## DRUG DISCOVERY USING MACHINE LEARNING

## A Capstone Project Report submitted in partial fulfilment of the requirements for the award of the degree of,

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)**

Submitted by:

|  |  |
| --- | --- |
| **N B Kundan Setty** | **BU21CSEN0500116** |
| **D Pranay Chandu** | **BU21CSEN0500120** |
| **O Veda Prabhas** | **BU21CSEN0500178** |

**Under the guidance of**

**Dr. Monica Jha**

**Assistant Professor**



**Department of Artificial Intelligence and Data Science,**

**GITAM SCHOOL OF TECHNOLOGY**

**GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT**

**(Deemed to be University)**

**Bengaluru Campus**

**March 2025**

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**GITAM SCHOOL OF TECHNOLOGY**

**GITAM**

**(Deemed to be University)**

**DECLARATION**

We, hereby, declare that the project report entitled **“DRUG DISCOVERY USING MACHINE LEARNING”** is an original work done in the **Department of Artificial Intelligence and Data Science, GITAM School of Technology, GITAM (Deemed to be University), Bengaluru,** submitted in partial fulfilment of the requirements for the award of the degree of **B.Tech.** in Artificial Intelligence and Data Science. The work has not been submitted to any other college or University for the award of any degree.

**Date:**

|  |  |  |
| --- | --- | --- |
| **Registration No(s).**  BU21CSEN0500116  BU21CSEN0500120  BU21CSEN0500178 | **Name(s)**  **N B Kundan Setty**  **D Pranay Chandu**  **O Veda Prabhas** | **Signature(s)** |

****

****

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**GITAM SCHOOL OF TECHNOLOGY GITAM**

**(Deemed to be University)**

**CERTIFICATE**

This is to certify that the project report entitled **“DRUG DISCOVERY USING MACHINE LEARNING”** is a Bonafide record of work carried out by **N B Kundan Setty (BU21CSEN0500116), D Pranay Chandu (BU21CSEN0500120), O Veda Prabhas (BU21CSEN0500178),** submitted in partial fulfilment of requirement for the award of the degree of **Bachelors of Technology in Computer Science and Engineering (Data Science)**.

|  |  |
| --- | --- |
| **Project Guide** | **Head of the Department** |
| **Dr. Monica Jha** | **Prof. A. Vadivel** |
| Assistant Professor | Professor |
|  |  |

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose consistent guidance and encouragement crowned our efforts with success.

We consider it our privilege to express our gratitude to all those who guided us in the completion of the project.

We express our gratitude to Director Prof. **Basavaraj Gundappa Katageri** for having provided us with the golden opportunity to undertake this project work in their esteemed organization.

We sincerely thank **Prof. A. Vadivel,** HOD, Department of Artificial Intelligence and Data Science, Gandhi Institute of Technology and Management, Bengaluru, for the immense support given to us.

We express our gratitude to our project guide, **Dr. Monica Jha,** Assistant Professor, Department of Artificial Intelligence and Data Science, Gandhi Institute of Technology and Management, Bengaluru, for their support, guidance, and suggestions throughout the project work.

|  |  |
| --- | --- |
| **Registration No(s).**  BU21CSEN0500116  BU21CSEN0500120  BU21CSEN0500178 | **Name(s)**  **N B Kundan Setty**  **D Pranay Chandu**  **O Veda Prabhas** |

**ABSTRACT**

Millions of people worldwide struggle with Parkinson's disease (PD), a progressive neurodegenerative disease that lowers quality of life and causes motor impairments and cognitive decline. Tremors, stiffness, and bradykinesia are some of the symptoms that are mainly caused by the degeneration of dopamine-producing neurons in the substantia nigra. Even with extensive research, Parkinson's disease is still incurable, and current therapies target symptoms rather than the underlying cause. Timely intervention depends on an early and accurate diagnosis, but traditional diagnostic techniques mainly rely on subjective clinical evaluations, which can cause delays in diagnosis. Furthermore, the intricacy of molecular interactions and the challenge of determining efficacious therapeutic targets make drug discovery for Parkinson's disease (PD) particularly challenging.

This study makes use of developments in machine learning (ML) and artificial intelligence (AI) to enhance the processes of drug discovery and biomarker identification in Parkinson's disease. Gene expression information was gathered from the "Parkinson's disease: substantia nigra (HG-U133B)" and SN133-B datasets in the Gene Expression Omnibus (GEO) database. Data preprocessing included standardizing the dataset, eliminating entries that lacked gene symbols, and using Genemania to transform gene expression data into network representations. Three clustering techniques—Graph Attention Autoencoder (GAE), Spectral Clustering, and Fuzzy C-Means (FCM)—were used to categorize genes according to their patterns of expression. The outputs from individual clustering techniques were then integrated using ensemble clustering via Evidence Accumulation Clustering (EAC), producing strong and biologically significant clusters. This method led to the identification of important hub genes CDK4 and EZH2 as possible PD biomarkers and treatment targets.

The GraphDTA model was used to predict binding affinity in order to investigate potential therapeutic avenues further. This deep learning framework uses hub genes as sequence embeddings and Graph Neural Networks (GNNs) to represent chemical compounds as molecular graphs. Protein sequences were one-hot encoded, and drug molecules were represented using SMILES strings that were then transformed into molecular graphs using RDKit. The Davis and KIBA datasets were used to train the GraphDTA model, which was then optimized using the Adam optimizer and Mean Squared Error (MSE) loss. The binding affinities (pKd values) between identified hub genes and potential drug molecules were predicted by the trained model.

The results of this study demonstrate how applying machine learning techniques to PD research can speed up drug discovery, increase biomarker identification, and improve diagnostic accuracy. Utilizing computational models can help researchers better understand the molecular mechanisms underlying Parkinson's disease (PD), speed up the creation of tailored treatments, and eventually enhance patient outcomes. Future research will concentrate on improving clustering algorithms, adding longitudinal data to datasets, and using clinical trials and lab testing to validate the computational predictions.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Title** | **Page No.** |
| Declaration | ii |
| Certificate | iii |
| Acknowledgement | iv |
| Abstract | v |
| Table of Contents | vii |
| List of Figures | ix |
| List of Tables | x |
| 1. Introduction | 1 |
| 2. Literature Survey | 2 |
| 3. Software and Hardware Specifications | 5 |
| 3.1 Hardware Requirement | 5 |
| 3.2 Software Requirement | 5 |
| 4. Problem Statement | 6 |
| 4.1 Objectives | 6 |
| 5. Methodologies | 7 |
| 5.1 Dataset Collection | 7 |
| 5.2 Data Pre-Processing | 7 |
| 5.3 Machine Learning Algorithms | 7 |
| 5.3.1 Graph Attention Autoencoder | 7 |
| 5.3.2 Spectral Clustering | 8 |
| 5.3.3 Fuzzy C-Means | 8 |
| 5.3.4 Evidence Accumulation Clustering (EAC) | 8 |
| 5.3.5 GraphDTA (Drug Target Affinity) | 8 |
| 6. Implementation | 9 |
| 6.1 Dataset Collection and Pre-processing | 9 |
| 6.2 Clustering Algorithms and Gene Classification | 10 |
| 6.2.1. Graph Attention Autoencoder (GAE) Clustering | 10 |
| 6.2.2. Spectral Clustering | 11 |
| 6.2.3. Fuzzy C-Means Clustering | 12 |
| 6.3 Ensemble Clustering - Evidence Accumulation Clustering (EAC) | 12 |
| 6.3.1 Load Clustering Results | 12 |
| 6.3.2 Construct the Co-Occurrence Matrix | 12 |
| 6.4 Validation | 13 |
| 6.5 Identification of Hub Genes | 14 |
| 6.5.1 Hub genes in Parkinson’s Disease Network | 14 |
| 6.5.1.2 Significance Assessment of Hub Genes | 15 |
| 6.5.1.3 Visualization of Hub Genes | 15 |
| 6.6 Chemical dataset | 16 |
| 6.7 Binding Affinity Prediction using GraphDTA | 16 |
| 6.7.1 Methodologies | 17 |
| 6.7.1.1 Drug Representation | 17 |
| 6.7.1.2 Protein Representation | 17 |
| 6.7.2 Model Development and Training | 17 |
| 6.7.2.1 Model Architecture | 18 |
| 6.7.2.2 Feature Representation | 18 |
| 6.7.3 Training Process | 18 |
| 6.7.3.1 Role of KIBA and Davis Datasets in Binding Affinity Prediction | 19 |
| 7. Results | 20 |
| 8. Limitation | 24 |
| 9. Conclusion | 25 |
| CO-PO-PSOOutcomes | 26 |
| References | 29 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure** | **Title** | **Page No.** |
|  | Work Flow Diagram | 9 |
|  | CDK4 Network Diagram | 15 |
|  | SMILES | 16 |
|  | Process for Converting Smiles to Graph | 17 |
|  | Fuzzy Clustering Algorithm results | 21 |
|  | GAE Clustering Algorithm results | 21 |
|  | Spectral Clustering Algorithm results | 22 |
|  | EAC Algorithm results | 22 |
|  | Validation for best cluster | 22 |
|  | Sorting of Hub Genes | 23 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table** | **Title** | **Page No.** |
|  | Dataset Details | 7 |
|  | Sample Data | 10 |
|  | Mean of P value and Q value for (GAE) Clustering [<0.5] | 11 |
|  | Mean of P value and Q value for Spectral Clustering [<0.5] | 11 |
|  | Mean of P value and Q value for Fuzzy C-Means Clustering [<0.5] | 12 |
|  | Degree Centrality Scores | 14 |
|  | Confirms the role of hub genes in target diseases. | 15 |
|  | Chemical dataset | 16 |
|  | KIBA and Davis Datasets | 18 |
|  | Binding Affinity Scores of CDK4 and EZH2 in Davis and KIBA Datasets | 20 |
|  | Potential Drug Candidates for Treating Parkinson’s Disease: Targeting CDK4 and EZH2 | 20 |

**1. INTRODUCTION**

Millions of people worldwide suffer from Parkinson's disease (PD), a progressive neurodegenerative illness that lowers quality of life and causes motor impairments and cognitive decline. Degeneration of dopamine-producing neurons in the substantia nigra is the main cause of Parkinson's disease (PD), which manifests as bradykinesia, rigidity, and tremors. For PD to be effectively treated, early and precise detection is essential. However, clinical evaluations are a major component of traditional diagnostic techniques, and they can be arbitrary and slow.

Recent developments in machine learning (ML) and artificial intelligence (AI) have revolutionized biomedical research by offering data-driven methods for diagnosing illnesses and creating treatments. In order to identify subtle patterns suggestive of Parkinson's disease early on, machine learning models analyze complex datasets, such as voice recordings, medical imaging, and biomarker data. In addition, machine learning (ML) is essential to drug discovery because it speeds up the process of finding possible therapeutic compounds by optimizing molecular structures, predicting drug-target interactions, and evaluating drug efficacy.

In order to ascertain how strongly a drug (ligand) binds to its target protein, binding affinity prediction is crucial in the drug discovery process. The GraphDTA model, a deep learning-based framework that depicts drug molecules as molecular graphs and proteins as sequence embeddings, is one of the most promising methods for predicting drug-target binding affinity. In order to predict affinity scores, the model integrates features from protein sequences with those extracted from molecular graphs using Graph Neural Networks (GNNs).

**2. LITERATURE SURVEY**

### A thorough analysis of recent studies using artificial intelligence (AI) and machine learning (ML) in drug discovery and early neurodegenerative disease detection, with an emphasis on Parkinson's disease (PD), is given in the literature review. This section describes important research, providing details on methods, results, and implications for disease progression tracking, drug-target interaction prediction, and biomarker identification.

### 2.1. **Artificial Intelligence and Machine Learning‐Aided Drug Discovery**

Vatansever et al. (2021)[15] Reviews of Medicinal Research Target identification, lead optimization, and preclinical development were the stages at which Vatansever et al. examined the incorporation of machine learning (ML) in the drug discovery pipeline. With a focus on their function in forecasting molecular characteristics, toxicity, and drug-likeness, the study examined models like random forests (RFs), support vector machines (SVMs), and deep neural networks (DNNs). The use of machine learning algorithms in virtual screening, which drastically lowers the time and expense associated with early-stage drug development, was a major highlight. Along with discussing issues like model interpretability and data scarcity, the authors suggested ensemble learning as a possible remedy.

2.2. **Applications of Machine Learning in Drug Discovery and Development**

Vamathevan et al. (2019)[8], Drug Discovery in Nature Reviews This thorough analysis explained how machine learning speeds up different stages of drug discovery. The authors looked at reinforcement learning for molecule optimization, unsupervised approaches for pattern recognition in chemical libraries, and supervised learning techniques for target identification. The study included case studies where ML models improved clinical trial predictions and identified new drug candidates. The necessity of high-quality datasets and interdisciplinary cooperation between pharmacologists and data scientists were among the difficulties mentioned.

2.3. **Identification of Endoplasmic Reticulum Stress-Related Biomarkers in Diabetic Nephropathy**

Su et al. (2023)[12], Endocrinology's frontiers Using bioinformatics and machine learning models to examine microarray data, Su et al. examined the part endoplasmic reticulum (ER) stress plays in diabetic nephropathy. They found important genes linked to ER stress using clustering algorithms like k-means and hierarchical clustering. Their techniques, which included pathway analysis and gene ontology, shed light on the role that ER stress pathways play in the development of the disease and provided a methodological basis for the identification of PD biomarkers.

2.4. **Early Detection of Alzheimer’s Disease Using Support Vector Machines (SVM)**

Eke et al. (2020)[11], The IEEE Journal of Health and Biomedical Informatics Eke et al. concentrated on creating an SVM-based model to identify non-amyloid blood biomarkers for Alzheimer's disease (AD). By evaluating six protein biomarkers (A2M, ApoE, BNP, Eot3, RAGE, and SGOT), they were able to detect AD in its early stages with high classification accuracy. This method offers similarities for PD early detection techniques and highlights the potential of machine learning in non-invasive diagnostics.

2.5. **Evaluation of Cerebrospinal Fluid Proteins for Early Parkinson’s Disease Diagnosis**

T. dos Santos et al. (2018)[5], One PLoS In order to find early PD biomarkers, this study assessed the proteins in cerebrospinal fluid (CSF). Two cohorts were examined: 30 patients and replication controls in the first cohort, which included 80 early-stage PD patients and 80 controls. Important results showed that tau and α-synuclein proteins varied significantly between groups, supporting their diagnostic utility.

2.6. **Classification of Neurodegenerative Disorders with Multiplex Blood Biomarkers**

Lin et al. (2020)[6], The International Journal of Molecular Sciences In order to differentiate between Parkinson's disease, Alzheimer's disease, and frontotemporal dementia, Lin et al. created an ML-based classification model for neurodegenerative diseases using multiplex blood biomarkers. With a 92% overall classification accuracy, their model showed a correlation between α-synuclein levels and the severity of dementia and Parkinson's disease.

2.7. **Machine Learning Algorithm for Neurological Disease Prediction Using Metabolic Biomarkers**

Han et al. (2024)[7], Chimica Clinica Acta In order to diagnose neurological diseases early, Han et al. presented a predictive machine learning algorithm that combines circulating metabolic biomarkers with clinical risk factors. When compared to conventional techniques, the model increased prediction accuracy by 9.8%, demonstrating the value of metabolic profiling in Parkinson's disease research.

### 2.8. **Imaging in Neurodegenerative Disorders**

Saba (2015)[9], Oxford University Press In her research, Saba focused on neuroimaging biomarkers, specifically FDG-PET and dopamine transporter (DAT) imaging, for the diagnosis of Parkinson's disease. Tracking the progression of a disease was shown to be more accurate when imaging data was integrated with machine learning algorithms.

### 2.9. **Inflammatory Mechanisms in Parkinson’s Disease: From Pathogenesis to Targeted Therapies**

Lee et al. (2021)[10], The neuroscientist Lee et al. looked into cytokines and C-reactive protein (CRP), two inflammatory biomarkers in Parkinson's disease. Correlations between biomarker levels and disease progression were found by the study using machine learning models to analyze data related to inflammation.

2.10. **Plasma Proteomics Identify Biomarkers Predicting Parkinson’s Disease Before Symptom Onset**

Hällqvist et al. (2024)[13], Communications in Nature According to this study, PD can be predicted using plasma proteomics and machine learning up to seven years before motor symptoms appear. After examining more than 1,200 plasma proteins, researchers discovered a panel of biomarkers with an 85% predictive accuracy.

2.11. **Study of Genes Associated with Parkinson’s Disease Using Feature Selection**

Rafieipour et al. (2020)[14], The Journal of Bioengineering Research Rafieipour et al. used feature selection methods, such as SVMs and HSIC-Lasso, to find genes linked to Parkinson's disease. With a prediction accuracy of 94.9%, the combined model demonstrated the importance of gene expression analysis in the early detection of Parkinson's disease.

**3. SOFTWARE AND HARDWARE SPECIFICATIONS**

For the successful execution of our Drug Discovery project using Machine Learning, the following Hardware and Software resources are required:

**3.1 Hardware Requirement**

**3.1.1 Minimum Hardware Requirements**

To run basic tasks and pre-process data locally before executing on cloud platforms, a system with the following specifications is recommended:

* Processor: Intel Core i5
* RAM: 8GB or above
* Storage: 512GB SSD
* Operating System: Windows 10/11
* GPU: NVIDIA GTX 1080 or higher (recommended for faster training)

**3.1.2 Cloud-Based Hardware (Google Colab)**

Since our project is executed on Google Colab, we utilize its built-in processor, CPU, and GPU, which help efficiently execute machine learning models and handle large-scale computations.

**3.2 Software Requirements**

**3.2.1 Programming Languages & Frameworks**

* Python (Version 3.8+) – Primary programming language for model development
* TensorFlow (Version 2.8+) – Used for training and deploying deep learning models
* PyTorch (Version 1.10+) – Alternative deep learning framework for experimentation

**3.2.2 Development Tools & Libraries**

* Google Colab – IDEs for writing and debugging code
* Scikit-learn – Machine learning library for evaluation metrics and preprocessing
* Matplotlib & Seaborn – Data visualization libraries for performance analysis

**4. PROBLEM STATEMENT**

There is currently no cure for Parkinson's disease (PD), a neurodegenerative condition for which there are only symptom management options. Determining how well a compound binds to and influences a biomarker after it has been identified is one of the main obstacles in the drug discovery process for Parkinson's disease. Finding an efficient molecular target is made more difficult by the disease's complexity, which is further increased by a number of contributing factors such as α-synuclein aggregation, mitochondrial dysfunction, and a drop in dopamine levels.

**4.1 Objective:**

The objectives of this project are threefold.

* To identify the biomarkers associated with Parkinson's disease and determine the most significant ones for drug targeting using machine learning techniques.
* To represent the chemical compounds and biomarkers in a graphical format. This step helps in understanding their structure and relationships.
* To Predict the interaction between these chemical compounds and the biomarkers through the application of machine learning models.

**5. METHODOLOGY**

* 1. **Dataset collection**
* The National Centre for Biotechnology Information (NCBI) provided us with a Gene Expression Omnibus (GEO) dataset.
* The medial and lateral Substantia Nigras (SNs) from post-mortem brain samples taken from patients with sporadic Parkinson's disease (PD) are analysed in this dataset. In PD, the SN shows significant tissue damage. This dataset contains conditions as attributes and genes as records.
* Samples from both Parkinson's disease patients and control subjects were obtained in various brain regions.

|  |
| --- |
| **Table 1**: Dataset Details |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **Dataset Title** | **Species** | **Size** | **Availability** |
| 1 | Parkinson's disease: substantia nigra (HG-U133A) | Homo Sapiens | 22283 \* 47 | http://www.ncbi.nlm.nih.gov/geo |

**5.2 Data Pre-processing**

* Not all of the gene symbols are here. A few are vacant. Therefore, those null value records must be deleted.
* To make sure the data values fall within a specific range, we standardized the dataset.
  1. **Machine Learning Algorithms**
     1. **Graph Attention Autoencoder for Clustering:**

A deep learning model called Graph Attention Autoencoder (GAE) [17] uses attention mechanisms to learn graph embeddings. By maintaining structural and feature information, it enhances clustering and captures intricate relationships between nodes. To improve clustering accuracy, the model reconstructs the graph and encodes node representations. It is extensively employed in biomarker identification and biological network analysis.

* + 1. **Spectral Clustering:**

In order to identify clusters, spectral clustering [18] converts data into a graph-based representation and uses eigenvalue decomposition. By employing the similarity between data points rather than distances, it successfully captures complex structures. Groups in non-linearly separable data can be found using this technique. It is frequently used in biomedical research to find patterns linked to disease.

* + 1. **Fuzzy C-Means:**

A soft clustering algorithm called fuzzy C-Means (FCM) assigns each data point to one of several clusters with differing degrees of membership. By iteratively updating cluster z according to weighted memberships, it reduces intra-cluster variance. When dealing with biological data uncertainty, this approach is helpful. It is frequently employed in biomarker clustering and medical image segmentation.

* + 1. **Evidence Accumulation Clustering**

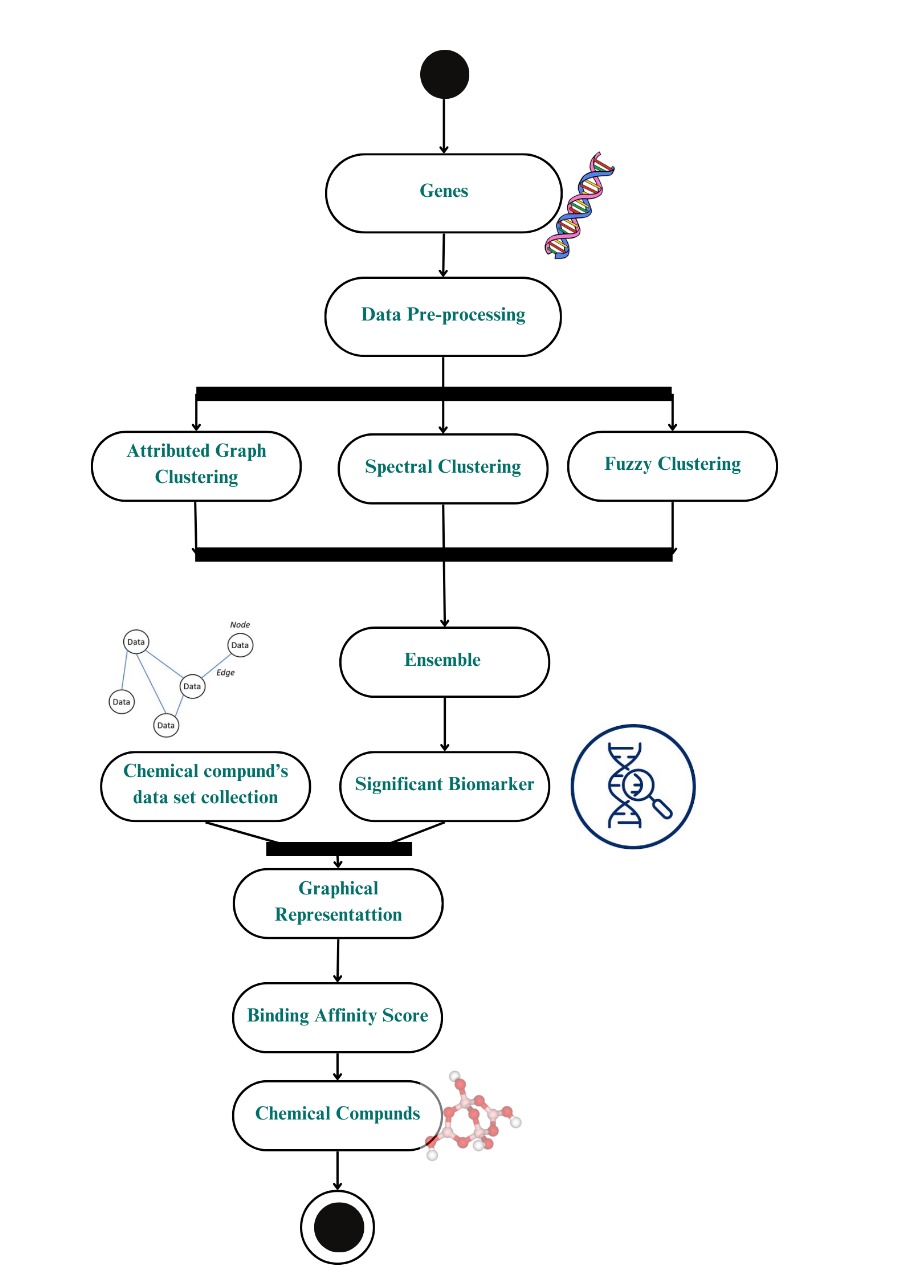
An ensemble clustering technique called EAC [16] combines several clustering results to create a final partition that is more resilient. To find groups, it uses hierarchical clustering and builds a co-occurrence matrix based on several clustering algorithms. This method increases the accuracy and stability of clustering. It helps with the integration of multi-omics and genomics data.

* + 1. **GraphDTA (Drug Target Affinity)**

GraphDTA [19] is a deep learning model that predicts the binding affinity between a drug and a protein. Instead of using traditional methods, it represents drugs as graphs (where atoms are nodes and bonds are edges) and processes proteins as sequences.

**6. IMPLEMENTATION**

The project's workflow is shown in Figure 1. The first step is gathering and preprocessing datasets about chemical compounds and biomarkers. To find important biomarkers, a variety of clustering techniques are used, including fuzzy clustering, spectral clustering, and attribute graph clustering. The outcomes of these techniques are then combined using an ensemble approach. These biomarkers and compounds are shown as graphs once chemical compound data has been gathered. To identify possible medication candidates for the treatment of Parkinson's disease, their binding affinity is then computed.



**Figure 1:** Work Flow Diagram

**6.1 Dataset Collection and preprocessing:**

The **"Parkinson's disease: substantia nigra (HG-U133A)"** dataset was taken from the **Gene Expression Omnibus (GEO) database (NCBI)**. It includes **gene expression** information from control groups and post-mortem brain samples of patients with Parkinson's disease (PD).

* Gene symbols that were missing were eliminated.
* To ensure uniformity, standardization was implemented.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **IDENTIFIER** | **GSM208622** | **GSM208623** | **GSM208624** | **GSM208630** | **GSM208631** |
| MIR4640 | 8.98544 | 7.74586 | 9.13177 | 9.31831 | 9.32942 |
| RFC2 | 4.43529 | 4.72744 | 4.70603 | 4.57991 | 4.73426 |
| HSPA6 | 4.02411 | 3.67353 | 3.94171 | 3.91261 | 4.02555 |
| PAX8 | 6.83235 | 7.30323 | 7.03826 | 7.18283 | 7.25938 |
| GUCA1A | 2.38701 | 2.51031 | 2.46309 | 2.43942 | 2.50131 |
| MIR5193 | 4.52135 | 4.61913 | 4.41992 | 4.72883 | 5.13425 |
| THRA | 4.16161 | 4.0081 | 3.91617 | 3.73846 | 4.20772 |
| PTPN21 | 2.68369 | 2.88248 | 2.77961 | 2.71665 | 2.82834 |
| CCL5 | 4.63534 | 5.06678 | 4.8123 | 4.87338 | 5.0046 |
| CYP2E1 | 3.17998 | 3.33354 | 3.16788 | 3.12864 | 3.23629 |
| EPHB3 | 2.2264 | 2.50322 | 2.48516 | 2.3504 | 2.30292 |
| ESRRA | 6.34694 | 6.40319 | 6.21305 | 6.0889 | 6.32575 |
| CYP2A6 | 4.42131 | 4.62017 | 4.51177 | 4.46517 | 4.91537 |
| GAS6 | 7.28692 | 7.25647 | 7.29879 | 7.6391 | 7.73095 |
| MMP14 | 4.31004 | 4.61842 | 4.40917 | 4.36184 | 4.42506 |

**Table 2:** Sample Data

**6.2 Clustering Algorithms and Gene Classification:**

Three clustering techniques were applied to classify genes:

**6.2.1. Graph Attention Autoencoder (GAE) Clustering**

* **Prepare the Data:** Use the node features and the graph's structure (nodes and edges) as input.
* **Train the Autoencoder:** To learn embeddings, run data through a 10-layer deep network with leaky ReLU activation.
* **Reconstruct the Graph:** The decoder uses an inner product function to predict the graph's structure.
* **Cluster Nodes:** To create five clusters, extract hidden embeddings and use K-Means.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Cluster Number** | **Mean Of P Value** | **Mean of q values** |
| Graph Attention Autoencoder | 4 | 0.000414675 | 0.017610242 |
| 3 | 0.000858615 | 0.014617858 |
| 2 | 0.001025525 | 0.001025525 |

**Table 3:** Mean of P value and Q value for (GAE) Clustering [<0.5]

**6.2.2. Spectral Clustering**

* **Create the Affinity Matrix:** Use a Gaussian function to transform the graph into a similarity matrix.
* To capture significant connections, normalize the graph by computing the Laplacian Matrix.
* **Extract Features:** To obtain a meaningful representation, use eigenvectors (apart from the first one).

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Cluster Number** | **Mean Of P Value** | **Mean of q values** |
| Spectral Clustering | 4 | 0.000437247 | 0.017802736 |
| 3 | 0.000624991 | 0.017720559 |

* **Cluster the Data:** To create groups, apply K-Means to the transformed data. **Table 4:** Mean of P value and Q value for Spectral Clustering [<0.5]

**6.2.3. Fuzzy C-Means Clustering**

* **Gene Data Input:** Normalized gene expression values should be used.
* To begin clustering, set the following parameters: 5 clusters, 1000 iterations, and fuzziness = 2. Every gene is allocated to clusters with varying probabilities in order to iterate and update.
* **Finalize** Clusters: When there are negligible changes (threshold = 0.005), stop.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Cluster Number** | **Mean Of P Value** | **Mean of q values** |
| Fuzzy C-Means | 2 | 0.000106828 | 0.041711525 |
| 4 | 0.001105345 | 0.023123916 |
| 3 | 0.001368292 | 0.022438834 |

**Table 5:** Mean of P value and Q value for Fuzzy C-Means Clustering [<0.5]

**6.3 Ensemble Clustering - Evidence Accumulation Clustering (EAC)**

An ensemble clustering method called Evidence Accumulation Clustering (EAC) [16] combines the results of several clustering algorithms with various parameters. It creates a consensus clustering that emphasizes each cluster's unique features.

**6.3.1. Load Clustering Results:**

* The dataset is subjected to several clustering algorithms (such as Spectral) or iterations of the same algorithm. Every run yields a unique clustering solution, or data partition.

**6.3.2. Construct the Co-Occurrence Matrix:**

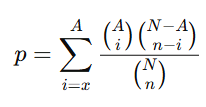
* To record how frequently two data points are grouped together across all clustering solutions, a co-association matrix is constructed.
* The Co-Occurrence matrix is clustered to produce the final consensus clusters.
* These clusters show strong patterns with biological significance.

Clustering outputs were successfully obtained by EAC, which found reliable and biologically significant clusters. Comparing cluster quality and biological relevance to individual algorithms, there is a noticeable improvement.

**6.4 Validation**

We verify the outcomes of different algorithms in order to assess the precision of the detected clusters. Gene Ontology-based functional enrichment is used for this validation. The best outcomes from various methods are compared after all three algorithms have been run. We use statistical metrics to assess each module's functional enrichment: q-value and p-value

The likelihood of discovering at least x genes linked to a particular annotation A in a module with n genes is represented by the P-value. The following formula is used to calculate it:



where A is the number of genes linked to a specific annotation, and n is the total number of genes in the module. To find gene interactions and investigate biological networks linked to Parkinson's disease, we employed Genemania [23].

The expected percentage of false positives among significant results is measured by the q-value, which stands for the False Discovery Rate (FDR).

The mean p-values and q-values for several clustering techniques, such as Graph Attention Autoencoder, Spectral Clustering, and Fuzzy C-means, are displayed in Tables 3, 4, and 5.

To make sure that only statistically significant clusters were taken into account, we set a threshold of p-value < 0.05. Clusters that failed to satisfy this criterion were eliminated from the final analysis.

The remaining clusters that made it past the cutoff are displayed in the table along with the matching p-values and q-values.

**6.5 Identification of Hub Genes**

Hub genes play a crucial role in biological processes because they are the most interconnected nodes in a gene network.

For Parkinson's disease, these genes are frequently important regulators, biomarkers, or possible therapeutic targets.

* Build a Gene Network: Using the EAC results, construct a network with nodes representing genes and edges representing interactions.
* Determine the Centrality Scores. Use Degree Centrality (number of connections) to identify key genes.

where n is the total number of genes in the module and is the gene's degree.

* Genes can be ranked by sorting them according to their Degree Centrality Scores.
* The following top hub genes have been identified:

['NONO', 'CDK4', 'TUBB', 'HNRNPF', 'MCM2', 'MCM7', 'MCM3', 'EZH2', 'RFC2', 'NUP205'].

**6.5.1 Hub genes in Parkinson’s Disease Network**

**Table 6:** Degree Centrality Scores

|  |  |
| --- | --- |
| **Gene** | **Degree Centrality** |
| NONO | 0.8409090909 |
| CDK4 | 0.8409090909 |
| TUBB | 0.7954545455 |
| HNRNPF | 0.7954545455 |
| MCM2 | 0.7954545455 |
| MCM7 | 0.7954545455 |
| MCM3 | 0.7727272727 |
| EZH2 | 0.7727272727 |
| RFC2 | 0.7518181818 |

**Table 6, shows the Degree Centrality score of key genes,**

Degree Centrality quantifies a gene's degree of network connectivity. A gene with a higher score interacts with more genes.

With the highest degree of centrality, NONO, CDK4 (0.8409090909) is an essential component of the gene interaction network.

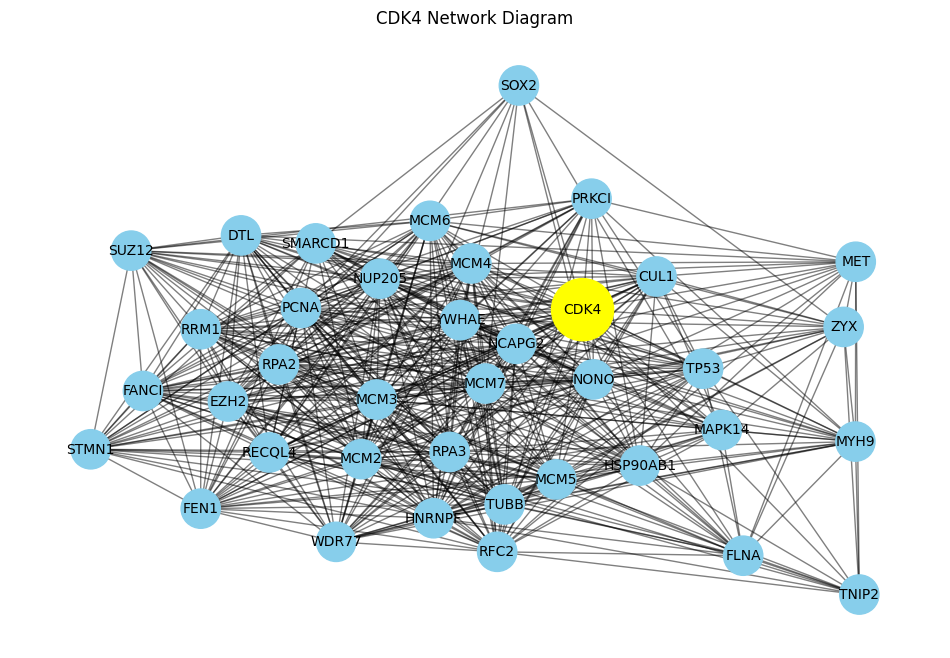
* + - 1. **Significance Assessment of Hub Genes**

To verify the involvement of hub genes in target diseases, we utilize a Gene Association Disease Literature Database like Disgenet [24].

**Table 7:** Confirms the role of hub genes in target diseases.

|  |  |
| --- | --- |
| **Gene Symbol** | **PubMed ID** |
| NONO | 38761794 |
| CDK4 | 32666227 |
| EZH2 | 31476446 |

**6.5.1.3 Visualization of Hub Genes**



**Figure 2:**  CDK4 Network Diagram

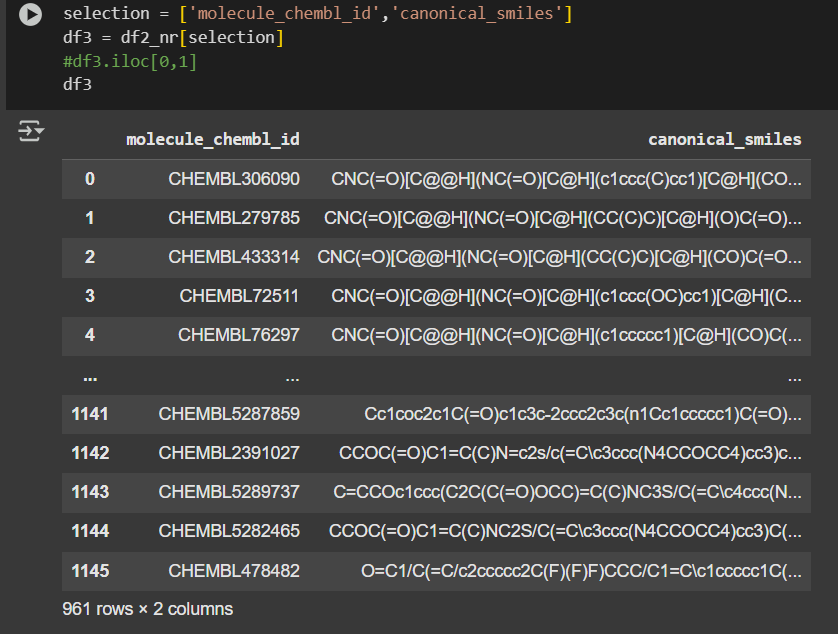
The central gene, **CDK4**, is represented by the **yellow node**; other genes that are connected to CDK4 are represented by the **blue nodes**; and interactions or relationships between these genes are shown by the **black lines (edges)**.

**6.6 Chemical dataset**

Using ChEMBL's APIs to retrieve chemical data, we were able to obtain SMILES (Simplified Molecular Input Line Entry System) representations, which make up the chemical dataset.

**Table 8:** Chemical Dataset

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Dataset Title** | **Species** |
| 1 | The chemical dataset that targets EZH2 from CHEMBL | Homo Sapiens |

****

**Figure 3:** SMILES (Simplified Molecular Input Line EntrySystem)

**6.7 Binding Affinity Prediction using GraphDTA**

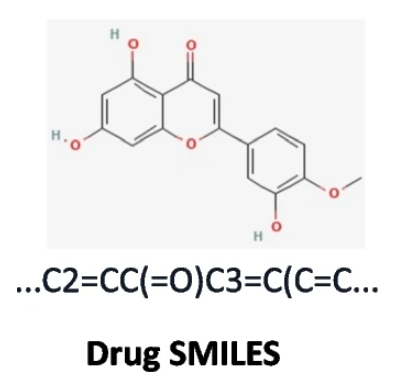
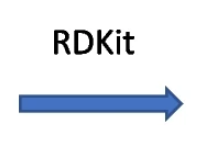
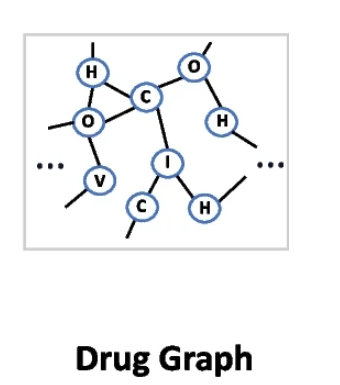
Binding affinity is a measure of how strongly a drug (small molecule) binds to its target protein. It tells us how well the drug can attach to the protein and stay connected. A higher binding affinity means a stronger interaction, which often leads to a more effective drug.

GraphDTA[19] is a deep learning model that predicts the binding affinity between a drug and a protein. Instead of using traditional methods, it represents drugs as graphs (where atoms are nodes and bonds are edges) and processes proteins as sequences.

**6.7.1 Methodologies**

**6.7.1.1 Drug Representation**

SMILES (Simplified Molecular Input Line Entry System), a text-based format that encodes molecular structures, is used to represent drugs.

Drugs are transformed into graphs, with bonds serving as edges and atoms as nodes. Atomic features and molecular graphs are extracted for deep learning models using the RDKit tool. **  **

**Figure 4 Process for Converting Smiles to Graph**

**6.7.1.2 Protein Representation**

One-hot encoding is used to represent proteins, in which a number is allocated to each amino acid. The amino acids that make up each protein sequence are denoted by distinct numbers and are extracted from UniProt[25].

Shorter sequences are filled with zeros, and the sequence is padded or truncated to 1000 residues. These sequences are converted into a 128-dimensional vector form by use of embedding layers.

Lastly, key features for binding prediction are extracted using max pooling and 1D convolutional layers

**6.7.2 Model Development and Training**

This section explains the process of designing, implementing, and training the model for binding affinity prediction. It covers how drug and protein representations are processed, the model’s architecture, the training procedure, and performance evaluation. The goal is to optimize the model for accurate prediction of drug-protein interactions.

**6.7.2.1 Model Architecture**

Protein sequences and drug graphs are processed by our Deep Learning Model (GraphDTA)[19] in order to predict binding affinity.

* Using graph neural networks (GNNs), which capture atomic interactions, drugs are represented as molecular graphs.
* One-hot encoding is used to represent proteins, while convolutional neural networks (CNNs) are used to process them.

The strength of the interaction between drugs and proteins is predicted by combining the outputs from both representations.

**6.7.2.2 Feature Representation**

* Drugs are converted into molecular graphs using RDKit, where atoms are nodes and bonds are edges. Each atom is described by a feature vector, including properties like atomic number, atomic mass, bonding and, aromaticity.
* Proteins are obtained from UniProt and encoded into fixed-length integer sequences of 1000 residues. These sequences are embedded into a 128-dimensional vector, capturing important protein structure details.

**6.7.3 Training Process**

The model is trained on datasets like KIBA[26] and Davis[27], which provide drug-protein binding affinity values. Training uses batch processing to handle large datasets while maintaining computational efficiency. The model learns patterns in drug-protein interactions through multiple training epochs with optimization techniques.

**Table 9:** KIBA and Davis Datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **No of proteins** | **No of Compounds** | **No of interactions** |
| Davis | 442 | 68 | 30056 |
| KIBA | 229 | 2111 | 118254 |

**6.7.3.1 Role of KIBA and Davis Datasets in Binding Affinity Prediction**

In binding affinity prediction, datasets like **KIBA** and **Davis** are used to train and evaluate models like **GraphDTA**. These datasets provide real-world information about drug-protein interactions, helping machine learning models learn how to predict binding affinity effectively.

**Davis Dataset**

The binding affinities for 442 targets and 68 medicines are included in the Davis Dataset. These affinities are expressed as Kd constants and range from 5.0 to 10.8 [21].

**KIBA Dataset**

KIBA includes the binding affinities for 229 targets and 2111 drugs, which are expressed as KIBA scores that range from 0.0 to 17.2 [22].

**7. RESULTS**

Identified key biomarkers **CDK4(Cyclin-Dependent Kinase 4) and EZH2(Enhancer of Zeste Homolog 2)** indicating potential targets for Parkinson’s treatment.

**Table 10:** Binding Affinity Scores of CDK4 and EZH2 in Davis and KIBA Datasets

|  |  |  |
| --- | --- | --- |
| **Gene Name** | **Davis Dataset** | **KIBA Dataset** |
| CDK4 | 3.2321908473968506 | 13.583166122436523 |
| EZH2 | 6.288619518280029 | 11.44206714630127 |

The binding affinity scores for CDK4 and EZH2 from the Davis and KIBA datasets are displayed in this table. A higher drug-target protein binding is indicated by a lower score.

* The Davis dataset indicates a stronger interaction, as evidenced by CDK4's binding affinities of 3.23 (Davis) and 13.58 (KIBA).
* Ezh2 binds more effectively in the Davis dataset than in the KIBA dataset, with binding affinities of 6.29 (Davis) and 11.44 (KIBA).

When creating treatments for Parkinson's disease, these scores aid in understanding how well the medications interact with their target proteins.

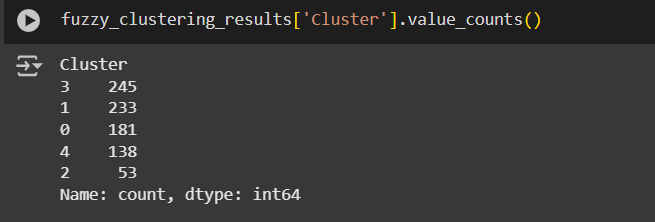
**Table 11:** Potential Drug Candidates for Treating Parkinson’s Disease: Targeting CDK4 and EZH2

|  |  |
| --- | --- |
| **Biomarker** | **IUPAC Name** |
| EZH2 - Enhancer of Zeste Homolog 2 | 4-(2-cyanophenoxy)-3-fluoro-N-(2,2,6,6-tetramethylpiperidin-4-yl)benzamide |
| N-[(4,6-dimethyl-2-oxo-1H-pyridin-3-yl)methyl]-6-[6-(4-methylpiperazin-1-yl)pyridin-3-yl]-1-propan-2-ylindazole-4-carboxamide |
| CDK4 - Cyclin-Dependent Kinase 4 | 3-(2,5-dichloro-N-[2-[4-[3-(dimethylamino)-2-hydroxypropoxy]anilino]pyrimidin-4-yl]anilino)propanenitrile |
| 1-[4-[[4-(2,5-dichloro-N-prop-2-ynylanilino)pyrimidin-2-yl]amino]phenoxy]-3-(dimethylamino)propan-2-ol |

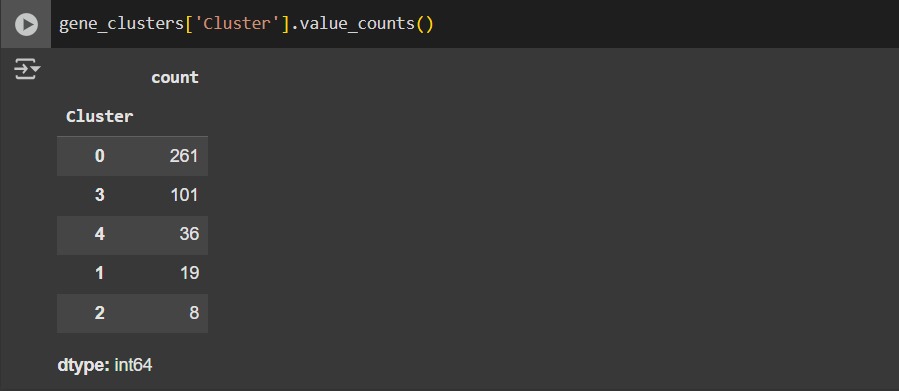
By focusing on the CDK4 and EZH2 proteins, this study identified four potential medications that may aid in the treatment of Parkinson's disease.

* **CDK4 Inhibitors:** These compounds may help protect brain cells by regulating cell growth and survival, potentially mitigating neurodegeneration in Parkinson’s disease.
* **EZH2 Inhibitors:** By influencing gene activity, these inhibitors could reduce nerve cell damage, which is a critical factor in the progression of Parkinson’s disease.
* Chemical compounds are uniquely identified by their molecular structure using the **International Union of Pure and Applied Chemistry (IUPAC)** Name, a standardized nomenclature system used worldwide.
* We can use PubChem[29], an online database that offers chemical information, including nomenclature, properties, and structures, to ascertain the IUPAC name from a SMILES representation.

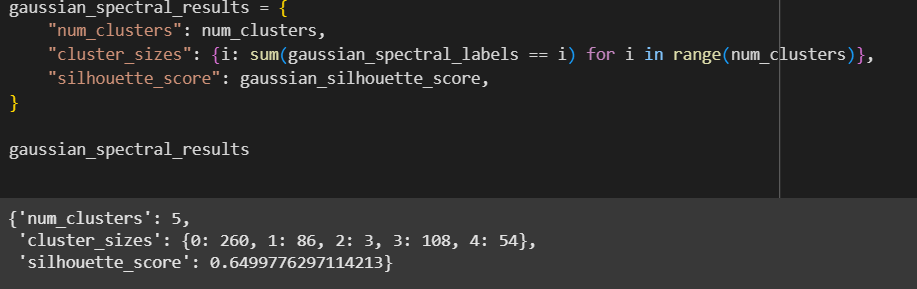
These drugs have special chemical structures that help them work better. They could be tested further to see if they can be used as real medicines in the future.

****

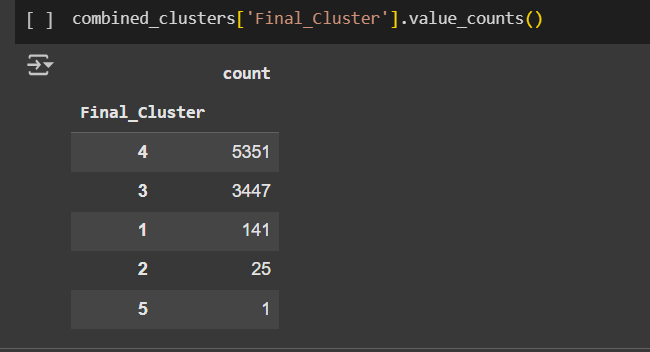
**Figure 5:** Fuzzy Clustering Algorithm results



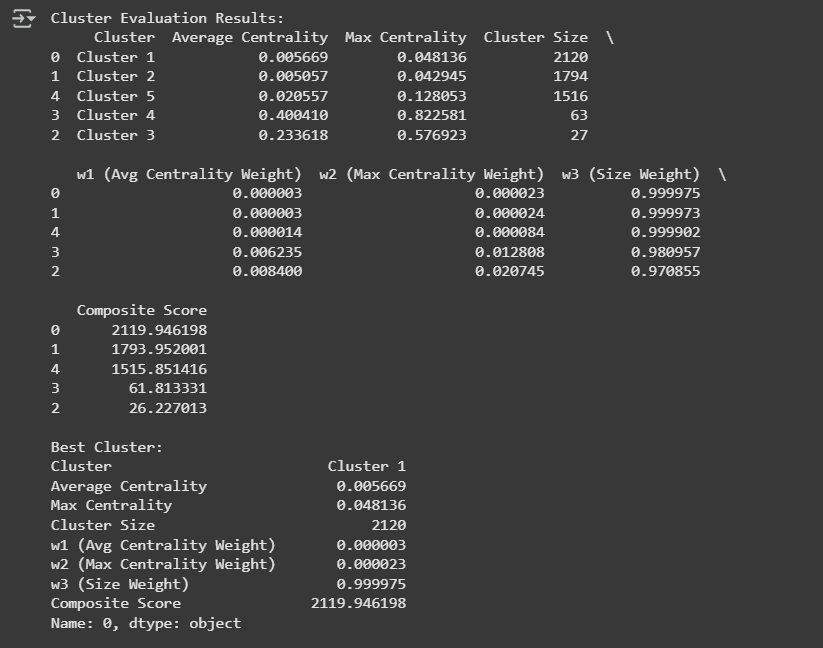
**Figure 6:** GAE Clustering Algorithm results

****

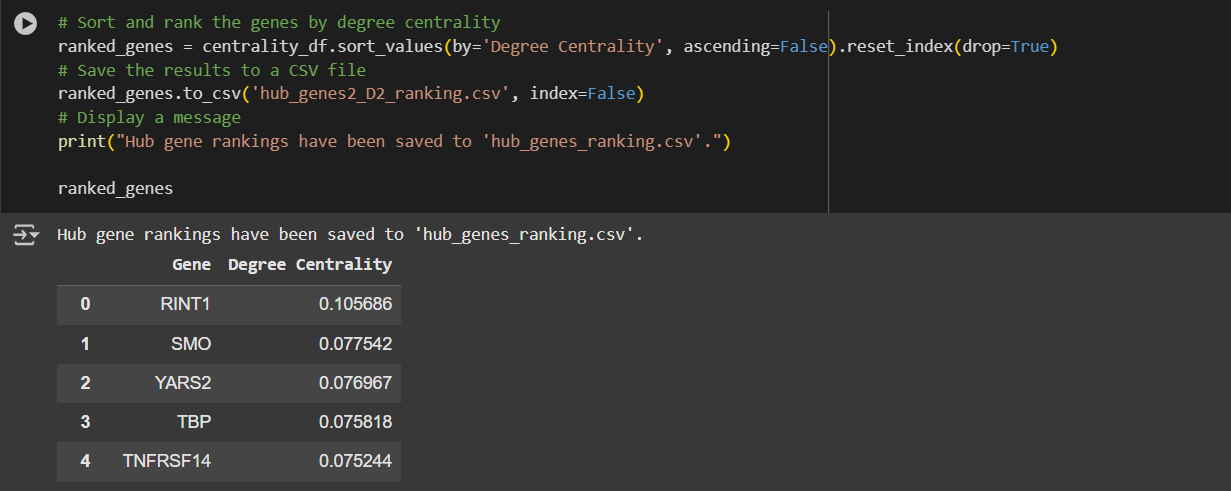
**Figure 7:** Spectral Clustering Algorithm results

****

**Figure 8:** EAC Algorithm results

****

**Figure 9:** Validation for best cluster

****

**Figure 10:** Sorting of Hub Genes

**8. LIMITATIONS**

Despite the success of the project, there are several limitations that need to be addressed:

* Limited Data Availability: There aren't as many high-quality labelled datasets available for drug discovery, particularly for Parkinson's disease.
* Selecting Key Features Is Difficult: It can be challenging to choose which drug properties and biomarkers to use to train the model. Making a poor decision can worsen forecasts.
* Selecting the Best Strategy Requires Time: Several strategies must be tried at every stage. We must try other approaches if one doesn't work well, which raises the possibility of success.
* Requires a Lot of Computer Power: Processing big datasets can take a while, and deep learning models like GraphDTA require powerful computers to run.

**9. CONCLUSION**

* ML in Drug Discovery: According to this study, machine learning aids in the quicker and more precise discovery of novel medications for Parkinson's disease.
* Saves Time and Money: When compared to conventional techniques, AI models speed up and lower the cost of drug discovery.
* GraphDTA's Function: By converting drug molecules into graphs, the GraphDTA model forecasts how well a drug will bind to a protein.
* Helpful Datasets: Training the model with datasets like Davis and KIBA improves predictions.
* Better Treatment Options: AI has the potential to improve patient care by enabling more accurate treatments for conditions like Parkinson's.

**9.1 FUTURE WORK**

* Using More Biological Data: Drug predictions can be enhanced by including data from genes, proteins, and metabolism.
* Making AI More Intelligible: By creating AI models that provide an explanation for their choices, researchers will be able to trust and enhance them.
* Investigating Quantum Computing: Better drug simulations can be achieved with the aid of emerging technologies such as quantum computing.
* Real-World Testing: Clinical trials and real-world experiments should be used to verify AI predictions.
* Developing New Drugs with AI: Drug discovery may become even more efficient if AI is used to create novel drug molecules.

**CO-PO-PSO OUTCOMES**

**Course Outcome:**

**CO1.** Perform literature search and / or patent search in the area of interest and formulate specific problem statements for ill-defined real-life problems with reasonable assumptions and constraints.

**CO2.** Conduct experiments / Design and Analysis / solution iterations and document the results.

**CO3.** Perform error analysis / benchmarking / costing.

**CO4.** Synthesis the results and arrive at scientific conclusions / products / solution.

**CO5.** Document the results in the form of technical report / presentation.

**Program Outcomes (POs) Attainment**

**PO1**. **Knowledge Application:** Utilize computational techniques and statistical analysis to study complex challenges in biomarker identification and drug-target interaction modelling.

**PO2.** **Analytical Thinking:** Assess and structure data-driven methodologies to enhance the prediction of drug efficacy, leveraging biomedical databases and computational modelling.

**PO3.** **Model Development:** Build and refine predictive models for clustering genes and evaluating drug-binding affinity, ensuring optimized decision-making in the drug discovery process.

**PO4.** **Research and Validation:** Conduct systematic investigations, feature extraction, dataset refinement, and statistical validation to establish accurate biomedical predictions.

**PO5.** **Technical Skills:** Apply data analysis tools and programming frameworks such as Python, TensorFlow, PyTorch, and RDKit for executing computational experiments efficiently.

**PO6.** **Societal Impact:** Examine the significance of technology-driven solutions in drug discovery, considering the economic, ethical, and healthcare benefits of AI-assisted research.

**PO7.** **Environmental and Ethical Responsibility:** Support the development of safe and sustainable medical solutions by promoting responsible data handling and unbiased model predictions.

**PO8.** **Integrity and Compliance:** Adhere to ethical guidelines in biomedical computing, ensuring the fair and unbiased selection of drug candidates in predictive analytics.

**PO9.** **Team Collaboration:** Work efficiently in interdisciplinary teams, coordinating with medical researchers, data scientists, and domain experts to achieve reliable and scalable outcomes.

**PO10.** **Communication Proficiency:** Deliver well-structured research findings through visual representations, scientific documentation, and oral presentations, tailored for different stakeholders.

**PO11.** **Project Execution:** Implement structured workflows in handling large-scale biomedical datasets, ensuring effective time and resource management in predictive research.

**PO12.** **Lifelong Learning:** Stay updated with emerging advancements in computational drug discovery, including biomedical informatics, deep learning models, and molecular simulations.

**Program Specific Outcomes (PSOs)**

**PSO1.** Implement gene expression clustering and predictive modelling techniques to enhance the understanding of molecular interactions in Parkinson’s disease.

**PSO2.** Develop computational models for drug discovery, integrating biological datasets with machine learning methodologies to improve therapeutic research.

**PSO3.** Acquire expertise in handling large-scale genetic and chemical datasets, ensuring compliance with biomedical research ethics and data security standards.

**CO-PO-PSO Mapping:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
| CO1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| CO2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| CO3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| CO4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| CO5 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

**PO’S ATTAINMENT:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Drug Discovery Using Machine Learning** | | | | | | | | | | | | | | |
| **Program Outcomes** | | | | | | | | | | | | **Program Specific Outcomes** | | |
| **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** | **PSO1** | **PSO2** | **PSO3** |
| ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ |

## ****REFERENCES****

1. Magare, A. C., & Patel, M. S. (2021). Biomarkers identification for Parkinson’s disease using machine learning. IEEE AIMV.
2. Wang, W., Lee, J., Harrou, F., & Sun, Y. (2020). Early detection of Parkinson’s disease using deep learning and machine learning. IEEE Access.
3. Li, T., & Le, W. (2020). Biomarkers for Parkinson's Disease: How Good Are They? Neuroscience Bulletin.
4. Hällqvist, J., Bartl, M., Dakna, M. et al. (2024). Plasma proteomics identify biomarkers predicting Parkinson’s disease up to 7 years before symptom onset. Nature Communications.
5. Saracchi, E., Fermi, S., & Brighina, L. (2014). Emerging candidate biomarkers for Parkinson’s disease: a review. Aging and disease, 5(1), 27.
6. T. dos Santos, M. C., Scheller, D., Schulte, C., Mesa, I. R., Colman, P., Bujac, S. R., … & Nogueira da costa, A. (2018). Evaluation of cerebrospinal fluid proteins as potential biomarkers for early stage Parkinson’s disease diagnosis. PLoS One, 13(11), e0206536
7. Lin, C. H., Chiu, S. I., Chen, T. F., Jang, J. S. R., & Chiu, M. J. (2020). Classifications of neurodegenerative disorders using a multiplex blood biomarkers-based machine learning model. International Journal of Molecular Sciences, 21(18), 6914.
8. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 18(6), 463-477.
9. Saba, L. (2015). Imaging in Neurodegenerative Disorders. Oxford University Press.
10. Lee, Stellina & Yates, Nathanael & Tye, Susannah. (2021). Inflammatory Mechanisms in Parkinson’s Disease: From Pathogenesis to Targeted Therapies. The Neuroscientist. 28. 107385842199226. 10.1177/1073858421992265.
11. Eke, C. S., Jammeh, E., Li, X., Carroll, C., Pearson, S., & Ifeachor, E. (2020). Early detection of Alzheimer's disease with blood plasma proteins using support vector machines. IEEE journal of biomedical and health informatics, 25(1), 218-226.
12. Su J, Peng J, Wang L, Xie H, Zhou Y, Chen H, Shi Y, Guo Y, Zheng Y, Guo Y, Dong Z, Zhang X and Liu H (2023) Identification of endoplasmic reticulum stress-related biomarkers of diabetes nephropathy based on bioinformatics and machine learning Front. Endocrinol. 14:1206154.doi:10.3389/fendo.2023.1206154
13. Hällqvist, J., Bartl, M., Dakna, M. et al. Plasma proteomics identify biomarkers predicting Parkinson’s disease up to 7 years before symptom onset. Nat Commun 15, 4759 (2024).
14. Rafieipour, H., Abdollah Zadeh, A., Moradan, A., & Salekshahrezaee, Z. (2020). Study of genes associated with Parkinson disease using feature selection. Journal of Bioengineering Research, 2(4), 1-11.
15. Vatansever, S., Schlessinger, A., Wacker, D., Kaniskan, H. Ü., Jin, J., Zhou, M. M., & Zhang, B. (2021). Artificial intelligence and machine learning‐aided drug discovery in central nervous system diseases: State‐of‐the‐arts and future directions. Medicinal research reviews, 41(3), 1427-1473.
16. Fred, Ana LN, and Anil K. Jain. "Combining multiple clusterings using evidence accumulation." IEEE transactions on pattern analysis and machine intelligence 27, no. 6 (2005): 835-850
17. Wang, Chun, et al. "Attributed graph clustering: A deep attentional embedding approach." arXiv preprint arXiv:1906.06532 (2019).
18. Yu, Kun, et al. "Determination of biomarkers from microarray data using graph neural network and spectral clustering." Scientific reports 11.1 (2021): 2382
19. Nguyen, Thin, et al. "GraphDTA: predicting drug–target binding affinity with graph neural networks." *Bioinformatics* 37.8 (2021): 1140-1147
20. Azuaje FJ. Selecting biologically informative genes in co-expression networks with a centrality score. Biology direct. 2014 Dec;9:1-23.
21. Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., ... & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, *29*(11), 1046-1051.
22. Tang, Jing, et al. "Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis." *Journal of chemical information and modeling* 54.3 (2014): 735-743.
23. <https://genemania.org/> - Genemania
24. <https://disgenet.com/> - Disgenet
25. <https://www.uniprot.org/uniprotkb> - Uniprot
26. <https://github.com/thinng/GraphDTA/tree/master/data/davis> - Davis Dataset
27. <https://github.com/thinng/GraphDTA/tree/master/data/kiba> - KIBA Dataset
28. <http://www.ncbi.nlm.nih.gov/geo> - Parkinson's disease: substantia nigra (HG-U133A)
29. <https://pubchem.ncbi.nlm.nih.gov/> - PubCheM