IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 October 2023, accepted 18 November 2023, date of publication 1 December 2023,
date of current version 13 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3338735

RESEARCH ARTICLE

Graph Convolutional Neural Network-Based Virtual Screening of Phytochemicals and In-Silico Docking Studies of Drug Compounds for Hemochromatosis

R. ANI[1] AND O. S. DEEPA[2]
[1]Department of Computer Science and Applications, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, Vallikavu 690525, India
[2]Department of Mathematics, Amrita School of Physical Sciences, Amrita Vishwa Vidyapeetham, Coimbatore 641112, India

Corresponding author: R. Ani (anir@am.amrita.edu)

ABSTRACT Machine learning based Virtual Screening has proved as an important intermediate process that helps in the field of drug discovery in reducing the cost and manpower of classical drug discovery process. This work proposes a deep learning based virtual screening model for the early discovery of drug compounds for the disease named Hemochromatosis, which is the excess absorption of iron in the human body. Our study focuses on finding possible drug compounds from medicinal plants to cure Hemochromatosis. The proposed method uses Graph Convolutional Neural Networks (GCN) for Ligand Based Virtual Screening (LBVS). Deep Learning algorithm, GCN outperformed all other experimented models in LBVS with an accuracy of 98.26% and F-score of 98% respectively. A small set of biologically active compounds was identified from the phytochemical dataset after performing the LBVS. The selected ligands after LBVS are taken for In-Silico Structure Based Screening (SBVS) called molecular docking and the best compounds that have high binding affinity towards the disease protein for Hemochromatosis is selected and recommended for in-vitro studies. Ablation studies are done with 12 different machine learning models including ensemble models. The proposed model exhibited a related percentage improvement of around 0.5% in accuracy and F-score, when compared to the tree based ensemble model, XGBoost. This study aims to suggest in-silico studies for ligand based and structure based screening to identify potential drug molecules from medicinal plants which can be tested in in-vitro analysis and studies.

INDEX TERMS Computational drug discovery, virtual screening, molecular descriptors, molecular fingerprints, graph convolutional neural networks.

I. INTRODUCTION

Drug Discovery is a time-consuming and cost expensive process faced by pharmaceutical industries. To address the challenges faced by the pharmaceutical industries, Computer Aided Drug Discovery (CADD) [1], [2] has been considered as a boon to overcome the challenges in pharmaceutical field in the drug discovery and development process. In the early stages of drug discovery, CADD helps to perform a computation based screening of compounds to reduce the large set of ligands that are to be experimented to check whether they are druggable or not. The drug discovery process employs in-silico strategies [3] that leverage state-of-the-art computational methods, including machine learning and deep learning algorithms. These algorithms have been widely applied for computer-assisted drug discovery [4], [5], [6], [7]. The advancement in chemoinformatics and the newest technologies in computing power have generated a tremendous rate of advancement in drug discovery procedures [8], [9], [10].

Virtual Screening includes many computational methods to screen a large number of small molecules in order to

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

identify hit ligands in the drug discovery process [11]. The two crucial virtual screening approaches towards drug discovery are Ligand Based Virtual Screening (LBVS) and Structure Based Virtual Screening (SBVS) [12], [13], [14], [15]. Many computational approaches are there to implement the two approaches. Machine learning and Deep learning based virtual screening are considered very useful in the initial screening of compounds.

The LBVS is a computational approach used in drug discovery to identify drug-like molecules when the structure of the target protein is not known. The screening of molecules based on the molecular similarity of known drug molecules is done in LBVS. It uses the information available in the ligand for the screening of molecules instead of information about structure of target protein. In LBVS, researchers may utilize molecular descriptors or molecular fingerprints to represent the molecular properties [16]. LBVS identify potential drug candidates which can be used to bind to a specific target protein, usually a disease related enzyme or receptor. The basic approach in LBVS is comparing the similarity of physicochemical properties and shape of molecules selected. It deals with searching through a compound database of small molecules called ligands and identifying those that have a similar 3D shape and chemical properties to a known, high affinity ligand for the target protein [13]. The LBVS works based on the assumption that if two molecules have similar shapes and chemical properties, then they are likely to bind to the same protein target. Depending on the type of virtual screening, non-drug molecules are removed from a particular data set using a variety of conditions and analytics. This is a fast and cost effective way of screening large numbers of compounds, reducing the number of candidates that need to be tested experimentally, and providing a starting point for further drug discovery studies. The entire screening process depends on how researchers compare the properties of a potential drug candidate with a given set of drug compounds or non-drug compounds [17]. It is needed to determine the optimal way to depict the molecular feature to create the perfect similarity based virtual screening system. Molecular fingerprints, which are the mathematical representations of molecular descriptor values are considered as best inputs in drug likeness prediction models.

In this study, a computational approach that includes ligand based and structure based screening methods is applied to identify a potential drug compound for the disease, Hemochromatosis [18]. The proposed approach is to identify a potential drug from medicinal plants to cure Hemochromatosis. A graph based molecular fingerprint is generated based on the SMILES [19] input using GCN and this information is used to predict the drug likeness of ligand molecule more accurately when compared to other simple and ensemble Machine Learning (ML) models. After the drug likeness prediction, 1000 drug compounds are identified for in-silico docking studies and the best 50 molecules were taken based on binding energy and 25 docking simulations are visualized for detailed understanding. The selected molecules after the in-silico studies can be suggested for the laboratory studies.

The organization of this article is arranged in a subsequent manner. Section I gives a brief introduction of the work. Section II describes drug discovery from medicinal plants. Section III describes the survey of literature based on different topics and technologies used for this work. Section IV reveals methods and materials, and Section V explains the proposed methodology. Section VI demonstrates the results and discussion. Section VI-A deals with the docking process and Section VII depicts the conclusion of the investigation.

II. DRUG DISCOVERY FROM MEDICINAL PLANTS

Recently, the significance of scientific research on medicinal plants is increasing in many developed countries. This need is being addressed by various research institutes, universities, pharmaceutical laboratories, and clinics. The research is primarily focused on two aspects: firstly, studying the bioactive molecules of plants that have been traditionally used for their therapeutic properties based on prior surveys and literature. Secondly, basic research has led to the identification of new medicinal plants containing novel bioactive molecules, bioactivity, and drugs from remote areas of the world. Numerous novel medications that have been found to be beneficial in treating ailments are derived from phytochemicals. Most ailments could be treated effectively with substances that are taken from plants. Several lead compounds have been found to be effective against diseases like AIDS, Alzheimer's, Diabetes, Malaria, Cancer, etc. When used as a medication to treat Influenza, the chemicals alone isolated from medicinal plants were quite effective [20]. Drug discovery from medicinal plants involves several difficulties, including material collection, active component identification, and selection, among others [21]. The importance of phytochemicals' biological activity has been demonstrated in several clinical investigations and pharmaceutical research. According to research on the chemical components of currently accessible pharmaceuticals, between 30 and 50 percent of them are derived from plants [22], [23].

Hemochromatosis is a genetic disorder happening in human body that causes the body to absorb and store too much of iron from food, leading to an excess deposit of iron. The iron deposits in various organs and tissues and the deposit of iron content damage the organs such as liver, pancreas, heart, and other vital organs, resulting in potentially life-threatening complications such as liver cancer, heart failure, and diabetes, arthritis, kidney failure etc [24], [25]. This is because of a mutation happening in the gene which is responsible for the absorption of iron from food into the body. Hemochromatosis can be caused by several different genetic mutations, but the most common cause is a mutation in the Homeostatic Iron Regulator(HFE)gene. This gene normally produces a protein called Hepcidin, that helps to regulate the absorption of iron in the body. When this protein is not functioning properly, too much iron can accumulate in the body. Many people suffer from the C282Y mutation happening to HFE gene which causes Hemochromatosis genetic disease vulnerability. Symptoms of Hemochromatosis can vary widely and may include fatigue, joint pain, abdominal pain, and weakness. If left untreated, hereditary Hemochromatosis can lead to morbidity and eventually death. Treatment may involve regular blood removal to lower the iron content levels in the body. Iron chelation therapy is used to remove iron from the organs.

Identifying an effective medicine from medicinal plants to cure hemochromatosis using computational methods is the main objective of this study. It can be discovered by efficient screening technologies using machine learning algorithms. The data set consists of more than 30,000 small molecules gathered from databases like NPASS (Natural Product Activity and Species), Super Natural II, Chembl, ZINC, etc… [26]. In the descriptor-based study, 48 molecular descriptor values are considered as features. In the molecular fingerprint approach, 166-bit size to 2048-bit sizes are considered features for various studies. The training set to test data ratio is 75:25.

III. LITERATURE REVIEW

Elbadawi et al. [27] reviews the use of advanced techniques to address limitations in machine learning (ML) for drug discovery. It discusses how ML can be used to improve classification performance and presents emerging techniques that could potentially expand its application. Examples from drug discovery are provided on how these approaches are being applied which results promising outputs. The review also looks at challenges such as needing large datasets, sparsity in data, lack of interpretability and retraining post deployment which may limit the effectiveness of ML algorithms when it comes to drug discovery applications. Finally, potential solutions using Bayesian neural networks (BNNs), explainable algorithms or other methods are discussed which aim to overcome these issues so that more effective usage is possible within this field.

Jiménez-Luna et al. [28] discusses the influences of Artificial Intelligence in chemoinformatics and its applications to drug discovery. The main focuses include quantitative structure-activity/property relationship and structure-based modelling, molecular design, and predictions of chemical synthesis. This discusses about deep-learning based applications which have been used to address some fundamental problems in drug discovery.

Atanasov et al. [29] presents some reviews about drug discovery from natural products that, the natural products are some of the major contribution to the pharmacotherapy, especially for cancer and many infectious diseases. There are many technical barriers like screening, isolation etc., in the discovery process. This causes to a decline in their pursuit by the pharmaceutical industry since 1990s. But recent technological developments like improved analytical tools and advanced datamining strategies have opened up new opportunities leading to research or interests in natural product-based drug discovery.

Machado et al. [30] proposes that machine learning-driven ligand based virtual screenings can be used to save time and money when looking for new treatments/inhibitors against HIV-1. In this study, Random Forest model combined with SMOTE was found to give good results in distinguishing between active or inactive compounds against the HIV-1 Integrase enzyme.

The study of Carpenter and Huang [31] presents reviews of ML-based methods used for Virtual Screening (VS) and applications to Alzheimer's Disease (AD) drug discovery. In this five Machine Learning techniques are discussed, and they are Naïve Bayes, k-Nearest Neighbors, Support Vector Machines, Random Forests and Artificial Neural Networks. All of these algorithms have found success in VS but neural networks -and more specifically Convolutional Neural Networks – may be preferred due to their accuracy when applied on unseen databases.

A workflow is proposed that can help researchers conduct ML based VS for potential therapeutics related to AD or other complex diseases with no known cure/prevention yet. Collaborations between AI companies & pharmaceuticals benefit from combining state-of-the-art hardware & technologies along with chemogenomics libraries which helps make drug development process faster & cost effective.

According to the review on Machine Learning in Drug Discovery by Dara et al. [32], Machine Learning (ML) tools and techniques can be used to accelerate the drug discovery process and reduce risk and expenditure in clinical trials. ML is being applied across various applications such as QSAR analysis, hit discoveries, de novo drug architectures etc., for accurate outcomes. Clinical trial data needs to generated accurately so that it helps tackle puzzles while validating ML techniques & improving decision-making processes during Drug Discovery activities.

Pinzi and Rastelli [33] reviews the importance and use of molecular docking in drug discovery. This paper discusses the use of docking in newer applications such as predicting the side effects, polypharmacology, drug repurposing, target fishing and profiling etc. This explores the future uses of molecular docking when combined with artificial intelligence. It explains that High-Throughput Screening (HTS) has been the standard for discovering biologically active hits but it is expensive to implement. So in silico approaches are being used more often due to their low cost and increased chances of finding desired drug candidates. It mentions various types of molecular modelling techniques used in structure based and ligand-based approaches.

Neves et al. [34] provides a review of the QSAR models and the advantages and disadvantages of QSAR based virtual screening in drug discovery. This study provides an overview about many applications where compounds having desired QSAR are identified using the computational approach.

This study emphasis the importance of QSAR than High Throughput Screening (HTS) in drug discovery domain. This paper says that QSAR based VS can be used to improve the hit rates of HTS.

Maia et al. [35] provides an overview of the challenges associated with CADD to perform SBVS. It compares common tools and techniques which are used for SBVS. A method called Consensus Virtual Screening (CVS) is introduced which helps to increase accuracy while reducing false positives obtained from the experiments. Homology modelling methodology is also discussed which allows prediction of 3D structure protein from its amino acid sequences.

Kimber et al. [36] reviews the applications and the advancement of machine learning and deep learning in virtual screening methods for active drug designs. Deep learning methods have been successfully implemented based on the availability of huge amount of chemical and bioactivity data. It discusses different encodings used to represent compounds as well as proteins along with various techniques such as Bioactivity data sets which help to train better while testing then against Benchmark Data Sets. The challenges faced are also discussed and the rise of deep learning due to novel technologies and increasing availability of chemical and bioactivity data are emphasised.

Andrianov et al. [37] proposes computational method to identify drug like compounds using in silico virtual screening methods. Molecular docking, quantum chemical calculations and molecular dynamics simulations are done on approved drug candidates to identify potential viral inhibitors of SARS-CoV2 main protease. They provide an overview of their research methodology for further exploration into this field.

Liu et al. [38] discusses the development of a user-friendly web server with integration of state-of-the-art deep learning algorithms. This web server is designed to help biologists and chemists perform virtual screening for chemical probes or drugs against specific targets. The DeepScreening tool allows user to analyse and to construct classification and regression models, generate target focused new libraries, as well as conduct virtual screenings on diverse chemical libraries in stock.

IV. MATERIALS AND METHODS

In this study for the ligand-based virtual screening of compounds, the input data is SMILES representation of compounds. The compounds are taken from natural compound databases and the format is converted into SMILES representation for processing. The features of molecules for the processing are the molecular descriptors. There are different approaches in giving the descriptor inputs to the system. First approach is calculating the different descriptor values that can be computed from molecular SMILES. The second approach is the representation of molecular descriptor values as 0's and 1's which is called molecular fingerprints. Another approach is using molecular graph and calculating the molecular fingerprints based on the graph. Here, SMILES representation is given as the basic input to the model.

In the second phase, an in silico docking study of selected bioactive compounds are carried out and the results are taken. Few compounds are taken for visualization and the results are analyzed.

A. SMILES REPRESENTATION

SMILES, which stands for Simplified Molecular Input Line Entry System, is a way of representing the structure of a molecule using a string of characters. SMILES gives a linear representation of chemical compound. SMILES notation consists of a series of characters containing no spaces in between In SMILES representations, atoms of compounds are represented using atomic symbols. The structure of the molecule is represented using string of characters. The atoms, bonds and functional groups are represented by fixed set of alphabets and some special alphanumeric characters. This notation is used to encode the structural information of a molecule in a concise and unambiguous way. The representation is case sensitive. The single, double, triple and aromatic bonds are represented as −, =, #, and: respectively. Parentheses indicate the location of functional groups on the ring structures. O, C and N are oxygen, carbon and nitrogen. Atoms in aromatic rings are specified by lower case letters [39], [40], [41].

B. MOLECULAR DESCRIPTORS

Different molecular descriptors are discussed in the following subsections.

1) DESCRIPTOR VALUES

Molecular descriptors are the unique features of a molecule which describes the molecule based on physicochemical and structural properties. The descriptors can be classified into 0D, 1D, 2D, 3D. The molecular formula, atom types, bond types are 0D descriptors., The counts of atom types, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rings, number of functional groups etc comes under 1D. Mathematical epresentations by graph theory or calculated values like lipophilicity, TPSA etc comes under 2D. 3D include all geometrical descriptors and Polar Surface Area [42].

2) MOLECULAR FINGERPRINT REPRESENTATION

The molecular fingerprint is a binary string that represents the presence or absence of specific structural and chemical features in the molecule. Once the molecule has been represented in a standardized format, a computational algorithm is used to generate a unique molecular fingerprint based on the molecular descriptors. Molecular fingerprints are vector representations of molecular properties. SMILES representation of molecules are used to generate the molecular fingerprint.

The generated binary feature vectors values will be imported to a machine learning model that will efficiently predict the given molecule is drug or non-drug. 2D fingerprints provides more structural information for the precise understanding of the molecule structure [16].

Major classification of molecular fingerprints include:

Substructure key based Fingerprints: Bits are generated in substructure key based fingerprints in accordance with the molecule substructures. Each position in the fingerprint represents the presence or absence of a substructure. Molecular ACCess Systems (MACCS) keys fingerprint and PubChem Fingerprints are the most common example of substructure key based fingerprints.

Path based fingerprints: All the molecule fragments that follow a linear path up to a predetermined number of bonds are evaluated to establish topological or path-based fingerprints. This method allows the creation of a reliable path-based fingerprint for any molecule and performs a validation process to ensure the fingerprint was created effectively. Topological fingerprints are more effective than other fingerprints at promptly identifying substructures. A specific bit is not associated with a single feature because path-based fingerprints are hashed fingerprints.

Circular Fingerprints: Circular fingerprints record each atom's environment up to a specific radius rather than searching for paths within the molecule. Due to the possibility of different contexts for the same fragment, they are thus inappropriate for substructure searches but are widely used for searching for complete structure similarity. Some of the circular fingerprints that are used the most frequently include Metaprint2D, Functional Class Fingerprints (FCFPs), Extended Connectivity Fingerprints (ECFPs), and so on.

Machine learning based in-silico LBVS studies using molecular descriptors are carried out. Ensemble method-based XGBoost is one of the best algorithms that can be used to implement LBVS based on molecular descriptor values and molecular fingerprints. Many base models are implemented to check the prediction accuracy and other performance measures. To improve the accuracy of the screening result and to boost the performance of the system, various state of the art ML based screening models were compared to conclude the effectiveness of this approach. There are defined set of rules to interpret the representations of molecular information in SMILES format. The chemical structure can be generated from a SMILES format representation of a molecule. The detailed examination demonstrates based on the findings, if the molecule is a feasible medication candidate, the framework will dock the molecule with the target protein to examine how it interacts with the target.

C. MACHINE LEARNING MODELS FOR DRUG CLASSIFICATION

In this study, 12 different machine learning models, including ensemble models, to predict drug likeness using molecular descriptor values and molecular fingerprints based on five different methods: MACCS166, Atompair, RDKit, Topological Torsion, and Morgan [43], [44], [45]. The workflow of the screening process based on molecular fingerprints and molecular descriptor values is shown in Fig.1.

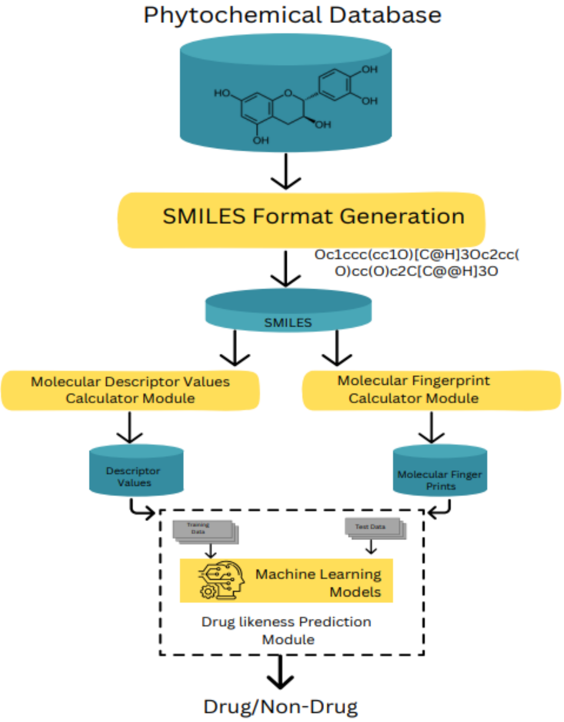


FIGURE 1. Workflow of ML based drug likeness prediction using molecular descriptors.

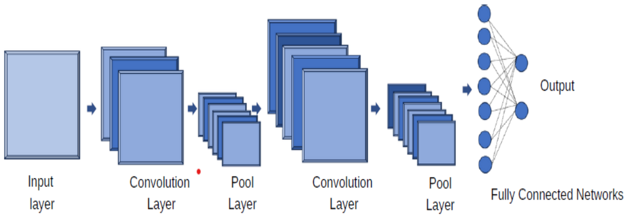


FIGURE 2. General structure of GCN.

The study aimed to compare the performance of the different models and methods in predicting drug likeness. The ensemble models [46] like SVM with Bagging, KNN with bagging, Random Forest, Rotation Forest with PCA, Rotation Forest with LDA, Adaboost, Gradient Boost, XGBoost, XGBoost and RFE, Stacking.

Extreme Gradient Boosting, or XGBoost, is an ensemble machine learning model built on decision trees that uses the gradient boosting technique [47], [48]. Gradient boosting is a special type of boosting technique that was introduced to make the system more optimized to handle data errors. Even though both Gradient boosting and XGBoost work based on boosting week learners using gradient decent technique the XGBoost has an upper hand on both performance and accuracy due to a much more optimized algorithm flow.

V. PROPOSED METHOD

GRAPH CONVOLUTIONAL NEURAL NETWORK (GCN)

A Graph Convolutional Network (GCN) based virtual screening model to predict the drug likeness of chemical compounds is proposed after detailed study and comparison with machine learning models [49]. Graph molecular fingerprints generated from GCN predict the drug likeness of compounds very precisely. The graph representations can be generated based on the representation of molecules in the SMILES format. Graph Convolutional Networks (GCNs) have solidified their position as the cutting-edge method for tackling drug-related tasks due to two critical advantages:

- GCNs excel at feature extraction by considering the inherent data structure, making them adept at capturing subtle relationships within the data.
- GCNs empower automatic feature extraction directly from raw inputs, eliminating the need for manually crafted features that could potentially overlook valuable information, influenced by the inherent biases of domain experts.

The architecture of GCN is the same as Convolutional Neural Networks (CNNs) and it consists of fully connected, pooling, and graph convolutional layers. The general structure diagram of GCN is shown in the Fig.2. The primary distinction between GCN and CNNs is the substitution of graph convolutional layers for convolutional layers. The convolutional layer, fully connected layer, and pooling layer are the three basic layers of CNN in the graph, the same as it is used with images. The pooling layer is used to downscale a graph, whereas the convolutional layer is used to learn receptive fields in graphs whose data points are not ordered in Euclidean grids. The output from the convolutional layer or final pooling is passed into the fully connected layer, where it is flattened before being applied. GCN's main principle is to implement convolution on a graph. It accepts a graph as its input rather than a 2-D array. In contemporary graph CNN research, the parameter combination of the graph filter spectrum and graph signal spectrum defines the convolution operation in the signal spectrum region. In order to restore the graph vertex domain, the outcome is modified. Because the drug molecule itself is represented as a graph, GCN is well suited for the drug identification. The architecture diagram of the proposed model is shown in the Fig.3.

The smiles of chemical compounds are given as the input to the GCN-based model. The GCN model accepts the SMILES input and generates Graph-based molecular fingerprint which is a representation of molecules which can be directly used to predict the properties and activities of compounds. The mathematical explanation for graph-based molecular fingerprints is:

Let **G = (V,E)** be a molecular graph where V is the set of nodes representing the atoms and E is the set of edges which is representing the bonds between atoms. Let **X** be the feature matrix representing the features of the molecule where each row represents an atom, and each column corresponds to a feature.

In the molecular graph, the Adjacency matrix is defined as:

$$A_{i,j} = \begin{cases} 1 & \text{if edge exists between i and j} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

GCNs perform convolutions on the graph by aggregating information from neighboring nodes. Each node aggregates and updates its features based on its neighbors' features. The new features of a node are computed by taking a weighted sum of its neighbors' features.

The mathematical representation of aggregation operation is:

$$h_i' = \sum_{j=1}^{N} \frac{1}{\sqrt{d_i d_j}} \cdot h_j \cdot W \tag{2}$$

where h_i' is the feature vector of node i, W is the weight matrix, and d_i and d_j are the degrees of nodes.

In traditional molecular fingerprints, molecules are represented by a fixed-length vector of binary or integer values, which encode molecular features such as the presence or absence of certain functional groups or substructures. Molecular fingerprints fail to capture the structural relationships between atoms and their local environments in a molecule. But GCN can learn to extract features directly from the molecular graph, which represents the molecular structure as a set of nodes (atoms) and edges (bonds). It uses convolutional filters to perform localized feature extraction on the graph, where the filters are learned from the graph structure itself and can be used to capture patterns and relationships in the molecular structure. The molecular graph is first constructed from the molecular structure, and then the GCN is applied to the graph to learn a set of features that are specific to the molecular structure. The learned features can then be used as input to the machine learning models, to predict various molecular activities and properties. GCN based molecular fingerprints have shown promise in a variety of applications, including drug discovery and material design. They are particularly effective for predicting properties and activities of molecules that have complex or flexible structures, where traditional molecular fingerprints may fail to capture the relevant information.

Graph convolutions operate on a graph and begin with a data vector for each node of the graph (for example, the chemical properties of the atom that node represents). Convolutional and pooling layers combine information from connected nodes (for example, atoms that are bonded to each other) to produce a new data vector for each node. For the implementation of graph based fingerprint generation,

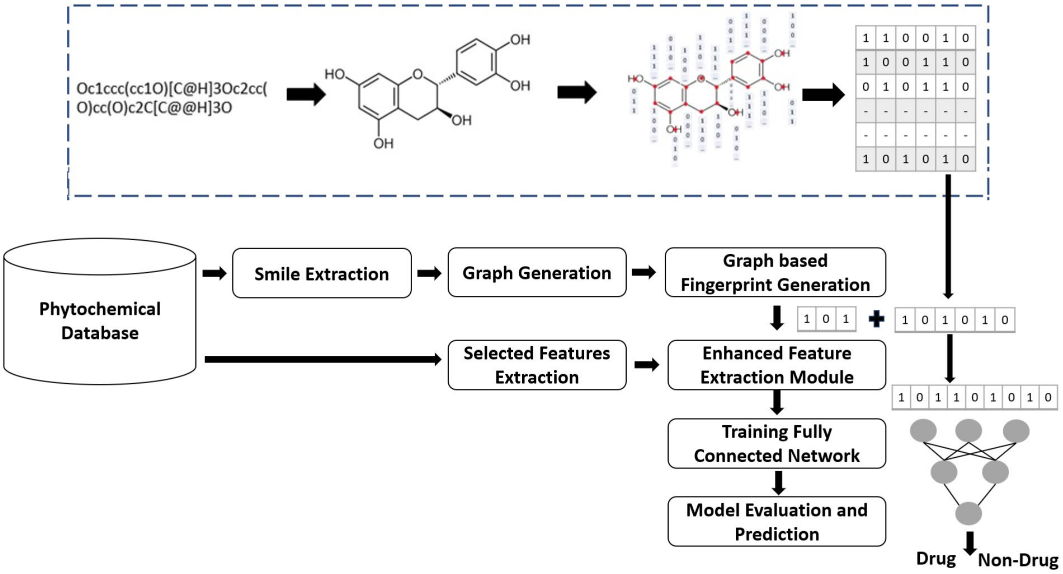

Oc1ccc(cc1O)[C@H]3Oc2cc(O)cc(O)c2C[C@@H]3O
HO
OH
OH
OH
OH
1 1 0 0 1 0
1 0 0 1 1 0
0 1 0 1 1 0
1 0 1 0 1 0
Phytochemical Database
Smile Extraction
Graph Generation
Graph based Fingerprint Generation
1 0 1
1 0 1 0 1 0
Selected Features Extraction
Enhanced Feature Extraction Module
1 0 1 1 0 1 0 1 0
Training Fully Connected Network
Model Evaluation and Prediction
Drug
Non-Drug

FIGURE 3. Architecture diagram of proposed model.

a python library, DeepChem[1] is being used. DeepChem aims to provide a high quality open source toolchain that democratizes the use of deep-learning in drug discovery, materials science, quantum chemistry, and biology. For the GCN model creation in the present study, the default parameter values (Dense layer size=512, Batch normalise=True, Batch size=64, Mode=Classification, No:of epochs=10) of DeepChem library for GCN are adopted. At first, an one hot encoding operation is done on the columns and removes the outliers by discarding rows that have features that have z score $\geq$ 3 std deviations away from the mean. Here 'Canonical SMILES' is the column used to model a Graph Convolution Network. In this study, DeepChem's ConvMolFeaturize is used to Featurize SMILE strings into Molecular graphs. A custom GraphConvModel is created by using the following layers from DeepChem.

A. GRAPHCONV LAYER

The function of this layer is to perform graph convolution, which involves combining feature vectors of individual nodes in a nonlinear manner with the feature vectors of neighboring nodes. This process effectively integrates information from local neighborhoods within the graph, resulting in a blended representation.

B. GRAPHPOOL LAYER

This layer performs max-pooling on the feature vectors of atoms within a neighborhood. It can be conceptualized as a counterpart to the max-pooling layer used in 2D convolutions, but adapted to operate on graphs instead.

C. GRAPHGATHER

Numerous Graph Convolutional Networks perform operations on feature vectors at the level of individual graph nodes. In the case of a molecule, each node could represent an atom, and the network would manipulate atomic feature vectors that provide a summary of the atom's local chemistry. Nevertheless, at the conclusion of the application, it is typically desirable to work with a feature representation at the molecule level. To achieve this, a graph gather layer is utilized, which combines all the node level feature vectors to create a single graph level feature vector. Subsequently, the output of the graph gather layer is concatenated with the remaining features in the dataset and passed through a dense layer.

D. TRAINING

The model is trained for 10 epochs after which the accuracy, precision, recall on the test dataset are calculated.

VI. RESULTS

The operations required to perform virtual screening are based on fingerprint similarity: a reference molecule, or at least one known active compound (s), a database of probable active molecule and software that can create to contrast fingerprints. The best fingerprint should then be selected after the reference molecules have been determined. The options offered by the programme being utilised typically restrict the options. The fingerprint accuracy can be obtained by choosing the best reference molecules. It is necessary to determine whether the database or the fingerprints used for screening the tautomeric forms, stereochemistry and the reference molecules. Screening databases should ideally be executed using stereochemistry-sensitive methods. The usage

[1] https://deepchem.readthedocs.io/en/latest/

TABLE 1. Comparison of performance measures of different ML algorithms in descriptor values based drug likeness prediction.

Algorithms	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost	97.750	97.753	97.753	97.753	0.977
RFE+XGBoost	97.578	97.773	98.172	97.377	0.972
Rotation Forest LDA	97.424	97.563	97.899	97.731	0.958
Stacking	97.400	97.404	97.404	97.404	0.974
Random Forest	97.310	97.361	97.744	97.552	0.973
Rotation Forest PCA	95.808	96.484	95.911	96.196	0.958
Decision Tree	95.227	95.530	95.904	95.716	0.951
Bagging+SVM	94.907	94.819	95.771	95.292	0.950
Gradient Boost	92.969	93.348	93.649	93.498	0.929
Bagging+KNN	91.424	93.384	90.842	92.095	0.915
Adaboost	91.072	93.342	90.607	91.955	0.911
LDA Classifier	90.921	93.062	90.473	91.749	0.910

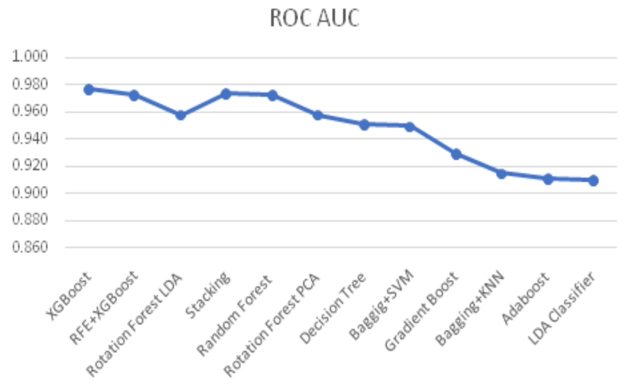


FIGURE 4. Comparison of AUC-ROC of different ML Algorithms using descriptor values in drug likeness prediction.

TABLE 2. Performance measures-MORGAN fingerprint.

	MORGAN				
Algorithms	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
XGBoost	97.500	97.497	97.496	97.496	0.975
Stacking	97.380	97.381	97.381	97.380	0.974
Random Forest	97.364	97.370	97.554	97.462	0.974
LDA Classifier	97.061	96.190	98.205	97.061	0.970
Bagging+SVM	96.629	96.652	96.970	96.811	0.966
Rotation Forest LDA	95.626	96.747	94.717	95.721	0.957
Gradient Boost	95.503	95.473	95.654	95.564	0.955
XGBoost+RFE	95.217	95.347	94.681	95.436	0.953
Adaboost	94.695	94.727	94.995	94.861	0.947
Rotation Forest PCA	94.463	94.786	94.573	94.680	0.945
Decision Tree	93.836	93.813	94.424	94.118	0.938
Bagging+KNN	68.972	62.862	99.963	77.186	0.673

of fingerprints that rely on conformations is made possible by their presence. Tautomerism of the molecules under study should also be considered because different tautomers of the similar molecule may have dissimilar fingerprints. The proposed algorithm would produce fingerprints for each reference molecule in the database before calculating the likeness coefficient among each reference molecule and each other molecule. The similarity coefficient can then be used to order the molecules in decreasing order. The top molecules in the rank should display behaviour that is comparable to that of the reference molecule.

Here, several machine learning algorithms are evaluated using parameters such as accuracy, F1-score, recall, precision, and ROC score. The parameters are calculated using eqs. (3)– (6) respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - score = \frac{2TP}{2TP + FP + FN} \tag{6}$$

A. ANALYSIS OF DIFFERENT LIGAND-BASED VIRTUAL SCREENING APPROACHES

In this section we present analysis of three different approaches for ligand based virtual screening viz., descriptor values based, finger print based and graph based.

1) LBVS USING DESCRIPTOR VALUES

Different ML algorithms are used to predict the drug likeness of molecules based on molecular descriptor values. The most accurate prediction result is given by XGBoost algorithm when the inputs are the different molecular values. So the tree based ensemble model which uses the boosting concept can be considered as a better prediction for drug likeness. The accuracy, precision, recall and F1 score values are given in Table1. The AUC of RoC curve is shown in Fig.4.

2) LBVS USING MOLECULAR FINGERPRINTS

- **MORGAN FP**

  Morgan fingerprint is primarily a reconfiguration of the Extended Connectivity Fingerprint (ECFP). Here, a number with a maximum based on bit number is hashed from each distinct path. The larger the fragments are encoded in the place where the value of the radius is increased. Therefore, a Morgan radius contains all pathways discovered in a Morgan radius along with the additional bits. The Morgan Fingerprint is used to compare different algorithms along with the metrics such as accuracy, recall, precision, F1-score and ROC score. Table 2 contains the results of different algorithms using Morgan Fingerprint. The values of the XGBoost algorithm are higher than the other algorithm.
- **MACCS166**

  Structure fingerprints known as MACCS keys are frequently employed to calculate molecular similarity. It is not limited to individual fingerprint vectors. The redundant vectors are eliminated using these fingerprints. From Table 3, it is observed that the proposed (XGBoost) algorithm performs better than other conventional algorithms.
- **ATOMPAIR FP**

  Atom fingerprints are often utilized to describe atoms immediate environs. It includes a structural search approach that reduces duplicate structures along with encoding information such as lengths of the link to nearby atoms or crystal organization of co-ordinating numbers. The energy invariance under

TABLE 3. Performance measures- MACCS166 fingerprint.

	MACCS166				
Algorithms	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
XGBoost	97.610	97.612	97.612	97.612	0.975
XGBoost+RFE	97.299	97.147	97.659	97.404	0.973
Stacking	97.090	97.092	97.092	97.090	0.971
Bagging+KNN	96.755	95.961	97.993	96.966	0.967
Baggig+SVM	96.478	96.439	96.759	96.599	0.965
Gradient Boost	96.433	96.113	97.101	96.605	0.964
Rotation Forest LDA	96.297	97.003	95.876	96.436	0.963
Random Forest	96.183	96.828	95.710	96.266	0.962
LDA Classifier	95.707	95.428	96.337	95.880	0.957
Rotation Forest PCA	95.689	96.871	94.764	95.806	0.957
Decision Tree	94.953	95.115	95.159	95.137	0.910
Adaboost	94.044	94.656	94.476	94.566	0.940

TABLE 4. Performance measures- ATOMPAIR fingerprint.

ATOMPAIR					
Algorithms	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
XGBoost	97.500	97.496	97.496	97.496	0.975
XGBoost+RFE	97.404	97.421	97.579	97.500	0.974
Random Forest	97.145	97.343	97.125	97.234	0.973
Bagging+SVM	96.425	97.157	96.072	96.612	0.964
Stacking	96.590	96.593	96.591	96.591	0.966
Rotation Forest LDA	95.459	96.396	94.929	95.657	0.955
Gradient Boost	96.088	96.176	96.283	96.230	0.961
LDA Classifier	96.066	95.651	96.792	96.218	0.960
Adaboost	95.709	94.236	95.979	95.890	0.941
Rotation Forest PCA	93.300	94.235	93.054	93.641	0.933
Decision Tree	92.520	91.815	93.832	92.813	0.925
Bagging+KNN	90.526	84.572	99.186	91.298	0.905

TABLE 5. Performance measures-topological torsion fingerprint.

TOPOLOGICAL TORSION					
Algorithms	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
XGBoost	96.880	96.881	96.881	96.879	0.969
Stacking	96.760	96.766	96.764	96.765	0.968
Random Forest	96.675	97.688	95.963	96.818	0.967
LDA Classifier	95.825	95.434	96.527	95.977	0.958
Bagging+SVM	95.772	96.076	95.754	95.915	0.958
Rotation Forest LDA	95.674	96.283	95.326	95.802	0.957
Gradient Boost	95.018	95.455	94.778	95.115	0.950
XGBoost+RFE	94.134	94.129	94.134	94.743	0.946
Rotation Forest PCA	93.243	94.607	92.158	93.366	0.933
Adaboost	92.171	92.579	92.325	92.452	0.922
Decision Tree	91.390	91.111	92.426	91.764	0.913
Bagging+KNN	77.647	69.867	99.963	82.248	0.768

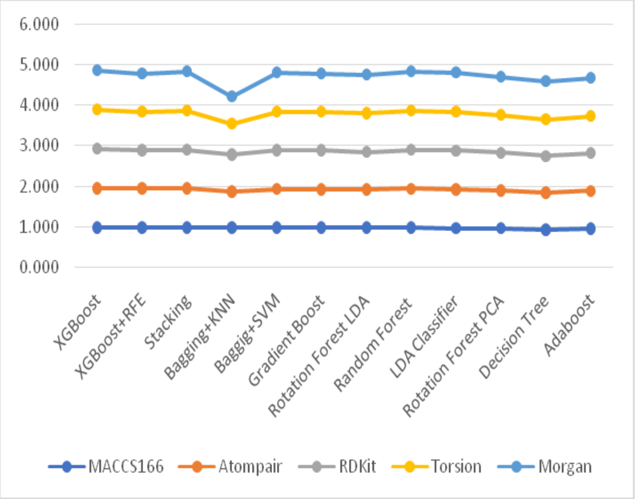


FIGURE 5. Comparison of AUC-ROC of 5 fingerprint models.

particular operations is not encoded by these types of fingerprints. It must hold true uniform translations, rotations, and permutations of the systems identical atoms. Furthermore, if two environments have identical fingerprints, they are guaranteed to be identical, hence fingerprints must be distinct. A fingerprint used as an input will give the same energy to two different non-degenerate structures if this requirement is not satisifed. Here, XGBoost algorithm performs efficiently with Atom fingerprints as described in Table 4.

- **Torsion FP**
  Molecule conformations are assessed and compared using the Torsion Fingerprint. It derives Torsion Fingerprints using a querying molecule and its produced conformations, weights them, and compares them while taking acyclic bonds and ring systems into consideration. Table 5 depicts the performance of the different algorithms using the TORSION fingerprints. The proposed algorithm obtains 96.8% of accuracy with TORSION fingerprints. Here, stacking and random forest are nearly similar but the proposed method outperforms in terms of accuracy, recall, precision, F1-score and ROC score.
- **RDKit**
  Most fundamental molecular functions are carried out by RDKit. Several functions can be used to transform single molecules into text. Usually, hydrogen atoms are implicit in the molecular graph when storing compounds in the RDKit. Using the various embedding techniques, many conformers can also be produced using the RDKit. In all situations, all that is required is to repeatedly execute the distance geometry calculation from several random starting positions. Table 6 showcases the performance of the different algorithms using the RDKit fingerprints. Here XGBoost algorithm outperforms other algorithms in terms of accuracy, recall, precision and ROC score. The proposed algorithm achieves 96.8% accuracy, recall, precision, F1-score and ROC score with RDKit.

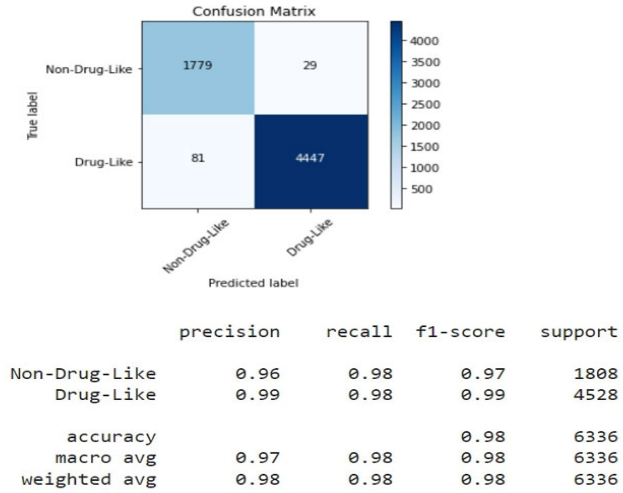
From multiple experiments, we noticed that the XGBoost-based models reported an average F-score of 97.27%. The comparison of AUC-ROC of five fingerprint models are shown in Fig.5.

3) LBVS USING GRAPH CONVOLUTIONAL NEURAL NETWORK

In this section, we present the evaluation of GCN based models. It is noticed from Fig.6 that the proposed approach reported an average F-score of 98% and a relative percentage improvement of 0.5 (Refer Table 7) with XGBoost.

TABLE 6. Performance measures-RDKit fingerprint.

Algorithms	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
XGBoost	96.880	96.880	96.880	96.880	0.969
Random Forest	96.675	97.688	95.963	96.818	0.967
Gradient Boost	96.401	96.064	96.657	96.359	0.961
Stacking	96.170	96.182	96.167	96.169	0.962
LDA Classifier	95.825	95.434	96.527	95.977	0.958
Baggig+SVM	95.625	95.092	96.445	95.764	0.956
XGBoost+RFE	94.798	94.821	93.216	94.763	0.938
Adaboost	93.435	93.995	93.290	93.641	0.934
Rotation Forest PCA	93.089	95.912	90.673	93.219	0.932
Rotation Forest LDA	92.048	93.975	86.384	91.815	0.922
Bagging+KNN	91.678	86.762	89.237	92.581	0.906
Decision Tree	90.975	90.485	92.477	91.470	0.909



```
              precision    recall  f1-score   support

Non-Drug-Like      0.96      0.98      0.97      1808
    Drug-Like      0.99      0.98      0.99      4528

     accuracy                          0.98      6336
    macro avg      0.97      0.98      0.98      6336
 weighted avg      0.98      0.98      0.98      6336
```

FIGURE 6. Performance measures of GCN.

TABLE 7. Performance analysis of three approaches used for LBVS in the present study.

Approach	Best Algorithm	Accuracy	Precision	Recall	F1 Score
Descriptor values based	XGBoost	97.75	97.753	97.753	97.753
Molecular Fingerprints based	XGBoost	97.50	97.496	97.496	97.496
Molecular Graph based	GCN	98.26	97.50	98.000	98.000

In summary, three approaches for drug likeness prediction are included in this study. One is molecular descriptor value-based prediction, second one is molecular fingerprint based prediction and third approach is molecular graph-based prediction. There are many ML algorithms implemented for the first two approaches and XGBoost gave a better performance. Third approach is accepting the molecular graph as input and learning the features by GCN and prediction is done using the fully connected network. The study gave the best accuracy compared to other two approaches.

B. STRUCTURE BASED IN-SILICO DOCKING STUDIES OF THE SELECTED COMPOUNDS

A docking study of the target protein (disease protein) with selected drug compounds is to be carried out to identify a potential drug for a disease. In the present study, we attempt to find a suitable drug for the disease, Hemochromatosis.

TABLE 8. Docking results of selected compounds.

SI No	Molecule Name	Binding Energy (kcal/mol)	RUN	Ki (nM)
1	Azadirachtin Q	-9.91	9	54.46
2	Anguidin	-9.9	8	55.35
3	Pannellin	-9.89	8	55.94
4	Tricolorin C	-9.88	1	57.22
5	Varixanthone	-9.87	9	57.78
6	Polygalacin D	-9.87	6	58.56
7	Hydroxysenkirkine	-9.86	1	59.66
8	Yuanhuadine	-9.85	9	60.73
9	Theograndin I	-9.82	7	63.57
10	Epicalyxin I	-9.81	3	64.05
11	Camelliaside B	-9.79	9	66.42
12	Nimonol	-9.78	4	67.53
13	Glycyrrhizin	-9.78	4	67.70
14	Cucurbitacin A	-9.71	9	75.79
15	11-Alpha,19-Dihydroxytelocinobufagin	-9.7	9	77.75
16	14 -Deoxycrassin	-9.69	8	78.97
17	Saikogenin D	-9.69	1	79.49
18	Moracin O	-9.67	3	81.60
19	Dactyloquinone B	-9.66	1	83.28
20	Proliferin B	-9.65	9	83.83
21	Isoadenolin J	-9.65	5	84.90
22	Landomycin B	-9.64	9	85.76
23	Colistin	-9.64	8	86.13
24	Rabdocoetsin B	-9.64	10	86.33
25	Saikosaponin A	-9.61	3	90.45
26	Pectolinarin	-9.61	7	90.75
27	Irinotecan	-9.6	1	91.37
28	Rabdokunmin A	-9.6	2	91.74
29	Fexofenadine	-9.59	1	93.80
30	Antiaroside F	-9.57	6	96.69
31	Ganoderenic Acid A	-9.56	3	98.30
32	Jasplakinolide R1	-9.55	9	99.17
33	Yuanhuacine	-9.54	1	101.85
34	Nabiximols	-9.52	10	105.38
35	Carotenoid	-9.52	7	105.65
36	Maslinic Acid Methyl Ester	-9.51	9	106.11
37	Petrosterol-3,6 -Dione	-9.51	7	107.19
38	Nelfinavir	-9.51	5	107.39
39	Diolein	-9.49	9	109.72
40	Homaxisterol A3	-9.48	2	111.66
41	Longipedlactone M	-9.47	2	113.52
42	Yuanhuaoate E	-9.46	10	115.64
43	Bis(3,5,5-trimethylhexyl) phthalate	-9.46	7	116.88
44	Griffipavixanthone	-9.44	5	119.65
45	Palodesangren A	-9.42	10	124.05
46	Communesin B	-9.4	1	128.28
47	Maslinic Acid	-9.4	8	129.14
48	Nepapakistamine	-9.4	3	129.83
49	Taiwaniaquinone G	-9.38	10	132.70
50	Plectrornatin A	-9.38	3	133.03

In the proposed work, the ligands were pre-processed using Autodock software [50]. This software aids in the discovery of potential drug candidates, substrates, and receptor binding. It is used to carry out the ligand docking procedure in order to identify the target proteins. More than 30,000 molecules were selected for investigation, among them a collection of

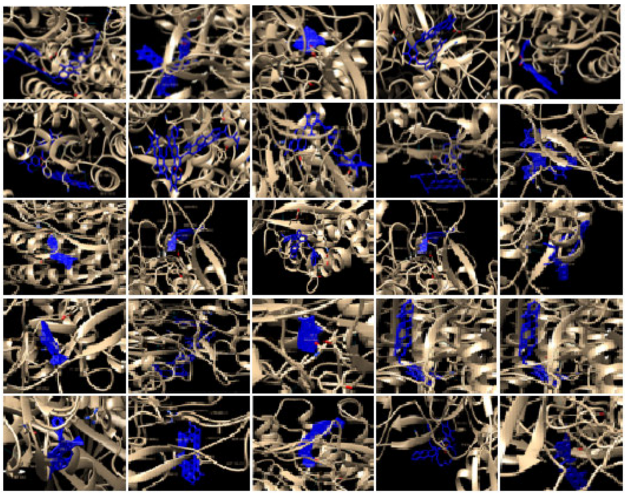


FIGURE 7. Visualistion of docking results.

1000 molecules were selected for docking studies. From the 1000 molecules, 50 molecules were selected, whose binding energy ranges from −9.9 to −8.9 (Table 8). For the visualization of the protein-ligand binding (Fig.7), 25 molecules were randomly selected. The visualization also shows that these selected compounds are likely to be the potential drug for Hemochromatosis.

VII. CONCLUSION

This study recommends the best 50 molecules for the in vitro analysis from a large database of more than 30,000 molecules. GCN model for the Ligand based Virtual screening is the suggested model for drug likeness prediction. The screened molecules based on their bioactive nature are taken for the in-silico docking process with the HFE protein structure 1A6Z. Autodock is the software platform used for docking studies. The initial study started with molecular descriptor values as the features used for the ML models. The study is extended with 5 different types of molecular fingerprint representation as features for the ML based classifications. In these studies, it is concluded that the ensemble model based LBVS based on XGBoost is one of the best model while using molecular descriptor values and fingerprint representations. The study is extended to deep learning based GCN model. A graph-based fingerprint generation by the GCN and prediction using the network is the suggested model for LBVS. The fixed length fingerprints require a huge volume of vectors to compute the sub-structures whereas the graph fingerprints utilize the encoding procedure for the relevant properties. It is revealed that graph fingerprints outperform molecular fingerprints. The experimental findings showed that the proposed graph-based fingerprint method such as Graph Convolutional Neural Network (GCN) outperforms existing methods with a greater accuracy of 98%. Hence, GCN based model is the recommended technique for drug discovery. This study selects a small set of phytochemicals after performing in silico LBVS screening from a large dataset for the in silico docking study which is carried out using Autodock software. After docking the molecules which give a binding energy between −8.9 to −9.9 kcal/mol are recommended for the laboratory experiments.

REFERENCES

[1] M. H. Baig, K. Ahmad, G. Rabbani, M. Danishuddin, and I. Choi, "Computer aided drug design and its application to the development of potential drugs for neurodegenerative disorders," *Current Neuropharmacol.*, vol. 16, no. 6, pp. 740–748, Jun. 2018.

[2] G. Schneider, "Automating drug discovery," *Nature Rev. Drug Discovery*, vol. 17, no. 2, pp. 97–113, Feb. 2018.

[3] R. S. Ani, B. Anand, and O. S. Deepa, "In silico prediction tool for drug-likeness of compounds based on ligand based screening," *Int. J. Res. Pharmaceutical Sci.*, vol. 11, no. 4, pp. 6273–6281, Oct. 2020.

[4] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, "Applications of machine learning in drug discovery and development," *Nature Rev. Drug Discovery*, vol. 18, no. 6, pp. 463–477, 2019.

[5] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, "Artificial intelligence in drug discovery and development," *Drug Discovery Today*, vol. 26, no. 1, pp. 80–93, Jan. 2021.

[6] A. Lavecchia, "Machine-learning approaches in drug discovery: Methods and applications," *Drug Discovery Today*, vol. 20, no. 3, pp. 318–331, Mar. 2015.

[7] P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F. J. Novoa, A. Carballal, V. Maojo, A. Pazos, and C. Fernandez-Lozano, "A review on machine learning approaches and trends in drug discovery," *Comput. Structural Biotechnol. J.*, vol. 19, pp. 4538–4558, Aug. 2021.

[8] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," *Drug Discovery Today*, vol. 23, no. 8, pp. 1538–1546, Aug. 2018.

[9] K. Martinez-Mayorga, A. Madariaga-Mazon, J. L. Medina-Franco, and G. Maggiora, "The impact of chemoinformatics on drug discovery in the pharmaceutical industry," *Expert Opinion Drug Discovery*, vol. 15, no. 3, pp. 293–306, Mar. 2020.

[10] J. Xu and A. Hagler, "Chemoinformatics and drug discovery," *Molecules*, vol. 7, no. 8, pp. 566–600, Aug. 2002.

[11] R. Ani, R. Manohar, G. Anil, and O. S. Deepa, "Virtual screening of drug likeness using tree based ensemble classifier," *Biomed. Pharmacol. J.*, vol. 11, no. 3, pp. 1513–1519, Sep. 2018.

[12] P. B. Jayaraj, S. Sanjay, K. Raja, G. Gopakumar, and U. C. Jaleel, "Ligand based virtual screening using self-organizing maps," *Protein J.*, vol. 41, no. 1, pp. 44–54, Feb. 2022.

[13] F. D. Botelho, M. C. dos Santos, A. D. S. Gonçalves, K. Kuca, M. Valis, S. R. LaPlante, T. C. C. França, and J. S. F. D. de Almeida, "Ligand-based virtual screening, molecular docking, molecular dynamics, and MM-PBSA calculations towards the identification of potential novel ricin inhibitors," *Toxins*, vol. 12, no. 12, p. 746, Nov. 2020.

[14] H. Li, K.-H. Sze, G. Lu, and P. J. Ballester, "Machine-learning scoring functions for structure-based virtual screening," *Wiley Interdiscipl. Rev., Comput. Mol. Sci.*, vol. 11, no. 1, p. e1478, 2021.

[15] C. Garcia-Hernandez, A. Fernández, and F. Serratosa, "Ligand-based virtual screening using graph edit distance as molecular similarity measure," *J. Chem. Inf. Model.*, vol. 59, no. 4, pp. 1410–1421, Apr. 2019.

[16] S. A. Hooshmand, S. A. Jamalkandi, S. M. Alavi, and A. Masoudi-Nejad, "Distinguishing drug/non-drug-like small molecules in drug discovery using deep belief network," *Mol. Diversity*, vol. 25, pp. 827–838, May 2021.

[17] I. Süntar, "Importance of ethnopharmacological studies in drug discovery: Role of medicinal plants," *Phytochemistry Rev.*, vol. 19, no. 5, pp. 1199–1209, Oct. 2020.

[18] P. Brissot, A. Pietrangelo, P. C. Adams, B. De Graaff, C. E. McLaren, and O. Loréal, "Haemochromatosis," *Nature Rev. Disease Primers*, vol. 4, no. 1, pp. 1–15, 2018.

[19] S. M. H. Mahmud, W. Chen, H. Jahan, Y. Liu, N. I. Sujan, and S. Ahmed, "IDTi-CSsmoteB: Identification of drug–target interaction based on drug chemical structure and protein sequence using XGBoost with over-sampling technique SMOTE," *IEEE Access*, vol. 7, pp. 48699–48714, 2019.

[20] S. Ben-Shabat, L. Yarmolinsky, D. Porat, and A. Dahan, "Antiviral effect of phytochemicals from medicinal plants: Applications and drug delivery strategies," *Drug Del. Translational Res.*, vol. 10, no. 2, pp. 354–367, Apr. 2020.

[21] U. Payyappallimana, K. Patwardhan, P. Mangalath, C. S. Kessler, R. Jayasundar, A. Kizhakkeveettil, A. Morandi, and R. Puthiyedath, "The COVID-19 pandemic and the relevance of ayurveda's whole systems approach to health and disease management," *J. Alternative Complementary Med.*, vol. 26, no. 12, pp. 1089–1092, Dec. 2020.

[22] U. Anand, N. Jacobo-Herrera, A. Altemimi, and N. Lakhssassi, "A comprehensive review on medicinal plants as antimicrobial therapeutics: Potential avenues of biocompatible drug discovery," *Metabolites*, vol. 9, no. 11, p. 258, Nov. 2019.

[23] C. Veeresham, "Natural products derived from plants as a source of drugs," *J. Adv. Pharmaceutical Technol. Res.*, vol. 3, no. 4, p. 200, 2012.

[24] B. R. Bacon, P. C. Adams, K. V. Kowdley, L. W. Powell, and A. S. Tavill, "Diagnosis and management of hemochromatosis: 2011 practice guideline by the American association for the study of liver diseases," *Hepatology*, vol. 54, no. 1, pp. 328–343, Jul. 2011.

[25] M. S. Chang and B. N. Smith, "Hereditary hemochromatosis," *DeckerMed Med.*, vol. 26, pp. 251–270, Nov. 2016.

[26] M. A. Miller, "Chemical database techniques in drug discovery," *Nature Rev. Drug Discovery*, vol. 1, no. 3, pp. 220–227, Mar. 2002.

[27] M. Elbadawi, S. Gaisford, and A. W. Basit, "Advanced machine-learning techniques in drug discovery," *Drug Discovery Today*, vol. 26, no. 3, pp. 769–777, Mar. 2021.

[28] J. Jiménez-Luna, F. Grisoni, N. Weskamp, and G. Schneider, "Artificial intelligence in drug discovery: Recent advances and future perspectives," *Expert Opinion Drug Discovery*, vol. 16, no. 9, pp. 949–959, Sep. 2021.

[29] A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, and C. T. Supuran, "Natural products in drug discovery: Advances and opportunities," *Nature Rev. Drug Discovery*, vol. 20, no. 3, pp. 200–216, Mar. 2021.

[30] L. A. Machado, E. Krempser, and A. C. R. Guimarães, "A machine learning-based virtual screening for natural compounds capable of inhibiting the HIV-1 integrase," *Frontiers Drug Discovery*, vol. 2, Oct. 2022.

[31] K. A. Carpenter and X. Huang, "Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: A review," *Current Pharmaceutical Design*, vol. 24, no. 28, pp. 3347–3358, Dec. 2018.

[32] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan, "Machine learning in drug discovery: A review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 1947–1999, 2022.

[33] L. Pinzi and G. Rastelli, "Molecular docking: Shifting paradigms in drug discovery," *Int. J. Mol. Sci.*, vol. 20, no. 18, p. 4331, Sep. 2019.

[34] B. J. Neves, R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov, and C. H. Andrade, "QSAR-based virtual screening: Advances and applications in drug discovery," *Frontiers Pharmacol.*, vol. 9, p. 1275, Nov. 2018.

[35] E. H. B. Maia, L. C. Assis, T. A. de Oliveira, A. M. da Silva, and A. G. Taranto, "Structure-based virtual screening: From classical to artificial intelligence," *Frontiers Chem.*, vol. 8, p. 343, Apr. 2020.

[36] T. B. Kimber, Y. Chen, and A. Volkamer, "Deep learning in virtual screening: Recent applications and developments," *Int. J. Mol. Sci.*, vol. 22, no. 9, p. 4435, Apr. 2021.

[37] A. M. Andrianov, Y. V. Kornoushenko, A. D. Karpenko, I. P. Bosko, and A. V. Tuzikov, "Computational discovery of small drug-like compounds as potential inhibitors of SARS-CoV-2 main protease," *J. Biomolecular Struct. Dyn.*, vol. 39, no. 15, pp. 5779–5791, Oct. 2021.

[38] Z. Liu, J. Du, J. Fang, Y. Yin, G. Xu, and L. Xie, "DeepScreening: A deep learning-based screening web server for accelerating drug discovery," *Database*, vol. 2019, Jan. 2019, Art. no. baz104.

[39] N. M. O'Boyle, "Towards a universal SMILES representation—A standard method to generate canonical SMILES based on the InChI," *J. Cheminformatics*, vol. 4, no. 1, pp. 1–14, Dec. 2012.

[40] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988.

[41] D. Weininger, A. Weininger, and J. L. Weininger, "SMILES. 2. Algorithm for generation of unique SMILES notation," *J. Chem. Inf. Comput. Sci.*, vol. 29, no. 2, pp. 97–101, 1989.

[42] D. Bajusz, A. Rácz, and K. Héberger, "Chemical data formats, fingerprints, and other molecular descriptions for database analysis and searching," in *Comprehensive Medicinal Chemistry III*, vol. 3. Amsterdam, The Netherlands: Elsevier, 2017, p. 8.

[43] [Online]. Available: https://datagrok.ai/help/domains/chem/fingerprints

[44] V. S. S. Kumar, K. Aparna, R. Ani, and O. S. Deepa, "Ensemble machine learning approaches in molecular fingerprint based virtual screening," in *Proc. 2nd Global Conf. Advancement Technol. (GCAT)*, Oct. 2021, pp. 1–6.

[45] M. Kokabi, M. Donnelly, and G. Xu, "Benchmarking small-dataset structure-activity-relationship models for prediction of Wnt signaling inhibition," *IEEE Access*, vol. 8, pp. 228831–228840, 2020.

[46] L. T. Afolabi, F. Saeed, H. Hashim, and O. O. Petinrin, "Ensemble learning method for the prediction of new bioactive molecules," *PLoS ONE*, vol. 13, no. 1, Jan. 2018, Art. no. e0189538.

[47] B. R. Manju and A. R. Nair, "Classification of cardiac arrhythmia of 12 lead ECG using combination of SMOTEENN, XGBoost and machine learning algorithms," in *Proc. 9th Int. Symp. Embedded Comput. Syst. Design (ISED)*, Dec. 2019, pp. 1–7.

[48] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms," *IEEE Access*, vol. 10, pp. 39700–39715, 2022.

[49] W. Torng and R. B. Altman, "Graph convolutional neural networks for predicting drug-target interactions," *J. Chem. Inf. Model.*, vol. 59, no. 10, pp. 4131–4149, Oct. 2019.

[50] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009.

R. ANI is currently an Assistant Professor (Sl. Gr.) and the Vice Chairperson of the School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri.

O. S. DEEPA is currently an Associate Professor with the Department of Mathematics, Amrita Vishwa Vidayapeetham, Coimbatore. Her research interests include statistical quality control, machine learning, and bioinformatics.