ISSN: 2348-2079

Volume-11 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Enhanced Drug Recognition based on Decision tree optimized Support vector machine for disease related recommendation

[1]E.Loganathan, [*2]M.Prakash, [3]G.Sivakumar

[1]Assistant Professor, Department Of Computer Engineering, Erode Sengunthar Engineering College (Autonomous), Perundurai, Erode – 638 057.
[2]Master of Engineering in Computer Science And Engineering*
Erode Sengunthar Engineering College (Autonomous)
Perundurai, Erode – 638 057.
[2]Head of the Department, Department of Computer Science and Engineering, Erode Sengunthar, Engineering College (Autonomous), Perundurai, Erode-638057.

***Corresponding Author: M. Prakash**

ABSTRACT

Drug analysis is important for recommendation disease oriented treatment. In recent days drug combination is improperly suggest to patients without knowing the disease factor. Due to non-relation drug patterns the patients are inse4cure to affect side effects. In most researchers implementing machine learning concepts based on big data analytics using different algorithms like, decision tree, random forest, and logical process. But in recognition to identify the relation is failed because of improper feature analysis and classifier value results inaccuracy. To resolve this problem we propose an efficient feature selection based on optimized decision tree model and implemented with support vector machine to classify the drug relation to make effective recognition class for patients. Initially the preprocessing was carried to normalize the dataset to reduce the Nosie. Then drug margin impact rate is estimated to find the relational margins. Then feature selection was done by decision and tree and classification was carried out by SVM. The classifier produce higher result in precision, recall rate, f-measure with low false measure. This proposed system achieves high performance compared to other systems.

Keywords: Drug analysis, prediction, feature selection, Support vector machine, classification, machine learning.

1. INTRODUCTION

Health professionals treat patients based on their symptoms. Doctors do not analyze each patient's genes, lifestyle, or genetics. Most diseases share the same symptoms, making it difficult to categorize the patient's disease. For example, the symptoms of Covid-19 like fever, body pain, cough, fatigue are more common in other diseases like Typhoid and other seasonal fevers which is being produced by various bacteria and viruses. Similar to the disease identification, treating the person with right set of treatment and medicines also more important

Towards drug compound analysis, there are number of approaches available. The frequency based approaches select the drug compound according to the number of times it has been given to different person. Similarly, the popularity based approach would select the compound according to the popularity of drug which is being measured based on the number of medical practitioner selects the drug. Similarly, there are number of other schemes available towards the problem and suffer to achieve higher performance. The recommendation systems are designed to support the medical practitioner in the selection of drugs to treat the patients. However, there exist several methods in the analysis of drug compound; [1] they suffer to achieve higher performance. To resolve above problem the Machine Learning (ML) and Deep Learning (DL) techniques plays a vital role in determining disease based recommending the best compound drug

Drug data prediction and analysis, there are number of approaches available. The frequency based approaches select the drug compound according to the number of times it has been given to different person [2]. Similarly, the popularity based approach would select the compound according to the popularity of drug which is being measured based on the number of medical practitioner selects the drug. Similarly, there are number of other schemes available towards the problem and suffer to achieve higher performance. The recommendation systems are designed to support the medical

practitioner in the selection of drugs to treat the patients. However, there exist several methods in the analysis of drug compound; they suffer to achieve higher performance.

The dimensionality is another issue while performing drug feature analysis. The methods would miss set of features and dimensions at the analysis and by incorporating the neural network, the problem of drug relation analysis can be performed effectively.

Contribution of the research is to find the feature relation based on disease relation using efficient feature selection based on optimized decision tree model and implemented with support vector machine to classify the drug relation to make effective recognition class for patients. Initially the preprocessing was carried to normalize the dataset to reduce the Nosie. Then drug margin impact rate is estimated to find the relational margins. Then feature selection was done by decision and tree and classification was carried out by SVM. The classifier produce higher result in precision, recall rate, f-measure with low false measure. This proposed system achieves high performance compared to other systems.

2. Related work

A Bayesian based machine learning technique to identify the target drug using various data types. Machine learning model analyze the performance of various drugs and their performance is compared with other approaches [4].

The Self-Organizing Map (SOM) and Principal Component Analysis (PCA) methods analyze biochemical data generated by screening programs to discover new pharmaceutically active compounds in microbial extracts. Classification of organisms can be considered a problem of pattern recognition [5]. The organism under investigation belonged to the genus Streptomyces. It was previously classified into one of the characteristics by numerical classification using a probability recognition matrix.

Existing methods still have limitations and are difficult to address the problem of multi-label molecular pathway prediction. Exploring the relationship between molecular structure and metabolic pathways is crucial for studying drug metabolism and pharmacokinetics. Therefore, people usually want to know what metabolic pathways a new synthetic drug compound [6].

Use Logistic Regression (LR) to combine a Random Forest (RF) with a support vector machine (SVM). These basic learners (RF and SVM) are generally considered the best way to solve complex classification problems on large datasets [7]. Aureus Proteome has tested our best classification model. Staphylococcus aureus is a major Gram-positive bacterial pathogen responsible for infectious diseases acquired by hospitals and communities.

The drug discovery process is far from optimal. Only one in ten compounds selected for development has succeeded in clinical trials. Fuzzy analysis techniques were performed on the peptide spectrum data of rat livers exposed to different doses of drugs to signto improve the efficiency of the process significantly [8].

Machine learning algorithms facilitate drug type classification and reduce lab costs. Prediction of supervised machine learning algorithms and their cancer treatments. Feature selection is applied to select the most suitable and relevant features. This will reduce the accuracy of the forecast [9].

The machine learning process is an effective and convenient way to build a data model based on a known data instance, using the known data instance to predict unknown data. The problem of class imbalance must be overcome to ensure the success of this method [10]. Class imbalance means that a classification task can compromise the classification performance of a standard classifier if the number of data instances in each class varies significantly [11]

Dimensionality reduction of dataset representations for constructing Quantitative Structure-Activity Relationship Classification (QASR) models is an important research topic for model interpretability and computational cost efficiency of classification algorithms like Gradient Boosting (GB), Logistic Regression (LR), and AdaBoost (ADA) [12]. After evaluating the classification results by a particular metric, the best combination of frameworks is obtained [13]. The feature selection method is appropriate because the classification process requires only a few related features.

From the literature, there are number of problems identified and listed in this section. The existing approaches consider only limited number of features in classification as well as prediction [14]. The existing drug compound analysis models consider only the frequency and popularity of drugs towards recommendation which introduces poor performance. The methods do not consider the curing rate in the performance analysis and produces higher false ratio [15]. The methods do not consider the impact or influence of drugs in the curing of any disease. The methods do not perform feature selection in the sense to reduce missing problem. The methods do not generate recommendations based on success rate and produces poor performance.

3. Proposed implementation

The intent of drug compound relation analysis provides an efficient feature selection model based on optimized decision tree model. The support vector machine to classify the drug relation to make effective recognition class for patients. Initially the preprocessing was carried to normalize the dataset to reduce the Nosie. Then drug margin impact rate is estimated to find the relational margins. Then feature selection was done by decision tree and classification was carried out by SVM. The classifier produce higher result in precision, recall rate, f-measure with low false measure. This proposed system achieves high performance compared to other systems.

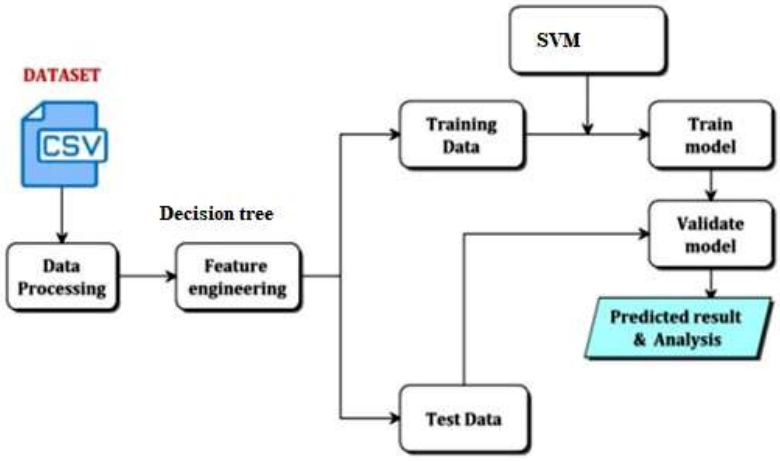

Fig 1: proposed architecture diagram (DT-SVM)

For any disease and the drug compounds identified, the method generates various patterns. Figure1 shows the proposed architecture diagram (DT-SVM). The drug recommendation is carried out through support vector machine. The testing and training validation was carried out by drug generative margins and classified with medical threshold pattern. The proposed system produce high performance based on the cumulative pattern generation depends on drug imitative usage for disease factor to recommend in high impact rate.

3.1 Data Pre-processing

The input collected drug compound data set given has number of missing features which are noisy to perform feature analysis. To remove the noise, the method reads the data set and identifies the molecule details and identifies the presence of data. If any of the log contains missing values of data, they are analyzed and then it will be removed from the data set. The noise removed data set has been used to perform clustering of related compound molecule preference under various classes of relative measures in the next stage.

Algorithm:
Input: Drug Data set Dwds
Output: Pre-processed data set Prds
Step 1: Start
Read Dwds.
For each Drug record Tc
Step 2: Verify the presence of compound molecules and values
Step 3: Check presence of data in all dimension in range of drug values
Step 4: compute the presence in relative weights
If $\int Tc \in DrugID \;\&\&\; Tc \in \boldsymbol{Molecule}\; ID \;\&\&\; Tc \in mean\; weig\square t$ then
Prds = $\sum(Tc \in Prds) \cup Tc$
End
End
Stop

The working principle of pre-processing algorithm is presented above which reads the Drug data set and identifies the presence of all the features like Drug power and relative molecule signs, drug data representation in each log. If any of them is missing, then it will not be added to the pre-processed data set.

3.2 Relational Drug pattern creation.

This stage create pattern for identifying the sequence of combination drug relation rated to set of support molecules depend son disease factor.

$$Hf(A) = \sum_{i}^{n} \begin{matrix} ni\ (Xi) == 1\ which\ is \\ \boldsymbol{equeivalent}\ another\ input\ series P(Ai) \log 2p(A2), \end{matrix}$$

And

$$Hf(B) = \sum_{j}^{m} \begin{matrix} mi(Xj) == 1\ which\ is \\ \boldsymbol{equeivalent}\ realtional\ input\ series P(Bi) \log 2p(B2) \end{matrix}$$

The joint features from sources A and B state are equivalent {A, B} = {$a_1b_1, a_1b_2, \cdots, a_nb_m$}, this creates joint distribution beyond the absolute weightage between the two states {A, B} is related to feature dependencies for directed disease support drugs.

3.3 Optimized feature selection Decision Tree (DT)

The feature selection is performed by measuring the Candidate Success Score (CSS) which is being measured for different pattern of drugs. The method first identifies the distinct profile users, which classify the features based on anatomic and health factors. For example, they are classified according to the chronic disease like diabetic, hypertension and so on. Now, for each class, the method estimates the CSS value for different drug pattern generated. The pattern with higher CSS value than the thresholds are selected and the features present in the patterns are selected. The selected patterns of drug and features are trained with different neurons at various layers.

In this sense, feature selection is clearer and less challenging in terms of interpretation. This property is centralized in a variety of valuable applications, such as finding the exact characteristics of a particular disease or developing an evaluative vocabulary to test hypotheses.

$$\frac{1}{2}w^Tw - v\rho + \frac{1}{N}\sum_{i=1}^{N}\xi_i$$

Subject to the constraints:

$$y_i(w^T\phi(x_i) + b) \geq \rho - \xi_i, \xi_i \geq 0, i = 1, \ldots\ldots\ldots \text{N and } \rho \geq 0$$

For this type of SVM the error function is:

$$\frac{1}{2}\mathrm{w^T w} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{N}}\xi_{\mathrm{i}} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{N}}\xi_{\mathrm{i}}$$

Which we minimize subject to:

$$\mathrm{w^T}\phi(\mathrm{x_i}) + \mathrm{b} - \mathrm{y_i} \leq \varepsilon + \xi_{\mathrm{i}}$$
$$\mathrm{y_i} - \mathrm{w^T}\phi(\mathrm{x_i}) - \mathrm{b_i} \leq \varepsilon + \xi_{\mathrm{i}}$$
$$\xi_{\mathrm{i}}, \xi_{\mathrm{i}} \geq 0, \mathrm{i} = 1, \ldots\ldots\ldots \mathrm{N}$$

Feature selection and extraction are usually displayed arbitrarily. Knowledge of the search algorithm uses the learning process to guarantee the extraction of features that reduce the dimensions of the data and improve the classification results.

DT comprises a set of 'rules' that provide the means to associate specific molecular features and/or descriptor values with the activity or property of interest. The DT approach has been applied to problems such as designing combinatorial libraries, predicting 'drug-likeness', predicting specific biological activities, and generating some specific compound profiling data. For splitting, the most popular criteria are "gini" for the Gini impurity and "entropy" for the information gain that can be expressed mathematically as,

$$Entropy\ E_p = -\sum g_i . log_2 . g_i$$

Gini $(G_I) = 1 - \sum g_i^2$

Where g_i denotes probability of class i

This method is used not only for the identification of substructures that discriminate activity from non-activity within a given compound database, but also for the classification of chemical compounds into drug and nondrug Instances are classified by checking the attribute defined by that node, starting at the root node of the tree, and then moving down the tree branch corresponding to the attribute value.

3.4 Classifier Support Vector Machine (SVM)

In machine learning, another common technique that can be used for classification, regression, or other tasks is a support vector machine (SVM). In high or infinite dimensional space, a support vector machine constructs a hyper-plane or set of hyper-planes. Intuitively, the hyper-plane, which has the greatest distance from the nearest training data points in any class, achieves a strong separation since, in general, the greater the margin, the lower the classifier's generalization error. It is effective in high-dimensional spaces and can behave differently based on different mathematical functions known as the kernel. Linear, polynomial, radial basis function (RBF), sigmoid, etc., are the popular kernel functions used in SVM classifier.

The purpose of SVM is to improve the generalization capability by extending classification gaps through a discriminant function. The hyperplane is shown in Figure 4. For all linear classification problem, let the training sample be $\{u_i, v_i\}, (i = 1,2,\ldots,m)$.

The mathematical expression of the optimal hyperplane is given in equation .

$$f(u) = \omega . \varphi(u) + a$$

Where;

$a \rightarrow$ Threshold value

$\omega \rightarrow$ Weight factor

The linear decision function based SVM classification can be written as follows;

$$f(u)=\text{sgn}\left(\sum_{i=1}^{m} v_i .b_i .r(u_i,u)+a\right)$$

Where;

b_i → Lagrange multiplier

a → Threshold value

u_i, v_i → Support vectors among any two classes

$r(u_i,u)$ → Kernel function

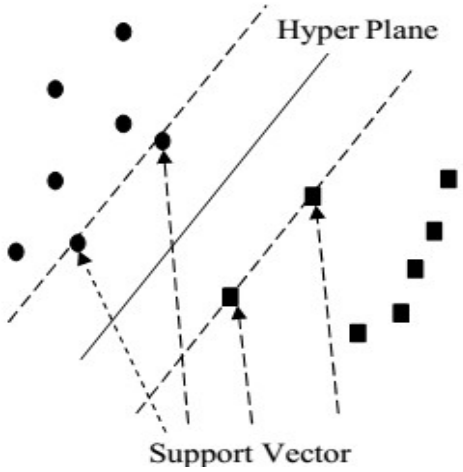

Fig 2: Support Vector Machine

Then, the error function between target output and classified output is defined using equation.

$$E = \varepsilon_{\max} + \varepsilon_{\min} + \frac{1}{N}\sum_{i=1}^{n} \left|O_{Class} - O_{Tar}\right|$$

Where, N represents the quantity of features, O_{Class} and O_{Tar} denote the classified output and target output respectively. $\varepsilon_{\min}$ and $\varepsilon_{\max}$ denote the minimum and maximum eigenvalue for weight vector and ε can be defined using equation .

$$\varepsilon = Eigen\left(W_{ij} * W_{ij}^{T}\right)$$

Besides,

$$\varepsilon_{\max} = \max(\varepsilon); \ \varepsilon_{\min} = \min(\varepsilon)$$

After the completion of training process, the SVM is used for classification purpose. The structure of the trained model is utilized for the testing or detection process.

4. RESULT AND DISCUSSION

The proposed drug component analysis and recommendation generation models are implemented and evaluated for their performance. The performance evaluation is carried in Python and the methods are evaluated on various metrics. The Drug Indications Database is collected from an online UCI repository. It contains 64 attributes: drug id, drug name, compound molecules (Chemical reaction) level, chemical id, and so on.

Table 1: Parameters settings

Parameters items	Values
Language	Python
Tool	Anaconda
Name of the dataset	Healthcare Drug Indications Database
Number of records	500000

The above table 1 shows the parameters settings for implantation of the proposed DT-SVM and existing methods carried out in Jupiter notebook in an anaconda environment. The total data is split into 70% training data and 30% test data. The confusion

matrix is used to calculate all the parameters such as precision, recall, false rate, and prediction performance.

Table 2: Analysis of Drug Success Rate Prediction performance

Drug Success Rate Prediction Performance in %			
Methods	100000 Records	200000 Records	500000 Records
LR	72	75	78
RF	76	80	83
FUZZY	80	83	87
DT-SVM	85	87	93

Above table 2 presents the comparison result of success rate prediction in %. The proposed algorithm has high performance compared with the previous algorithm.

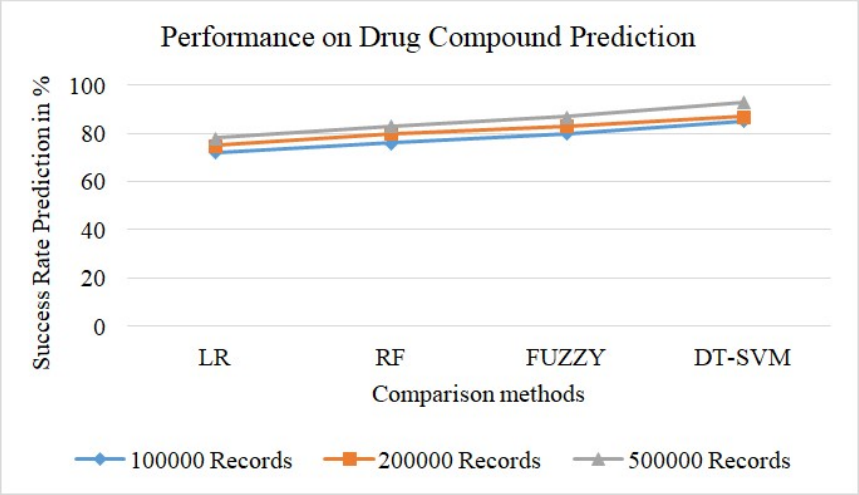

Fig 3: Analysis of Drug success rate prediction

Figure 3 presents an analysis of drug success rate prediction performance in percentage. The x-axis shows comparison methods, and the y-axis presents prediction performance. The proposed Deep Spectral Neural Classification (DT-SVM) algorithm obtained prediction result is 93% for 500000 Records.

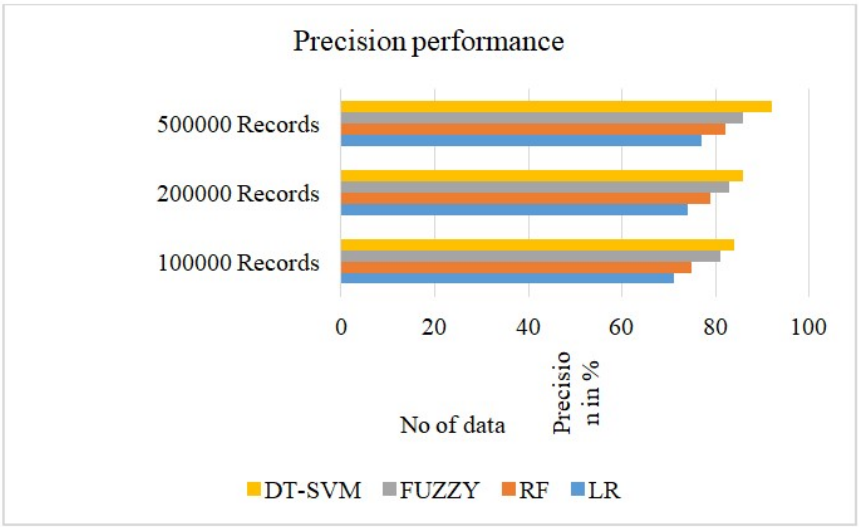

Fig 4: Analysis of Precision performance

Figure 4 illustrates the analysis of precision performance the recommended and existing results comparison. The x-axis shows the number of data, and the y-axis shows precision performance in percentage.

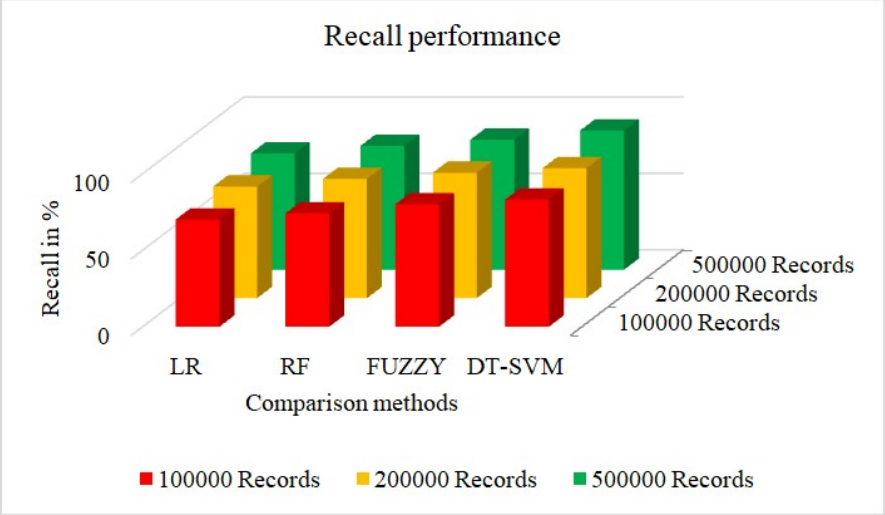

Fig 5: Analysis of Recall performance

Analysis of recall performance the proposed and existing algorithm comparison results are presented in figure 5. The proposed DT-SVM algorithm recall result has 91%, likewise, the existing algorithm results FUZZY are algorithm has 76%, RF algorithm has 81%, and PrOCTOR has 85%.

Table 3: Analysis of false rate performance

The false rate in %			
Methods	100000 Records	200000 Records	500000 Records
LR	28	25	22
RF	24	20	17
FUZZY	20	17	13
DT-SVM	15	13	7

Analysis of false rate performance comparison result present in table 3. The proposed algorithm provides low false classification performance.

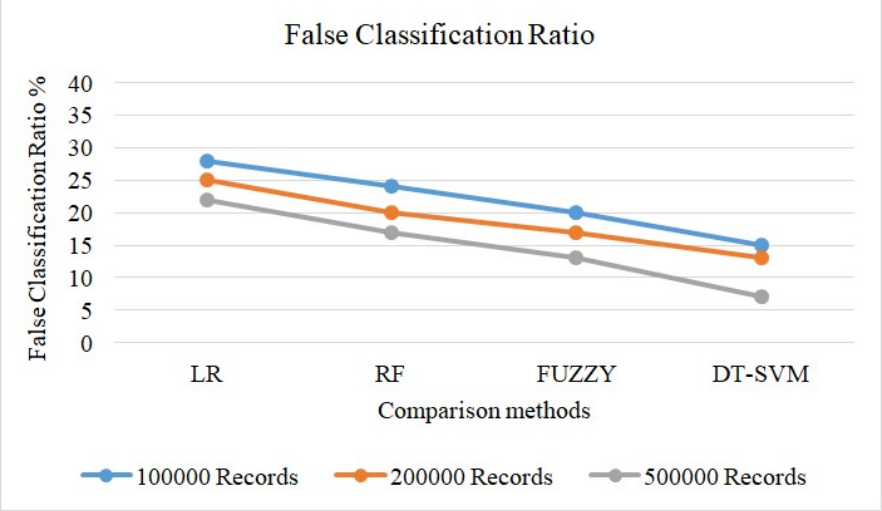

Fig 5: Analysis of false classification ratio performance

Analysis of false rate classification performance results is shown in figure 5. The proposed Deep Spectral-DT-SVM false rate performance is 7%.

Table 4: Analysis of time complexity performance

Time Complexity performance in sec			
Methods	**100000 Records**	**200000 Records**	**500000 Records**
LR	57	77	91
RF	44	65	79
FUZZY	36	54	70
DT-SVM	20	25	31

Above, table 4 presents the time complexity for drug success rate prediction performance. The proposed algorithm provides low time compared with existing algorithms.

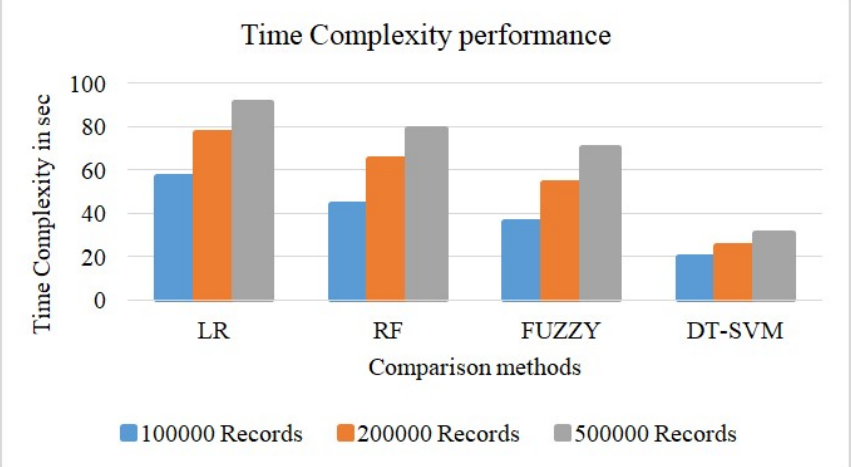

Fig 6: Analysis of Time Complexity performance

Analysis of time complexity performance results is shown in figure 6. The proposed Deep Spectral Neural Classification (DT-SVM) drug success rate prediction time complexity result is 31sec for 500000 Records.

5. CONCLUSION

The conclusion of the proposed method produce best performance. The proposed method identifies the set of drugs provided for different diseases. The method generates various patterns for any disease and the drug compounds identified. Then drug margin impact rate is estimated to find the relational margins. Then feature selection was done by decision and tree and classification was carried out by SVM. The classifier produce higher result in precision, recall rate, f-measure with low false measure. This proposed system achieves high performance compared to other systems. In the future, drug analysis, compound prediction, and recommendation performance can be improved by incorporating genetic and lifestyle features in measuring the success rate of different drugs

REFERENCES

1. Saad E, Din S, Jamil R, Rustam F, Mehmood A, Ashraf I, et al. Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums. IEEE Access. 2021;9:85721-37. doi: 10.1109/ACCESS.2021.3088838.
2. Zhang C, et al. Assignment optimization of pandemic influenza antiviral drugs in Urban pharmacies. J Ambient Intell Humanit Comput. 2019;10:3067-74. https :. doi: 10.1007/s1265 2-018-0872-6.
3. Nyabadza F, Coetzee L. A systems dynamic model for drug abuse and drug-related crime in the Western Cape Province of South Africa. Comput Math Methods Med. 2017;2017:4074197. doi: 10.1155/2017/4074197, PMID 28555161.
4. Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. Cell Chem Biol. 2016;23(10):1294-301. doi: 10.1016/j.chembiol.2016.07.023, PMID 27642066.
5. Madhukar NS, Khade PK, Huang L, Gayvert K, Galletti G, Stogniew M et al. A Bayesian machine learning approach for drug target identification using diverse data types. Nat Commun. 2019;10(1):5221. doi: 10.1038/s41467-019-12928-6, PMID 31745082.
6. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019;11(1):4. doi: 10.1186/s13321-018-0325-4, PMID 30631996.
7. Saad E, Din S, Jamil R, Rustam F, Mehmood A, Ashraf I et al. Determining the efficiency of drugs under special conditions from users. Rev Healthc Web Forums, IEEE Access. 2021;9:85721-37.

8. Lo AW, Siah KW, Wong CH. Machine learning with statistical imputation for predicting drug approvals. Harv Data Sci Rev. 2019;1(1):1-42.
9. Mittapally R (2023). Intelligent Framework Selection: Leveraging MCDM in Web Technology Decisions. J Comp Sci Appl Inform Technol. 8(2): 1-8.
10. Nyabadza F, Coetzee L. A systems dynamic model for drug abuse and drug-related crime in the Western Cape Province of South Africa. Comput Math Methods Med. 2017;17:1-13.
11. Hu Qiwan, Feng Mudong, Lai L, Pei J. Prediction of drug-likeness using deep autoencoder neural networks. Front Genet. 2018;9:585. doi: 10.3389/fgene.2018.00585, PMID 30538725.
12. Pu Y, Li J, Tang J, Guo F. DeepFusionDTA: drug-target binding affinity prediction with information fusion and hybrid deep-learning ensemble model. IEEE Trans Comp Biol Bioinformatics. 2021;2021:1-.
13. Zhimiao, Yu, Jiarui, Lu, Yuan, Jin et al. KenDTI: an ensemble model for predicting drug-target interaction by integrating multi-source information. IEEE Trans Comp Biol Bioinformatics. 2021;18(4):1305-14.
14. Hu Q, Feng M, Lai L, Pei J. Prediction of drug-likeness using deep autoencoder neural networks. Front Genet. 2018;9:585. doi: 10.3389/fgene.2018.00585, PMID 30538725.
15. Song Q. Emergency drug procurement plFuzzying based on big-data driven morbidity prediction. IEEE Trans Ind Inform https :. 2019;79. doi: 10.1109/tii.2018.28708.
16. Zhimiao Yu KD. An ensemble model for predicting drug-target interaction by integrating multi-source information, IEEE. Trans Comp Biol Bioinformatics. 2021;18(4):1305-14.