Predicting New Drugs for the Treatment of Chagas Diseases using Machine Learning and Docking

Kavya Singh*
Department of Biotechnology, Indian Institute of Technology, Roorkee, Haridwar, Uttarakhand, India.
*Corresponding Author
Mail id – Ksingh@bt.iitr.ac.in
Kavyasingh585@gmail.com

ABSTRACT

Chagas otherwise also known as as American trypanosomiasis is a hazardous disease caused by the protozoan parasite Trypanosoma*cruzi (T. cruzi)*. In today's world, only two drugs are available for the treatment of this fatal disease. So, there is an urgent requirement of some medications that will going to protect people from this disease caused by *T.cruzi.* In this research paper, we have used the Machine Learning model on the basis of the C4.5 algorithm which has an accuracy of around 65%. This model predicts the drugs which could be used for the treatment of the Chagas disease. Around 280 drugs were predicted which included the available ones also. These drugs were docked for the validation of their accuracy and effectiveness using AutoDock 4.2 and visualized by PyMol software. Among all, we suggest that the Lomitapide (Drug Bank ID – DB08827) would be probably one of the most effective on the basis of the selected criterions. Hence, we could expect progressing the theoretically proved drugs to be soon at the investigational stages and then finally avail to the required patients of the Chagas Disease.

Index Terms: Docking**,** C4.5 algorithm, Machine Learning**,** Naïve Bayes, PyMOL, Random Forest**,** Support Vector Machine

INTRODUCTION

Chagas otherwise also known as as American trypanosomiasis, is a hazardous disease caused by the protozoan parasiteTrypanosoma*cruzi (T. cruzi)*[1,2].It is spread mostly by insects named as Triatominae, or "Kissing bugs". Symptoms caused by this disease changes over a period of time. Symptoms are typically either not present or mild and may include headache, fever or local swelling at the site of bite. During the chronic phase, the parasites are hidden mainly in the heart and digestive muscles .*T.cruzi* have many mode of transmission: consuming food contaminated with *T.cruzi* through, for example making contact with urine or faeces of infected parasites[3,4]. Up to 10% of people develop digestive, neurological or mixed alterations and around 30% of chronically infected people develop cardiac alterations. In some later stages this fatal disease can lead to the sudden deaths due to progressive heart failure or cardiac arrhythmias, which is caused by destruction of heart muscle and nervous system[4,5].

It has been estimated that around 6 million to 7 million people are affected by this *Trypansosomacruzi*, the parasite that is responsible for causing the Chagas disease[6]. This fatal disease can be prevented through a number of methods. Several experimental treatments have shown promise in animal models. These include inhibitors of oxidosqualenecyclase and squalene synthase, cysteine protease inhibitors, dermaseptins collected from frogs in the genus Phyllomedusa (P. oreades and P. distincta), the sesquiterpene lactone dehydroleucodine (DhL), which affects the growth of cultured epimastigote-phase Trypanosomacruzi, inhibitors of purine uptake, and inhibitors of enzymes involved in trypanothione metabolism. Megazol seems to be more active against Chagas than Benznidazole, but it had not been studied in humans. Hopefully, new drug targets may be revealed following the sequencing of the *T. cruzi* genome. Chagas vaccine (TcVac3) has been only studied in mice[7–9].

It is not only present in Latin America but it is increasingly spreading among different countries such as Europe, Japan, Australia, North America mainly due to migration. This parasite *T.cruzi* is responsible for causing 20,000 annual deaths, infecting more than 10 million people globally. The treatment cost for curing this particular disease remains substantial. In Colombia alone, the annual cost of medical care for all patients with the disease was estimated to be about US$ 267 million in 2008. Spraying insecticide to control vectors would cost nearly US$ 5 million annually – less than 2% of the medical care cost[3,7].

This disease is prevailing since last few decades. It was initially spread across the Latin America and caused very low life expectancy rate because of it. Today with the help of several advanced techniques, the life expectancy has been increased dramatically and also several cure available for this disease[6,9]. In

this new era, people have found several techniques to fight against the disease and also they have overcome the attached social issues with the disease.

In this research paper, we are applying machine learning to predict several new potential drugs and to validate the drugs predicted, we are using Autodock4.2 software and PyMOL software for visualization purposes. Using machine learning we are training our model with the inhibitors of the disease caused by *T.CRUZI.* The approved and investigational drugs are taken from the Drug bank as a test model to predict the new drugs for the treatment of Chagas Disease. These new drugs are again validated using docking method to ensure that the drugs match with the same active site on protein as previously accepted drugs do.

MATERIAL AND METHODS

Dataset

In this research, the compounds of the dataset are tested in the cell based system using plate reader and then their results are stored as Bioassay Dataset in the Pubchem. These datasets contain around 10822 compounds as two activity sets and they are the inhibitors of *T.cruzi* replicating in the cells. These datasets are stored in the section Bioassay of PubChem database of National Centre for Biotechnology Information (NCBI), and they have the identification AID number as AID 651739 and AID 651740[10]. These corresponding bioassays are belong to the Broad Institute Inhibition of *T.cruzi*replication in culture Inhibitor Probe Project. The compounds are classified under three distinct categories as actives, inactives and inconclusive. So, here the compounds which are significantly suppressing the luminescence, and thusly b-galactosidases articulation will be recognized as hits in the screen. Compounds that inhibit luminescence activity may kill *T. cruzi*, inhibit *T. cruzi* invasion or inhibit development of the parasite within the host cell and hence these are classified under the Active section and the compounds which do not show effectiveness are classified under Inactive section. These complete datasets were downloaded in the form of SDF (Structure Data File).

The DrugBank is an online database which contains detailed data about various medications.Today, it has been widely used to facilitate in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction and general pharmaceutical education. This database of more than 4900 Drugs is categorized into many different types as Trial stages Drugs, Approved Drugs and Withdrawn Drugs. In this database, more than 45% of drugs are approved for various medication purposes[11]. In this research, we have focused only on the approved drugs which are around 2388. These drugs were downloaded in the form of SDFs and after several processing the description generated were taken as the test model for the train model which was made on the basis of database containing the inhibitors of the *T.cruzi*. Then the model has predicted few of the potential drugs.

The NCBI Protein Database were used in the process of getting the FASTA sequence of the desired protein (Cruzain). The FASTA sequence was in use for the modeling of the Protein 3D Structure and on the basis of this structure the docking of the known and predicted drugs have carried on[12].

Processing Dataset

These datasets are in the form of SDFs. So to train the model, we need to generate the attributes present in the SDFs. In this case, we have used the PowerMV Description Generator Software. Here, the information present in the SDFs are generated as CSV files which are used as the training dataset and test dataset for preparing the Machine Learning models[13]. This CSV files containing both the actives and inactives are split into 80% as training dataset and 20% as test dataset. This entire splitting process is random. This process is done by self-written python code to split as per the conditions.

The following FASTA sequence of the Cruzain was modeled using the SWISS – MODEL[14–17]. This provided the predicted structures for the Cruzain on the basis of its FASTA sequence. Then, with reference to the QMEAN Score4, we took the best possible prediction among all the predicted structures.

Machine Learning

Today, it has been almost impossible to think of analyzing the large datasets without using the Artificial Intelligence. Machine Learning is a part of Artificial Intelligence which allows us to predict the some important features of the datasets after training the model[18,19]. In this case, there is no need of explicitly programming the methods of analysis, which has been one of the most important advantages of implementing Machine Learning. Using Machine learning, we can think of both the classification and regression. It is dependent only upon the algorithm, we are implementing to it and the presentation of the datasets. Here, in this case, we have implemented the classification algorithms.

Classification Algorithm

The classification is a type of supervised learning in which the computer system can learn from the dataset which contains the detail and practical results. The algorithmic procedure of the classification is to assign an input value according to the description in the datasets[20]. So, for this it requires a mathematical

classifier which can assign certain class(Actives and Inactives) labels to instances defined by the attributes. In this process, the training model is made to learn according to a dataset where the classification is already assigned and on the basis of which it is able to run on different datasets to classify them according to the present instances[21]. In machine learning, there are several number of algorithms used for the classification purpose. In this study, we have compared the results from the classifiers that are Naïve Bayes, Random Forest, SMO and C4.5. The different features of these classifier are presented below: -

Naive Bayesian classification algorithm is very easy and simple by assuming that its classification attributes are independent and they don't have any correlation with each other[22,23]. It is a type of classifier that depends on Bayes' hypothesis. Naive Bayes does work best in two cases: complete independent feature (as expected) and functionally dependent features(as expected) and is a widely tested method for probabilistic induction. This classifier does tremendously well and has advantages over many other induction algorithms. It is nothing but easy to work as it has no entangled iterative parameter that makes it work for vast data sets.

This algorithm is more useful than any other induction algorithms because of its computation speed and reliability. It can be useful both the binary classification as well as multi-classification. But one of the main disadvantages which make it less popular is while determining the several relations between the different attributes[24].

Random Forest is a stream classification algorithm that users use the same techniques of traditional random forest to build steaming decision trees. This algorithm was first ever created by the Tim Kam Ho[25]. It is counted in the category of Supervised Learning. It considers two parameters that are number of trees to be built and tree window. The decision tree is the basic building block of Random forest. This decision tree is used in the process for interpretation of accurate results. In an understandable way of saying, Decision tree works on the principal of the series of questions. A flowchart is produced to minimize our range of answers from which we work our way towards the prediction we want to make[26]. With no extra prior knowledge, Random forest learns about the framework of the required object through the help of provided dataset bit by bit and creates a flowchart, by the help of questions, where it tries to minimize the error percentage and give the best possible outcome. The model adapts any connections between the information (features) and the qualities we need to predict (target)[27].

Other advancement in this random forest algorithm has been done such as semi-supervised random forest, rotation forest, fuzzy random forest. There are many advantages of using this algorithm use of it in both the cases of Classification and Regression, then the way it handles the missing values in the datasets and finally the most important thing in this is the greater number of trees you put in, the better predictions you get. But there is a big disadvantage, i.e., over fitting arises very easily, and very difficult to determine too.

Sequential Minimal Optimization (SMO) is a machine learning algorithm that is used for training support vector machines. This algorithm was proposed by Platt. So, it requires the solution of large quadratic programming (QP) optimization problem that actually consumes a lot of computational time and also power for training a support vector machines. SMO works by breaking this large QP problem into series of smallest QP problems[28]. This helps in saving a lot of computational time which makes it easy for also to run large datasets in the SVMs. In the case SMO, it at first divides the QP problem into several sub-problems and in every step, it aims to solve the smallest possible optimization problem. Using the two lagrange multipliers in every small step, this SMO technique avoids completely the original path for solving the large QP problem. Finally, after the process is done the SMO uses the Osuna's theorem to ensure the convergences of the several problems are done perfectly[29].

It can even work on large datasets as it contains many optimized designs within it. By evaluating SVM, SMO computational time can be dominated. This algorithm has a good speed and is faster than any other algorithms.

C4.5 is an extension of the ID3 (Iterative Dichotomiser 3). It is similar to other algorithms generating decision tree. It was at first developed by Ross Quilan (Developer of ID3)[30]. Sometimes, the C4.5 is also known to be as statistical classifier because of the reason that the generated decision trees are used for the classification. This uses the concept of information entropy to train the model. The measuring of the information entropy has been associated with each possible data value in the dataset and the negative logarithm of the probability mass function for the data value[31].

This algorithm has improved much over its precursor such as with each attribute and also finds the normalized information gain ratio on splitting it. Then, it also selects the highest information gain attribute on basis of which the decision node is split[32].

Training Model

The training model is prepared by the 80% of the original dataset. The dataset is completely classified from where the computer learns and finds the relations

among various attributes. The cross-validation is used along with the algorithm to train the model. In this case, suppose the cross-validation is n set with n-folds, then it will divide the training dataset into n parts, then the n-1 parts will be used as training data and the other one will be used to validate the rest. This process of iteration goes on for n iteration times. In this research, we have used 10-fold and it is chosen as per the size of the dataset[33,34].

Generally, the datasets containing binary classification on the basis of several attributes are imbalance, so as in this case, the same has seemed to be observed. These imbalance datasets are not possible to be handled by the normal classifiers since they give importance to each of the attribute equally which mead lead misclassification errors cost equally and because of which the accuracy might get decreased of the trained model[35,36]. So, in this case, we have used the misclassification cost where the trained model become cost sensitive and able to find the lowest expected cost. This use is actually much randomized because it neither depends upon the number of attributes neither on the minority class ration rather it depends on the base classifier[37].

Here, we have two methods to introduce the misclassification cost with the imbalance dataset. The first method is to make the classifying the algorithm into the cost-sensitive one and proceed with the rest settings[21,38]. The other is the use of a wrapper which helps in the base classifiers into cost sensitive ones. This second one is mainly known as the Meta Learning. In this case, at first using bootstrap aggregating on the decision trees, it estimates the reliable probability[39]. On the basis of that, it relabels the attributes in the training model. Then, these are used in building cost-insensitive classifier. This also helps in avoiding the algorithm to over fit in the dataset and even it reduces the variances of the dataset. Among all the classifier, we have taken the use of Meta Learning in only C4.5 because for two main reasons: one is that it tends to get over fit with the datasets sometimes and other is that the Meta Cost works best with the unprimed trees[39].

In the case of Naïve Bayes, Random Forest and SMO, we have used the CostSensitiveClassifier which uses the cost insensitive algorithm to predict the probability estimations of the test instances and then using this it predicts class labels for the examples of the test dataset. In our report, we have classified our datasets into two classes i.e. active and inactive. So, we use the 2X2 matrix which is generally used for the binary classification. In the matrix sections are True Positives (Active classified as Active), False Positives (Inactive classified as Active), False Negatives (Active classified as Inactive) and True Negatives (Inactive classified as Inactive). In this case, the percent of False Negatives are more important than the percent of False Positives and the upper limit for False Positives were set to 20%[40]. In this process, we increase the misclassification up to the set percent which also help in the increasing of the True Positives.

Independent Validation

There are various methods for the validation of the binary classifiers. The True Positive Rate is the ratio of the actual actives to the predicted positives and this can be obtained as (TP/TP+FN). The False Positive Rate is the ratio of the predicted false actives to actual inactives and this can be obtained as (FP/TN+FP). Accuracy shows the model's performance relative to the real values and this can be calculated as (TN+TP/TN+TP+FP+FN). The Sensitivity shows the model's ability to identify the positive results and this is calculated as (TP/FN+TP) and the Specificity shows the model's ability to identify the negative results and this is calculated as (TN/TN+FP)[41]. A model with high specificity and sensitivity has a low error rate. The Balanced Classification Rate (BCR) is the mean of the sensitivity and specificity which provides the accuracy of the model applied on the imbalanced dataset. This BCR can be calculated as 0.5*(specificity+sensitivity). Apart from the BCR, the Mathews Correlation Coefficient (MCC) is also used whose range varies from -1 to 1. The Receiver Operating Characteristic (ROC) curve is the visualization of the ratio of FPR to TPR. In this case, the FPR is placed on the x-axis and the TPR on the y-axis. The Area under curve shows the probability prediction of the classifier and its ability to classify the randomly chosen instance into the correct class[42].

Docking of the Predicted Drugs

Around 280 drugs were predicted by our machine learning model which can be effective for the treatment of Chagas Disease. The two available drugs i.e., Benzimdazole and Nifurtimox for this disease were predicted accurately by our model with a confidence of 80%[43,44].

The predicted compounds with above 80% of confidence were docked using AutoDock4.2. It is the molecular modeling simulation software and also it is one of the most cited software in this area. It is most effective for the study of protein interactions with other compounds. This software is maintained mainly by the Scripps Research Institute[45–47].

Then, after the docking, the most important is the visualization. The visualization is carried using the PyMOL software. In this case, we have looked upon the pictorial representation of each docked compounds[48–50].

RESULTS AND DISCUSSIONS

In this research, we have at first taken the inhibitors of the *T.cruzi*, which doesn't allow them to replicate in the host. These were used as the main component for the modelling of the training model using Machine Learning. The 914 attributes were taken under consideration for more than 10,000 compounds. Here, we have not used unsupervised learning to filter out the dataset, because it might make the dataset much weaker than creating it better. As mentioned, we have used for Classifying algorithm to train the model and the best among them was further used for the testing and predicting of the drugs from the DrugBank.

Among the four algorithms C4.5 was the best against the dataset. It has shown accuracy of approximately 64.96%. Against this model, we have used the drugs from the DrugBank to get predicted for the identification of the potential drugs which can be used for the treatment of Chagas Disease. Then following this algorithm, the SMO algorithm comes with the 63.72% accuracy when tested against the test set which is 20% of the full dataset. After the SMO, the algorithm Random Forest and Naïve Bayes follow this testing pattern with 60.48% and 58.17% respectively.

The C4.5 algorithm has presented around 453 False Positives and 733 True Positives when tested with the test set of 2164 compounds. The C4.5 has also the highest MCC with 0.282 and maximum BCR i.e., 64.63%. As per the independent validation, the algorithm with the lowest possible False Positives and highest possible True Positives can be said to be the most effective model for the prediction of the drugs from the DrugBank. In all the cases, the compounds for the False Positives were set to 23%. When these results are compared with the rest of the dataset, it is found to be way better because it has satisfied both criteria and the rest of the algorithms has not reached the marked that has been achieved by C4.5.

The model created with C4.5 algorithm predicted around 280 drugs out of all the 2388 approved drugs. These 280 drugs contain all the drugs which are under investigational and also the approved ones for the treatment of Chagas Disease.

Here in this research, we have predicted several drugs and out of which, we have docked and suggested the top 10 drugs keeping align both the machine learning accuracy and the docking result. The docking results has kept in mind binding energy and its effectiveness to bind with the compound.

At first, we have docked the approved drug i.e., Benzimidazole with the predicted structure of Protein Cruzain. This predicted structure is taken from the FASTA Sequence availed from the Protein Database of NCBI and prediction of 3D modelling of Protein was carried by the Swiss Model. Benzimidazole drug is used to treat infection caused by a protozoan parasite Trypanosoma cruzi (*T.cruzi*). This drug has been approved and is easily available in market. The docking of the Cruzain protein with the approved drug Benzimidazole (DrugBank ID – DB11989), has a binding energy: -4.47eV is revealing that it had numerous steric clashes with the adjoining strands and thus highlighting its potential to inhibit Chagas disease. Overall, the docking analysis suggests that the inhibitor effectively hinders the binding of strands, thus inhibiting the disease.

In keeping view to this docking of the present drug, we have docked several other approved drugs available in the DrugBank and which are predicted by our Machine Learning with a Confidence Level of above 80%. With the reference to that we have found the best drug predicted by the model with the inclusive of all the parameters taken under consideration for the machine learning prediction is the Lomitapide (Drug Bank ID – DB08827). This drug has around binding energy of -9.23 eV which reveals that it had more steric classes than the present drug for the treatment of Chagas Disease with the adjoining strands and this highlights its potential effectiveness towards the treatment of Chagas Disease if carried.

Along with the Lomitapide, we have suggested 9 other compounds which can be more effective than the Benzimidazole.

CONCLUSION

Today, it takes more than 15 years to bring a drug from the investigational stages to market availability. It is because of the trial-end error process or the so-called Edisonian Approach, where we went on trying several compounds to find the best possible one. These days with the inclusion of Artificial Intelligence, this time span has been reduced to a great extent and people are able to approach in a rational manner for the drug discovery process. In this research paper, we have targeted for the new drugs for the treatment of Chagas Disease. At first, we started with the inhibitors for the *T.cruzi*, then with respect to that we found several drugs that can be used as an inhibitor predicted by our Machine Learning Model. These drugs were docked to validate their accuracy and effectiveness.

In this process, we have found that machine learning model created on the basis of the C4.5 algorithm is the most effective one with the accuracy of almost equal to

65%. The drugs predicted by this model has shown immense effectiveness which was shown by the docking process. So, we can expect that the drugs predicted and validated by this paper theoretically, would come soon to the investigational process and in future could show immense potential for treatment of Chagas Disease.

REFERENCES

1. Matza, D. & Foucault, M. Madness and Civilization: A History of Insanity in the Age of Reason. *Am. Sociol. Rev.* (1966). doi:10.2307/2090782
2. Prata, A. Clinical and epidemiological aspects of Chagas disease. *Lancet Infectious Diseases* (2001). doi:10.1016/S1473-3099(01)00065-2
3. Coura, J. R. & Borges-Pereira, J. Chagas disease: 100 years after its discovery. A systemic review. *Acta Tropica* (2010). doi:10.1016/j.actatropica.2010.03.008
4. Andrade, D. V., Gollob, K. J. & Dutra, W. O. Acute Chagas Disease: New Global Challenges for an Old Neglected Disease. *PLoS Negl. Trop. Dis.* (2014). doi:10.1371/journal.pntd.0003010
5. Marin-Neto, J. A., Cunha-Neto, E., Maciel, B. C. & Simões, M. V. Pathogenesis of chronic Chagas heart disease. *Circulation* (2007). doi:10.1161/CIRCULATIONAHA.106.624296
6. Dias, J. C. P. Chagas disease (American trypanosomiasis). in *Arthropod Borne Diseases* (2016). doi:10.1007/978-3-319-13884-8_17
7. WHO. WHO | Chagas disease (American trypanosomiasis) Factsheet. *WHO* (2008). doi:10.1098/rspb.2009.1480
8. Carod-Artal, F. J. American trypanosomiasis. *Handb. Clin. Neurol.* (2013). doi:10.1016/B978-0-444-53490-3.00007-8
9. Barr, S. C. Canine Chagas' Disease (American Trypanosomiasis) in North America. *Veterinary Clinics of North America - Small Animal Practice* (2009). doi:10.1016/j.cvsm.2009.06.004
10. Wang, Y. *et al.* PubChem's BioAssay database. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gkr1132
11. Wishart, D. S. *et al.* DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* (2008). doi:10.1093/nar/gkm958
12. Agarwala, R. *et al.* Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1071
13. Liu, K., Feng, J. & Young, S. S. PowerMV: A software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. *J. Chem. Inf. Model.* (2005). doi:10.1021/ci049847v
14. Biasini, M. *et al.* SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gku340
15. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* (1997). doi:10.1002/elps.1150181505
16. Kiefer, F., Arnold, K., Künzli, M., Bordoli, L. & Schwede, T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* (2009). doi:10.1093/nar/gkn750
17. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* (2003). doi:10.1093/nar/gkg520
18. Mitchell, T. M. *Machine Learning. Annual Review Of Computer Science* (1997). doi:10.1145/242224.242229
19. Spring, M. L. *Machine Learning in Action. ... for Engineering and ...* (2015). doi:10.1007/978-0-387-77242-4
20. Ng, A. 1. Supervised learning. *Mach. Learn.* (2012). doi:10.1111/j.1466-8238.2009.00506.x
21. Wang, J., Zhao, P. & Hoi, S. C. H. Cost-Sensitive Online Classification. *IEEE Trans. Knowl. Data Eng.* (2014). doi:10.1109/TKDE.2013.157
22. Friedman, N. *et al.* Bayesian Network Classifiers *. *Mach. Learn.* (1997). doi:10.1023/A:1007465528199
23. Rish, I. An empirical study of the naive Bayes classifier. *Empir. methods Artif. Intell. Work. IJCAI* (2001). doi:10.1039/b104835j
24. Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. *Proc. 23rd Int. Conf. Mach. Learn.* (2006). doi:10.1145/1143844.1143865
25. Breiman, L. Random Forests. *Mach. Learn.* (1999). doi:10.1023/A:1010933404324
26. Segal, M. R. Machine Learning Benchmarks and Random Forest Regression. *Biostatistics* (2004).
27. Breiman, L. Random forests. *Mach. Learn.* (2001). doi:10.1023/A:1010933404324
28. Platt, J. C. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Advances in kernel methods* (1998). doi:10.1.1.43.4376
29. Cortes, C. & Vapnik, V. Support vector machine. *Mach. Learn.* (1995). doi:10.1007/978-0-387-73003-5_299
30. Quinlan, J. R. *C4.5: Programs for Machine Learning. Morgan Kaufmann San Mateo California* (1992). doi:10.1016/S0019-9958(62)90649-6
31. Quinlan, J. R. Bagging, boosting, and C4.5. *Proc. Thirteen. Natl. Conf. Artif. Intell.* (2006). doi:10.1212/NXI.0000000000000092
32. Ruggieri, S. Efficient C4.5. *IEEE Trans.*

Knowl. Data Eng. (2002). doi:10.1109/69.991727
33. Browne, M. W. Cross-validation methods. *J. Math. Psychol.* (2000). doi:10.1006/jmps.1999.1279
34. Refaeilzadeh, P., Tang, L. & Liu., H. 'Cross-Validation.' in *Encyclopedia of database systems* (2009). doi:10.1007/978-0-387-39940-9_565
35. Thai-Nghe, N., Gantner, Z. & Schmidt-Thieme, L. Cost-sensitive learning methods for imbalanced data. in *Proceedings of the International Joint Conference on Neural Networks* (2010). doi:10.1109/IJCNN.2010.5596486
36. Sun, Y., Kamel, M. S., Wong, A. K. C. & Wang, Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.* (2007). doi:10.1016/j.patcog.2007.04.009
37. Kukar, M. & Kononenko, I. Cost-sensitive learning with neural networks. *13th Eur. Conf. Artif. Intell.* (1998). doi:10.1.1.13.8285
38. Sen, P. & Getoor, L. Cost-sensitive learning with conditional Markov networks. *Data Min. Knowl. Discov.* (2008). doi:10.1007/s10618-008-0090-5
39. Domingos, P. MetaCost: A General Method for Making Classifiers Cost-Sensitive. *Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* (1999). doi:10.1145/312129.312220
40. Masnadi-Shirazi, H. & Vasconcelos, N. Cost-sensitive boosting. *IEEE Trans. Pattern Anal. Mach. Intell.* (2011). doi:10.1109/TPAMI.2010.71
41. Jamal, S. & Scaria, V. Cheminformatic models based on machine learning for pyruvate kinase inhibitors of Leishmania mexicana. *BMC Bioinformatics* (2013). doi:10.1186/1471-2105-14-329
42. Periwal, V., Kishtapuram, S. & Scaria, V. Computational models for in-vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. *BMC Pharmacol.* (2012). doi:10.1186/1471-2210-12-1
43. Bermudez, J., Davies, C., Simonazzi, A., Pablo Real, J. & Palma, S. Current drug therapy and pharmaceutical challenges for Chagas disease. *Acta Tropica* (2016). doi:10.1016/j.actatropica.2015.12.017
44. Romanha, A. J. *et al.* In vitro and in vivo experimental models for drug screening and development for Chagas disease. *Mem. Inst. Oswaldo Cruz* (2010). doi:10.1590/S0074-02762010000200022
45. Garrett M. Morris, David S. Goodsell, Michael E. Pique, William "Lindy" Lindstrom, Ruth Huey, Stefano Forli, William E. Hart, Scott Halliday, R. B. and A. J. O. AutoDock Version 4.2. *Citeseer* (2012).
46. Huey, R. & Morris, G. Using AutoDock 4 with AutoDockTools: A Tutorial. *Scripps Res. Institute, USA* (2008).
47. Trott, O. & Olson, A. J. Autodock vina. *J. Comput. Chem.* (2010). doi:10.1002/jcc
48. DeLano, W. L. The PyMOL Molecular Graphics System, Version 1.8. *Schrödinger LLC* (2002). doi:citeulike-article-id:240061
49. Objects, E. M., Torsions, R. B., Scenes, V., Shows, C. & Coordinates, S. Introduction to PyMOL Introduction to PyMOL. *Proteomics* (2010). doi:10.1213/ANE.0b013e3181e9c3f3
50. Stockwell, G. PyMOL tutorial. *Biochemistry* (2003).

FIGURES AND TABLES

ALGORITHMS	SPECIFICITY	SENSITIVITY	ACCURACY	BCR	MCC	ROC
SMO	56.59%	71.3%	63.72%	63.94	0.282	0.639
RANDOM FOREST	56.68%	64.53%	60.48%	60.6	0.213	0.663
NAÏVE BAYESIAN	55.96%	60.53%	58.17%	58.24	0.165	0.642
C4.5	59.37%	69.9%	64.96%	64.63	0.294	0.697

Table No.1 – Shows the comparison of the effectiveness of the applied Algorithms for the preparation of Machine Learning Model.

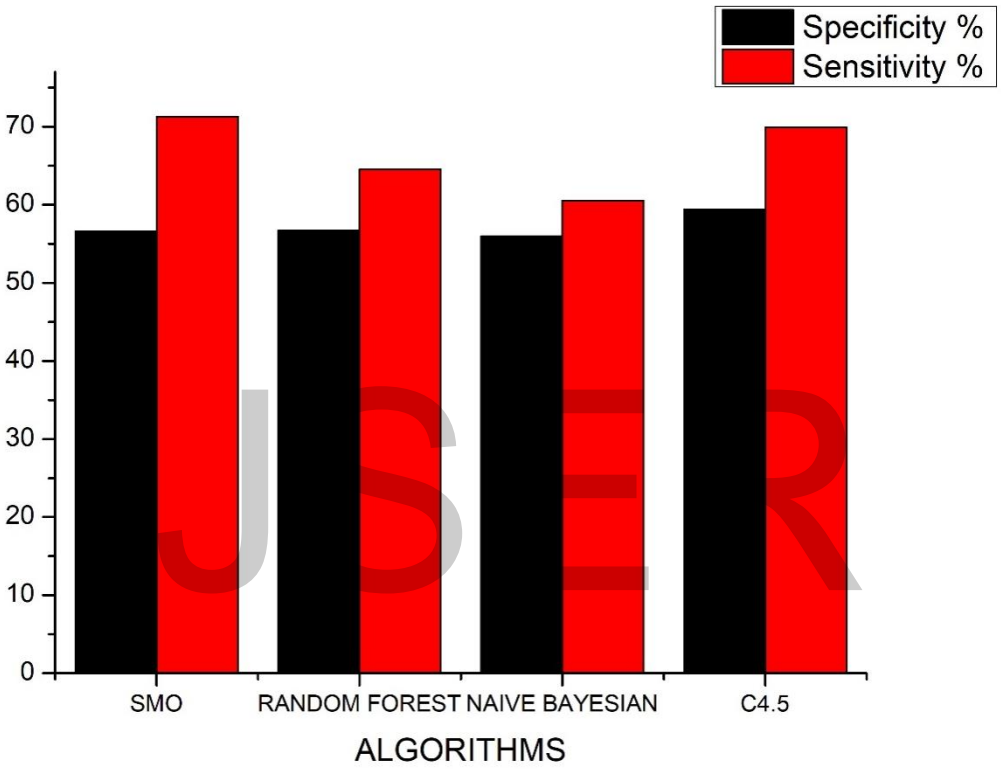

Fig 1 – Shows the Comparison of Algorithm's Specificity and Sensitivity used for the Preparation of ML Model.

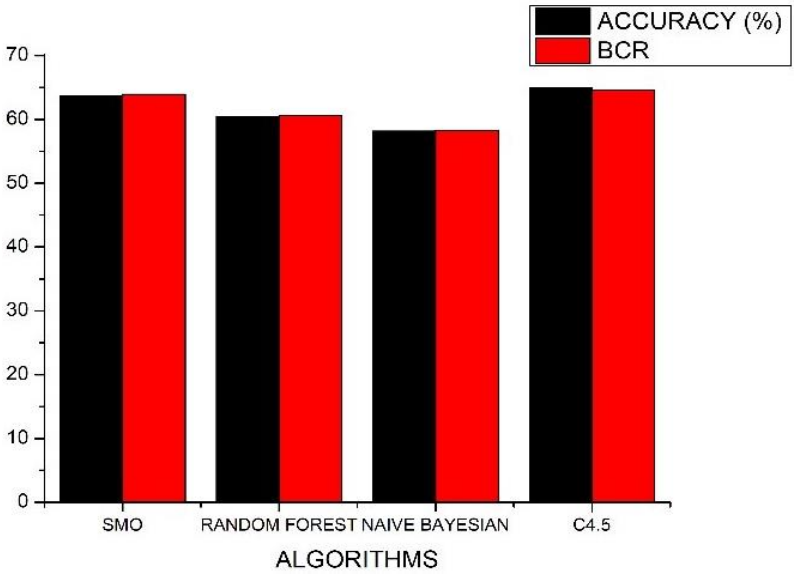

Fig 2 – Shows the Comparison of Algorithm's Accuracy and BCR used for the Preparation of ML Models.

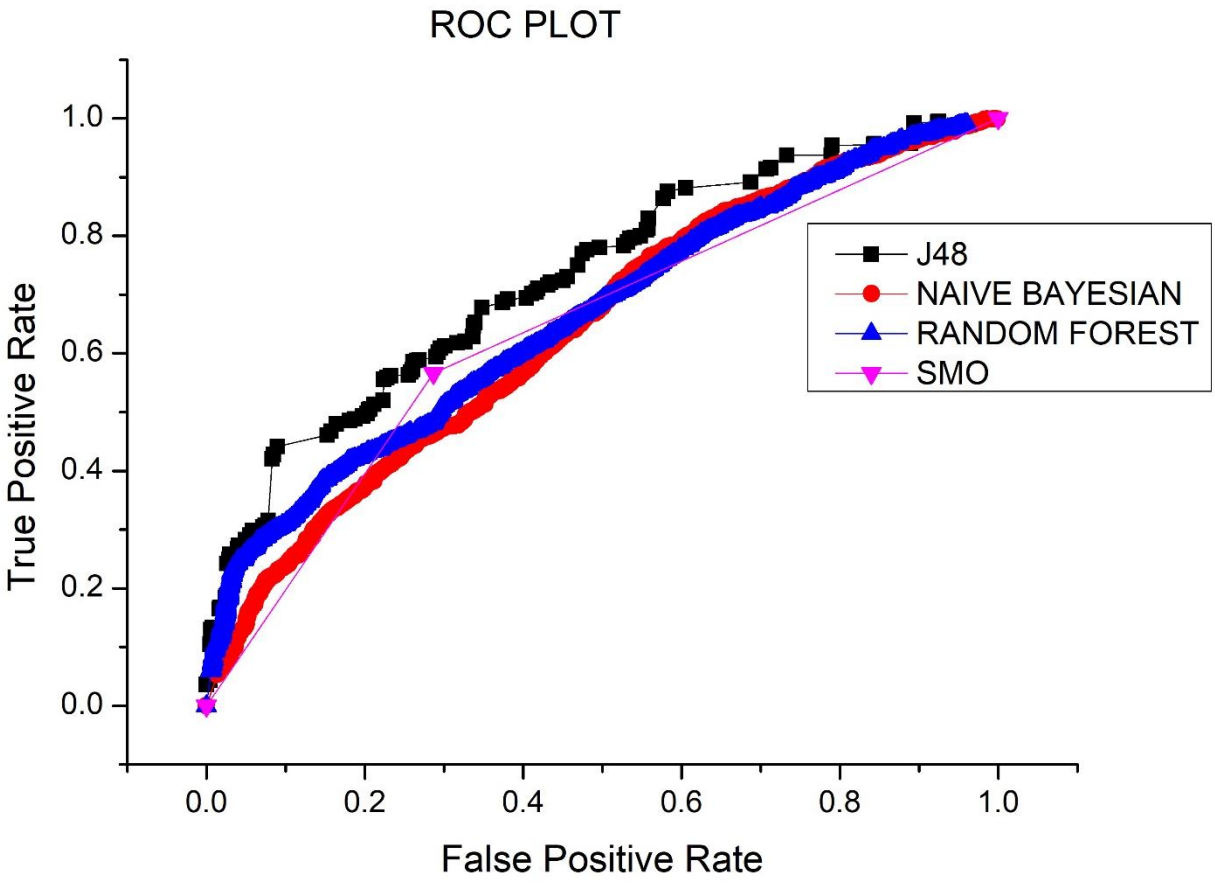

Fig-3: Shows the comparison of the False Positive Rate and True Positive Rate for all Algorithms used in case of preparation of Machine Learning models.

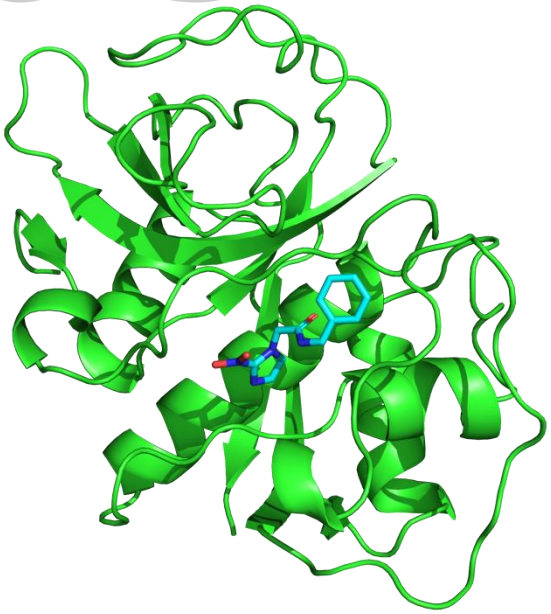

Figure – 4 - The docking of the Cruzain protein with the approved drug Benzimidazole (Drug Bank ID – DB11989) with binding energy of -4.47eV.

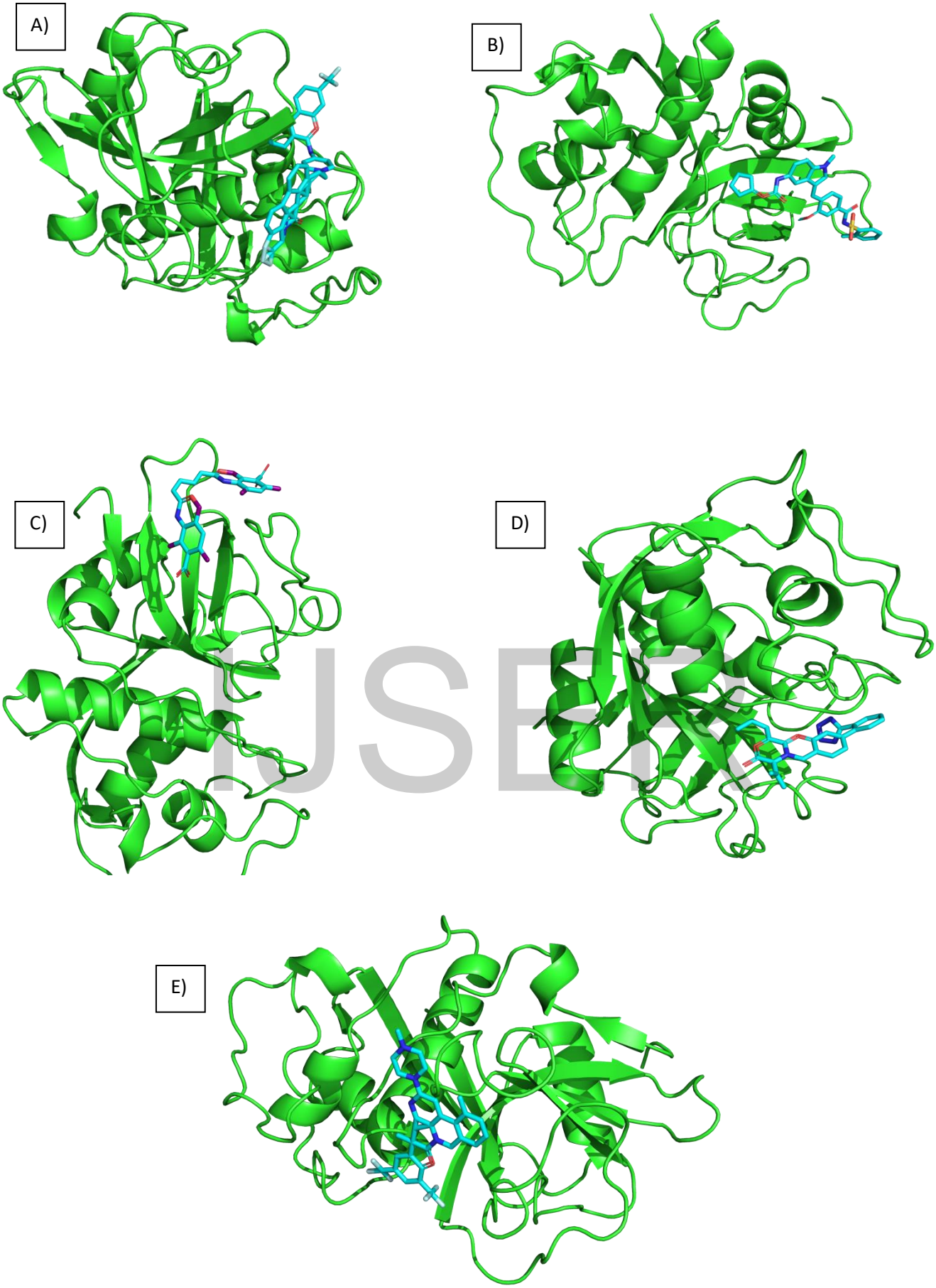

Figure – 5– Representation of the docking of Cruzain proteins with the Predicted Drugs from the Machine Learning Model. A) Lomitapide(DrugBank ID – DB08827) B) Zafirlukast (DrugBank ID – DB00549) C) Lodipamide (DrugBank ID- DB04711) D) Salmon Calcitonin (DrugBank ID – DB00017) E) Netupitant (DrugBank ID – DB09048.