	0.94	0.91	0.97

Table3, describes the performance of proposed gene-ICD based feature extraction and classification model to the conventional biomedical document classification models for F-measure computation. As given in figure, proposed framework has better F-measure than the conventional approaches for large biomedical document sets.

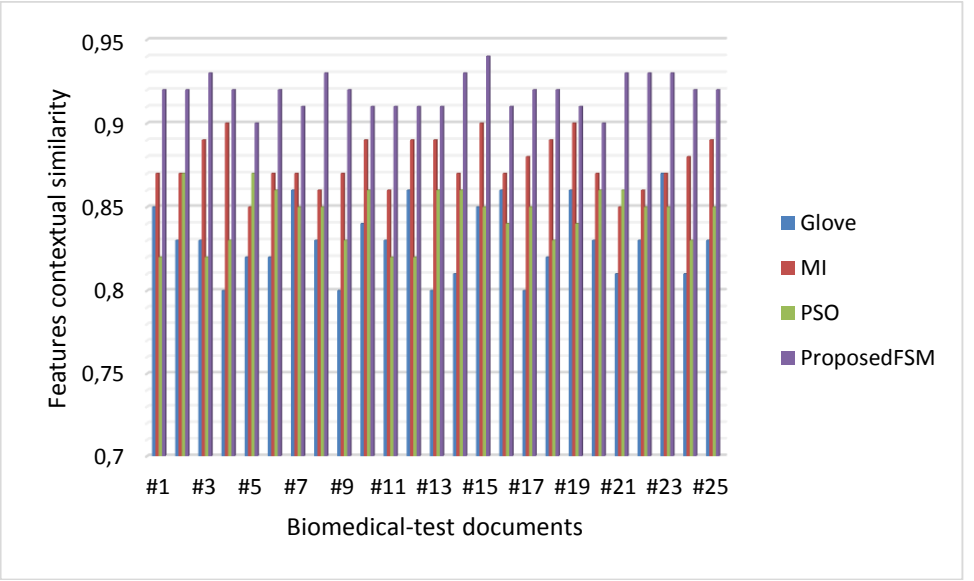

Figure 6. Performance of proposed gene-ICD based feature extraction model to the conventional biomedical document feature extraction measures

Figure 6, describes the feature extraction count of proposed approach to the state-of-art approaches on different biomedical document sets. In this figure, features count is evaluated using the k measure and genes, drugs and its contextual similarities.

Conclusion

Gene disease based contextual biomedical documents are used to find the essential key information for the decision making systems. Most of the conventional systems are difficult to extract the integrated gene and ICD based key features for the large biomedical document sets. Conventional word embedding models such as gain ratio, entropy, Glove, chi-square and probabilistic measures are used to find the essential key terms and its relationships using static gene disease databases. The main objective of the proposed work is to optimize the word embedding model along with the key-phrase ranking and classification. Most of the biomedical applications use pre-trained gene-disease database with limited number of gene names for key phrase ranking and classification process. In this work, an integrated gene-disease database and ICD drug database codes are used to train the model using the optimized SVM classification model and Bayesian estimation model. In this work, an ensemble learning model is integrated to CNN framework to classify the document sets using the gene disease database and ICD codes. Experimental results are implemented on biomedical documents using ICD codes and gene disease databases with different statistical metrics. Results show that the present framework has better true negative rate, error rate and precision than the stat of art algorithms.

References

1. M. Almagro, R. Martínez, S. Montalvo, and V. Fresno, “A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation,” Journal of Biomedical Informatics, vol. 94, p. 103207, Jun. 2019, doi: 10.1016/j.jbi.2019.103207.
2. M. Amith, Z. He, J. Bian, J. A. Lossio-Ventura, and C. Tao, “Assessing the practice of biomedical ontology evaluation: Gaps and opportunities,” Journal of Biomedical Informatics, vol. 80, pp. 1–13, Apr. 2018, doi: 10.1016/j.jbi.2018.02.010.
3. Y. Balarajan and M. R. Reich, “Political economy of child nutrition policy: A qualitative study of India’s Integrated Child Development Services (ICDS) scheme,” Food Policy, vol. 62, pp. 88–98, Jul. 2016, doi: 10.1016/j.foodpol.2016.05.001.
4. C. Cabot, S. Darmoni, and L. F. Soualmia, “Cimind: A phonetic-based tool for multilingual named entity recognition in biomedical texts,” Journal of Biomedical Informatics, vol. 94, p. 103176, Jun. 2019, doi: 10.1016/j.jbi.2019.103176.
5. D. Dinh, L. Tamine, and F. Boubekeur, “Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies,” Artificial Intelligence in Medicine, vol. 57, no. 2, pp. 155–167, Feb. 2013, doi: 10.1016/j.artmed.2012.08.006.
6. F. Duarte, B. Martins, C. S. Pinto, and M. J. Silva, “Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text,” Journal of Biomedical Informatics, vol. 80, pp. 64–77, Apr. 2018, doi: 10.1016/j.jbi.2018.02.011.
7. 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.
8. G. Harerimana, J. W. Kim, and B. Jang, “A deep attention model to forecast the Length Of Stay and the in-hospital mortality right on admission from ICD codes and demographic data,” Journal of Biomedical Informatics, vol. 118, p. 103778, Jun. 2021, doi: 10.1016/j.jbi.2021.103778.
9. M. A. Ibrahim, M. U. Ghani Khan, F. Mehmood, M. N. Asim, and W. Mahmood, “GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification,” Journal of Biomedical Informatics, vol. 116, p. 103699, Apr. 2021, doi: 10.1016/j.jbi.2021.103699.
10. A. G. Jácome, F. Fdez-Riverola, and A. Lourenço, “BIOMedical Search Engine Framework: Lightweight and customized implementation of domain-specific biomedical search engines,” Computer Methods and Programs in Biomedicine, vol. 131, pp. 63–77, Jul. 2016, doi: 10.1016/j.cmpb.2016.03.030.
11. I. Kamkar, S. K. Gupta, D. Phung, and S. Venkatesh, “Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso,” Journal of Biomedical Informatics, vol. 53, pp. 277–290, Feb. 2015, doi: 10.1016/j.jbi.2014.11.013.
12. A. Khalifa et al., “A qualitative investigation of biomedical informatics interoperability standards for genetic test reporting: benefits, challenges, and motivations from the testing laboratory’s perspective,” Genetics in Medicine,

vol. 23, no. 11, pp. 2178–2185, Nov. 2021, doi: 10.1038/s41436-021-01301-y.
13. L. Li et al., “Real-world data medical knowledge graph: construction and applications,” Artificial Intelligence in Medicine, vol. 103, p. 101817, Mar. 2020, doi: 10.1016/j.artmed.2020.101817.
14. J. Noh and R. Kavuluru, “Improved biomedical word embeddings in the transformer era,” Journal of Biomedical Informatics, vol. 120, p. 103867, Aug. 2021, doi: 10.1016/j.jbi.2021.103867.
15. O. Rouane, H. Belhadef, and M. Bouakkaz, “Combine clustering and frequent itemsets mining to enhance biomedical text summarization,” Expert Systems with Applications, vol. 135, pp. 362–373, Nov. 2019, doi: 10.1016/j.eswa.2019.06.002.
16. J. Sankhavara, R. Dave, B. Dave, and P. Majumder, “Query specific graph-based query reformulation using UMLS for clinical information access,” Journal of Biomedical Informatics, vol. 108, p. 103493, Aug. 2020, doi: 10.1016/j.jbi.2020.103493.
17. A. Sonabend W et al., “Automated ICD coding via unsupervised knowledge integration (UNITE),” International Journal of Medical Informatics, vol. 139, p. 104135, Jul. 2020, doi: 10.1016/j.ijmedinf.2020.104135.
18. L. Wang, P. J. Haug, and G. Del Fiol, “Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository,” Journal of Biomedical Informatics, vol. 69, pp. 259–266, May 2017, doi: 10.1016/j.jbi.2017.04.014.
19. 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.
20. X. Zhan, M. Humbert-Droz, P. Mukherjee, and O. Gevaert, “Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases,” Patterns, vol. 2, no. 7, p. 100289, Jul. 2021, doi: 10.1016/j.patter.2021.100289.
21. D. Zhao, J. Wang, Y. Chu, Y. Zhang, Z. Yang, and H. Lin, “Improving biomedical word representation with locally linear embedding,” Neurocomputing, vol. 447, pp. 172–182, Aug. 2021, doi: 10.1016/j.neucom.2021.02.071.