Cairo University
Faculty of Computers and Artificial Intelligence

Protein-Protein Interactions Binding Sites Prediction using Deep Learning

Supervised by

Prof. Hesham Hassan
Dr. Ahmed Farouk
TA. Sarah Hassan

Implemented by

20178056	*Mohamed Khaled Hamed*
20178036	*Mohamed Ali Mohamed*
20178003	*Ahmed Hatem*

Graduation Project Academic Year 2020-2021
Documentation

Abstract

Protein-protein interaction (PPI) sites play a key role in the formation of protein complexes, which is the basis of a variety of biological processes. BCL-2 protein is known to be associated with cancer diseases. It is quite often that their precise functional role in disease remains unclear. A strategy to gain a better understanding of the function of BCL-2 protein to make use of a combination of different aspects of proteins data type prediction protein-protein interaction as therapeutic targets for anticancer drug discovery Bcl-2 is acting as apoptosis suppressor gene and the protein in cancer cells may block or delay the onset of apoptosis, by selecting and maintaining long-living cells So if we know the binding site of protein we will able to know the hotspots that can drug attack to the protein on it to regulate or increases protein secretion to avoid cancer spreading to other healthy cells as experimental methods to solve PPI sites are expensive and time-consuming, which has led to the development of different kinds of prediction algorithms. We propose a convolutional neural network for prediction binding site of PPI and use residue binding propensity to improve the positive samples.

Proposed models obtains a remarkable result of the area under the curve (AUC) = 0.89 using CNN. And accuracy 90% using RFC and 84% using Hybrid Model (Feature extraction using 1DCNN-Transform)

In addition, it yields much better results on samples with high binding propensity than on randomly selected samples. This suggests that there are considerable false-positive PPI sites in the positive samples defined by the distance between residue atoms.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

PPI: Protein-Protein interaction

ML: Machine Learning

DL: Deep learning

CNN: Convolutional Neural Network

SVM: Support Vector Machine

SVC: Support Vector Classifier

KNN: K-Nearest Neighbors

LOO-XVE: leave-one-out cross validation error

GAN: Generative adversarial networks

BA-HPC: High-Performance Computing (HPC) cluster

Chapter one: Introduction

1.1 Introduction

Currently, almost everything is computerized, consequently, programs are needed even for the biology field of study. In other words, this project faces topics of both the biology and computer science fields.

It is rare that a set of measurement and analytic techniques can revolutionize biomedical research and clinical practice. It is precise because the excitement and the expectations surrounding this field are so high that we are compelled to do this project [2].

Protein-protein interactions are critical to nearly all aspects of cellular function, such as regulation of metabolic and signaling pathways, immunological recognition, DNA replication and gene translation, as well as protein synthesis. Identifying the binding sites between two interacting proteins provides important clues to the function of a protein and the structural elucidation of protein complexes, thus helps to identify pharmacological targets and guides drug design. Hence, solving the puzzle of predicting the interaction sites is of great significance to molecular recognition [4].

Protein-protein interactions has a huge impact in all biological processes Where most of proteins need to bind to another protein to fully perform their functions interactions can also cause the inhibition of the functions of the protein to fully understand the functions of these proteins the study of protein-protein interactions is needed but studying them is both time and money Consuming if they are studied experimentally so these interactions are computed to save much time [7].

1.2. Motivation

Currently, almost everything is computerized Consequently, programs are needed even for the biology field of study. In other words, this project faces topics of both the biology and computer science fields.

It is rare that a set of measurement and analytic techniques can revolutionize biomedical research and clinical practice. It is precisely because the excitement and the expectations surrounding this field are so high that we are compelled to do this project.

Here, we provide a source of challenge, problem, and dataset that will simulate basic development while furthering important goals in biological discovery. Therefore, this project will use underlying computer science technique (machine learning).

So, this brings an exciting field of study for Machine Learning researchers. In addition to this, noise and variability of the data make this domain more exciting.

This project faces what should be done when having a huge amount of data. How to choose features in data that will giveyou as good or better accuracy whilst requiring less data. Lastly, this will provide users with a way to identify the most effective genes in causing many diseases.

1.3 Beneficiary

Who is this intended for? Answering this question has served as our aim in this project. There are three audiences in that we havehad in mind.

1. Experienced Pharmacist with limited experience using Binding sites Generated.
2. Experienced Compute Scientist with limited experience analyzing Binding sites data.
3. Students entering the field of Bioinformatics.

How is this beneficiary for them?

•Saves time and effort for doctors and scientist to understanding functionality of genes that may cause cancer or not.

• Future data will be less, but more informative.

• Doctors can aid patients before disease progresses by knowing their gene history.

1.4 Problem Definition

BCL-2 is believed to be an apoptosis suppressor gene meaning that the cell Can kill itself and the BCL-2 can suppress that gene action. Whether there is correlation between the predominance and the gene expression of BCL-2 in cases of primary cancer or not is controversial. Some studies showed that there is no correlation, Other studies proved that there is significant correlation, these studies also found that there is high level of BCL-2 and absence of apoptotic cells in areas of cancer, so our test data is BCL-2 Many of the existing studies focus on the identification of protein-protein interaction sites with specific physicochemical and geometric characteristics.[6] Binding sites have been widely observed to be more hydrophobic, planar, globular, and protruding than outer surfaces. Different amino acid compositions have also been found among the interaction sites of homo-permanent complexes, homo-transient complexes, hetero-permanent complexes, and hetero-transient complexes. Interfaces have a significant number of polar residues, where usually the interactions are less permanent. Through alanine-scanning mutagenesis, it has been observed that the binding free energy is not distributed equally across these protein interfaces. Residues of interface, protein core and non-interface surface are found significantly different in sequence entropy and secondary structure [13].

The main application for protein-protein interactions is drug discovery Where protein complexes (proteins interacted together forms a complex) may be of harmful nature or oppress a activity and a small molecule is offered to the complex to disrupt it.

So, what exactly is the problem?

To summarize:

1. A large microarray with a lot of data to observe
2. How to find the useful aspects of the data used.
3. How to reduce the data, to only use the useful information it provides.
4. What to do if the data is or is not classified.
5. How accurate is our classification of this data.

1.5 Problem Objective

Interactions between proteins play a crucial part in cellular function and form the backbone of almost all biochemical processes. While many interacting protein pairs have been identified through large-scale experiments on whole genomes, the residues involved in these interactions are generally not known and the vast majority of the interactions remain to be characterized structurally [15]. Identifying key players and their interactions is fundamental for understanding biochemical mechanisms at the molecular level. Information about residues that form the interacting surface of a protein are useful for a wide range of applications such as the design of mutants for experimental verification of the interactions, the development of drugs that target protein–protein interactions, understanding the mechanism of the molecular recognition and as an aid to predicting complexes through docking and homology modelling[18]. The ever-increasing number of alternative ways to detect protein-protein interactions [16]. So, the experimental determination of protein–protein complexes are an expensive and time-consuming process. One alternative to prediction by comparative modelling is protein–protein docking but this method are hampered by a lack of a complete understanding of the forces involved and by the conformational changes that often take place upon protein–protein binding.[15] As the number of proteins with known atomic resolution has grown more groups have addressed the issue of extracting basic features of interacting protein complexes such as shape complementarity, chemical complementarity and combinations of the two Matching chemistry and shape [17].

So, what exactly are ours goals?

To summarize:

1. To obtain an active site (hotspots) for protein-protein interaction that can be classified as binding site or not.

2. Reduce the dimensionality

3. To optimize the classification result

1.6 The used tools in the project

BA-HPC - Bibliotheca Alexandrina as Bibliotheca Alexandrina (BA) offers researchers in Egypt merit-based access to a High-Performance Computing (HPC) cluster

This project does need High powerful GPU and accept as research from **BA-HPC**.

1.PSAIA – Protein Structure and Interaction Analyzer: PSAIA – Protein Structure and Interaction Analyzer [19].

2.Tensorflow

3.Keras

4. Rotation Forest Classifier

There are many different types of IDEs that could implement this project. Such as:

- Python
- C / C++
- Java
- R

The best IDEs for this project would be python or R

because they include libraries that can benefit us in this project

Therefore, it was concluded that python IDE would best suit the implementation for this project.

1.7 Project development methodology

Primary methods will include the following:

• Search for Datasets that is related to our project.

To search common online sites such as DBD, Prank Web, Site Out …etc. for datasets related to protein-protein interaction. And as mentioned to the used dataset for this research is a supervised Prediction Binding site.

•Enter the name of protein-protein interaction network in the program

After downloading the related data. We enter the data to our preferred IDE to start our research.

•Pre-process the microarray data

To eliminate many noisy and redundant data represent as (ex: disorder proteins). Selects k best methods generally used as a preprocessing step.

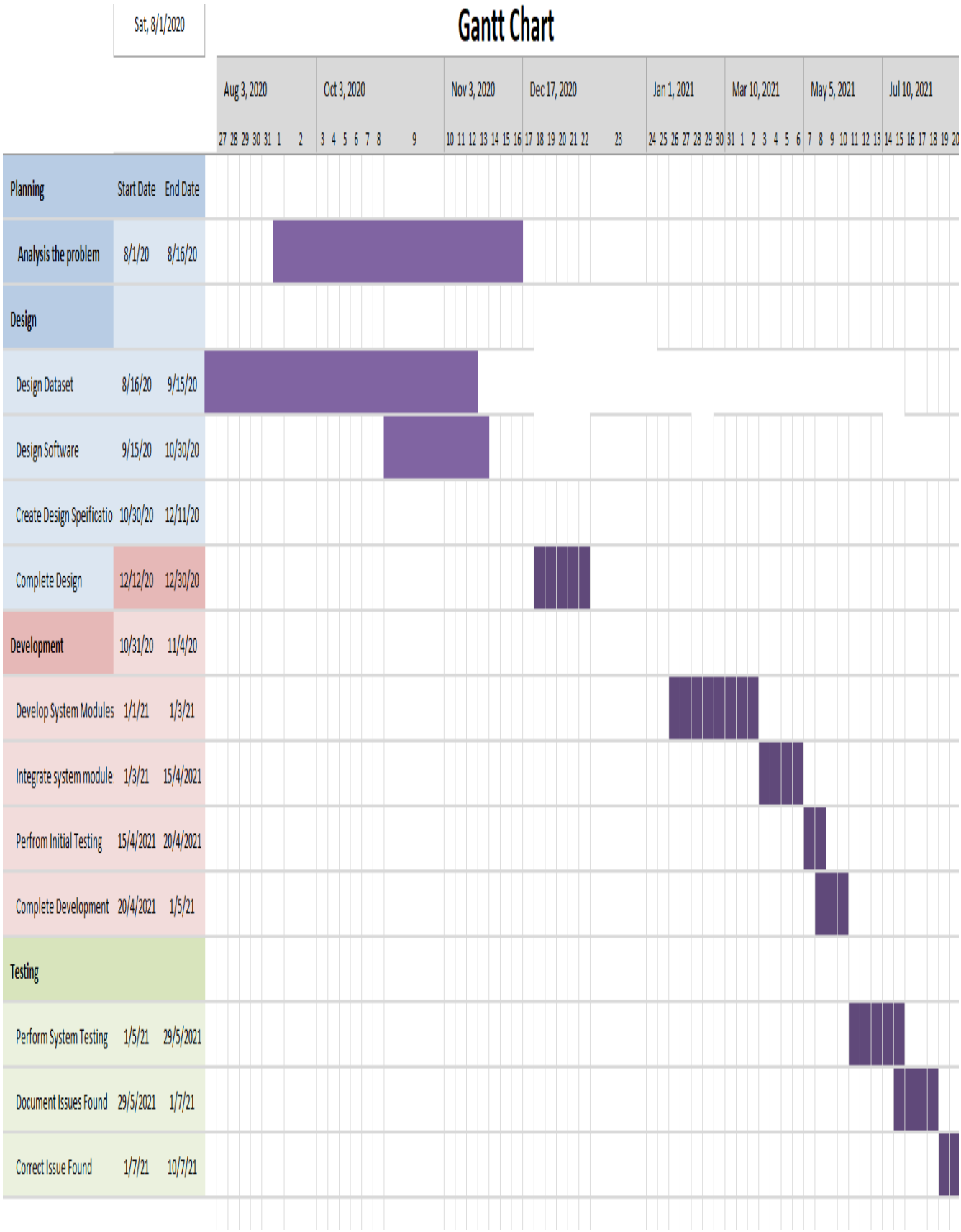
•The CNN program runs

A CNN uses a system much like a multilayer perceptron that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers, and normalization layers

•Evaluate accuracy of program through testing phase

To take our output data from the CNN program and test its accuracy by running the output in a machine learning technique called K-fold Validation

1.8 Gantt chart



Figure(1):Grantt Chart

Task	Task Title	Description	Task status (completed/expected in time)
Planning	BCl-2 Dataset	Collection a data set that suits our project according to Benchmark DBD	Completed
Planning	Convolutional Neural Network (**CNN**)	An algorithm that will determine the binding sites between receptor and ligand proteins	Completed
Planning	Programmable tools to docking	Use docking tools to verify the output from CNN	Complete
Implementation	Dataset preprocessing	Fixing and cleaning dataset tobe processed in program	Completed
Implementation	Determine the active site	Determine active site that in which predictions of whether a pair of residues from two different proteins interact or not.	Completed
Implementation	Feature Extraction	Sequences Feature and Structure	Completed
Implementation	CNN	1D and 2D CNN and hybrid model	Completed
Implementation	Feature-Selective Rotation Forest	effectively reduce the data dimension and remove the noise information in the data, thus improving the prediction accuracy and speed of the classifier	Completed
Testing	Integration	Putting all implemented parts Together	Completed
Testing	Testing and verification	Testing the entire program and verifying the final output	Completed

Table 1.1: Task completion, Project timeline

1.9 Report Organization

Chapter Two: Related Work

In chapter two, we will establish other work associated with our research. There, we show that different ways to achieve our goals.
We will declare the authors of these methods, the dates theywere founded and their explanation.

Chapter Three: System Analysis

In chapter three, we will clarify project specification. Where a good project specification is a simple but complete description of a software's functionality and purpose. It contains descriptionsof how the software will be used from a user perspective and performance details such as usability, reliability and stability. In addition, illustrate a use case diagram that emphases our program.

Chapter Four: System Design

In chapter four, we will portray our research diagrams. The purpose of a component diagram is to show the relationship between different components in a system. The term "component" refers to a module of classes that represent independent systems or subsystems with the ability to interface with the rest of the system. Also, a sequence diagram is a type ofinteraction diagram because it describes how—and in what order—a group of objects works together.

Chapter Five: Model proposed

In chapter Five, we will demonstrate step by step our implementation methods. We will give an explanation beginning from how our data looks like, how it is represented and constructed. Also, how it is preprocessed. In addition, we will demonstrate our program functions, how it affects our data and such. Moreover, we will explain the machine learning and Deep learning technique that was used in the model.

Chapter Six: Model Setup and Results

In chapter Six, we will show our parameter settings and describeit. We will also show our model results accordingly. And lastly, give a conclusion to our research.

Chapter Two: Related Work

The following are some research examples that are most related to our project objective:

1) Dechang Pi and Chishe Wang, The Prediction of Protein-Protein Interaction Sites Based on RBF Classifier Improved by SMOTE (hindawi.com)
2) Tzu-Hao Kuo and Kuo-Bin Li, China, 2018
3) Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest Wang, L., Wang, HF., Liu, SR 2019
4) Prediction of Protein–Protein Interaction Sites Using Convolutional Neural Network and Improved Data Sets (Zengyan Xie * , Xiaoya Deng and Kunxian Shu 2021)

Related work	technique	Accuracy	Data
1) Dechang Pi and Chishe Wang	RBF	-	Structural based features
2) Tzu-Hao Kuo and Kuo-Bin Li, China, 2018)	Mchine Learning	0.583	Sequence based features
3)Wang, L., Wang, HF., Liu, SR 2019	CNN +FSRF	0.97	Sequence based
4) Zengyan Xie * , Xiaoya Deng and Kunxian Shu 2021	2D CNN	0.91	Integrated Data
Proposed Model	1D CNN and Hybrid	0.90	Integrated Data

Table (2) Related works

- **The Prediction of Protein-Protein Interaction Sites Based on RBF Classifier Improved by SMOTE**

  In this approach, (Dechang Pi and ChisheWang ,China , 2014) propose similarity is that they have imbalanced dataset just like ours meaning that the number of negative samples is very high which affect the accuracy negatively oversampling to maximize the positive dataset or under sampling to minimize the negative class however this cause data loss therefore they used a method called SMOTE algorithm to balance the dataset.

- **The Main Difference between Propsed approach and** (Dechang Pi and ChisheWang China, 2014).

  The difference is in the model they have used RBF model to classify the binding sites RBF is a feed forward neural network has input layer equals to number of samples (input) and hidden layer that is nonlinear and a linear output layer we are going to use CNN which is a special type of neural network and was proven to have high efficiency in protein-protein interactions projects.

- **Predicting Protein–Protein Interaction Sites Using Sequence Descriptors and Site Propensity of Neighboring Amino Acids**

  In this approach, (Tzu-Hao Kuo and Kuo-Bin Li, China , 2018) propose alternative to similar approaches requiring structural information, the proposed method takes all of the input from protein sequences. In addition to typical sequence features, our method takes into consideration that interaction sites are not randomly distributed over the protein sequence. We characterized this positional preference using protein complexes with known structures, proposed a numerical index to estimate the propensity and then incorporated the index into a learning system.

- **The Main Difference between proposed approach and**, (Tzu-Hao Kuo and Kuo-Bin Li, China, 2018)

  This approach only used sequence-based features coming from a sequence input this approaches a non-acceptable accuracy AUC (0.675) but with usage of structural features the AUC of model we are working on is more likely to be more the result may provide insight into areas, such as mutant design and the investigation of protein interaction networks. In view of the difficulties of obtaining 3D structures for protein complexes, we adopted a sequence-based approach, in which features for the learning process are exclusively derived from protein sequences. In addition to classical sequence features, such as amino acid conservation and physicochemical properties.

- **Predicting binding sites from unbound versus bound protein structures (**Jordan J. Clark, Zachary J. Orban & HeatherA. Carlson**) [20]**

  This approach based on comparison study for prediction binding sites in two conformational:

  a) bound conformation (complex state)

  b) unbound conformation (single state)

  As the number of available protein structures increases, structural alignment-based algorithm becomes the dominant approach for protein-binding sites prediction. However, the present algorithms underutilize the ever-increasing numbers of three-dimensional protein–ligand complex structures (bound protein), and it could be improved on the process of alignment, selection of templates and clustering of template. Herein, we built so far, the largest database of bound templates with stringent quality control.

- **The Main Difference between proposed approach and**, (Jordan J. Clark, Zachary J. Orban & HeatherA. Carlson)

  This approach used PDB in unbound conformation to Determine active sites and compute geometric parameters for large sets of protein structures and perform the same operation on bound conformation. But our approach prefix_l_u.pdb File for the ligand in unbound conformation.

  Features are computed from this file

prefix_r_u.pdb File for the receptor in unbound conformation.

Features are computed from this file

prefix_l_b.pdb File for the ligand in bound conformation. Residue contacts are computed from this file

prefix_r_b.pdb File for the receptor in bound conformation. Residue contacts are computed from this file

Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest (Wang, L., Wang, HF., Liu, SR 2019**)**

we propose a novel approach, namely CNN-FSRF, for predicting PPIs based on protein sequence by combining deep learning Convolution Neural Network (CNN) with Feature-Selective Rotation Forest (FSRF). The proposed method firstly converts the protein sequence into the Position-Specific Scoring Matrix (PSSM) containing biological evolution information, then uses CNN to objectively and efficiently extracts the deeply hidden features of the protein, and finally removes the redundant noise information by FSRF and gives the accurate prediction results. When performed on the PPIs datasets Yeast and Helicobacter pylori, CNN-FSRF achieved a prediction accuracy of 97.75% and 88.96%. [22]

- **The Main Difference between proposed approach and**, (Wang, L., Wang, HF., Liu, SR 2019)

  This approach only used sequence-based features coming from a sequence input this approach depend only on Position-Specific Scoring Matrix (PSSM) as feature to prediction Binding site. but **proposed approach** used sequence-based features coming from a sequence input and Structure input represent in secondary structure of protein and three-dimensional protein structure. beside that our approach depends on sequence features, such as amino acid conservation, PSSM, PSFM and physicochemical properties.

  Structure features: Residue attributes for analysis - Accessible Surface Area, Relative ASA, Depth Index, Protrusion Index, and Hydrophobicity.

Prediction of Protein–Protein Interaction Sites Using Convolutional Neural Network and Improved Data Sets(Zengyan Xie * , Xiaoya Deng and Kunxian Shu 2021**)**
This approach uses convolutional neural network for PPI site prediction and use residue binding propensity to improve the positive samples. This method obtains a remarkable result of the area under the curve (AUC) = 0.912 on the improved data set. In addition, it yields much better results on samples with high binding propensity than on randomly selected samples. This suggests that there are considerable false-positive PPI sites in the positive samples defined by the distance between residue atoms.

- **The Main Difference between proposed approach and**, (Zengyan Xie * , Xiaoya Deng and Kunxian Shu 2021)

  This approach uses only 2D CNN, but our approach used 1D,2D and Hybrid model of CNN. this approach results of the area under the curve (AUC) = 0.912 on the improved data set. But our approach results of the area under the curve (AUC) = 0.9541 on the normal data set. This approach does not detect intrinsically disordered protein (IDP) that a protein that lacks a fixed or ordered three-dimensional structure. But our approach detects regions of disordered protein that Distinguishing disordered regions from ordered regions in protein sequences facilitates the exploration of protein structures and functions. This approach uses only DBD4.0 database and determine the active site using PairPred that obsolete now. Our approach uses DBD5.5 database and determine the active site using BIPSPI

Chapter Three: System Analysis

3.1 Project Specifications

In chapter two, many different methods exist for each different function. In this chapter, we will mention the methods, we had used for each function in our project.

Here, we will explain how the user will interact and the outputof the program.

3.1.1 System Architecture

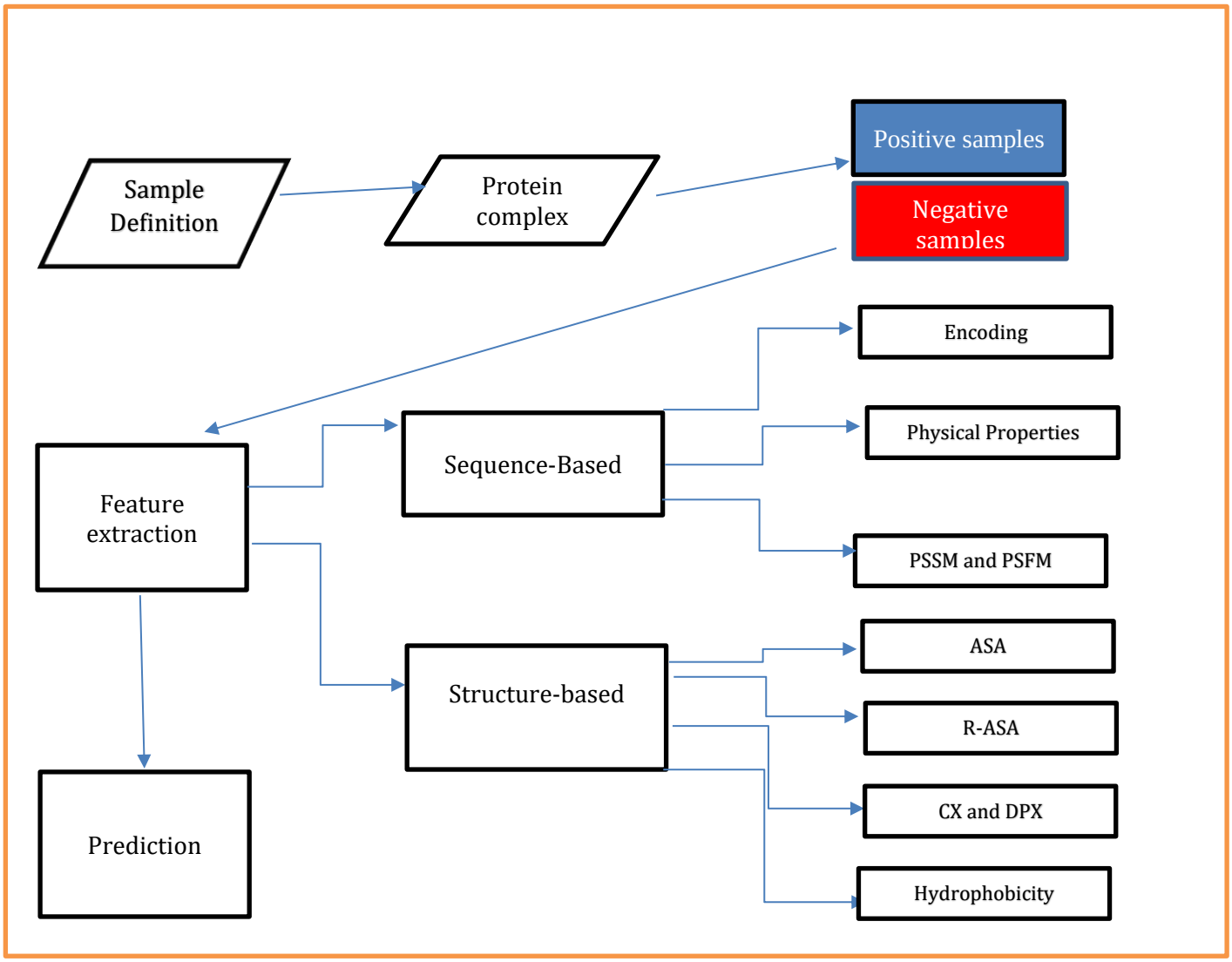


Figure (2): System Architecture

3.2 Functional Requirements

The aim of our research is to find binding regions of protein from a large dataset of protein that will determine if a residual is classified hotspot or not. This final output is obtained by analyzing Feature-Selective Rotation Forest. In other words, we want to do a feature selection. We use the CNN to algorithm to do this.

5.3. Non-functional Requirements

- **Performance**

  Features of proteins (Generated data sets) subset selection works by removing features that are not relevant or are redundant. The subset of features selected should give the best performance according to some objective function. In this proposed method, we examine the use of a CNN Algorithm to do so.

- **Usability**

  The use of Deep learning techniques to automatically analyzedata is becoming increasingly widespread. However, the size of the data to be processed has increased the past 5 years and therefore feature selection has become a requirement before any kind of classification takes place. Our model makes it easier for users to classify the binding sites accurately efficiently for protein-protein interaction.

- **Reliability**

  After applying a features of protein dataset with the resulted subset of binding sites of proteins to a classifier (e.g., Feature-Selective Rotation Forest classifier) the results are supposed to show that the proposed method has **excellent classification** performance which can yield 100% classification accuracy using only a small number of proteins.

3.4 Class Diagram

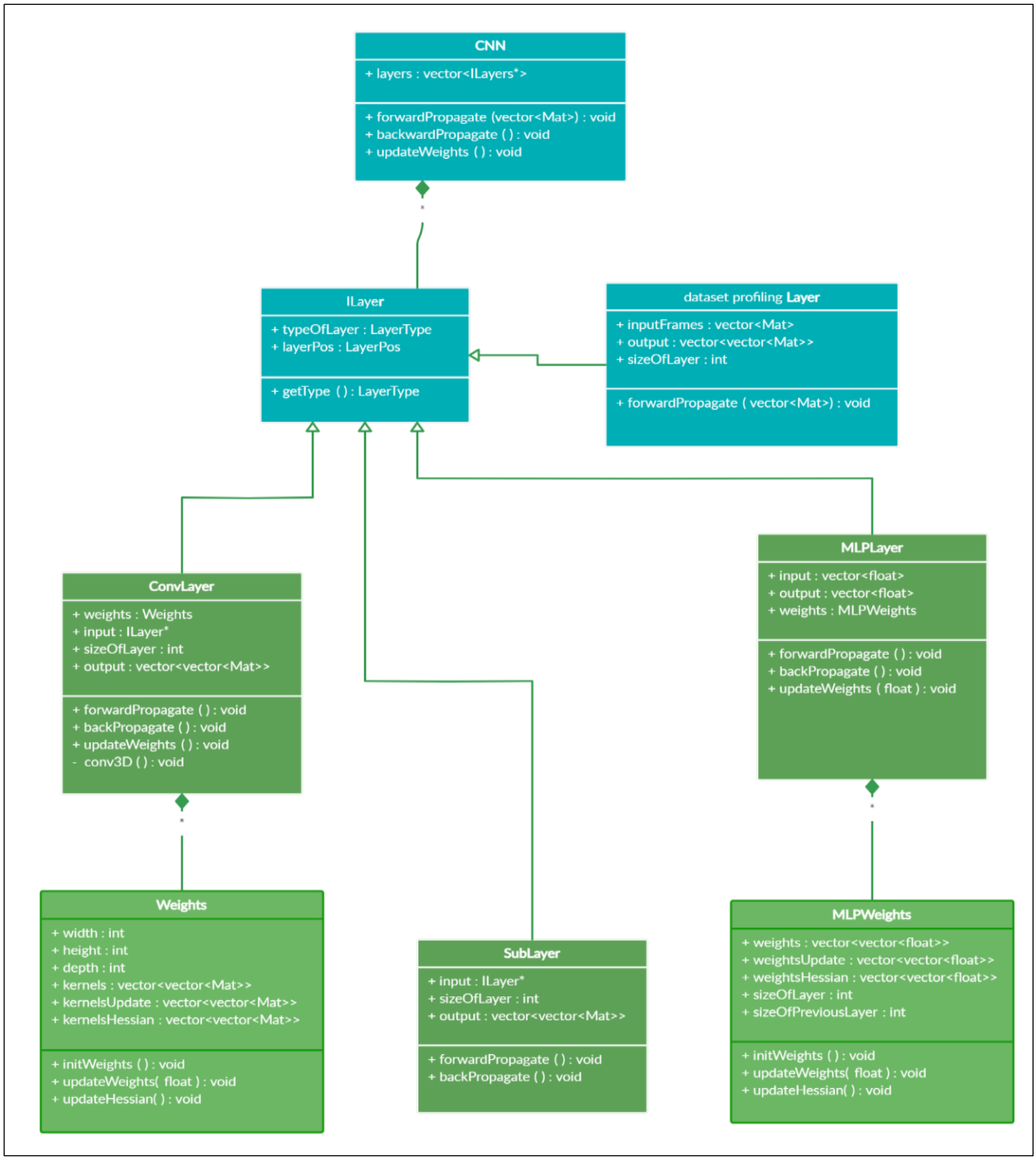


Figure (2): Class Diagram

3.5 Use Case Diagram

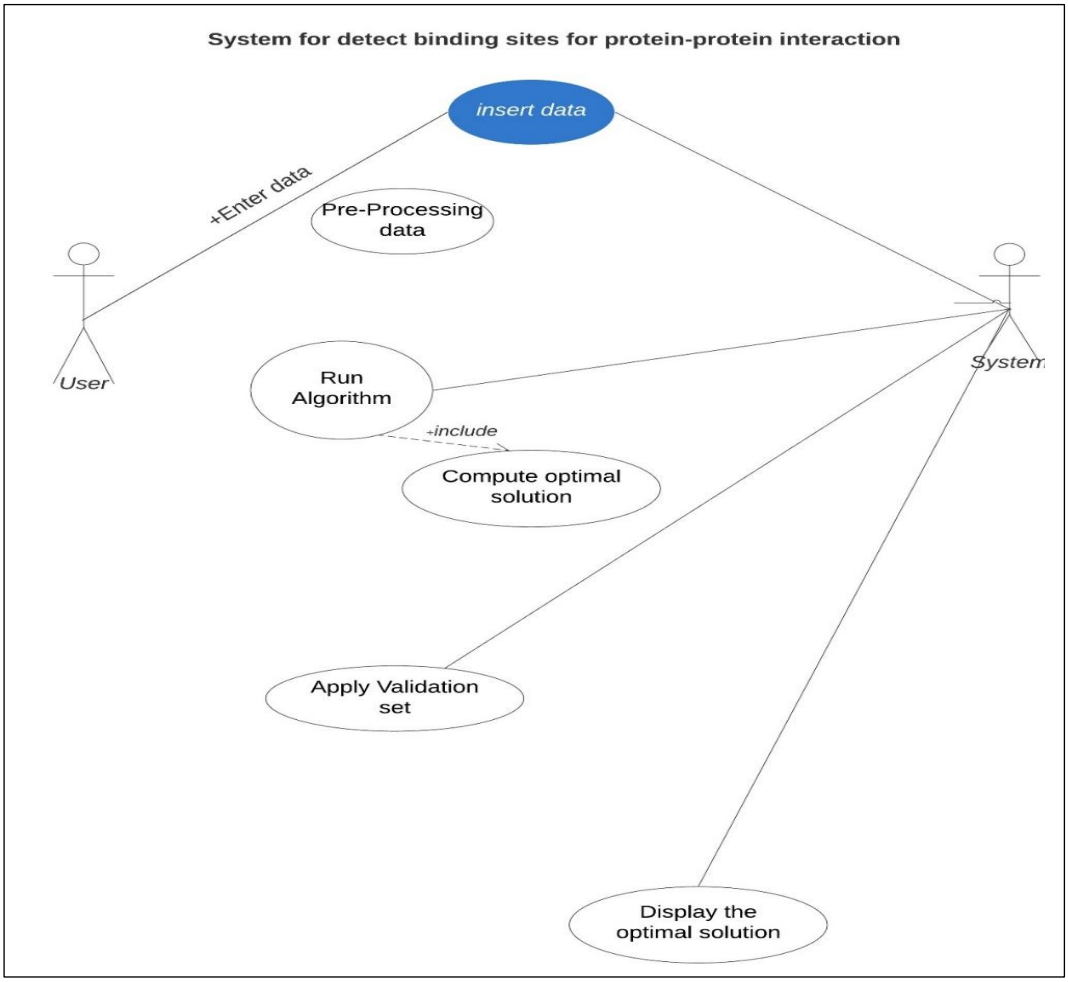


Figure (3): Use Case Diagram

Chapter Four: System Design

4.1 System Component Diagram

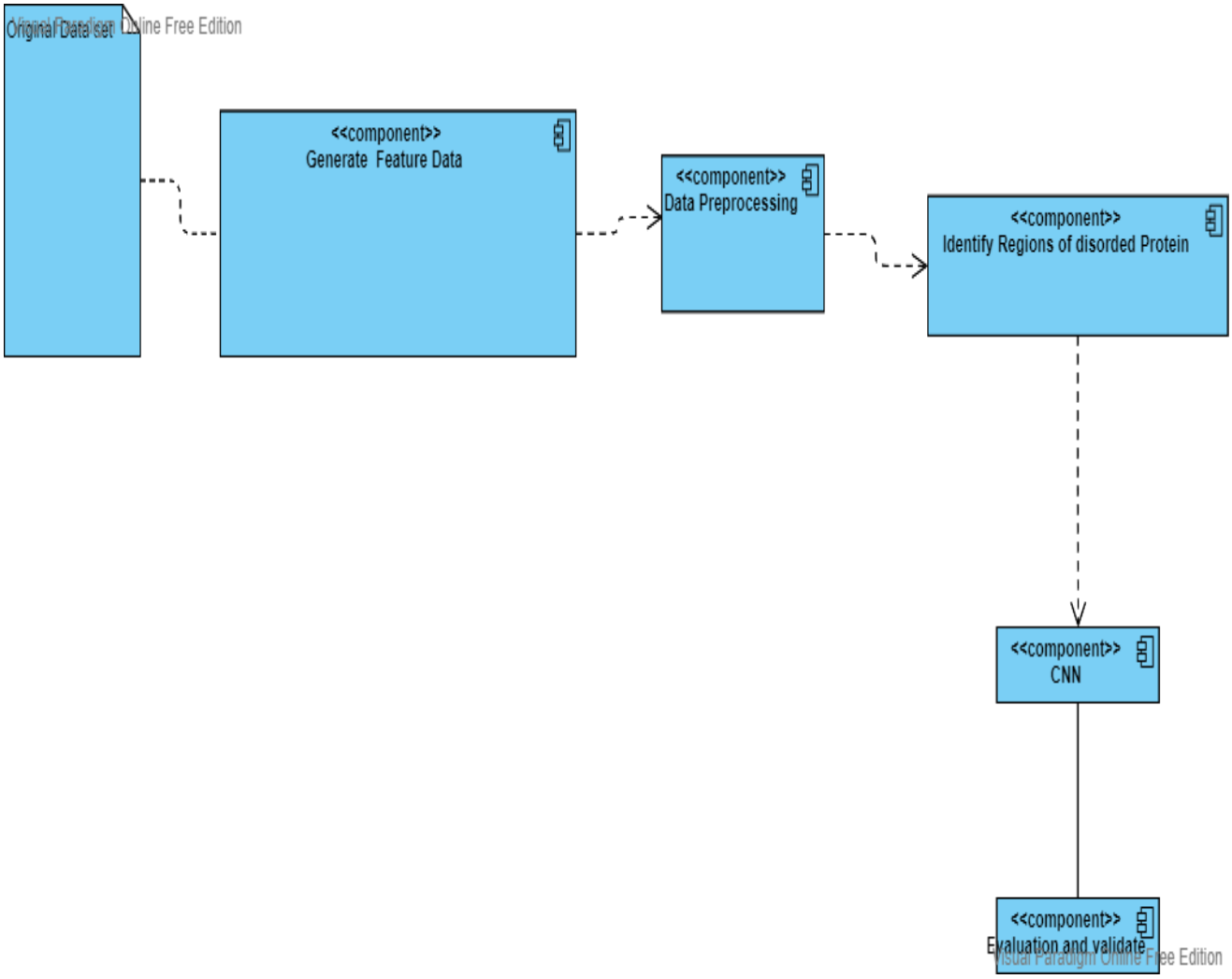


Figure (4): System component diagram

4.2 Sequence Diagram

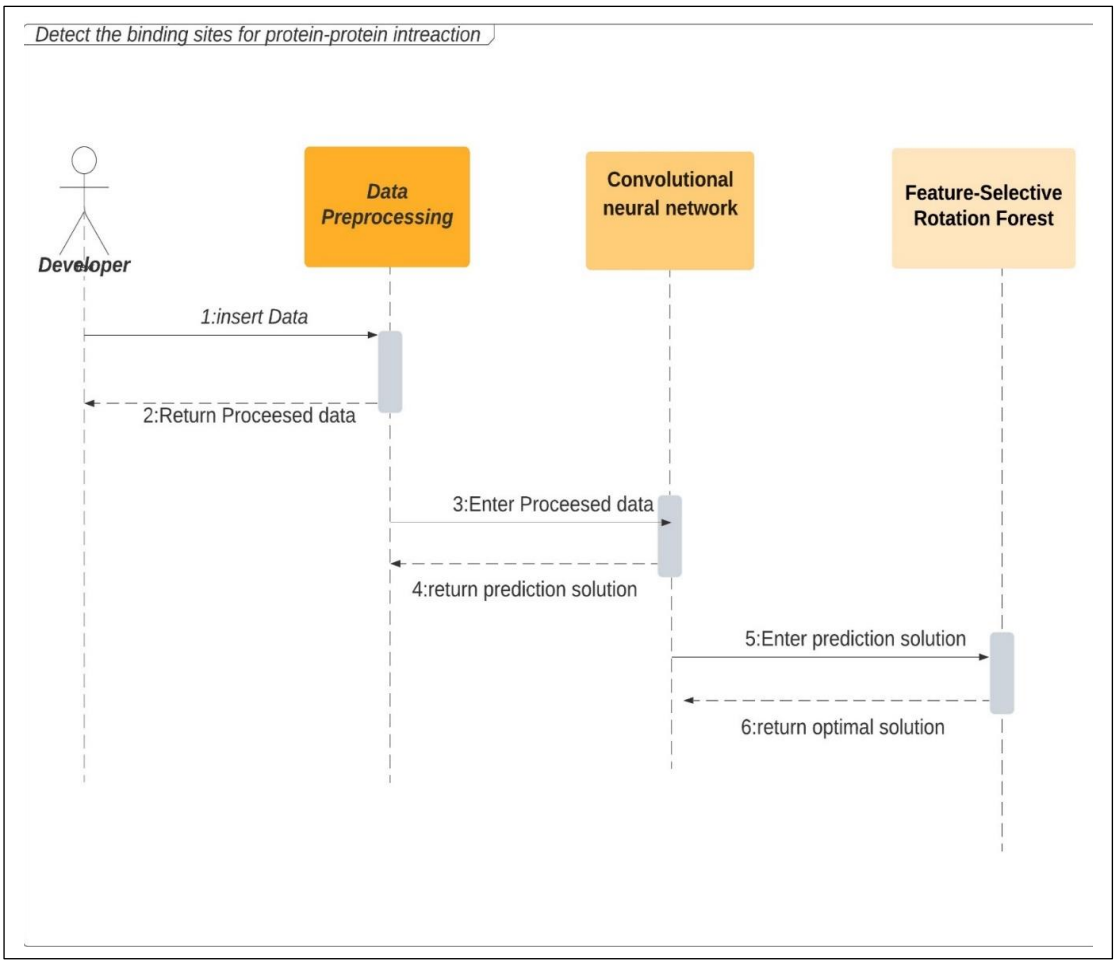


Figure (5): Sequence Diagram

Chapter Five: Model Proposed

5.1 Generate Feature Data

As Popular PPI sites prediction methods can be sorted into three groups according to the information they are based on:

1. Sequence-based methods. Methods based on sequence information use features extracted from protein sequences to predict protein interaction sites.
2. Structure-based methods. Knowledge of the three-dimensional (3D) structure of the protein complex provides much valuable information on the protein interaction sites.

3. Methods based on integrated information. Three-dimensional structure of proteins are far more difficult and expensive to elucidate than protein sequences, so its magnitude in protein structure databases such as the Protein Data Bank (PDB) is remarkably smaller compared to that of sequences in protein sequence databases like UniProt. Therefore, most methods use a combination of structural and sequence information for the prediction of PPI sites.

So, our approach based on Methods based on integrated information. So, from PDB that represent three-dimensional structure of protein from DBD 5.5 contains 270 non-redundant dimers with characterized bound and unbound X-ray crystallography structures. Two interaction protein chains of a dimer are from different families defined by Structural Classification of Proteins (SCOP) with sequence identity less than 30%.

Beside that sequence information use features extracted from protein sequences to predict protein interaction sites. as the position-specific scoring matrix (PSSM) and amino acid composition and achieves an area under the receiver operating characteristic (ROC) curve (AUC)

5.1.2 Features

5.1.2.1 Sequence Features

The following are some of the sequence features:

a) Amino Acid Encoding:

It represents the Twenty amino acids were coded as one-hot encoding

No.	Amino acid	Coding
1	A	10000000000000000000
2	L	01000000000000000000
3	I	00100000000000000000
4	V	00010000000000000000
5	G	00001000000000000000
6	K	00000100000000000000
7	R	00000010000000000000
8	D	00000001000000000000
9	E	00000000100000000000
10	H	00000000010000000000
11	N	00000000001000000000
12	Q	00000000000100000000
13	S	00000000000010000000
14	T	00000000000001000000
15	C	00000000000000100000
16	M	00000000000000010000
17	Y	00000000000000001000
18	W	00000000000000000100
19	F	00000000000000000010
20	P	00000000000000000001

Table (3): amino acid encoding

b) Sequence Features Profile Features Position specific scoring matrix (PSSM) and position specific frequency matrix (PSFM). There are reflect the conservation of residues at specific positions of protein chains based on evolutionary information.

Each row of the PSSM or PSFM is a 20-dimensional vector. PSSM and PSFM were computed by running 3 iterations of PSIBLAST [66] against the UniProtKB/Swiss-Prot database for a given protein with E-value set to 0.001. PSSM and PSFM columns were taken within a length 3 window centered at a residue of the protein to obtain a 3 × 40 matrix.

c) Amino Acid Physicochemical Properties:

The physicochemical properties of a protein are determined by the analogous properties of the amino acids in it.

Twenty-four physicochemical properties of amino acids [67] are used in this our approach. Twenty amino acids are divided into three groups according to these properties and each group is encoded using one-hot encoding, thus each amino acid is represented as a 72-dimensional vector. e.g., alanine (A) is encoded as follows:

A = [0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0]

5.1.2.2 Structure Features Residues

Structure Features Residues that play important roles in protein function generally appear at the surface of proteins.

a) The accessible surface area (ASA) or solvent-accessible surface area (SASA)

   It is the surface area of a biomolecule that is accessible to a solvent. as it defined the locus of the center of the solvent molecule as it rolls over the van der Waals surface of the protein.

b) Protrusion index (CX)

   It simple and fast algorithm is described that calculates a measure of protrusion (cx) for atoms in protein structures, directly useable with the common molecular graphics programs. as cx calculated sphere of predetermined radius is centered around each non-hydrogen atom, and the volume occupied by the protein and the free volume within the sphere (internal and external volumes, respectively). Atoms in protruding regions have a high ratio (cx) between the external and the internal volume [23].

c) Relative accessible surface area (RASA)

   It used to identify whether a residue is at the surface of a protein. (RASA) of a residue indicates a degree of residue solvent exposure. It can be calculated by normalizing the total accessible surface area (ASA) of the residues in a protein structure by the ASA of the residues in the most exposed state to a solvent molecule [24]. The geometric properties of the protein surface can affect the interaction between proteins.

d) depth index (DPX)

Atom depth, originally defined as the distance between a protein atom and the nearest water molecule sur-rounding a protein, is a simple but valuable geometrical descriptor of the protein interior. It can be easily computed from the 3D structure of a protein, thus complementing the information provided by the calculation of the solvent accessible surface area and buried surface area. Depth has been found to be correlated with several molecular, residue and atomic properties, such as average protein domain size, protein stability, free energy of formation of protein complexes, amino acid type hydrophobicity, residue conservation.[25]

e) hydrophobicity

hydrophobic-hydrophilic (HP) interactions happen when a protein is folding into its tertiary structure. The amino acids with hydrophobic side chains move to the core of the protein to be away from water, and the hydrophilic side chains move towards the outside of the protein because they have an affinity with water.

Globular proteins fold by minimizing the nonpolar surface that is exposed to water, while simultaneously providing hydrogen-bonding interactions for buried backbone groups, usually in the form of secondary structures such as α-helices, β-sheets, and tight turns.

the “hydrophobicity of the amino acids in these regions has been quite weak that many amino acid side chains contain considerable nonpolar sections, even if they also contain polar or charged groups. which plays an important role in PPIs, protein folding and unfolding.

These five structure-based features were computed using PSAIA [19], which was developed for calculation of the geometric parameters of large protein structures and the prediction of protein interaction sites.

Definition of Interacting Residue Pairs A pair of residues from two proteins are considered to have interaction if the Euclidean distance between any two atoms from each of the two residues in the bound state is less than or equal to 6 Å. According to this definition,622,331positive samples (interacting residue pairs) and 13,661,568 negative samples (non-interacting residue pairs). The number of negative samples is much larger than that of positive samples. This imbalanced data made it difficult to train the model using Machine learning techniques such as the under-sampling might lead to information loss, so we prefer used the Ensemble algorithm to build the training set with equal positive and negative samples.

5.2 Data Preprocessing

5.2.1 Preprocessing of file components

First, we had our Dataset file preprocessed, removing all Thun unneeded information attached to the dataset in file.

5.2.2 Preprocessing of Dataset

The second step taken is that of preprocessing our data, where a feature selection technique is applied to reduce the amount of input variables that are to be used in the CNN. The aim of preprocessing step is that it generally to give CNN Pattern to understand this numerical feature dataset as image in general use of CNN algorithm

5.2.2.1 Protein features representation

Each protein (ligand and receptor) has a set of amino acids, each amino acid has a set of features that distinguish it from the rest, so each row contains the amino acid and its's features, thus, the protein file will have the following shape:

- Amino acid |23 columns physical Feat |21 columns chemical Feat |40 for PSSM+PSFM

Therefore, a total of 85 column

- So, each protein whether in the ligand state or the receptor state will have 85 columns
- 23 columns representing physical features extracted from PSAIA tool
- Physical features based on the location of each amino acid inside the 3d structure of protein

- 21 columns of chemical features extracted computationally where each amino acid has chemical features that distinguish it from the rest regardless of the location
- 40 PSSM +PSFM matrix from EMBL website

5.2.2.2 Protein-protein interaction computational representation

- The amino acids of the ligand protein are checked if they interact with amino acids of the receptor protein as follows:
- each amino acid in receptor is checked with all amino acids of ligand once ,meaning if the ligand protein has 100 A.A and receptor has 50 so, the ligand protein will be repeated 50 times each time corresponding to one amino acid of the receptor while each amino acid in the receptor will be repeated 100 times each time corresponding to the whole length of the ligand protein giving a total of 100*50=5000 rows this is similar to cross joining .

- Proteins are found in the bound state (complex) and unbound state (single)
- Features are extracted from proteins in the unbound state
- Positive and negative samples are extracted from proteins in bound state

5.2.2.3 Disorder between proteins in bound and unbound state:

- 3D structure of protein in the bound state is different than the structure of the same protein in the unbound state
- The features are generated from the protein in the unbound state
- Positive and negative samples are generated from the bound state of protein
- The disorder between two proteins must be removed so that the positive and negative samples represent the interaction between two proteins

remove disorder between two proteins
global alignment is a technique that aligns two strings together and outputs the best similar string between them

Since features are generated, positive and negative samples are generated so the unnecessary data must be removed from them to get the similarity by:

1. perform global alignment on the ligand protein in the unbound state and the ligand protein in the bound state.
2. Identify regions to be removed represented by an astrik.
3. Insert the astrik into the two proteins to be removed.
4. Remove the regions in question.

These steps are performed in the ligand proteins and in the receptor proteins.

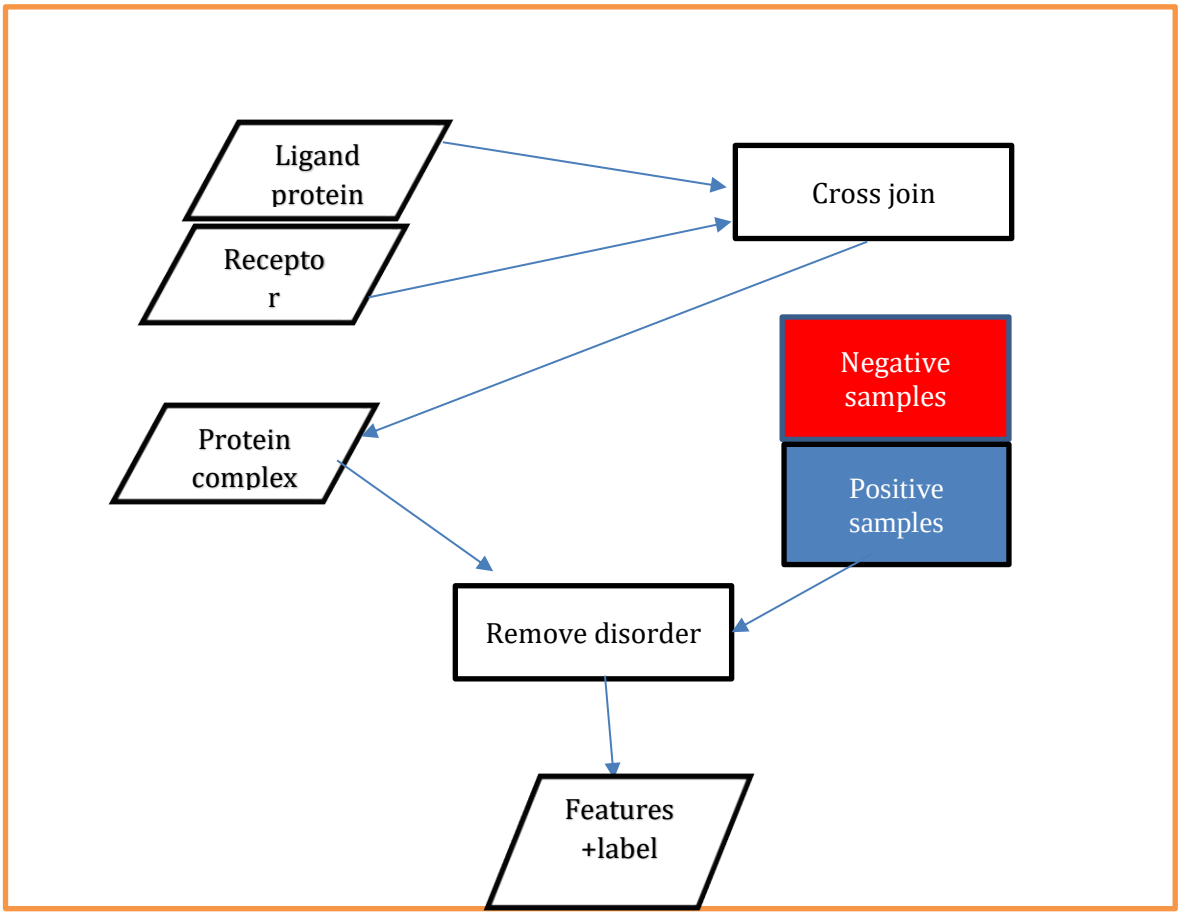


Figure 5.2.1 data Preprocessing

5.3 Convolutional Neural Network (ConvNet/CNN)

CNN is a Deep Learning algorithm which can take in an input assign importance (learnable weights and biases) to various aspects/objects in the input and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNet have the ability to learn these filters/characteristics.
Which of these are reasons for Deep Learning recently taking off?

As we have access to a lot more computational power and have access to a lot more data. Beside that Deep learning has resulted in significant improvements in important applications such as online advertising, speech recognition, and image recognition.

Convolutional neural Network
Convnets contain one or more of each of the following layers**:**

1. Convolution layer
2. Pooling layer
3. Batch Normalization layer
4. ReLU (rectified linear units) layer (element wise threshold)
5. Dense /Fully connected layer
6. Loss layer (during the training process)

- **Convolution layer**

a convnet processes an input using a matrix of weights called filters (or features) that detect specific attributes such as diagonal edges, vertical edges, etc. Moreover, as the input progresses through each layer, the filters able to recognize more complex attributes.

- **Pooling layer**

the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to decrease the computational power required to process the data through dimensionality reduction. Furthermore, it is useful for extracting dominant features which are rotational and positional invariant, thus maintaining the process of effectively training of the model. There are two types of Pooling: Max Pooling and Average Pooling. Max Pooling returns the maximum value from the portion of the input covered by the Kernel. Average Pooling returns the average of all the values from the portion of the input covered by the Kernel.

- **Batch Normalization layer**

Batch normalization is a technique for training very deep neural networks that standardizes the inputs to a layer for each mini batch. This has the effect of stabilizing the learning process and dramatically reducing the number of training epochs required to train deep networks

- **ReLU (rectified linear units) layer**

The ReLU (short for rectified linear units) layer commonly follows the convolution layer.

• The addition of the ReLU layer allows the neural network to account for non-linear relationships,

• The ReLU function takes a value y and returns 0 If y is negative and y if y is positive

Advantage of Relu:

Simplifies backprop - Makes learning faster - Make feature sparse

- **Dense/fully connected layer**

The dense layer is a neural network layer that is connected deeply, which means each neuron in the dense layer receives input from all neurons of its previous layer. The dense layer is found to be the most used layer in the models.

In the background, the dense layer performs a matrix-vector multiplication. The values used in the matrix are parameters that can be trained and updated with the help of backpropagation.

The output generated by the dense layer is an 'm' dimensional vector. Thus, dense layer is basically used for changing the dimensions of the vector. Dense layers also applies operations like rotation, scaling, translation on the vector.

- **Loss layer (during the training process)**

The function we want to minimize or maximize is called the objective function or criterion. When we are minimizing it, we may also call it the cost function, loss function, or error function. In calculating the error of the model during the optimization process, a loss function must be chosen. This can be a challenging problem as the function must capture the properties of the problem and be motivated by concerns that are important to the project and stakeholders.

5.4 Validation Function

Finally, after obtaining the subset of binding sites a validation step needs to be taken, that is passing our list of binding sites obtained to a validation function, to test the actual efficiency of these binding discrimination power across some given classified-dataset. Validation function used in our approach is K-Fold Cross-Validation and K-Nearest-Neighbors.

5.3.1 k-Fold Cross-Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=5 becoming 5-fold cross-validation.

That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    1. Take the group as a hold out or test data set
    2. Take the remaining groups as a training data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

5.3.2 K-Nearest-Neighbors Validation:

Validation function used in our case is K-Nearest-Neighbors (KNN) classifier function, a supervised machine learning model.
K-NN models work by taking a data point and lookingat the ‘k’ closest labeled data points. The data point is then assigned the label of the majority of the ‘k’ closest points.

Algorithm ––

- Let **m** be the number of training data samples. Let **p** bean unknown point.

1. Store the training samples in an array of data points **arr[].** *This means each element of this array representsa tuple (x, y).*
2. for i=0 to m:
3. Calculate Euclidean distance d(arr[i], p).
4. Make set S of K smallest distances obtained. Each of these distances corresponds to an already classified datapoint.
5. Return the majority label among **S**.

Chapter Six: Model Setup and Results

6.1 Parameter settings

The setting of CNN parameters is very important in the CNN and hybrid model method. For CNN parameters, if the dataset size is too small, it is difficult to get the best resolution.

This goes back to the idea of understanding what we are doing with a convolution neural net, which is basically trying to learn the values of filter(s) using backprop. In other words, if a layer has weight matrices, that is a “learnable” layer.

CNN Parameters	Benchmark Data 5.5
Conv1D	4 layers
Padding	Max
pooling	7x7
Stride	3

Table 6.1: CNN Algorithm parameter settings

- **batch_size** determines the number of samples in each mini batch. Its maximum is the number of all samples, which makes gradient descent accurate, the loss will decrease towards the minimum if the learning rate is small enough, but iterations are slower. Its minimum is 1, resulting in stochastic gradient descent: Fast but the direction of the gradient step is based only on one example, the loss may jump around. batch_size allows to adjust between the two extremes: accurate gradient direction and fast iteration. Also, the maximum value for batch size may be limited if your model + data set does not fit into the available (GPU) memory.
- **steps_per_epoch** the number of batch iterations before a training epoch is considered finished. If you have a training set of fixed size you can ignore it but it may be useful if you have a *huge* data set or if you are generating random data augmentations.
- **validation_steps** similar to steps_per_epoch but on the validation, data set instead on the training data. If you have the time to go through your whole validation data set I recommend to skip this parameter.

Tensor board:

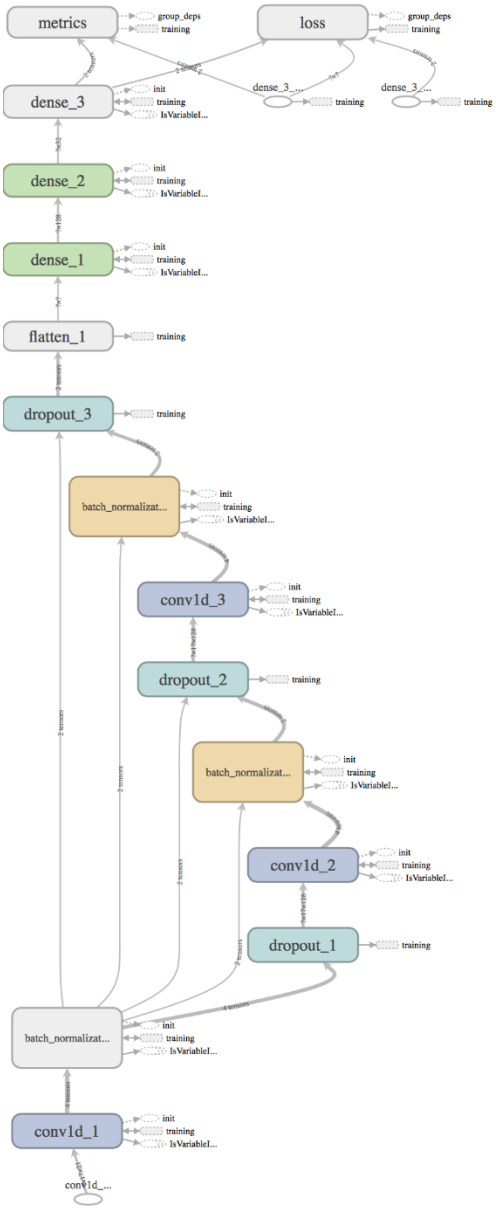
It provides the visualization and tooling needed for machine learning experimentation as:

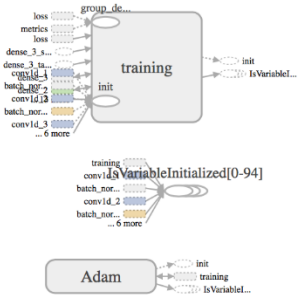
a) Tracking and visualizing metrics such as loss and accuracy

b) Visualizing the model graph (ops and layers)

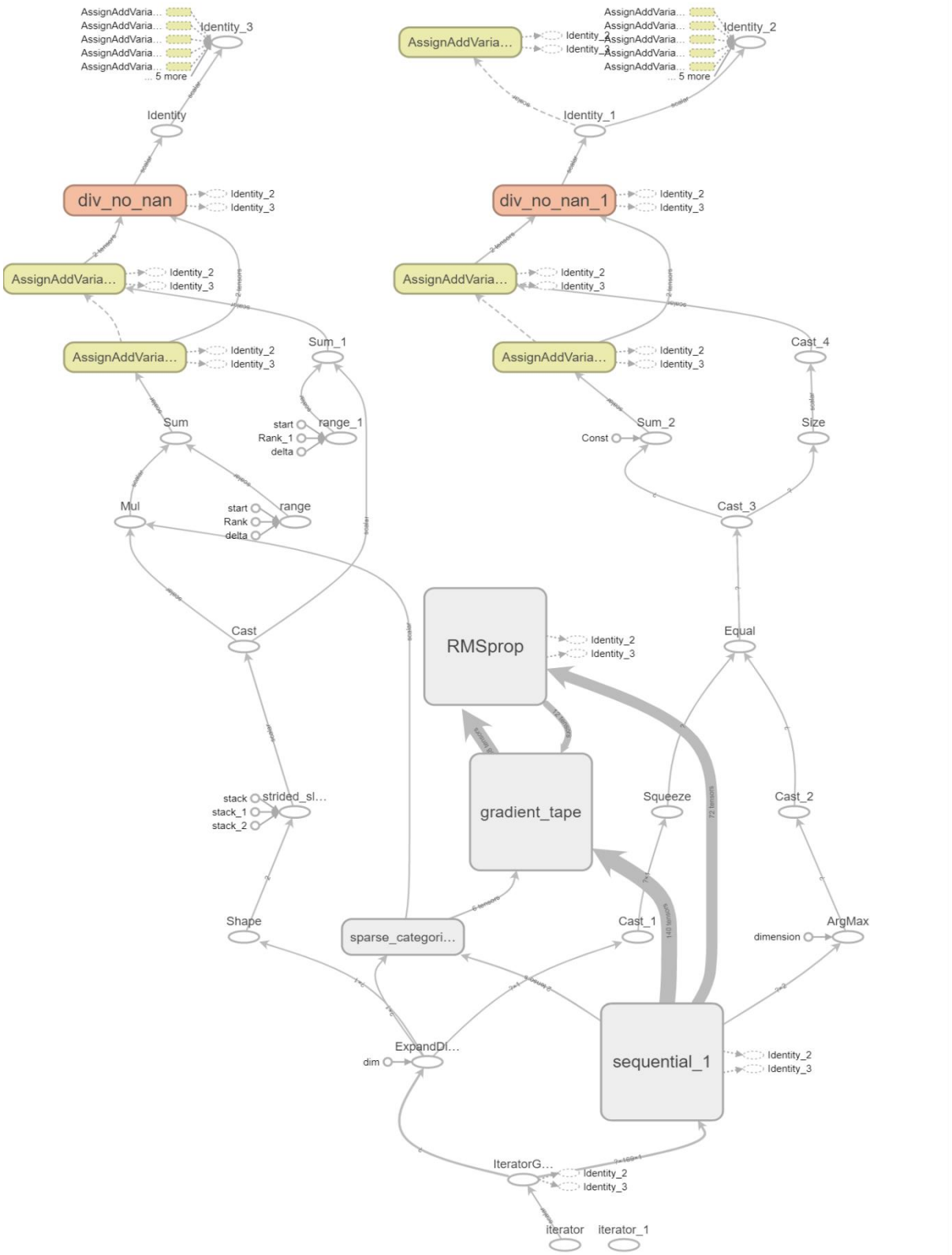
The next figures Will Provide the Architecture of each Deep learning model

First CNN Architecutre

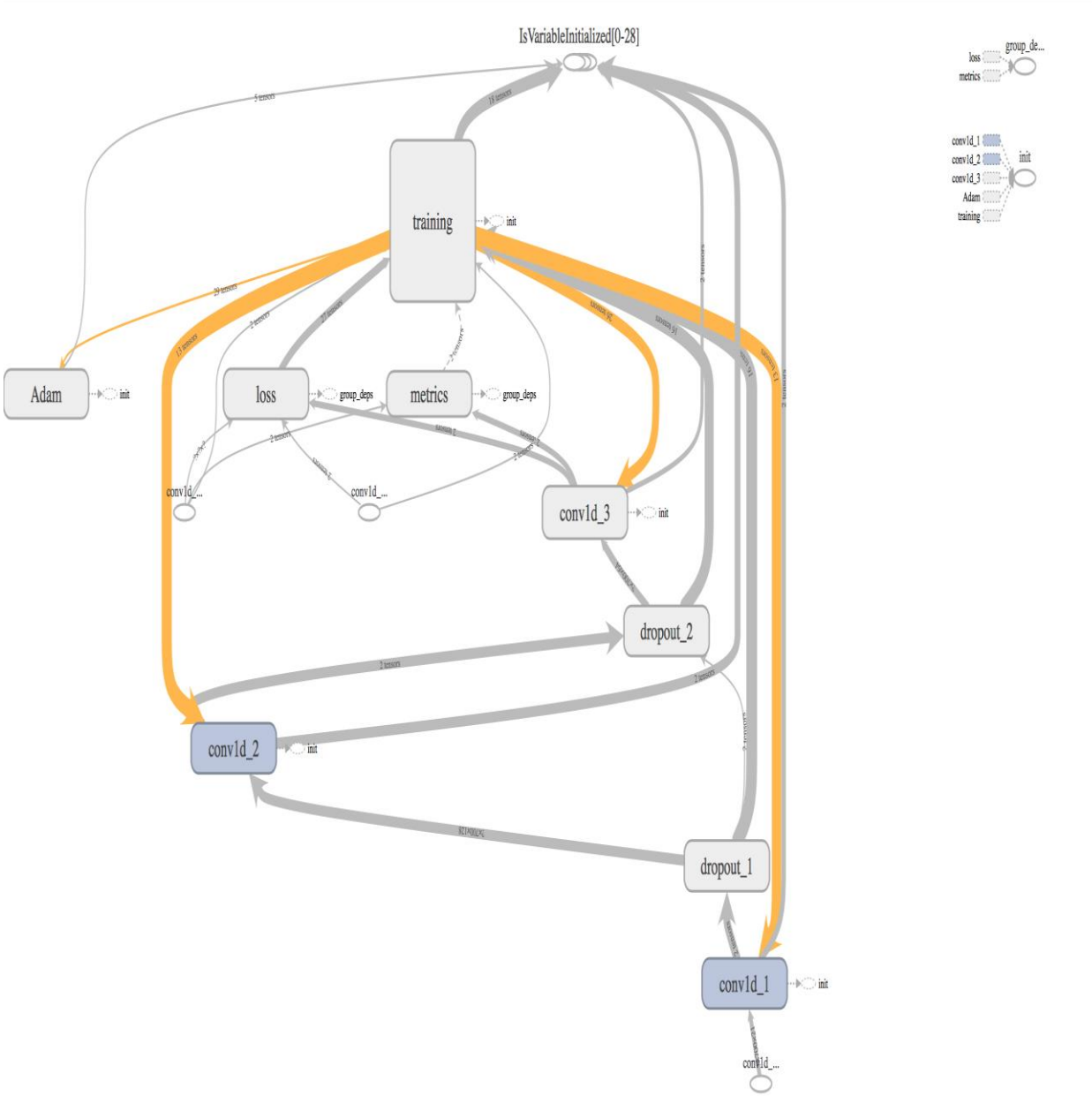




Second CNN Architecutre:



Third CNN Architecture:



6.2 Model Results

The process of CNN was repeated many times by try and justify deep learning models and the intersection between results is considered as the best result,the Area Under the Curve Is 0.97 that reflect that the highest accuracy is about 92

First CNN Model:

CNN parameters	Benchmark data 5.5
Conv1D	2 layers
Padding	Max
Pooling	2x2
Stride	1

Table (4) :FIRST CNN model parameter

Model results:

AUC of the model on the threshold 0.9 dataset is represented in this figure

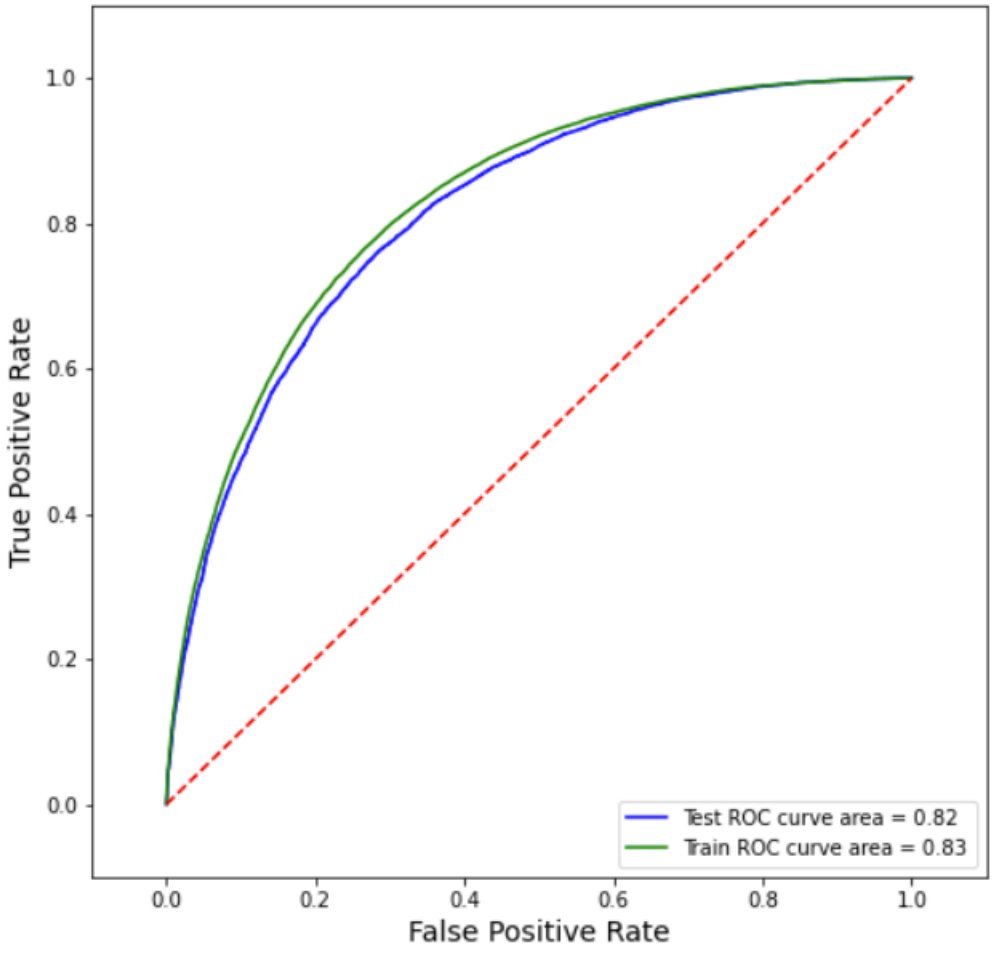


Figure 7: AUC-First model of CNN

Model loss

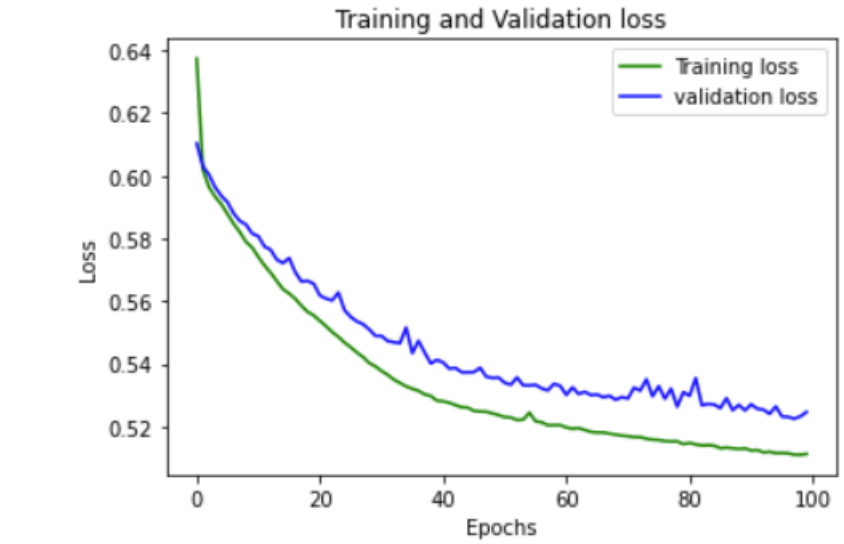


Figure 8: training and validation loss of first CNN

The loss decreases gradually as the number of epochs increases however even at epoch 100 The loss is still considerably high

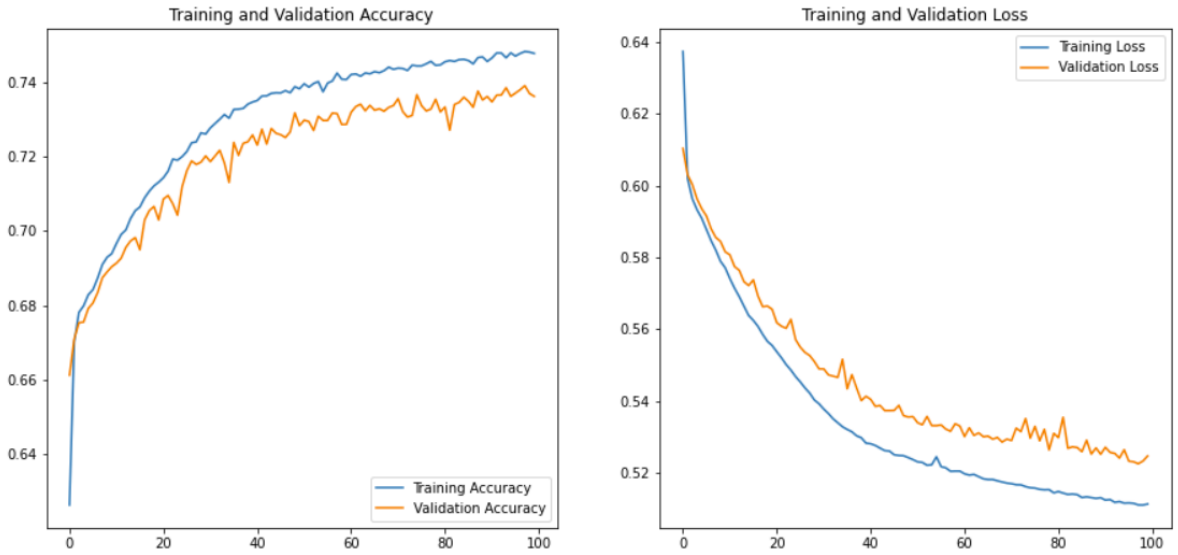


Figure 9: accuracy and loss comparison of first model

Table1 . Accuracy of our model in different threshold.

Threshold	Accuracy	model
0.5	0.70	CNN1
0.6	0.71	CNN1
0.7	0.73	CNN1
0.8	0.73	CNN1
0.9	0.74	CNN1

Table (4): accuracy cross threshold of First CNN Model

Second: Inception_CNN model

CNN parameters	Benchmark data 5.5
Conv1D	5layers
Padding	Max
Pooling	2x2
Stride	5
Batch size	128
Dense Layer	3 layer

Model results:

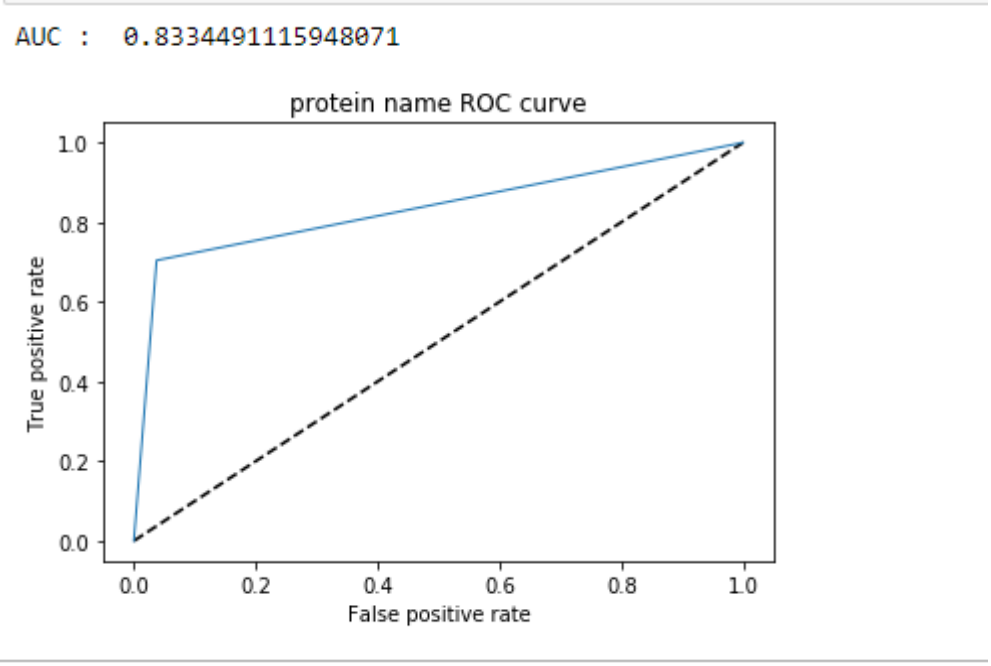


Figure 10: AUC Inception CNN model

Table . Accuracy of our model in different threshold.

Threshold	Accuracy	model
0.5	0.74	CNN2
0.6	0.745	CNN2
0.7	0.75	CNN2
0.8	0.76	CNN2
0.9	0.82	CNN2

Table (5) accuracy cross threshold of Second CNN Model

Third: Res-net CNN

CNN parameters	Benchmark data 5.5
Conv1D	5layers
Padding	Max
Pooling	2x2
Stride	5
Batch size	16

Table (6) res-net CNN parameters

Model results:

AUC of the model on the threshold 0.9 dataset is represented in this figure

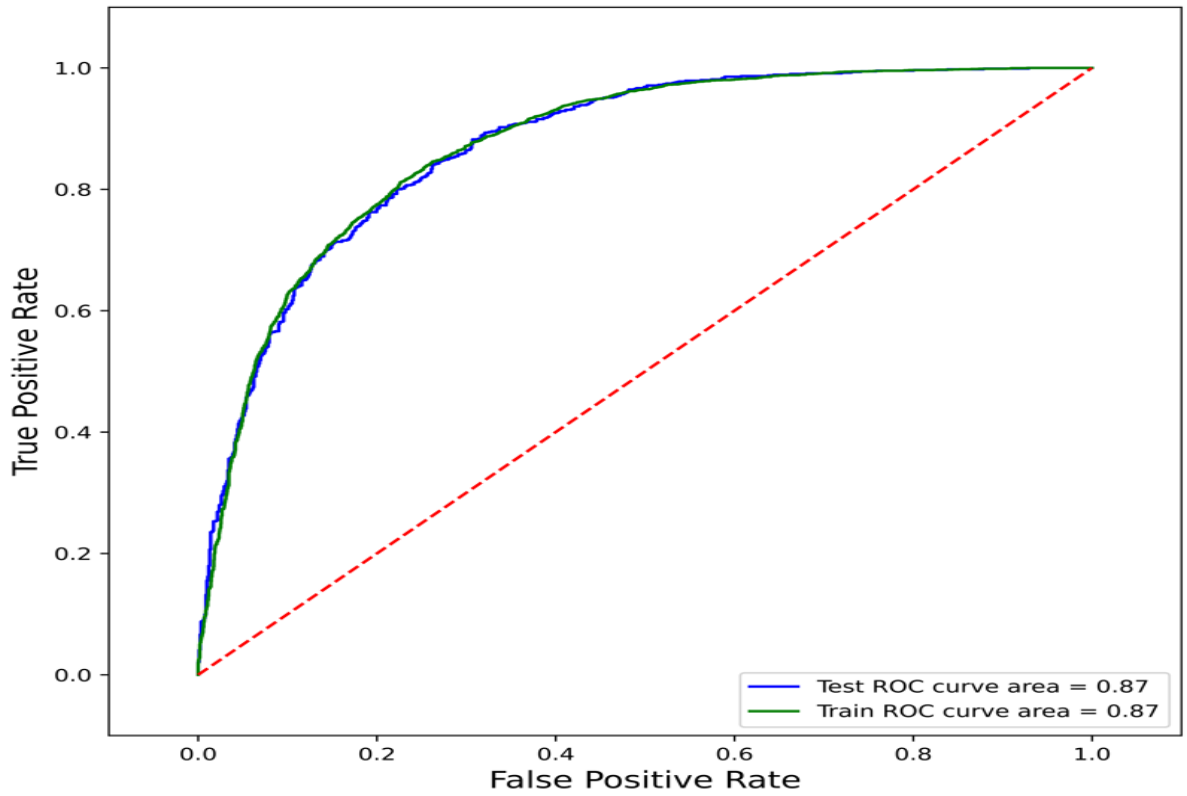


Figure 11: AUC RES NET CNN

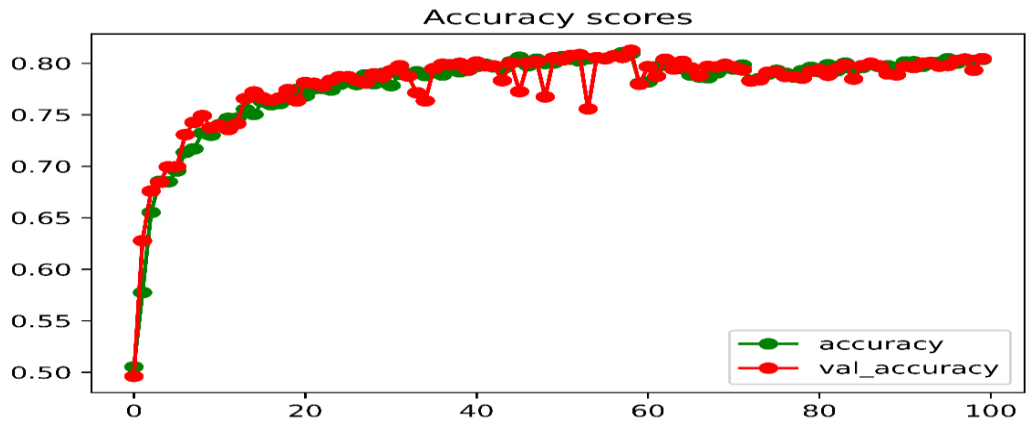


Figure 12: accuarcy of RES NET CNN

Table2 . Accuracy of our model in different threshold.

Threshold	Accuracy	model
0.5	0.76	CNN3
0.6	0.77	CNN3
0.7	0.77	CNN3
0.8	0.81	CNN3
0.9	0.84	CNN3

Table (7) accuracy cross threshold of Second CNN Model

Four Model :U-net CNN

CNN parameters	Benchmark data 5.5
Conv1D	3layer forward and 2 backward
padding	Max
pooling	same
Stride	3
Batch size	128

Table 8 u-net parameters CNN model

Model results:
AUC of the model on the threshold 0.9 dataset is represented in this figure

as a validation step we ran the K Nearest Neighbors and K-Fold Cross Validation subset and it gave 99% accuracy.

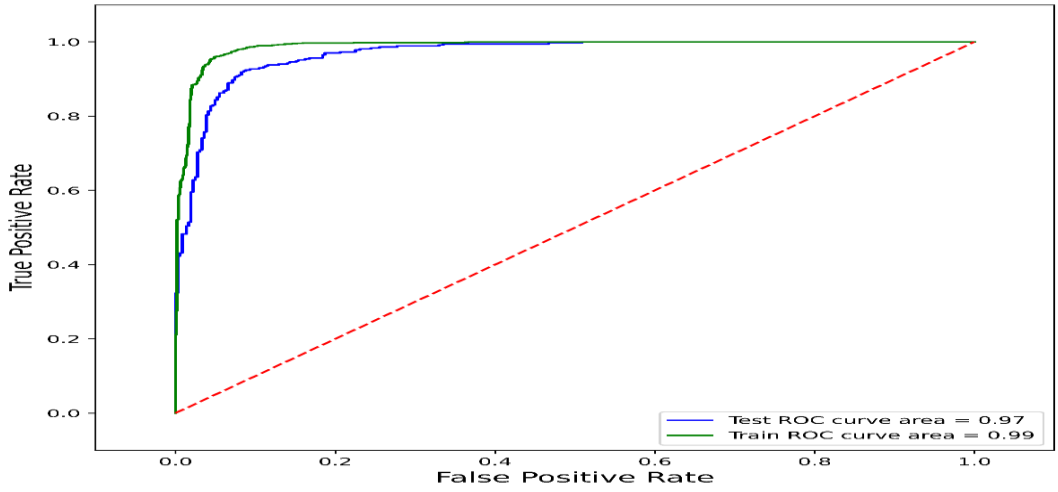


Figure 13 AUC OF U- model

Table . Accuracy of our model in different threshold.

Threshold	Accuracy	model
0.5	0.76	CNN4
0.6	0.77	CNN4
0.7	0.77	CNN4
0.8	0.85	CNN4
0.9	0.89	CNN4

Table (8) accuracy cross threshold of Second CNN Model

Five Hybrid Model:

This are use 1D CNN to extract features and Transform Classifier to train model

CNN parameters	Input Shape
Conv1D	3 layer
padding	Max
pooling	Same
Stride	3
Transform Layer	2 (CNDNNLSTM)
Batch size	128

Model results:
Accuracy of the model on the threshold 0.9 dataset is represented in this figure

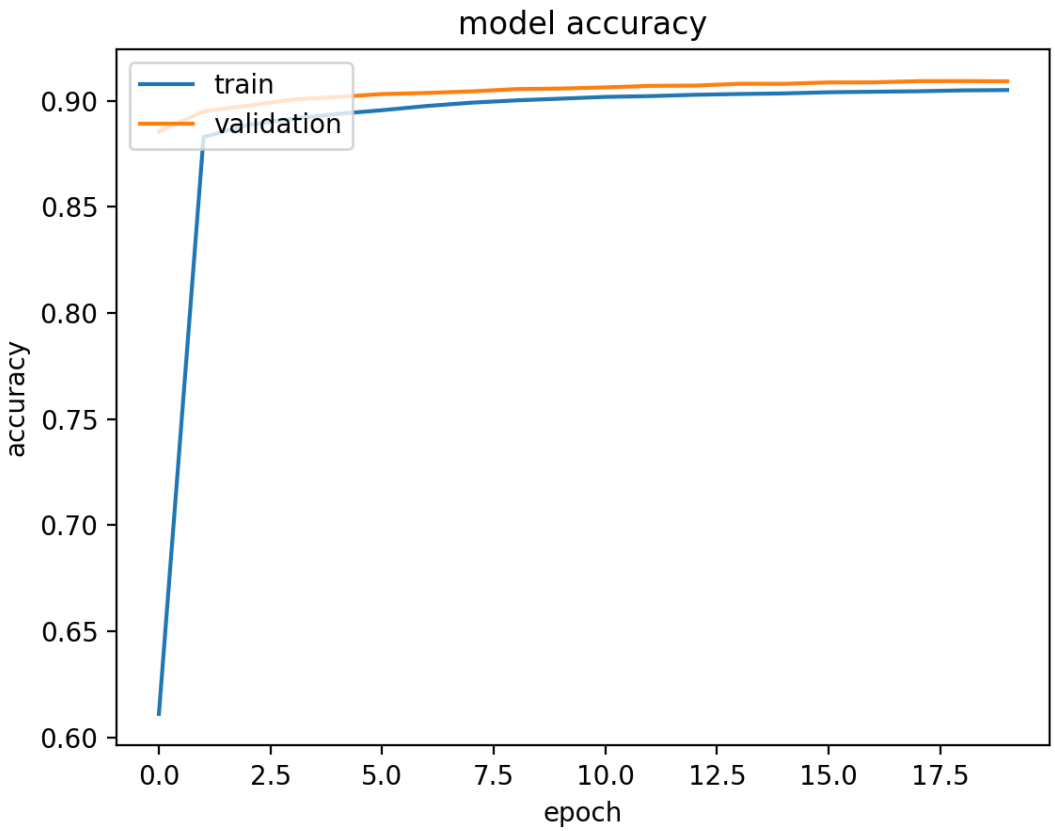


Figure 14: accuracy of hybrid model

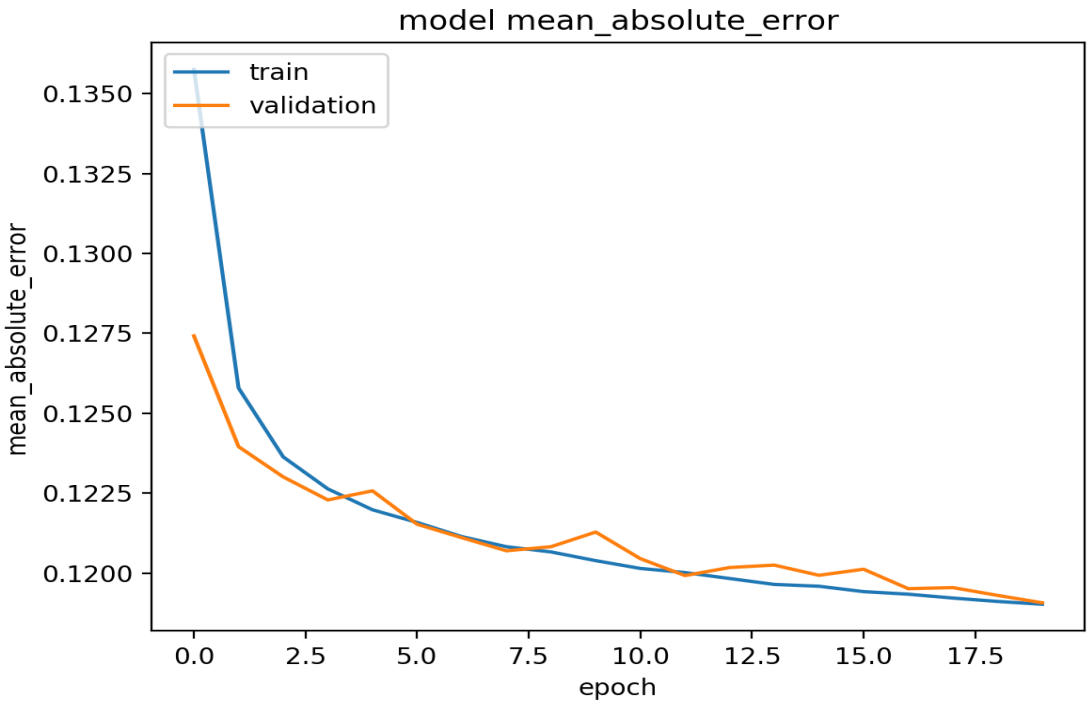


Figure 15: hybrid model mean error

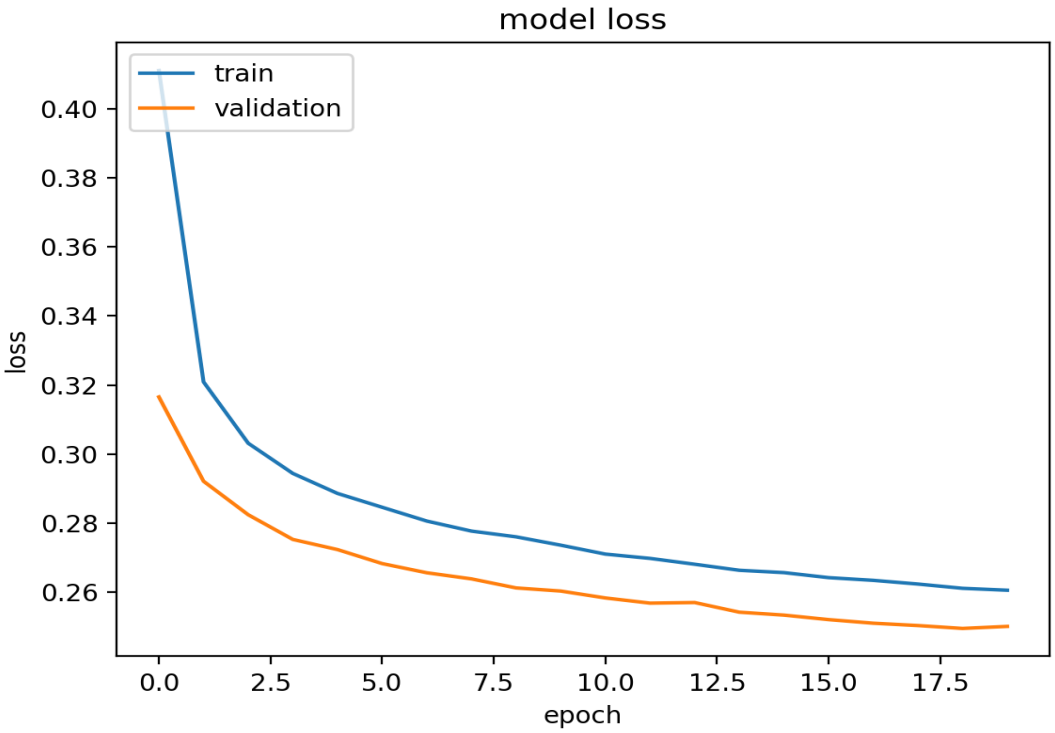


Figure 16: hybrid model loss

Table. Accuracy of our model in different threshold.

Threshold	Accuracy	model
0.5	0.84	TRANS+CNN
0.6	0.81	TRANS+CNN
0.7	0.83	TRANS+CNN
0.8	0.89	TRANS+CNN
0.9	93	TRANS+CNN

Table (9) accuracy cross threshold of Hybrid Model

RFC : applies multiple decision trees to dataset

This approach is convenient with large datasets and heterogenous features

Table2 . Accuracy of our model in different threshold.

Threshold	Accuracy	model
0.5	0.87	RF
0.6	0.88	RF
0.7	0.885	RF
0.8	0.895	RF
0.9	0.9	RF

Table (10) accuracy cross threshold of RFC Model

6.3 Conclusion

In this project, we describe a novel computational method for predicting protein interaction sites. The result may provide insight into areas, such as mutant design and the investigation of protein interaction networks. In view of the difficulties of prediction of Binding sites without consider secondary and tertiary structures for protein complexes, we adopted a structure-based approach, in which features for the learning process are exclusively derived from protein sequence and protein structure. In addition to classical sequence features, such as amino acid conservation and physicochemical properties, we considered the positional preference of interacting sites within a neighboring region in the protein sequence and converted it into sequence features. Experimental results show that a perfect prediction is obtained. At the same time, the effects of different CNN parameters. And random forest classifiers on classification accuracy are analyzed. Finally, the Most binding sites are listed and their expression levels in different samples shows clear separation between the two classes.

We also demonstrated that incorporating B-factor data into our pipeline may further improve the prediction performance.

It is anticipated that the findings derived from this investigation will provide useful clues for further in-depth studies in the problem of predicting Protein-Protein interaction sites.

Accuracy comparison across all models

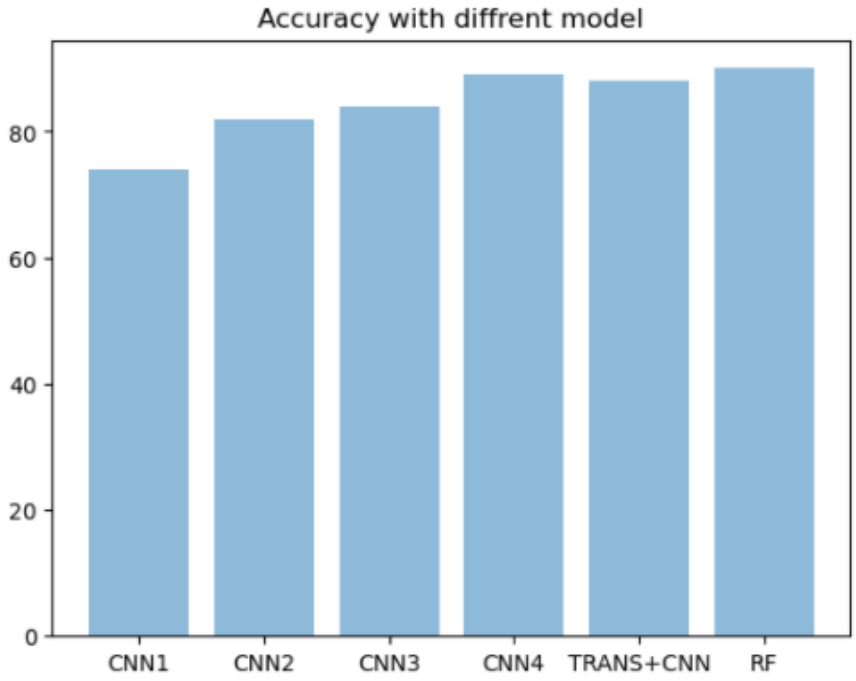


Figure 17: comparative bar plot of all models

References

- 1. James R. Bradford, David R. Westhead, Improved prediction of protein–protein binding sites using a support vector machines approach, *Bioinformatics*, Volume 21, Issue 8, Pages 1487–1494,

- 2. Zhang, L., Yu, G., Guo, M. *et al.* Predicting protein-protein interactions using high-quality pairs. *BMC Bioinformatics* 19, 525 (2018). https://doi.org/10.1186/s12859-018-2525-3

- 3. Kuo TH, Li KB. Predicting Protein-Protein Interaction Sites Using Sequence Descriptors and Site Propensity of Neighboring Amino Acids. *Int J Mol Sci*. 2016;17(11):1788. Published 2016 Oct 26. doi:10.3390/ijms17111788

- 4. Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins. 2005 Oct 1;61(1):21-35. doi: 10.1002/prot.20514. PMID: 16080151.

- 5. Dechang Pi and Chishe Wang, The Prediction of Protein-Protein Interaction Sites Based on RBF Classifier Improved by SMOTE (hindawi.com)

- 6. van Delft MF, Huang DC. How the Bcl-2 family of proteins interact to regulate apoptosis. Cell Res. 2006 Feb;16(2):203-13. doi: 10.1038/sj.cr.7310028. PMID: 16474435.

- 7. Sebastián Maldonadoa ,Richard Weberb , Fazel ,Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines - ScienceDirect

- 8. Xie Z, Deng X, Shu K. Prediction of Protein-Protein Interaction Sites Using Convolutional Neural Network and Improved Data Sets. *Int J Mol Sci*. 2020;21(2):467. Published 2020 Jan 11. doi:10.3390/ijms21020467

- 9. LeiWang , Hai-FengWang , San-Rong Liu, XinYan& Ke-Jian Song Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest | Scientific Reports (nature.com)

- 10. Ruben Sanchez-Garcia, C O S Sorzano, J M Carazo, Joan Segura, BIPSPI: a method for the prediction of partner-specific protein–protein interfaces, Bioinformatics, Volume 35, Issue 3, 01 February 2019, Pages 470–477, https://doi.org/10.1093/bioinformatics/bty647

- 11. Cunliang Geng, Yong Jung, Nicolas Renaud, Vasant Honavar, Alexandre M J J Bonvin, Li C Xue , iScore: a novel graph kernel-based function for scoring protein–protein docking models | Bioinformatics | Oxford Academic (oup.com)

- 12. Xiao Wang,Protein docking model evaluation by 3D deep convolutional neural networks | Bioinformatics | Oxford Academic (oup.com)

- 13. Deng, L., Guan, J., Dong, Q. *et al.* Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinformatics* **10,** 426 (2009). https://doi.org/10.1186/1471-2105-10-426

- 14. Wang, L., Wang, HF., Liu, SR. *et al.* Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest. *Sci Rep* **9,** 9848 (2019). https://doi.org/10.1038/s41598-019-46369-4

- 15. Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia, Michael L. Tress, Progress and challenges in predicting protein–protein interaction sites, *Briefings in Bioinformatics*, Volume 10, Issue 3, May 2009, Pages 233–246, https://doi.org/10.1093/bib/bbp021

- 16. Xing S, Wallmeroth N, Berendzen KW, Grefen C. Techniques for the Analysis of Protein-Protein Interactions in Vivo. Plant Physiol. 2016 Jun;171(2):727-58. doi: 10.1104/pp.16.00470. Epub 2016 Apr 25. PMID: 27208310; PMCID: PMC4902627.

- 17. Warwicker J. Investigating protein-protein interaction surfaces using a reduced stereochemical and electrostatic model. J Mol Biol. 1989 Mar 20;206(2):381-95. doi: 10.1016/0022-2836(89)90487-7. PMID: 2541255.

- 18. Techniques for the Analysis of Protein-Protein Interactions in Vivo. Plant Physiol. 2016 Jun;171(2):727-58. doi: 10.1104/pp.16.00470. Epub 2016 Apr 25. PMID: 27208310; PMCID: PMC4902627

- 19. Mihel, J., Šikić, M., Tomić, S. *et al.* PSAIA – Protein Structure and Interaction Analyzer. *BMC Struct Biol* **8,** 21 (2008). https://doi.org/10.1186/1472-6807-8-21

- 20. Clark, J.J., Orban, Z.J. & Carlson, H.A. Predicting binding sites from unbound versus bound protein structures. *Sci Rep* **10,** 15856 (2020).

- 21. Ruben Sanchez-Garcia, C O S Sorzano, J M Carazo, Joan Segura, BIPSPI: a method for the prediction of partner-specific protein–protein interfaces, *Bioinformatics*, Volume 35, Issue 3, 01 February 2019, Pages 470–477, https://doi.org/10.1093/bioinformatics/bty647
- 22. Wang, L., Wang, HF., Liu, SR. *et al.* Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest. *Sci Rep* **9,** 9848 (2019). https://doi.org/10.1038/s41598-019-46369-4
- 23. Alessandro Pintar, Oliviero Carugo, Sándor Pongor, CX, an algorithm that identifies protruding atoms in proteins, *Bioinformatics*, Volume 18, Issue 7, July 2002, Pages 980–984,
- 24. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science. 1985 Aug 30;229(4716):834-8. doi: 10.1126/science.4023714. PMID: 4023714.
- 25. Pintar, Alessandro & Carugo, Oliviero & Pongor, Sándor. (2003). Atom depth in protein structure and function. Trends in biochemical sciences. 28. 593-7. 10.1016/j.tibs.2003.09.004.