**Compound Similarity Prediction**

**Using Machine Learning**

1,3M.JYOTHSNA, 1,4K.ASHOKKUMAR, 1,5P.JAHNAVI, 1,6S.NIKHILA

2,7 P. SRINIVAS RAO

1 MVGR College of Engineering, Chinthalavalasa, Vizianagaram

2Associate Professor, CSE Department, MVGR College of Engineering, Chinthalavalasa, Vizianagaram

3[jyothsnamukhee03@gmail.com](mailto:jyothsnamukhee03@gmail.com), 4deva4926464@gmail.com, 5[jahnavipathivada123@gmail.com,](mailto:jahnavipathivada123@gmail.com,1d) 6nikhilasweety762@gmail.com.

7 psr.sri@gmail.com.

**ABSTRACT**

In the field of Medicine, Bio Technologies and Pharmacology, The drug industry is also one of the major player that guides the development. Drug discovery is a process where all the drugs are discovered and designed with some similarities. It is a process which aims at identifying a compound therapeutically useful in curing and treating disease. Development of pipelines and the Drug Discovery can be said to be as a long, complex and also depend on various factors. Drug designers mine chemical data from large databases to extract chemical compound that becomes the lead compound in drug discovery. The drug target designing combine machine learning and deep learning algorithm that improve the quality of drugs discovered. Drug discovery involves seven step process that includes disease selection, target hypothesis, lead identification, lead optimization, pre- clinical trial, clinical trial, pharmacogenic identification. Machine learning can be applied to identify the drug targets and in optimization of lead compound. Both supervised and unsupervised algorithms when applied to the databases increases the efficacy of identifying a new target and optimize the lead compound. The process failure rate when machine learning is applied will be very low compared to the traditional drug discovery process. Here we present a Compound Similarity Prediction by using k-means clustering and using dbscan.

Through this compound similarity prediction, the compounds are clustered according to their similarity. So, if the reaction of the new compound is unknown, we could know its reaction through this compound similarity prediction. The main advantage of compound similarity prediction is to know the reaction of the new compound. If the reaction of the compound is known then we can speed up the process of drug discovery.

**Keywords:** Data Mining, Clustering, Density Based Clustering, DBSCAN, K-Means,

Similarity Prediction, Lead compound Identification

**INTRODUCTION**

Drug discovery, which is a complicated process where it involves an investment of huge time and money. On average, it takes 6 to 12 years with an investment of 500 million to 1 billion (in US dollars) to identify a drug for fighting against a target. After that, even we struggled a lot, the success rate is very low. Many long-term research projects may end up fruitless resulting in wastage of enormous efforts. Blockbuster drugs are the drugs that are prescribed for the common medical problems like cold, diabetes, high blood pressure, asthma and flu. These acts as a very profitable in the pharmaceutical type of industries. They bring revenues greater than 1 billion per year and a profit of more than 1 million a day (in dollars). However, it can also result in problems for the company if the drug shows any side effects. Generally, the drugs also have patents that will expire resulting in competition from less cost or expensive equivalents. So we can say that the process of drug discovery is very complicated and high risk activity. But this is motivated always with the benefits it could do to millions of people suffering from various diseases [1][3].

To compare the compound similarities it is mandatory to generate smiles [2][4] and from smiles we produce chemical fingerprint generation [6][5] using jupyter notebook and we find similarity between compounds and based on similarity we perform clustering. Drug discovery using traditional methods takes years to discover a new drug [7][9]. Using machine learning algorithms, we can reduce the time taken to discover a new drug. This motivates us to take drug discovery project. All know that the biomedical data are very complex, by using algorithms to design new drugs makes more possible than it has ever been. Machine learning can enhance many stages of the drug discovery process [10][11] preliminary but crucial stages including designing a drug’s chemical structure. By investigating the drug’s effect, both the basic preclinical research and clinical trials, biomedical data is produced a lot. The newly finded patterns in those kind of data can be facilitated using machine learning. There are different kinds of data, including genetic and imaging ones[12][15]. These kind of data can be analyzed using machine learning and also further used to build the great novel solutions for drug discovery. First of all, we need to know the reaction of the new compound in order use that compound in any process so Compound similarity prediction is used to speed up the process.

The different compounds are taken and generated smiles from the structure of the compound. Now similarity is found among the smiles. The SMILES whose abbreviation is Simplified Molecular-Input Line-Entry System is a specification in the form that is of a line notation for describing the structure of chemical series using short ASCII strings. SMILES strings can be imported by most molecule editors for conversion back into two-dimensional drawings or three- dimensional models of the molecules[13][16]. Commonly, a number of valid SMILES strings which are equally, can be written for a molecule. For example, CCO, OCC and C(O)C all specify the structure of ethanol. Algorithms have been developed to generate the same SMILES string for a given molecule; of the many possible strings, these algorithms choose only one of them. This SMILES column is unique for each structure of drug, although its dependent on the canonicalization algorithm which is used to generate it, and is termed as the canonical SMILES. All these algorithms first convert the SMILES to an internal representation of the molecular structure; an algorithm then examines that structure and produces a unique SMILES string. Various algorithms for generating canonical SMILES have been developed and include those by Daylight Chemical Information Systems, OpenEye Scientific Software, MEDIT, Chemical Computing Group, MolSoft LLC, and the Chemistry Development Kit[14][19].

**RELATED WORK**

Hakime Ozturk et al[1][3], explained an approach on offensive content clustering in compound similarity in Drug Discovery using Machine learning. the Information about the design and discovery of drugs with the column called SMILES. SMILES(Simplified Molecular Input Entry Specification) consists of all the chemical formula of each and every compound. In this agglomative clustering, is used to combine the similar compounds. With this review, they have summarized the impact of NLP on bio/cheminformatics to encourage this already interdisciplinary field to take advantage of recent advances. Hakime Ozturk et al[2][4], Explained the study of SMILES-based Compund similarity functions for drug-target interaction prediction. By using SMILES, Molecular structures can be represented as strings of special characteristics . Since each molecule is represented as a string, the similarity between compounds can be computed using SMILES-based string similarity functions. The 2D-based compound similarity kernels such as SIMCOMP was used by most previous studies on drug-target interaction prediction. Martin Vogt et al[5][7] Described Python package for modelling Tanimoto similarity

value distributions”. It described the information about how to use the ccbmlib Python package which is a collection of modules for modelling similarity value distributions. From this paper invoked that this is used to assess the statistical significance of tanimoto coefficients and evaluate how molecular similarity is reflected when different fingerprint representations are used and conditional significance score to estimate where a test compound would be ranked in a search and the resulting models have been evaluated for RDKit fingerprints, taking a collection of ChEMBL compounds as a reference data set. B. Zagidullin et al[6][8] describes the Application of machine and deep learning methods in drug discovery and cancer research has gained a considerable amount of attention in the past years. As the field grows, it becomes crucial to systematically evaluate the performance of novel computational solutions in relation to established techniques. To this end, in prediction of drug combination sensitivity and drug synergy scores we compare rule-based and data-driven molecular representations using standardized results of 14 high-throughput screening studies, consisting 64 200 unique combinations out of 4153 molecules tested in 112 cancer cell lines. By adapting the Centered Kernel Alignment metric,we quantify the similarity of molecular representations and evaluate their clustering performance . Our work demonstrates that to identify an optimal molecular representation type, it is necessary to supplement quantitative benchmark results with qualitative considerations, such as model interpretability and robustness, which may vary between and throughout preclinical drug development projects. B. Zagidullin et al[9][11] compare rule-based and data-driven molecular representations in prediction of drug combination sensitivity and finding drugsynergy scores using standardized results and later on to evaluate the clustering performance of molecular representations and quantify their similarity by adapting the Centred Kernel Alignment metric (identify an optimal molecular representation type (CKA)). Daniel Probst et al[10][12] explained that MAP4 is a new molecular fingerprint suitable for drugs and can be adopted as a universal fingerprint to describe and search chemical space so here it takes MinHasing. Ravi Manne et al[13][15] examined machine learning and deep learning techniques which helppharma industry in all stages of drug discovery which includes target validation, prognostic bookmarks, technical traits, in which these are totally seven phases in the process of drug discovery.Suresh Dara et al[14][16] Described the feasible literature on drug discovery through ML tools and techniques that are enforced in every phase of drug development to accelerate the research process.

Target validation, prognostic biomarkers, digital pathology are considered as the problem statements where In clinical trials [17][19] , absolute and methodological data must be generated to tackle many puzzles in verifying Machine Learning techniques, , promoting awareness in ML approaches, improving decision-making and also can reduce the reduce the risks in drug discovery.

**METHODOLOGY**

As the problem statement is defined, to predict the compounds which are similar, can be estimated using clustering algorithms such as K means clustering and Hierarchial Clustering.

**3.1 k-means clustering** -

A) Procedure of K-mean Algorithm K-mean distributes all objects to K number of clusters at random;

1. Let us choose or assume the number of clusters as K and obtain the data points of that clusters.

2. Let we place or let we assume that c1, c2, ..... ck as centroids randomly.

3. Repeat the steps 4 and 5 again until the convergence occur or until the end of a fixed number of iterations that was given.

4. for each data point xi:

- find the nearest centroid of all the clusters (c1, c2 .. ck)

- assign the point to that cluster which is exactly or almost in center.

5. for each cluster in j = 1..k

- new centroid = mean (average distance) of all points that are assigned to that cluster

6. End or stop the algorithm.

E = ∑i=0k ∑ p ∊Ci | p - mi|2

Here, E is said to be as the total square error of all the objects in that particular data cluster, where p is given data object, where mi is the mean (statistical mean) value of cluster named Ci (note : p and m both are multi-dimensional).

Tanimoto are the values which represents the percentage of compound similarity with the generated fingerprints.

**3.2 Hierarchical clustering** - A classification method that builds a hierarchy of clusters by agglomerative clustering e.g., merging smaller clusters or divisive clustering e.g., splitting a large cluster to smaller ones. Hierarchical clustering is also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters.

Algorithm:

1. Begin (Start) the initialization of clusters c, c1 = n, Di = {xi}, i = 1,…,n ‘
2. Do c1 = c1 – 1
3. Find nearest clusters which are close to each other, say, Di and Dj
4. Merge (Combine both clusters into one) Di and Dj.
5. Do same process until the clusters got equal c = c1
6. Return all those c clusters
7. End or stop the algorithm.

This algorithm is always starts with any no. of clusters more than 1 , lets say n clusters initially where each and every data point is a cluster ( n data points are considered as n clusters). By reaching each iteration, the number of clusters are reduced by merging or combining by 1 as the 2 nearest clusters get merged. This process continues until the number of clusters reduces to the predefined value or the number of clusters reduces to one cluster.

**3.3 Principal component analysis** - A statistical method that uses orthogonal procedure to transform a set of correlated features to new independent variables called principal components. Principal Component Analysis (PCA) is definied as an unsupervised machine learning technique that attempts to derive a set of low-dimensional set of features from a much larger set while still preserving as much variance as possible. Also we can see the two main applications of PCA, those are: Variable selection and Visualizing High-Dimensional. Independent component analysis is a statistical method which separates a multivariable output into statistical independent additional components. Independent Component Analysis (ICA) is a machine learning technique to separate independent sources from a mixed signal. Unlike principal component analysis which focuses on maximizing the variance of the data points, the independent component analysis focuses on independence, i.e., independent components.

**Evaluation Metrics:**

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

So, to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

4.1 STANDARDIZATION

Standardization can be measured by subtracting the mean, after that divide the result with standard deviation for each and every value of every variable.

Z = (Value - Mean ) / Standard Deviation

all the variables are transformed to the same scale , Once the standardization is completed,

4.2 COVARIANCE MATRIX COMPUTATION

It can be used to understand how the variables of the input data set are different from the mean with respect to each other,

in alternative words, to examine if there's any relationship between them. as a result of generally, variables area unit extremely correlate in such the way that they contain redundant info.

So, so as to spot these correlations, we compute the variance matrix.

given that the sample covariance matrix of X is

Note: S ∈ R p×p and for each 1 ≤ j, k ≤ p

Sj,k = (1/n) ∑ n i=1 xij xik = S(xj , xk) is the sample covariance of features j and kn

The sample covariance matrix of X is given by S = (1 / n) Xt X = (1 /n) ∑i=1  Xi Xit

Note: S ∈ R p×p and for each 1 ≤ j, k ≤ p

Sj,k = (1/n) ∑ n i=1 xij xik = S(xj , xk) is the sample covariance of features j and k

4.3 COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS

Principal components are new variables which can be built as linear mixtures or combinations of the intial variables. These mixtures are carried out in this sort of manner that the brand new variables (i.e., Principal components ) are uncorrelated and maximum of the data withinside the initial variables is squeezed or compressed into the Principal components . So, the concept is 10- dimensional information offers you 10 Principal components , however PCA attempts to position most viable data withinside the first component, then most closing data withinside the 2nd and so forth till having some thing like proven withinside the screen plot below.

Determinent of a matrix = Product of Eigen Values

Trace of a matrix = Sum of Eigen Values

Ax = λx

λ = Eigen value

**5. SYSTEM MODEL**

Fig - 1

MACHINE LEARNING

TARGET

HYPOTHESIS

DISEASE SELECTION

LEAD-IDENTIFICATION

LEAD OPTIMIZATION

PHRAMACOGENIC

IDENTIFICATION

PRE- CLINICAL TRAIL

CLINICAL TRAIL

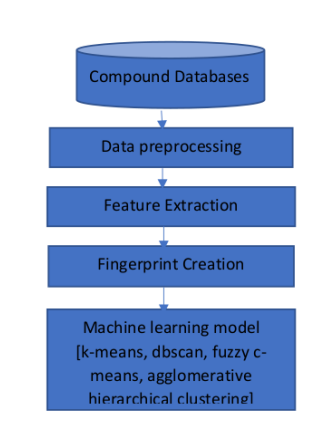
****

Fig - 2

Fig – 1 explaines:

1. Pre-Drug Discovery Process (Disease identification)

2. Modern Drug Discovery Process: The discovery process includes four important processes such as, target identification and validation, lead identification, lead optimization and pre-clinical trials.

a) Target Identification & Validation

It is the process of identifying target molecule which can be either protein or gene

b) Lead Identification

This process also helps to view which molecules were strongly binded to the target.

c) Lead Optimization

In this phase ,we discover the drug candidate from the lead identified compound. The goal is a process of improving the chemical structure of a confirmed. Hit to improve its drug characteristics.

d)Pre- Clinical Trial

an important phase to check whether the compound is working correctly or not

e) Clinical Trial

This is the primary phase which will be fastest as well as safest way to identify treatments.

Trials can be done in five ways such as, prevention trials, screening trials, diagnostic trials, treatment trial and quality of life trials.

**ENVIRONMENTAL SETUP**

**RESULTS**

Consider the Tanimoto values of all the five generated fingerprints.

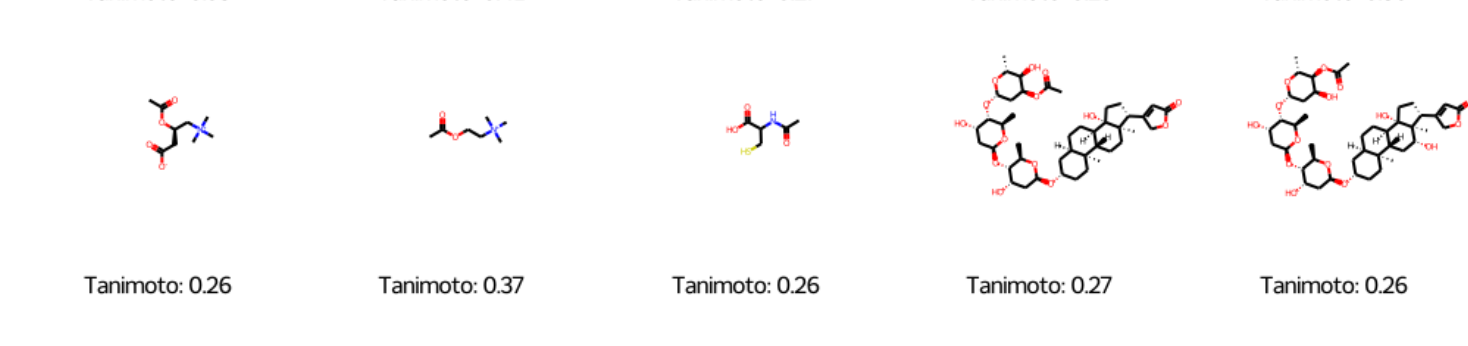


Fig – 3(a)

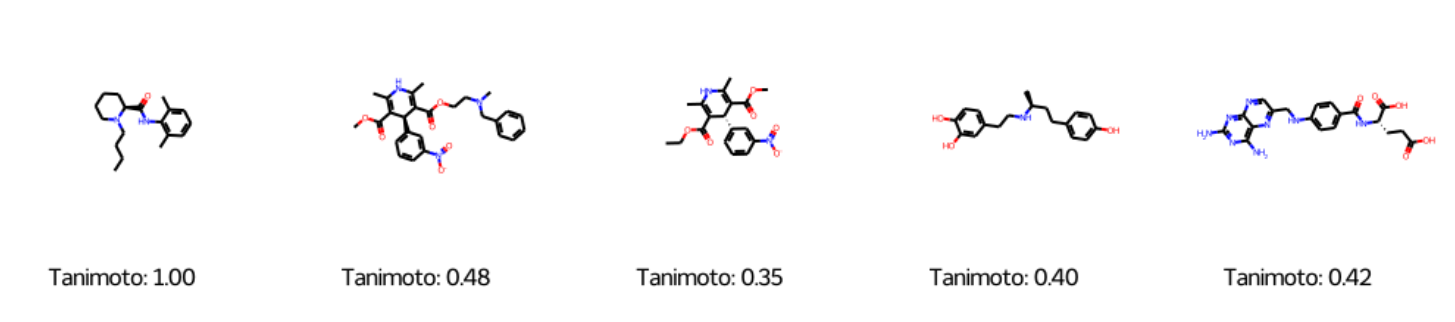


Fig – 3(b)

Fig - 4

Fig(4) is an Elbow method to get the number of clusters Efficiently. It gives a graph between Error and No. of Clusters.

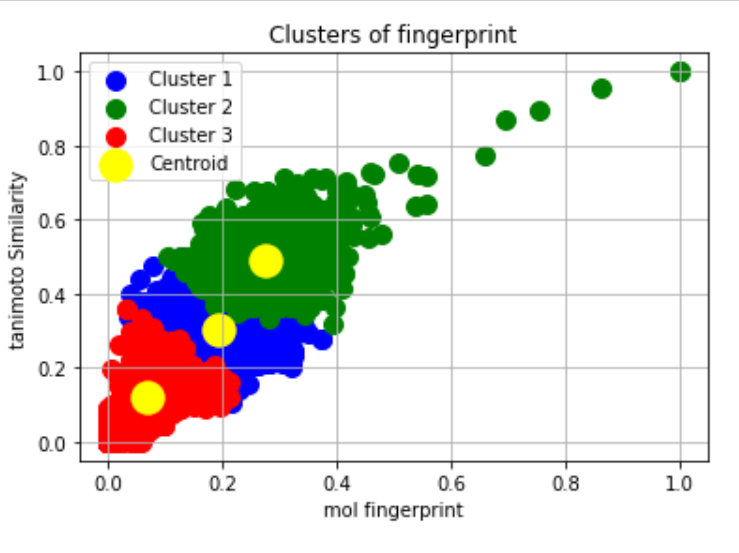


Fig - 5

Fig(5 ) shows Fingerprints are clustered, According to the fingerprints, similar compounds are got together.

Fig – 6(a)

Fig – 6(b)

Fig – 6(c)

**Conclusion and Future Scope**

Drug discovery is a long process and takes years to discover a new drug. So in order to reduce time we need machine learning algorithms to speed up the process of drug discovery. We also find many drugs but we don’t know the reaction of each drug. If we know the reaction of drug then we could save some time in discovering new medicine.so we have done compound similarity prediction. We used k-means clustering for compound similarity prediction and also, we used density-based clustering algorithm. Finally, we clustered the compounds based on their similarity. We have successfully implemented compound similarity prediction and finally we can know the reaction of the new drug. We also found the Tanimoto Values of compounds based on their similarity compared to generated fingerprints.

Based on the similar compounds, we can estimate the amount of drug required to a patient with certain conditions. Using classification methods one can separate the conditions and can estimate the drug. Ex: A patient with diabetes of age 55 is suffering from fever so he need 650 mg of DOLO.

A patient of age 16 (Healthy) is suffering from fever so he need 500 mg of DOLO.

According to the condition of the patient one can take the medicine.

**References:**

[1] Hakime Ozturka , Arzucan Ozgura , Philippe Schwallerb , Teodoro Lainob, Elif Ozkirimlic,d, “Exploring Chemical Space using Natural Language Processing Methodologies for Drug Discovery.”,arXiv peprint 2020.

[2] Hakime Öztürk, Elif Ozkirimli and Arzucan Özgür,“ A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction.” arXiv peprint, 2020.

[3] Pradeep Bheemavarapu, P S Latha Kalyampudi and T V Madhusudhana Rao, “An Efficient Method for Coronavirus Detection Through X-rays using deep Neural Network”, Journal of Current Medical Imaging,2020.

[4] S Satyanarayana ,“Privacy Preserving Data Publishing Based On Sensitivity in Context of Big Data Using Hive”,Journal of Bigdata(Springer), Volume:5,Issue:20, ISSN: 2196-1115, July 2018.

[5] Martin Vogt , Jürgen Bajorath Department of Life Science Informatics, B-I University of Bonn, Endenicher Allee ccbmlib – ” a Python package for modeling Tanimoto similarity value distributions [version 1; peer review: 2 approved] “ , 19c, Bonn, NRW, 53115, Germany,2019.

[6] B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen and J. Tang “Comparative analysis of molecular fingerprints in prediction of drug combination effects.“,brefings in bioinformatics,22(6),2021.

[7] P.Mahesh Kumar,P. Srinivasa Rao, “Frequent Pattern Retrieval on Data Streams by using Sliding Window”, EAI Endorsed Transactions on Energy web,Volume:5,issue:35,2021.

[8] T.V. Madhusudhana Rao, Suresh Kurumalla, Bethapudi Prakash, “[Matrix Factorization Based Recommendation System using Hybrid Optimization Technique](https://eudl.eu/doi/10.4108/eai.19-2-2021.168725), EAI Endorsed Transactions on Energy Web,Volume:5,issue:35,2021.

[9] Daniel Probst and Jean‑Louis Reymond “One molecular fngerprint to rule them all: drugs, biomolecules, and the metabolome Alice Capecchi.”, jouneral of cheminformatics,2020.

[10] Ravi Manne, ” Machine Learning Techniques in Drug Discovery and Development “,all research, 2021.

[11] T.V. Madhusudhana Rao, P Srinivasa Rao, P.S. Latha Kalyampudi, “Iridology based Vital Organs Malfunctioning identification using Machine learning Techniques”, International Journal of Advanced Science and Technology, Volume: 29, No. 5,PP: 5544 – 5554,2020.

[12] S.Vidya sagar Appaji, P. V. Lakshmi, P. Srinivasa Rao, “Maximizing Joint Probability in Visual Question Answering Models”, International Journal of Advanced Science and Technology Vol. 29, No. 3, pp. 3914 – 3923,2020.

[13] Suresh Dara1,Swetha Dhamercherla1,Surender Singh Jadav2,CH Madhu Babu1, Mohamed Jawed Ahsan3 , “Machine Learning in Drug Discovery: A Review”, springer Nature B.V. 2021.

[14] Geeks for geeks and tutorial point, Anaconda navigator installation jupyter notebook environment setup.

[15] Vidya sagar Appaji setti ,P Srinivasa Rao , “A Novel Scheme For Red Eye Removal With Image Matching”, Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 13-Special Issue, 2018.

[16] P Srinivasa Rao, Krishna Prasad, P.E.S.N, “A Secure and Efficient Temporal Features Based Framework for Cloud Using MapReduce”, springer, 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017),Volume:736,pp:114-123, ISSN 2194-5357 Held in Delhi,India, December 14–16, 2017.

[17] John B. O. Mitchell∗ “Machine learning methods in chemoinformatics”, WIREs Compute Mol Sci, 2014.

[18] N. Priya1 and G. Shobana2 “Application of Machine Learning Models in Drug Discovery: A Review”, Research Trend , 2019.

[19] Madhusudhana Rao, T.V., Srinivas, Y, "A Secure Framework For Cloud Using Map Reduce",Journal Of Advanced Research In Dynamical And Control Systems(IJARDCS), Volume:9,Sp-14,Pp:1850-1861,ISSN:1943-023x,Dec, 2017.

[20] P Srinivasa Rao,Sushma Rani N, “An Efficient Statistical Computation Technique for Health Care Big Data using R”, Scopus, IOP Conference Series: Materials Science and Engineering, Volume: 225, ISSN:1757-8981,ISSUE NO :012159,2017.

[21] M Ester, H-P. Kriegel. J. Sander, and X, Xu. 1996. “A density-based algorithm for discovering clusters in large spatial databases”. KDD’96.

[22] M.Parimala, Daphne Lopaz, N.C. Senthilkumar, “Survey on Density based Clustering Algorithm for mining large spatial databases”, IJAST 2011.

[23] L. Duan, L. Xu, F. Guo, J. Lee and B. Yan, “A local-density based spatial clustering algorithm with noise,” Information Systems, vol. 32, pp. 978-986, 2007

[24] Mohammed T. H. Elbatta and Wesam M. Ashour, “A Dynamic Method for Discovering Density Varied Clusters,” Int. Journal of Signal Processing, Image Processing, and Pattern Recognition, vol. 6, no. 1, pp. 123-134, 2013.

[25] Vinod CSS, Anand Hareendran S Machine learning: a practitioner’s approach. PHI Learning Pvt Ltd, Delhi,2021

[26] Visibelli A, Bongini P, Rossi A, Niccolai N, Bianchini M A deep attention network for predicting amino acid signals in the formation of [formula: see text]-helices. J Bioinform Comput Biol:2050028, 2020