European Journal of Clinical Pharmacy
Print ISSN: 2385-409X **Online ISSN**: 3105-0409
Website: https://farmclin.com/

Original Research Article

An efficient Gene-ICD cluster based classification framework for heterogeneous micro-array databases

Article History:

Received: 05-09-2025
Revised: 25-10-2025
Accepted: 04-11-2025
Published: 18-12-2025

Name of Author:
Araja Raja Gopal,[1], Dr. M.H.M. Krishna Prasad [2]

Affiliation: *[1]Research Scholar, Department of Computer Science, Jawaharlal Nehru Technological University Kakinada , Andhra Pradesh, India*
[2]Professor, Department of Computer Science and Engineering, University College of Engineering Kakinada, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

Corresponding Author:
Araja Raja Gopal

Abstract: With the continuous rise in genes and diseases, determining key feature sets has become increasingly complex due to the large size of the data and associated sparsity challenges. As the number of biomedical entities and their chemical ICD classifications increases, the task of finding and extracting critical key phrases for microarray disease genes becomes more challenging due to the high computational and memory requirements. The high dimensionality of candidate sets poses one of the primary obstacles in extracting contextual key phrases. In this work, a hybrid biomedical document extraction model is designed to extract microarray disease patterns from extensive databases. A significant number of biomedical documents are extracted from real-time biomedical datasets to facilitate contextual key entity extraction. A hybrid word embedding technique, along with a similarity metric, is utilized to enhance the contextual similarity between document sets, leading to better key entity extraction. These extracted biomedical key entities are employed to map microarray data classification patterns for gene-to-ICD mapping. A hybrid classification model is applied to microarray cancer datasets to identify essential patterns for the biomedical document mapping process. Experiments conducted on different sets of biomedical documents from microarray datasets show that the gene-chemical disease cluster-based classification model achieves better optimization than conventional document embedding methods, similarity metrics, and classification approaches.

Keywords: large data, ICD, disease pattern, biomedical, micro-array, classification, document mapping

INTRODUCTION

As microarray datasets grow larger, identifying key features in these extensive feature spaces becomes increasingly complex due to the challenges of data size and sparsity. For scientific and biomedical researchers, the major challenge lies in ranking and classifying microarray features, which are characterized by high-dimensional feature spaces and limited sample sizes. Each microarray includes numerous identical DNA molecules, which help detect gene-related diseases. Techniques such as feature transformation, feature ranking, and data classification are crucial in effectively classifying high-dimensional data with high accuracy. Feature transformation normalizes data within specified ranges, improving feature ranking in large feature spaces. Traditional feature transformation methods, like log transformation and min-max normalization, generally ignore data distribution and outliers. Machine learning allows classifiers to learn decision-making rules from expert-labeled data, reducing costs and

improving scalability compared to fully manual systems. Research on medical data classification largely focuses on binary classifiers, where a classifier is built from positive and negative examples to predict class membership. For multiclass datasets, separate binary classifiers are often constructed for each class, and their results are combined. Classification can be fully automated or use a hybrid approach involving human intervention. Microarray datasets, particularly in chronic diseases, evolve gradually, from mild symptoms to severe disease and even death. Medical datasets often consist of cancer patches and their related diseases, which are difficult for doctors to detect in high-risk patients. A patient's medical history and previous knowledge also aid in disease detection. Structural changes in the airways can result in airway remodeling, reducing luminal diameter and contributing to conditions like emphysema, a chronic respiratory disease where alveolar walls are destroyed without fibrosis.

The PubMed database currently holds around 29 million articles, and approximately 1 million new articles are added each year. A substantial amount of essential information related to proteins, drugs, diseases, and chemicals is available in unstructured formats. This exponential increase in the volume of documents makes it difficult to manually collect and organize biomedical information such as protein-protein, drug-drug, and chemical-protein interactions. Biomedical information extraction is a process designed to automatically detect biomedical concepts and their relationships using advanced language processing and machine learning tools. Each year, researchers and healthcare professionals publish large numbers of biomedical research articles, most of which are accessible online. These articles are valuable not only for biomedical scientists in their research but also for healthcare professionals in their clinical work. The volume of information continues to grow across various domains due to the increase in distributed biomedical repositories. Document preprocessing is used to reduce peer documents into summaries by selecting essential information from the source. However, this has resulted in information overload. To address this, multi-document clustering and feature extraction can reduce inter-cluster variation. This research employs a feature extraction strategy and key phrase clustering and pattern discovery to minimize redundancy across multiple original documents. Text classification is a technique that helps automatically assess the significance of a document. Additionally, identifying synonyms and abbreviations is prioritized, adding complexity to biomedical literature. Hence, more efficient methods are needed to extract biomedical information from the ever-growing pool of resources. A suitable mining approach is essential for uncovering various types of knowledge from biomedical literature, where term variation is common, including numbers, capital letters, hyphens, and other special characters within terms.

Currently, the PubMed database contains roughly 29 million articles, with around 1 million new ones being added every year. A significant amount of important knowledge about proteins, drugs, diseases, and chemicals is stored in an unstructured form. The rapid growth of these documents makes manually collecting and organizing information, such as protein-protein, drug-drug, and chemical-protein interactions, highly complex. Biomedical information extraction refers to the automated process of identifying biomedical concepts and their relationships using advanced natural language processing and machine learning methods. Every year, biomedical researchers and healthcare professionals publish large quantities of biomedical research articles, the majority of which are available online. These articles aid both scientists in their research and healthcare professionals in their practice. The sheer volume of information continues to grow across various domains due to the expansion of distributed biomedical repositories. Document preprocessing is used to reduce peer-reviewed documents into summaries by selecting critical information. However, this has created the problem of information overload. To mitigate this, multi-document clustering and feature extraction are used to reduce inter-cluster variation. This research focuses on feature extraction techniques combined with key phrase clustering and pattern discovery to eliminate redundancy in multiple original documents. Text classification automatically assesses a document's importance. Additionally, emphasis is placed on identifying synonyms and term abbreviations, adding complexity to biomedical literature. As a result, there is a growing need for more efficient approaches to extract biomedical information from large sources. A suitable mining approach must be implemented to uncover diverse types of knowledge from biomedical literature, where there is significant variation in terms, including numbers, capital letters within words, hyphens, and special characters.

In the past, the accuracy of medical disease prediction has been significantly reduced when training datasets are small, due to class imbalance and high-dimensional data. In this method, each attribute is checked for missing values. A pattern mining and classification model was introduced for disease prediction using microarray datasets, where pathways were ranked, and disease-related patterns were identified. The random forest classification model was employed to filter and classify co-related disease patterns, but this approach requires considerable computational power and memory for larger datasets. Biomedical text data are a rich source of information. This paper described the use of the MapReduce method, a parallel and distributed programming paradigm, to mine relationships among various biomedical concepts extracted from literature. Initially, biomedical concepts are extracted using text matching with the Unified Medical Language System (UMLS), the most widely adopted standard biomedical

database. The MapReduce method is then implemented to evaluate specific interestingness measures. He et al. introduced a graph-based model for unstructured biomedical text, which can be applied in several real-world applications. To address these challenges, two new approaches, RT-TNG, were proposed. Topic N-Grams plays an important role in assigning these models. The researchers also examined semantic shifts in biomedical literature, analyzing word semantic changes in the biomedical domain. They identified representative words based on frequency and topic probability distributions, showing how words can cluster with their semantic neighbors and coevolve, or drift apart over time. Functional annotation of genes is a critical process, as it helps in understanding gene relationships. Multiple gene functions are described using standardized vocabularies, known as bio-ontologies. Assigning bio-ontology terms to genes is done through approaches based on data mining and machine learning methods, such as maximum entropy and support vector machines. The goal of this study is to propose an alternative method for annotating genes, including the development of efficient classification schemes, validation models, and graphical representations of outcomes. Reducing dataset dimensions is also a key concern, with classification schemes relying on linear discriminant analysis and validation models based on statistical interpretations.

2. Related works

Biswal et al. proposed an innovative approach that integrates medical entity relationships with keyword-based search, extending the personalized PageRank algorithm. In this extended keyword-based search system, user preferences act as filters, limiting relations in linked data. In their research, they presented the PRRank algorithm, which uses all relations present in linked data, with suggestions for further expansion. They also proposed a document-based graphical method, which clusters sentences based on relationships and applies ranking at the document level. A query-sensitive rating method was introduced for graph-based classification, improving traditional models that only accounted for queries to sentence nodes. This method estimates sentence-to-sentence edges and calculates responses on demand in the graph model. A variation of TextRank was provided, using the shortest path for generating summaries. Initially, the graph model was designed to represent documents and their connected sentence entities with meaningful relationships. A weighted graph system was also introduced to rate phrases and sentences for document classification. The word embedding model implemented here relies on preprocessed patterns and strict sentence identification to highlight significant subjects across various document lengths. A three-phase feature extraction method was used, consisting of preprocessing, soft clustering, and feature extraction. The clustering algorithm in this model involves four steps: vector space model initialization, similarity rank matrix determination, parameter initialization, and iterative development. The document extraction model combined document clustering with feature extraction. Preprocessing involved comparing sentences, identifying document features, feature positions, and similarity functions. Interest in word embedding has led to comparative studies in recent years. Scheepers et al. compared Word2Vec, FastText, and GloVe, but noted bias due to differences in training datasets. They also compared these models using the BLEU score without conducting individual evaluations, focusing on whether semantic relationships are preserved. Beam et al. created large word embeddings using medical data, focusing on Word2Vec and GloVe. Their benchmark involved statistical co-occurrence of concepts. Similarly, Huang et al. examined Word2Vec on three medical systems but did not focus on semantic connections. Wang et al. compared word embedding training for medical NLP tasks, focusing on models trained with the same data format. Word embeddings are used for intrinsic and extrinsic bio-NLP tasks: intrinsic tasks assess semantic similarity between words, while extrinsic tasks include relation extraction and text classification. Chiu et al. found that smaller windows benefit extrinsic tasks, while larger windows are better for intrinsic tasks, a finding confirmed in our tests. HAL generates semantic co-occurrences from phrases, while GLSA computes term and document vectors using an LSA-based method

3. Proposed Model

The overall architecture of the proposed model is represented in fig 1. Initially, each microarray gene disease dataset is processed to find the synonym of the gene feature for efficient gene-symbol to gene-name mapping. The overall architecture of the proposed model is represented in fig 1. Initially, each microarray gene disease dataset is filtered to fill the sparsity problem and missing values of the gene featured. Here, a hybrid data transformation approach is used to transform the feature values using the gaussian transformation measure. Each value is normalized to improve the balancing property of each feature and its class. In the initial phase, each feature is transformed using the gaussian transformation process. In the second phase, essential features are extracted using the hybrid PCA approach. Finally, an optimized decision tree classifier is designed to find the essential cancer patterns for prediction process. Proposed filtered based IPCA are integrated to improve the classification rate of the ensemble classification model with weak classifiers on high dimensional feature selection as shown in figure 1. Most of the ensemble classification technique is designed and implemented using the set of weak classifiers to optimize the overall classification rate and to minimize the error rate.

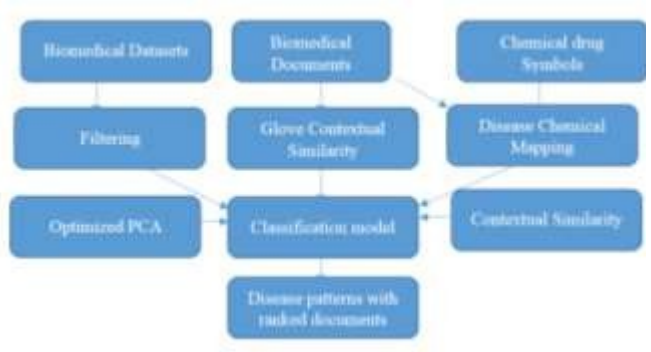

Figure 1: Multi-level Gene-Disease-Chemical drug classification and Ranking Framework

In the proposed approach, biomedical documents, gene disease database and chemical drug names are taken as input for biomedical document processing. Initially, biomedical documents are filtered using the Stanford parser in order to remove the noise and tokenization. Here, each document is converted to word2vector data for data normalization. A novel rank similarity model is used to find the essential gene to disease patterns for chemical drug mapping. Each microarray training dataset is pre-processed using the data transformation function to remove the variation among the data distribution.

High dimensional data transformation:

Input *:* Training dataset D, F(S): Feature sets.

Output*:* Non-linear normalization values.

Procedure:

1. Input dataset D with feature space F(S).
2. To each feature in the feature set F(S)
3. perform
4. Compute non-linear normalization to each feature values in the feature space as

$$\phi = \sqrt{2.\Pi.(1+e^{-F[i]^2})}$$

$$\text{Nonlinearnormalization}(F[i]) = \eta = (\frac{1}{(1+\phi)} \frac{1}{\cos(\sum \log(F[])})$$

5. If(η >0.75)
6. Then
7. Normalize each feature using Min-max normalization [η ,1]
8. Else
9. Normalize each feature using Min-max normalization with lower and upper bounds as 0 and 1.

$$x' = \frac{x - \min\_(x)}{\max\_(x) - \min\_(x)} * (R2 - R1) + R1$$

10. End if
11. Done

Proposed Algorithm 1: Biomedical document filtering

Input *:* Chemical drugs CD, Gene-Disease pattern GDP, Documents D.

Phase 1: Data Filtering on the gene-disease patterns and Biomedical documents.

Read gene-disease patterns GDP.
Read biomedical documents BD.
Read Chemical drugs CD.
for each gene-disease pattern g[i] in GDP
Do
To each gene-chemical document di in D
Do
Tok[]=NLPOTokenizer(di)
To each gene-chemical tokens in Tok[]
Do
Apply Stanford NLP stopword removal , stemming and other text preprocessing.
Gt[]={RemoveStopWords(g[i]),Remove NonspecialChars[g[i],Tokenizer(g[i])
BDt[]={RemoveStopWords(g[i]),RemoveNonspecialChars[g[i],Tokenizer(g[i])
Mapping (Gt,BDt) to DCi

DC_1	$Sim(Gt_1,BDt_1)$
DC_2	$Sim(Gt_2, BDt_2)$
….	……
DC_n	$Sim(Gtn, BDtn)$

Done
Done
Done

Phase 1, describes the data preprocessing of the gene tokens and biomedical documents. Stanford NLP parser is used to filter the input documents. Stemming, stop word removal and tokenization are performed on the input documents for similarity computation. Proposed similarity computation is used to find the contextual relationship among the genes and disease patterns to the chemical symbols .

Contextual Rank Similarity for the biomedical documents:

GloVe encodes significance in embedded space as vector offsets. In this Glove vector model, word co-occurrences are taken as vectors to find the main and contextual key word vectors for biomedical gene-disease relationships.

Proposed Glove Optimization algorithm:

Step-1: Let WCM is the word co-occurence matrix and F_{ij} is the probability estimation of i^{th} word that occur in context $j^{th}word$.

M_w Represents the Glove main word

C_w Represents the Glove contextual word

b_m defines the main word bias

b_c defines the contextual word bias value

$\Phi = \text{Min}\ \{ b_m\ b_c \}.\ D(M_w . b_m,\ C_W . b_c)$

Step2 : Cαonstruct the word pair constraints by using cost function and word weight as

CostContraint(CC)= a + b + c

$$\text{Where, } a = b_m M_w^T C_w$$

$$b = b_c M_w^T C_w$$

$$c = \phi.e^{\alpha} - (\log(F_{ij}) / \max\{b_m .\| M_w \|, b_c .\| C_w \|\}$$

$$\psi = \text{weight\_function}$$

$$WF(F_{ij}) = \begin{cases} (\max\{b_m, b_c\}.\frac{F_{ij}}{(F_{ij})_{max}})^{\alpha} & \text{if } b_m.F_{ij} < b_c.F_{MAX} \\ 1 & \text{otherwise} \end{cases}$$

Step -3: *Hybrid Glove cost function is defined* as, *Optimal Cost function* is shown below

$$i.e., \sum_{i=1}^{V}\sum_{j=1}^{V} \psi.(a+b+c / \max\{b_m .\| M_w \|, b_c .\| C_w \|\})^2$$

$$a = b_m M_w^T C_w$$

$$b = b_c M_w^T C_w$$

$$c = \phi.e^{\alpha} - (log(F_{ij})$$

2. Bio-Gene Rank similarity measure for Glove Key-phrase extraction

Input : Glove scientific corpus main vector SC, Glove scientific corpus contextual features .

Step 1: Read Optimized glove feature values.

Step 2*: Let* $GM_{(i)} \leftarrow (w_1, w_2, \ldots w_i)$ *represents glove key main feature i.*

$GC_{(j)} \leftarrow (w_1, w_2, \ldots w_j)$ *represents glove contextual feature j.*

Type equation here.

$$GM(i) = \sqrt{GM(w_1)^2 + GM(w_2)^2 \ldots . GM(w_i)^2}$$

$$GC(\text{j}) = \sqrt{GC(w_1)^2 + GC(w_2)^2 \ldots . GC(w_\text{j})^2}$$

$$GM(i).GC(\text{j}) = GM(w_1).GC(w_1) + GM(w_2).GC(w_2) \ldots + GM(w_i).GC(w_\text{j})$$

Biomedical ranked document dissimilarity index (BRDDI) is given as

$$BRDDI = \frac{Log(GM(w_i).GC(w_\text{j})) * \sqrt[3]{(|GM(i)| + |GC(j)|)}}{(|GM(i)| + |GC(\text{j})|)}$$

Where $i \neq j$

Contextual BRDDI is given as $CBRDDI = 1 - BRDDI;$

In this bio-gene rank similarity measure , the similarity between the each document and the optimal glove features are computed to each document. Here, each document with highest similarity value is taken as highest ranked document in the context of genes/chemicals/proteins.

Algorithm :TopKPCA (TKPCA):

Input: Microarray-training data.

Output: Principal components for feature selection.

Step-1*:* Read input pre-processed data D'.

Step-2: Evaluate the co-variance computation on each feature to its co-related features as

$$CV(F[]) = \frac{\sum_{i=1}^{n}(F[x_i] - \mu_{F[x_i]})(Fy_i] - \mu_{F[y_i]})}{(n-1)}$$

Step-3: Find the ranked composite Eigen score to each feature in the feature list as

Eigen_scores[]= $\text{Det}(\lambda I - CV(F[])) = 0$ -

$$\text{OptimalEScore} = \frac{\sum F[])}{\{\max\{EV[]\}, 0.5 * (\text{Min}\{\text{mean}(F[i]), \text{mean}(F[j])\}\}}$$

Step-4: Highest ranked features are selected as principal components.

Improved EM Gene-disease clustering approach :

In the expectation maximization model, two phases are implemented on the training data to predict the best clustered features for the gene-disease prediction.

Expectation phase(E-phase) : In the expectation phase, model parameters are estimated using the hybrid probabilistic measure. This probabilistic measure is used to find the essential key features for the gene-disease clustering.

Let $\hat{\theta}$ represent the novel posterior estimation parameter used to predict the occurrence of gene disease pattern in the given large number of training samples.

$\text{Prob}(\hat{\theta}/c)$ is the occurrence of gene-disease new patterns in the large category of disease classes.

In the expectation phase, the maximization of the gene-disease patterns in all the training real-time datasets are given as :

$$\text{Prob}(\hat{\theta}/c_i) = \frac{\max\{\sigma_k(GD_i, C_i), \log(GD_i . C_j)\} . \sum_{i=1}^{N} P(c_i / D(GD)_i)}{|N| . P(c_i / GD)}$$

Maximization phase: In the maximization phase, model parameters are estimated using the gene-disease and its class patterns. In the maximization step, the probability of data occurrence in the given disease class is given as

$$\text{Prob}(c / GD_i, \hat{\theta}) = \frac{P(GD_i / c_i).\log(|c|).P(\hat{\theta})}{\text{Prob}(c/D_i)} .$$

These two phases are repeated until the number of maximum iterations or no change in the error rate.

Proposed Classification Algorithm

1: Read pre-processing gene-disease patterns, training cancer datasets, gene database. All these input patterns are partitioned 'm' clusters based on EM approach.

2: To each clustered

3: do

4: Apply proposed ensemble decision tree model on each cluster data.

5: In the proposed classification model, a novel gene-disease feature selection measure is implemented on each cluster.

Proposed decision tree gene-disease feature selection measure

Statistical Hoeffding entropy measure

Let D_p represents the clustered gene-disease probabilistic patterns for the decision tree classification problem. The hoeffding entropy of the gene-disease pattern analysis is given by

$$n = \sum D_p[i] . \log(D_P[i])$$

$$\text{Ent}(D_p) = \frac{(n + D_p[i] . \min\{D_p\}\{(\sum \log (D_p [i])}{\sqrt[3]{(\sum \log (D_p[i]) . \mu_{D_p[i]})}}$$

2. Proposed Random forest Micro-array feature selection

$$\text{HRTASM} = \frac{-n * \sqrt[3]{\text{Ent}(D_p)}}{(D_p[i] . \text{cheby}(D_p))^3}$$

4. Experimental Results

Experimental results are performed on real-time medline biomedical documents and micro-array datasets. Proposed feature selection-based ensemble methods increase the efficiency of the F-measure , recall and accuracy on high dimensional datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore, the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models. Simulation results represent the proposed ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features. Different types of cancer datasets and its types are presented in below table.

Experimental results are implemented in java environment with third party libraries for word embedding and NLP pre-processing. The minimum hardware configuration include amazon AWS large instance with 32GB RAM.

Micro array Datasets	**Gene sets**	**Data-Type**
Prostate	2136	Continuous/Numeric
Lymphoma	5000	Continuous/Numeric
DLBCL-Stanford	4000	Continuous/Numeric
Breast cancer	24481	Continuous/Numeric
Leukemia	7129	Continuous/Numeric

Table 1: Microarray datasets used

In this experimental study, Glove word embedding model is developed to improve the efficiency of biomedical disease prediction. Glove model contains 5 billion vocabulary tokens in order to find and extract the key terms in the biomedical disease document sets. Glove provides the words in vector format.

The following table represents the different gene-disease patterns and its probabilistic scores.

Probabilistic Measure :===> 1) debrisoquin, 2) dextromethorphan. 3) 0.016757323111589797
Probabilistic Measure :===> 1) (CYP2D1) 2) and/or 3) 0.031250243772462445
Probabilistic Measure :===> 1) humans. 2) and 3) 0.01704255330910507

Probabilistic Measure :===> 1) accumulation 2) significant 3) 0.03402329107350634
Probabilistic Measure :===> 1) in 2) liver 3) 0.0341996286373331
Probabilistic Measure :===> 1) prepared 2) O-demethylation 3) 0.026918506928464828
Probabilistic Measure :===> 1) CYP2D6 2) also 3) 0.04208105902246065
Probabilistic Measure :===> 1) recirculating 2) perfusion 3) 0.03852092261602163
Probabilistic Measure :===> 1) CYP2D1. 2) the 3) 0.035517394512509924
Probabilistic Measure :===> 1) and 2) perfused 3) 0.029370642784192734
Probabilistic Measure :===> 1) recirculating 2) a 3) 0.02237212773524969
Probabilistic Measure :===> 1) recirculating 2) and 3) 0.01697450410974184
Probabilistic Measure :===> 1) dextromethorphan 2) role 3) 0.02551078103008535
Probabilistic Measure :===> 1) different 2) in 3) 0.009957266362554885
Probabilistic Measure :===> 1) Lewis 2) we 3) 0.009711337717649466
Probabilistic Measure :===> 1) competitively 2) metabolism 3) 0.014289138825513468
Probabilistic Measure :===> 1) during 2) concentrations 3) 0.03071846144710269
Probabilistic Measure :===> 1) was 2) microsomes 3) 0.027411477372297703
Probabilistic Measure :===> 1) clearance 2) difference 3) 0.020220480479656293
Probabilistic Measure :===> 1) after 2) a 3) 0.02237214156633587
Probabilistic Measure :===> 1) experiment. 2) same 3) 0.02835809646084604
Probabilistic Measure :===> 1) no 2) significance 3) 0.04545931276339734
Probabilistic Measure :===> 1) was 2) time 3) 0.0045656736620106105
Probabilistic Measure :===> 1) There 2) accumulation 3) 0.03285413761420369
Probabilistic Measure :===> 1) vs. 2) clearance 3) 0.01740911483253628
Probabilistic Measure :===> 1) and 2) rats 3) 0.041312535208022555
Probabilistic Measure :===> 1) but 2) 1) 3) 0.044154364972400544
Probabilistic Measure :===> 1) experiments 2) clearance 3) 0.017483435208418104
Probabilistic Measure :===> 1) metabolism 2) in 3) 0.009957280419487218
Probabilistic Measure :===> 1) rat 2) liver 3) 0.034364430110329074
Probabilistic Measure :===> 1) 1) 2) 0.05 3) 0.0019160794148305705
Probabilistic Measure :===> 1) microsomes 2) from 3) 0.013371495253140126
Probabilistic Measure :===> 1) 1) 2) a 3) 0.022372118523637224

Probabilistic Measure :===> 1) role 2) O-demethylation. 3) 0.019050595490488043
Probabilistic Measure :===> 1) difference 2) repeat 3) 0.040407050521106186
Probabilistic Measure :===> 1) the 2) pathways 3) 0.030752212770722303
Probabilistic Measure :===> 1) 1.61 2) 0.27 3) 0.006091829903285804
Probabilistic Measure :===> 1) the 2) CYP2D1. 3) 0.004713617801154348
Probabilistic Measure :===> 1) liver 2) during 3) 0.01273991715751172
Probabilistic Measure :===> 1) 2) whether 3) 0.02706645111133696
Probabilistic Measure :===> 1) dextromethorphan 2) kinetics 3) 0.01606715063602239
Probabilistic Measure :===> 1) during 2) clearance 3) 0.01748341453648163
Probabilistic Measure :===> 1) or 2) 4-hydroxydebrisoquin 3) 0.024327187086899
Probabilistic Measure :===> 1) significance 2) no 3) 0.012513772458955916
Probabilistic Measure :===> 1) in 2) debrisoquin 3) 0.026905057597800396
Probabilistic Measure :===> 1) to 2) 1.61 3) 0.03610630646277621
Probabilistic Measure :===> 1) that 2) pathways 3) 0.03075219893613819
Probabilistic Measure :===> 1) recirculation, 2) and 3) 0.016974505583242618
Probabilistic Measure :===> 1) that 2) isozyme 3) 0.03541973365076106
Probabilistic Measure :===> 1) 4-hydroxydebrisoquin 2) accumulation 3) 0.03334901614898441
Probabilistic Measure :===> 1) on 2) not 3) 0.028836521968342934
Probabilistic Measure :===> 1) an 2) have 3) 0.02980160681314787
Probabilistic Measure :===> 1) +/- 2) returned 3) 0.019857838662197348
Probabilistic Measure :===> 1) used 2) we 3) 0.009367745834280098
Probabilistic Measure :===> 1) human 2) 600 3) 0.011798184322111372
Probabilistic Measure :===> 1) of 2) accumulation 3) 0.03318439704632819
Probabilistic Measure :===> 1) drop 2) clearance 3) 0.01746026623611775
Probabilistic Measure :===> 1) 2) 2) (p 3) 0.03707077673088432
Probabilistic Measure :===> 1) but 2) during 3) 0.013355795898422257
Probabilistic Measure :===> 1) 2) oxidative 3) 0.015174759750541516
Probabilistic Measure :===> 1) (clearance 2) to 3) 0.00668491611749443
Probabilistic Measure :===> 1) to 2) metabolism 3) 0.014476406005253475
Probabilistic Measure :===> 1) clearance 2) +/- 3)

0.04051783763292883
Probabilistic Measure :===> 1) 0.27 2) 2) 3) 0.04394924042132791
Probabilistic Measure :===> 1) and/or 2) pathways 3) 0.03104330901142606
Probabilistic Measure :===> 1) in 2) O-demethylation. 3) 0.01888269410679679
Probabilistic Measure :===> 1) activity 2) the 3) 0.03502834128096228
Probabilistic Measure :===> 1) human 2) livers 3) 0.0039017528235280366
Probabilistic Measure :===> 1) 4-hydroxydebrisoquin, 2) oxidative 3) 0.01543822022227091
Probabilistic Measure :===> 1) 2) studies 3) 0.04509673843608276
Probabilistic Measure :===> 1) CYP2D6; 2) active 3) 0.008704046535286426
Probabilistic Measure :===> 1) to 2) 3.21 3) 0.02753624854057324
Probabilistic Measure :===> 1) rats 2) cytochrome 3) 0.031727762942233076
Probabilistic Measure :===> 1) to 2) by 3) 0.015650391417226888
Probabilistic Measure :===> 1) 0.27 2) 1.61 3) 0.03610628624319749
Probabilistic Measure :===> 1) livers 2) rat 3) 0.03625282887952381
Probabilistic Measure :===> 1) in 2) Lewis 3) 0.01799540863826695
Probabilistic Measure :===> 1) 1) 2) significance 3) 0.04545927294073733
Probabilistic Measure :===> 1) in 2) O-demethylation 3) 0.02756515869231954
Probabilistic Measure :===> 1) observed 2) a 3) 0.022558871688294763
Probabilistic Measure :===> 1) CYP2D1 2) by 3) 0.016158980617806344
Probabilistic Measure :===> 1) rat 2) Lewis 3) 0.017995405572371236
Probabilistic Measure :===> 1) 4-hydroxydebrisoquin 2) dextromethorphan 3) 0.0036419783779478867
Probabilistic Measure :===> 1) 4-hydroxydebrisoquin 2) suggest 3) 0.045164547036482185
Probabilistic Measure :===> 1) rat 2) human 3) 0.030542241076616967
Probabilistic Measure :===> 1) to 2) ml/min 3) 0.0074281717418085026
Probabilistic Measure :===> 1) of 2) livers 3) 0.004601257825472651
Probabilistic Measure :===> 1) has 2) CYP2D6 3) 7.651902383048919E-4
Probabilistic Measure :===> 1) of 2) period 3) 0.02387548703107695
Probabilistic Measure :===> 1) of 2) that 3) 0.03599273343326564
Probabilistic Measure :===> 1) vs. 2) 3, 3) 0.013342389324211059
Probabilistic Measure :===> 1) pathways 2) that 3) 0.036510900272514316
Probabilistic Measure :===> 1) on 2) inhibitory 3) 0.0452700738269872
Probabilistic Measure :===> 1) rat 2) and/or 3) 0.03125021589434711
Probabilistic Measure :===> 1) have 2) debrisoquin 3) 0.0269345156779795
Probabilistic Measure :===> 1) using 2) perfusion 3) 0.03835793237676289
Probabilistic Measure :===> 1) no 2) ml/min 3) 0.007676688112971564
Probabilistic Measure :===> 1) of 2) a 3) 0.02255885264224118
Probabilistic Measure :===> 1) liver 2) human 3) 0.030117243650341137
Probabilistic Measure :===> 1) a 2) of 3) 0.008324291813649439
Probabilistic Measure :===> 1) liver 2) using 3) 0.04000352092450912
Probabilistic Measure :===> 1) other 2) (a) 3) 0.02926171962477493
Probabilistic Measure :===> 1) to 2) was 3) 0.006869843685163282
Probabilistic Measure :===> 1) and/or 2) 4-hydroxydebrisoquin 3) 0.024327168580840462
Probabilistic Measure :===> 1) debrisoquin. 2) metabolites 3) 6.386882842580482E-4
Probabilistic Measure :===> 1) from 2) human 3) 0.03005711400226832
Probabilistic Measure :===> 1) To 2) 3) 0.003031766505710992
Probabilistic Measure :===> 1) perfusion 2) the 3) 0.03518417956626145
Probabilistic Measure :===> 1) and 2) livers 3) 0.00459507860133359
Probabilistic Measure :===> 1) 2) 2) 0.27 3) 0.006405365999146619
Probabilistic Measure :===> 1) from 2) and 3) 0.01753118257694162
Probabilistic Measure :===> 1) liver 2) recirculation, 3) 0.0013786651370134612
Probabilistic Measure :===> 1) Cytochrome 2) also 3) 0.042080925603021874
Probabilistic Measure :===> 1) and 2) inhibition 3) 0.008872587415141
Probabilistic Measure :===> 1) by 2) to 3) 0.006831600284393899
Probabilistic Measure :===> 1) studies, 2) 3) 0.003031766535118025
Probabilistic Measure :===> 1) after 2) 1) 3) 0.04490958518272891
Probabilistic Measure :===> 1) Results 2) 3) 0.003358008285907023
Probabilistic Measure :===> 1) human 2) from 3) 0.013489559129636503
Probabilistic Measure :===> 1) (clearance 2) no 3) 0.012302025636921371
Probabilistic Measure :===> 1) plays 2) role 3) 0.02552132273456194

Probabilistic Measure :===> 1) clearance 2) recirculation, 3) 9.208618457334544E-4
Probabilistic Measure :===> 1) perfused 2) from 3) 0.013489560254334483
Probabilistic Measure :===> 1) (CYP2D1) 2) rat 3) 0.036555683597528955
Probabilistic Measure :===> 1) perfusion 2) liver 3) 0.03446153573252217
Probabilistic Measure :===> 1) +/- 2) 3.21 3) 0.027536248911729372
Probabilistic Measure :===> 1) (a) 2) suggest 3) 0.04511170601675249
Probabilistic Measure :===> 1) (clearance 2) 0.57 3) 0.001365409879737494
Probabilistic Measure :===> 1) of 2) and 3) 0.017145990595662845
Probabilistic Measure :===> 1) and 2) during 3) 0.013533491548454485
Probabilistic Measure :===> 1) 4-hydroxydebrisoquin, 2) sequentially 3) 0.02085524514472007
Probabilistic Measure :===> 1) to 2) time 3) 0.004565683886104804
Probabilistic Measure :===> 1) 0.57 2) 3.27 3) 0.02453791946594544
Probabilistic Measure :===> 1) and/or 2) (a) 3) 0.02882743826051393
Probabilistic Measure :===> 1) inhibition 2) rat 3) 0.0370137146828839
Probabilistic Measure :===> 1) (CYP2D1) 2) that 3) 0.036629129642514545
Probabilistic Measure :===> 1) plays 2) also 3) 0.042093095888918894
Probabilistic Measure :===> 1) effect 2) on 3) 0.039743297473654855
Probabilistic Measure :===> 1) significance 2) clearance 3) 0.017464627347461195
Probabilistic Measure :===> 1) the 2) due 3) 0.042131494774954925
Probabilistic Measure :===> 1) system 2) in 3) 0.008647082106755102
Probabilistic Measure :===> 1) that 2) (CYP2D1) 3) 0.025737918389834972
Probabilistic Measure :===> 1) but 2) dextromethorphan. 3) 0.01729590504054165
Probabilistic Measure :===> 1) dextromethorphan 2) of 3) 0.009956901182787012
Probabilistic Measure :===> 1) human 2) different 3) 0.0077085880245435535
Probabilistic Measure :===> 1) dextromethorphan 2) 4-hydroxydebrisoquin 3) 0.02431117708048031
Probabilistic Measure :===> 1) 4-hydroxydebrisoquin, 2) catalyzed 3) 0.03513077685316922
Probabilistic Measure :===> 1) in 2) drop 3) 0.03891456872588973
Probabilistic Measure :===> 1) competitively 2) dextromethorphan 3) 0.0038314394595462885
Probabilistic Measure :===> 1) perfusion 2) and 3) 0.017145978082207344
Probabilistic Measure :===> 1) in 2) by 3) 0.01626249414161305
Probabilistic Measure :===> 1) nonrecirculating 2) fell 3) 0.013214952151228457
Probabilistic Measure :===> 1) nonrecirculating 2) a 3) 0.02200754727444746
Probabilistic Measure :===> 1) suggest 2) These 3) 0.04077125347459667
Probabilistic Measure :===> 1) microsomes 2) in 3) 0.008647081121832025
Probabilistic Measure :===> 1) to 2) clearance 3) 0.017297767813539647
Probabilistic Measure :===> 1) the 2) dextromethorphan 3) 0.0033472118426880813
Probabilistic Measure :===> 1) primarily 2) not 3) 0.028481948352137534
Probabilistic Measure :===> 1) to 2) of 3) 0.009095839677300654
Probabilistic Measure :===> 1) 0.05 2) than 3) 0.030520326881962625
Probabilistic Measure :===> 1) (b) 2) CYP2D6; 3) 0.03452007646189757
Probabilistic Measure :===> 1) site 2) of 3) 0.008466478608269068
Probabilistic Measure :===> 1) have 2) of 3) 0.008466460459625553
Probabilistic Measure :===> 1) inhibition 2) livers 3) 0.004585182184622583
Probabilistic Measure :===> 1) difference 2) clearance 3) 0.017485616467484394
Probabilistic Measure :===> 1) due 2) accumulation 3) 0.033540751633634895
Probabilistic Measure :===> 1) but 2) debrisoquin, 3) 0.004125857199692327
Probabilistic Measure :===> 1) in 2) with 3) 0.023493595189035304
Probabilistic Measure :===> 1) using 2) nonrecirculating 3) 0.034113581780193504
Probabilistic Measure :===> 1) from 2) perfused 3) 0.029313501832792056
Probabilistic Measure :===> 1) dependent 2) activity 3) 0.043057778969426766
Probabilistic Measure :===> 1) experiments 2) liver 3) 0.03402933508460086
Probabilistic Measure :===> 1) 4-hydroxydebrisoquin, 2) used 3) 0.007849993133374214
Probabilistic Measure :===> 1) In 2) preliminary 3) 0.012262104174821764
Probabilistic Measure :===> 1) period 2) 30-min 3) 0.004919045977369762
Probabilistic Measure :===> 1) +/- 2) 0.46 3) 0.007935173139009797
Probabilistic Measure :===> 1) rat 2) specificity 3) 3.440825133190347E-4
Probabilistic Measure :===> 1) and 2) 4-hydroxydebrisoquin. 3) 0.0239700819843943
Probabilistic Measure :===> 1) (b) 2) CYP2D1 3) 0.017229034554826037

Probabilistic Measure :===> 1) whether 2) determine 3) 4.767718864389431E-4
Probabilistic Measure :===> 1) on 2) the 3) 0.03527805494148346
Probabilistic Measure :===> 1) kinetics 2) the 3) 0.03527802652498218
Probabilistic Measure :===> 1) other 2) of 3) 0.00838791759913072
Probabilistic Measure :===> 1) active 2) human 3) 0.030057114825388847
Probabilistic Measure :===> 1) CYP2D6; 2) CYP2D1 3) 0.017325916605261088
Probabilistic Measure :===> 1) clearance 2) a 3) 0.021336769877735133
Probabilistic Measure :===> 1) perfusate 2) liver 3) 0.03400411587060772
Probabilistic Measure :===> 1) debrisoquin, 2) experiments 3) 0.038602277096001725
Probabilistic Measure :===> 1) or 2) microsomes 3) 0.02774128890875547
Probabilistic Measure :===> 1) a 2) showed 3) 0.002315140542431605
Probabilistic Measure :===> 1) 0.05 2) vs. 3) 0.013904641550385373
Probabilistic Measure :===> 1) +/- 2) ml/min 3) 0.007950261808892285
Probabilistic Measure :===> 1) this 2) change 3) 0.03265489676861767
Probabilistic Measure :===> 1) data 2) that: 3) 0.032548268227874884
Probabilistic Measure :===> 1) returned 2) +/- 3) 0.04041172700243104
Probabilistic Measure :===> 1) 3.21 2) to 3) 0.007444625689640301
Probabilistic Measure :===> 1) of 2) sequentially 3) 0.020879162472258968
Probabilistic Measure :===> 1) by 2) 4-hydroxydebrisoquin 3) 0.02495033735321166
Probabilistic Measure :===> 1) dependent 2) primarily 3) 0.013818884332927389
Probabilistic Measure :===> 1) CYP2D6 2) substrate 3) 0.03072516578515834
Probabilistic Measure :===> 1) due 2) the 3) 0.0352779769231048
Probabilistic Measure :===> 1) clearance 2) rat 3) 0.03681433198398012
Probabilistic Measure :===> 1) +/- 2) to 3) 0.007444622600670097
Probabilistic Measure :===> 1) clearance 2) vs. 3) 0.013904649753307342
Probabilistic Measure :===> 1) metabolites 2) other 3) 0.04489370268045357
Probabilistic Measure :===> 1) less 2) than 3) 0.03052032627933622
Probabilistic Measure :===> 1) metabolism 2) human 3) 0.030358718412542066
Probabilistic Measure :===> 1) by 2) human 3) 0.030358688633611648
Probabilistic Measure :===> 1) 3.27 2) in 3) 0.009376713886608542

Probabilistic Measure :===> 1) from 2) rat 3) 0.03651830795202204
Probabilistic Measure :===> 1) by 2) 4-hydroxydebrisoquin, 3) 0.018969474170877493
Probabilistic Measure :===> 1) 2) of 3) 0.008387919140109548
Probabilistic Measure :===> 1) clearance 2) experiments 3) 0.037931416032908255
Probabilistic Measure :===> 1) human 2) by 3) 0.01575995267417451
Probabilistic Measure :===> 1) in 2) presence 3) 0.04065580692834076
Probabilistic Measure :===> 1) are 2) the 3) 0.03498161404334103
Probabilistic Measure :===> 1) human 2) CYP2D6 3) 9.27822036768218E-4
Probabilistic Measure :===> 1) used. 2) nonrecirculating 3) 0.034658986829697855
Probabilistic Measure :===> 1) perfused 2) Lewis 3) 0.01772371367144879
Probabilistic Measure :===> 1) debrisoquin 2) of 3) 0.008368780298343516
Probabilistic Measure :===> 1) clearance 2) 1) 3) 0.044909564630458046
Probabilistic Measure :===> 1) CYP2D6 2) 3) 0.0031462357435050915
Probabilistic Measure :===> 1) that 2) and/or 3) 0.029972550779882468
Probabilistic Measure :===> 1) dependent 2) on 3) 0.0397432898142729
Probabilistic Measure :===> 1) 30-min 2) a 3) 0.02134882527603823
Probabilistic Measure :===> 1) specificity 2) rat 3) 0.03667167321183359
Probabilistic Measure :===> 1) microsomes 2) microM. 3) 0.019524058132239813
Probabilistic Measure :===> 1) showed 2) liver 3) 0.03400409238907634
Probabilistic Measure :===> 1) 2) studied 3) 0.034212829924723354
Probabilistic Measure :===> 1) 1.61 2) to 3) 0.007444646107128894
Probabilistic Measure :===> 1) ml/min 2) (clearance 3) 0.020096984272389116
Probabilistic Measure :===> 1) liver 2) metabolism 3) 0.014476405962326739
Probabilistic Measure :===> 1) debrisoquin 2) perfused 3) 0.029996820458973415
Probabilistic Measure :===> 1) undergoes 2) oxidative 3) 0.016415529134825538
Probabilistic Measure :===> 1) change 2) in 3) 0.00930380533137157
Probabilistic Measure :===> 1) recirculation, 2) perfusate 3) 0.025529834982747364
Probabilistic Measure :===> 1) oxidative 2) 3) 0.0025599523967240303
Probabilistic Measure :===> 1) returned 2) but 3) 0.008129145061787033
Probabilistic Measure :===> 1) and 2) primarily 3) 0.013818881947958345

Probabilistic Measure :===> 1) 4-hydroxydebrisoquin 2) microsomal 3) 0.035968887542210365
Probabilistic Measure :===> 1) cytochrome 2) in 3) 0.012147676062186996
Probabilistic Measure :===> 1) and 2) CYP2D6; 3) 0.03452005227123776

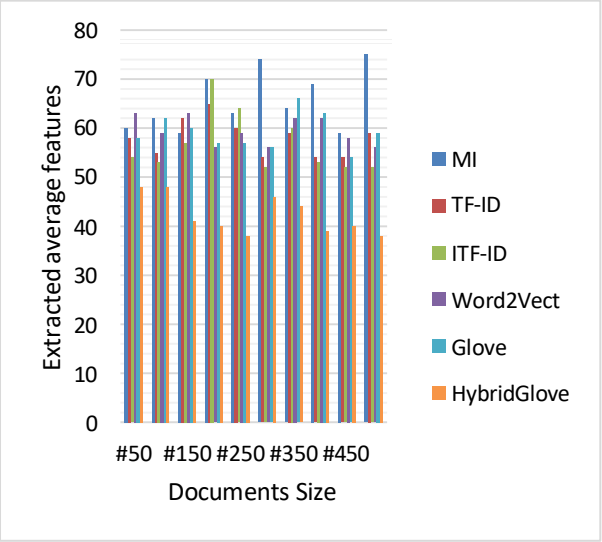

Figure 2: Comparison of proposed Glove features to the conventional feature extraction models on ChEBI **biomedical document sets.**

Figure 2, illustrates the performance of feature extraction using the proposed Glove approach on ChEBI document sets. From the figure 2, it is noted that the proposed feature extraction has high filtering rate as compared to the existing approaches.

Doc_ Size	MI	TF-ID	ITF-ID	Word2 Vect	Glove	Hybrid Glove
#50	68	54	61	61	61	48
#100	59	62	60	63	59	49
#150	68	53	61	61	62	44
#200	65	53	70	65	66	42
#250	60	62	66	58	64	48
#300	61	57	61	55	60	45
#350	66	63	62	63	57	45
#400	70	60	50	59	67	36
#450	63	64	69	61	66	41
#500	72	55	60	62	57	39

Table 2: Candidate features extraction using the proposed Glove model on PHAEDRA dataset

Table2, illustrates the performance of feature extraction using the proposed glove approach on PHAEDRA datasets. From the table2, it is clearly shown that the present feature extraction procedure has high filtering rate as compared to the existing approaches.

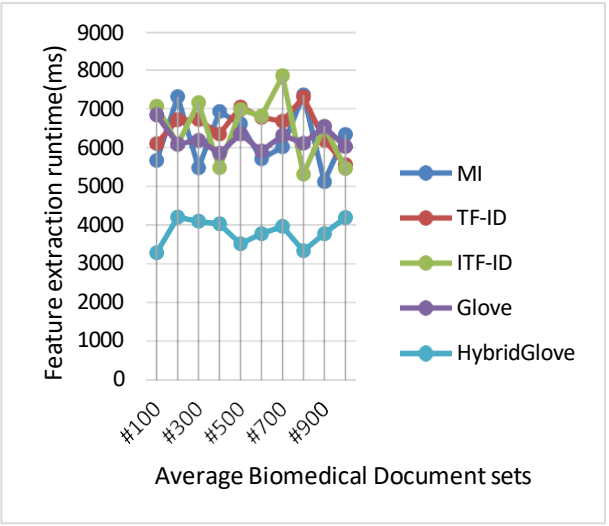

Figure 3: Performance analysis of average computational runtime(ms) with different traditional feature extraction models

Figure 3 describes the performance of the various feature extraction model's average computer runtime(s) for the proposed approach on all datasets. Figure 3 shows clearly that the current function extraction process has low calculation runtime in comparison with existing approaches.

Figure 4: Proposed real-time application of gene-disease mapping and its probabilistic scores

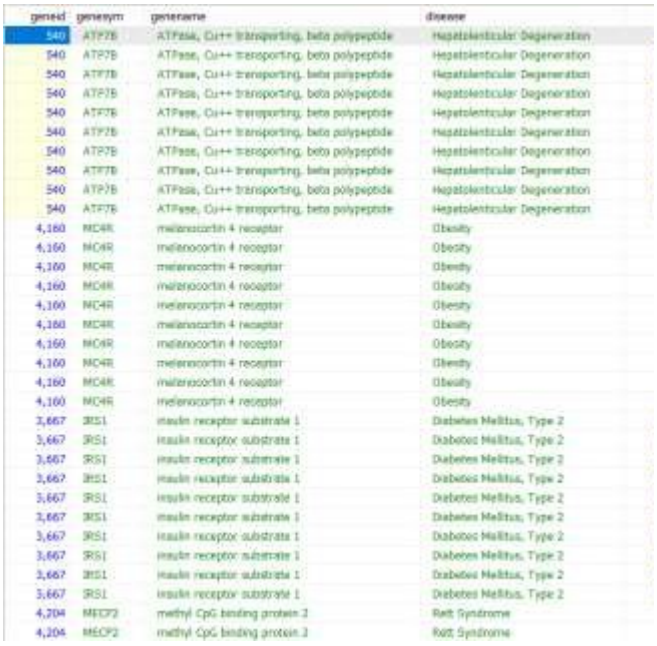

Figure 5: Proposed real-time application of all gene-disease mapping and its probabilistic scores

```
HG2868-HT3012_s_at < 171.5 : 1 (4/0)
HG2868-HT3012_s_at >= 171.5
|  U41315_rna1_s_at < 130.5
|  |  X90828_at < -155.5 : 0 (6/0)
|  |  X90828_at >= -155.5
|  |  |  J05257_at < 101
|  |  |  |  U09367_at < 14 : 1 (1/0)
|  |  |  |  U09367_at >= 14 : 0 (8/0)
|  |  |  J05257_at >= 101
|  |  |  |  S69189_at < 95
|  |  |  |  |  M60891_s_at < -82
|  |  |  |  |  |  U19796_at < 261.5
|  |  |  |  |  |  |  X97324_at < 90 : 1 (1/0)
|  |  |  |  |  |  |  X97324_at >= 90 : 0 (3/0)
|  |  |  |  |  |  U19796_at >= 261.5 : 1 (6/0)
|  |  |  |  |  M60891_s_at >= -82 : 0 (4/0)
|  |  |  |  S69189_at >= 95 : 1 (7/0)
|  U41315_rna1_s_at >= 130.5
|  |  S85655_at < 528 : 0 (16/0)
|  |  S85655_at >= 528
|  |  |  U28251_cds2_at < -120 : 0 (2/0)
|  |  |  U28251_cds2_at >= -120 : 1 (2/0)
```

Leukemia Data

```
X17042_at < 2892
|  U09851_s_at < 16 : AML (2/0)
|  U09851_s_at >= 16 : ALL (40/0)
X17042_at >= 2892
|  M10058_at < -691.5 : AML (14/0)
|  M10058_at >= -691.5
|  |  D89501_at < -6.5 : ALL (4/0)
|  |  D89501_at >= -6.5
|  |  |  Z26256_at < 209.5
|  |  |  |  S77356_at < 531 : ALL (1/0)
|  |  |  |  S77356_at >= 531 : AML (9/0)
|  |  |  Z26256_at >= 209.5 : ALL (2/0)
```

Table. 3 Datasets and Its Characteristics

Features set	PCA+RF	IG+SVM	PCA+NN	T-test++NaiveBayes	ProposedModel
Disease Genes50	0.93	0.92	0.87	0.93	0.98
Disease Genes100	0.94	0.93	0.86	0.93	0.98
Disease Genes150	0.92	0.94	0.86	0.93	0.97
Disease Genes200	0.93	0.93	0.85	0.95	0.98
Disease Genes250	0.94	0.94	0.86	0.93	0.97
Disease Genes300	0.93	0.94	0.87	0.93	0.98
Disease Genes350	0.94	0.92	0.87	0.94	0.98
Disease Genes400	0.93	0.93	0.88	0.93	0.98
Disease Genes450	0.95	0.93	0.88	0.95	0.98
Disease Genes500	0.92	0.94	0.87	0.92	0.98
Disease Genes550	0.94	0.95	0.86	0.95	0.98
Disease Genes600	0.94	0.93	0.87	0.94	0.98
Disease Genes650	0.93	0.94	0.86	0.93	0.97
Disease Genes700	0.93	0.93	0.87	0.94	0.97
Disease Genes750	0.93	0.93	0.87	0.95	0.98
Disease Genes800	0.94	0.92	0.87	0.94	0.97
Disease Genes850	0.94	0.92	0.86	0.95	0.98
Disease	0.93	0.94	0.87	0.93	0.97

Genes900					
Disease Genes950	0.94	0.93	0.86	0.93	0.98
Disease Genes1000	0.92	0.93	0.88	0.92	0.97

Table 4: Comparative analysis of proposed caner classification approach to the traditional approaches by using accuracy on various realtime cancer gene-disease patterns evaluation.

Table 4, describes the performance of the proposed model on cancer gene-disease realtime datasets. Here, all the cancer datasets are evaluated using the proposed model to find the average true positive rate and precision rate on the high dimensional datasets. From the table, it is visualized that the present approach has better true positive rate and precision over the existing models.

Features set	PCA+RF	IG+SVM	PCA+NN	T-test++NaiveBayes	ProposedModel
Disease Genes50	0.93	0.94	0.86	0.95	0.98
Disease Genes100	0.93	0.93	0.85	0.93	0.98
Disease Genes150	0.94	0.93	0.86	0.93	0.98
Disease Genes200	0.94	0.92	0.88	0.94	0.97
Disease Genes250	0.95	0.94	0.86	0.94	0.97
Disease Genes300	0.94	0.94	0.87	0.95	0.97
Disease Genes350	0.94	0.92	0.87	0.92	0.98
Disease Genes400	0.94	0.92	0.87	0.93	0.98
Disease Genes450	0.94	0.92	0.86	0.93	0.98
Disease Genes500	0.95	0.95	0.88	0.94	0.98
Disease Genes550	0.93	0.92	0.88	0.94	0.97
Disease Genes600	0.93	0.95	0.88	0.93	0.98
Disease Genes650	0.94	0.93	0.87	0.94	0.98
Disease Genes700	0.94	0.94	0.86	0.92	0.98
Disease Genes750	0.93	0.93	0.86	0.93	0.98
Disease Genes800	0.95	0.95	0.86	0.92	0.98
Disease Genes850	0.93	0.93	0.86	0.93	0.98
Disease Genes900	0.94	0.95	0.87	0.94	0.98
Disease Genes950	0.92	0.93	0.87	0.95	0.99
Disease Genes1000	0.92	0.94	0.87	0.93	0.98

Table 5: Performance comparison of proposed classification accuracy to the conventional approaches by using accuracy on various average accuracy of DLBCL, Prostate, Lymphoma, BreastCancer gene-disease patterns.

Table 5, describes the performance of the proposed model on all gene-disease real-time datasets. Here, all the datasets are evaluated using the proposed model to find the average true positive rate and precision rate on the high dimensional datasets. From the table, it is visualized that the present approach has better true positive rate and precision over the existing models.

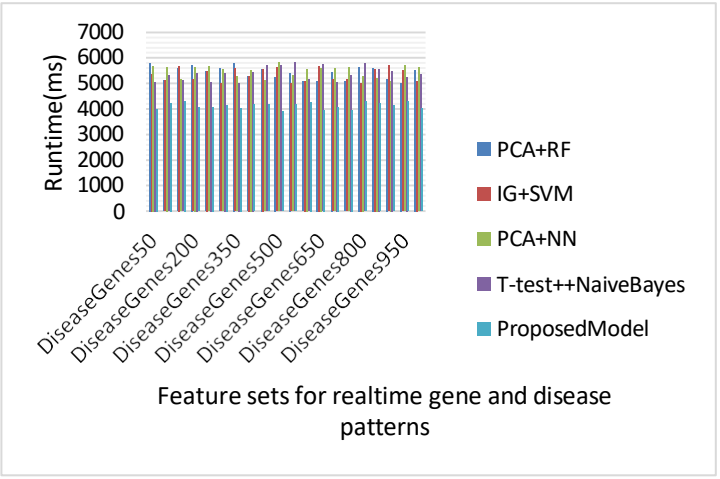

Table 6: Comparative runtime of present approach to the traditional approaches by using accuracy on various real-time average accuracy of DLBCL, Prostate, Lymphoma,

BreastCancer gene-disease patterns.

ICD-Gene related somatic cancer relations:

A31	9	A319	Mycobacte	Mycobacte Infection due to other mycobacteria
A32	0	A320	Cutaneous	Cutaneous Listeriosis
A321	1	A3211	Listerial m	Listerial m Listerial meningitis and meningoencephali
A321	2	A3212	Listerial m	Listerial m Listerial meningitis and meningoencephali
A327		A327	Listerial se	Listerial se Listerial sepsis
A328	1	A3281	Oculoglan	Oculoglan Other forms of listeriosis
A328	2	A3282	Listerial er	Listerial er Other forms of listeriosis
A328	9	A3289	Other form	Other form Other forms of listeriosis
A329		A329	Listeriosis,	Listeriosis, Listeriosis, unspecified
A33		A33	Tetanus ne	Tetanus ne Tetanus neonatorum
A34		A34	Obstetrica	Obstetrica Obstetrical tetanus
A35		A35	Other teta	Other teta Other tetanus
A36	0	A360	Pharyngea	Pharyngea Diphtheria
A36	1	A361	Nasopharyı	Nasopharyı Diphtheria
A36	2	A362	Laryngeal (	Laryngeal (Diphtheria
A36	3	A363	Cutaneous	Cutaneous Diphtheria
A368	1	A3681	Diphtheriti	Diphtheriti Other diphtheria
A368	2	A3682	Diphtheriti	Diphtheriti Other diphtheria
A368	3	A3683	Diphtheriti	Diphtheriti Other diphtheria
A368	4	A3684	Diphtheriti	Diphtheriti Other diphtheria
A368	5	A3685	Diphtheriti	Diphtheriti Other diphtheria
A368	6	A3686	Diphtheriti	Diphtheriti Other diphtheria
A368	9	A3689	Other diph	Other diph Other diphtheria
A369		A369	Diphtheria	Diphtheria Diphtheria, unspecified
A370	0	A3700	Whooping	Whooping Whooping cough due to Bordetella pertus

The above table describes the somatic germline related ICD codes and its description for the classification problem. Here, these ICD codes are used to find the relationships between the genes and the disease. The following table describes the somatic germline mutations for cancer prediction.

```
@relation SomaticVsGermline

@attribute inExAct {true, false}
@attribute dbSNP {true, false}
@attribute CNT numeric
@attribute fre numeric
@attribute VAF numeric
@attribute mutAss
{'neutral','low','medium','high','stopgain','stoploss'}
@attribute pattern {'CG', 'CA', 'CT', 'TA', 'TC', 'TG'}
@attribute SeqContent
{'ATT','CTT','GTT','TAT','AAA','CAA','AAC','CAC','GAA','
','GAC','GAG','TGA','TGC','TCA','AAT','TCC','TGG','CAT',
T','TGT','TTA','TTC','TCT','TTG','TTT','AGA','CGA','AGC'
CA','CCA','GGA','ACC','AGG','CCC','CGG','GGC','GCA','ACG
GCC','GGG','GCG','AGT','ATA','ATC','CGT','CTA','ACT','CT
'CCT','GGT','GTA','TAA','CTG','GTC','TAC','GCT','GTG','T
@attribute isFlanking numeric
@attribute polyphen {'benign', 'probably', 'possibly'}
@attribute isSomatic {true, false}

@data
true,true,0,0.01,40.06,neutral,TA,GAT,0,benign,false,
false,false,1,0.01,36.36,low,TC,TAC,0,?,true,
false,false,1,0.01,14.08,?,CA,AGG,1,?,true,
true,false,1,0.01,18.41,?,CA,AGG,1,?,true,
true,true,1,0.38,49.789,neutral,TC,TTA,0,?,false,
true,true,0,0.01,47.92,low,TC,CAA,?,possibly,false,
false,false,1,0.01,43.14,neutral,CA,CGG,0,probably,true,
true,true,0,0.01,50.58,neutral,TC,GTA,0,?,false,
true,false,0,0.01,41.11,low,TC,CTA,?,?,false,
true,true,0,0.01,42.86,neutral,TC,ATG,0,?,false,
true,true,0,0.01,57.66,medium,TG,CAA,?,?,false,
false,false,1,0.01,31.2,high,CA,GGA,0,?,true,
false,false,1,0.01,34.15,medium,CA,AGT,?,?,true,
false,false,1,0.01,52.54,low,TC,GAT,?,possibly,true,
false,false,1,0.01,48.82,neutral,CA,CGC,?,benign,true,
false,false,1,0.01,33.85,stopgain,CA,TGA,?,?,true,
true,false,0,0.01,42.86,low,CG,CGT,0,?,false,
false,false,1,0.01,16.12,medium,CA,GGA,?,?,true,
true,true,0,0.01,47.3,low,TC,ATA,0,?,false,
false,false,0,0.01,50,neutral,CA,AGG,0,benign,true,
true,true,0,0.34,32.504,neutral,TC,ATG,0.029,?,false,
false,false,1,0.01,28,low,TC,AAT,?,?,true,
```

The following patterns describes the somatic mutations for caner disease prediction using the ICD codes and the gene patterns. Here, each pattern defines the somatic feature and its relationship with the cancer disease class label for decision making.

```
=== Stratified cross-validation ===

Correctly Classified Instances       18137               99.9724 %
Incorrectly Classified Instances         5                0.0276 %
Kappa statistic                          0.9994
Mean absolute error                      0.0003
Root mean squared error                  0.0122
Relative absolute error                  0.0637 %
Root relative squared error              2.4498 %
Total Number of Instances            18142

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      0.999     1.000     1.000     true
                 1.000    0.000    1.000      1.000    1.000      0.999     1.000     1.000     false
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      0.999     1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 9070    1 |    a = true
    4 9067 |    b = false
1.11 seconds

=== Error on test data ===

Correctly Classified Instances       17830               98.2802 %
Incorrectly Classified Instances       312                1.7198 %
Kappa statistic                          0.9656
Mean absolute error                      0.0182
Root mean squared error                  0.1299
Relative absolute error                  3.635  %
Root relative squared error             25.9708 %
Total Number of Instances            18142

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.018    0.982      0.983    0.983      0.966     0.986     0.979     true
                 0.982    0.017    0.983      0.982    0.983      0.966     0.986     0.978     false
Weighted Avg.    0.983    0.017    0.983      0.983    0.983      0.966     0.986     0.979

=== Confusion Matrix ===

    a    b   <-- classified as
 8913  153 |    a = true
  159 8512 |    b = false
```

In the above results, experimental results are performed on the somatic germline cancer mutation dataset using ICD and gene sets. From the experimental results it is observed that the proposed somatic germline mutation

cancer prediction has better efficiency nearly 99% accuracy than the conventional cancer detection models with ICD and gene feature selection.

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 to problem of gene based biomedical document ranking. a hybrid word embedding method and similarity metric is 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 to ICD mapping process. In this work, a hybrid classification model is implemented on micro-array cancer datasets to find the essential patterns for the biomedical document mapping process. Experimental results are simulated on different biomedical document sets on the microarray datasets. Proposed results proved that the gene-chemical disease clustered based classification framework has better optimization than the conventional biomedical document embedding methods, similarity metrics and classification models.

REFERENCES

[1]B. Hosseini and K. Kiani, "A big data driven distributed density based hesitant fuzzy clustering using Apache spark with application to gene expression microarray," Engineering Applications of Artificial Intelligence, vol. 79, pp. 100–113, Mar. 2019, doi: 10.1016/j.engappai.2019.01.006.

[2]M. Daoud and M. Mayo, "A survey of neural network-based cancer prediction models from microarray data," Artificial Intelligence in Medicine, vol. 97, pp. 204–214, Jun. 2019, doi: 10.1016/j.artmed.2019.01.006.

[3]A. K. Shukla, P. Singh, and M. Vardhan, "A two-stage gene selection method for biomarker discovery from microarray data for cancer classification," Chemometrics and Intelligent Laboratory Systems, vol. 183, pp. 47–58, Dec. 2018, doi: 10.1016/j.chemolab.2018.10.009.

[4]J. Lee, I. Y. Choi, and C.-H. Jun, "An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data," Expert Systems with Applications, vol. 166, p. 113971, Mar. 2021, doi: 10.1016/j.eswa.2020.113971.

[5]X. Liu, S.-C. Lee, G. Casella, and G. F. Peter, "Assessing agreement of clustering methods with gene expression microarray data," Computational Statistics & Data Analysis, vol. 52, no. 12, pp. 5356–5366, Aug. 2008, doi: 10.1016/j.csda.2008.06.004.

[6]O. Stoss and T. Henkel, "Biomedical marker molecules for cancer – current status and perspectives," Drug Discovery Today: TARGETS, vol. 3, no. 6, pp. 228–237, Dec. 2004, doi: 10.1016/S1741-8372(04)02459-4.

[7]F. Zhu et al., "Biomedical text mining and its applications in cancer research," Journal of Biomedical Informatics, vol. 46, no. 2, pp. 200–211, Apr. 2013, doi: 10.1016/j.jbi.2012.10.007.

[8]Y. He, W. Pan, and J. Lin, "Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data," Computational Statistics & Data Analysis, vol. 51, no. 2, pp. 641–658, Nov. 2006, doi: 10.1016/j.csda.2006.02.012.

[9]R. Wang, L. Scharenbroich, C. Hart, B. Wold, and E. Mjolsness, "Clustering analysis of microarray gene expression data by splitting algorithm," Journal of Parallel and Distributed Computing, vol. 63, no. 7, pp. 692–706, Jul. 2003, doi: 10.1016/S0743-7315(03)00085-6.

[10]S. Bruhn et al., "Combining gene expression microarray- and cluster analysis with sequence-based predictions to identify regulators of IL-13 in allergy," Cytokine, vol. 60, no. 3, pp. 736–740, Dec. 2012, doi: 10.1016/j.cyto.2012.08.009.

[11]B. S. Biswal, A. Mohapatra, and S. Vipsita, "Ensemble Neighborhood Search (ENS) for biclustering of gene expression microarray data and single cell RNA sequencing data," Journal of King Saud University - Computer and Information Sciences, Dec. 2019, doi: 10.1016/j.jksuci.2019.11.011.

[12]S. R. Kumaran, M. S. Othman, L. M. Yusuf, and A. Yunianta, "Estimation of Missing Values Using Hybrid Fuzzy Clustering Mean and Majority Vote for Microarray Data," Procedia Computer Science, vol. 163, pp. 145–153, Jan. 2019, doi: 10.1016/j.procs.2019.12.096.

[13]C. L. Clayman, S. M. Srinivasan, and R. S. Sangwan, "K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes," Procedia Computer Science, vol. 168, pp. 97–104, Jan. 2020, doi: 10.1016/j.procs.2020.02.265.

[14]M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions," International Journal of Cognitive Computing in Engineering, vol. 1, pp. 78–97, Jun. 2020, doi: 10.1016/j.ijcce.2020.11.001.

[15]J. Cui, N. Zhang, Y. Liu, L. Zhang, C. Gao, and S. Liu, "Microarray gene expression profiling provides insights into functions of TIPE2 in HBV-related apoptosis," Molecular Immunology, vol. 131, pp. 137–143, Mar. 2021, doi: 10.1016/j.molimm.2020.12.031.

[16]D. R. Rhodes et al., "ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform," Neoplasia, vol. 6, no. 1, pp. 1–6, Jan. 2004, doi: 10.1016/S1476-5586(04)80047-2.

[17]S. Sarbazi-Azad, M. S. Abadeh, and M. E. Mowlaei, "Using Data Complexity Measures and an Evolutionary Cultural Algorithm for Gene Selection in Microarray Data," Soft Computing Letters, p. 100007, Oct. 2020, doi: 10.1016/j.socl.2020.100007.

[18]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.
[19]D. Zhao, J. Wang, Y. Chu, Y. Zhang, Z. Yang, and H. Lin, "Improving Biomedical Word Representation with Locally Linear Embedding," Neurocomputing, Mar. 2021, doi: 10.1016/j.neucom.2021.02.071.
[20]M. Moradi, M. Dashti, and M. Samwald, "Summarization of biomedical articles using domain-specific word embeddings and graph ranking," Journal of Biomedical Informatics, vol. 107, p. 103452, Jul. 2020, doi: 10.1016/j.jbi.2020.103452.
[21]D. Dimitriadis and G. Tsoumakas, "Word embeddings and external resources for answer processing in biomedical factoid question answering," Journal of Biomedical Informatics, vol. 92, p. 103118, Apr. 2019, doi: 10.1016/j.jbi.2019.103118.

