Gene-disease and chemical contextual similarity based multi-class SVM and deep learning framework

JOSE MARY GOLAMARI[1], D. HARITHA[2]

[1]Research Scholar, Department of Computer Science and Engineering, , Koneru Lakshmaiah Education Foundation,Vaddeswaram, Guntur, Andhra Pradesh.

[2]Professor, Department of computer science and engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh

ABSTRACT

Background: Text feature ranking is an essential step in biomedical data analysis which directly affects the text prediction models. Traditional feature extraction methods such as mutual information, probabilistic information gain, statistical chi-square and inverse document frequency are used to find the essential key terms and its relationships in document sets of limited size. Semantic relationships plays a vital role in large document sets is a challenging task for text mining applications. In the biomedical applications, predicting the gene-disease based chemical drug is difficult due to large number of training datasets and sparsity issue. In order to overcome these issues, a multi-layered graph based deep learning framework is designed and implemented on the large biomedical gene-disease and chemical drug data. In this framework, an extended version of particle swarm optimization and Multiclass Support Vector Machine (IPSO-MSVM) model are developed to extract the key terms for document analysis.

Results: Experimental results are tested on different biomedical document sets using the proposed gene based chemical drug database. Results show that the present framework has better true positivity and accuracy than the stat of art algorithms .

Conclusion: Here, a non-linear support vector machine is used in CNN framework to classify the gene disease patterns using the medical datasets. After the analysis of data the gene patterns are identified, based on this the drugs are discovered. So it will more useful for the clinical experts for decision making about the patients and helps to predict the diseases they are prone to and for discovering the drugs based on their genes.

Keywords: iPSO, CNN, MSVM, Deep Learning, Genes, Drug Discovery

1.Introduction:

The discovery of new medications and bioactive substances is a complex and resource-intensive process that poses significant challenges to researchers and healthcare providers alike. Our team has developed a state-of-the-art machine-learning framework that revolutionizes the drug discovery process by prioritizing compounds based on their pharmacological impacts. This cutting-edge approach expedites the development of safe and effective medications, ultimately benefiting patients and healthcare providers alike.

Furthermore, the emergence of computer-aided drug design as a distinct field has propelled progress in this area to unprecedented levels. By leveraging advanced computational tools, researchers can now design drugs with unparalleled efficiency and precision, further streamlining the drug development process and reducing costs.As the field of bioinformatics continues to expand, it has led to the creation of invaluable databases containing nucleic sequences, protein patterns, and structures. Our team remains dedicated to advancing this field through innovative solutions, recognizing the immense potential for the discovery of new medications and bioactive substances. By staying at the forefront of research and development, we aim to address the pressing healthcare challenges of our time and improve patient outcomes.

The availability of sparse biomedical data has made it an invaluable resource for researchers seeking publicly accessible and computationally applicable information. However, extracting this data using computers can be a complex and error-prone task [2]. Identifying and analyzing drug actions require extensive research to pinpoint their specific targets, and with the exponential growth of potential drug process sources, the need for accurate and efficient data analysis has become even more critical. Fortunately, the emergence of bioinformatics has revolutionized the field by enabling scientists to unravel and organize the intricate nucleotide structures of genes, providing a deeper understanding of the data and the ability to identify potential drug targets with greater precision and efficacy. This breakthrough has significantly advanced drug discovery and development, offering new hope for the treatment of various diseases. The identification of target proteins and exploration of novel drug development pathways are critical pursuits that require a rigorous analysis of human genome data. Advancing the drug discovery process is a top priority for drug designers, who are constantly striving to incorporate diverse and potent qualities, making extensive research an essential prerequisite. Fortunately, bioinformatics is an effective tool that empowers drug developers to locate and scrutinize natural drug targets, thereby expanding the scope of potential drugs in the pharmaceutical pipeline. This presents a unique opportunity to prioritize these initiatives and pave the way for revolutionary drug development, leading to the discovery of groundbreaking therapies that can transform human health. The intricate nature of Proteomics and Genomics presents significant challenges to the development of new drugs. These fields offer immense potential for revolutionizing drug discovery, particularly through the exploration of complex human gene mechanisms. However, traditional drug discovery methodologies have heavily relied on computational science and cheminformatics, which may not fully capture the complexity of these fields. Thankfully, the Protein Data Bank has been a valuable resource for the past three decades, providing researchers with access to a vast collection of molecular structure data, spanning proteins, DNA, and RNA. To ensure the success of future drug development, it is crucial for researchers to recognize the profound impact of Proteomics and Genomics and integrate them into their drug discovery efforts. By doing so, they can unlock new insights and pave the way for more effective medications. In recent times, the production of sequence data has experienced an unprecedented surge, marked by an exponential increase in volume. However, this growth is accompanied by a marked escalation in complexity, leading to new and formidable challenges for data analysis. The advent of the Internet of Things (IoT) has also triggered a surge in medical data, with information collected from the human body being stored in the cloud. This has prompted pharmaceutical companies to intensify their efforts to discover new drugs, leveraging the vast data available to them. To make sense of this data, state-of-the-art machine learning and deep learning algorithms are employed to process and analyze gene patterns, enabling the identification and diagnosis of various human conditions. The extraction of relevant features from this data is meticulously carried out to further refine it, unlocking new insights and possibilities in the process. The effective utilization of this data holds tremendous potential for the advancement of biomedical research and the development of new therapies.A meticulous examination of the intricate gene patterns within the structure of genes enables clinical pharmacologists to create groundbreaking drugs through DNA sequencing. This cutting-edge approach to disease detection holds immense potential for significantly reducing disease incidence [5]. The proposed system serves as a highly perceptive instrument, offering insightful medication recommendations to both patients and their attentive caregivers. The integration of this innovative method into clinical practice has the potential to revolutionize the field of pharmacology and improve patient outcomes.

2.Background

The synergistic integration of machine learning and bioinformatics has revolutionized our understanding of the intricate mechanisms underlying cellular processes. Within the field of biomedical research, bioinformatics has emerged as a critical tool, empowering clinicians to meticulously scrutinize and evaluate clinical data obtained from their experiments. The systematic collection and analysis of data for future utilization, commonly referred to as "Clinical Informatics", is an essential component of this field, enabling researchers to extract meaningful insights and drive scientific progress. In the drug development process, identifying and validating a target is a crucial initial step, followed by synthesizing a compound that can seamlessly interact with it. The rich history of neural networks and machine learning in bioinformatics is noteworthy, as they have played a pivotal role in advancing our understanding of cellular mechanisms, paving the way for innovative therapeutic interventions.

The unrelenting pursuit of discovering new drugs, developing innovative drug compounds, and unraveling the complexities of gene interactions [8] has led to the replacement of established methodologies such as Naive Bayesian Classification and SVMs with more sophisticated techniques. This transition was motivated by the realization that these conventional approaches are prone to over or under-training and are often criticized for producing output without a clear comprehension of the underlying methodology [9]. However, recent advancements in neural systems have exhibited the potential to overcome some of these limitations and provide valuable insights for identifying potential drug candidates. Nevertheless, the application of deep neural network models requires a scrupulous examination and a comprehensive definition of their respective domains of applicability. The ultimate objective is for these models to furnish practical guidelines for drug discovery and gene analysis.

The development of confidence measures that are specifically tailored to neural system classifiers has been met with significant success by researchers, and these measures can be readily implemented [10]. As users continue to deepen their understanding of the intricate sub-atomic features that underlie pattern recognition and classification (Meissner et al., 2006), a comprehensive evaluation of deep neural network strategies is expected to emerge. To facilitate this process, there exists a wide range of atomic representations that effectively capture the diverse properties of chemical elements. These representations are then subjected to a rigorous scoring or measurement plan that effectively compares the encoded atoms with each other [11]. Ultimately, a sophisticated machine-learning algorithm is employed to identify the relevant features that can be effectively leveraged to subjectively or quantitatively differentiate the molecular structures and their properties. This approach holds significant promise for advancing our understanding of neural network classifiers and enhancing their practical applications.

The emergence of advanced technologies, such as machine learning and deep learning, has reinvigorated previously sluggish composites, endowing them with a dynamic and potent character [12]. These advancements have equipped the sub-atomic box with a multitude of potentially transformative strategies [13]. Innovative research has demonstrated that sophisticated deep learning designs will be pivotal for comprehensive data analysis in pharmaceutical research, toxicity prediction, genome mining, and chemogenomic applications in the future [14]. These capabilities have the potential to revolutionize personalized healthcare, but it is crucial to have a thorough understanding of the advantages and limitations of deep learning techniques. It is not advisable to solely rely on these methods without incorporating linear techniques to achieve the most precise and reliable results. Therefore, it is essential to exercise caution and recognize the significance of integrating linear techniques with deep learning to attain optimal outcomes.

Our primary objective is to conduct a rigorous and comprehensive investigation, utilizing advanced research methodologies, to identify and meticulously scrutinize the commonalities that exist among a diverse array of chemical techniques. Through this research, we aim to contribute to the advancement of scientific knowledge and enhance our understanding of chemical processes.

3.Proposed Model (IPSO-MSVM)

In the proposed approach, biomedical documents, gene disease database and chemical drug names are taken as input for deep learning biomedical document processing (Figure 1). 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 graph-based similarity model is used to find the essential gene to disease patterns for chemical drug mapping. A set of chemical drug symbols and gene names are used as training data for document feature extraction in CNN deep learning framework. CNN deep learning framework is used to extract the gene disease based chemical drug features in large biomedical document sets.

In the proposed model, the biomedical data collected from large databases, the data is preprocessed (Figure 1). The optimized PSO approach is used to optimize the parameters for classification of the data and gives as input to SVM classifier. The classified features are given to the CNN framework.

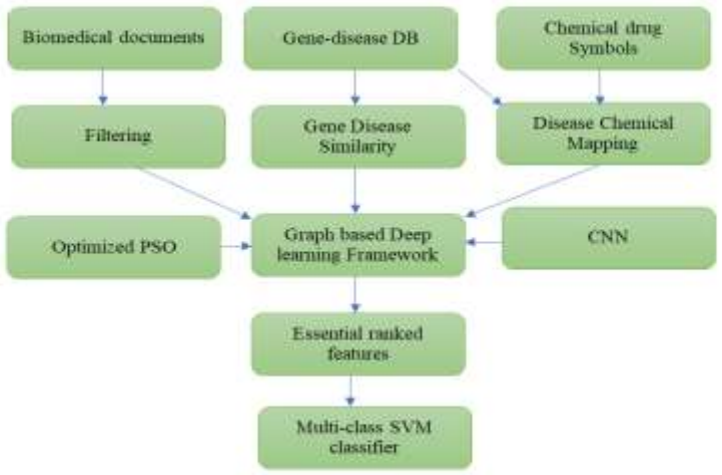

Figure1. Deep learning framework for Gene-Disease to Chemical drug classification

In the figure1, different biomedical document sets are taken as input for data filtering , gene disease ranking and disease to drug mapping process. Here, each gene and chemical drug is mapped with contextual similarity. Since, in the existing models, similarity is computed based on the gene and disease patterns. In this work, a hybrid contextual similarity between the gene terms and drug is used to find the key features in the CNN framework.

Phase 1: Data filtering:

In this phase, input data are filtered using the Stanford NLP library. Here, all the non-special noisy symbols , tokenization, stemming and stopword removal are performed on the input biomedical document sets.

Phase 2: Gene disease and Drug mapping Contextual similarity

```
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
Repeat to each token t in document sets D
Do
        Gt[]={RemoveStopWords(g[i]),RemoveNonspecialChars[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, stopword 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 in the phase 2.

Context Similarity of genes and diseases.

p represents the number of genes of the document M[i] that contain disease di

q denotes the number of genes of the document MD[i] that do not contain disease di

r represents the number of genes that do not belong to the document MD[i] but contain disease di

and s is the number of genes that do not belong to the document M[i] and do not contain disease di

$$Sim(Gt_i, BDt_j) = \frac{\Pr ob(\mathrm{CD[j]}/(\mathrm{p} \cap s)) \times \left(p_{ij}s_{ij} - q_{ij}r_{ij}\right)^2}{\left(p_{ij} + q_{ij}\right) \times \left(p_{ij} + r_{ij}\right) \times \left(q_{ij} + s_{ij}\right) \times \left(r_{ij} + s_{ij}\right)}$$

Context Similarity of genes and drug.

Let r1 represents the number of gene sets in the document set D.

Let r2 represents the number of drug entities in the document set D.

The contextual similarity between the genes and drug is computed by using the following formula

$$\mathrm{Sim}(\mathrm{t}[], \mathrm{r1}[], \mathrm{r2}[]) = \Pr(\mathrm{t}/\mathrm{r1}).\Pr(\mathrm{t}/\mathrm{r2}) / \Pr(\mathrm{t} / (\mathrm{r1} \cap \mathrm{r2}))$$

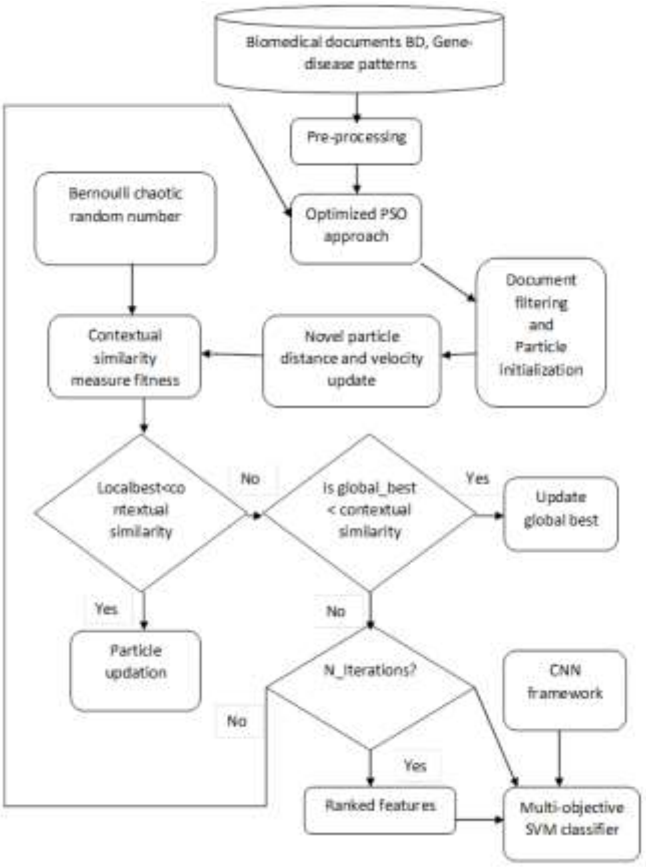
Biomedical documents BD, Gene-disease patterns
Pre-processing
Bernoulli chaotic random number
Optimized PSO approach
Document filtering and Particle initialization
Contextual similarity measure fitness
Novel particle distance and velocity update
Localbest<contextual similarity
No
Is global_best < contextual similarity
Yes
Update global best
Yes
No
Particle updation
N_Iterations?
CNN framework
No
Yes
Ranked features
Multi-objective SVM classifier

Figure 2. Propose Model for identification genes

Phase 3: Proposed PSO particle initialization and model update.

Phase 3, describes the improved particle swarm optimization algorithm for feature extraction process in biomedical document sets. In the figure 2, all the particles are initialized using the biomedical document tokens and gene tokens. After the initialization the particles, each particle position and velocity is updated using the following equations.

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1}$$

$$v_{id}^{k+1} = \xi_{\mathrm{p}}(wv_{id}^{k}) + c_1 \times \phi_{\mathrm{m1}} \times \left(p_{id} - x_{id}^{k}\right) + c_2 \times \phi_{\mathrm{m2}} \times \left(p_{gd} - x_{id}^{k}\right)$$

ξ_{p} *is the convergent factor formulated as*

$$\xi_{\mathrm{p}} = \frac{1}{|\,2\pi - \left(\tau_{\mathrm{m1}} + \tau_{\mathrm{m2}}\right) - \sqrt{\left(\tau_{\mathrm{m1}} + \tau_{\mathrm{m2}}\right)^2 - 4\left(\tau_{\mathrm{m1}} + \tau_{\mathrm{m2}}\right)}\,|}$$

where $\tau_{\mathrm{m1}}, \tau_{\mathrm{m2}} \in \psi$.

$$\psi_{n+1} = \begin{cases} \dfrac{\psi_n}{2} & \psi_n < 0.5 \\ \dfrac{1}{3}\psi_n\left(1-\psi_n\right), & \text{otherwise} \end{cases} \qquad \psi_n \in (0,1)$$

Proposed activation functionfor deep network CNN layer is computed as

$$Fitness_i = \tau_{m1}.\mathrm{MSVM}_i + \tau_{m2} \cdot \left(1 - \frac{\sum_{i=1}^{|F|} |\mathrm{sf}_i|}{N_f} \right) where\ \tau_{m1}, \tau_{m2} \in \psi$$

$|sf_i|$ represents the selected features count

MSVM_i represents the classification accuracy of the multi-objective SVM classifier

Phase 4: Graph based CNN Deep learning framework

Initially, document dependency graph DDG →(V, E) is constructed using the edge set and vertex set as E and V. Here vertex set V is represented asgene, disease and chemical drug vector sets and edge set E is represented as weighted rank between the disease gene sets and chemical drug symbols.

The weighted rank between the chemical drug symbols and disease gene patterns are computed in following equation.

$$Edge\_weight : \mathrm{w}(\mathrm{t_i}, \mathrm{t_j}, t_k) = \sum_{\mathrm{k,j}} \frac{\Pr ob\left(\mathrm{t_i}, \mathrm{t_j}\right)}{\Pr ob\left(\mathrm{t_i}\right) + \Pr ob\left(\mathrm{t_j}\right) + \Pr ob\left((\mathrm{t_i}, \mathrm{t_j}) / \mathrm{t}_k\right)}$$

$where\ \mathrm{t_i}, \mathrm{t_j}, t_k \in v_i, v_j; t_i : gene$ terms, $\mathrm{t}_j : disease$ terms ,$\mathrm{t}_k : chemical$ drug names

$\Pr ob\left(\mathrm{t_i}, \mathrm{t_j}\right)$: Probability of both terms occur together

$\Pr ob\left((\mathrm{t_i}, \mathrm{t_j}) / \mathrm{t}_k\right) : Conditional$ probability of $(\mathrm{t_i}, \mathrm{t_j}) given\ \mathrm{t}_k$.

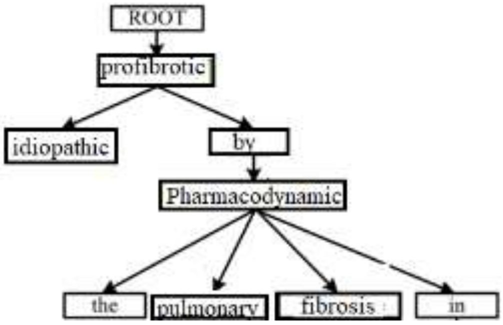

Figure 3. Sample graph based disease pattern extraction

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

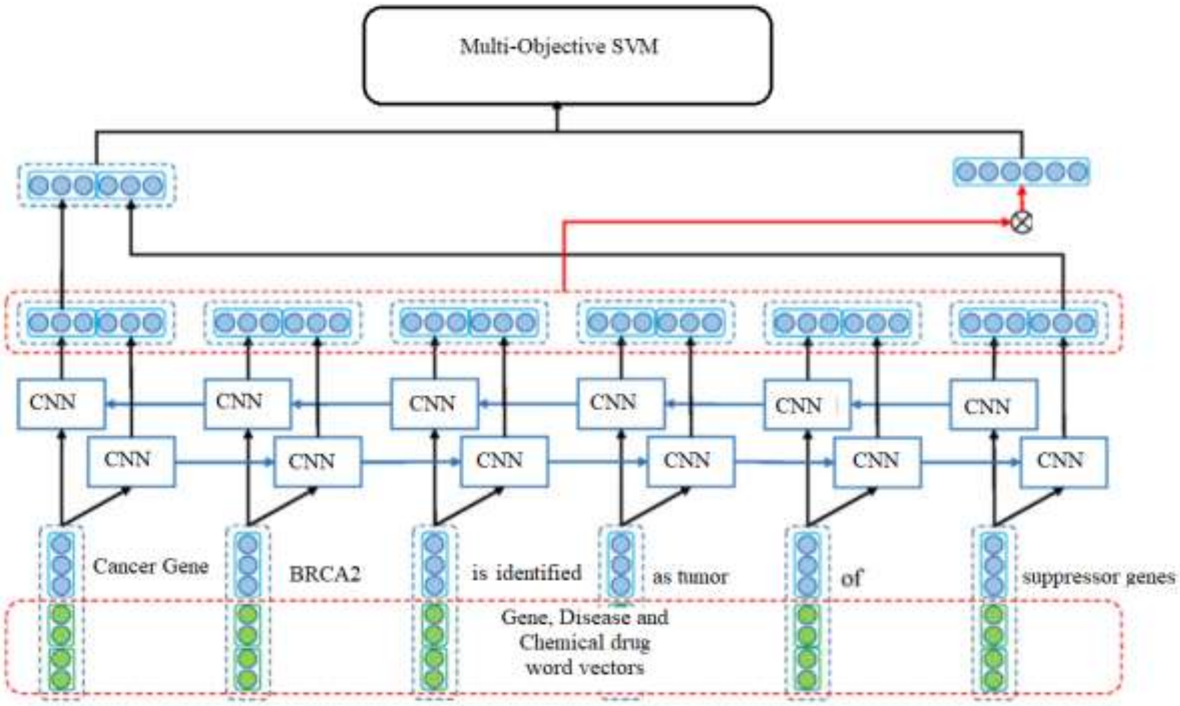

Figure 4. Proposed Graph based CNN framework for disease2drug prediction

Different filters with varying window sizes (here, only the height is different) slide over the full E rows in the convolution layer, i.e. the filter width is usually the same as the E width. Each filter conducts E-convolution, generating various function maps. A max-over-time pooling operation is applied to the elements located in the same feature map to extract the most important feature. Max-over-time pooling operation is performed on the PSO features to find the essential key relationships among the gene disease based chemical drug terms as shown in Figure 4. In the proposed CNN framework, a multi-objective SVM is proposed to the classify the input tokens for chemical drug prediction based on the gene disease patterns.

Input the CNN features for data classification.

```
For each feature set do
for each disease in GDP.
    do
        Apply SVM multi-class optimization models as
```

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

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

$$\xi_i^n>0$$

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

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

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

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

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

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

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

Experimental setup and results

The proposed model is simulated in Amazon AWS cloud server with 48GB of RAM. In the AWS server, different deep learning frameworks are selected to find the best processing speed and accuracy on large datasets. In the proposed work, java based deep learning framework is implemented in the large AWS instance to filter the essential key patterns and classification models. In order to improve the classification speed, proposed model is simulated in Hadoop based deep learning framework on large features space. We evaluate the performance of our frameworks on different Gene dataset, Chemical drug dataset and biomedical documents. Here, different metrics such as accuracy , Precision, and Area under the Curve (AUC) are used to evaluate the performance of gene chemical prediction.

Table1: Comparative result of proposed gene,disease and chemical similarity based CNN and MSVM framework to the conventional approaches for accuracy measure.

TestSample	Chisquare+RandomForest	MI+CNN	PSOCNN	IPSO+CNN+MSVM
#1	0.88	0.92	0.93	0.96
#2	0.17	0.92	0.92	0.98
#3	0.28	0.91	0.91	0.98
#4	0.28	0.91	0.92	0.97
#5	0.44	0.92	0.9	0.96
#6	0.29	0.91	0.92	0.97
#7	0.58	0.91	0.91	0.97
#8	0.34	0.93	0.9	0.97
#9	0.7	0.9	0.92	0.97
#10	0.46	0.94	0.92	0.97
#11	0.83	0.94	0.92	0.97
#12	0.68	0.92	0.92	0.97
#13	0.27	0.91	0.9	0.98
#14	0.59	0.94	0.9	0.98
#15	0.88	0.92	0.92	0.97
#16	0.34	0.94	0.93	0.97
#17	0.21	0.91	0.91	0.98
#18	0.36	0.93	0.9	0.98
#19	0.41	0.93	0.92	0.97
#20	0.37	0.94	0.9	0.98
#21	0.88	0.94	0.91	0.97
#22	0.37	0.94	0.91	0.97

#23	0.38	0.93	0.92	0.97
#24	0.41	0.94	0.9	0.98
#25	0.66	0.92	0.91	0.98

Table 1, describes the accuracy comparison of proposed MSVM-CNN framework to the state-of-art approaches on different biomedical document sets. In this table, accuracy measure is compared by using the genes, drugs and its contextual similarities to classify the document sets with deep learning framework.

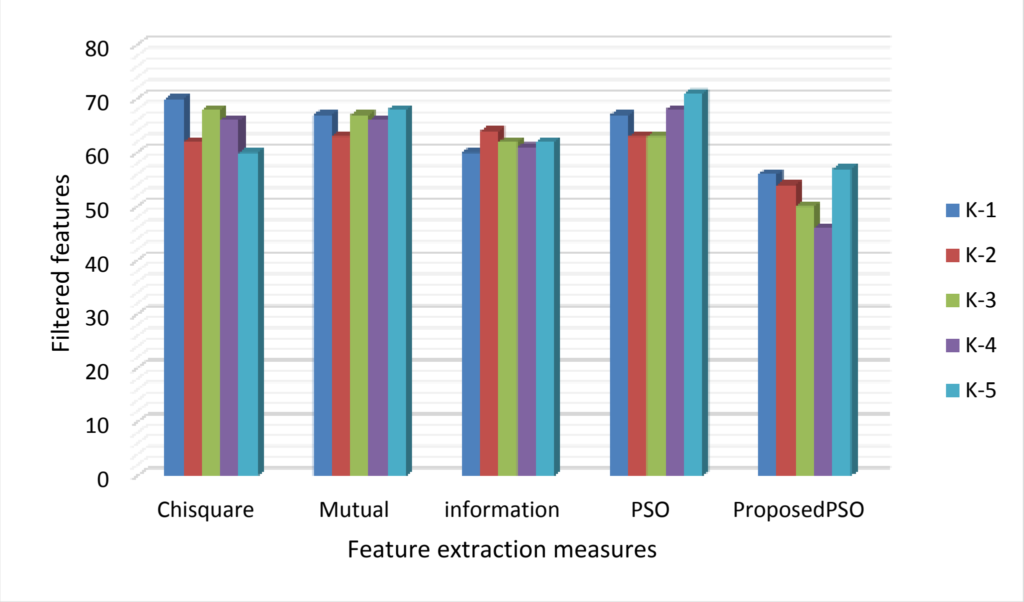

Figure 5: Comparison of proposed filter based approach to the conventional models on integrated database.

Figure 5, describes the feature extraction count of proposed approach to the state-of-art approaches on different biomedical document sets. In this figure , features count is evaluated using the k measure and genes, drugs and its contextual similarities .

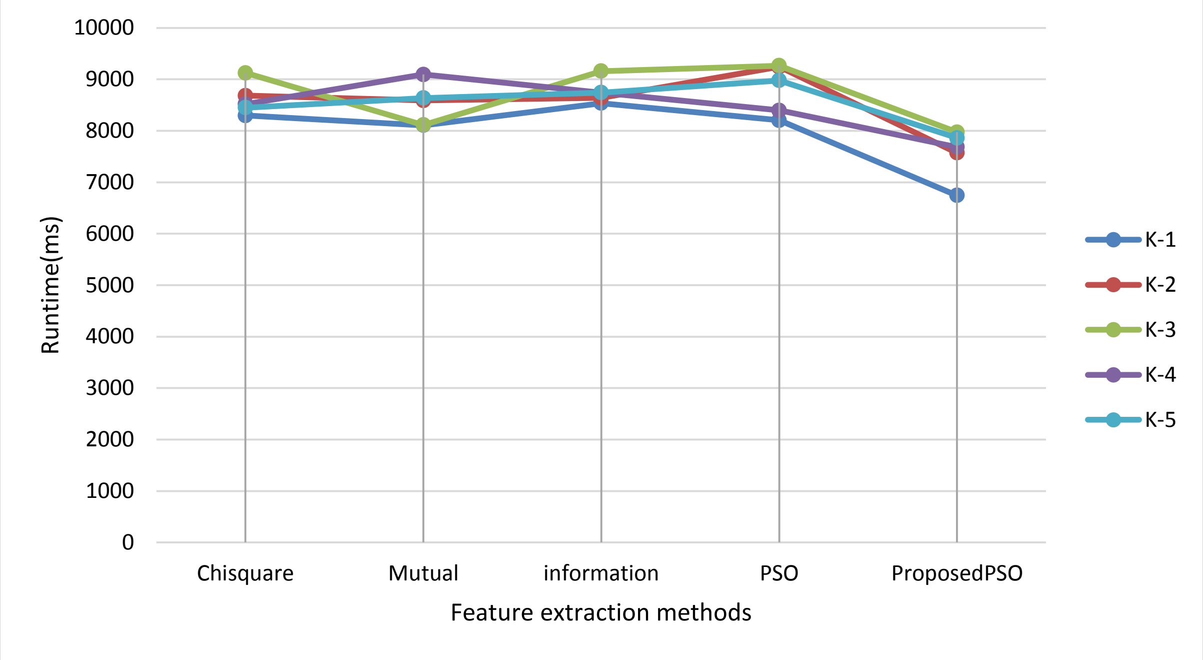

Figure 5: Comparative analysis of proposed contextual gene, chemical drug and disease features selection in CNN framework w.r.t runtime(ms).

Figure 5, describes the runtime of proposed approach to the state-of-art approaches on different biomedical document sets. In this figure , runtime count is evaluated using the k measure and genes, drugs and its contextual similarities .

Table 2: Comparative analysis of proposed contextual gene, chemical drug and disease classification accuracy using CNN framework on cross domain database.

Test	Chisquare+CNN	Mutualinformation+CNN	InformationGain+CNN	PSO+CNN	ProposedPSO+CNN+SVM
K-1	0.906	0.893	0.933	0.913	0.962
K-2	0.905	0.884	0.937	0.911	0.953
K-3	0.916	0.891	0.909	0.916	0.957
K-4	0.904	0.92	0.905	0.932	0.946
K-5	0.9	0.896	0.915	0.931	0.97

Table 2, illustrates the comparative analysis of proposed optimized pso approach and CNN framework with SVM classification model on cross domain integrated database . From the table, it is noted that the proposed ipso based CNN classifier has better gene, chemical and disease with drug contextual classification accuracy on integrated database than the conventional models.

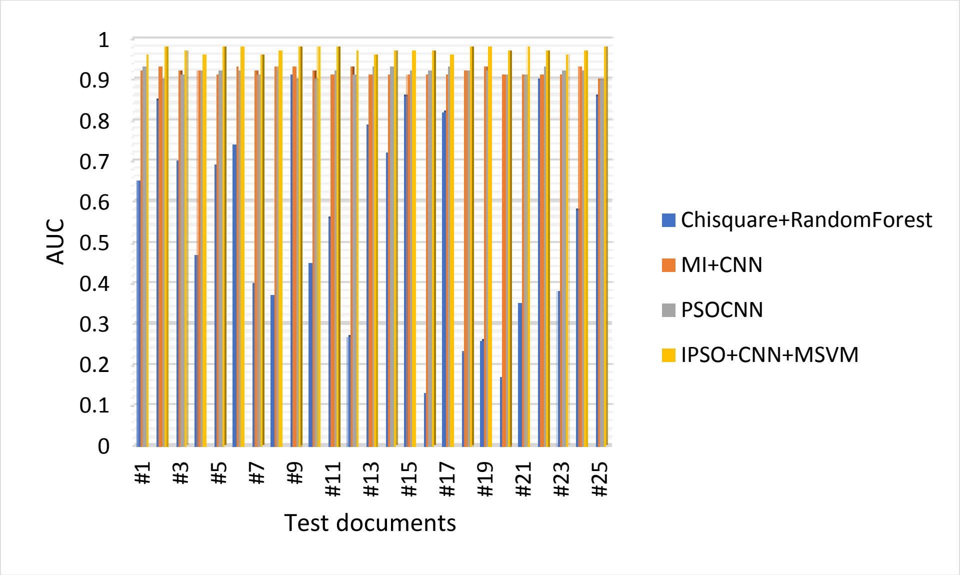

Figure 6: Comparative result of proposed gene,disease and chemical similarity based CNN and MSVM framework to the conventional approaches for AUC measure.

Table 1, describes the AUC comparison of proposed MSVM-CNN framework to the state-of-art approaches on different biomedical document sets. In this table, accuracy measure is compared by using the genes, drugs and its contextual similarities to classify the document sets with deep learning framework.

Table 3: Comparative result of proposed gene,disease and chemical similarity based CNN and MSVM framework to the conventional approaches for precision measure.

TestSample	Chisquare+RandomForest	MI+CNN	PSOCNN	IPSO+CNN+MSVM
#1	0.54	0.94	0.93	0.97
#2	0.58	0.95	0.93	0.97
#3	0.36	0.92	0.92	0.97
#4	0.26	0.92	0.92	0.97
#5	0.42	0.94	0.92	0.98
#6	0.56	0.93	0.91	0.97
#7	0.65	0.92	0.91	0.97
#8	0.76	0.9	0.9	0.97
#9	0.36	0.94	0.91	0.96
#10	0.49	0.9	0.92	0.96
#11	0.62	0.92	0.93	0.97
#12	0.39	0.92	0.92	0.98
#13	0.44	0.91	0.93	0.97

#14	0.47	0.92	0.93	0.97
#15	0.26	0.91	0.92	0.97
#16	0.61	0.93	0.93	0.97
#17	0.91	0.9	0.91	0.97
#18	0.74	0.93	0.93	0.97
#19	0.71	0.92	0.92	0.97
#20	0.77	0.93	0.92	0.97
#21	0.51	0.95	0.91	0.96
#22	0.43	0.92	0.92	0.97
#23	0.8	0.93	0.9	0.97
#24	0.51	0.93	0.92	0.97
#25	0.4	0.91	0.9	0.97

Table 3, describes the precision comparison of proposed MSVM-CNN framework to the state-of-art approaches on different biomedical document sets. In this table, accuracy measure is compared by using the genes, drugs and its contextual similarities to classify the document sets with deep learning framework.

5.Conclusion

The drug discovery plays crucial role due to the increase of diseases and population. The virus of the disease affects the antibodies of human body and leads to sickness and death of a person. So the advanced studies identified the drug for particular diseases based on the patterns of the genes. Thousands of organic chemicals combined and analysed for screening. After the evolution of bioinformatics, the combination of DNA driven drug discovery added new combinations from old medicines. The genetic engineering is used in the production of antibodies and proteins made immunogenic. The proposed model is used to identify the features from the medical datasets and analyse the pattern of the genes. This is very helpful to the scientists to discover the drug. The proposed method is identifying less numbers of relevant features with very low computational time than all other algorithms and having an accuracy of 98%, which is very high. The confusion matrix metrics are compared with the proposed algorithm to the traditional algorithms.

References

1. Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*. https://doi.org/10.15252/msb.20156651
2. Bikku, T., Nandam, S. R., & Akepogu, A. R. (2018). A contemporary feature selection and classification framework for imbalanced biomedical datasets. *Egyptian Informatics Journal*. https://doi.org/10.1016/j.eij.2018.03.003
3. Bikku, Thulasi, & Paturi, R. (2019). A novel somatic cancer gene-based biomedical document feature ranking and clustering model. *Informatics in Medicine Unlocked*. https://doi.org/10.1016/j.imu.2019.100188
4. Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with

deep learning. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2016.03.028

5. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. In *Nature Medicine*. https://doi.org/10.1038/s41591-018-0316-z
6. Faris, H., Mafarja, M. M., Heidari, A. A., Aljarah, I., Al-Zoubi, A. M., Mirjalili, S., & Fujita, H. (2018). An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2018.05.009
7. Gardner, C. R., Walsh, C. T., & Almarsson, Ö. (2004). Drugs as materials: Valuing physical form in drug discovery. In *Nature Reviews Drug Discovery*. https://doi.org/10.1038/nrd1550
8. Gong, M., Liu, J., Li, H., Cai, Q., & Su, L. (2015). A multiobjective sparse feature learning model for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2015.2469673
9. Kahn, M. G., & Weng, C. (2012). Clinical research informatics: A conceptual perspective. *Journal of the American Medical Informatics Association*. https://doi.org/10.1136/amiajnl-2012-000968
10. Kanwal, S. (2016). Towards a novel medical diagnosis system for clinical decision support system applications. In *PQDT - UK & Ireland*.
11. Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. https://doi.org/10.1038/89044
12. Li, N., Katz, S., Dutta, B., Benet, Z. L., Sun, J., & Fraser, I. D. C. (2017). Genome-wide siRNA screen of genes regulating the LPS-induced NF-κB and TNF-α responses in mouse macrophages. *Scientific Data*. https://doi.org/10.1038/sdata.2017.8
13. Meissner, M., Schmuker, M., & Schneider, G. (2006). Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-7-125
14. Scheeder, C., Heigwer, F., & Boutros, M. (2018). Machine learning and image-based profiling in drug discovery. In *Current Opinion in Systems Biology*. https://doi.org/10.1016/j.coisb.2018.05.004