An Efficient Model for Medical Data Classification using Gene Features

Ensemble Model for Text Classification (EMTC)

Kosaraju Chaitanya[1], Rachakonda Venkatesh[2]
Department of CSE
Vignan's Nirula Institute of Technology and Science for Women, Peda Palakaluru, Guntur-522009
Andhra Pradesh
India

Thulasi Bikku[3]
Department of CSE
Vignan's Nirula Institute of Technology and Science for Women, Peda Palakaluru, Guntur-522009
Andhra Pradesh
India

***Abstract*—In the medical field to solve the new issues, the novel approaches for managing relevant features by using genomes are considered; using the sub-sequence of genes the outcome of interest is analyzed. In this implementation part of the model, we have given the MEDLINE and PubMed archives as inputs to the proposed model. A large number of MESH terms with gene and protein are utilized to characterize the patterns of a large number of medical documents from a large set of records. Standard datasets with different characteristics are used for examination study. The characteristics and inadequacies of different techniques are noted. Feature selection techniques are given in perspective of data composes and region traits by applying proper rules. Feature context extraction through name element distinguishing proof is an essential errand of online therapeutic report grouping for learning disclosure databases. The parameters are identified to compare with other models implemented on these datasets and the results prove that the proposed method is very effective than existing models. The primary point of the proposed ensemble learning models is to characterize the high dimensional information for gene/protein-based disease expectation in light of substantial biomedical databases. The proposed model uses an efficient ranking algorithm to select the relevant attributes from a set of all attributes; the attributes are given to the classifier to improve the accuracy based on the users' interest.**

Keywords*—Classification; Hadoop framework; biomedical documents; feature selection; gene features; medical datasets

I. Introduction

In the bioinformatics field, the Dimensionality reduction procedures have turned into a need. The relevant features are used for effective reduction of training time and overfitting of the model. In the medical area, a colossal measure of information is produced and put away in the medicinal archives. The reports or archives contain the data about the indications of a patient and furthermore numerous medical test reports that might be produced. The dominant part of genuine classification issues require algorithm of supervised learning in which the hidden class probabilities and conditional probabilities are ambiguous, which are not clearly mentioned, and each instance in test data is related with a class name. A candidate feature is neither superfluous nor repetitive to the objective idea; an irrelevant component does not influence the objective idea at all, and irrelevant attributes or repeated attributes does not add anything new to the objective idea. In many applications, due to the large datasets, the learning algorithms never give accurate results because of irrelevant and spurious attributes, so it is essential to segregate the relevant features from large data repositories. Diminishing the quantity of superfluous/repetitive features definitely decreases the algorithm's training time and yields a negative impact on the objective function [1]. The feature selection methodology does not make novel attributes, it only chooses unique features. One of the difficulties comprise for the development of model is to choose the features when selection is required. Selections of relevant features are measured to check the properties of the data might be required to settle on a decision.

In the meantime, feature classification, selection, and feature construction can be used together to improve the execution of the objective function and lessen the dimensionality of the datasets. The model construction can be in three distinctive ways: 1) before constructing the model feature classification and selection are performed; 2) feature construction should be done; and 3) in the meantime performing feature classification, selection, and construction. These aides in showing signs of improvement knowledge into the basic idea of a true classification issue. These days, the development of the high-throughput innovations has brought about exponential development in the reaped information as for both dimensionality and test data size. Feature selection techniques endeavor to choose a subset of attributes that are applicable to the objective idea. With the quick growth of biomedical text documents, automatic text analyzing systems is increasing at an exponential rate to find relevant patterns from the distributed biomedical data [2]. Data mining and machine learning models have been successfully implemented in various bio-medical domains such as cancer detection from pathology reports, disease diagnosis and predicting complex diseases [3]. Irrelevant features not only results high true negative rate but also lead to inaccurate classification results. Text classification and feature prediction are the important issues of biomedical text documents in a large number of distributed biomedical applications [4]. Eliminating irrelevant features from the large bio-medical databases facilitates noise reduction and optimizes the classification accuracy. Distributed data mining is a functioning exploration region for structure high computational and for proficient decision-making design patterns for learning based applications in the bioinformatics [5].

Because of the exponential development of models in the distributed environment, a large number of server farms have been shaped and collected huge measures of the information. It is important to find hidden patterns of knowledge to improve decision making from those databases in the bioinformatics [6]. The programmed arrangement of restorative reports into predefined classes is developing quickly on online biomedical repositories [7]. The current arrangements that require earlier information of classification accuracy for different sorts of relevant attributes, which is difficult to acquire day by day. The commitment of the proposed research work for biomedical databases is feature selection and classification and Hadoop-based decision tree model; which is used to construct the patterns based on the user requirements in large data sets. In this scenario, one of the most concerning issues persuaded by the exponential development of repositories about biomedical data is to help researcher in finding valuable data from the databases of MEDLINE utilizing PubMed web search tool in making decisions including functions about gene, patterns of diseases using genes, associations among the genes and MeSH knowledge discovery patterns [8].

The search results using the synonyms of drugs, genomes, disease names, and medical terms browsed by the users in the internet are shown in Fig. 1. The user searches the medical documents based on the MeSH terms or general medical terms in the web, based on the user interface the medical documents are extracted from the biomedical repositories. The document extraction is based on the ranking algorithms and synonyms of the medical terms based on the user requirement.

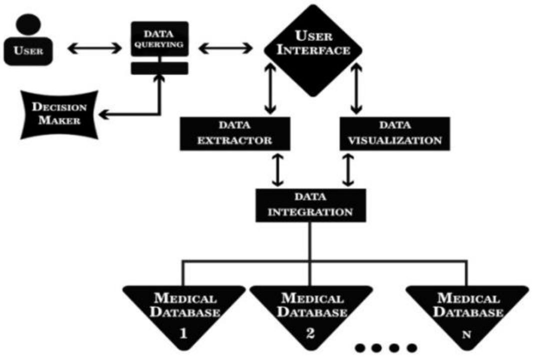

Fig. 1. Users Search for Medical Documents.

Traditional Hadoop based distributed ensemble models such as Random forest, Fuzzy c-means and Multi-objective ensemble classifier; Bayesian ensemble models were implemented to discover cluster based decision patterns with a few feature vectors. In this dynamic model, an ensemble medical document classification model was developed to discover the clustered feature vectors with patterns in the distributed medical datasets. As the traverse of the online biomedical databases is creating well ordered, analyzing interesting patterns in a fundamental structured or unstructured biomedical applications have ended up being more troublesome with traditional algorithms [9]. Additionally, because of the accessibility of various biomedical reports in the PubMed and MEDLINE documents, it is all the more testing to investigate information, foresee patterns and translate the documents data utilizing the customary information mining by using classification and clustering.

A typical framework of a distributed approach works as follows: it performs the analysis of the local dataset at each distributed peer node, and then the discovered information is migrated to a central site where the integration of the distributed local nodes is performed. The computed results are returned to the distributed data sets so that all nodes contain the updated information. Peer-to-peer (P2P) medical system involves a large number of peers in the network without any centralized control, in which each peer node exchanges and share data throughout the network. Each peer node is linked directly to a large number of peers within the overlay network. It is impractical to gather all the data distributed in the P2P network into a centralized node or site and then perform the conventional data mining techniques. P2P overlay networks emerged as a promising mechanism to manage and share the high dimensional data in the client-server architecture [10]. A high-performance Hadoop framework has become extremely important for most of the large scale data applications such as Medline and Pubmed [11].

A typical framework of a distributed approach works as follows: it performs the analysis of the local dataset at each distributed peer node, and then the discovered information is migrated to a central site where the integration of the distributed local nodes is performed. The main ideology behind implementing the distributed environment in bioinformatics is easy and effective to implement, reliable, fault tolerance, scalable in different dimensions. To process the large amounts of data, the single cannot run in stipulated time, it has many limitations like processing speed, storage etc., so we go for distributed environment to retrieve the knowledge patterns. The biomedical datasets contain heterogeneous data, the data from multiple sources must be integrated and the data is shared between the repositories. The primary sources of medical databases are GENEBANK, EMBL, NCBI, NDB and SWISS-PROT. The DNA information of the living being is encoded by genomes, to know about their heritage of gene, so that it will be easy to predict their diseases and health conditions of the patient.

II. Literature Survey

The literature survey on biomedical repositories contains the information about the preprocessing techniques of the medical documents. The preprocessed data converts the raw input into output as useful structured format by avoid missing data and to provide clean data [12]. Bioinformatics is the field of study about biological data involving gene ontologies and genomics. It contains the data of DNA and protein sequences, gene expression data, images which are incomplete in nature. The bioinformatics data can be stored, managed, analyzed using algorithms and statistical methods and tools of computer science. The biomedical data is organized and stored in the databases; the new entries of data are also stored in the large repository.

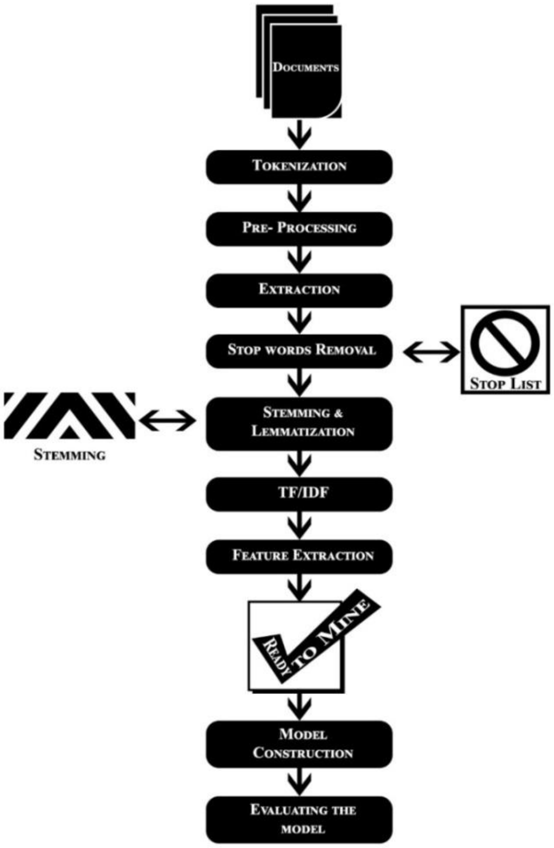

Fig. 2. The Preprocessing Steps.

The identification of gene can be done by using the name of the gene or its equivalent word in the title part of the record in the datasets [13]. The mining used to discover patterns for the discovery of genes can be performed in two stages: Initially, every one of the terms identified with genes out of a MEDLINE database is distinguished and extricated. Here, TF-IDF (Term Frequency-Inverse Document Frequency) and Z-scores methodologies are actualized for the identification of related genes. In the subsequent advance, the report preprocessing approach is executed to frame the gene features shown in Fig. 2.

The data repository stores the information of different formats like genomics, medical images, incomplete data, which should be cleaned in the preprocessing stage. Thus the data obtained having many dimensions having irrelevant data. The feature selection of data is not an easy procedure, which would not be possible manually. High quality data gives good results, removing the irrelevant data may not affect the accuracy of the classifier. The medical documents may contain the stop words i.e. they are commonly used words like articles and prepositions, sometimes contain symbols like ‘,’, ‘.’, ‘_’, ‘-’ are removed. The stemming is the process of removing the words having the extensions like ‘ed’, ‘ing’. The tf/ idfare used to identify the key terms of the medical document based on the frequency of the terms used in the document. Now the preprocessed data is ready to mine based on the user required queries.

The traditional Relational Database Management System (RDBMS) and Object-Oriented Database Management System (OODBMS) are not suitable to retrieve the genome data due to the complexity of the data. So the researches are done to discover method to store, manage the data based on the users’ requirement by extending the existing methods. The pre-analysis of information consequently chooses or unselects various alternatives and gives valuable data to the client dependent on his prerequisite. It begins by checking if all the gene expression patterns have a similar number of conditions and a valid identifier. It rectifies them generally by altering the number of components in each row and including a unique identifier when necessary.

Map-Reduce are a framework that enabled the processing of large-scale databases for the first time [14]. This framework processes datasets in a distributed platform with introducing two phases, Map and Reduce functions. Named entities in the medical text documents are structured and unstructured due to its boundary detection and feature classification. Pietari Pulkkinen and Zhihua Wei, 2008 implemented name entity identification approaches from biomedical databases using support vector machine classifier [15]. Junshan Zhang, 2013, proposed a probabilistic model of name entity identification with decision tree approach. This model proposes the unsupervised system to design an index based clinical documents classification from the large corpus [16]. Cheng Hua Li, 2014 implemented a topic modeling using principal component analysis to achieve correlation between the feature vectors in different databases [17]. A number of biomedical documents have been used in the development of parallel computing and data mining models on the MapReduce framework. Typical features of textual machine learning models include iterative data scanning and feature extraction, which require high performance and computation over the distributed data.

By using the Biomedical Name Entity Recognition (BNER) used to filters the various biomedical entities such as genes, proteins, cell-type, etc. by using. The taggers which are commonly used are GENIA, NLProt, ABNER, and BANNER [18]. GENIA predicts a biomedical name entity using Natural Language Processing (NLP) tasks such as tokenizing, chunking, part of speech (POS) tagging and machine learning measures for entity detection. Both ABNER and BANNER outlined and built up an algorithm for data learning model utilizing contingent irregular fields with relevant attributes [19]. Nonetheless, the parameter estimation for contingent arbitrary fields is costly, when the complexity of space and time are concerned. This issue winds up harder while training on numerous biomedical databases with constrained processing resources and assets. At long last, GENETAG is used to filter the channels of gene or protein labels from biomedical records by word reference or pattern matching coordinating models of GenBank substances.

Information extraction in the medical domain includes extraction of proteins, genes, and extraction of the relationship between these entities. Herman implemented a system that retrieves and visualizes domain knowledge from literature and database using gene names with high recall and low precision and it is difficult to detect the semantic relationships among the entities [20]. Another approach is a rule-based technique, which aims to detect and extract bimolecular relations using inhibit relation achieved 91% of precision and 56% of recall.

The supersets of all the terms in the bioinformatics are stored in the medical database known as Unified Medical Language System (UMLS). The searching operation is performed on the UMLS to select the terms used to retrieve the medical text files. If a term is detected in the UMLS, that specific term is considered as a biomedical term and included in the document. Else, the terms are ignored and deleted from the document.

A different decision tree-based multi-objective estimation of conveyance algorithm (DT-MEDA) for improvement concerns through uninterrupted features is generated. Decision tree-based probabilistic models are utilized to convert contingent conditions stuck between factors in DT-MEDA [21]. As of late, a substantial number of statistical machine learning strategies are connected to the content classification framework. The utilization of the most punctual machine learning strategy is naive Bayes classification (NBC). Along these lines, all the essential machine learning calculations have been connected to the field of content classification, for instance, *K* nearest neighbour (KNN), neural network system (NNS), support vector machine (SVM), decision tree, kernel learning, and some others [22]. The expectation maximization (EM) algorithm is used to distinguish themes in unaligned biopolymer groupings. Late examination have given opposing outcomes when contrasting Bayesian estimators with Expectation Maximization (EM) for unsupervised Hidden Markov Model (HMM) Parts Of Speech (POS) tagging, and we demonstrate that the distinction in detailed outcomes is to a great extent because of contrasts in the training data size and the number of states in the Expectation Maximization - Hidden Markov Model (EMHMM). In the proposed model (EMTC), a large number of documents are processed with high true positive and precision. Experimental results prove that the proposed approach has a high recall, precision, and less standard error compared to conventional ensemble approaches. The proposed models used for biomedical classification using the Hadoop framework can be extended further for semantic entity mining and deep learning as follows.

III. Proposed Model

The proposed method mainly concentrates on Ensemble Model for Text Classification (EMTC) shown in Fig. 3. Let document D is segmented into a set of medical phrases or sentences in a peer document set. Let D= {p1, p2, p3,....,pn} and T={t1,t2,t3,...., tm} represents all the term occurs in a document set, where n is the number of phrases and m is the number of terms in a document D. The major challenge in the document classification is a physical representation of the document set in the distributed environment, which will not work properly in a high dimensional vector space model.

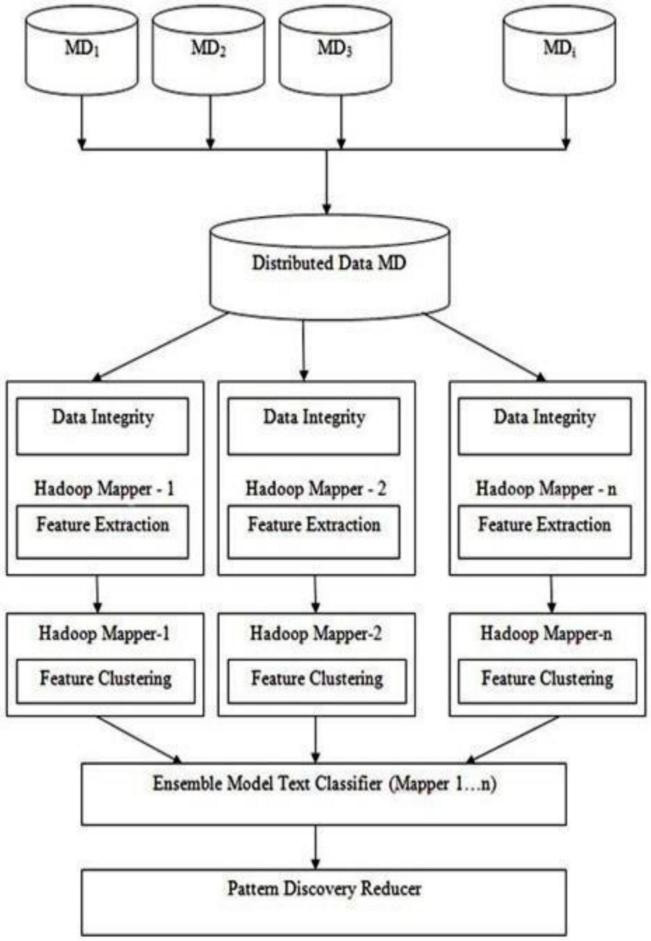

Fig. 3. Proposed Method (EMTC) for Feature Clustering.

Similarity measures are used in text related document repositories and applications such as information retrieval, language processing, text mining, and text clustering and prediction process. An enhanced MapReduce framework was used to discover the textual patterns from a large collection of structured or unstructured medical document sets. Also, multiple Hadoop mapper interfaces are assigned to each distributed data source for domain knowledge identification. Medical documents are integrated and preprocessed using the feature extraction algorithm of the Hadoop's Mapper interface; each Mapper phase is responsible to prune the structured, unstructured and inconsistent features.

After the feature extraction process, an ensemble cluster-based classification model was applied to discover the interesting patterns for medical documents. Ensemble integrates the results of multiple classifiers; thus assists to overcome the possible issues of classifiers in medical databases. The proposed model is evaluated on two datasets, namely GENETAG and Medline distributed datasets.

A. Combining Multiple Distributed Data Integration

```
Input: Heterogeneous Medical datasets
Output: Single repository with integrated data
Procedure: The medical repositories with names BD_1 and BD_2
for attribute-m in BD_1 do
for attribute-n in in BD_2 do
if(attribute-m≠NULL&&attribute-n≠NULL) then
if(type(attribute-m)==type(attribute-n)) then
Similarity (attribute-m, attribute-n) = (Parent(attribute-m/attribute-n)*Parent(attribute-n))*Correlation(attribute-m,attribute-n);
// the similarity between two attributes are measured
The correlated attributes are mapped using
Map((attribute-m, attribute-n), similarity(attribute-m, attribute-j));
Done;
To integrate the attributes, we merge the attributes having maximum similarity.
Select attributes having maximum similarity pair (attribute-m, attribute-n)
Repository contains attributes integrated with attribute-m and attribute-n mapped with map ((attribute-m, attribute-n), similarity (attribute-m, attribute-n));
Done;
```

B. Proposed Model (EMTC)

```
Input: Dynamic Document sets, number of data sources
Output: Preprocessed integrated dataset
Mapper1 (1...m, index)//Hadoop mappers indexing from 1 to m
Step1: for each document 'd'in the data source DS do
//for feature selection, where did is document id, tf is term frequency, itf is inverse term frequency and wt is weight measure
<did,index,tf,itf,wt>=featureselection(d);
Document representation should be done as <did, index,tf,itf,wt>
MD(1…n)=addDocument(<did,index,tf,itf,wt>);
done;
Step2: for each documented in MD(1…n)
do
Remove stop words
Remove special characters
List the medical features‘’ in the document using Medical Subject Headings (MeSH) terms.
Extract the synonyms’ ‘to the document features‘’ using GENETAG database.

Step3:find the feature clustering using the equation (1) as do
Conditional probabilityCP (tfi, θ j)=Σ(prob(tfi∩ θ j)/prob(tfi))*wtj;
Fclust( ρ i,fj)=Σj=1Cov( θ , ρ j)*CP(tfi, θ j);
//Covariance provides a measure of the strength of random variables between two or more sets
List<id,Fclust>
Mapper2<1...m,list<id,Fclust>>done
Step4: Estimate the decision parameters using co-decision matrix D from the instances set of all classifiers. For each data, source computes the patterns using the following models do
Apply Neural Network NN, Random Forest RF, Ensemble learning model, and Genetic algorithm based support vector machine GA_SVM and HBEC models.
Reduce<model, patternlist>done
Reducer<model, patterns>
Step5: For each model in the list do
Integrate and visualize the top k patterns in the pattern list;
done;
Given a set of medical document features
Feature, F={feature-1,feature-2…feature-n},the set contains n features.
For each document term t in d do
MI(t,d)=log(p(t/d)/p(t))=log(p(t,d)/p(t).p(d));
if(MI(t,d)>0)
wt=log(N/tf)*itf;
//tf-term frequency, itf-inverse term frequency and N-number of documents
Else continue;
done;
```

C. Distributed Databases (PubMed and Medline) used to Extract the Top Most Genes using Ranking Algorithm

```
Input: Database contains gene synonyms (GENETAG-DB), Gene Documents document sets from PubMed and Medline.
Output: Top ranked Distributed Gene-Synonym Documents
Procedure:
For each synonym used to describe genomes in GENETAG-DB do
For each Gene Documents document set do
GeneDocumentlist[i] =Gene Documents;
For each token t in GeneDocumentlist[i] do
getGeneSynonym[]=GeneSynonym(t);
PgetGeneSynonymDocuments[]=PubMed(Url(disease,getGeneSynonym[]);
MgetGeneSynonymDocuments[]=Medline(Url(disease,getGeneSynonym[]);
To check the mutual independence of two genes in the given contextual document use mutual information.
The common probability of the Document i, Pro(Di∩ getGeneSynonym[j])
The sum of all probability of Document i, Pro(Di∩ getGeneSynonym[j])
Probability of the total gene synonym documents Di Pro(Di )
Thres=0.5;// user-defined threshold for ranking
If(MFR>Thres)
Then
Add GESynonymDocumentList (MFR, PgetGeneSynonymDocuments, EgetGeneSynonymDocumentss);
End if
Done
Done;
```

IV. RESULTS

After The Medline and GENETAG datasets are used for the experimental results. The data consist of a large collection of medical documents with different features. A total of one million medical documents were processed for training and testing in these experimental results. In this implementation part of the model, we have given the MEDLINE and PubMed archives to the proposed model. To implement the model, we utilize the present Apache Hadoop system with Amazon AWS server. The design of Amazon AWS server consists of cluster nodes ranging from 10-50, each with 10 CPU cores and 24 GB RAM is connected to the mapper node. A large number of MESH terms with gene and protein are utilized to characterize the patterns of a large number of medical documents from a great many records. These compared algorithms are done using Map/Reduce framework using Java Programming on Cluster Nodes.

From Table I, it is clear that the percentage of accuracy is achieved by the proposed model (~97%) are higher than the traditional ensemble model on different document set. As the size of the document increases the average runtime is decreased for the proposed model as compared to traditional models.

Document classification measures such as true positive rate, recall, precision are summarized in Table II. Also from the table, it is clearly observed that the proposed model has high computational efficiency (at least 5-10%) than the traditional models.

TABLE. I. ACCURACY TERMS OF DOCUMENT PROCESSED TIME

Algorithm	**Accuracy**		**Documents Size in MB**	**Average Running Time (secs)**
	40K	100K		
Multi-objective DT	72.75	76.35	5	235
Multi-text Bayesian Classifier	83.1	80.84	6	199
EMHMM	86.34	83.5	7	178
Neural Networks	90.28	88.33	8	204
Proposed Model (EMTC)	96.86	96.98	10	135

TABLE. II. TRUE POSITIVE, PRECISION AND RECALL OF PROPOSED METHOD COMPARED WITH TRADITIONAL METHODS

Algorithm	**True Positive Accuracy**	**Precision**	**Recall**
Multi-objectiveDT	73.65	71.65	75.25
Multi-text Bayesian Classifier	86.36	84.76	84.08
EMHMM	87.83	89.43	90.43
Neural Networks	89.64	90.53	91.87
Proposed Model (EMTC)	95.75	94.98	96.37

From Fig. 4, it is clear that the true positive rate scores achieved by the proposed model (~95.75%)are higher than the traditional ensemble model on different node configurations with different metrics.

From Table III, it is clear that the proposed model has high computational efficiency in terms of time and memory is concerned.

The proposed model (EMTC) completely eliminates the duplicate documents in the distributed databases (PubMed and Medline) using Top k-ranking algorithm.

From Fig. 5, it is clear that the proposed model performance in terms of time and memory is concerned. Proposed model completely eliminates the duplicate documents in the distributed databases (PubMed and Medline) using Top k-ranking algorithm. The proposed model is far better than traditional algorithms.

Table IV describes the feature extraction process of the gene to disease datasets using the proposed feature extraction models to the existing parameters. From the table, it is clearly observed that the proposed feature extraction model has high computational efficiency and less average runtime than the traditional gene-disease extraction measures.

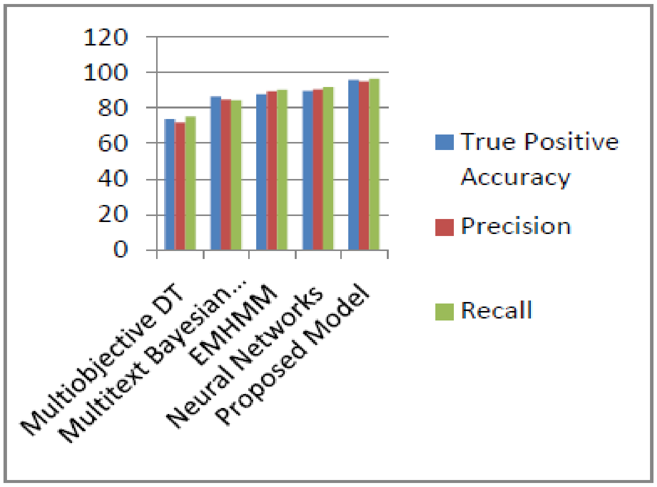

Fig. 4. Comparing the Proposed Model with Traditional Model.

TABLE. III. PERFORMANCE ANALYSIS OF TOP RANKING GENE-BASED METHODS

Algorithm	**Duplicate Filtered (Documents)**	**Time (ms)**	**Memory (Kb)**
Multi-objective DT	6524	12763	25.76
Multi-text Bayesian Classifier	6935	11863	21.86
EMHMM	5987	12865	28.34
Neural Networks	5709	10273	23.86
Proposed Model (EMTC)	3876	8963	17.95

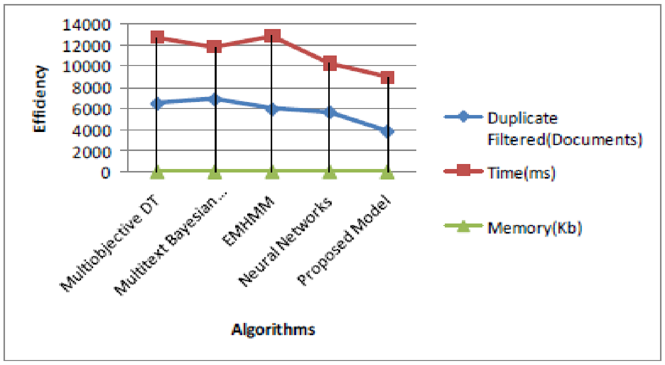

Fig. 5. Performance Analysis of Top Ranking Gene-based Documents.

TABLE. IV. GENE FEATURE EXTRACTIONS AND ITS ACCURACY

Parameters	**Feature extraction Avg Accuracy**	**Avg Runtime(ms)**
Information Gain	87.45	9743
Gain Ratio	89.29	8974
Chi-square	91.73	8689
Correlation	85.87	9217
Rough set	92.75	8248
Gene mutual information	97.53	7936
Gene Chi-square Measure	95.25	8013

In the proposed model (EMTC) is compared with different traditional algorithm models like Multi-objective DT, Multi-text Bayesian Classifier, EMHMM and Neural Networks, a large number of documents are processed with high true positive and precision. Experimental results prove that the proposed approach has a high recall, precision, and less standard error compared to conventional ensemble approaches. The proposed models used for biomedical classification using the Hadoop framework can be extended further for semantic entity mining and deep learning also. The proposed work can also be reached out to locate the perplexing qualities DNA-ailments designs utilizing parallel semantic mining model. Sem MedDB, a repository of semantic predictions used to find the relationship between the DNA sequence to disease, gene-disease and its clinical trials using the Hadoop framework. A novel Hadoop based semantic mining model on biomedical databases for multi-pattern evaluation using gene-disease patterns. An extension of parallel semantic mining model is used to find the relationship between two semantic entities (e.g. genedisease1, gene--disease2, etc.) and to check the existence of the disease patterns in more than three repositories. The proposed work can be extended to work on the disease to DNA or Protein sequences using distributed databases. The proposed work can be extended to deep learning models convolution neural networks for biomedical text classification. The research can be advanced to the disease-related patterns and its auto clinical decision system for medication. This work can be extended to optimize computing resources such as memory and time, when the data size is large (>30million records).

In future work, we will develop a multi-objective ensemble classification model to improve accuracy and outliers.

V. CONCLUSION

Here we propose a novel model, which uses a different classification and feature selection approaches makes it possible to select relevant genes with high confidence. A new feature selection and classification algorithm EMTC is implemented and evaluated by comparing with related feature selection algorithms. Our proposed model demonstrates its efficiency and effectiveness in dealing with high dimensional data for classification. As a result, text mining has evolved in the field of biomedical systems where text mining techniques and machine learning models are integrated using high computational resources. The main purpose of the work is to enhance the techniques used for feature selection and classification of high dimensional datasets of biomedical repositories and compared with different traditional algorithms. The results prove that the proposed model is efficient in selecting the relevant features and attains the outcomes as per users' interest with very low runtime. The accuracy of any model using relevant features is better than the complete features in the entire dataset.

ABBREVIATIONS

BOW: bag-of-words
MEDLINE: Medical Literature Analysis and Retrieval System Online
MeSH: Medical Subject Headings
XML: extensible markup language
BNER: Biomedical Named Entity Recognition
NLP: Natural Language Processing
POS: parts of speech
SVM: support vector machines
NMF: Nonnegative Matrix Factorization
sLDA: supervised Latent Dirichlet Allocation
UMLS: Unified Medical Language System
NER: Named Entity Recognition
DT-MEDA: decision tree-based multi-objective estimation of conveyance algorithm
NBC: naive Bayes classification
KNN: *K* nearest neighbour
NNS: neural network system
EM: expectation maximization
HMM: Hidden Markov Model
EMTC: Ensemble Model for Text Classification

ACKNOWLEDGMENT

We thank our college management, especially Dr. Lavu Rathaiah Garu, Chairman of Vignan's Group of Institutions and our beloved Principal Prof. Paturi Radhika Garu for supporting us, who put their faith in us. This work is mainly based on biomedical data extraction from the corpus, so we thank the referees in this wide area. We thank each and every one who supported us to accomplish this work successfully. This material has not been published in whole or in part elsewhere and not funded by any organization or committee. The author(s) received no specific funding for this work.

References

[1] Kamath, U., De Jong, K., & Shehu, A. (2014). Effective automated feature construction and selection for classification of biological sequences. PloS one, 9(7), e99982. https://doi.org/10.1371/journal.pone.0099982.

[2] Bikku, T., Nandam, S. R., & Akepogu, A. R. (2018). A contemporary feature selection and classification framework for imbalanced biomedical datasets. Egyptian Informatics Journal, 19(3), 191-198. https://doi.org/10.1016/j.eij.2018.03.003.

[3] Wagholikar, K. B., MacLaughlin, K. L., Kastner, T. M., Casey, P. M., Henry, M., Greenes, R. A., & Chaudhry, R. (2013). Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. Journal of the American Medical Informatics Association, 20(4), 749-757. https://doi.org/10.1007/s10916-011-9780-4.

[4] Chen, C. M., Lee, H. M., & Chang, Y. J. (2009). Two novel feature selection approaches for web page classification. Expert systems with Applications, 36(1), 260-272. https://doi.org/10.1016/j.eswa.2007.09.008.

[5] Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics, 43(1), 15-23. https://doi.org/10.1016/j.jbi.2009.07.008.

[6] Bikku, T., & Paturi, R. (2019). A novel somatic cancer gene based biomedical document feature ranking and clustering model. Informatics in Medicine Unlocked, 100188. https://doi.org/10.1016/j.imu.2019.100188.

[7] Hu, C., Xu, Z., Liu, Y., Mei, L., Chen, L., & Luo, X. (2014). Semantic link network-based model for organizing multimedia big data. IEEE Transactions on Emerging Topics in Computing, 2(3), 376-387. https://doi.org/10.1109/TETC.2014.2316525.

[8] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., .& Gocayne, J. D. (2001). The sequence of the human genome. science, 291(5507), 1304-1351.

[9] Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter, 14(2), 1-5. https://doi.org/10.1145/2481244.2481246.

[10] Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of data grids for distributed data sharing, management, and processing. ACM Computing Surveys (CSUR), 38(1), 3. https://doi.org/10.1145/1132952.1132955.

[11] Bikku, T., Rao, N. S., & Akepogu, A. R. (2016). Hadoop based feature selection and decision making models on big data. Indian Journal of Science and Technology, 9(10). https://doi.org/10.17485/ijst/2016/v9i10/88905.

[12] Feldman, R., & Sanger, J. The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data. https://doi.org/10.1017/CBO9780511546914.

[13] Piech, C., Sahami, M., Koller, D., Cooper, S., & Blikstein, P. (2012, February). Modeling how students learn to program. In Proceedings of the 43rd ACM technical symposium on Computer Science Education (pp. 153-160). ACM. https://doi.org/10.1145/2157136.2157182.

[14] Bikku, T. (2017, August). A Novel Multi-Class Ensemble Model for Classifying Imbalanced Biomedical Datasets. In IOP Conference Series: Materials Science and Engineering (Vol. 225, No. 1, p. 012161). IOP Publishing. https://doi.org/10.1088/1757-899X/225/1/012161.

[15] Pulkkinen, P., & Koivisto, H. (2008). Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms. International Journal of Approximate Reasoning, 48(2), 526-543. https://doi.org/10.1016/j.ijar.2007.10.004.

[16] Qian, D., Zheng, D., Zhang, J., Shroff, N. B., & Joo, C. (2013). Distributed CSMA algorithms for link scheduling in multihop MIMO networks under SINR model. IEEE/ACM Transactions on Networking (TON), 21(3), 746-759.

[17] Zhang, L., Luo, Y., Tao, F., Li, B. H., Ren, L., Zhang, X., ... & Liu, Y. (2014). Cloud manufacturing: a new manufacturing paradigm. Enterprise Information Systems, 8(2), 167-187.

[18] Raja, K., Patrick, M., Gao, Y., Madu, D., Yang, Y., & Tsoi, L. C. (2017). A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries. International journal of genomics, 2017. https://doi.org/10.1155/2017/6213474.

[19] Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In Biocomputing 2008 (pp. 652-663). https://doi.org/10.1142/9789812776136_0062.

[20] Leroy, G., McDonald, D. M., Su, H., Xu, J., Tseng, C. J., & Chen, H. (2005). Gene Pathway Text Mining and Visualization. https://doi.org/10.1007/0-387-25739-X_18.

[21] Zhong, X., & Li, W. (2007, December). A decision-tree-based multi-objective estimation of distribution algorithm. In 2007 International Conference on Computational Intelligence and Security (CIS 2007) (pp. 114-11). IEEE. https://doi.org/10.1109/CIS.2007.136.

[22] Shih, B. (2011). Target sequence clustering (Doctoral dissertation, Ph. D. Thesis, Machine Learning Department, Carnegie Mellon University).G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.