DRUG TARGET INTERACTION PREDICTION

This project is submitted to

St. Vincent Pallotti College of Engineering & Technology

(An Autonomous Institution Affiliated to Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur)

in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in

COMPUTER ENGINEERING

Submitted by

CHAITREYA SHRAWANKAR

DEEP KHAUT

OMKAR INGOLE

SHREEYASH GAIKI

Under the guidance of

DR. DIPAK W. WAJGI
Associate Professor

Academic Year 2024-25

ARISE & SHINE

Department of Computer Engineering

ST. VINCENT PALLOTTI

COLLEGE OF ENGINEERING AND TECHNOLOGY

Gavsi Manapur, Wardha Road, Nagpur -441108

DRUG TARGET INTERACTION PREDICTION CE 2024-25

Drug Target Interaction Prediction

This project is submitted to

St. Vincent Pallotti College of Engineering & Technology

(An Autonomous Institution Affiliated to Rashtrasant Tukadoji Maharaj Nagpur University,Nagpur)

In partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in

COMPUTER ENGINEERING

Submitted by

CHAITREYA SHRAWANKAR

DEEP KHAUT

OMKAR INGOLE

SHREEYASH GAIKI

Under the guidance of

DR. DIPAK W. WAJGI
Associate Professor

Academic Year 2024-25

Department of Computer Engineering

ST. VINCENT PALLOTTI
COLLEGE OF ENGINEERING AND TECHNOLOGY
Gavsi Manapur,Wardha Road, Nagpur -441108

CERTIFICATE

Certified that this project report "**DRUG TARGET INTERACTION PREDICTION**" is the bonafide work of "**CHAITREYA SHRAWANKAR, DEEP KHAUT, OMKAR INGOLE & SHREEYASH GAIKI**" who carried out the project work under my supervision in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in **Computer Engineering** of **St. Vincent Pallotti Collegeof Engineering & Technology,** *(An Autonomous Institute affiliated to Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur.*)

Dr. Dipak W. Wajgi	**Prof. Komal Gehani**	**Dr. Sunil M. Wanjari**
Associate Professor	**Assistant Professor**	**Associate Professor**
Project Guide	**Project Coordinator**	**Head of Department**
Computer Engineering	**Computer Engineering**	**Computer Engineering**

PRINCIPAL

ST. VINCENT PALLOTTI

COLLEGE OF ENGINEERING AND TECHNOLOGY

Gavsi Manapur, Wardha Road, Nagpur – 441108

(An Autonomous Institute affiliated to RTMNU, Nagpur)

ACKNOWLEDGEMENT

This project work is one of the major milestones in my journey of learning. We would like to sincerely thank **Dr. Dipak W. Wajgi, project guide**, for her/his guidance at every stage of the project and for her prompt and insightful input.

We would like to thank **Dr. Sunil M. Wanjari, Head, Department of Computer Engineering** and all our faculty members who reviewed our work and pointed out the shortcomings, their valuable insights and recommendations propelled our project forward, helping us overcome challenges and complexities along the way.

Also, we are very grateful to the Institute **Management** and **Principal** for their overwhelming support in providing us the facilities of the computer lab and other required infrastructure.

PROJECT MEMBERS:

CHAITREYA SHRAWANKAR
DEEP KHAUT
OMKAR INGOLE
SHREEYASH GAIKI

CONTENTS

ABSTRACT

The identification of drug-target interactions is a crucial step in drug discovery and development [2]. Experimental methods for identifying these interactions are often costly, time-consuming, and labour- intensive. Computational approaches, such as the one implemented in this project, provide a faster and more efficient alternative to experimental techniques. By predicting potential drug-target interactions, such models can support the early stages of drug discovery and reduce the reliance on extensive biological testing [2].

This project focuses on building a deep learning-based model to predict drug-target interaction (DTI) binding affinities. The model is designed using a Multilayer Perceptron (MLP) and uses 1D sequence data from proteins and drugs as input. Drugs are represented using their SMILES strings, which describe their molecular structures, while proteins are represented using their amino acid sequences. The goal is to predict the binding affinity between a given drug and target protein based on these inputs.

The MLP model developed in this project captures patterns in drug and protein sequence data and predicts their interaction strength. This project uses pre-processed numerical representations of drug and protein sequences to train the model. The aim is to contribute to the growing field of computational drug discovery by providing a tool to assist in identifying promising drug candidates.

LIST OF FIGURES

1. INTRODUCTION

The project aims to develop a machine learning model to predict drug-target binding affinity using their 1D representations. This involves utilizing SMILES strings for drugs and amino acid sequences for proteins as inputs to train a predictive model, enabling efficient and accurate identification of potential drug-target interactions.

1.1 DEFINITION

The prediction of drug-target interactions (DTIs) is a critical task in drug discovery, as it identifies how effectively a drug binds to a protein target, influencing therapeutic outcomes. Traditional methods for determining binding affinities rely on experimental techniques like high- throughput screening, which are time-consuming, expensive, and limited in scale. With the rapid growth of biological data, computational approaches have emerged as efficient alternatives. This project focuses on developing a machine learning-based framework to predict DTIs using simple 1D representations of drugs (SMILES) and proteins (amino acid sequences). By leveraging molecular fingerprints and sequence embeddings, the model aims to accelerate drug discovery, reduce research costs, and support the identification of promising drug candidates.

1.2 PURPOSE & OBJECTIVE

Drug discovery faces challenges such as the high cost, time-consuming nature of experimental methods, and limited scalability of traditional approaches for identifying drug-target pairs.

1. **Predict Binding Affinities**: Develop a machine learning model to predict the interaction strength between drugs and human protein targets using molecular fingerprints and protein sequence embeddings.
2. **Efficient Data Integration**: Collect and process diverse datasets, including drug properties, protein sequences, and binding affinity measurements, for effective model training.
3. **Accelerate Drug Discovery**: Enable rapid identification of potential drug candidates, reducing the dependency on experimental screening and minimizing research costs.
4. **Support Drug Repurposing**: Identify approved drugs with potential efficacy against emerging diseases or for alternative therapeutic uses, optimizing drug development pipelines.
5. **Scalable Framework**: Create a computational system that can be extended to larger datasets and integrated into future drug discovery workflows.

2. LITERATURE REVIEW

The prediction of drug-target binding affinity is a critical challenge in drug discovery, as it helps identify how strongly a drug interacts with a protein, which directly impacts its therapeutic effectiveness. Computational methods for drug-target interaction (DTI) predictionhave evolved significantly over the years, transitioning from traditional statistical approaches to modern deep learning-based models [2].

DeepDTA: Hakime Öztürk et al. (2018) proposed DeepDTA, a pioneering model in DTI prediction. It utilized convolutional neural networks (CNNs) to process the 1D sequence information of drugs (SMILES strings) and proteins (amino acid sequences) [1]. This approach moved away from reliance on 3D structural data, leveraging simpler representationsto predict continuous binding affinity values. DeepDTA was evaluated on the Davis and KIBA datasets, achieving Mean Squared Error (MSE) scores of 0.262 and 0.194, respectively[1]. Its success highlighted the potential of deep learning for DTI prediction, outperforming classical methods like KronRLS and SimBoost.

Models like MFDR used Sparse Auto-Encoders (SAEs) for feature abstraction, training support vector machines (SVMs) for improved accuracy. Similarly, **DL-CPI** utilized domain-specific binary vectors to represent proteins.

The Davis dataset comprises 68 drugs and 442 proteins, with 30,056 drug-target interaction pairs measured using dissociation constants (Kd). Other approaches, such as WideDTA and GraphDTA, have extended the principles introduced by DeepDTA. WideDTA integrates text-based protein representations with SMILES features, while GraphDTA employs graph convolutional networks for more detailed molecular modeling. These studies emphasize the importance of data representation and encoding in DTI prediction [3].

Popular databases supporting these predictions include DrugBank, PubChem BioAssays, and BindingDB, which provide extensive experimental DTI data. Curated negative datasets, like those from MATADOR, further refine model training by balancing interactions with non-interacting pairs.

In this project, we leverage the findings of DeepDTA but adapt the architecture to use a Multilayer Perceptron (MLP) model. This approach simplifies computational requirements while maintaining robust predictive performance. By integrating molecular fingerprints for drugs and sequence embeddings for proteins, our work aims to make DTI prediction more accessible and efficient. Drug3D-DTI incorporated 3D molecular spatial information, leveraging atom proximities for accurate binding affinity prediction.

Recent advancements in deep learning and feature abstraction have significantly improved DTI prediction models. Incorporating raw sequence data and advanced architectures like CNNs and SAEs has enabled greater accuracy and applicability in drug discovery tasks. Future work may integrate graph-based models or attention mechanisms to further enhance predictive performance and generalization.

	Proteins	Compounds	CI (std)	MSE
KronRLS (Pahikkala *et al.*, 2014)	S–W	Pubchem Sim	0.871 (0.0008)	0.379
SimBoost (He *et al.*, 2017)	S–W	Pubchem Sim	0.872 (0.002)	0.282
DeepDTA	S–W	Pubchem Sim	0.790 (0.009)	0.608
DeepDTA	CNN	Pubchem Sim	0.835 (0.005)	0.419
DeepDTA	S–W	CNN	0.886 (0.008)	0.420
DeepDTA	CNN	CNN	0.878 (0.004)	0.261

Fig 2.1 The average CI and MSE scores of test set trained on five different training sets for the Davis dataset

The table shows till date Model developed on Drug-Target interaction prediction along with their encoders/transformers and their respective MSE score.

3. PROJECT PLANNING & SCHEDULING

Figure 3.1 shows the planning and scheduling process for the project using a Gantt chart.

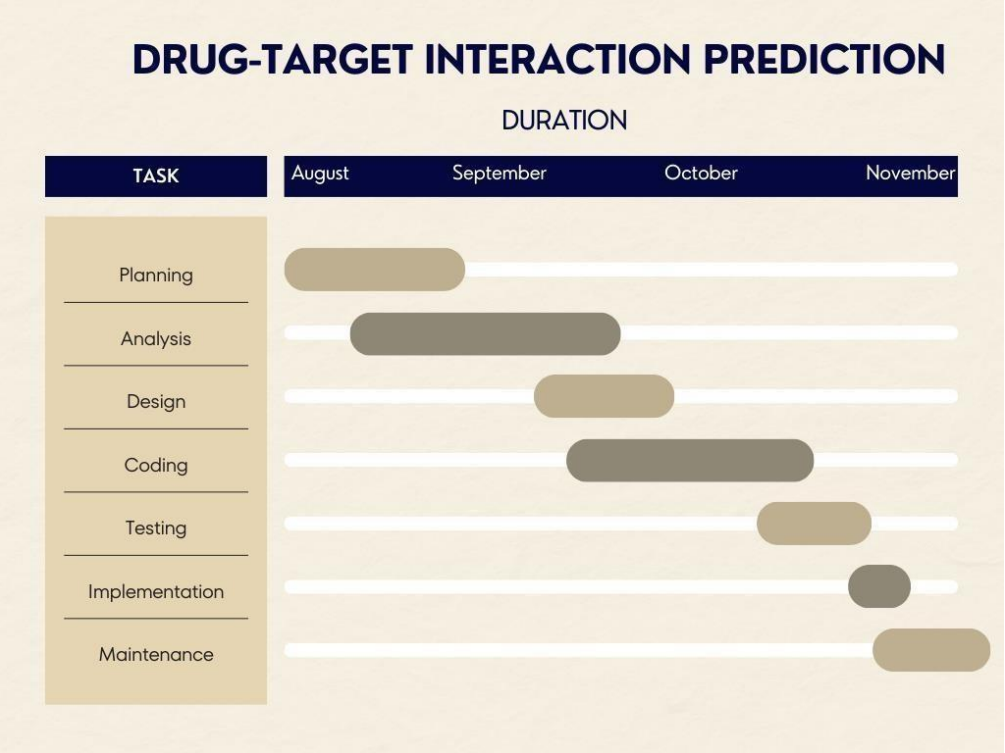

Fig. 3.1 Gantt Chart

Planning (August): This phase involves outlining the project's scope, identifying objectives, gathering requirements, and finalizing resources and technologies.

Analysis (August-September): Detailed exploration of drug-target datasets, selection of reliable data sources, and defining input/output specifications for the machine learning model.

Design (September): Creation of the architecture for the machine learning pipeline, including pre-processing modules, model selection (MLP), and integration flow for user-friendly predictions.

Coding (September-October): Implementation of the machine learning model using Python libraries, feature engineering, data cleaning, and optimization techniques.

Testing (October): Rigorous validation of the model's performance using metrics like RMSE, MAE, and correlation coefficients on training and test datasets.

Implementation (October-November): Deployment of the trained model in a controlled environment to simulate real-world applications and identify potential issues.

4. REQUIREMENT ANALYSIS

Requirement analysis is the first and most important phase of the software testing activities developing a project effectively. We have started to list all the functions that our system can provide for the user.

4.1 FUNCTIONAL REQUIREMENTS

1. **Input Processing:** Accept SMILES notation for drugs and amino acid sequences for proteins. Preprocess the data to remove duplicates and inconsistencies.
2. **Feature Transformation:** Generate molecular fingerprints for drugs and embeddings for proteins.
3. **Model Prediction:** Trained deep learning model will take the transformed input to predict its binding affinity value.
4. **Model Evaluation:** Evaluate model performance using metrics such as MSE, accuracy, or F1-score.

4.2 NON-FUNCTIONAL REQUIREMENTS

1. **Performance:** Ensure the model predictions are accurate and efficient for the large datasets.
2. **Scalability:** Support the integration of new drug and protein datasets.
3. **Reliability:** Maintain consistent performance under varying input conditions.

5. SYSTEM DESIGN

This chapter includes detailed descriptions of the data flow, module-level designs, and the overall structure of the implemented model. It explains the transformations applied to the input data, the functionality of each layer in the machine learning model, and the techniques used for optimizing performance. Additionally, the chapter provides diagrams and illustrations to visualize the design, ensuring a clear understanding of how the system processes inputs and generates outputs.

5.1 SYSTEM OVERVIEW

The implemented model is a **Multi-Layer Perceptron (MLP)** designed to predict drug- target binding affinities using 1D sequence representations of drugs and proteins. The data preprocessing involves encoding drug molecules through **SMILES (Simplified Molecular Input Line Entry System)** strings and protein sequences using embeddings, ensuring the model receives meaningful input features.

The training phase utilizes a curated dataset comprising drug-protein pairs with experimentally measured binding affinities, effectively split into training and validation setsto evaluate model performance.

This approach eliminates the reliance on computationally expensive 3D molecular structures, providing a faster and scalable alternative. Additionally, the model's ability to process simple sequence-based representations enables it to address challenges in drug discovery, including repurposing existing drugs for new therapeutic targets and accelerating the identification of potential drug candidates

5.2 SEQUENCE DIAGRAM

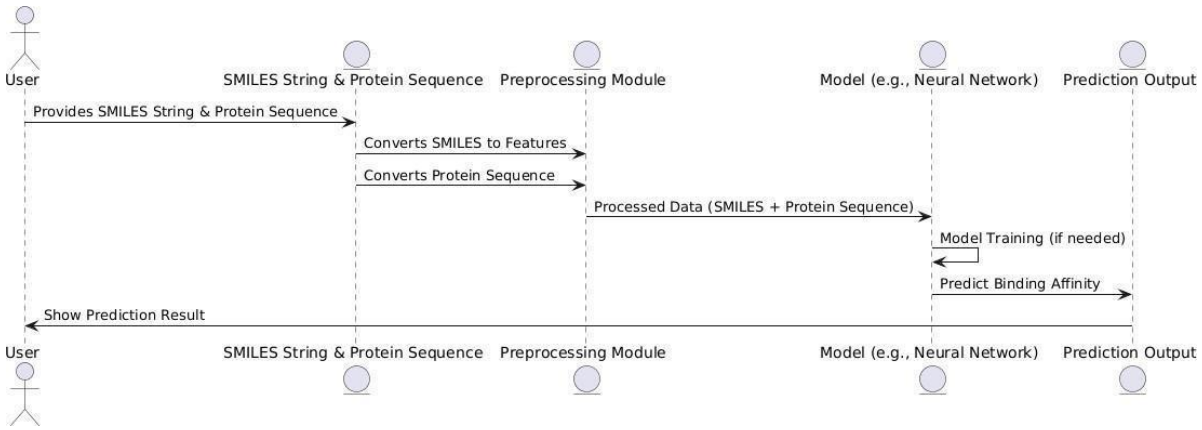

Fig 5.1 Sequence Diagram

5.3 DATA ARCHITECTURE

Fig 5.2 shows the detailed dataflow in the system.

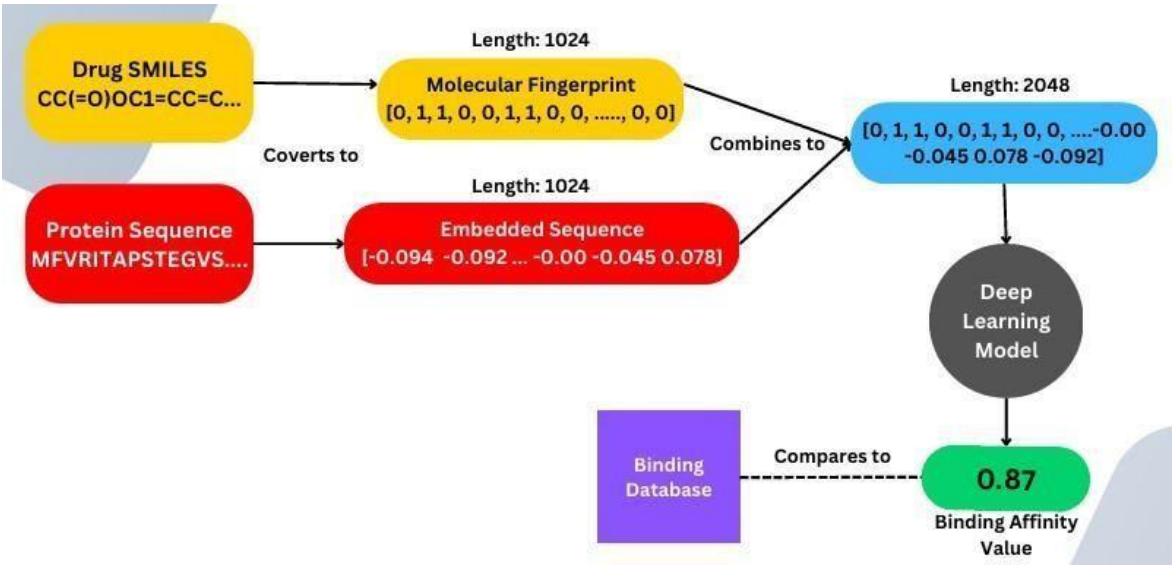

Fig 5.2 Dataflow Diagram

User Input

Drug is input as a SMILES (Simplified Molecular Input Line Entry System) string, which encodes the molecular structure of the drug. Protein is input as a sequence of amino acids that represents theprotein target.

Data Transformation

Drug SMILES string is processed using tokenization and embedding techniques to convert it into a numerical molecular fingerprint format suitable for the model. This might involve one-hot encoding, word embeddings, or pre-trained SMILES-based encoders. Protein amino acid sequence is similarly transformed using sequence embedding techniques, such as one-hot encoding, biochemical property-based encoding, or pre-trained sequence models.

Concatenation

The transformed drug and protein representations are concatenated into a single feature vector. This step ensures that the model receives a unified input that captures both drug and protein features.

Model Processing

The concatenated feature vector is passed through the Multi-Layer Perceptron (MLP)model, which consists of several fully connected layers. Each layer applies weights, biases, and activation functions to extract hierarchical patterns and correlations between the drug and protein features. Dropout and normalization techniques are applied during training to prevent overfitting and improve generalization.

Prediction

The final output layer of the MLP generates a continuous value, representing thebinding affinity between the drug and the protein target. This value indicates the likelihood or strength of the interaction between drug compound and the target protein.

5.4 MODEL ARCHITECTURE

The MLP architecture comprises the following layers:

Input Layer takes a concatenated vector combining the drug encoding (2048bits) and protein encoding.

Dense Layer 1 with 2048 neurons, ReLU activation, and L2 regularization. Includes batch normalization and dropout (0.3).

Dense layer 2 with 1024 neurons, ReLU activation, and L2 regularization. Includes batch normalization and dropout (0.3).

Dense layer 3 with 512 neurons, ReLU activation, and L2 regularization. Includes batch normalization and dropout (0.3).

Dense layer 4 with 256 neurons, ReLU activation, and L2 regularization. Includes batch normalization and dropout (0.2).

Dense layer 5 with 128 neurons, ReLU activation, and L2 regularization. Includes batch normalization and dropout (0.2).

Output Layer with single neuron with linear activation function for predicting the binding affinity score. Multiple dense layers allow the model to capture complex relationships between the drug and protein features. Batch normalization stabilizes training by normalizing inputs to each layer, reducing internal covariate shifts. Dropout is used to prevent overfitting by randomly disabling neurons during training. L2 regularization penalizes large weights to improve generalization.

Figure 5.3 shows the model architecture and training process.

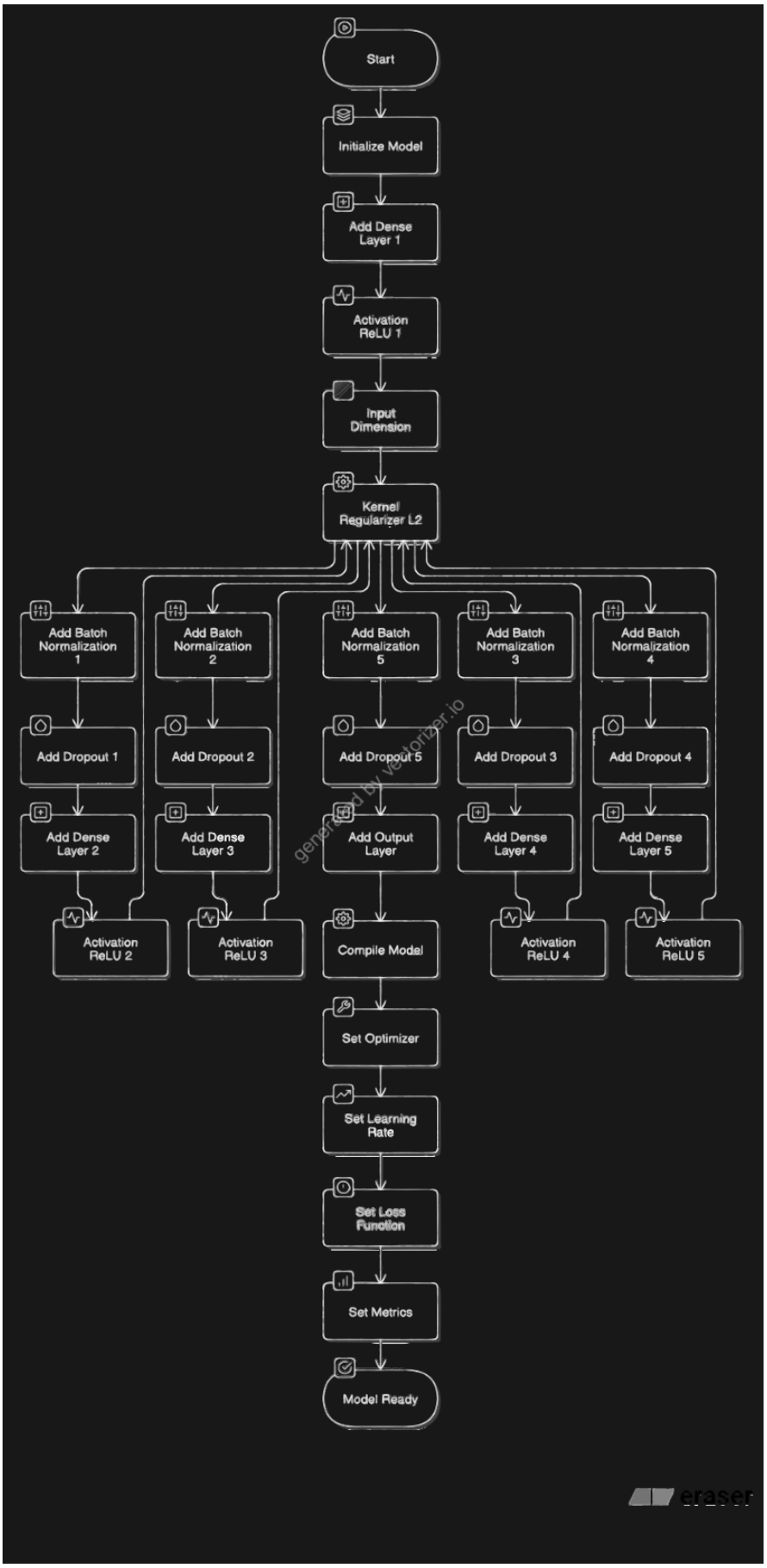

Fig 5.3 MLP Model Architecture

6. IMPLEMENTATION

This chapter delves into the practical aspects of building the Drug-Target Interaction (DTI) prediction system. This chapter covers the dataset used, including its structure, preprocessing steps, and transformation methods applied to prepare it for the machine learning model. Additionally, the chapter provides an analysis of the training process, highlighting key results and insights gained during model evaluation.

6.1 DATASET

The Davis dataset contains selectivity assays of the kinase protein family and the relevant inhibitorswith their respective dissociation constant (Kd) values. It comprises interactions of 442 proteins and68 ligands so in total it contains 30,056 interactions. The compound SMILES strings of the Davis dataset were extracted from the Pubchem compound database based on their Pubchem CIDs. For the compounds of the Davis dataset, the maximum length of a SMILES is 103, while the average length is equal to 64. Both SMILES and protein sequences have varying lengths. Hence, in order to create an effective representation form, we decided on fixed maximum lengths of 85 for SMILES and 1200 for protein sequences for Davis.

	SMILES	Target_seq	Binding_Affinity
2	CC1=CN=C(N=C1NC2=CC(=CC=C2)S(=O)(=O)NC(C)	MEGAAAPVAGDRPDLGLGAPGSPREAVAGATAALEPRKPHGVKRHHHKI	1.531478917
3	CN1CCN(CC1)C(=O)C2=CC3=C(N2)C=CC(=C3)Cl	TMPPRPSSGELWGIHLMPPRILVECLLPNGMIVTLECLREATLITIKHELFKE/	4
4	CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN	MEGDGVPWGSEPVSGPGPGGGGMIRELCRGFGRYRRYLGRLRQNLRET	4
5	CN1CCN(CC1)CCCOC2=C(C=C3C(=C2)N=CC(=C3N	MPALARDGGQLPLLVVFSAMIFGTITNQDLPVIKCVLINHKNNDSSVGKSS	3
6	CN1C=C(C=N1)C2=NN3C(=NN=C3SC4=CC5=C(C=C	MASSSVPPATVSAATAGPGPGFGFASKTKKKHFVQQKVKVFRAADPLVGV	3.51851394
7	C=CC(=O)NC1=C(C=C2C(=C1)C(=NC=N2)NC3=CC(=	MAGGPGPGEPAAPGAQHFLYEVPPWVMCRFYKVMDALEPADWCQFAA	2.653212514
8	CC1=C(C=C(C=C1)C(=O)NC2=CC(=CC(=C2)N3C=C(I	MSRSKRDNNFYSVEIGDSTFTVLKRYQNLKPIGSGAQGIVCAAYDAILERNV	2.653212514
9	CCN(CC)CCNC(=O)C1=C(NC(=C1C)C=C2C3=C(C=C	MATCIGEKIEDFKVGNLLGKGSFAGVYRAESIHTGLEVAIKMIDKKAMYKAG	2.278753601
10	CN1C2=C(C=C(C=C2)OC3=CC(=NC=C3)C4=NC=C(N	MRGARGAWDFLCVLLLLLRVQTGSSQPSVSPGEPSPPSIHPGKSDLIVRVC	2.929418926
11	CNC(=O)C1=NC=CC(=C1)OC2=CC=C(C=C2)NC(=O)I	MDLSMKKFAVRRFFSVYLRRKSRSKSSSLSRLEEEGVVKEIDISHHVKEGFEI	4
12	CNC(=O)C1=NC=CC(=C1)OC2=CC=C(C=C2)NC(=O)I	MSLLQSALDFLAGPGSLGGASGRDQSDFVGQTVELGELRLRVRRVLAEGC	4
13	CC1(CNC2=C1C=CC(=C2)NC(=O)C3=C(N=CC=C3)N	MAGRGSLVSWRAFHGCDSAEELPRVSPRFLRAWHPPPVSARMPTRRWA	4
14	C1=CC=C2C(=C1)C(=NN=C2NC3=CC=C(C=C3)Cl)C(	MDKYDVIKAIGQGAFGKAYLAKGKSDSKHCVIKEINEEKMPIQEKEASKKEVL	4

Fig 6.1 Davis Dataset

6.2 DATA TRANSFORMATION

It processes input data, such as SMILES strings and protein sequences, through feature transformation pipelines before feeding them into a Multi-layer Perceptron network.

Drug Compound

SMILES Conversion: SMILES strings (e.g., CCO, representing ethanol) encode themolecular structure as a string of characters. Each character or meaningful substructure (e.g., C, O, =, etc.) is separated. The tokens are converted into numerical molecular fingerprint vectors using one-hot encodingor pre-trained embeddings trained on large SMILES datasets.

Example: Input SMILE: COO

Tokenized: C, O, O

Encoded vector: [0.8, 0.2, 0.0, 0.0] for C and [0.1, 0.9, 0.0, 0.0] for O.

Protein Sequence

The amino acid sequence (e.g., MKT) is tokenized into individualresidues (M, K, T). Each amino acid is mapped to a numeric vector using biochemicalproperty encodings or pre-trained sequence models.

Example: Input Sequence: MKT

Tokenized: M, K, T

Encoded vector: [1.0, 0.5, 0.2] for M, [0.8, 0.3, 0.7] for K.

6.3 MODEL PREDICTION

The encoded molecular fingerprint vectors of the drug compound and the protein sequence embedding vector are concatenated into a unified vector, capturing the joint properties of both entities. The concatenated vector is passed through multiple fully connected layers. Each layer extracts higher-order patterns, identifying the likelihood ofinteraction between the drug and the protein. The final layer of the MLP outputs a binding affinity score (e.g., 7.5kcal/mol), which quantifies the strength of the interaction.

This process makes the model efficient in leveraging the 1D sequence representation of drugs and proteins, bypassing the computational complexity of 3D molecular modelling while achieving accurate predictions.

6.4 MODEL TRAINING

The training process for the Drug-Target Interaction (DTI) prediction model involved a supervisedlearning approach with the collected dataset of drug SMILES and protein sequence representations.Training began with an initialized Multi-Layer Perceptron (MLP) model, where the weights were updated iteratively based on the calculated loss function.

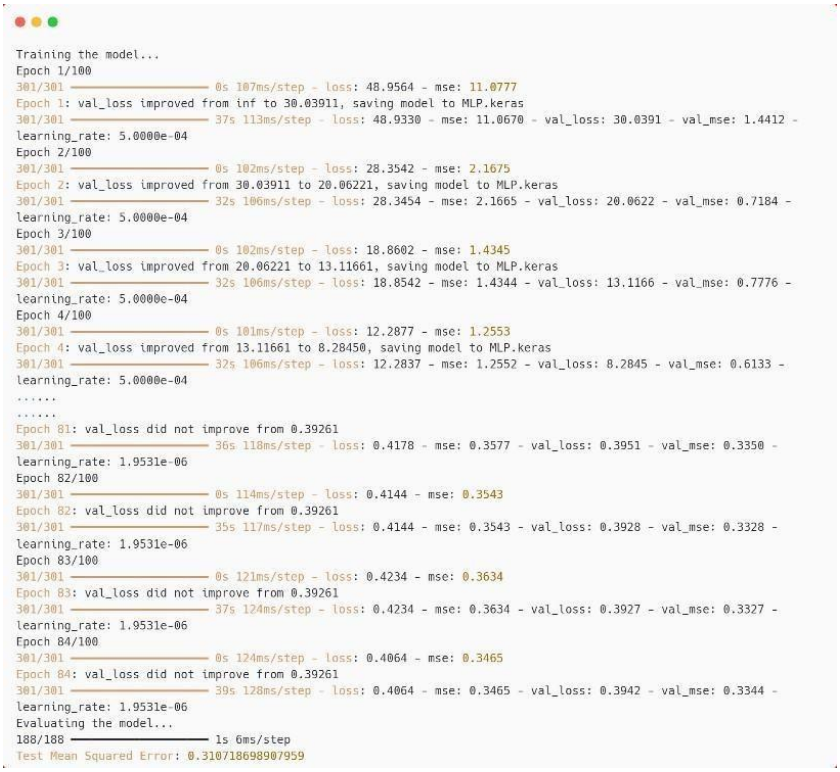

Fig 6.2 Training Epochs

Initially, during the early epochs the training began with higher loss and mean squared error (MSE)values as the model initialized its weights. Epoch 1 started with a training loss of 48.93 and an MSE of 11.07, while the validation loss was 30.03. The learning rate was set to 0.0005. This phase showedrapid progress as the model began learning meaningful patterns from the data.

In middle epochs, the model exhibited gradual improvements. While the training loss continued todecline, the rate of reduction in validation loss slowed as the model approached convergence. By Epoch 74, the validation loss hovered around 0.39, and further training yielded negligible improvements. The learning rate was reduced to as low as 1.9531e-06, leading to minimal weight adjustments. At Epoch 84, the training loss stabilized at 0.4064 with a final MSE of 0.3465, and thevalidation loss concluded at 0.3926. The final test evaluation reported an **Test Mean Squared Error:0.310718698907959**, indicating that the model successfully generalized to unseen data. The MLP model demonstrated a clear learning trajectory, with rapid initial improvements followed bygradual optimization and final convergence.

7. TESTING & IMPROVEMENT

This chapter evaluates the performance of the implemented model using various metrics, such as Mean Squared Error (MSE), and visualizes the results through graphs for loss, accuracy, and validation trends across epochs. Furthermore, this chapter identifies potential enhancements to improve model accuracy, such as modifications in architecture, hyperparameter tuning, and dataset expansion.

7.1 PERFORMANCE EVALUATION

The graph illustrates the trend of the **Mean Squared Error (MSE)** over training epochs for both the training and validation datasets. This pattern is representative of a well-converging model and is indicative of effective learning.

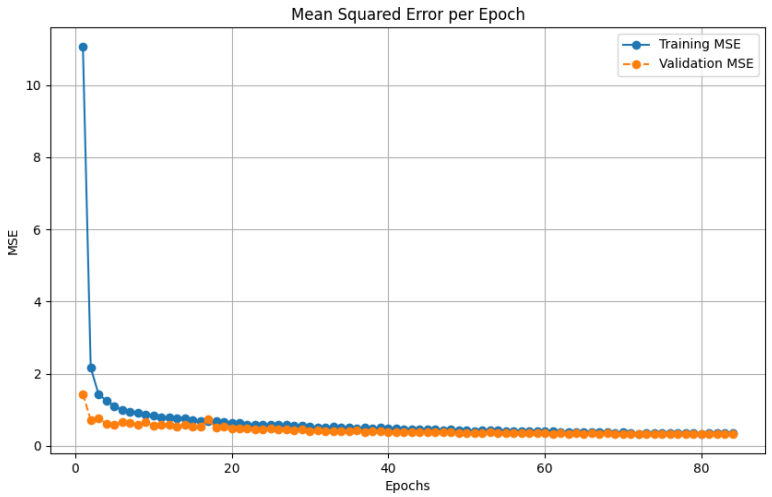

Fig 7.1 MSE Graph

In the first few epochs, the training MSE decreases sharply, reflecting the model's rapid adjustment to the data during the early stages of training. The validation MSE also drops quickly and begins to follow a similar pattern to the training MSE. This suggests that the model generalizes well to unseen data during initial epochs. After about 20 epochs, both training and validation MSE values plateau, indicating that the model's learning rate has stabilized, and further improvement is minimal. This reflects thatthe model has likely achieved its optimal fit.

The parallel trend between training and validation MSE suggests there is no significant overfitting, as both curves maintain similar values without divergence. Accuracy increases rapidly in early epochs, plateaus as learning stabilizes, and remainsconsistent between training and validation, indicating good generalization.

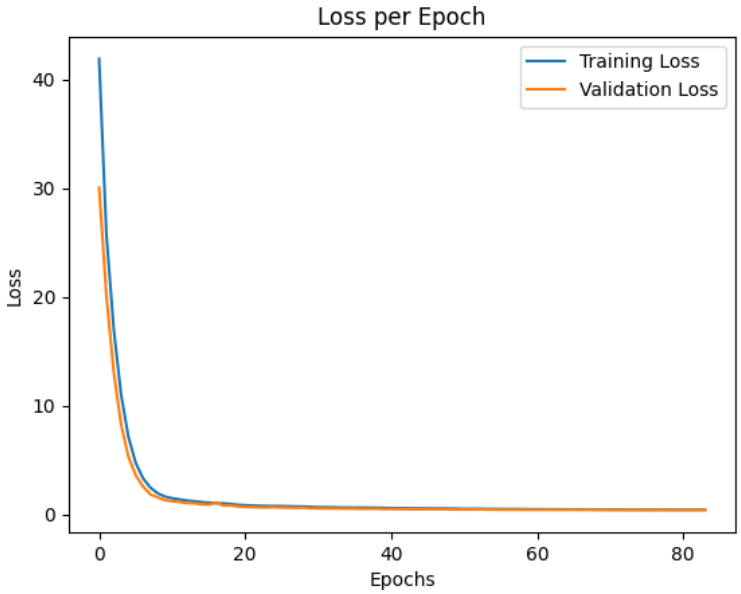

Fig 7.2 Loss per Epoch Graph

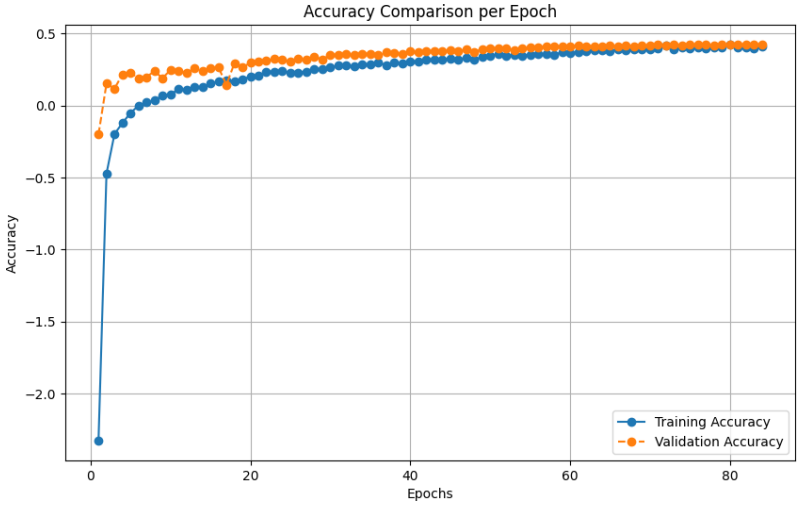

Fig 7.3 Accuracy per Epoch Graph

7.2 IMPROVEMENTS

Early Stopping

Based on the graph, the validation MSE stabilizes around epochs. Implement early stopping to halt training once the validation performance stops improving, avoiding overfitting and saving computational resources.

Learning Rate Adjustment

Reduce the learning rate dynamically (learning rate scheduler) after 10 epochs when the validation curve flattens, allowing the model to converge more effectively.

Regularization

Introduce dropout or L2 regularization to reduce the risk of overfitting observed in later epochs. These techniques can help maintain a balance between training and validation performance.

Advance Architecture

Experiment with hybrid architectures like combining MLP with RNN layers to capture sequential dependencies in protein sequences.

Collaborative Improvements

Test the model across multiple DTI datasets (e.g., KIBA, BindingDB) to ensure generalization.

8. CONCLUSION AND FUTURE SCOPE

This chapter summarizes the project, emphasizing its objectives, methodologies, and outcomes. The chapter also explores the potential applications of this project in healthcare and pharmaceutical research. It outlines future enhancements to improve the model's accuracy, scalability, and real-world applicability, ensuring its continued relevance and effectiveness.

8.1 CONCLUSION

This project successfully implemented a Multi-Layer Perceptron (MLP) model for predicting drug-target binding affinities using one-dimensional representations of drugs (SMILES strings) and proteins (amino acid sequences). y transforming these inputs into numerical vectors through feature extraction techniques, the model effectively captured complex biochemical relationships.

Designing and training a robust machine learning model with layered architectures and regularization techniques to achieve a balance between accuracy and generalization. Achieving a significant reduction in training and validation errors, with a final Mean Squared Error (MSE) of **0.3107**, indicating strong predictive performance. Demonstrating the practical utility of the model in drug repurposing, offering a computationally efficient alternative to traditional experimental methods.

The project not only provides a foundation for faster and cost-effective drug discovery but also underscores the potential of integrating machine learning into pharmaceutical research workflows. By leveraging computational tools, researchers can significantly accelerate the identification of promising drug candidates, especially in addressing emerging diseases or optimizing existing drugs.

8.2 FUTURE SCOPE

Data Enhancement

Incorporate larger and more diverse datasets, including rare drug-protein interactions, to improve model robustness.

Advanced Algorithms

Exploring deep learning models like Graph Neural Networks (GNNs) could further enhance the predictive accuracy for DTI, given their capacity to handle complex molecular structures.

Integration of Pretrained Models

Use pretrained embeddings (e.g. DeepPurpose CNN & RNN encoders) for better feature representation and learning efficiency.

Real-Time Prediction System

Developing a real-time web interface for the model can enable researchers and pharmaceutical companies to input new drug candidates and predict their interactions with target proteins quickly.

Expansion to Multi-Target Analysis

Extend the model to handle Poly pharmacology scenarios, predicting interactions between multiple drugs and multiple targets.

REFERENCES

1. Hakime Ozturk, Arzucan Ozgur and Elif Ozkirimli, “DeepDTA: deep drug–target binding affinity prediction,” Department of Computer Engineering and Department of Chemical Engineering, Bogazici University, Istanbul 34342, Turkey.
2. Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, Jimeng Sun, “DeepPurpose: a deep learning library for drug–target interaction prediction Bioinformatics, Volume 36, Issue 22-23, December 2020, Pages 5545–5547.
3. Tian,K. et al. (2015) Boosting compound–protein interaction prediction by deep learning. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA. IEEE, pp. 29–34.
4. Xing Chen, Chenggang Clarence Yan, Xiaotian Zhang, Xu Zhang, Feng Dai, Jian Yin, Yongdong Zhang, “Drug–target interaction prediction: databases, web servers and computational models,” Briefings in Bioinformatics, Volume 17, Issue 4, July 2016, Pages 696–712.
5. Tudor I. Oprea, Julie E. Bauman, Cristian G. Bologa, Tione Buranda, “Drug repurposing from an academic perspective.” Christopher A. Lipinski – Scientific Advisor, Melior Discovery, Waterford, CT 06385-4122, USA.
6. Davis dataset for drug-target interaction prediction Available at: https://github.com/dingyan20/Davis-Dataset-for-DTA-Prediction
7. Drug-target interaction prediction based on protein features, using wrapper feature selection by Hanegame Abbasi Mesrabadi, Karim Faez & Jamshid Pirgazi.
8. How to approach machine learning based prediction of drug/compound-target interactions by Heval Atas Guvenilir & Tunca Dogan.
9. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper by Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, Kayvan Najarian.

PROJECT TEAM MEMBER'S INFORMATION

Name: Chaitreya Shrawankar

Contact Number: 7755963369

Email ID: chaitreyashrawankar70@gmail.com

Name: Deep Khaut

Contact Number: 8857003108

Email ID: deepkhaut1234@gmail.com

Name: Omkar Ingole

Contact Number: 9356829277

Email ID: omkar.i1708@gmail.com

Name: Shreeyash Gaiki

Contact Number: 7028536832

Email ID: shreeyashgaiki1339@gmail.com

PROJECT GUIDE INFORMATION

Name: Dr. Dipak W. Wajgi

Academic Designation: Associate Professor

Contact Number: 7030600096

Email ID: dipak.wajgi@gmail.com