

MINING SERENDIPITY FROM DRUG REVIEWS USING SUPPORT VECTOR MACHINES

¹K. Swathi, ^{*2}V. V. S. Srikar, ³K. Shrushti, ⁴G. Radhanjali

ABSTRACT--Serendipity is all about the Positivity or Goodness. Sentiment analysis is that the process of determining whether an editorial is positive, negative or neutral. Analysing opinions allows data analysts on large companies to gauge vox populi, conduct nuanced research, monitor brand and merchandise reputation and understand customer experience. The proposed model is trained and evaluated with a UCI ML dataset that describes drug use. If the patient's drug use effect might be computationally identified, it could help generate and validate drug repositioning hypotheses; Measures are often taken for that drug, both in production and in use. We investigate supervised machine learning models to extract positivity in drug use. Furthermore, the proposed model is compared with Naive Bayes. Finally, a machine learning model using Natural Language Processing techniques on drug usage review dataset is implemented to find the sentiment of drugs and results were presented.

Keywords--Machine Learning, Sentiment Analysis, Serendipity

I. INTRODUCTION

In the past decade, invasive social media websites have reached a crucial mass of patient discussions regarding diseases and medicines primarily within the type of unstructured, casual human language. This knowledge covers numerous medication outcomes like effectiveness, adverse effects thanks to medication, adherence, and cost. Such info can be useful for generating and confirming drug-repositioning hypotheses if these statements can be computationally developed victimization sentiment analysis and totally different classification models. Typically there are two main approaches for sentiment analysis: a machine learning approach (or an applied math text mining approach) and a linguistic approach (or a linguistic communication process approach). Since clauses are quite short and don't contain several subjective words, the machine learning approach typically suffer from knowledge sparsity downside. Conjointly the machine learning approach cannot handle advanced grammatical relations between words in a very clause.

Classifiers supervised learning is made to form prediction, given AN unforeseen input instance. A supervised learning algorithmic program takes a identified set of input dataset and its identified responses to the information (output) to be told the regression/classification model. A learning algorithmic program then trains a model to come

¹ Professor, NRI Institute of Technology

² Scholars, NRI Institute of Technology, srikarvvs@gmail.com

³ Scholars, NRI Institute of Technology

⁴ Scholars, NRI Institute of Technology

up with a prediction for the response to new knowledge or the take a look at dataset. Supervised learning uses classification algorithms and regression techniques to develop prognosticative models.

Unlike the computational pipeline that we used in the current study and our previous research, this application takes only drug-review comments and the drug review as inputs, making it adaptive to broader sources of patient-generated health data.

The proposed system gathers data from WebMD site in the form of huge data set. It uses all the steps for using supervised machine learning algorithms. The final outcome of this work is to compute three classes or polarities (positive, negative and neutral) from sentiment analysis done on drug review posts.

The proposed system is going to be developed with supervised learning with usage of well-known classifiers i.e., SVM (Support Vector Machine) and Naïve Bayes Machine Learning Algorithms. The review text always gives Sentiment of reviewer. Here it gives sentiment of reviewer. The review text sentiment analysis is done with NLP techniques. NLP and machine-learning methods are integrated into an automated workflow. Serendipity provides a user interface for scientists working in drug discovery and development who have limited programming experience, and it also has an application programming interface (API) for software developers who want to integrate Serendipity into other programs. Machine learning learns from historical data. It gives more accuracy.

NAIVE BAYES'S: Naive Bayes classifier is a kind of Classification technique that supported Bayes' Theorem with Associate in nursing assumption of independence among predictors. In straightforward terms, a Naive Bayesian classifier assumes that the presence of a selected feature in a very class is unrelated to the presence of the other operate. Naive Bayes model is accessible to make and significantly helpful for in depth datasets. Multinomial Naive Bayes classification algorithmic program tends to be a baseline resolution for sentiment analysis task. The fundamental plan of this technique is to search out possibilities the possibilities the chances of categories assigned to texts by victimization the joint probabilities of words and categories.

LINEAR SVM: A Support Vector Machine may be a kind of Classifier, during which a discriminative classifier formally outlined by a separating hyperplane. This classifier works making an attempt to make a line that divides the dataset departure the larger margin as attainable between points known as support vectors. The algorithmic program outputs associate in nursing best hyperplane that categorises new examples. In 2 dimensional area, this hyperplane may be a line dividing a plane into 2 components whereby every category lay on either facet.

Natural Language processing using Text Blob: Natural Language process (NLP) refers to AI technique of human action with associate degree intelligent systems employing a language like English. Processing of language is needed once you need associate degree intelligent system like mechanism to perform as per your directions, once you need to listen to call from a dialogue primarily based clinical professional system, etc. The field of IP involves creating computers to perform helpful tasks with the natural languages humans use. The input associate degree output of an IP system is often Speech or written language.

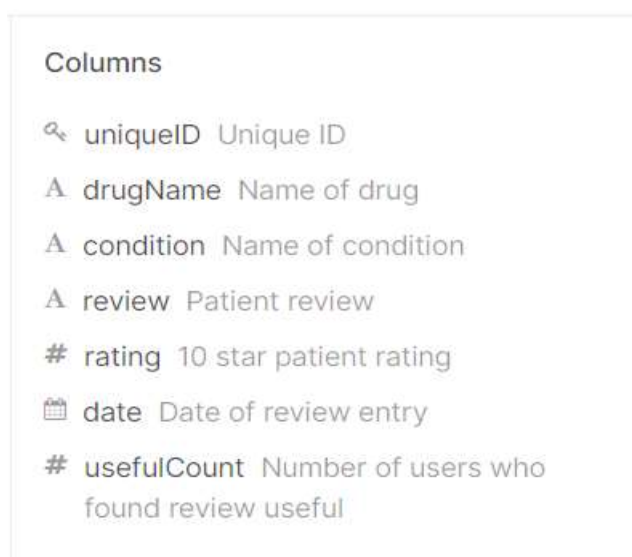
Text Blob package for Python is a convenient way to do a lot of Natural Language Processing (NLP) tasks. For example the English phrase "not a very great calculation" has a polarity of about -0.3, meaning it is slightly negative, and a subjectivity of about 0.6, meaning it is fairly subjective. There will be a question "Where do these numbers come from and this question is solved by going to the <https://github.com/sloria/TextBlob>. The lexicon it refers to is in en-sentiment.xml, an XML document that includes the following four entries for the word "great".

1. <word form="great" cornetto_synset_id="n_a-525317" wordnet_id="a-01123879" pos="JJ" sense="very good" polarity="1.0" subjectivity="1.0" intensity="1.0" confidence="0.9" />
2. <word form="great" wordnet_id="a-01386883" pos="JJ" sense="relatively large in size or number or extent" polarity="0.4" subjectivity="0.2" intensity="1.0" confidence="0.9" />
3. <word form="great" wordnet_id="a-01278818" pos="JJ" sense="of major significance or importance" polarity="1.0" subjectivity="1.0" intensity="1.0" confidence="0.9" />
4. <word form="great" wordnet_id="a-01677433" pos="JJ" sense="remarkable or out of the ordinary in degree or magnitude or effect" polarity="0.8" subjectivity="0.8" intensity="1.0" confidence="0.9" />

When calculating sentiment for a single word, TextBlob uses a sophisticated technique known to mathematicians as “averaging”. TextBlob goes along finding words and phrases it can assign polarity and subjectivity to, and it averages them all together for longer text. Here we consider only the polarity value based upon which the positive, negative and the neutral labels are given.

About Data Set

The UCI ML Drug Review dataset contains patient reviews on specific drugs along with related conditions and a 10-star overall patient scoring system reflecting patient satisfaction.



Columns	
uniqueID	Unique ID
drugName	Name of drug
condition	Name of condition
review	Patient review
rating	10 star patient rating
date	Date of review entry
usefulCount	Number of users who found review useful

Figure 1: Attribute Information

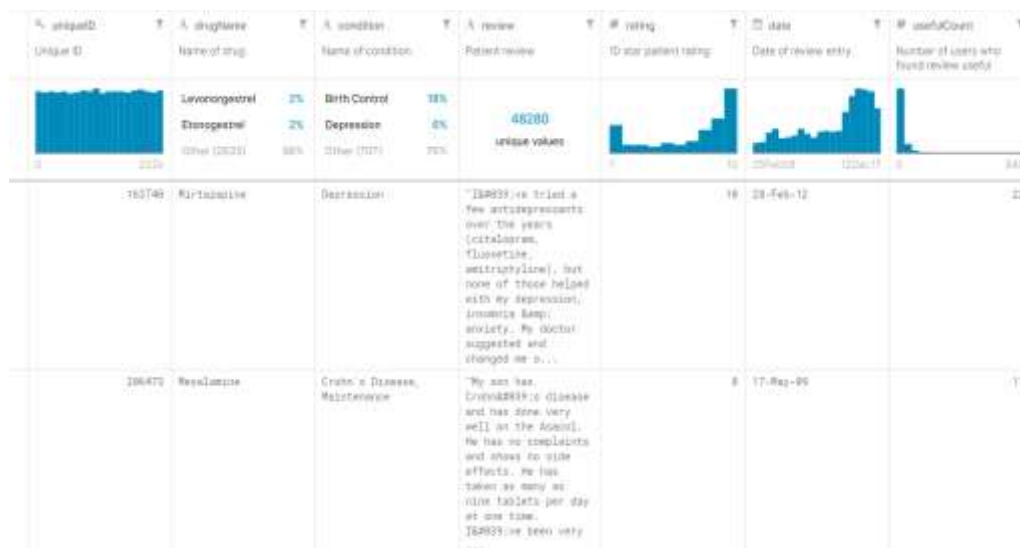


Figure 2: Few reviews of UCI ML data set

II. RELATED WORK

The authors T. T. Ashburn and K. B. Norse deity [1] Repositioning existing medication for brand spanking new indications may deliver the productivity will increase that the business wants whereas shifting the locus of production to biotechnology firms. Additional firms are scanning the existing repositioning candidates and the success rate is rapidly increasing.

The process of finding new uses outside the scope of the initial medical indication for existing medication is known as drug repositioning. The authors J. T. Dudley, T. Deshpande, and A. J. Butte [2] have projected that process ways for locating are most likely the foremost economical thanks to yield novel indications for these medication and no matter whether or not process ways become the quality for drug locating, it's clear that several alternative undiscovered uses of medicine do exist. Finding these new uses is a crucial and necessary step towards reducing the burden of malady.

The authors [11] have proposed that RBFN perform higher in operate approximation and yields higher accuracy in prediction. In general, neural network primarily based approaches perform higher than the applied mathematics based approach. The PNN (Precision = 88.6%, recall = 88.3%, f-score = 88.7%), RBFN(Precision = 93.8%, recall = 90.8%,f-score = 94.0%) are the models used.

In [12] the authors have used the Incremental approach and achieves an 82% accuracy on training corpus and 78% accuracy on testing corpus of health reviews. This study is based on the previous work, and we are improving the approach by increasing the accuracy of the model.

III. METHODOLOGY

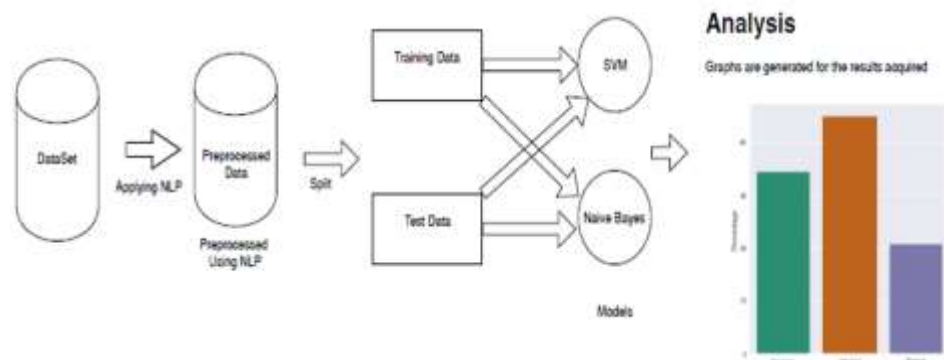


Figure 3: Process Diagram

Initially the data is collected from the kaggle i.e., a drug data set which is UCI ML Drug Review dataset.

	uniqueID	drugName	condition	review
	Unique ID	Name of drug	Name of condition	Patient review
1	163748	Mirtazapine	Depression	"I've tried a few antidepressants over the years (citalopram, fluoxetine, amitriptyline), but none of those helped with my depression, insomnia & anxiety. My doctor suggested and changed me o...
2	286473	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's disease and has done very well on the Asacol. He has no complaints and shows no side

Figure 4: UCI ML Drug Data Set

In this preprocessing step the missing values are removed. Then the special symbols present in the text or the review are replaced with the empty space and stop words are removed and the words obtained are stored respectively in the next column. The words are then joined as a sentence and then the polarity is found using the Text Blob module and stored in the other column.

Then the polarity is labelled by positive, negative, neutral. If the polarity values is greater than zero then it is labelled as Positive. If the values is less than zero then it is labelled as negative and if it is zero then it is labelled as Neutral.

In the Next Step two models are built i.e., using Naïve Bayes and Support Vector Machine Classifier. Sixty Percentage (60%) of the Data is used for the Training and the Remaining Data is used as the testing data. Only the sentiment and score are trained to model. Then the Training Data is used to build the SVM model and Naïve Bayes model. Then the testing data (40%) is used to find the accuracy of the models and for the further predictions.

Then the Graphs are plotted for the distribution in the test split for Support Vector Machine and Naïve Bayes algorithm.

IV. RESULTS AND DISCUSSIONS

After implementation it is observed that the SVM has shown highest accuracy. Figures 5 and 6 show the number of total positive, negative and neutral predictions of SVM and Naïve Bayes classifiers respectively. Where as Figure 7 and Figure 8 are presenting the confusion matrix for both SVM and Naïve Bayes classifications. It is observed that SVM is predicting with an accuracy of 98.44% (Percentage).Naive Bayes is Predicting with a accuracy of 98.12% (Percentage).

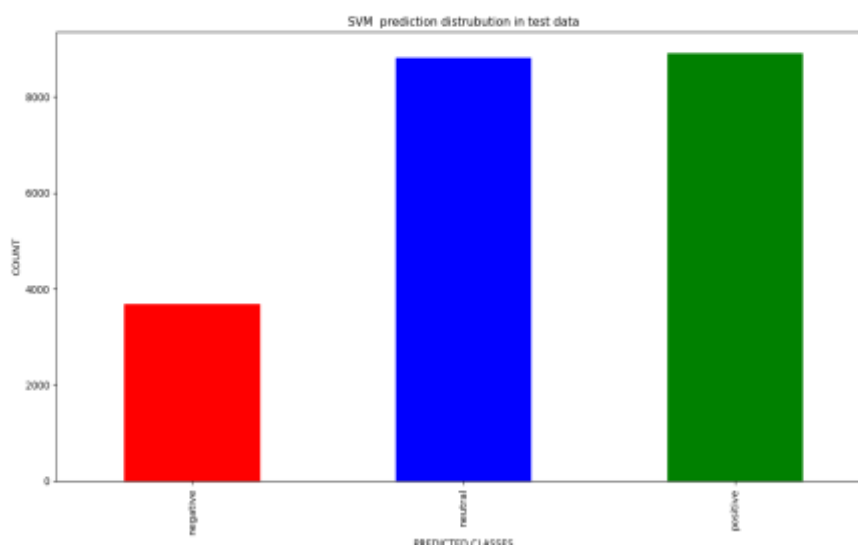


Figure 5: SVM graph

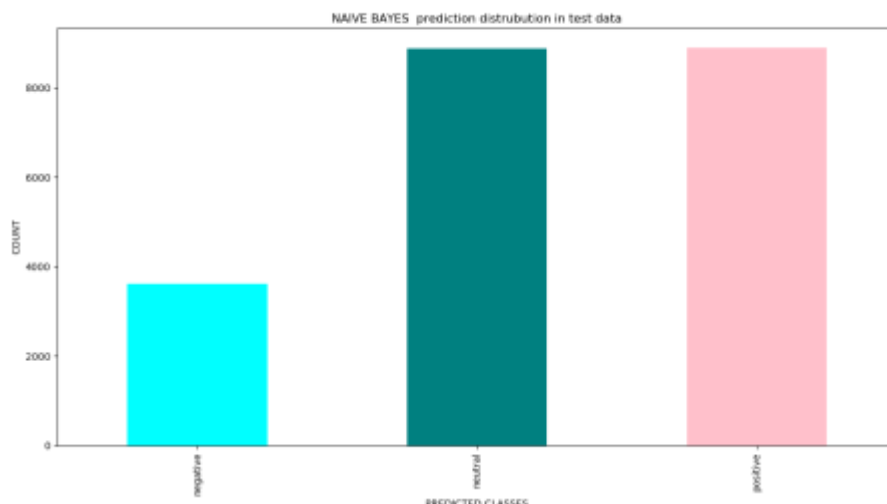


Figure 6: Naive Bayes Graph

The Confusion Matrix for the SVM and Naïve Bayes models is as follows

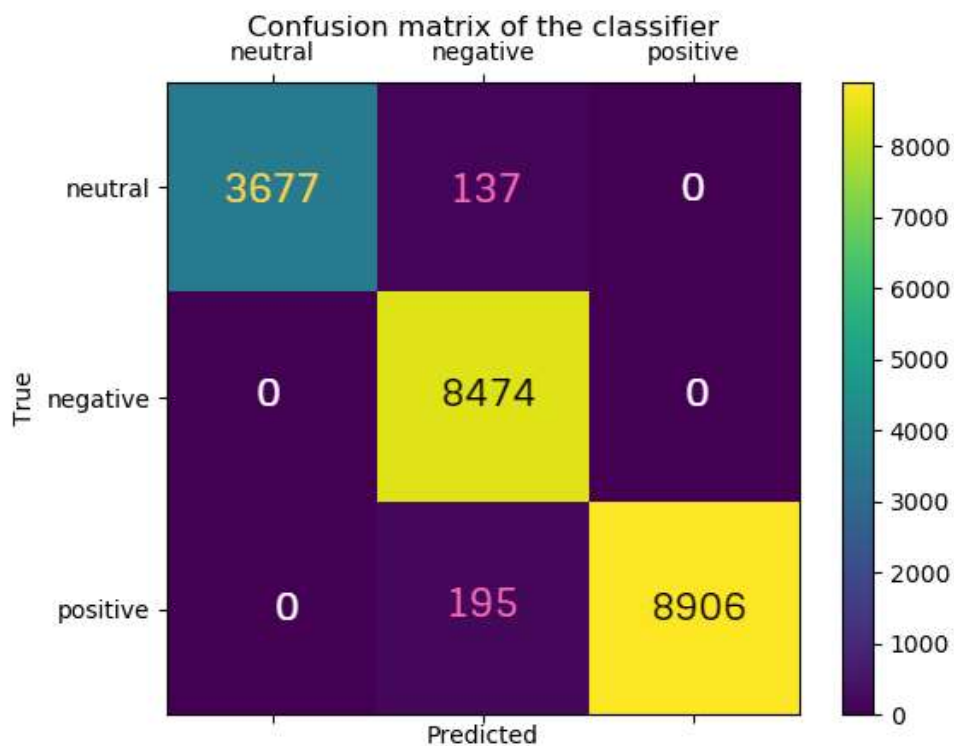


Figure 7: Confusion Matrix of SVM classifier

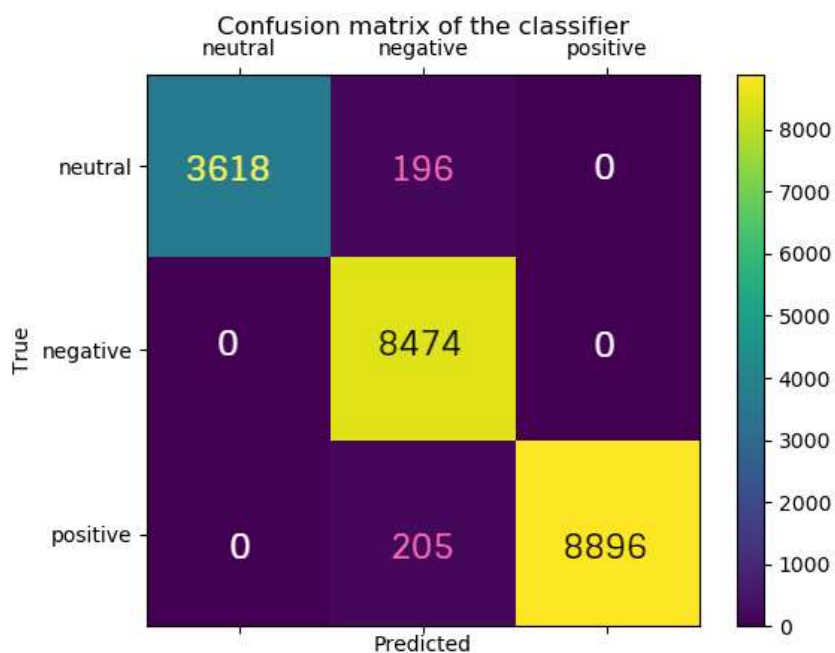


Figure 8: Confusion Matrix for Naive Bayes classifier

There will be many drugs which can be used for the particular symptom or condition. We can able to predict the top 5 list based upon the condition. Assume that if the condition is given as Depression then there are 24 Drugs related to that condition and among them the top 5 list is extracted based upon the Positivity Percentage in the Drug reviews as shown in Figure 9. Figure 10 shows a graphical representation of top 5 drugs that has highest positive percentage.

score	negative	neutral	positive
drugName			
Abilify	7.69	34.62	57.69
Aripiprazole	15.69	29.41	54.90
Mirtazapine	16.00	29.33	54.67
Trintellix	16.39	29.51	54.10
Wellbutrin	11.96	33.70	54.35

Figure 9: Top 5 Drug list for Depression

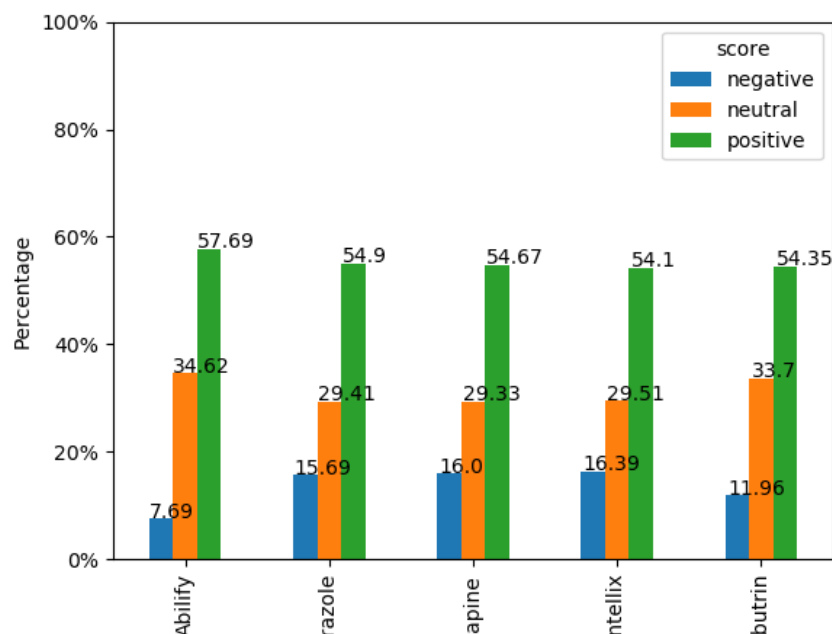


Figure 10: Graph of the Top5 Drug list for Depression

V. CONCLUSIONS

The application uses machine-learning models i.e., Support Vector Machine, Naïve Bayes and NLP. In several classification applications, SVM's have proved to be extremely performing and straightforward to handle classifiers with superb generalization. It provides a GUI to take user inputs for the Drug he need and can visualize analytic results. However, these efforts are just the initial phase of software development. The Accuracy Obtained using the SVM is 98.44%.

The SVM is compared with the Naïve Bayes and the SVM performs better than the NB as the accuracy obtained is about 98.12%. This implementation can be improved in several ways. In future the UI can be changed that the user can interact and acquire the results on their own. The new data acquired from the users of different drugs can be added to the model and that can help in increasing the accuracy.

REFERENCES

1. T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature Review Drug Discovery*, vol. 3, pp. 673-683, 2004.
2. J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug-disease relationships for computational drug repositioning," *Briefings in Bioinformatics*, vol. 12, pp. 303-311, 2011.
3. L. Yao, Y. Zhang, Y. Li, P. Sanseau, and P. Agarwal, "Electronic health records: Implications for drug discovery," *Drug Discovery Today*, vol. 16, pp. 594-599, 2011.
4. C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis, "Literature mining, ontologies and information visualization for drug repurposing," *Briefings in Bioinformatics*, vol. 12, pp. 357-368, 2011.

Variable	CNN	LSTM	CLSTM	CNN+FCN	LSTM+FCN	CLSTM+FCN	Input dimensions
No. of	10,000	10,000	10,000	10,195	10,195	10,195	

CNN filters 96 -- 384 768 -- 192 No. of FCN filters 128 -- 32 480 -- 480 No. of LSTM units -- 10
135 -- 35 120 No. of weights to train 345,889 8,665 380,561 1,532,833 51,297 446,561 IEEE
Transactions on NanoBioscience, Volume:18,Issue:3,Issue Date:July.2019

5. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. Abbas, S. J. Hufeisen, et al., "Predicting new molecular targets for known drugs," *Nature*, vol. 462, pp. 175-181, 2009.
6. P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, et al., "Use of genome-wide association studies for drug repositioning," *Nature Biotechnology*, vol. 30, pp. 317-320, 2012.
7. A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology*, vol. 7, p. 496, 2011.
8. J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner, "Knowledge discovery by automated identification and ranking of implicit relationships," *Bioinformatics*, vol. 20, pp. 389-398, 2004.
9. L. Yao, "In silico search for drug targets of natural compounds," *Current pharmaceutical biotechnology*, vol. 13, pp. 1632-1639, 2012.
10. H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, et al., "Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality," *Journal of the American Medical Informatics Association*, vol. 22, pp. 179-191, 2014.