iJRASET

International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VI Month of publication: June 2025

DOI: https://doi.org/10.22214/ijraset.2025.72272

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Integration of Network Biology and Machine Learning for Identification and Prioritisation of Breast Cancer Targets

B. Spandana[1], B. Sindhuja[2], Ansu Sinsodhiya[3], A. Sai Tejaswini[4], Prof. M.Sumakanth[5]
[1, 2, 3, 4]Student, Department of Pharmaceutical Chemistry, RBVRR Women's College of Pharmacy, Osmania University, Hyderabad, India
[5]Principal,RBVRR Women's College of Pharmacy,Osmania University,Hyderabad,India

Abstract:** Breast cancer is a complex disease involving multiple genes and proteins. Identifying key proteins and their interactions is crucial for understanding the disease mechanisms and developing targeted therapies. This study employs a network-based approach to analyze protein-protein interaction (PPI) data related to breast cancer, utilizing the PageRank algorithm and random forest classifier. Breast cancer-related PPI 984data was obtained from the STRING database and processed using Python libraries such as pandas and networkx. Topological analysis was performed to identify central proteins based on degree, betweenness, closeness, and eigenvector centrality measures. The PageRank algorithm was applied to rank proteins by their importance in the network. A random forest classifier was trained using the PageRank scores and known cancer relevance labels to predict the cancer relevance of proteins. Additionally, molecular docking simulations were conducted using AutoDock Vina to evaluate the binding affinities of PARP inhibitors (Niraparib, Olaparib, Veliparib, and Rucaparib) to the PARP1 protein. The docking results were rescored using the DeltaVina RF scoring function, which combines the Vina scoring function with a random forest approach. The study identified key proteins involved in breast cancer, with the top-ranked proteins being ENSP00000418960, ENSP00000260947, and \ENSP00000278616. The random forest classifier achieved perfect accuracy in predicting cancer relevance based on PageRank scores. Molecular docking and rescoring revealed Niraparib and Veliparib as the most promising PARP inhibitors. This study demonstrates the utility of combining network analysis, machine learning, and molecular docking techniques to identify potential drug targets and evaluate drug candidates for breast cancer treatment.

Keywords: Breast cancer, DeltaVina, Network biology, Protein-Ligand Scoring

I. INTRODUCTION

The integration of computational algorithms into biological research has enabled the systematic exploration of complex biological networks and drug discovery processes. In this study, multiple algorithmic approaches are employed to prioritize key DNA repair proteins associated with cancer, evaluate their relevance using machine learning, and predict their druggability through molecular docking. The following algorithms are employed in the research.

A. Page Rank Algorithm

PageRank is a node ranking algorithm used to identify important nodes (e.g., genes or proteins) based on their LTW1vnetwork connectivity and influence. Rather than just counting the number of connections a node has (degree),PageRank considers the quality of those connections— giving higher importance to nodes connected to other highly ranked nodes.

The PageRank score $P(i)$of a node i is given by:

$$PR(i) = \frac{1-c}{n} + c \sum_{j \in M(i)} \frac{PR(i)}{L(j)}$$

Where:
N =total number of nodes in the network
C = damping factor (typically 0.85), representing the probability of continuing a random walk
(i) =set of nodes linking to node i
(j)number of outbound links from node j [1],[2]

Functions

1) Identifying key nodes

PageRank can pinpoint central genes or proteins in gene expression or protein-protein interaction (PPI) networks.

2) Functional Module Detection

PageRank identifies densely connected sub-networks(modules) that likely participate in shared biological functions and useful in discovering mechanisms behind gene regulation, signaling pathways, or disease mechanisms.

It has been successfully applied to:

- Prioritize candidate disease genes
- Identify regulatory modules
- Analyze signal transduction pathways
- Detect drug targetsin infectious diseases (e.g., ThyX in tuberculosis)[3],[4],[5],[6]

B. Random Forest Algorithm [RF]

A supervised learning technique for classification and regression problems .In order to generate forecasts, it builds several decision trees and combines their outputs. The RF algorithm gets around the drawbacks of individual decision trees, including low bias, large variation, and over-fitting.[7],[8],[9]

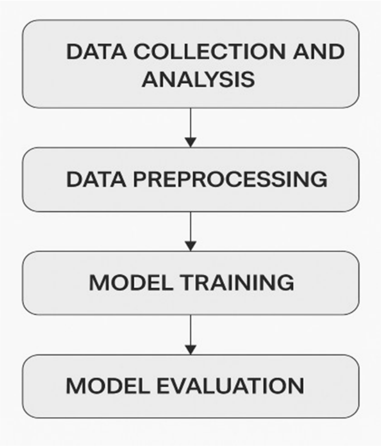


Fig.1:Steps involved in Random Forest Algorithm

C. Deltavina Rf Scoring Function

The Delta vina along with RF is a new protein-ligand scoring function .It employs a RF approach to parameterize corrections to the Auto Dock.The RF (random forest) score is a machine learning-based protein-ligand scoring function This allows it to combine the excellent docking power of Vina with the improved scoring accuracy from random forest.The Deltavina and RF scoring function developed in this work is important because it can achieve superior performance compared to traditional scoring functions in all power tests, including scoring, ranking, docking and screening power tests, for both the CASF-2013 and CASF-2007 benchmarks. [10],[11]

II. METHODOLOGY

The methodology for breast cancer-related target analysis can be given as follows:

Data collection involved retrieving protein-protein interaction data from the STRING database, focusing on breast cancer-related genes in humans. The data was downloaded in TSV format and processed using Python libraries such as pandas and networks. Key steps included data cleaning, handling missing values, and exploratory analysis of gene interactions.

Network analysis was performed using various topological metrics, including degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. These measures helped to identify influential proteins within the network. Visualization techniques were employed to represent centrality measures, aiding in the interpretation of results.

The PageRank algorithm was applied to the protein-protein interaction network to prioritize potential breast cancer targets. This involved creating a graph representation of the network, computing PageRank scores for each protein, and ranking them based on their importance within the network.

Finally, a machine learning approach using Random Forest was implemented to classify proteins as cancer-relevant or not. This involved preprocessing the data, combining PageRank scores with known cancer relevance labels, splitting the data into training and testing sets, and training the Random Forest model to predict cancer relevance based on network features.

Molecular Docking and Machine Learning Rescoring Workflow by DeltaVina RF scoring approach for PARP1 and PARP inhibitors

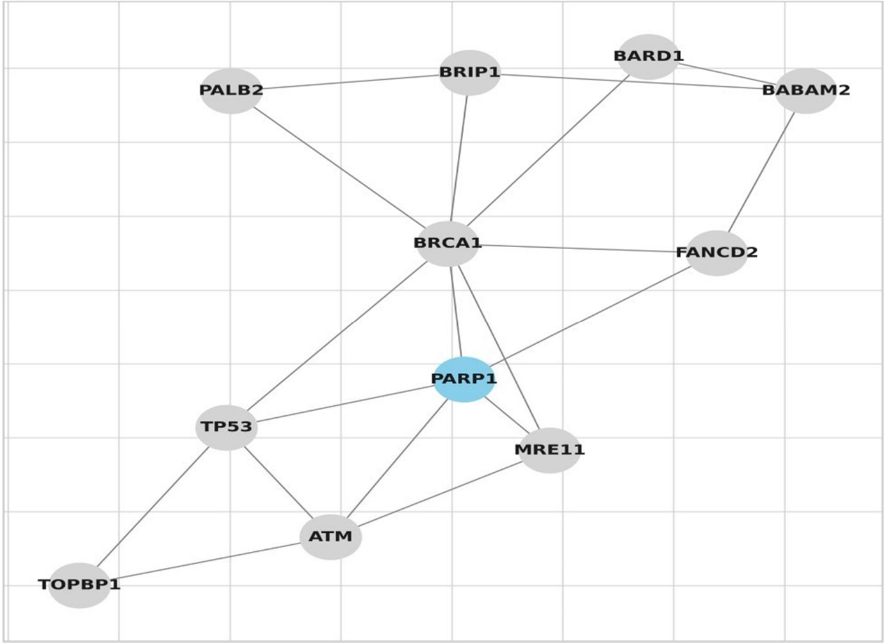


Fig.2: PARP1-Centric Protein–Protein Interaction Network Illustrating Its Connectivity with 10 Key Genes involved in breast cancer

The computational workflow for evaluating the binding affinities of PARP1 inhibitors through molecular docking, Random Forest (RF) score prediction, and deltaVina rescoring encompasses a structured, multi-stage process. Protein The three-dimensional crystal structure of human PARP1 (Poly (ADP-ribose) polymerase 1)(PDB ID:7KK4| pdb_00007kk4) was obtained from the Protein Data Bank, representing the catalytic domain essential for inhibitor binding.preparation involved the removal of crystallographic water molecules and co-crystallized ligands, the addition of hydrogen atoms, the assignment of partial atomic charges, and energy minimization to optimize the structural conformation for docking simulations. A set of clinically approved PARP1 inhibitors, including Olaparib, Rucaparib, Niraparib, and Veliparib, were selected for docking based on their therapeutic relevance, and their structures were geometry-optimized prior to molecular docking.

The binding site for docking was defined by identifying the catalytic active site of PARP1, followed by the construction of a grid box encompassing the relevant region to restrict docking simulations to the biologically significant domain. Molecular docking was subsequently performed using AutoDock or a comparable software package. For each inhibitor, multiple binding poses were generated, and initial scoring and ranking were conducted based on the software's internal scoring functions.

Post-docking analysis involved the extraction of key protein-ligand interaction features from the docked complexes, which were utilized as input to a pre-trained Random Forest (RF) regression model to predict binding affinities, producing the RF score for each pose. In parallel, deltaVina rescoring was performed by re-evaluating the docking poses to derive alternative binding affinity estimates, enabling a comparative assessment against the original docking scores.

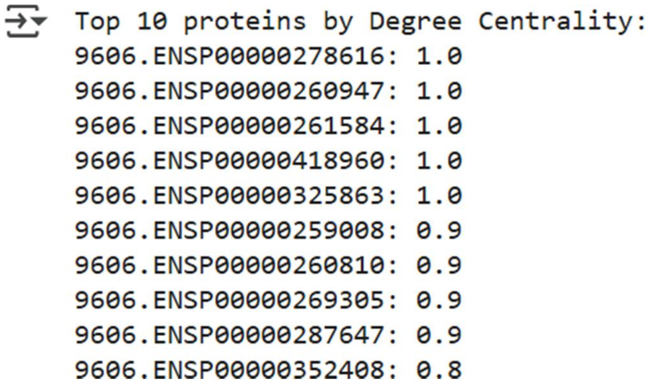
Finally, detailed visualization and interpretation were performed by analyzing the binding conformations of top- ranked inhibitors and comparing their autodock,Rf and Deltavina scores. Collectively, this integrated computational approach synergizes molecular docking with advanced machine learning-based scoring techniques to facilitate a more reliable and comprehensive evaluation of PARP1 inhibitors, thereby strengthening predictive drug discovery pipelines.

III. RESULTS AND DISCUSSION

This section presents the detailed analysis of the network biology-based prioritization of key DNA repair proteins implicated in breast cancer. A combination of centrality metrics, machine learning models and molecular docking simulations was employed to systematically evaluate the biological importance and druggability of the identified proteins and the results are follows:

A. *Node Centrality Scores*

1) *Degree Centrality*



```
Top 10 proteins by Degree Centrality:
9606.ENSP00000278616: 1.0
9606.ENSP00000260947: 1.0
9606.ENSP00000261584: 1.0
9606.ENSP00000418960: 1.0
9606.ENSP00000325863: 1.0
9606.ENSP00000259008: 0.9
9606.ENSP00000260810: 0.9
9606.ENSP00000269305: 0.9
9606.ENSP00000287647: 0.9
9606.ENSP00000352408: 0.8
```

Fig 3:Top 10 Proteins by Degree Centrality in the Protein–Protein Interaction Network

2) *Visualising Degree Centrality*

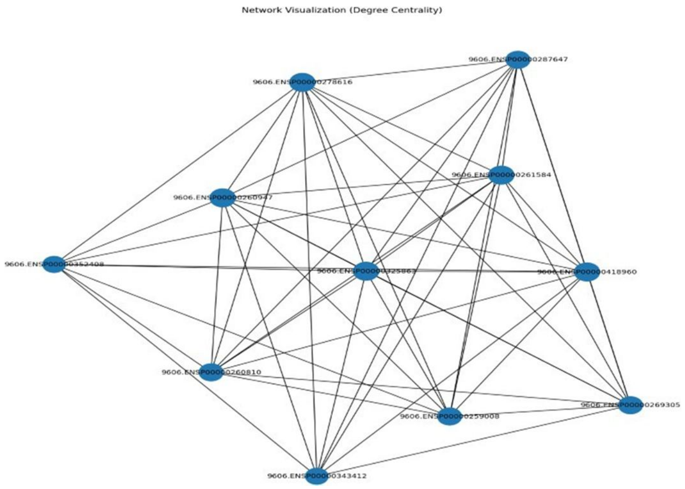
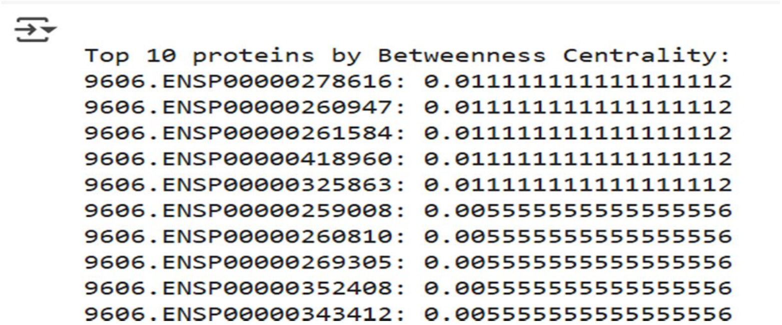


Fig.4:Visualisation of top 10 proteins by Degree Centrality in the Protein–Protein Interaction Network

B. *Betweenness Centrality*

Betweenness centrality identifies the most important nodes for communication in the network.



```
Top 10 proteins by Betweenness Centrality:
9606.ENSP00000278616: 0.011111111111111112
9606.ENSP00000260947: 0.011111111111111112
9606.ENSP00000261584: 0.011111111111111112
9606.ENSP00000418960: 0.011111111111111112
9606.ENSP00000325863: 0.011111111111111112
9606.ENSP00000259008: 0.005555555555555556
9606.ENSP00000260810: 0.005555555555555556
9606.ENSP00000269305: 0.005555555555555556
9606.ENSP00000352408: 0.005555555555555556
9606.ENSP00000343412: 0.005555555555555556
```

Fig.5:Top 10 Proteins by Betweenness Centrality in the Protein–Protein Interaction Network

Visualising Betweenness Centrality

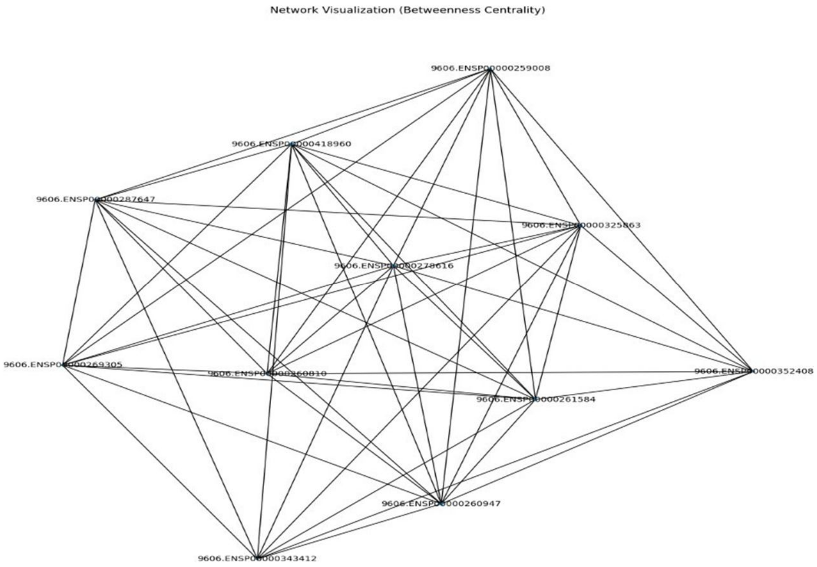
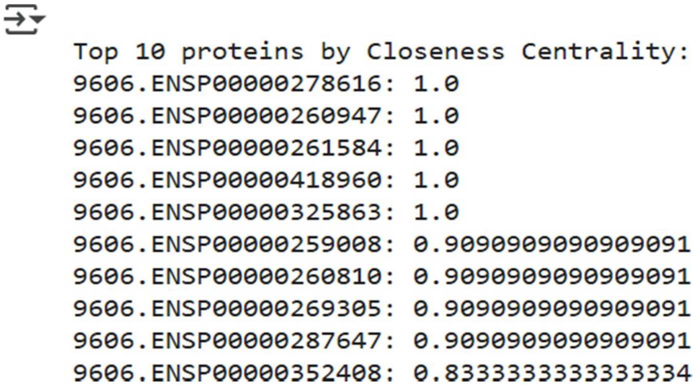


Fig.6:Top 10 Proteins by Betweenness Centrality in the Protein–Protein Interaction Network

C. *Closeness Centrality*

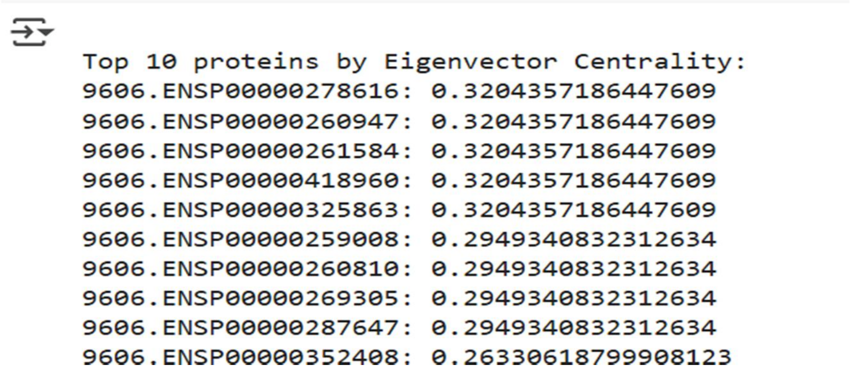


```
Top 10 proteins by Closeness Centrality:
9606.ENSP00000278616: 1.0
9606.ENSP00000260947: 1.0
9606.ENSP00000261584: 1.0
9606.ENSP00000418960: 1.0
9606.ENSP00000325863: 1.0
9606.ENSP00000259008: 0.9090909090909091
9606.ENSP00000260810: 0.9090909090909091
9606.ENSP00000269305: 0.9090909090909091
9606.ENSP00000287647: 0.9090909090909091
9606.ENSP00000352408: 0.8333333333333334
```

Fig.7:Top 10 Proteins by Closeness Centrality in the Protein–Protein Interaction Network

D. *Eigenvector Centrality*

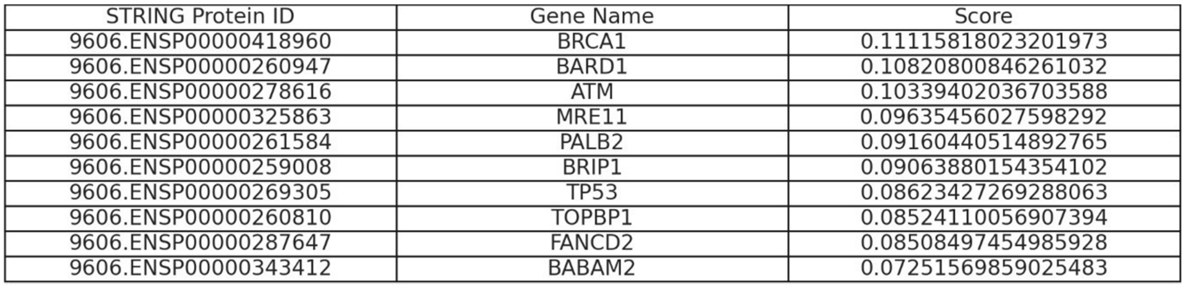
Eigenvector centrality takes into account not just the number of connections, but also the quality of those connections.



```
Top 10 proteins by Eigenvector Centrality:
9606.ENSP00000278616: 0.3204357186447609
9606.ENSP00000260947: 0.3204357186447609
9606.ENSP00000261584: 0.3204357186447609
9606.ENSP00000418960: 0.3204357186447609
9606.ENSP00000325863: 0.3204357186447609
9606.ENSP00000259008: 0.2949340832312634
9606.ENSP00000260810: 0.2949340832312634
9606.ENSP00000269305: 0.2949340832312634
9606.ENSP00000287647: 0.2949340832312634
9606.ENSP00000352408: 0.26330618799908123
```

Fig.8: Top 10 Proteins by Eigenvector Centrality in the Protein–Protein Interaction Network

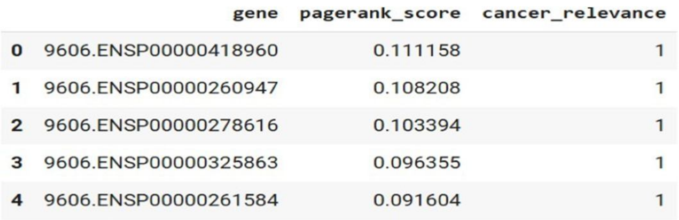
- PageRank Algorithm Scores



STRING Protein ID	Gene Name	Score
9606.ENSP00000418960	BRCA1	0.11115818023201973
9606.ENSP00000260947	BARD1	0.10820800846261032
9606.ENSP00000278616	ATM	0.10339402036703588
9606.ENSP00000325863	MRE11	0.09635456027598292
9606.ENSP00000261584	PALB2	0.09160440514892765
9606.ENSP00000259008	BRIP1	0.09063880154354102
9606.ENSP00000269305	TP53	0.08623427269288063
9606.ENSP00000260810	TOPBP1	0.08524110056907394
9606.ENSP00000287647	FANCD2	0.08508497454985928
9606.ENSP00000343412	BABAM2	0.07251569859025483

Fig.9: PageRank-Based Prioritization of DNA Repair Proteins and Their Cancer Relevance

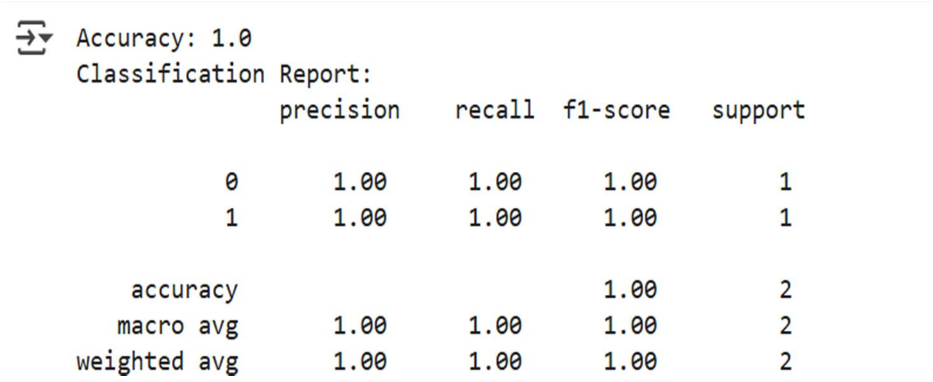
- Cancer Relevance Scores by Random Forest Algorithm



	gene	pagerank_score	cancer_relevance
0	9606.ENSP00000418960	0.111158	1
1	9606.ENSP00000260947	0.108208	1
2	9606.ENSP00000278616	0.103394	1
3	9606.ENSP00000325863	0.096355	1
4	9606.ENSP00000261584	0.091604	1

Fig.10: Top Five Genes Identified by PageRank Algorithm with Confirmed Cancer Relevance by Random Forest algorithm

- Accuracy and Classification Report

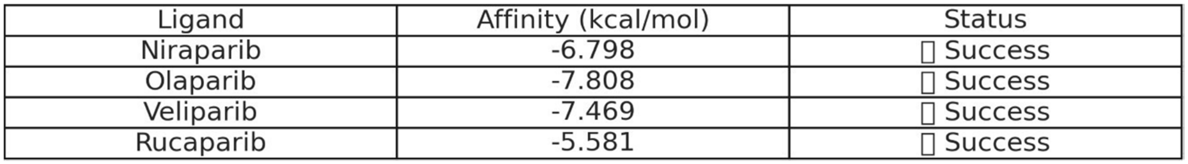


```
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       1.00      1.00      1.00         1

    accuracy                           1.00         2
   macro avg       1.00      1.00      1.00         2
weighted avg       1.00      1.00      1.00         2
```

Fig.11: Classification Report for Random Forest-Based Cancer Relevance Prediction

The Random Forest model performed perfectly on this small dataset, with an accuracy of 1.0 (100%) and all the classification metrics (precision, recall, and F1-score) being 1.0 for both classes.

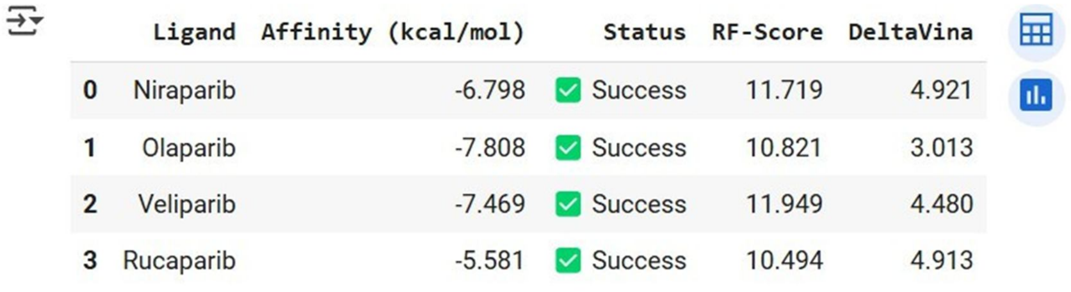
- Autodock Scores



Ligand	Affinity (kcal/mol)	Status
Niraparib	-6.798	□ Success
Olaparib	-7.808	□ Success
Veliparib	-7.469	□ Success
Rucaparib	-5.581	□ Success

Fig.12: Binding Affinity of PARP Inhibitors Predicted by AutoDock Vina

- Rf Scores and DeltaVina Scores



	Ligand	Affinity (kcal/mol)	Status	RF-Score	DeltaVina
0	Niraparib	-6.798	☑ Success	11.719	4.921
1	Olaparib	-7.808	☑ Success	10.821	3.013
2	Veliparib	-7.469	☑ Success	11.949	4.480
3	Rucaparib	-5.581	☑ Success	10.494	4.913

Fig.13: Comparative Scoring of PARP Inhibitors Using AutoDock Vina, RF-Score, and DeltaVina

- Visualisation of Autodock,Rf and Deltavina score:

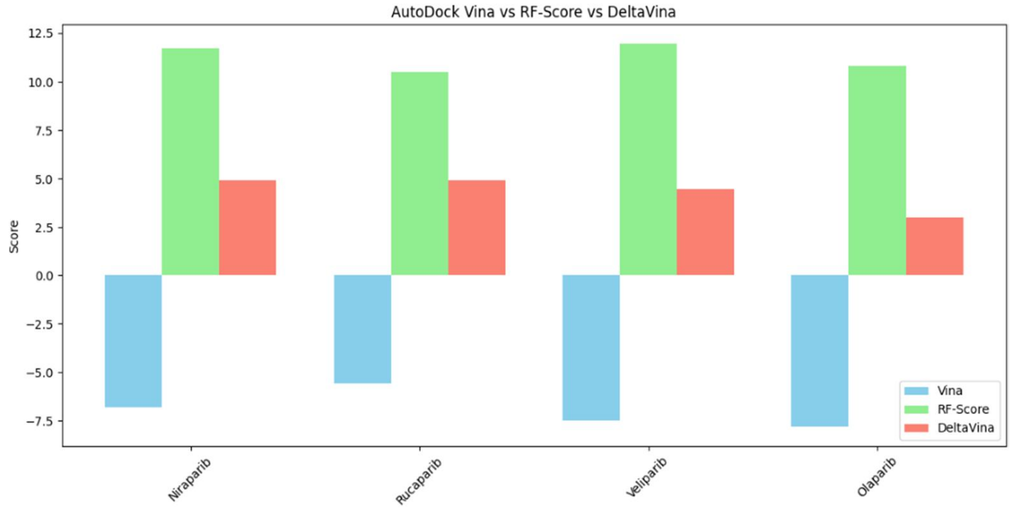


Fig.14: Visual Comparison of Binding Scores from AutoDock Vina, RF-Score, and DeltaVina for Four PARP Inhibitors

- Evaluation of Results

Table 1: Evaluation of Molecular Docking and Machine Learning Scoring for PARP1 Inhibitors

Inhibitor	Vina Score	RF Score	DeltaVina Score	Key Observations	Conclusion
Niraparib	-6.798	11.719	4.921	Strong ML and good docking balance	Highly promising candidate with high predicted binding affinity
Olaparib	-7.808	10.821	3.013	Best Vina score but moderate ML support	Strong docking binder; ML flags potential limitations
Veliparib	-7.469	11.949	4.480	Highest RF score, strong docking	Standout candidate with excellent ML features
Rucaparib	-5.581	10.494	4.913	Poor docking but good ML score	Chemically promising despite weak docking pose

Among the four PARP1 inhibitors evaluated, Niraparib and Veliparib showed the most promising profiles. Niraparib demonstrated a balanced performance with a Vina score of -6.798, a high RF score of 11.719, and the highest deltaVina score of 4.921, indicating both favorable binding energy and strong machine learning-based potential. Veliparib achieved the highest RF score (11.949) and a strong Vina score (-7.469), supported by a deltaVina score of 4.480, suggesting excellent chemical features and docking compatibility.Olaparib recorded the best Vina score (-7.808), indicating strong raw binding affinity. However, its lower RF score (10.821) and a significantly reduced deltaVina score (3.013) suggest possible structural limitations, despite its clinical relevance. Rucaparib had the weakest Vina score (-5.581) but a respectable RF score (10.494) and a high deltaVina score (4.913).The strong RF and deltaVina scores indicate that, despite a poorer docking pose, Rucaparib possesses favorable chemical and structural features such as key atom types and interaction patterns that are recognized by the machine learning model.

In summary, Niraparib and Veliparib appear as the most compelling inhibitors, combining favorable docking energies and machine learning-based affinity predictions. Olaparib remains strong due to its clinical validation, while Rucaparib shows potential that may not be fully captured by traditional scoring alone.

IV. CONCLUSION

The research integrated network biology algorithms and molecular docking techniques to identify and evaluate potential drug targets for breast cancer. The study applied the PageRank algorithm to a protein-protein interaction network derived from breast cancer-related genes, identifying top-ranked proteins as potential drug targets. A Random Forest classifier trained on PageRank scores achieved 100% accuracy in predicting cancer relevance of proteins. Molecular docking using AutoDock Vina was performed with four PARP inhibitors (Niraparib, Olaparib, Veliparib, Rucaparib) on the PARP1 receptor(7KK4 | pdb_00007kk4) revealing favorable binding for all ligands, with Olaparib showing the best binding affinity. A Random Forest regression model (RF-Score) was applied to rescore the docked ligands, and DeltaVina scores were calculated by combining Vina and RF-Score results. Niraparib emerged as the most promising candidate with a high RF-Score and the highest DeltaVina score, while Olaparib showed a lower DeltaVina score when compared to traditional docking results.This comprehensive approach combining network analysis, molecular docking, and machine learning rescoring provides a robust framework for identifying and evaluating potential drug targets and ligands for breast cancer treatment.

V. FUTURE SCOPE

The research integrates multiple approaches, identifies novel drug targets using network biology, and combines molecular docking with machine learning for improved predictions. This comprehensive approach provides a solid foundation for further investigations in breast cancer drug discovery and personalized treatment strategies.

VI. ACKNOWLEDGEMENT

We gratefully acknowledge Prof. M. Sumakanth, Principal, RBVRR Women's College of Pharmacy, Hyderabad for her intellectual guidance, thoughtful mentorship and expert insights which greatly contributed to the depth and quality of this research.

REFERENCES

[1] G. Van and V. Grolmusz, "When the Web meets the cell: Using personalized PageRank for analyzing protein interaction networks," Bioinformatics, vol. 27, no. 3, pp. 405–407, Feb. 2011.

[2] J. Goll and P. Uetz, "The elusive yeast interactome," Genome Biol., vol. 7, p. R56, Feb. 2006.

[3] J. Li and P. X. Zhao, "Mining functional modules in heterogeneous biological networks using multiplex PageRank approach," Front. Plant Sci., vol. 7, p. 903, Jun. 2016.

[4] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista et al., "NCBI GEO: Mining tens of millions of expression profiles–database and tools update," Nucleic Acids Res., vol. 35, pp. D760–D765, 2007, doi: 10.1093/nar/gkl887.

[5] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," Mol. Syst. Biol., vol. 3, no. 1, p. 140, Oct. 2007.

[6] X. Zhu, M. Gerstein, and M. Snyder, "Getting connected: Analysis and principles of biological networks," Genes Dev., vol. 21, no. 9, pp. 1010–1024, May 2007.

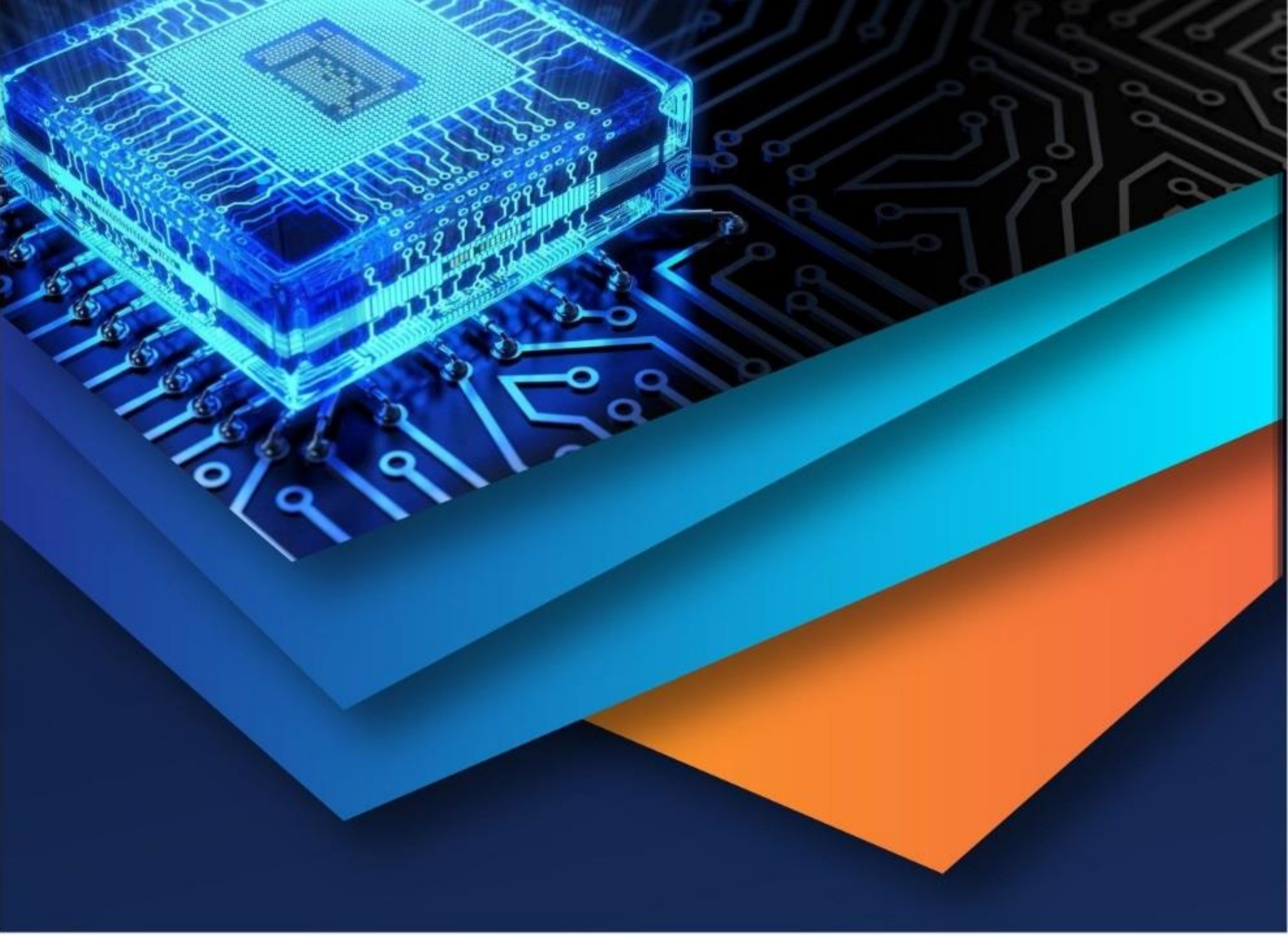
[7] A. Batool and Y. C. Bun, "Breast cancer classification using random forest algorithm," in Proc. J. Phys.: Conf. Ser., vol. 2559, no. 1, p. 012002, Aug. 2023.

[8] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," J. Algorithms Comput. Technol., vol. 12, no. 2, pp. 119–126, Jun. 2018.

[9] F. Ahmed, J. W. Lee, A. Samantasinghar, Y. S. Kim, K. H. Kim, I. S. Kang et al., "SperoPredictor: An integrated machine learning and molecular docking-based drug repurposing framework with use case of COVID-19," Front. Public Health, vol. 10, p. 902123, Jun. 2022.

[10] C. Wang and Y. Zhang, "Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest," J. Comput. Chem., vol. 38, no. 3, pp. 169–177, Jan. 2017.

[11] P. D. Lyne, "Structure-based virtual screening: An overview," Drug Discov. Today, vol. 7, pp. 1047–1055, 2002.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

TOGETHER WE REACH THE GOAL

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)