International Journal of Control and Automation
Vol. 13, No. 4, (2020), pp. 707 - 717

Affinity Prediction Models from Protein-Protein Interaction of SCA Using Ensemble Learning

[1]P. R. Asha, [2]M. S. Vijaya

[1]*Research Scholar, PSGR Krishnammal College for Women, Coimbatore*

[2]*Associate Professor, PSGR Krishnammal College for Women, Coimbatore*
[1]*ashamscsoft@gmail.com,* [2]*msvijaya@psgrkcw.ac.in*

Abstract

Remedy of inherited disorder like spinocerebellar ataxia (SCA) is a challenge and a necessary task in biomedical research. There are number of approaches available for affinity prediction through various scores and features in a standard computational framework hence it is significant to depict the binding affinity for drug identification. Affinity prediction is incredibly significant for drug discovery and it involves numerous steps like active site identification and docking. Identification of active site is obligatory for proteins to interact either with ligands or protein. The main focus of this work is to utilize the ensemble learning methods to build model for affinity prediction through 3d protein structures and interactive properties of that structures which can be used for predicting binding affinity. Protein-protein interaction is performed and the binding affinity is calculated from the interacted complex. Features like physio-chemical properties, energy calculations, interfacial and non-interfacial properties are extracted from the interacted complexes to construct enhanced predictions. Ensemble learning scheme is meta algorithms that coalese numerous machine learning techniques into one predictive model in order to lessen variance, bias, or improve predictions. The Random forest regressor in forest of randomized trees performs better by combining many algorithms. Experiments discovered the dominance of random forest regressor in forest of randomized trees when compared to other ensemble learning methods.

Keywords: *binding affinity prediction, rigid-docking, ensemble learning, protein-protein interaction*

1. Introduction

Spinocerebellar ataxia (SCA) is a traditional chaos reveals the declination in the brain and spinal cord. Change in the gene of SCA causes different types of mutations [1]. Mutation fabricates differentiation in the structure of protein. The types of spinocerebellar ataxia that is the origin of repeat mutation are SCA type1, SCA type2, SCA type3, SCA type6, SCA type7, SCA type8 and SCA type10 [2].

Protein-protein interactions is significant in numerous aspects of the structural and purposeful organization of the cell, and their elucidation is crucial for a better understanding of processes like metabolic management, signal transduction, and factor regulation. There are several protein-protein interaction databases and repositories [3].

Rigid tying up is performed for protein-protein interactions. Many ways supported tying up to review supermolecule complexes have conjointly been well developed over the past few years. Most of those approaches don't seem to be driven by experimental knowledge however it supports a mixture of energetics and form complementarity [4].

The affinity of a compound is significant where conformational changes occur due to binding. When structural changes occur the behavior of the protein also gets changed. Durable unit force of attraction ends up in high bonding affinity substance binding, whereas the substance binding of low-affinity involves lower and weak unit force between ligands and their receptors.

ISSN: 2005-4297 IJCA

Many research works are carried with the databases available for protein-protein interaction. In this work, it is projected to generate a dataset from where the binding affinity is predicted. Totally 626 protein structures interacted and affinity is predicted from the interacted complexes. Some of the literature studies were reviewed and described below:

Tammy Man-Kuang Cheng et al., planned a scheme of protein-protein interaction. The correct grading of rigid-body moorage orientations represents one among the key difficulties in protein–protein moorage prediction. They explored a method referred to as pyDock for rigid moorage. it's supported Coulombic physical science with distance dependent insulator constant, and implicit desolvation energy with atomic association parameters antecedently adjusted for rigid-body protein–protein moorage. the tactic is in a position to discover a near-native resolution from twelve,000 moorage poses and place it inside the a hundred lowest-energy moorage solutions in fifty six of the cases, in an exceedingly utterly unrestricted manner and with none different further data [5].

Solène Grosdidier and Juan Fernández-Recio, proposed a technique for distinguishing supermolecule hot spots. they need applied machine tying up approach known as normalized interface propensity values derived from rigid-body tying up with natural philosophy and desolvation evaluation for the prediction of interaction hot-spots. This parameter achieves upto eightieth positive prophetic price aside from existing strategies. The NIP values derived from rigid-body tying up will dependably determine a number of hot-spot residues whose contribution to the interaction arises from natural philosophy and desolvation effects. Our methodology will propose residues to guide experiments in complexes of biological or therapeutic interest, even in cases with no on the market 3D structure of the complex [6].

Pedro J. Ballester, John B. O. Mitchell proposed a technique for predicting binding affinity using computational approach. They proposed a novel scoring function called RF-score. Dataset which was used by them was PDBbind benchmark. Intermolecular interaction features are extracted and random forest is used for regression. They have obtained the root mean squared error as 1.52 by calculating RF Score. They also considered distance dependent features to be calculated in future for better prediction rate [7].

Jacob D. Durrant and J. Andrew McCammon proposed a neural network based model and nn scoring function is used for evaluation. Protein structures of x-ray crystal and magnetic protein formations were taken from PDB which has kd values. Binding affinity values considered for protein-ligand complexes from the database MOAD and PDBbind-cn [8].

From the background study it is proven that there is need for affinity prediction from protein-protein complexes. In existing work, affinity is used from the databases like PDBbind, MOAD and also the features like scoring functions are used. This requires for more work on binding affinity prediction from protein-protein interaction with the complexes that are not given in the database. In this work, interacting protein structures are used based on gene cards and the protein structures are taken from the curated database. Affinity is calculated from the interacted complex. Features like physio-chemical properties and energy calculations helps in making accurate affinity predictions than the scoring functions.

2. Materials and Methods

This work explores ensemble learning methods that are capable of combining several algorithms under a single roof and thereby predicting binding affinities. The proposed architecture is defined according to the ensemble methods depicted in Fig 1. Features are extracted from the complex, where the model works with hyperparameters optimization to provide better prediction.

In a rare genetic disorder like SCA, a mutation in the genes causes the change in the part of brain and spinal cord. SCA gene is mostly affected by repeat mutation, affinity

prediction helps in drug development and chemists can find out the curable medicine for rare genetic disorders. The work is alienated into the following stages: dataset creation, model building and evaluation.

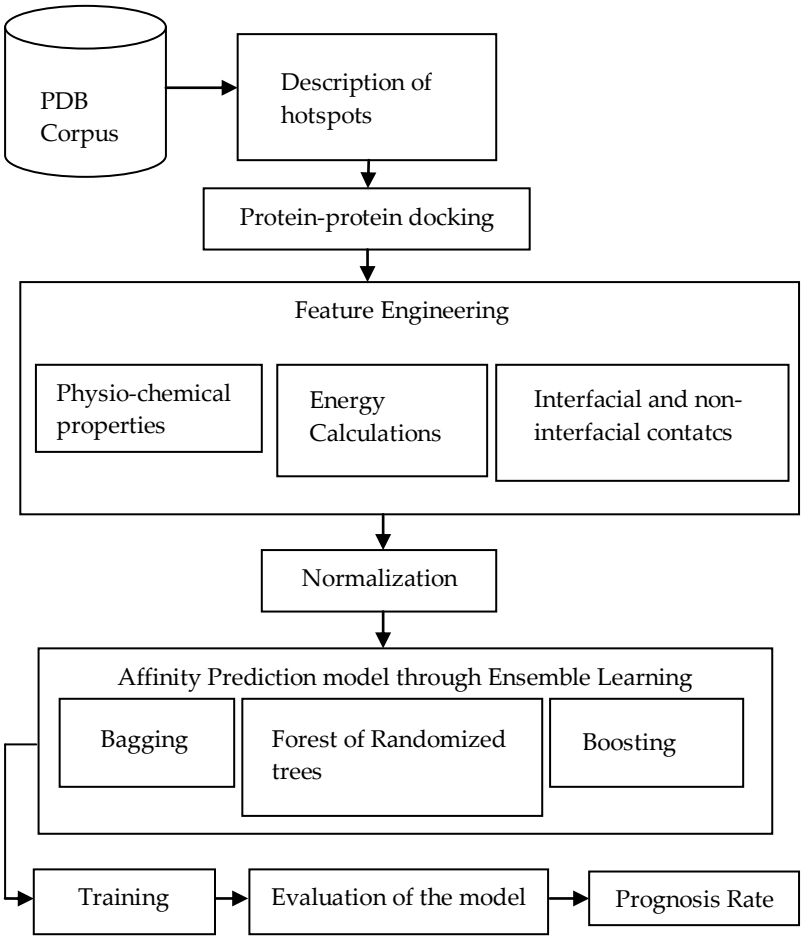

Fig 1. System Architecture

2.1 Corpus Preparation

In this system data's are collected from PDB corpus. Five types of ataxia namely sca1, sca2, sca3, sca6 and sca10 which commonly occur due to repeat mutations, are chosen. Proteins corresponding to those five types of ataxia are ataxia-1, ataxia-2, ataxia-3, cacna1a, ataxin-10 respectively. Protein interaction profile is referred from literatures and gene cards. Five types of protein have sixteen structures. Each protein has number of structures and it is interacted with proteins listed in the interaction profile. Interaction profile for each protein is given in Table 1. Totally 313 complexes are interacted for binding affinity prediction.

2.2 Identification of Hotspots

Hotspot is very important because the drug or protein that needs to bind in that hotspot, either to inhibit or reduce the growth of the disease. Hotspot or active site is very essential for docking. Hotspot is not necessary for flexible docking, but for rigid docking active site is necessary for both the proteins that are going to interact. Hotspots are identified by passing each protein's 3d structure into convolutional neural networks [9]. Threshold value is set as 0.5 and the protein which obtains below 0.5 is considered to be very low for drug binding. The protein which obtains 0.9 is best for consideration of binding and it is best suited for interaction. The protein which obtains more than 0.5 also can be considered. Hotspot for hotspot 1j46 protein is obtained by passing the protein

structure in neural network and it is shown in Fig 2.

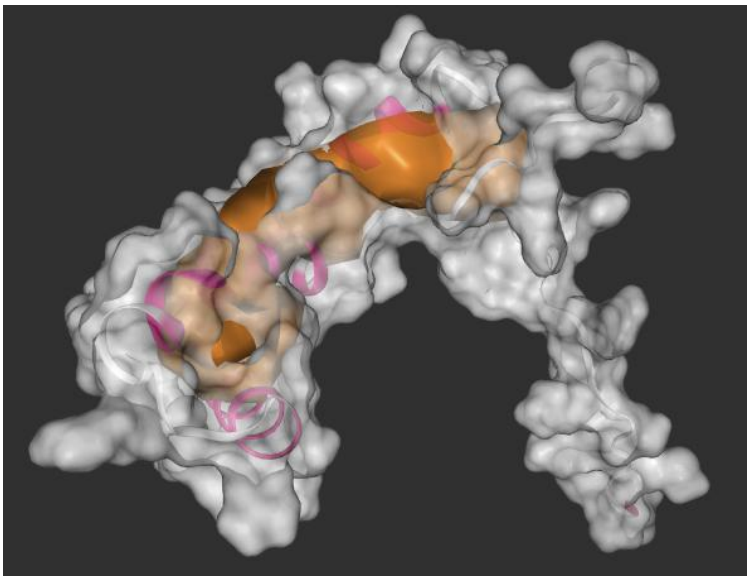

Fig 2. Hotspot identification of 1j46

2.3 Protein-protein Docking

Protein-protein interaction helps to understand physical and functional properties between molecules. It is vital because it helps biologists to determine protein's purpose and actions. It helps in predicting the unknown function involved in the protein. Interaction helps to exemplify protein compound and conduit [3]. There are many databases available for molecular interaction like STRING, IntACT, BioGRID etc [10]. Once hotspot is known, protein-protein interaction is performed. Flexible docking in protein-protein interaction is not possible, because of spatial transformation during docking. So rigid docking is performed, where the molecular structures are rigid that cannot their spatial shape during docking. Affinity prediction is calculated from the interacted complex. In this research work, initial proteins for five types of SCA are analyzed from gene cards. Sixteen structures are analyzed for five types of spinocerebellar ataxia. Each type of protein has unique set of proteins to be interacted and that information is available in Table 1. Proteins are interacted in haddock software. Rigid docking is performed by specifying amino acid position obtained through hotspot. In haddock software the proteins are interacted by specifying the hotspot in each protein. The interacted protein comes as a result of many clusters. Cluster is chosen based on lowest intermolecular energies [4].

Table 1. Interaction profile

Protein	Interacted Proteins
Ataxin-1	4j2l,4j2j,2m41,2gzk,1j46,1yqb,2jy6,2knz,4xos,4kdi,2pjh,1s3s,5ftn,5ftj, 5c19,3cfo,1u8f,2xxn,2f1x,2f1z,1nbf,5fwi,2kbr, 5jtv,4pyz,3u3o,4wpi,4y0c,3bzh,1y61
Ataxin-2	31py,2cqb,2r99,4pjo,3jcr,5mf9,1d3b,1n54,1h2t, 1h6k,1h2v,1n52,3p8b,2ckk,3fe2,4pxa,4lk2, 2i4i,4kbg,4kbf,3kx2,1n52,1cbj,2k8g,5ifn,1cvj,4f02,2xa6,5elt,5vl3,2bl5,3qhe
Ataxin-3	5ijo,4zol,4tv9,5fnv,5iy4,3vht,4kdi,2pjh,1s3s,5ftn,5ftj,5c19,3cfo,4v3l, 3u3o,4ksl,1gjz,5gjq,3b08,3low,2w9n,5b83,2znv, 3zn2,2qho,5gjq,5hpl,5koy,4k2x,4uq5,3o65,4xkh,2kl2,2mkg,4wth,4kbq, 3q4a,2c2l,2oxq,1p1a,1oel,1ify,2f4m,2qsf,1dvo0, 1iyf,5c1z,4inf,2jm0,4p50,2n7k,2brf,3zvn,3zvl,4rck,1jey,1jeq,1jjr,1e17,2k86

CACNA1A	4l9m,2vrw,4l9u,5cm8,1xd2,5kbt,1nvv,2yuu,4dex,3dvk,ebxl,3bxk,3dvj, 2ws7,3w14,3w11,1g7a,4oga,1jk8,2kqp,1toc,4y19, 4qsz,2w44,1b9y,1m56,5kd0,4q5q,1aqg,3mpx,1xd4,4f7z, 3c5h,3h5h,2ee5,3ah8,2bcj,3pvu,2rmk,5hzh,1x86,5c2k, 3cx8,3ab3,1zca,3uzs
Ataxin-10	2bcj,3uzs,1xhm,3ny8,3a8y,1xqs,1yuw,4wv7,4po2,3lof,1hx1,3c7n,1ckr,4kbq,2p32

2.4 Feature Engineering

Features like energy calculations, physio-chemical properties and interfacial contacts are extracted from the complex which obtained from interacted protein. Energy calculations are calculated from haddock software. Physio-chemical properties are calculated using R. Number of features involved in physical and chemical properties like amino acid composition, molecular weight, number of amino acids, theoretical PI, aliphatic index, positively charged and negatively charged etc [11]. Interfacial contacts, non-interacting surfaces and binding affinity constants are calculated using prodigy software [12]. Energy calculations, physio-chemical properties, interfacial and non-interfacial properties are measured as independent variables and Binding affinity is considered as response variable. There are 313 instances and 56 attributes. Features and their description are listed in Table 2.

Table 2. Feature and its Description

Features	Description
Haddock Score	The haddock score is performed according to the weighted sum of the following terms: van der Waals intermolecular energy.
Cluster size	Specifies the size of cluster
RMSD	Calculated as pairwise matrix for lowest energy as the structure
Desolvation energy	It is the static and/or van der Waals energy and measures interaction lose between substance and compound
Van der waals energy	Term used to define the attraction of intermolecular forces between molecules
Electrostatic energy	It is long term interaction that occurs between charged atoms of interacting proteins
Z-score	The z-score represents the standard deviations of the HADDOCK score
Buried surface area	Predict different measures of flexibility
Binding affinity	Strength of attraction between molecule and ligand
Dissociation constant	Ratio of dissociated ions to original acid
Physio-chemical properties	Includes physical and chemical properties of a protein
Interfacial contacts	Calculate number of interface residue pair wise contacts for each complex
NIS properties	It includes percentage of polar, apolar and charged residues

ISSN: 2005-4297 IJCA

2.5 Dataset

Response variable is binding affinity and other variables are independent variables. Response variable is extracted from haddock software. Independent variables like energy calculations are extracted from haddock software and other variables like physio-chemical properties are extracted using R coding. Once features are extracted from the dataset, they are normalized. In this research work pre-processing is done using min-max normalization. Normalization is necessary because to reduce and eliminate data redundancy and it is the process of organizing data [13].

2.6 Model Building

Ensemble models are built by passing the feature vectors. In this work ensemble models like bagging, boosting and randomized trees are used. Each model is built by optimizing the hyper parameters. Parameters of ensemble learning models are n_estimators, max_depths and max_features.

Parameter tuning leads to better prediction. In this work, three parameters are used namely n_estimators, max_features and max_depth. Number of trees in the forest is described as n_estimators and number of features measured for dividing a node as max_features. Number of stages in decision tree is described in max_depth. Default parameter settings are used in ensemble models which gives better prediction.

3. Experiments and Results

Ensemble learning is implemented with three techniques for building the binding affinity prediction models in scikit learn and coded in python. The advantage of protein-protein interaction in this experiment is to attain the stable complex through which binding affinity is determined. The training dataset with 313 instances related to six categories of spinocerebellar ataxia, i.e., spinocerebellar ataxia type1, spinocerebellar ataxia type2, spinocerebellar type3, cacna1a, spinocerebellar ataxia type 8 and spinocerebellar ataxia type10 has been used to train the model.

Evaluation measures considered for this work are Explained_variance score, mean_squared error, RMSD (Root mean squared error), MAE (Mean absolute error), MEAE (Median absolute error) and R2_score. These are the regression metrics considered most vital for evaluating the model. Bagging uses bootstrap sampling to get the information subsets for coaching the bottom learners. For aggregating the outputs of base learners, sacking uses vote for classification and averaging for regression.

Table 3. Performance criteria of bagging

Explained_variance score	0.70
MSE	0.32
RMSD	0.57
MAE	0.37
MEAE	0.23
R2_score	0.70

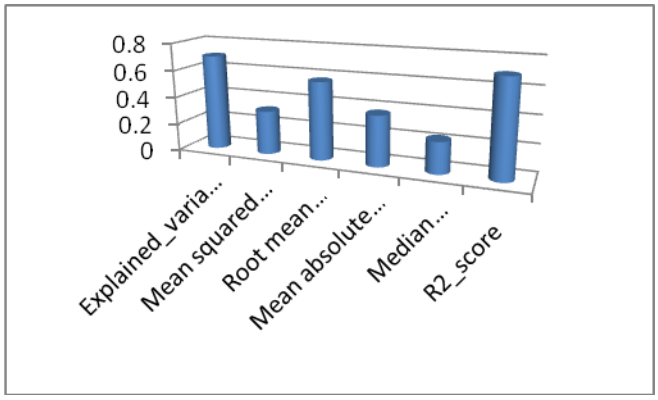

Fig 3. Evaluation measures of bagging

Bagging regressor achieves the explained score of 0.70 and the error rate as 0.32 and the results of bagging regressor is listed in Table 3 and evaluation measures are given in Fig 3.

In random forests, every tree within the ensemble is constructed from a sample drawn with replacement (i.e. a bootstrap sample) from the coaching set. Additionally, rather than victimisation all the options, a random set of options is chosen, any randomizing the tree. As a result, the bias of the forest will increase slightly, however thanks to the averaging of less correlate trees, its variance decreases, leading to Associate in Nursing overall higher model.

Table 4. Performance criteria for forest of randomized trees

Regression models	Explained_variance score	R2_score	MSE	RMSE	MAE	MEAE
Random forest regressor	0.90	0.90	0.12	0.34	0.15	0.6
Extremely randomized trees(decision tree)	0.85	0.85	0.20	0.20	0.10	0.02
Extremely randomized trees(extra tree)	0.86	0.86	0.2	0.44	0.35	0.15
Extremely randomized trees(random forest tree)	0.84	0.84	0.2	0.44	0.35	0.15

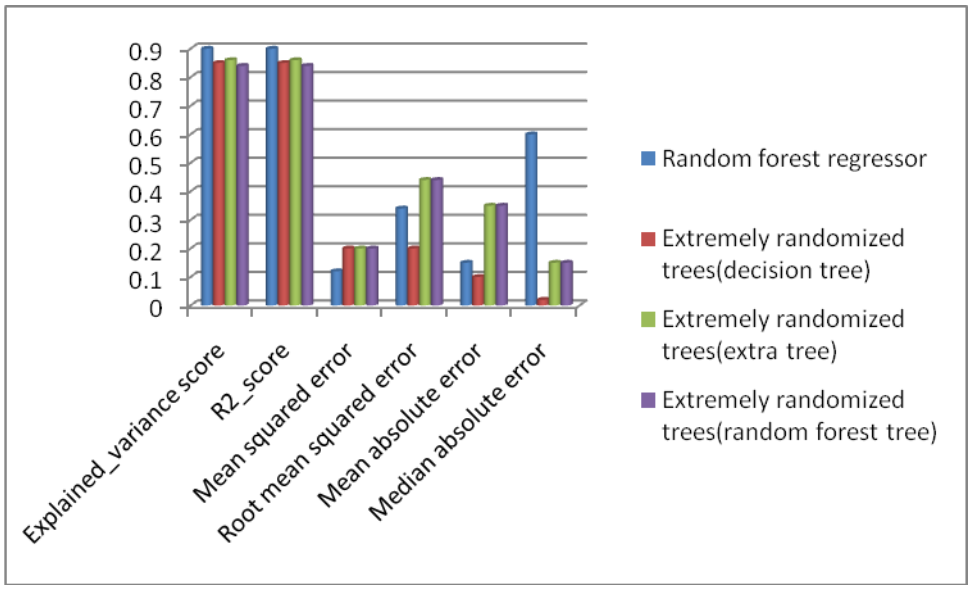

Fig 4. Evaluation measures for forest of randomized trees

Among the forest of randomized trees, random forest regressor produces 0.90 of explained_variance_score and the mean squared error is 0.12. The error rate is minimized in random forest regressor and the variance score is very high when compared with other regressor in forest of randomized trees. Results of random forest regressor are given in Table 4 and evaluation measures are given in Fig 4.

Boosting is to suit a sequence of weak learners− models that unit exclusively slightly on top of random plan, like very little decision trees− to weighted versions of the knowledge. Extra weight is given to examples that were misclassified by earlier rounds. The predictions unit then combined through a weighted majority vote (classification) or a weighted add (regression) to provide the last word prediction. The principal distinction between boosting and thus the committee ways that, like textile, is that base learners unit trained in sequence on a weighted version of the knowledge. Gradient boosting gives the variance score as 0.86 and the error as 0.2. Gradient boosting performs better when compared to Adaboost regressor and the results of boosting are given in Table 5 and Figure 5.

Table 5. Performance criteria for boosting

Regression models	Explained_variance score	R2_score	MSE	RMSE	MAE	MEAE
Adaboost	0.82	0.81	0.2	0.44	0.15	0.06
Gradient boosting	0.86	0.86	0.2	0.44	0.35	0.15

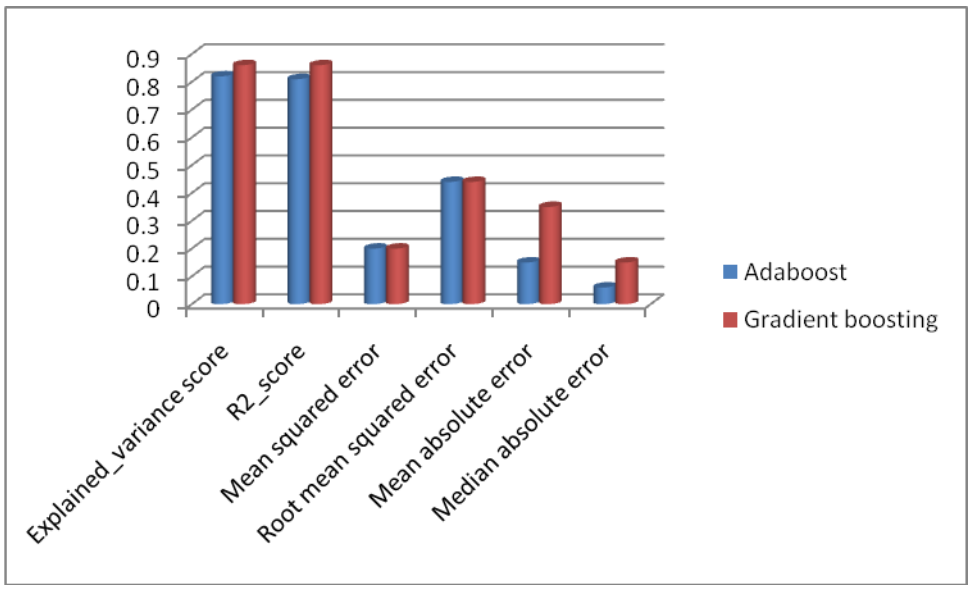

Fig 5. Evaluation measures of boosting

The performance of the projected work is contrast with the ensemble learning random forest regressor. The ensemble models with bagging, boosting and forest of randomized trees are calculated with various evaluation metrics and the results prove that random forest regressor in forest of randomized trees outperformed other ensemble models. It is evident that the random forest regressor in forest of randomized trees achieved comparable performance than the other ensemble learning models.

Among all ensemble learning models random forest regressor in forest of randomized trees performs better with variance score of 0.90 and with the error rate of 0.12. Table 6. depicts the comparison of random forest regressor and other ensemble learning methods and Fig 6. shows the evaluation measures of the ensemble learning models like random forest regressor, bagging and gradient boosting.

ISSN: 2005-4297 IJCA

Table 6. Comparison of Ensemble models

Regression models	Explained_variance score	R2_score	MSE	RMSE	MAE	MEAE
Random forest regressor	0.90	0.90	0.12	0.34	0.15	0.6
Gradient boosting	0.86	0.86	0.2	0.44	0.35	0.15
Bagging	0.70	0.70	0.32	0.57	0.37	0.23

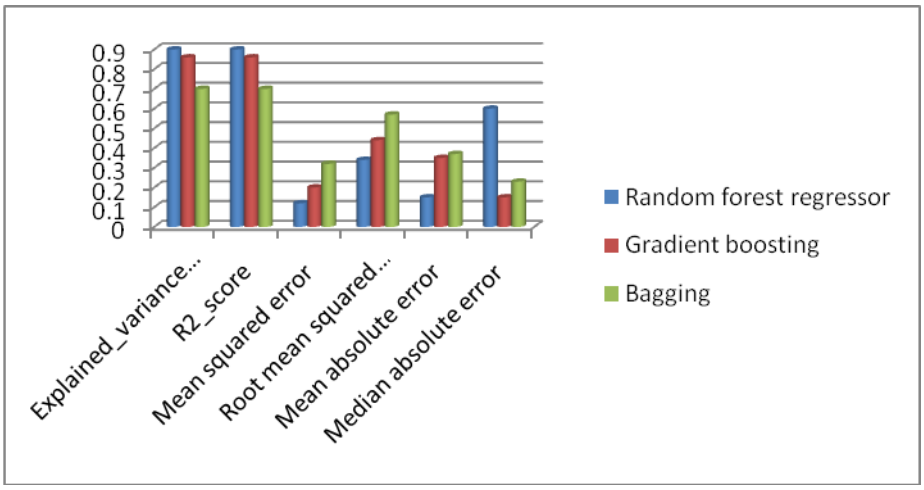

Fig 6. Comparison of Random forest regressor with bagging and gradient boosting

4. Discussion and Findings

From the above experiments, it is evidently tacit that the energy calculations like desolvation energy, electrostatic energy and vanderwaals energy, scores from haddock and also physio-chemical properties projected in this work facilitates in discerning the affinity prediction and civilizing the recital of the model. It is inveterate that the model erected using forest of randomized trees conquer intensified results for affinity prediction from protein-protein interaction complexes as fortunate results are achieved. This work attains prominent explained_variance score. The mean squared error is diminished where the consistency of the system is enhanced.

Ensemble learning executed using python library of scikit learn alters data to array format. It is apparent that these methods are apposite in prediction of binding affinity of spinocerebellar ataxia and also with features like physio-chemical properties, energy calculations, interfacial and non-interfacial properties binding affinity can be predicted for any type of rare genetic disorder. From this work it is evident that affinity prediction using forest of randomized trees attains the better results with variance score of 0.90 and the error rate of 0.12, than the other models that are built using ensemble models like bagging and boosting.

5. Conclusion

In this paper, Ensemble models are proposed with bagging, boosting and forest of randomized trees to predict affinity of SCA using PPI. The various hyperparameters of the ensemble models namely max_depths, max_features and n_estimators were investigated. The ensemble model with random forest regressor proved its strength at default parameter settings. Experimental results exhibit that the ensemble model with random forest regressor achieved the better prediction rate. Moreover, the recognition result is investigated by comparing the bagging, boosting and random forest regressor. It is verified that random forest regressor in forest of randomized trees showed better results outperforming the other ensemble models. This comparison of results offers a baseline for further research, and it is expected that it can provide a better result with more number of feature vectors and by using varying algorithms.

ISSN: 2005-4297 IJCA

References

[1] Thomas C. Weiss., Ataxia Spinocerebellar: SCA Facts and Information, 2010.
[2] Thomas D Bird, MD., Hereditary Ataxia Overview, March 3, 2016.
[3] Javier De Las Rivas and Celia Fontanillo, Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks, PLOS Computational biology, june 24, 2010.
[4] Cyril Dominguez, Rolf Bolens and Alexandre M. J. J. Bonvin, HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information, Journal of American Chemical Society, January 21, 2003, pp 1731-1737
[5] Tammy Man-Kuang Cheng, Tom L. Blundell, Juan Fernandez-Recio, pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking, Proteins, april 19, 2007.
[6] Solène Grosdidier and Juan Fernández-Recio, Identification of hot-spot residues in protein-protein interactions by computational docking, BMC Bioinformatics, October 21, 2008
[7] Pedro J. Ballester, John B. O. Mitchell A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking, Bioinformatics, Vol 26, Issue 9, pp 1169-1175, May 2010.
[8] Jacob D. Durrant and J. Andrew McCammon, NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein−Ligand Complexes, Journal of chemical information and modelling, 2010 Oct 25; 50(10): 1865–18
[9] Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G, DeepSite: protein-binding site predictor using 3D-convolutional neural networks, Bioinformatics, October 1, 3036-3042
[10] Damian Szklarczyk, Lars Juhl Jensen, Protein-Protein Interaction Databases, International Journal of Oral Science, vol 1278, pp 39-56
[11] Ingemar Björk, Bengt-Åke Petersson, John Sjöquist, Some Physicochemical Properties of Protein A from Staphylococcus aureus, European journal of biochemistry banner, September 1972
[12] Anna Vangone, Alexandre MJJ Bonvin, Contacts-based prediction of binding affinity in protein–protein complexes, Structural biology and molecular biophysics, July 20, 2015
[13] Yogendra kumar jain, santhosh kumar bhandare, Min Max Normalization Based Data Perturbation Method for Privacy Protection, Google scholar
[14] Dietterichl, Thomas G. (2002). Ensemble learning. In M. Arbib (ed.), The Handbook of Brain Theory and Neural Networks. MIT Press. pp. 405--40

Authors

P. R. Asha, Ph. D Research Scholar,

Completed M.Sc[SE] in Bannari Amman Institute, M.phil in krishnammal college for women and currently pursuing my Ph.D in krishnammal college for women.

Areas of Interest: Computational biology, data mining and statistics. I'm interested in finding new drugs for the disease which has no drug to cure the disease. Currently pursuing doctorate degree in the field of research titled predicting binding affinity for spinocerebellar ataxia. Four papers has been presented and published in conference proceedings. Papers published in **JARDCS (scopus indexed)** journal entitled "Deep Neural Networks for Affinity Prediction of Spinocerebellar Ataxia Using Protein Structures", **IJEAT (scopus indexed)** Journal entitled "Affinity Prediction of Spinocerebellar Ataxia Using Protein-Ligand and Protein-Protein

Interactions with Functional Deep Learning", **IJAST (International Journal of Advanced Science and Technology)** Journal entitled "Affinity Prediction of Spinocerebellar Ataxia using Protein-protein Interactions and Deep Neural Network with User-Defined Layer".

M. S. Vijaya, Associate Professor & Head

Completed masters in PSG college of Technology and did Ph.D in Amirta university Coimbatore and currently working as associate professor in krishnammal college for women. She is a head of computer science department.

Area of Specialization: Data Mining, Machine Learning, Support Vector Machine, Pattern Recognition, Bioinformatics. She has guided many M.phil research scholars and five Ph.d research scholars had completed under her guidance.

ISSN: 2005-4297 IJCA