International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Random Forest Algorithm for Enhanced Prediction of Drug Target Interactions

Shikha Mehta, Sparsh Sharma, Sagar Anand, Shreshth Sharma, Aditi Goel

Abstract: ***Identification of drug-target interaction (DTI) is an important challenge for research and development in the pharmaceutical industry. Biomedicine researchers have stepped from in vitro and in vivo experiments to in-silico methods for fast results. In the recent past, machine learning algorithms have become very popular for DTI predictions. This paper presents an ensemble approach- Random forest algorithm for DTI predictions. The performance of proposed approach is evaluated with respect to Matrix factorization, genetic algorithm, Support vector machines, K-nearest neighbor, Decision Trees and Logistic Regression over 4 benchmark datasets with diverse properties. The algorithm is evaluated over Accuracy and average ranking. Results establish that random forest algorithm is more suitable or DTI predictions as compared to other algorithms.***

Keywords: Random Forest, Drug target Interactions, Chemogenomic, Genetic Algorithm, Ensemble Approach.

I. INTRODUCTION

The discovery of new drugs and repositioning of old drugs has always been the key problem and challenge in the field of biomedicine. Drugs interact with target proteins to trigger or restrain a target's biological process. Therefore, determination of target genes that interact with the drugs that is drug target interactions (DTI) is an important step in drug discovery. In last few years, in silico based DTI prediction techniques have become more prominent as in vitro and in vivo experiments are time consuming, costly and laborious. In literature, computational methods used for DTI prediction are divided into 3 main categories namely ligand based methods [1], docking simulations [2] and chemogenomic methods [3].

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Shikha Mehta*, Department of Computer and Science Engineering, Jaypee Institute of Information Technology, Noida, India. E-mail: mehtshikha@gmail.com
Sparsh Sharma, Department of Computer and Science Engineering., Jaypee Institute of Information Technology, Noida, India. E-mail: mehtshikha@gmail.com
Sagar Anand, Department of Computer and Science Engineering., Jaypee Institute of Information Technology, Noida, India. E-mail: mehtshikha@gmail.com
Shreshth Sharma, Department of Computer and Science Engineering., Jaypee Institute of Information Technology, Noida, India. E-mail: mehtshikha@gmail.com
Aditi Goel, Department of Computer and Science Engineering, Jaypee Institute of Information Technology, Noida, India. E-mail: mehtshikha@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Ligand based methods are based on the concept that similar molecules mostly bind to the same group of proteins like Quantitative Structure Activity Relationship (QSAR) [5]. Docking Simulations are time consuming as they require three-dimensional (3D) structures of proteins. Moreover, these simulations cannot be performed where 3d structures of proteins are not available or are too complex to obtain e.g membrane proteins like G-Protein Coupled Receptors (GPCRs), ion channel etc. Chemogenomic [4] methods have been used successfully in drug discovery in the last few years. These approaches are based on the integration of chemical space, pharmacological space or genomic space for DTI predictions. Chemogenomic approaches are further categorized as machine learning-based techniques, graph-based techniques and network-based techniques [4]. Among these, machine learning approaches have become more popular due to availability of massive data, computational power and reliability of predictions.

This work primarily focuses on machine learning algorithms. These are shallow learning techniques which are more suitable for structured data in contrast to deep learning models which are more suitable for perceptual applications. Considering the DTI prediction as a binary classification problem several algorithms such as support vector machines (SVM) [6] and regularized least square (RLS) [7] have been used for DTI prediction. Bleakley and Yamanishi[8] applied bipartite local models (BLMs) for DTI prediction. Xia et al.[9] developed semi-supervised machine learning approach called as Laplacian regularized least square (LapRLS). Liu et al.[10] presented neighborhood regularized logistic matrix factorization (NRLMF) for DTI prediction. It was based on logistic matrix factorization in which drugs-specific and targets-specific properties are represented as latent vectors. Cobanoglu et al. used probabilistic technique for matrix factorization (PMF) to predict [11] unknown DTIs.

This work presents Random forest algorithms for predicting the drug target interactions. Random forest algorithms are ensemble machine learning techniques whose efficacy has already been established in various medical applications such as lung cancer [12-13], Parkinson diseases[14], Alzheimer's disease[15] etc. However, as per literature review, this algorithm is not applied for predicting drug target interactions. Performance of random forest algorithm is evaluated with respect to 6 popular machine learning algorithms namely Matrix factorization, genetic algorithm, Support vector machines, Decision Trees and Logistic Regression over 4 datasets with varied properties.

Remaining the paper is arranged as follow- Related work is present in Section 2. Proposed random forest algorithm is presented in section 3 followed by experiments and results in section 4. Section 5 presents conclusion and future work.

II. RANDOM FOREST ALGORITHM FOR DTI PREDICTION

It is an ensemble learning algorithm. Ensemble models are composite methods of classification which combine multiple classifiers. These algorithms include multiple learners known as base learners for developing classification model. Base learners may be generated using

- Different algorithms, e.g Decision Tree, Bayes algorithm etc
- Different hyper parameters but same algorithm.
- Different representations/ modalities
- Different Training sets.

Final target label is based on the combined decision of all learners based on majority voting, weighted or unweighted voting. An ensemble technique combines a number of n individual classifiers C_1, C_2, :::, C_n, to create an enhanced composite model M. For the same, given dataset D is divided into n training sets D_1, D_2,:::, D_n, where D_i ($(1 \le i \le n-1)$ is used to generate classifier C_i. For the new data sample, each base classifier predicts the class label. The ensemble methods use majority voting, weight voting methods etc for final class label prediction. These methods have become more popular in the recent past due to their reliability and considerable accuracy of prediction as compared to individual classifiers. In[] authors developed two ensemble approach –decision tree ensemble and Kernel Ridge regression ensemble. Results indicae KRR ensemble performed better than Decision Tree ensemble. This decision tree ensemble was not random forest. In contrast to bagging and boosting methods, random forest is an ensemble of decision tree classifier. Numbers of trees combine to form the "forest". To decide the split for all nodes of the individual decision trees in the forest, attributes are selected randomly. Random forests do not depend on the number of attributes considered for each split and use inherent estimate of attribute importance. Many small decision trees may be created in parallel on different CPUs and then combined to form a single strong learner. Therefore, they are very efficient and suitable for large databases. For classification, most popular class is assigned based on voting of each tree. RF algorithm is mainly based on the concept that modeling small decision trees with few attributes is computationally less expensive. The accuracy of RF is comparable to AdaBoost, yet it is more robust to errors and outliers.

The pseudo code of Random Forest (RF) as given in Algorithm 1 is divided into two parts

(a) Creation of Random Forest

(b) Prediction using RF classifier created in Step (a)

The working of RF algorithm is as follows: For every tree in the forest, a bootstrap sample S is selected where $S^{(j)}$ denotes the jth sample of the bootstrap. Thereafter learning model of decision-tree is developed using a modified decision-tree algorithm. The modifications in the algorithm are as follows: Rather than evaluating all the possible attribute splits at every node of the tree, a small subset of features(f) are randomly selected such that $f \subseteq F$ where F is the set of features. Subsequently the decision tree is learned based on this small set f which considerably reduces the complexity of the algorithm as f is very very small then F. node

Algorithm 1 Random Forest

Input:
A training set ST with features F
Number of Trees-T
----------.
1 Select k features from F randomly, where k<<F
2 Compute the best split node "n" from k features
3 Create child nodes by splitting the node using best split.
4 Repeat steps 1-3 until desired number of nodes.
5 Repeat steps 1-4 to build the forest with desired number of trees "n"

For Prediction:

1. ***Use the test set features and predict the outcome using the various rules generated by the decision trees created randomly.***
2. ***Store the predicted outcome.***
3. ***Count the votes for every predicted class.***
4. ***The target with highest votes as assigned as final target for prediction.***

III. EXPERIMENTS AND RESULTS

A. Dataset

In the presented work, meticulous experiments were performed to evaluate the performance proposed random forest algorithm based prediction of drug target interactions. The algorithm was evaluated on 4 benchmark protein datasets namely Ion Channels, Enzymes, ion channels, nuclear receptors, and GPCRs. The numbers of drugs which are known to the target are 233, 445, 210 and 54 in GPCRs, enzymes, ion channels, and nuclear receptors datasets respectively. The counts of various proteins that are known to be targeted by the drugs are 664, 204, 95, and 26 FOR enzymes, ion channels, and nuclear receptors datasets respectively. These datasets were used as the gold-standard datasets by Yamanishi et al [3] These datasets are freely available from KEGG BRITE [19], BRENDA [21], SuperTarget [18], and Drug Bank [20] databases.

B. Evaluation Parameters

Accuracy- Accuracy is a performance evaluation measure for supervised learning algorithms to assess the ability to correctly classify or predict labels. It is defined as the ratio of correctly predicted data items to the total number of data items. Higher the accuracy means good algorithm.

Average ranks- Average ranking[22] is a conventional and simple technique to rank the algorithms. In this method, accuracy obtained by every algorithm for each dataset is sorted and assigned the ranks.

The algorithm with highest value is ranked 1, second highest is ranked 2 and so on for all datasets independently. Then overall average rank of each algorithm is computed by taking mean of ranks on all datasets. Let $rank^n_m$ be the m^{th} algorithm rank of n^{th} dataset. Then average rank of each algorithm is computed as

$$rank_m = \frac{\sum_{n=1}^{k} rank_m^n}{k} \quad (1)$$

C. Analysis of Results

The performance of proposed Random forest algorithm is evaluated with respect to 6 algorithms namely Genetic algorithm(GA), Matrix factorization(MF), K-nearest neighbor(KNN), Support vector machine(SVM), Decision trees(DT) and Logistic regression(LR). These algorithms have been widely used for predicting drug target interaction in the literature. Figure1 - Figure 6 depict the results obtained by all the algorithms over ion channel dataset, GPCR dataset, nuclear receptor dataset and enzyme dataset respectively. It can be observed from these results that Random forest algorithms are

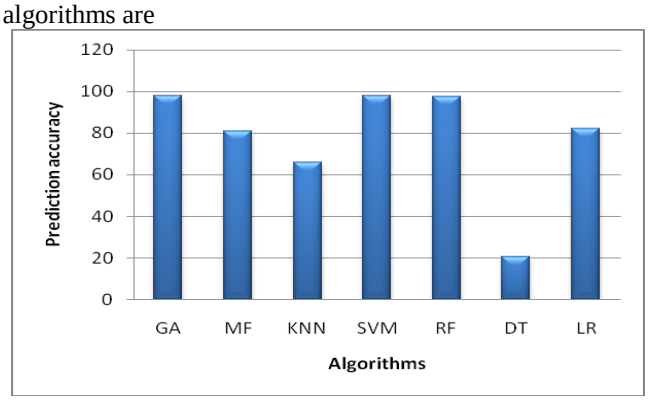

Figure 1: ION Channel Dataset

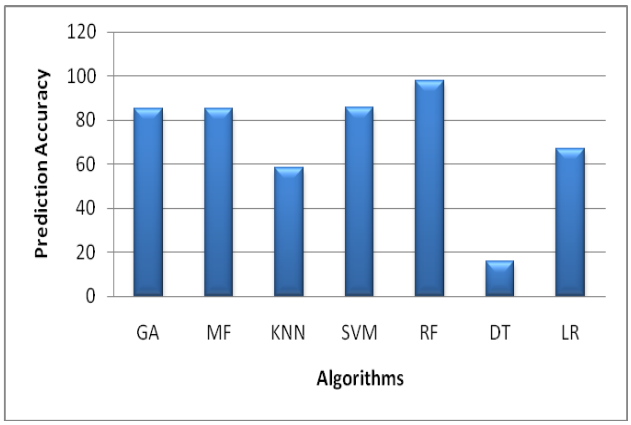

Figure 2: GPCR dataset

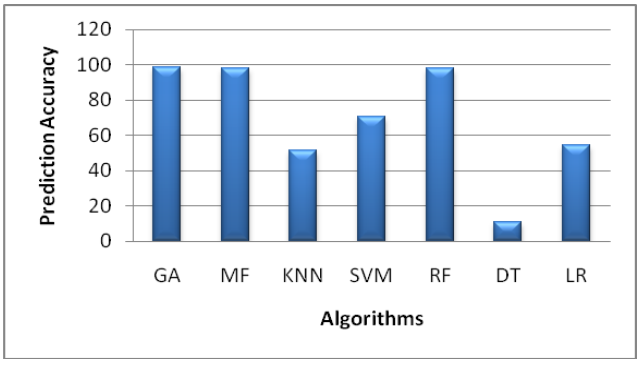

Figure 3: Nuclear Receptor

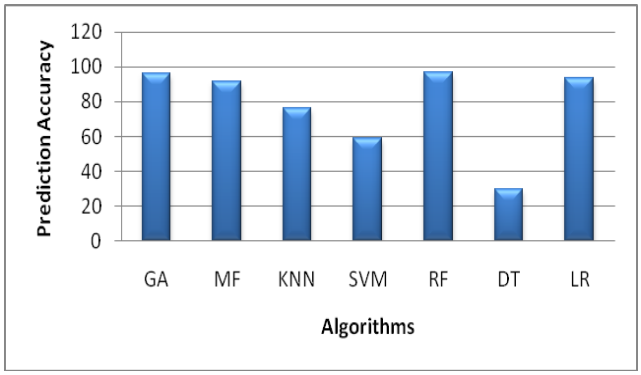

Figure 4: Enzyme

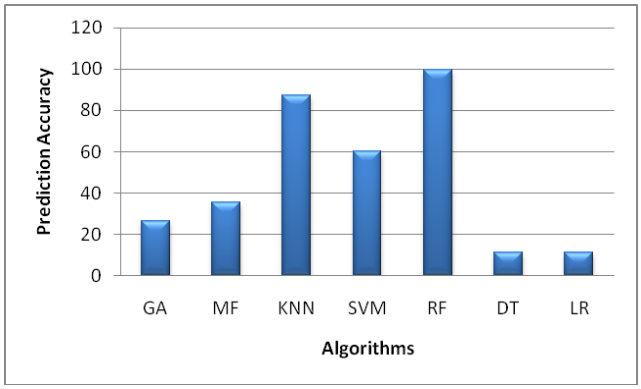

Figure 5: Protein dataSet -1

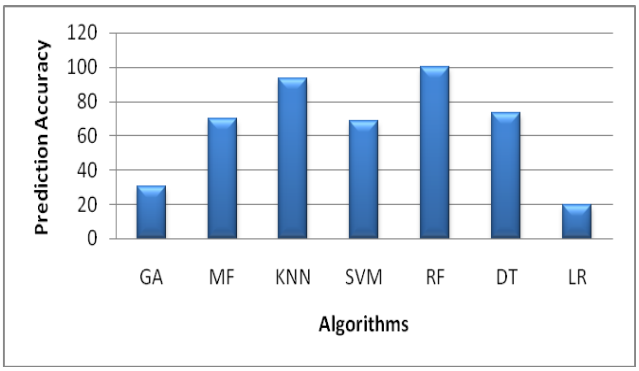

Figure 6: Protein dataSet -2

Depicting highest accuracy and decision trees are showing the least accuracy in all datasets. Since other algorithm like genetic algorithm, matrix factorization and support vector machines are also showing considerable performance, these algorithms are ranked over all datasets. Average ranking help in assessing the overall performance of datasets over datasets with varied characteristics as shown in Table 1.

Table1: Average Ranking of algorithms over 4 datasets.

Datasets \Algorithms	Ion Channel	GPCR	Nuclear Receptor	Enzyme	Average Rank
GA	1	2	1	1	1
MATRIX	3	2	1	3	2
KNN	4	4	4	4	4
SVM	1	2	2	5	3
RF	1	1	1	1	1
DT	6	5	7	6	6
LR	2	3	3	2	3

It can be inferred from Table 1 that Random forest and genetic algorithm are able to obtain same rank due to their similar prediction accuracy for all datasets. As genetic algorithm is a non deterministic meta-heuristic algorithm, it may not be suitable for large datasets. Therefore, Random forest may be preferred over GA. Second Rank is obtained by Matrix factorization technique, whose computational complexity is quite high. Logistic Regression and support vector machine stand third, even they are suitable for small datasets only. From these results it can be concluded that RF, GA, SVM and LR portray good performance on the considered datasets. For large datasets, random Forest Algorithm may be preferred over others as it is fast, computationally less expensive and provides good prediction accuracy.

IV. CONCLUSION

Convergence of information and communication technologies has given a boost to the drug discovery in pharmaceutical industry as well. Prediction of drug-target interaction (DTI) is a challenging task in drug discovery process. To cope with the issues, machine learning algorithms are used very prominently to get better results at a faster pace. This work presented Random Forest for predicting drug target interactions. It is an ensemble learning algorithm which is less complex and provides faster results as compared to its contemporary counterparts. The experimental evaluation of proposed algorithm over accuracy and average ranking metrics for various datasets establish that Random forest algorithm is able to attain considerable accuracy and is better than compared algorithms.

REFERENCES

1. Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K. Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 2007, 25, 197–206.
2. Arola, L.; Fernandez-Larrea, J.; Blay, M.; Salvado, M.J.; Blade, C.; Ardevol, A.; Vaque, M.; Pujadas, G. Protein-ligand docking: A review of recent advances and future perspectives. Curr. Pharm. Anal. 2008, 4, 1–19.
3. Yamanishi, Y. Chemogenomic approaches to infer drug–target interaction networks. In Data Mining for Systems Biology: Methods and Protocols; Mamitsuka, H., DeLisi, C., Kanehisa, M., Eds.; Humana Press: Totowa, NJ, USA, 2013; Volume 939, pp. 97–113. ISBN 978-1-62703-107-3.
4. Mousavian, Z.; Masoudi-Nejad, A. Drug-target interaction prediction via chemogenomic space: Learning-based methods. Expert Opin. Drug Metab. Toxicol. 2014, 10, 1273–1287. [CrossRef] [PubMed]
5. Li, J.; Zheng, S.; Chen, B.; Butte, A.J.; Swamidass, S.J.; Lu, Z. A survey of current trends in computational drug repositioning. Brief. Bioinform. 2016, 17, 2–12. [CrossRef] [PubMed]
6. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press; 2000.
7. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters. 1999;9(3):293–300.
8. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics. 2009;25(18):2397–2403.
9. Xia Z, Wu LY, Zhou X, Wong ST. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol. 2010;4(suppl 2):S6
10. Liu Y, Wu M, Miao C, Zhao P, Li X-L (2016) Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. PLoS Comput Biol 12(2): e1004760. https://doi.org/10.1371/journal.pcbi.1004760
11. Salakhutdinov R, Mnih A. Probabilistic matrix factorization. Adv Neural Inf Process Syst 21. 2008; p. 1257–1264.
12. D, Rovelli C (2019) Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PLoS ONE 14(1): e0208737. https://doi.org/10.1371/journal.pone.0208737
13. 13. Lee, K., Jeong, H., Lee, S. et al. CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network. Sci Rep 9, 16927 (2019) doi:10.1038/s41598-019-53034-3
14. A. Rahmim et al., "Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images", NeuroImage Clin., vol. 16, pp. 539-544, Aug. 2017.
15. Random forest prediction of Alzheimer's disease using pairwise selection from time series data Moore P.J., Lyons T.J., Gallacher J. (2019) PLoS ONE, 14 (2) , art. no. e0211558
16. Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. PLoS Comput Biol, 2012, 8(5): e1002503.[for datasets]
17. Meng, F.-R.; You, Z.-H.; Chen, X.; Zhou, Y.; An, J.-Y. Prediction of Drug–Target Interaction Networks from the Integration of Protein Sequences and Drug Chemical Structures. Molecules 2017, 22, 1119.[dataset details]
18. Günther S. SuperTarget and Matador: Resources for exploring drug-target relationships. Nucleic Acids Res. 2008;36:919–922. doi: 10.1093/nar/gkm862.
19. Kanehisa M., Goto S., Hattori M., Aokikinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., Hirakawa M. From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res. 2005;34:354–357. doi: 10.1093/nar/gkj102.
20. Wishart D.S. DrugBank: A knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36:D901–D906. doi: 10.1093/nar/gkm958
21. Schomburg I., Chang A., Ebeling C., Gremse M., Heldt C., Huhn G., Schomburg D. BRENDA, the enzyme database: Updates and major new developments. Nucleic Acids Res. 2004; 32:431–433. doi: 10.1093/nar/gkh081.
22. Brazdil P. B., Soares C., "A Comparison of Ranking Methods for Classification Algorithm Selection," in European conference on machine learning, 2000, pp. 63–75.
23. Ali Ezzat, Min Wu, Xiao-Li Li, Chee-Keong Kwoh, Drug-target interaction prediction using ensemble learning and dimensionality reduction, Methods, Volume 129, 1 October 2017, Pages 81-88.
24. Jinjian Jiang,1,2 Nian Wang,1 Peng Chen,3 Jun Zhang,4 and Bing Wang5, DrugECs: An Ensemble System with Feature Subspaces for Accurate Drug-Target Interaction Prediction, BioMed Research International,olume 2017 (2017), Article ID 6340316, 10 pages https://doi.org/10.1155/2017/6340316

AUTHORS PROFILE

Dr Shikha Mehta, Associate Professor, Department of Computer Science Engineering, Jaypee Institute of Information Technology, Noida, attained her doctorate in the domain of Artificial Intelligence from University of Delhi, Delhi. She has over 18 years of teaching experience at both undergraduate and postgraduate level. Dr Shikha has organized many special sessions in conferences and is in the program committee of various conferences of repute. She has addressed diverse audience as a keynote speaker in UGC and TEQIP sponsored FDPs and workshops. With many national and international research publications to her credit, she is also the chief editor of a book published by reputed international publisher. She is an active researcher in the field of Machine Learning, Nature Inspired computing and Social networks. She has guided various M tech thesis, more than 100 Btech projects and Presently she is guiding three PhD scholars and one has been already awarded PhD under her guidance.

Sparsh Sharma, is currently pursuing his B.Tech Computer Science from Jaypee Institute of Information Technology, Noida. He is also the Training and Placement coordinator for the batch 2016-2020. His interest is in the field of Nature inspired algorithms .Currently he is working on increasing the lifetime /efficiency of Wireless Sensor Network using Nature inspired algorithms.

Shreshth Sharma , Student Of Jaypee Institute of Information Technology, Computer Science Department. He is having interest in the field of Machine Learning and Artificial Intelligence. Learning and exploring in the field of Genetic Network Programming, Neural network and its applications. Looking forward for research opportunities in the field of Artificial Intelligence and harnessing its power to help the needy.

Sagar Anand, Student of Jaypee Institute of Information Technology, Computer Science Department. He is having interest in Machine Learning and Artificial Intelligence. Practicing and exploring the fields of Machine Learning and Deep Learning and its applications. Researching in the field of Genetic Network Programming to explore new openings it might offer.

Aditi Goel, Student of Jaypee Institute of Information Technology, Computer Science Department. She is researching in the fields of machine learning and exploring its various applications. She has keen interest in Drug Target Interaction Prediction field. Also have good hand in image recognition and analysis. She is also exploring in the field of artificial intelligence and Cloud Computing.