Original Article

Investigation of Information gain and Chi test feature selection methods in dimensionality reduction using Machine learning for drug discovery

Dogga Aswani[1], Mamatha Vayelapelli[2], Uppada Gautami*[3]

[1,2]Assistant Professor, Dept. Of CSE, Anil Neerukonda Institute of technology and sciences, Sangivalsa, Visakhapatnam, AP, India

[3]Assistant Professor, Dept. Of IT, Anil Neerukonda Institute of technology and sciences, Sangivalsa, Visakhapatnam, AP, India

Email: ugautami.it@anits.edu.in

DOI: 10.47750/pnr.2022.13.S01.58

Abstract

Despite all the recent improvements made in the pharmaceutical industry, especially in the area of cancer research, there is still much room for development. The process researchers use to find new drugs haven't really changed. It costs money to take a drug from its discovery to market availability. The Tufts Center conducted research that indicates it takes at least 13 years and costs about $2.6 billion to develop a new drug. To reduce drug discovery timeline, machine learning plays a major role. Machine learning typically uses feature selection as a preprocessing step. The performance of learning algorithms is frequently enhanced by the elimination of duplicate and unnecessary data. Comparison on Information gain and Chi test in drug discovery is presented in this paper, mainly concentrated on dimensionality reduction and used SVM classifier to categorize chemical compounds.

Keywords: Information gain, Chi test, dimensionality reduction, features, unrelevant data, drug discovery, pre-processing, SVM classifier, classification, chemical compounds.

I. INTRODUCTION

Despite all the advancements made recently in the pharmaceutical business, particularly in the field of cancer research, there is still much room for growth. Since the 1920s, we haven't really changed how we discover new drugs.

1.1 Drug discovery process

From inspiration to development through approval, drug discovery comprises several stages and procedures. The process of discovering new drugs is a lengthy one that can last up to 13 years. Generally, screening for possibly active compounds is the first step in the initial drug discovery process. Following their discovery, these substances are tested for safety and efficacy. They must have a therapeutic impact on the intended ailment.

The process of bringing a medication from discovery to commercialization is expensive. According to a research done by the Tufts Center, the cost of finding a new medicine is projected to be roughly $2.6 billion [23]. There are several stages and steps involved in process of drug discovery. "Early Drug Discovery, Pre-Clinical Phase, Clinical Phases, and Regulatory Approval" are typically the four primary phases. Figure 1 shows the timeline for each process.

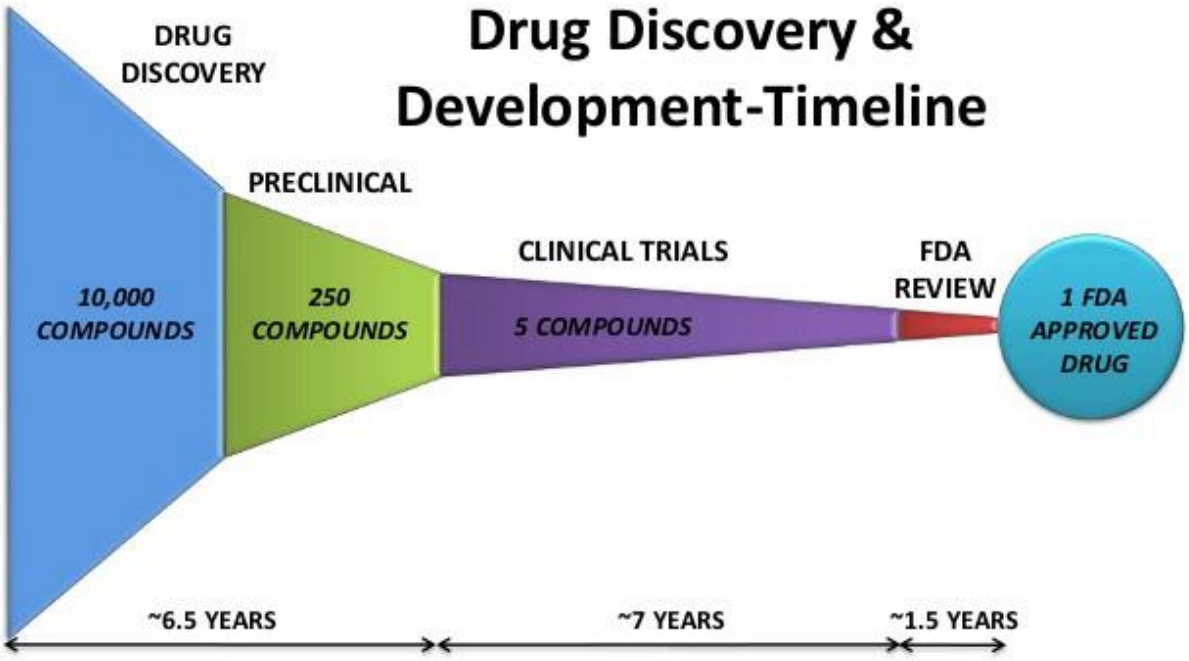

Figure 1. Drug discovery and development process (Image source: [24])

1.1.1. Early Drug Discovery

The early stages of drug discovery entail a number of procedures and examinations. In order to discover and improve prospective lead to certain objective, researchers work together. To potentially treat the disease, scientists essentially need to have a desired impact on a particular biological target. Currently, various animal models, cell cultures, biochemical testa and silico platforms are used in the laboratory to conduct research.

1.1.2. Pre Clinical Phase

Compounds discovered from earlier phase were thoroughly investigated in a lab setting as well as using animals/substitute models throughout this phase. Prior to Human trials starting, there had to be sufficient evidence of effectiveness and safety. The proper doses to test in people can then be determined after this is verified. Prior to the start of clinical trials, it needs to be ensured that new drug will be available in necessary quantities.

1.1.3. Clinical Phase

Clinical trials are divided into phases I, II, III, and IV. First, a relatively limited number of healthy people will be used to assess the drug candidate's safety and tolerability. Phases IIa and IIb have begun in order to assess the efficacy, acceptability, and dose in a bigger cohort. IIb phase study, goals of determining proper dosage, while IIa studies focus largely on the therapeutic idea. 100 to 500 adult research participants are typically included in phase II investigations.

Before the drug can be accepted as medicine, doctors test it on thousands of people in last phase to ensure that safety and effectiveness can be established in the diverse population. Additionally, drug interactions are examined. Phase II and III studies in which few patients receive newer drug while a second group receives pre-existing conventional medication.

1.1.4. Regulatory Approval

Data is then gathered and evaluated when a clinical trial for an active drug is complete. The necessary authorities can then evaluate it after it has been filed. Vaccination or drug cannot be sold without the permission of the centralized system/national regulatory body. In the end, just one drug out of the huge number examined survives “clinical study phases and regulatory testing”. Due of this, only one chemical may be used to make a drug or vaccine.

The main aim of this research work is to investigate Information gain and Chi test feature selection methods in dimensionality reduction for drug discovery.

2. Related work

Drug discovery has also utilized a number of additional machine learning (ML) approaches, such as unsupervised learning and multiple ML approaches.

Authors' [1] study focused on the identification of pharmacological side effects, particularly how cellular components respond to a medicine. The scientists used a 3D matrix made up of elements relating to the cell's content, test will be performed. This matrix showed every test which would be necessary for understanding how a medication would affect every protein target and kind of cell in the body. The Perturbagen Effect Hyper-Rectangle matrix was designed to avoid time-consuming experiments. These kinds of active learning models address the demand for an immediate answer, choose the tests that are most pertinent and helpful, and choose data point in matrix which are ambiguous and unexplored.

A graph-theoretic method to drug discovery was put out by researchers [2]. It is suggested that because to the criteria for the kernels, this technique was compatible with learning algorithms including Gaussian processes, ridge regression, support vector regression, support vector machine but not with NNs.

To determine which approach would be most effective in discovering and forecasting newer targets (proteins), newer medications, and newer interaction with target and medications, researches [3] introduced nearest neighbor similarity based approach. The first step included projecting target space and drug using kernels on 2 lower dimension regions. Next, estimates of target-drug interaction are made in lower dimension space. It is inefficient since, made use of 3 matrices, each of which had a random initialization.

While just carrying out 29 percent of all feasible tests, researchers [4] explored using active learning to understand the impact of "48 chemical compounds on subcellular localization of 48 proteins". In order to prevent the medicine from being tested on people it before underwent costly clinical trials, study showed the big data-based method for identifying harmful side effects. This system, called PrOCTOR, was inspired by the widely used Moneyball strategy in baseball. Researchers trained their computer to perform all of these tasks automatically by teaching it a set of 48 distinct criteria to evaluate every drug's safety for clinical usage.

Another popular approach in this area is the neural network (NN) technique. According to authors [5], employing machine learning in conjunction with Computer-Aided Drug Design (CADD) methods helps us understand how antibiotics combat bacteria that have developed a resistance to them. Antibiotics that have undergone experimental validation were identified using Decision Trees and NNs, two cutting-edge nonparametric machine learning approaches. They were subsequently put to use in order to anticipate receptor and ligand base binding strategies. These predictions were produced by the NN which primarily recreates cellular structure of brain.

The complexity of drug design was discussed by authors [6], who also examined the efficacy of various methods such support vector machines, counter-propagation neural networks, Bayesian neural network, multi layer perceptrons, self-organizing maps.

NN application for forecasting how specific malignancy would react to the therapeutic therapy was described by authors [7]. According to genetic features of cancerous cell lines and chemical characteristics of prescribed medicine, NN projected how the cancer cell lines would react to the drug therapy. Data were evaluated using a feed-forward multilayer perceptron and 8 cross fold validation. Back propagation is employed to train data model.

Drug research has made extensive use of SVM supervise learning approach. Even though tuberculosis (TB) is treatable, research [8] is challenging for predicting antibiotic resistance. To better forecast resistance, they discussed machine learning-based strategies based on well-known curative chemicals. Direct association, logistic regression, SVM was the machine learning methods employed in their research. After doing fivefold cross validation, discovered 3 techniques and models performed well across the majority of datasets. When it came to classifying resistance to isoniazid, the crucial Mycobacterium tuberculosis antibiotic, their algorithm ultimately achieved 93% accuracy (MTB).

The effectiveness of an SVM and NN approach for classifying drugs against non-drugs was examined by researchers [9]. The study offered a simple experiment via which it was demonstrated that the SVM model outperformed the NN models, although with a small margin of improvement [17][19].

In order to combine molecules from sizable collection for target molecules as rapidly and with fewest repeats or iterations as feasible, researchers [10] attempted this. The most crucial tactic was based on the Support Vector Machines-generated maximum margin hyperplane. Overall, their research has demonstrated the value of SVMs in distinguishing important information from irrelevant data. In active learning settings, the data was restored throughout trials for aiding in improving outcomes, the same holds true[18].

The Bayes theorem, which is extensively utilized in the field of machine learning, is also commonly employed in the search for new medicines. By analyzing datasets with machine learning, researchers [11] discovered the number of FDA approval drugs

which are effective on “Ebola Virus (EBOV) in vitro”. They employed Bayesian models to their methods, which selected substances for testing in order to forecast molecule inhibitors. They digitally screened chemical libraries after prioritizing compounds using models created from the higher throughput screen data. By the usage of EBOV replication assay and viral pseudotype entrance assay data, the researchers created Bayesian machine learning models. They examined the models on various dataset, using 5 cross fold validation, discovered recursive partitioning forest and bayesian model are accurate at forecasting, around 85% accuracy, whilst SVM had about accuracy 76 percent. Finally, three compounds with anti-EBOV capabilities were discovered, and the three most potent molecules were put to in vitro testing[19][21].

Sean Ekins [12] analyzed dosage responsive data for “vero cell cytotoxicity and whole-cell antitubercular activity” to develop new TB therapies. To handle datasets with single-point, dual-event dosage responses, twelve models were developed. Researchers evaluated dataset sizes ranging from 1,000 to 345,000 chemicals to see whether bigger datasets are superior for ML. Discovered that the dataset size made little effect, indicating that knowledge may be extrapolated from smaller datasets. As those are good models for merely dual-event dosage response, the authors are still worried about data dependencies with dual-event and single-point dosage response. Continue to be dubious about using these methods to treat other illnesses[20].

“Ligand-Based virtual screening (LBVS)” and the usage of ML models were described by authors [13]. Their studies have been effective to identify compounds, forecasting new active molecule due to development and expansion in large-scale, public-domain chemical and biological data. The biological characteristics and toxicity of chemicals, as well as protein targets and bioactivity classes, were predicted using the Naive Bayesian approach. This technique had proved effective in determining whether an active chemical was present[22].

In order to choose which chemicals to test, researches [14] recommended using three dual event Bayesian models in conjunction with bioinformatics and chemoinformatics. Then, these substances were examined, verified, and improved. This method might be used to discover cures for ailments like malaria.

3. Methodology

The main aim of this research work is to perform information gain and Chi test feature selection for drug discovery on SVM machine learning classifier to classify chemical compounds. Over view of the system is shown in figure 2.

3.1 Data collection

Dataset from KDD Cup [15] was utilized to generate the data for this research. “DuPont Pharmaceutical Research Laboratories” provided the findings of laboratory tests that evaluated organic compounds 1909 (train set) into whether those get combined to thrombin for this categorization competition. 42 compounds produced the favourable outcome. Every chemical is defined by the single feature vector that had 139 351 binary characteristics that represented the compound's three-dimensional properties and a class value (inactive (I), active (A)). 150 out of the 634 substances in the test set for the KDD cup were active.

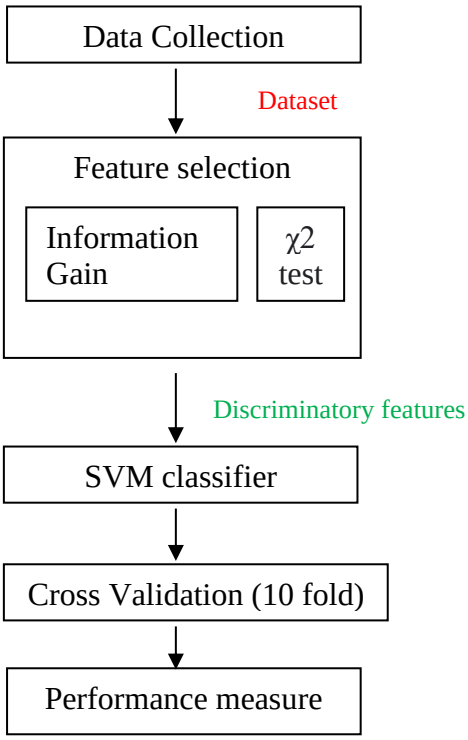

Figure 2: System overview

3.2 Feature selection

Machine learning uses the feature selection approach to pick the most relevant subset of the features that are present in a dataset. By doing this, the data noise may be reduced and a more accurate prediction or classification can be made. Each feature is given a score that indicates how significant or beneficial it is expected to be for training and classification by χ2 test and information gain (IG) approach.

3.2.1 IG: By knowing whether a feature is present or absent, information gain measures the quantity of information bits obtained for categorization prediction.

$$\text{Information gain } (f_k) = \sum_{c\varepsilon(ci,\,\hat{}ci)} \sum_{f\varepsilon(fk,\,\hat{}fk)} Pr(f,c) * \log \frac{Pr(f,c)}{Pr(f)*Pr(c)} \qquad (1)$$

Here,

^fk : k feature absence

fk : k feature presence

3.2.2 χ2 test :

A feature f and a category c's lack of independence is measured by CHI.

$$Chi(fk,ci) = \frac{N*(Pr(fk,ci)*Pr(\hat{}fk,\hat{}ci)-Pr(fk,\hat{}ci)*Pr(\hat{}fk,ci))2}{Pr(fk)*Pr(\hat{}fk)*Pr(ci)*Pr(\hat{}ci)} \qquad (2)$$

Here,

ci : active category

^ci : inactive category

3.3 SVM classifier

Classification and regression problems are resolved using SVM supervise learning technique. It is mostly used, nevertheless, in ML Classification problems. In order to swiftly categorize new data points, SVM algorithm aims to define the best line or decision boundary which can split n-dimensional spaces into categories. The name of this best choice boundary is hyperplane.

SVM is a hybrid of instance-based learning and linear modelling. An SVM creates the linear discriminatory function which indeed differentiates the crucial border samples of each category as feasible. The "kernel" approach would be utilized for dynamically injecting train data into the greater space, train divider in the space, if linear separation is not practicable [16].

When two feature vector groups are linearly separable, “by dividing them by maximum margin, or the separation between the nearest training vector and separating hyperplane”, SVM constructs a hyperplane. Finding a w (second vector) and b parameter which minimize ||w||2 ,meets requirements below allowed for the creation of the hyperplane.

$$w*x_i + b <= -1, \text{ for } y_i = -1 \text{ inactive (category 2)} \quad (3)$$

$$w*x_i + b >= +1, \text{ for } y_i = +1 \text{ active (category 1)} \quad (4)$$

3.4 Performance measure

Inactivity predictivity, active predictivity, specificity, sensitivity are the four measures used to test the performance of SVM classifier on the feature from feature selection methods IG and Chi test.

Inactive predicitivity : proportion of compounds whose predicted inactive was accurate.

Active predicitivity : proportion of projected active compounds which really were active.

Specificity: proportion of inactive substances which were appropriately identified.

Sensitivity: proportion of active substances which were categorized properly.

4. Experimental results

Based on the ratings that the IG and Chi test awarded, the features were sorted. And for classification purposes, only the highest-ranking features are considered. There were 100, 200, 500, 1000, 5000, 10000, 15000, 50000, 100000, and 139351 features tested. Table 1 and Figure 3 shows the feature selection method IG results for active and inactive predicitivity, table 2 and figure 4 shows the Chi test results for active and inactive predicitivity.

Table 1. Inactive and active predicitivity effect for IG feature selection method on SVM classifier

SVM	Predicitivity	
	Inactive	**Active**
IG-100	0.90	0.62
IG-200	0.95	0.81
IG-500	0.96	0.82
IG-1000	0.97	0.82
IG-5000	0.96	0.84
IG-10000	0.97	0.84
IG-15000	0.97	0.86
IG-50000	0.96	0.87
IG-100000	0.97	0.88
IG-139351	0.97	0.89

Table 2. Inactive and active predicitivity effect for Chi test feature selection method on SVM classifier

SVM	Predicitivity	
	Inactive	**Active**
Chi-100	0.90	0.40
Chi-200	0.92	0.50
Chi-500	0.93	0.55
Chi-1000	0.94	0.23
Chi-5000	0.94	0.62
Chi-10000	0.93	0.72
Chi-15000	0.93	0.73

Chi-50000	0.94	0.84
Chi-100000	0.95	0.84
Chi-139351	0.97	0.84

As the preprocessing step for machine learning, feature selection worked well to decrease dimensionality, eliminate unnecessary data, boost learning accuracy, and enhance result comprehensibility. This study examined and assessed the information gain and chi test feature selection approaches. “Sensitivity, specificity, active predictivity, and inactive predictivity” were employed as assessment metrics since they are often used in machine learning to assess the impact of feature selection on the quality of SVM classifier.

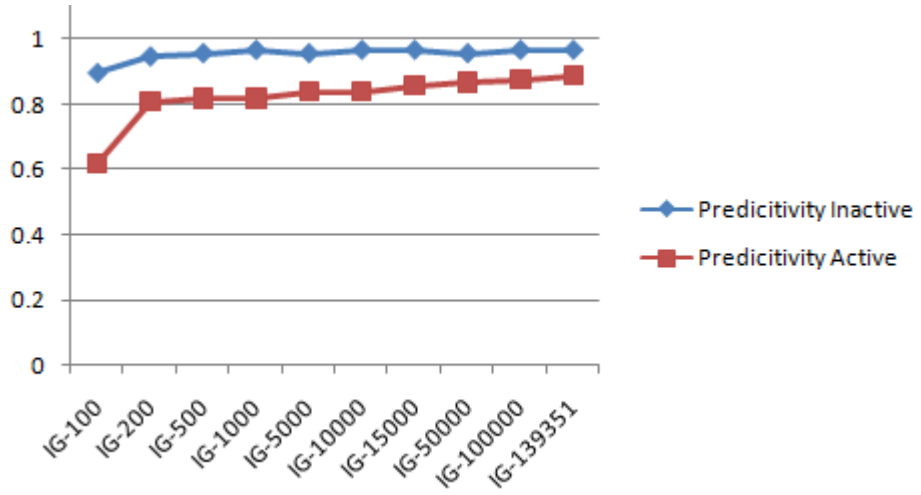

Figure 3. Information gain effect for active and inactive predicitivity on dimensionality reduction

SVM responded to the shrinkage of feature space considerably less. Only a little percentage was lost in terms of sensitivity “from 57.7% to 51.5%” and specificity "from 97.4% to 96.2%" as number of features is decreased by 99% "from 139351 to 100". Compared Chi test, Information gain gave better results in dimensionality reduction for drug discovery.

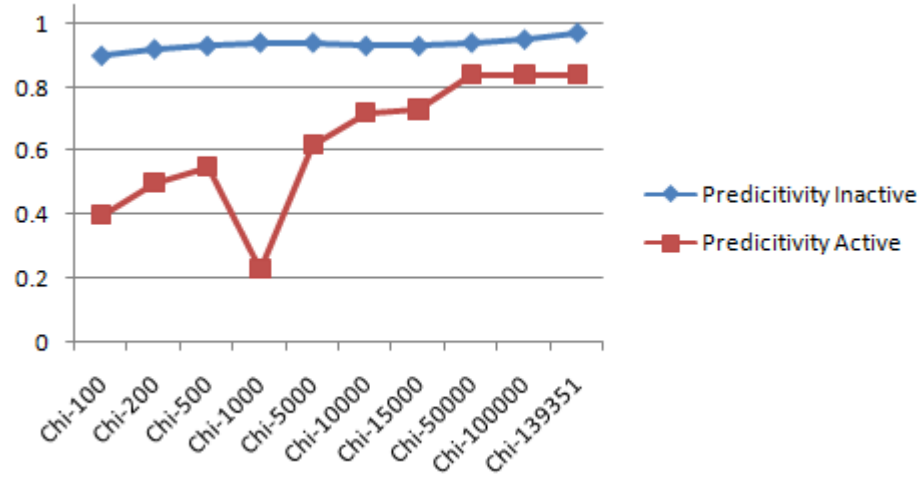

Figure 4. Chi test effect for active and inactive predicitivity on dimensionality reduction

5. Conclusion

The feature selection method is used by machine learning to identify the most pertinent subset of the features that are available in a dataset. In doing so, it's possible to lessen data noise and get predictions or classifications that are more precise. By using a Chi test and information gain (IG) strategy, each feature is given a score that represents how important or advantageous it is anticipated to be for training and classification. SVM classifier is used to classify chemical compounds. SVM reacted to the

reduction in feature space much less strongly. As the number of features was reduced by 99%, “from 139351 to 100”, just a little proportion of sensitivity “from 57.7% to 51.5%” and specificity “from 97.4% to 96.2%” were lost. Information gain produced superior outcomes in dimensionality reduction for drug discovery when compared to the Chi test.

REFERENCES

[1] Murphy, Robert F. "An active role for machine learning in drug development." Nature chemical biology 7.6 (2011): 327-330.
[2] Giguere, Sébastien, et al. "Machine learning assisted design of highly active peptides for drug discovery." PLoS computational biology 11.4 (2015): e1004074.
[3] Ding, Hao, et al. "Similarity-based machine learning methods for predicting drug–target interactions: a brief review." Briefings in bioinformatics 15.5 (2014): 734-747.
[4] Naik, Armaghan W., et al. "Active machine learning-driven experimentation to determine compound effects on protein patterns." Elife 5 (2016): e10047.
[5] Durrant, Jacob D., and Rommie E. Amaro. "Machine-learning techniques applied to antibacterial drug discovery." Chemical biology & drug design 85.1 (2015): 14-21.
[6] Lavecchia, Antonio. "Machine-learning approaches in drug discovery: methods and applications." Drug discovery today 20.3 (2015): 318-331.
[7] Menden, Michael P., et al. "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties." PLoS one 8.4 (2013): e61318.
[8] Niehaus, Katherine E., et al. "Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis." IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2014.
[9] Byvatov, Evgeny, et al. "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification." Journal of chemical information and computer sciences 43.6 (2003): 1882-1889.
[10] Warmuth, Manfred K., et al. "Active learning with support vector machines in the drug discovery process." Journal of chemical information and computer sciences 43.2 (2003): 667-673.
[11] Ekins, Sean, et al. "Machine learning models identify molecules active against the Ebola virus in vitro." F1000Research 4 (2015).
[12] Ekins, Sean, Joel S. Freundlich, and Robert C. Reynolds. "Are bigger data sets better for machine learning? Fusing single-point and dual-event dose response data for Mycobacterium tuberculosis." Journal of chemical information and modeling 54.7 (2014): 2157-2165.
[13] Lavecchia, Antonio. "Machine-learning approaches in drug discovery: methods and applications." Drug discovery today 20.3 (2015): 318-331.
[14] Ekins, Sean, et al. "Combining metabolite-based pharmacophores with bayesian machine learning models for mycobacterium tuberculosis drug discovery." PloS one 10.10 (2015): e0141076.
[15] Cheng, Jie, et al. "KDD Cup 2001 report." ACM SIGKDD Explorations Newsletter 3.2 (2002): 47-64.
[16] Liu, Huiqing, Jinyan Li, and Limsoon Wong. "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns." Genome informatics 13 (2002): 51-60.
[17] Natarajan, V. A., Babitha, M. M., & Kumar, M. S. (2020). Detection of disease in tomato plant using Deep Learning Techniques. International Journal of Modern Agriculture, 9(4), 525-540.
[18] Gampala, Veerraju, E. Fantin Irudaya Raj. "Deep learning based image processing approaches for image deblurring." Materials Today: Proceedings (2020).
[19] Mohamed Yasin Noor Mohamed. "Segmentation of nuclei in histopathology images using fully convolutional deep neural architecture." In 2020 International Conference on computing and information technology (ICCIT-1441), pp. 1-7. IEEE, 2020.
[20] Macha Babitha, ,”Machine Learning Based Identification of Covid-19 From Lung Segmented CT Images Using Radiomics Features”, Biosc. Biotech. Res. Comm. Special Issue, 14(07), 350-355.
[21] Naresh Tangudu , "Analysis of Groundwater Level Fluctuations and its Association with Rainfall Using Statistical Methods", JOURNAL OF ALGEBRAIC STATISTICS, Vol. 13 No. 3, PP: 1895-1904, 2022.
[22] V Tamizhazhagan, " Forecasting of Wind Power using LSTM Recurrent Neural Network", Journal of Green Engineering (JGE) Volume-10, Issue-11, November 2020.
[23] https://csdd.tufts.edu/tufts-csdd-cost-study
[24] https://medium.com/@niran_jan/drug-discovery-what-goes-behind-the-scene-6c0d5c0537a