ORIGINAL RESEARCH

Drug Side Effect Prediction Using Supervised Machine Learning Techniques

M. Pushpanjali . A Srinivasan . Y Sreeraman

Department of Computer Science and Engineering
Sreenivasa institute of technology and management studies, Chittoor, India.

Received: 08 November 2022 / Revised: 28 November 2022 / Accepted: 10 December 2022

Abstract – In the pharmaceutical sector, drug combinations are now a popular research issue, but experiment-based approaches are quite expensive in terms of both time and money. Many computational techniques have been put out to tackle these issues by beginning with current pharmacological combinations. To properly evaluate medication combinations, however, only chemical structure information is often provided, covering a rather narrow range of pharmacological properties. By combining the neighbour recommender approach with ensemble learning algorithms, we incorporated similarity-based multi-feature drug data to increase prediction accuracy. We chose the best medication characteristics by feature evaluation analysis, achieving an accuracy of 0.801 in the ensemble models. . Results of the study demonstrated that ensemble models outperformed conventional machine learning techniques including support vector machines (SVM), random forests (RF), and logistic regression (GLM). Additionally, we predicted 10 potential medication combinations for the medicine paclitaxel, and we were able to effectively confirm that two of the anticipated combinations have positive outcomes. The goal of this project take the input data from user and predict the side effects accurately automatic.

Index Terms – Supervised Learning, Machine Learning, Logistic Regression, Linear SVM, Random Forest

I. INTRODUCTION

The pharmaceutical sector has not invested much in research and development during the previous few decades [1]. The FDA (US Food and Medication Administration) does not approve the majority of drug candidates for commercialization because of significant negative effects. A hurdle in the drug-development process is identifying the unwanted off-target behaviours of prospective medications that are skilled of producing side effects and, in turn, leading to drug discovery disappointments. Some of the significant adverse effects are documented during the post approval surveillance, even though the majority are discovered during preclinical and clinical studies. Pharmaceutical firms, as well as consumers, are worried about the unknown side effects of new medications since they constitute a healthiness risk and may smooth be fatal [2].

The current computational techniques for pharmacological side effect prediction make the assumption that medications with similar organic and biological features, such as targets and structures, will have similar side effects. While Mistunes et al. invented a technique based on chemical structures of drugs and target proteins, Paywalls et al. used a sparse canonical correlation analysis model to predict probable medication side effects [3, 4]. Drugs with equivalent chemical structures have been shown to have comparable biological effects [5]. Even in drug design studies that forecast the characteristics of chemical compounds, reviewing a sizable number of chemical databases including the structures of accessible chemicals is a crucial step [6]. It makes sense to assume that shared pharmacological targets that provide comparable therapeutic benefits also result in comparable signaling cascades and, hence, comparable side effects.

After considerable attention was paid to chemical and biological features, earlier researches on side effect prediction was expanded to phenotypic traits. Along with chemical and biological information, Liu et al. included drug-phenotypic info in the features for machine learning and showed a noticeable increase in the forecast outcomes [7]. In calculation to medication alternatives, chemical structures, and targets, Zheng et al. used beneficial data. The premise behind these two investigations was that medications with similar beneficial benefits may also have similar adverse effects [8]. The great majority of earlier research, however, did not fully take use of what was known about medications. This study focuses on the hypothesis that a side effect's phenotypic manifestation may resemble that of a disease. Drug side special effects cannot be solely anticipated by their chemical features since they are focus to numerous factors such as metabolic changes and other pharmacokinetic transformations when they are metabolized and physiologically distributed [9]. Therefore, it was hypothesized that molecular similarities employed in medication repositioning may be used to the forecast of side effects. In addition, different similarity assessments may be useful in enhancing the model's predictive powers. Single nucleotide polymorphisms (SNPs) and drug-drug interactions (DDIs) that were not utilized by previous side effect forecast studies were utilized since both have been used in drug repositioning studies, where they have exhibited extraordinary results [10, 11].

By utilizing a variety of data sources on drug and side effect possessions, including (1) drug-drug connections from Drug Bank (DDIs-D), (2) drug-drug interactions from networks (DDIs-N), (3) SNPs, (4) chemical structures, (5) signs, (6) marks, and (7) side effect functional grading, this study proposes a supervised machine learning approach for the identification of potential drug side effects. To create a drug-side effect pair, a specified set of seven characteristics and several machine learning techniques were used. The findings demonstrated that integrating novel features suggested in this training in calculation to chemical, symptom, and target factors enhanced the side effect prediction model's capacity for prediction. Additionally, the detection of unlabeled adverse effects of prescription medications was a result of improving the machine learning model to attain the highest prediction performance. The research talks about the following four well-known medications: dasatinib, sitagliptin, vorinostat, and clonidine.

II. RELATED WORK

Public health forums emphasise the tension and the urgent need for research in this exciting period of contemporary medicine (Chee, Berlin, & Schatz, 2011). The idea behind precision medicine is to discover the patient's best, most effective treatments with the fewest negative effects. Since the discipline of translational bioinformatics is developing quickly, it is conceivable to conduct cutting-edge insilico drug discovery. (Butte, 2008) Predicting pharmacological side effects is a significant and essential step in the direction of customised medicine. Machine learning techniques have been utilised in scientific research in the past to predict drug-side effect 4 (W. Zhang, Liu, Luo, & Zhang, 2015). (T.-B. Ho et al., 2016).

This section examines computational strategies for discovering pharmaceutical side effects. Only a few of the earlier methodologies described here include cluster analysis, supervised deep learning methodology, factor analysis, causality analysis, network analysis, genome wide association studies (GWAS), enrichment analysis for result validations, and data-mining strategy. One of the data sources used to analyse side effects in various studies is chemogonomic information on both drugs and therapeutic targets. The DrugClust tool (Dimitri & Lió, 2017) is one recent development. It is a R tool that forecasts side effects using machine learning. The pipeline for analysis consists of two primary steps: cluster analysis and enrichment analysis. The data analysis pipeline initially groups the medications based on shared characteristics. While performing this cluster analysis, Bayesian priors are taken as given. In order to get a more biological interpretation of the created clusters, enrichment analysis is carried out for the clusters as a second step. The pathway enrichment analysis aids in examining interactions between drug clusters with complimentary profiles, which refer to medications that interact with comparable pharmacological targets, biological processes, and side effects. A measure called the Rand Index is used to determine if a cluster is statistically significant. Several publicly accessible datasets have been used to demonstrate the prediction performance.

Bresso et al. explained the adverse effects of medications using an integrated approach. The Drugbank and SIDER databases were used to collect the data. Drug targets descriptors and drug fingerprints are used to conduct drug clustering of comparable medications. Inductive-logic programming outperformed decision trees when the two machine learning techniques were compared, both in terms of performance and in terms of further elucidating the functional relationships between drug targets and pharmaceuticals' metabolic pathways. (2013) Bresso et al. Niu et al. used an intriguing strategy in which they ranked the medications using side effects as penalty scores. The average scores from simulation trials were used to rank the medications after scores were generated at random. For the study, three separate data sources—drug targets, chemical descriptors of pharmaceuticals, and drug therapy indications—were combined. In order to assign different weights to medications based on their various side effects, intended targets, and therapeutic indications, ensemble machine learning models were deployed. 2017 (Niu & Zhang) Giving out scores is a concept that has its roots in the gaming industry. It is used in this project to elaborate significant connections between drug-disease associations, drug and drug-side effects that are frequently brought on by treatments medications, and it can help researchers in pharmaceutical companies to develop hypotheses for drug discovery.

By focusing on 244 pharmacogenes from the PharmGKB database that are linked to side effects of 176 different medications, a connection between pharmacogenomics and side effects has been shown. The FDA has found 28 genes that are linked to a higher risk of adverse consequences. 2015's (Zhou et al.) Liang et al. used a different cutting-edge deep learning approach for genome-wide association studies (GWAS) to take use of patient phenotypic responses and pharmacogenomics data. Single nucleotide polymorphisms (SNPs), pharmacokinetic data, and side-effects data are all used in this supervised deep

learning process. This paradigm, in particular, organises single nucleotide polymorphism (SNP) with negative implications. This model makes use of stochastic organisations and step capabilities derived from markov chains. This approach outperformed the k-Closest Neighbor process and rope relapse. Liang, Huang, Zeng, and Zhang (2016). A large scope network focusing on natural pathways and incidental effects needs pathway information, target information, and phenotypic information. The findings of this investigation revealed that there are links between natural routes and secondary effects. Scheiber et al., 2009 Furthermore, the use of information mining has benefited pharmacovigilance. Hauben, Horn, and Reich (2007) define pharmacovigilance as the assessment and reporting of reported prescription adverse effects in light of purchaser drug observation critique.

III. PROPOSED METHODOLOGY

Dataset: - The benefits of user feedback for safe and effective medication usage are now being demonstrated. Consumer perceptions of their ailments and previously used medications are included in this data collection. Companies like 1mg may find this product useful in providing detailed ratings of the product's side effects on their website. Checking the adverse effects of the medications before purchasing them may also be beneficial for people who purchase medications online. We collected tis data from kaggle. In this dataset 1, 61,297 data points and 7 different columns are there. In this 7 columns we are going to use text and target columns only.

Data preprocessing: - Once we collected the data do data preprocessing concept. Check for some duplicate data points using pandas library using repeated data points there is no use and get some good accuracy and while deploying it we will loss those accuracy need to drop all the duplicates. Machine learning models won’t accept null values if we use null values we will get error need to remove it because we are working on text data. If we have categorical or numerical data by using mean or median or most frequent word we can replace the null values.

Text Preprocessing: - Whenever we have a text data need do apply text processing and clean it. In this text preprocessing first step punctuation symbols removal. First step want to remove some punctuation removal there is no using this symbol and get create some high dimensionality. Second step remove the stop words define or import the stop words from NLTK tool and remove the all stop from the each data point and then apply the tokenization. In this step split the sentence into words and apply stemming. Stemming is nothing but convert the word into base form for example beautiful, beauty, betaines the base form is beauty. By using stemming concept we can reduce the dimensionally also. By doing the all the text preprocessing steps we will step preprocessed text. Apply text featuraization concept on preprocessed text.

Bag of Word: - Bag of Word id on the most used text featuraization technique. It will create dictionary of unique words first and create the dimensionality for all the unique words then go to each data point and each word if the word already present in that directionality make it 1 not present make it 0. It will create a sparse matrix. Sparse matrix means less non zero elements. The machine learning algorithms don't accept this text data or categorical data directly convert into mathematical or numeric format using BOW.

Logistic Regression: - Logistic Regression is one of the mostly used and simplest Machine learning and supervised algorithm. Logistic regression widely used in most of the internet applications and low latency system. The time complexity of the logistic regression is very low training time complex to other machine learning algorithms. It works for categorical data with high categories and linear data also. In this algorithm sigmoid function used as cost function. Sigmoid function predict the values in probabilistic between 0 and 1. In this we are working on text data. First need to convert the text data into numerical data using BOG. Split the as train and test fit the logistic regression model on train data and predict it using the test data. In logistic regression simple linear model can't handle the outliers and outlier impacts also very but using sigmoid function we can handle it. Sigmoid function is robust model handle outliers very easily.

$$J(\theta) = -\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i))) + \left(1-y^{(i)}\right)\log(1-h\theta(x(i)))\right]$$

FIG. 1: LOGISTIC REGRESSION MODEL COST FUNCTION

Linear SVM model: - Liner SVM model is one of the most used and common supervised machine learning algorithm. Linear SVM can do classification and regression also. For classification we use support vector classifier and regression support vector regression. In the classification to distinguish or predict the target variable data we use SVM. The goal of SVM is draw a line intersect with positive data points and negative data points those lines are called positive support vectors and negative support vector. We need to maximize the distance between these two lines if the distance is high the accuracy is high and predict the class label very accurately.

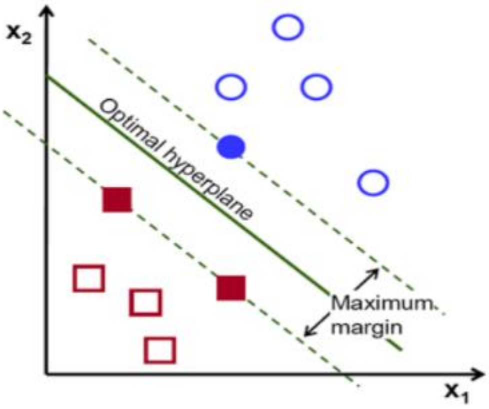


FIG. 2: SVM MODEL ARCHITECTURE

The fundamental goal of the SVM classifier is to find the plane with the greatest margin between the two data classes.

Random Forest: - Finally building random forest. In random forest 2 hyper parameters are need to tune. Using cross validation technique or grid search or random search tune the hyper parameter. If we pick the wrong hyper parameter might be get overfitting or under fitting. In random forest using IG (Information Gini) value build the tree. Random forest model time complexity is little bit high but give best results compared to another machine learning technique.

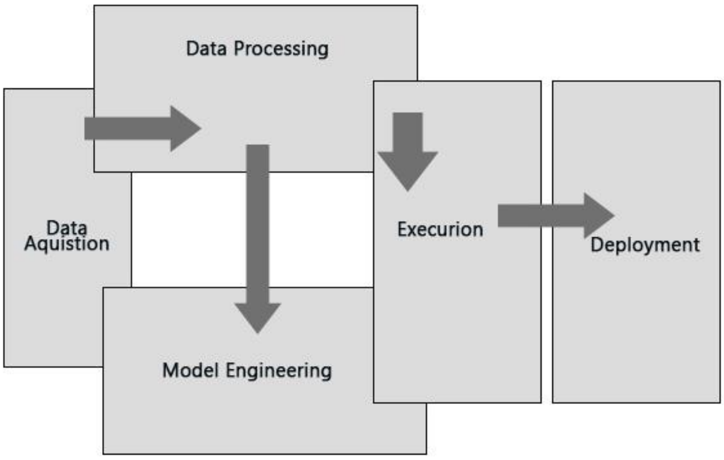
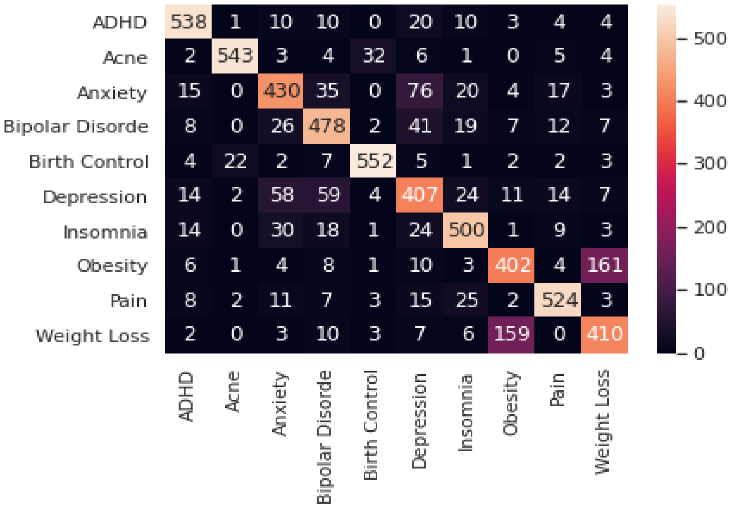


FIG. 3: PROPOSED WORK ARCHITECTURE

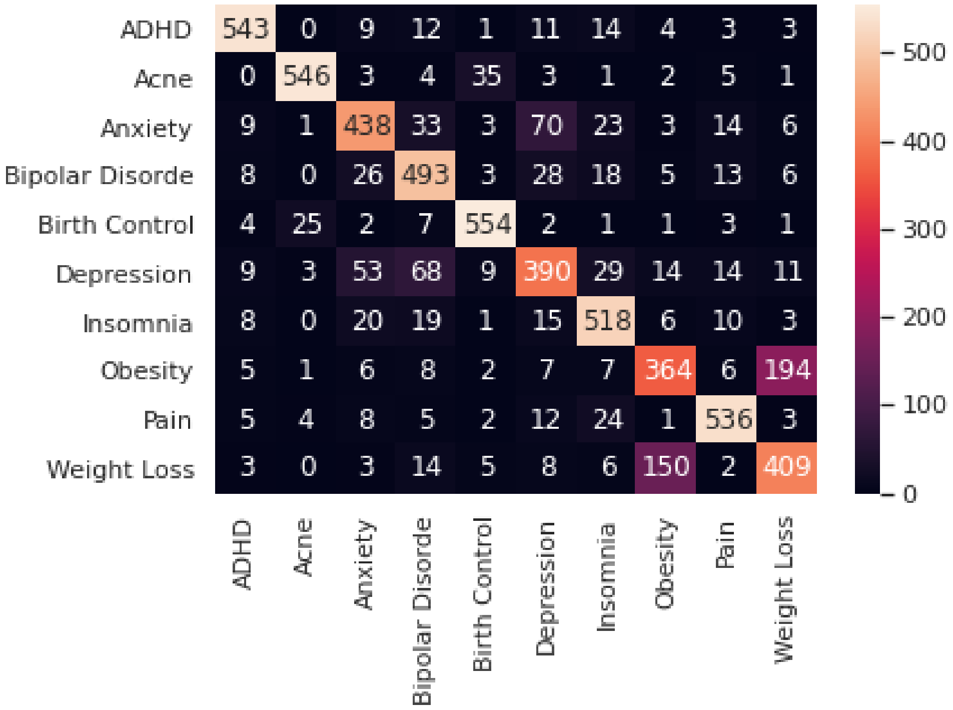
IV. RESULT AND ANALYSIS

This is multiclass classification problem total 10 classes are there. Our project complete goal take one new data point and classify or predict the class label. First we tried logistic regression model. By using this model we got 79.83 % accuracy and plotted confusion metric also. You can see the result in fig 4.



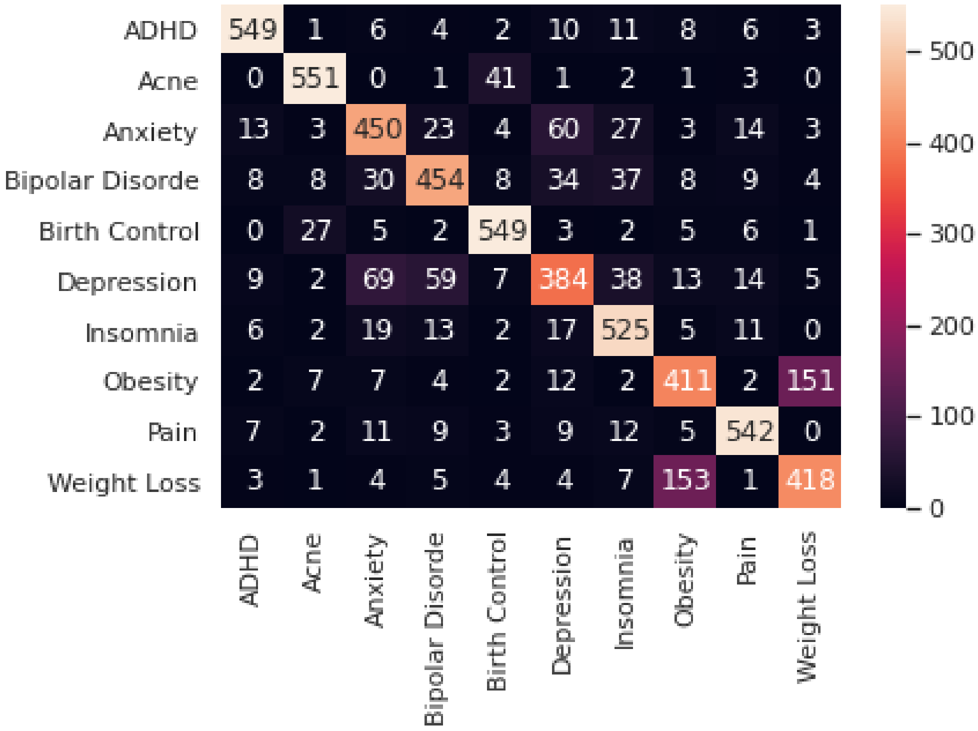
	ADHD	Acne	Anxiety	Bipolar Disorde	Birth Control	Depression	Insomnia	Obesity	Pain	Weight Loss
ADHD	538	1	10	10	0	20	10	3	4	4
Acne	2	543	3	4	32	6	1	0	5	4
Anxiety	15	0	430	35	0	76	20	4	17	3
Bipolar Disorde	8	0	26	478	2	41	19	7	12	7
Birth Control	4	22	2	7	552	5	1	2	2	3
Depression	14	2	58	59	4	407	24	11	14	7
Insomnia	14	0	30	18	1	24	500	1	9	3
Obesity	6	1	4	8	1	10	3	402	4	161
Pain	8	2	11	7	3	15	25	2	524	3
Weight Loss	2	0	3	10	3	7	6	159	0	410

FIG 4: LOGISTIC REGRESSION MODEL CONFUSION MATRIX



	ADHD	Acne	Anxiety	Bipolar Disorde	Birth Control	Depression	Insomnia	Obesity	Pain	Weight Loss
ADHD	543	0	9	12	1	11	14	4	3	3
Acne	0	546	3	4	35	3	1	2	5	1
Anxiety	9	1	438	33	3	70	23	3	14	6
Bipolar Disorde	8	0	26	493	3	28	18	5	13	6
Birth Control	4	25	2	7	554	2	1	1	3	1
Depression	9	3	53	68	9	390	29	14	14	11
Insomnia	8	0	20	19	1	15	518	6	10	3
Obesity	5	1	6	8	2	7	7	364	6	194
Pain	5	4	8	5	2	12	24	1	536	3
Weight Loss	3	0	3	14	5	8	6	150	2	409

FIG. 5: LINEAR SVM MODEL CONFUSION MATRIX



	ADHD	Acne	Anxiety	Bipolar Disorde	Birth Control	Depression	Insomnia	Obesity	Pain	Weight Loss
ADHD	549	1	6	4	2	10	11	8	6	3
Acne	0	551	0	1	41	1	2	1	3	0
Anxiety	13	3	450	23	4	60	27	3	14	3
Bipolar Disorde	8	8	30	454	8	34	37	8	9	4
Birth Control	0	27	5	2	549	3	2	5	6	1
Depression	9	2	69	59	7	384	38	13	14	5
Insomnia	6	2	19	13	2	17	525	5	11	0
Obesity	2	7	7	4	2	12	2	411	2	151
Pain	7	2	11	9	3	9	12	5	542	0
Weight Loss	3	1	4	5	4	4	7	153	1	418

FIG 6: RANDOM FOREST MODEL CONFUSION MATRIX

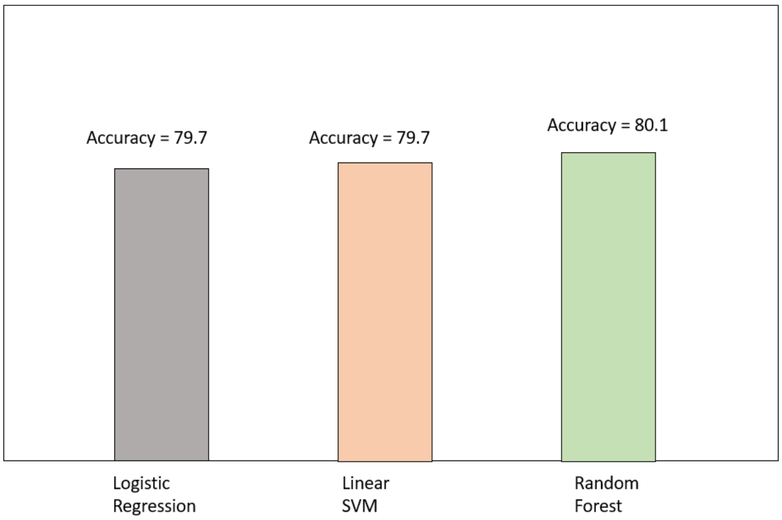


FIG. 7: RESULTS COMPARISON TABLE

By using logistic regression some miss classified data points there we can see it fig 4 confusion metric then try Linear SVM model with hyper parametric tuning. Using SVM also there is no much change we got almost same accuracy 79.8%. We are not satisfying with this result we are expecting better results. We tried linear models only so decided to build tree based model. Because of the data tree based model also not performing well we got 80.1% accuracy. There is no that much of different between linear models and random forest model. Finally random forest model giving good result compared to linear models 80. % accuracy we can see it fig 5 and fig 6. Finally we can conclude random forest is our proposed model see the results in fig 7. We deploy the project with random forest model because compare to three models it is giving slightly better result.

V. CONCLUSION

Drug development is being revived by ML-based methods. These techniques are based on several applications in the identification of targets, lead compounds, synthesis, protein-ligand interactions, etc. Algorithm-enhanced data query, analysis, and creation are being made possible by machine learning (ML) applications. One such instance is the use of ML to target identification, which mainly relies on the analysis and exploration of pre-existing omics and medical data. Viable targets may be determined utilizing data clustering, regression, and classification from enormous omics datasets and sources through the integration of AI employing ML approaches. Finally we are predicting with 80.1 % accurately using random forest model. In future by using deep learning models we can increasing the accuracy and using large number data set or increasing the train and test data size we can increase the accuracy.

REFERENCES

1. Fayyad, J., Jaradat, M. A., Gruyer, D., & Najjaran, H. (2020). Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, *20*(15), 4220.

2. Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060-1089.
3. Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.
4. Joachims, T., & Radlinski, F. (2007). Search engines that learn from implicit feedback. *Computer*, *40*(8), 34-40.
5. Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., & Greyson, D. (2011). The cost of drug development: a systematic review. *Health policy*, *100*(1), 4-17.
6. Ng, H. W., Zhang, W., Shu, M., Luo, H., Ge, W., Perkins, R., ... & Hong, H. (2014, December). Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. In *BMC bioinformatics* (Vol. 15, No. 11, pp. 1-15). BioMed Central.
7. Ng, H. W., Shu, M., Luo, H., Ye, H., Ge, W., Perkins, R., ... & Hong, H. (2015). Estrogenic activity data extraction and in silico prediction show the endocrine disruption potential of bisphenol A replacement compounds. *Chemical research in toxicology*, *28*(9), 1784-1795.
8. Hong, H., Neamati, N., Winslow, H. E., Christensen, J. L., Orr, A., Pommier, Y., & Milne, G. W. A. (1998). Identification of HIV-1 integrase inhibitors based on a four-point pharmacophore. *Antiviral Chemistry and Chemotherapy*, *9*(6), 461-472.
9. Hong, H., Tong, W., Xie, Q., Fang, H., & Perkins, R. (2005). An in silico ensemble method for lead discovery: decision forest. *SAR and QSAR in Environmental Research*, *16*(4), 339-347.
10. Hong, H., Fang, H., Xie, Q., Perkins, R., Sheehan, D. M., & Tong, W. (2003). Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. *SAR and QSAR in Environmental Research*, *14*(5-6), 373-388.
11. Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, *23*(8), 1538-1546.
12. Talevi, A., Morales, J. F., Hather, G., Podichetty, J. T., Kim, S., Bloomingdale, P. C., ... & Conrado, D. J. (2020). Machine learning in drug discovery and development part 1: a primer. *CPT: pharmacometrics & systems pharmacology*, *9*(3), 129-142.
13. Gertrudes, J. C., Maltarollo, V. G., Silva, R. A., Oliveira, P. R., Honorio, K. M., & Da Silva, A. B. F. (2012). Machine learning techniques and drug design. *Current medicinal chemistry*, *19*(25), 4289-4297.
14. Agarwal, S., Dugar, D., & Sengupta, S. (2010). Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, *50*(5), 716-731.
15. Rodrigues, T., & Bernardes, G. J. (2020). Machine learning for target discovery in drug development. *Current Opinion in Chemical Biology*, *56*, 16-22.
16. Gao, D., Chen, Q., Zeng, Y., Jiang, M., & Zhang, Y. (2020). Applications of machine learning in drug target discovery. *Current Drug Metabolism*, *21*(10), 790-803.
17. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, *18*(6), 463-477.
18. Zoffmann, S., Vercruysse, M., Benmansour, F., Maunz, A., Wolf, L., Blum Marti, R., ... & Prunotto, M. (2019). Machine learning-powered antibiotics phenotypic drug discovery. *Scientific reports*, *9*(1), 1-14.
19. Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., ... & Clark, A. M. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature materials*, *18*(5), 435-441.
20. Kumar Raja, D. R., Hemanth Kumar, G., Basha, S. M., & Ahmed, S. T. (2022). Recommendations based on Integrated Matrix Time Decomposition and Clustering Optimization. *International Journal of Performability Engineering*, *18*(4).
21. Kumar, S. S., Ahmed, S. T., Xin, Q., Sandeep, S., Madheswaran, M., & Basha, S. M. Unstructured Oncological Image Cluster Identification Using Improved Unsupervised Clustering Techniques.
22. Ahmed, S. T., Singh, D. K., Basha, S. M., Abouel Nasr, E., Kamrani, A. K., & Aboudaif, M. K. (2021). Neural network based mental depression identification and sentiments classification technique from speech signals: A COVID-19 Focused Pandemic Study. *Frontiers in public health*, *9*, 781827.