SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V10I12P107

Volume 10 Issue 12, 73-82, December 2023
© 2023 Seventh Sense Research Group®

Original Article

A Hybrid CNN-Multi-Class SVM Framework for Biomedical Document Gene-Disease Datasets Classification

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

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

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

Received: 09 October 2023 Revised: 21 November 2023 Accepted: 11 December 2023 Published: 23 December 2023

***Abstract* -** *Healthcare investigators and clinicians need biomedical document classification to organize and handle the large volume of biomedical literature. Conventional classification methods use manually designed features, which may be time-consuming and may not represent biomedical text complexity. Biomedical data's high dimensionality and sparsity may also challenge current approaches. For big datasets, CNNs are computationally costly. Increasing feature extraction efficiency reduces training and inference durations. The proposed method intends to improve the accuracy of document classification in the biomedical sector considerably. It functions in two stages: feature extraction and classification. The proposed method employs a hybrid approach to biomedical document classification, focusing on the intricate interactions between genes, diseases, and chemical treatments via the use of a CNN Multi-class Support Vector Machine (M-SVM) model. CNN is utilized to extract features, while M-SVM is employed as a classifier. This work discusses Improved CNNs, which may extract more discriminative and informative features from input data, resulting in a more accurate representation of underlying patterns and connections. Error-Correcting Output Coding (ECOC) based on M-SVM is used to manage noisy data by merging the outputs of many binary classifiers, enabling it to recover from faults in individual classifiers and thereby lowering the risk of overfitting. The study's results show that the proposed model is successful, with an accuracy of 99.28% and an F1-score of 99.84% across biomedical document datasets.*

***Keywords* -** *Biomedical documents, Gene data, Feature extraction, Classification, CNN, M-SVM.*

1. Introduction

Many biomedical entity relations exist in the literature. Automatically and properly extracting these links and structuring information helps biological areas improve. Studying disease molecular pathways and finding solutions requires gene-disease and chemical connections. Due to the complicated interactions between genes, illnesses, and chemicals, extracting these links from the enormous biomedical literature is difficult. Manually maintained databases may be time-consuming and may not include the latest results. Biomedical data's high dimensionality and sparsity may also challenge current approaches.

The FSVM model has surpassed traditional SVM in classifying biomedical data, which often contains sparse, incomplete, and imbalanced features. To tackle these challenges, various models, such as the hybrid cluster-based Bayesian model, Adaboost, bagging, and fuzzy SVMs, have been introduced. These models have shown enhanced accuracy and true positive rates in simulations with different training datasets. As biomedical data grows in complexity, these models are becoming increasingly vital for extracting insights and data. Continuous exploration and development of new methods are essential for biomedical data analysis. The remaining data serves as a background for these methods, which have shown superior accuracy and efficiency compared to conventional methods [1, 2].

In medical diagnosis and treatment, ranking and clustering techniques are crucial. A novel ranking algorithm combining network-based and sequence-based features to identify cancer drug targets, showing higher accuracy and sensitivity. A clustering method combining K-means and principal component analysis to identify cancer subtypes based on gene expression, yielding clinically relevant results [3]. Biomedical data analysis is an emerging field where machine learning and data mining enhance medical diagnosis and treatment. Ranking and clustering techniques are vital for identifying drug targets and grouping data points. One-class classification is promising for analyzing medical data with limited training and testing data. In biomedical research, predicting diseases using ranking and clustering techniques is growing, but a class imbalance in datasets poses challenges. Researchers use over-sampling and under-sampling to balance datasets. The linear combinational model helps identify suitable hyperplanes for classification with tailored margin distance measures.

SVM techniques like maximum and soft margin classifiers are essential in classification. The goal is to develop advanced tools for analyzing large datasets and providing accurate medical evidence for diagnosing and treating heart disease. Clustering methods group patients with similar profiles, identifying high-risk subgroups for effective disease management. Ranking techniques prioritize key terms for disease identification. Mortality risk models benefit from these techniques, identifying risk factors and predicting outcomes. Machine learning and big data analytics enhance these models, but further validation in real-world scenarios is needed.

In biomedical documents, clustering and ranking methods analyze large datasets, identifying patterns and extracting critical information. A graph representation with nodes and edges helps in identifying common patterns by extracting frequent subgraphs [4]. Determining the relevance of documents to specific topics is vital in biomedical document analysis. Various algorithms, including the well-known PageRank used in web search engines, have been adapted for this purpose.

However, the unique challenges of biomedical documents, such as intricate interrelations and specialized language, complicate the application of PageRank in this context. Clustering and ranking are indispensable in analyzing biomedical records, helping researchers identify patterns and extract critical information. The success of these methods hinges on data quality and clarity, necessitating the development of new algorithms to handle the complexity of biomedical data [5].

In managing sub-clusters, they choose leaders for their superior ranking and clustering abilities. This selection relies on assessing the similarity between nodes and their subgraphs. A dynamic ranking system within this framework evaluates each node’s relevance to its subcluster and the broader network. The employed graph-based clustering technique adeptly identifies subgraphs with significant similarity measures.

Additionally, the framework incorporates security protocols to safeguard the confidentiality and integrity of data in Peer-to-Peer (P2P) networks, offering a secure and effective method for clustering biomedical documents [6]. Hierarchical clustering is a robust method for analyzing gene expression data. It uses a probabilistic approach to determine the number of gene clusters and their interrelations. This method allows for accurate clustering based on expression patterns, revealing biological processes and pathways. Their relevance can rank clusters to particular diseases or conditions, facilitating focused analysis and hypothesis formulation.

Clustering and ranking are critical in exploring extensive biomedical datasets, offering insights into complex biological systems. Hierarchical clustering, therefore, is highly recommended for researchers working with gene expression data, as it can enhance the understanding of biological mechanisms [7].

The proposed multi-class classification system for biomedical document feature extraction and classification combines the capabilities of Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) to tackle these difficulties. While Convolutional Neural Networks (CNNs) are great at extracting local patterns and features from sequential data (like text), Support Vector Machines (SVMs) are great at handling high-dimensional data. The main goals of this study are to:

1. Construct a multi-class classification system that uses Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) to classify and extract features from biomedical articles efficiently.
2. To improve the feature extraction procedure, examine other text representation approaches, such as word embedding methods.
3. Test the proposed framework’s scalability to large-scale datasets and compare its performance to that of standard classification approaches on benchmark biomedical document datasets.

2. Background and Related Work

In biomedical research, accurately predicting gene and protein behaviours is crucial. Model-based methods have proven effective for precise data predictions. These advancements have transformed selecting cluster numbers in data from random to calculated decisions. Post-modeling, it’s possible to assess the likelihood of two genes being in the same cluster using a mathematical threshold linked to an infinite number of components in a standard finite mixture connected to the Dirichlet process prior.

The infinite Gaussian mixture model negates the need for arbitrary decisions about the number of clusters. Hidden Markov chain Monte Carlo methods aid in inferring these Bayesian models. While the infinite mixture model involves several parameters, the parameters for a fixed number of mixture components need to be explicitly shown. Different groups have independently established values based on Dirichlet Process Mixtures (DPMs) [8-10]. Protein sequence clustering employs a class-based approach. In unstructured Peer-to-Peer (P2P) overlay structures, efficient searching and indexing are crucial. Semantic-based indexing, which has recently gained prominence, relies on semantic document preprocessing at the peer node using the document vector representation model. The topology-based method assesses node similarity [11].

In patent analysis within biomedical research, algorithms play a vital role. Ranking and clustering are key: ranking identifies relevant patents based on set criteria while clustering groups similar patents for more accessible analysis. The success of these processes depends on the quality of the algorithms and the data set size. Continuous refinement and rigorous testing of these algorithms are essential for the precise and efficient analysis of biomedical documents and patents.

A comprehensive review and statistical analysis showed that the Artificial Neural Network (ANN) model outperforms others in accuracy and efficiency, making it helpful in clustering patients in biomedical documents. For patent clustering, algorithms like PageRank, TF-IDF, and BM25 are crucial for identifying relevant patents based on term frequency, document length, and link structure. Clustering techniques such as k-means, hierarchical clustering, and spectral clustering group similar patents by content and relevance. Clustering is a foundational technique in biomedical data analysis, grouping similar data points to reveal patterns and relationships for further analysis and decision-making.

Researchers use various clustering algorithms, including k-means, hierarchical, and fuzzy clustering, depending on the nature and research objectives. Employing classification and clustering techniques is essential in biomedical data analysis, as they help identify insights and patterns that can improve patient outcomes and healthcare delivery [12-15]. The MC4.5 algorithm marks a significant progression in biomedical data classification, introducing an advanced gain ratio formula that boosts classification accuracy and efficiency. Its innovative clustering technique, grouping similar instances, results in a refined decision tree model.

The algorithm excels in optimizing cluster numbers by minimizing the within-cluster sum of squares and maximizing the between-cluster sum of squares, leading to a more stable and accurate decision tree. This method benefits complex biomedical data analysis, where precision is paramount [16].

The algorithm's iterative approach in selecting the best training set further enhances its reliability and performance, making it superior to traditional methods for ranking and clustering biomedical data. The MC4.5 algorithm, with its clustering feature, stands as a robust tool for biomedical researchers, offering precise data classification and clustering capabilities essential for advanced analysis. Biomedical research, a field reliant on extensive data analysis, often employs ranking and clustering techniques to discern patterns and connections in data. Traditional methods, however, can need help with missing values or large, distinct attributes, affecting analysis accuracy.

The MC4.5 algorithm addresses these issues, adeptly managing missing values and diverse attributes. Its enhancements have shown it to be more effective than previous techniques, affirming its importance in biomedical research. Looking ahead, there is significant potential for further refinement. Exploring the C4.5 algorithm for handling large, different-valued attributes and a hybrid approach for multidimensional data with vast intervals could lead to more precise, comprehensive biomedical data analysis. Such advancements promise to aid researchers in making more informed decisions, paving the way for the best discoveries and advancements in the field.

The learning algorithm in focus optimizes feature selection, determining the best subset of features. It then validates the model with a new test data wrapper approach mechanism. While wrapper methods like sequential forward selection and sequential backward elimination are user-friendly, they come with risks: the former may lead to overfitting, and the latter can be computationally intensive. Filter and wrapper methods differ from embedded methods. Filter techniques use the learning algorithm only for feature extraction, not affecting the classification process.

In contrast, wrapper methods leverage the learning algorithm for both feature extraction and classification. Embedded feature selection, integrating classification with feature selection, tends to be more efficient than wrapper methods. Decision trees and weighted Naive Bayes models are examples of embedded feature selection algorithms [17-20]. Traditional filtering schemes need help to measure relationships between genes effectively.

In biomedical diagnostics, gene expression data, often containing numerous genes, are pivotal. Recent studies suggest that using fewer, more relevant genes can enhance cancer diagnosis accuracy. Thus, gene selection in processing microarray data is challenging but crucial for data reduction.

In the analysis of large databases for breast cancer detection, mining algorithms aim for high accuracy. Classification, a common algorithm, focuses on selecting critical cancer patches to minimize noise. Researchers explore various methods, combining classifiers for optimal feature selection, employing filters or wrappers, or both. Bayesian estimation models are instrumental in gene activity prediction and drug discovery. They apply Bayesian probability theory, balancing new data with prior knowledge of Gene inhibitors' characteristics and activities. The model predicts the likelihood of a molecule's activity against a Gene using molecular features, adjusting probability estimates during training. Once trained, it can predict novel chemical activities, improving accuracy and resilience by incorporating uncertainty [21].

Bayesian modelling pairs with machine learning to predict Gene inhibitor activities, offering probabilistic activity estimates rather than binary categorizations. They prioritize compounds with high posterior probabilities for further testing. Bayesian estimation, based on Bayes' theorem, modifies initial assumptions with new data.

The probability function and prior distribution shape the posterior distribution, which is used for event forecasting and uncertainty assessment in parameter estimations. Bayesian modelling's probabilistic estimates are advantageous in selecting effective inhibitors for further experimentation.

3. Proposed Modelling

The proposed framework introduces an innovative approach to classify biomedical documents by employing a combination of multi-class Support Vector Machine (SVM) and deep learning-based Convolutional Neural Networks (CNNs), explicitly targeting the complex relationships found in gene-disease and chemical contexts. At its core, the framework operates in two main stages: feature extraction and classification. In the feature extraction stage, the method departs from traditional techniques like bag-of-words or simple word frequency analysis.

Instead, it harnesses the power of pre-trained word embeddings and a sentence similarity matrix to capture the nuanced relationships between genes, diseases, and chemicals. These embeddings effectively map words into a vector space, enabling the model to understand and use the semantic and contextual relationships inherent in the biomedical text. The classification stage then employs a multi-class SVM coupled with a deep learning model.

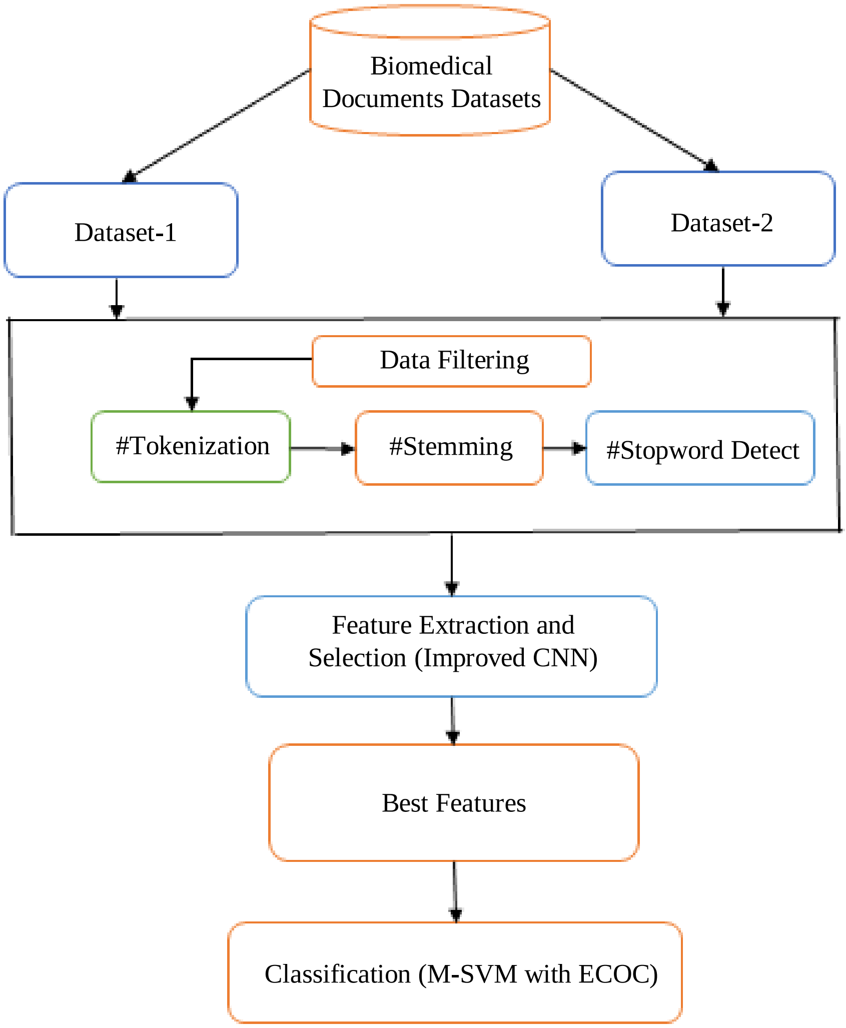
Biomedical Documents Datasets
Dataset-1
Dataset-2
Data Filtering
#Tokenization
#Stemming
#Stopword Detect
Feature Extraction and Selection (Improved CNN)
Best Features
Classification (M-SVM with ECOC)

Fig. 1 Proposed framework

3.1. Biomedical Document Datasets

These are crucial for training and evaluating machine learning algorithms in various biomedical research domains, including disease diagnosis, drug discovery, and biomedical literature analysis. These datasets provide researchers with a collection of annotated biomedical documents, where each record is labelled with one or more relevant categories or concepts. All these datasets are collected from PubMed, ClinicalTrials.gov, Medical Subject Headings (MeSH), and Kaggle.

3.2. Data Filtering

Data filtering of biomedical document collections entails extracting valuable papers from a mountain of scientific research. Many activities in biomedical research rely on this procedure, such as literature reviews, finding new data, and retrieving previously stored data. Investigators are able to filter through a massive volume of biomedical literature in search of relevant information when they use efficient data filtering strategies in Algorithm 1.

Contextual similarity in gene-disease and drug mapping is a computational strategy that uses contextual ties among genes, diseases, and medications to infer novel linkages among them. This technique is attracting interest in biomedical research because it provides a viable path for finding novel therapeutic targets, adapting current medications for new purposes, and explaining the molecular processes driving the complicated diseases discussed in Algorithm 2.

3.2.1. Algorithm 1: Data Filtering

Input: Biomedical document set

Output: A Collection of biomedical documents

1. Begin by creating a list of filtered documents that are empty.
2. For all of the documents that are loaded into the system:
   a. Identify and extract important document properties, including entities, connections, and keywords.
   b. Use the extracted features as a basis for filtering, such as removing documents that include specific keywords or those that have poor significance scores.
   c. Upon successful validation, include the document in the list of filtered records.
3. Get the documents that were filtered back.

3.2.2. Algorithm 2. Gene-Disease and Drug Mapping Contextual Similarity

Contextual similarity is used in the gene-disease and drug mapping contextual similarity technique to find and map genes to diseases and medicines to diseases.

Input: Gene-disease and drug-disease interaction networks, biomedical literature

Output: Gene-disease and drug-disease mapping scores

1. Create a function. ContextualSimilarity Calculator (gene1, gene2, disease):
   - Extract contextual information from the biomedical literature for gene1, gene2, and illness.
   - Applying word embedding algorithms presents the contextual information as vectors.
   - Determine the cosine similarity of gene1, gene2, and disease contextual vectors.
2. Identify related gene-disease pairings based on contextual similarities for each gene-disease pair.
   - Compute the average CSS for the gene-disease pairings that are comparable.
3. Identify related drug-disease couples based on contextual similarities for each drug-disease pair.
   - Compute the average CSS for the drug-disease pairings that are comparable.
4. Use the derived contextual similarity scores to map genes to illnesses and medicines to disorders.
5. Use suitable metrics to evaluate the gene-disease and drug-disease mapping results.

3.3. Feature Extraction and Selection

CNNs have emerged as valuable tools for feature extraction and selection in a variety of applications, including biomedical document classification. Their ability to collect local patterns and contextual information from text input enhances their capacity to extract relevant features for classification tasks. Improved CNN structures will become of growing significance in extracting significant features from biomedical documents as the field of biomedical informatics evolves, leading to improved classification accuracy, interpretation, and valuable insights into the field of biomedical data.

3.3.1. Improved Convolutional Neural Networks (CNNs)

Step 1: Long-range relationships in data are captured via initially dilated convolutions. This qualifies them for tasks like feature extraction from biomedical documents, which frequently comprise complex associations between words and ideas. The following is the mathematical equation for a dilated convolution:

$$y[i] - \sum_{k} w[k] * x[i + d * k] \tag{1}$$

Where $y[i]$ denotes the output value at position I, $w[k]$ is the weight value at position k, $x[i]$ is the input value at position I, and d is the dilation rate; The dilation rate determines how closely the input data are sampled. A dilation rate of one indicates that the input values are sampled consecutively. With a dilation rate of 2, every other input value is sampled, and so on.

Step 2: This residual link, also known as a skip connection, connects several levels of a CNN. They help in the reduction of vanishing gradients, which may occur in deep CNNs. The following is the mathematical equation for a residual connection:

$$y[i] = h(x[i] +)x[i] \tag{2}$$

Where $h(x[i])$ is the output of the layer that is being connected. The residual connection adds the layer's output to the input value. This ensures that the gradient may return to the network even if the layer itself does not learn anything.

Step 3: The attention methods enable CNNs to concentrate on the most critical sections of the input data, which is especially beneficial for tasks such as feature extraction from biomedical texts, where the essential information may be spread out over many regions of the document. The following is the mathematical equation for an attention mechanism:

$$\alpha_i = softmax\left(w^T * h(x_i)\right) \tag{3}$$

Where α_i is the attention weight for the ith input value, w is a weight vector and $h(x_i)$ is the output of the encoder for the ith input value; The attention weight is a scalar value indicating the significance of the ith input value. The attention weights are then utilized to construct a weighted sum of the input values, which serves as the attention mechanism's output.

Step 4: For feature extraction, Gated Recurrent Units (GRUs) are often employed in combination with CNNs. GRUs can detect long-term relationships in data and are less prone to gradient loss than typical RNNs. The mathematical equations for GRUs are as follows:

$$z_t = \sigma\left(W_z * x_t + U_z * h_{(t-1)}\right) \tag{4}$$

$$r_t = \sigma\left(W_r * x_t + U_r * h_{(t-1)}\right) \tag{5}$$

$$h_t = tanh\left(W_h * \left(r_t * h_{(t-1)}\right) + x_t\right) \tag{6}$$

Where x_t is the input, h_t is the hidden state, z_t is the update gate, r_t is the reset gate at time step t, respectively. W_z, U_z, W_r, and U_r are weight matrices. The update gate determines how much information is preserved from the previous hidden state. The reset gate determines how much data from the current input value is used.

The proposed approach involves a complex multi-step process to analyze biomedical documents, mainly focusing on the relationships between genes, diseases, and chemical drugs. The procedure begins with the construction of a Document Dependency Graph (DDG). In this graph, each vertex represents either a gene, a disease, or a chemical drug, and these vertices are interconnected through edges. The edges are not just simple connections but are weighted to reflect the strength or significance of the relationship between the entities they connect. It mainly focused this weighting on the rank between disease gene sets and chemical drug symbols, capturing how closely related a particular gene and disease are to a specific chemical drug. Once the DDG is constructed, the next crucial step involves computing the weighted rank between chemical drug symbols and disease gene patterns. This computation is fundamental in establishing the involved relationships within the biomedical data, providing a basis for further analysis.

3.4. Methods Used for Classification

3.4.1. Support Vector Machines (SVMs)

SVMs are a general-purpose machine learning technique that can deal with both linear and non-linear classification problems. They accomplish non-linearity by transforming the input data into a higher-dimensional space where linear separation is more successful. The linear kernel, polynomial kernel, and Radial Basis Function (RBF) kernel are examples of standard kernel functions.

SVMs in biomedical document classification seek an ideal hyperplane with the most significant margin of separation between documents belonging to distinct classes. The margin is defined as the distance between the hyperplane and the nearest support vectors from each category. SVMs successfully capture the basic framework of the data while minimizing the chance of overfitting by maximizing the margin.

3.4.2. Linear Support Vector Machines (LSVMs)

LSVMs are a subset of SVMs that are only used for linear classification problems. They do not need kernel functions since they presume that the data may be divided by a linear hyperplane. This constraint limits the flexibility of LSVMs compared to ordinary SVMs, but it also makes them more efficient and interpretable.

3.4.3. SVM for Multi-Class Classification

Because of their efficacy, robustness, and adaptability, SVMs have emerged as a viable tool for multi-class classification in biological document categorization. Their capacity to handle a high number of classes and their inclination to account for noisy input makes them well-suited for many biomedical classification tasks. This method solves a multi-class problem directly [22] by changing the binary class objective function and adding a constraint for each class. Multi-class classification is calculated simultaneously using the updated objective function:

$$\min_{x,a,\xi}\left[\frac{1}{2}\sum_{j=1}^{N}\|x\|^2 + b\sum_{j=1}^{l}\sum_{t\neq t_j}\xi_j^s\right] \tag{7}$$

Subject to the constraints,

$x_{p_j}.t_j + a_s \geq x_s, t_j + a_s + 2 - \xi_j^s$ for $\xi_j^s \geq 0$ for j=1,,,,,,l

Where t_j∈{1,..., N} denotes multiclass data vector labels and s∈{1,...,N}.

A user should examine the required accuracy, computing time, available resources, and the nature of the issue. For instance, because of the high memory requirements and very lengthy computing time, the multi-class objective function technique may not be appropriate for a problem with a large number of training samples and classes.

3.4.4. Proposed Error-Correcting Output Codes (ECOC)

The Error-Correcting Output Codes (ECOC) approach is a multi-class classification approach that is increasing in popularity in a variety of applications, notably biomedical document classification. It entails breaking down a multi-class issue into a series of binary classification jobs and using error-correcting methods to improve classification accuracy.

Consider a dataset D containing N biomedical documents, each represented by a feature vector $d \in D$, to illustrate the ECOC technique for biomedical feature classification mathematically. The purpose is to classify each document into one of C different groups.

A feature vector xd∈RF represents each document d∈D, where F is the number of features. The document d's matching class label is marked as y_d∈ {1, 2, ..., C}. Then ECOC transforms the multi-class classification issue into P binary classification jobs, where P is the number of error-correcting codes.

A binary classifier $h_k\colon R^F \rightarrow \{0,1\}$ is trained for each code k = 1, 2, ..., P to differentiates documents belonging to a given codebook C_k from those pertaining to all additional codebooks. This C_k is formed by splitting the set of classes 1, 2,..., C into P disjoint subsets. Each codebook C_k comprises a subset of classes, and each document is allocated to a certain codebook depending on the label of its class.

A feature extraction function $f_k\colon R^F \rightarrow R^{D_k}$ is used for each binary classifier h_k to extract relevant features from the input feature vector x_d. The extracted features are represented as $f_k(x_d) \in R^{D_k}$, where D_k is the dimension of the extracted feature space for the kth binary classifier. Each binary classifier h_k is trained to use the extracted features $f_k(x_d)$ and the binary labels assigned by the codebook. The training method entails optimizing the classifier's parameters in order to minimize classification error.

The ECOC technique predicts the class label of a new document d with feature vector x_d by merging the outputs of the binary classifiers. Various decoding algorithms, such as majority voting or Hamming distance-based approaches, may be used. The majority vote within the binary classifier outputs is used to predict the class label y_d. The document's class label is the class with the most projected '1's. Identifying the class with the shortest Hamming distance to the codeword vector $C(d)$ yields the class label y_d. The Hamming distance represents the number of points where the relevant elements vary between two codewords.

4. Results and Discussions

In this study, three publicly available multi-label biomedical datasets incorporating various types of biomedical literature and clinical notes are employed for carrying out tests. For datasets related to biomedical literature, a 10-fold cross-validation approach was adopted, which uses nine different folds iteratively for training, while the remaining single fold serves as the test set. Conversely, for the clinical notes dataset, all the proven practices in the literature are divided into training, validation, and test subsets.

This paper uses gene sets, protein sets, disease databases, and ICD codes to identify crucial key phrases for the document classification process. The experimental simulations for document classification are conducted within a Java-based computational environment. Several tests were carried out to assess the proposed architecture's performance in terms of accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC) using Equations (8) to (12).

$$Accuarcy\ (Acc) = \frac{TP+TN}{TP + FN + FP + TN} \quad (8)$$

$$Precision\ (Pre) = \frac{TP}{TP + FP} \quad (9)$$

$$Recall\ (Re) = \frac{TP}{TP + FN} \quad (10)$$

$$F1\ score = 2.\frac{Precision \cdot Recall}{Precision + Recall} \quad (11)$$

$$MCC = \frac{(TP*TN-FP*FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (12)$$

4.1. Analysis on Dataset 1: Online Mendelian Inheritance in Man (OMIM)

This is a comprehensive, authoritative database of human genes and genetic traits that is publicly accessible and updated on a regular basis. It is a valuable tool for academics, physicians, and everyone interested in human genetics and genetic diseases. OMIM has information on over 15,000 genes and over 7,000 genetic disorders. The performance of the proposed approach for biomedical document Online Mendelian Inheritance in Man dataset classification is compared with CNN-SVM and CNN-LSVM methods in terms of accuracy, precision, recall, F1 score, and MCC, which are listed in Table 1.

Table 1. Classification performance for biomedical documents OMIM dataset

Method	Accuracy	Precision	Recall	F1-Score	MCC
CNN-SVM	0.967	0.964	0.960	0.958	0.894
CNN-LSVM	0.971	0.965	0.962	0.961	0.901
Proposed	0.988	0.987	0.985	0.965	0.912

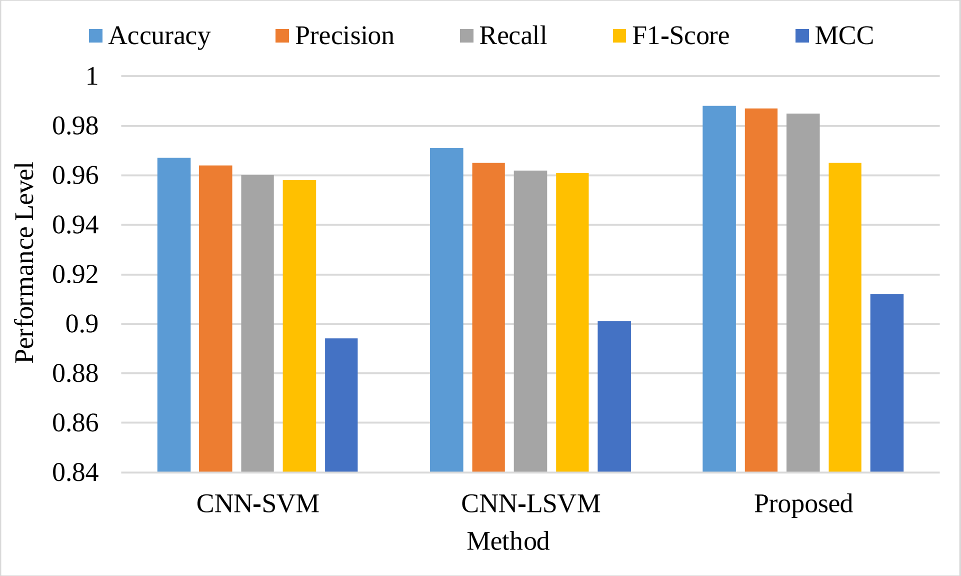

Fig. 2 Analysis comparing the proposed method with CNN-SVM and CNN-LSVM on OMIM dataset

Figure 2 shows the proposed method is compared with CNN-SVM and CNN-LSVM methods. It is evident that the proposed method achieves the highest accuracy of 98.8% and F1-score of 96.5%, while the lowest performance is obtained for the CNN-SVM method. Also, the suggested method prediction is good, with an MCC of 91.2%.

4.2. Analysis on Dataset 2: ClinVar (Clinically Varified Pathogenic Variants)

ClinVar is a publicly accessible, public repository of clinically relevant variant interpretations that is growing in popularity throughout the world. Its goal is to collect data on the link between human genetic variants and their impact on human health. ClinVar enables researchers, physicians, and patients to share information regarding genetic variants and their clinical importance. ClinVar has many different types of genomic variants, such as Single Nucleotide Variants (SNVs), Insertions and Deletions (INDELs), Copy Number Variations (CNVs), and Structural Variants (SVs).

The performance of the proposed approach for the biomedical document Clinically Varified Pathogenic Variants dataset classification is compared with CNN-SVM and CNN-LSVM methods in terms of accuracy, precision, recall, F1 score, and MCC, which are listed in Table 2.

Figure 3 shows the proposed method is compared with CNN-SVM and CNN-LSVM methods. It is evident that the proposed method achieves the highest accuracy of 98.9% and F1-score of 97.4%, while the lowest performance is obtained for the CNN-SVM method. Also, the suggested method prediction is good, with an MCC of 92.3%.

Table 2. Classification performance for biomedical documents ClinVar (Clinically Varified Pathogenic Variants) dataset

Method	Accuracy	Precison	Recall	F1-Score	MCC
CNN-SVM	0.973	0.975	0.982	0.987	0.875
CNN-LSVM	0.979	0.969	0.972	0.976	0.912
Proposed	0.989	0.989	0.987	0.974	0.923

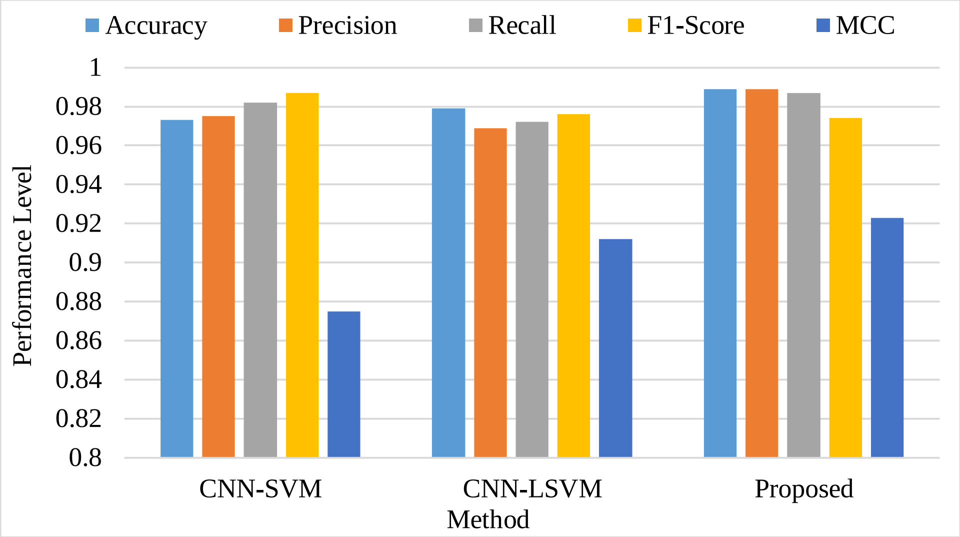

Fig. 3 Analysis comparing the proposed method with CNN-SVM and CNN-LSVM on ClinVar dataset

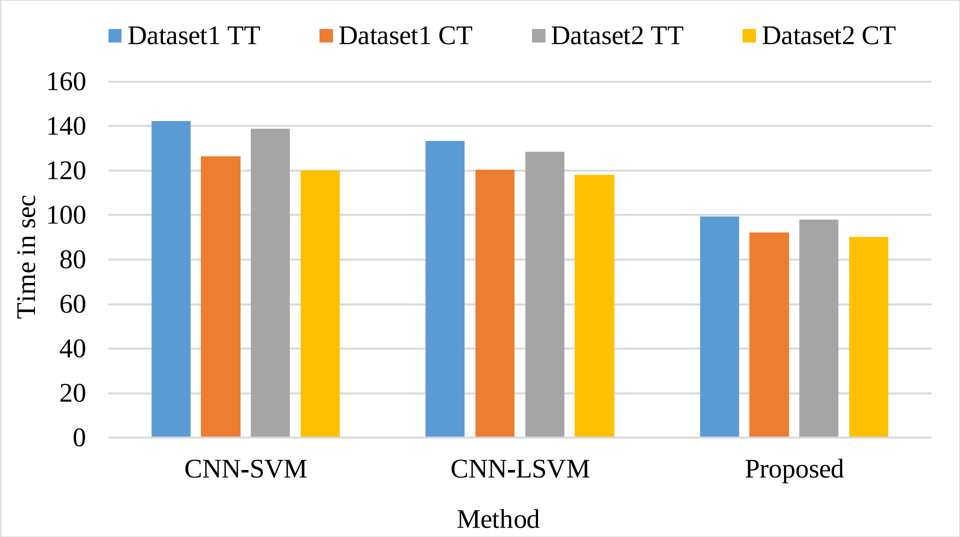

Fig. 4 Analysis comparing training time and computation time of two datasets of the proposed method with CNN-SVM and CNN-LSVM

The training time and computation time of the two datasets are depicted in Figure 4. It is evident that the proposed method has less Training Time (TT) of 99.34 sec for dataset1 and 97.35 sec for dataset2 respectively and less Computation Time (CT) of 92.13 sec for dataset1 and 90.13 sec for dataset2 respectively, as related to the other two methods.

5. Conclusion

The proposed framework successfully integrates advanced machine-learning techniques to address the challenges of biomedical document classification. By constructing a Document Dependency Graph (DDG) and utilizing improved CNNs for feature extraction, the framework effectively captures and utilizes the complex relationships between genes, diseases, and chemical drugs.

The subsequent classification through a multi-class SVM with ECOC, optimized for handling the multi-class nature of biomedical data, demonstrates a significant improvement over traditional methods. The proposed model has been evaluated on two datasets, OMIM and ClinVar, and achieved an accuracy of 98.8% to 98.9% and a reduced training time of 99.34 sec to 92.13 sec, respectively. Overall, this framework represents a significant advancement in biomedical data analysis, providing a more accurate, efficient, and contextually aware method of document classification. The hybrid combination of CNNs for feature extraction and M-SVMs for classification, along with the incorporation of gene-disease and chemical contextual similarities, marks a substantial step forward in the application of machine learning techniques to biomedical research.

References

[1] Martín Pérez-Pérez et al., “A Novel Gluten Knowledge Base of Potential Biomedical and Health-Related Interactions Extracted from the Literature: Using Machine Learning and Graph Analysis Methodologies to Reconstruct the Bibliome,” *Journal of Biomedical Informatics*, vol. 143, 2023. [CrossRef] [Google Scholar] [Publisher Link]

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

[3] Xiaofeng Liu et al., “A Syntax-Enhanced Model Based on Category Keywords for Biomedical Relation Extraction,” *Journal of Biomedical Informatics*, vol. 132, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Evan French, and Bridget T. McInnes, “An Overview of Biomedical Entity Linking Throughout the Years,” *Journal of Biomedical Informatics*, vol. 137, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Kairui Guo et al., “Artificial Intelligence-Driven Biomedical Genomics,” *Knowledge-Based Systems*, vol. 279, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Esmaeil Nourani, and Vahideh Reshadat, “Association Extraction from Biomedical Literature Based on Representation and Transfer Learning,” *Journal of Theoretical Biology*, vol. 488, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Elizabeth S. Chen et al., “Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study,” *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 87-98, 2008. [CrossRef] [Google Scholar] [Publisher Link]

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

[9] Alberto G. Jácome, Florentino Fdez-Riverola, and Anália 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, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[10] Martín Pérez-Pérez et al., “Boosting Biomedical Document Classification through the Use of Domain Entity Recognizers and Semantic Ontologies for Document Representation: The Case of Gluten Bibliome,” *Neurocomputing*, vol. 484, pp. 223-237, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Laura Plaza, “Comparing Different Knowledge Sources for the Automatic Summarization of Biomedical Literature,” *Journal of Biomedical Informatics*, vol. 52, pp. 319-328, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[12] Hyunjin Shin et al., “Comparing Research Trends with Patenting Activities in the Biomedical Sector: The Case of Dementia,” *Technological Forecasting and Social Change*, vol. 195, 2023. [CrossRef] [Google Scholar] [Publisher Link]

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

[14] Saranya Muniyappan, Arockia Xavier Annie Rayan, and Geetha Thekkumpurath Varrieth, “EGeRepDR: An Enhanced Genetic-Based Representation Learning for Drug Repurposing Using Multiple Biomedical Sources,” *Journal of Biomedical Informatics*, vol. 147, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Xu Ling et al., “Generating Gene Summaries from Biomedical Literature: A Study of Semi-Structured Summarization,” *Information Processing & Management*, vol. 43, no. 6, pp. 1777-1791, 2007. [CrossRef] [Google Scholar] [Publisher Link]

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

[17] Nichola Foster et al., “IBM Watson AI-Enhanced Search Tool Identifies Novel Candidate Genes and Provides Insight into Potential Pathomechanisms of Traumatic Heterotopic Ossification,” *Burns Open*, vol. 7, no. 4, pp. 126-138, 2023. [CrossRef] [Google Scholar] [Publisher Link]

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

[19] Tomoki Tsujimura, Makoto Miwa, and Yutaka Sasaki, “Large-Scale Neural Biomedical Entity Linking with Layer Overwriting,” *Journal of Biomedical Informatics*, vol. 143, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Tommaso Mario Buonocore et al., “Localizing In-Domain Adaptation of Transformer-Based Biomedical Language Models,” *Journal of Biomedical Informatics*, vol. 144, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Hermenegildo Fabregat et al., “Negation-Based Transfer Learning for Improving Biomedical Named Entity Recognition and Relation Extraction,” *Journal of Biomedical Informatics*, vol. 138, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Bernhard Schölkopf, and Alexander J. Smola, *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond*, The MIT Press, 2018. [CrossRef] [Google Scholar] [Publisher Link]