Feature Based Method for Predicting Pharmacological Interaction

Ansa Baiju, Linda Sara Mathew, Neethu Subash

Abstract: ***Prediction of drug target interaction is an extrusive domain of drug discovery and repositioning of drugs. Most conventional studies are carried out in early years in the wet laboratory, but it is very expensive and time consuming. So nowadays, the use of machine learning techniques to predict drug target pairs. A new method of interaction targeting drugs is introduced in this paper. Use the Pseudo Position Specific Scoring Matrix (PsePSSM) is used to represent the target, which generate features that describe the original information of protein. The drug chemical structure information can be extracted through FP2 molecular fingerprint which describe the molecular structure information. Then a drug target interaction network is constructed using bipartite graph where in which each node represents a target or drug and each link indicates a drug target interaction. From the above stages, the data contains some noise and redundant data which have a negative impact on the prediction output. So, LASSO (Least Absolute Shrinkage and Selection Operator) method is handle it and reduce the dimension of the extracted feature information of original data. But drug target pair samples have some imbalanced, then cost-sensitive ensemble method is used to address the imbalanced problem between positive and negative samples, and learns about the minority class by assigning higher costs and optimizing their cost error. Finally, the processed data is given as input to the extreme gradient boosting classifier algorithm for predicting new drug target interaction pairs. This method can significantly improve the prediction accuracy of drug target interaction.***

Keywords: Drug Target Interaction, Lasso, Extreme gradient boosting classifier, Pseudo Position Specific Scoring Matrix

I. INTRODUCTION

Drug or medicine are chemical syntheses that are used to cure, prevent and treat various diseases and diseases. The drug that interacts with our body with a certain biomolecule and treats for some diseases. Targets are primarily biomolecules, some of which are proteins, enzymes, channels of ions, nuclear receptors, etc. Therefore, the prediction of the drug target interaction essentially refers to the drug's interaction with the target protein and that produces an impact in the human body. The prediction of drug target pairs is a noteworthy area in drug discovery [1] i.e. identifying a novel drug for an existing protein target and also discover the target. If the drug that interacts with some target as macromolecule and performs a specific function. Some studies have been incorporated in the early years to identify the drug target pairs in a wet laboratory, but it is very expensive and time consuming. So, some form of computation method is adopting and evaluate the interaction pairs.

Some conventional computational methods can be categorized as three methods to predict drug target pairs: ligand-based approaches, docking approaches, and chemogenomic approach [2]. A system analogous to a molecule that interacts with identical targets and also shows similar properties is the ligand-based approaches. This approach is like the Quantitative Structure Activity Relationship (QSAR), which compares certain ligand to known ligand only by suiting other methods of machine learning. But this approach has some drawback, it's not considering target protein sequence information. These approaches have less performance, because it only considers less recognized protein ligands. The docking method uses 3D protein structures but it is inapplicable because the 3D structure of some protein is not available. And their prediction of interaction is a very challenging task. The chemogenomic approach which simultaneously uses both the chemical space of the drug and the genomic space of the target. These methods which integrate drug and target information into unified space. It is also divided into network-based method, graph-based method, and machine learning process-based features. All of these approaches are using the drug and target function and implementing a form of machine learning and predicting the interactions.

Machine learning approaches are being implemented for the prediction in the last few years. Support vector machine [3], K-nearest neighbors [4] and random forest [5] etc. are the widely used methods for predicting the drug target interaction. The high dimensional data are very difficult to process in the identification of the drug target interaction pairs. So, use a method of reducing dimensionality that applies to the dataset. The dimensionality reduction is classified into feature selection and feature extraction. The feature selection means which omit all redundant and irrelevant features from the samples. The feature extraction, which maps all features into a useful information and into low dimensional features. So, a lasso technique was used in the paper to handle high-dimensional data and extract main features. Unbalancing data may have a negative effect on results. It happens when the number of one class is substantially greater than the other class. Many methods may be used to control data imbalances, such as data level and algorithm level [6]. The data level mainly data preprocessing and re balance the data by resampling methods like oversampling and under sampling.

Manuscript received on January 06, 2021.
Revised Manuscript received on January 15, 2021.
Manuscript published on January 30, 2021.
* Correspondence Author
Ansa Baiju*, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. Email: ansabaiju96@gmail.com
Linda Sara Mathew, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. Email: lindasaramathew@gmail.com
Neethu Subash, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. Email: Neethu.subash@gmail.com

However, resampling techniques can some risk like major lose in essential information of majority samples when it under sampled and in oversampling, there is overfitting in minority samples. In algorithm level, involves existing classifiers learning algorithms to minority classes such as cost sensitive learning and ensembles schemes. In the proposed methods, a support vector machine based on self-adaptive cost weights should be used to manage the imbalanced samples. Cost sensitive ensemble learning that learns from the minority class and assigns the highest cost to samples of the minority class. Thus, a new approach is used to identify new drug target interaction pairs in this paper. Firstly, represent the target protein from the position score matrix and create a pseudo position score matrix and also represent the drug compound using the FP2 fingerprint. Then construct a bipartite graph by using the information of drug target pairs information on the extracted features. Then, select the features from the drug target pairs and also reduces the redundant information by using Lasso method. The cost-sensitive approach is used to f i x data imbalanced issues. Finally, all the data are fed to the extreme gradient boosting and predicting the new drug target pairs.

The rest of the paper is as follows. Section 2 describes the related works in this area. Section 3 briefly explains the proposed work. Section 4 describes obtained results and the last section concludes the paper.

II. RELATED WORK

In 2009 Kevin Bleakley [7], a novel supervised interference method is proposed to predict unknown DTI from chemical information and genomic sequence. The method which is formalization of bipartite graph interference which has a set of separate local supervised learning problems. All the problems which predict new drug compounds, new target compounds which are associate to target protein. The known bipartite local model which predict the protein target for a given drug and also predict the drug target for the given protein. By training the local models which solve the bipartite graph interference problem and predict a new edge linking drug nodes with the target. Then make a list of all known target and drug in bipartite network and labels as +1 and -1 then apply a classification rule to +1 data from -1 label from the available genomic sequence and chemical structure of drug. To predict the DTI by using SVM.

In 2012, Cheng F [8], is implement a novel method for predicting the new DTI and drug repositioning. The method is derived from recommendation algorithm. To predict the interaction, there are three inference methods are developed: Drug-Based Similarity Interference (DBSI), Target Based Similarity Interference (TBSI). In DBSI, the idea is drug which interact with a target, then other drug similar to drug will recommends the target. In TBSI, the drug which interact with a target, then drug will be recommended to another target. NBI method is used to predict the new DTI in the drug target bipartite network. In 2016, Zaynab Mousavian [9], a new learning model is proposed to predict the new DTI pairs from the evolutionary information of protein. From the position scoring specific matrix firstly extract the bigram features. Extract the positive and negative samples from the bipartite graph. Balance the samples by randomly select the negative samples from the unknown DTI pairs until the samples can reaches to the positive samples. Then concatenate the fingerprint of drug and target by encoding the target and drug. Then finally SVM is used to predict the new DTI pairs.

In 2017, Fan-Rong Meng [10], a new computational method is proposed to predict the new DTI pairs based on the protein sequence i.e., Predicting Drug Targets with Protein Sequence (PDTPS). It combines the bigram probabilities, Position Specific Scoring Matrix (PSSM), and Principal Component Analysis (PCA) with Relevance Vector Machine (RVM). The bigram features which is counting of bigram frequencies occurrences in PSSM. Based on the PCA, it reduces the redundant sample and convert into low dimensional samples. Finally, to predict the new DTI pairs by using RVM (identifiable function of SVM). But the disadvantage of the prediction is probabilities.

In 2017, Rawan S. Olayan [11], efficient computational method to discover a new DTI pair. This method is based on the heterogenous graph which contains known DTI with multiple similarities between drug and multiple similarities between target. It combines the different similarities by applying a non linear similarity fusion. Before fusion, it performs a preprocessing step which selecting the best similarities. Then normalized the similarity measure of drug and target by the range [0,1]. Finally, applies random forest model to the heterogenous graph and predict the new DTI pairs.

III. PROPOSED WORK

The proposed method is to predict the novel drug target pairs and has feature extraction phase, reducing the dimensionality, balancing the dataset and finally predict it. The figure 1 shows the proposed method.

In the proposed approach, use a gold standard dataset and it contains four data's like enzymes, ion channels (IC), G protein-coupled receptors (GPCR) and nuclear receptors NR). It's obtained from KEGG BRITE, BRENDA, Super Target and Drug Bank and drug approved datasets. Datasets containing different details, such as drug component, target protein and drug target interactions. This information may be used to correctly classify the different new experiences.

A. Represent Target

The proposed approach, first, generates a pseudo-position specific score matrix to characterize the target protein from a position specific score matrix. In the first step of constructing a position weight matrix, a simple Position Frequency Matrix (PFM) is generated by counting the occurrences of each nucleotide at each position. A Position Probability Matrix (PPM) can now be generated from the PFM by dividing the former nucleotide count by the number of sequences at each position, thus normalizing the values. Based on the extraction of amino acid sequences of proteins on enzymes, ion channels, GPCRs, and nuclear receptors, in order to convey the characteristic information in

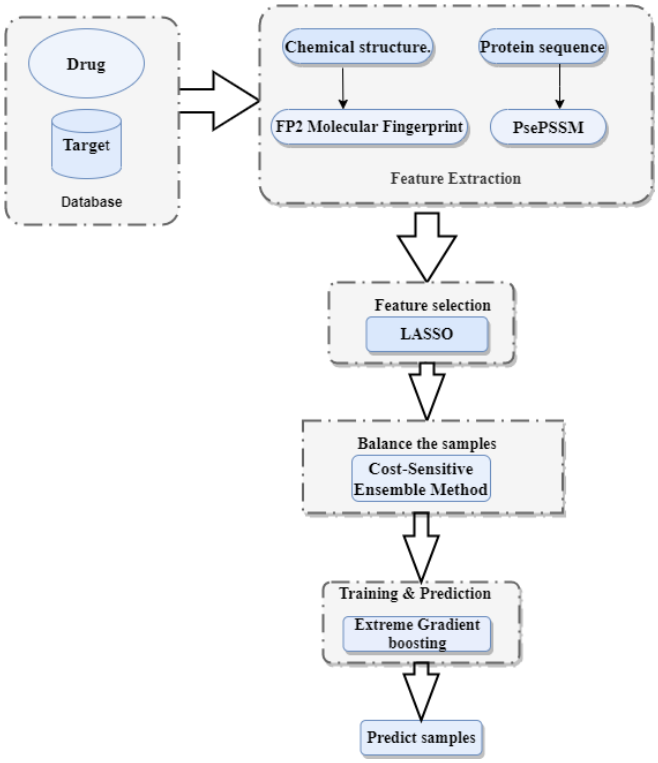

Fig 1. Proposed Model

the amino acid sequence as accurately as possible, the pseudo-position specific matrix (PSSM) [12] characteristics are used to describe the evolution and sequence details of the protein sequence.

The protein target sequence P with L amino acid residues it and dimension is L X 20 (20 amino acid). Firstly calculate the score ($E_{i\to j}$)of the residue of the i-th position in the amino acid sequence being mutated to the j-th 11 amino acid residue, which is searched using PSI-BLAST[13].The positive score indicates that the resulting residue is more often mutated and the negative is just the opposite. The maximum number of iterations for multiple search in PSI-BLAST is 3. Then normalize the matrix into the interval (0,1) by using Equ. 1.

$$\overline{E}_{i\to j} = \frac{1}{1+\exp\left(-E_{i\to j}\right)} \quad (1)$$

Where $\overline{E}_{i\to j}$ is the score of the i-th position in the amino acid sequence changed to the j-th amino acid after the normalization. But some loss in sequence information so use the concept of Pseudo amino acid composition (PseAAC). Then generate 20+20 X dimensional feature vector. After feature extraction, the different length of protein sequences is converting into same dimension by using this method.

B. Represent Drug

To represent the drug compounds by using FP2 molecular fingerprint [14]. Different types of drugs have different properties and descriptors. Molecular fingerprint is a way of encoding the structure of a molecule. The most common types of fingerprint are a series of binary digits that represent the presences and absences of particular substructure in the molecule. To extract the molecular

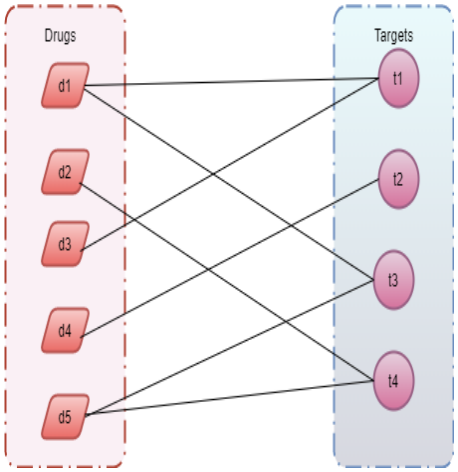

Fig 2. Bipartite Graph

fingerprint from the drug molecules by the using the software Open Babel [15]. Open Babel is a software, a chemical expert system mainly used to convert chemical file formats. A mol file format of the given drugs containing extensive details on the chemical structure of the drug.

Users can then use the program to transfer data between different formats and also allow users to scan, convert and store different formats. Various types of fingerprints can be produced in the toolkit. Fp2 fingerprint is used in the proposed process. The mol format of the drug compound which converted into Fp2 molecular fingerprint file format. But the sequence of the fingerprint is hexadecimal digit of length 256 which convert into the decimal sequence between the 0 and 15 of drug compounds.

C. Constructing Graphs

The known DTI can be represented by bipartite graph as shown figure 2. Bipartite graph is a graph with vertices that can be separated into two distinct and different sets such that each edge connects the vertex to one. Each node in the graph which represent the target protein or drug molecule and each side which represent the known DTI pairs. All the known D T I pairs is represented as positive samples and non-DTI are considered as negative samples. The number of non-interacting drug target pairs is greater than the number of interacting drug target pairs i.e. the number of negative samples is higher than positive Samples creating an unbalanced problem, using a cost-sensitive set process.

D. Reduce the Dimensionality

In the drug target pairs has a few noises and redundant facts which is have an effect on the output of last model. To reduce the noise, useless samples and also discover the main feature from the drug target pairs so use, Lasso technique [16] is used to lessen the dimension of the drug target pairs because processing an excessive dimensional information is very complex issue. Lasso is a regression analysis and compression estimation approach and is often used for the selection of variables. The key principle of this approach is to cause the sum of the absolute value of the regression coefficients to be less than the fixed value and to set the coefficients to zero.

Reducing the model parameter values to impose a penalty against complexity. Minimizes the penalized residual number of squares.

Given a linear regression with sample dataset X_{ij} and centered response values y_i for i=1, 2,..,N and j=1,2,...,p, the lasso solves the L1 -penalized regression problem of finding w= w_j to minimize in the Equ.2.

$$\sum_{i=1}^{N}(y_i - \sum_{j} X_{ij}\ w_j)^2 + \sum_{j=1}^{P}|w_j| \quad (2)$$

where w is the regression coefficients. This method also improves prediction accuracy and reduced the irrelevant information

E. Balancing the Samples

Unbalanced data can have an effect on the efficiency of the prediction model. It happens when the number of negative samples is greater than the positive sample. A self-adaptive, cost-sensitive ensemble approach [17] is used to do this. In this step, the Ad boost algorithm is used to handle and use the support vector machine as a base learner. In the ad boost build base learner sequentially and also give more weight to misclassified samples by base learners and less weight is allocated to correctly classified samples. Every iteration then updates the weight to misclassified samples based on the output of the previous classifier. The updation minimizes the cumulative training error of the combined classifiers. In Ad boost algorithm the determination of wight updation and optimization of classifiers are improve the performance of ensemble.

If given the unbalanced data samples firstly train the cost sensitive SVM base learner and distribute the weight and cost. The update the weight and normalize the samples in Equ.3

$$D^{(t+1)}(i) = \frac{c_i^{(t)}(i)\exp(-\propto_t h_t(x_t)y_i)}{Z_t} \quad (3)$$

This method is which handle the unbalancing data efficiently

F. Predicting the DTI pairs

Finally, all the processed data are fed into the classification algorithm and predict the drug target interaction pairs. Extreme Gradient Boosting (XGBoost)[18] is a machine method to predict the DTI pairs. XGBoost is optimized ensemble algorithm and advanced implementation of Gradient Tree Boosting (GTB). It is ensemble tree boosting approach which follows the same rules as GBT. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. XGBoost follows the same process as the GTB algorithm with a minor adjustment to the regularized goal to increase model performance. In the dataset, the result of the prediction is the sum of the predicted scores of the K trees in the Equ.4.

$$y^{-} = \sum_{k}^{k=1} f_k(x_i) \quad (4)$$

needs to be defined. There must always be two sections to an objective function: training loss and regularization. The term regularization penalizes the complexity of the model. XGBoost includes regularization, thus controlling the complexity of the model and preventing overfitting.

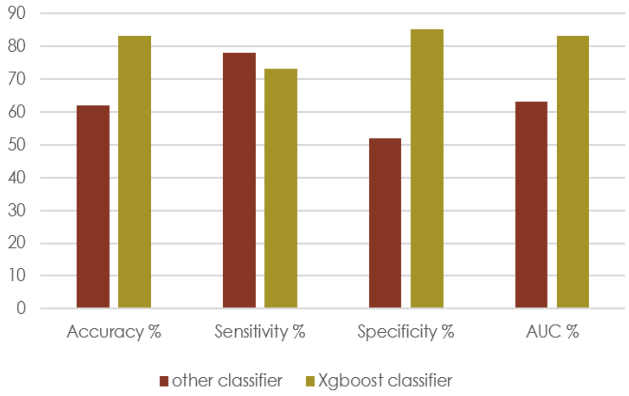

Fig 3. Comparison Graph

IV. RESULT AND DISCUSSION

In the proposed method, the first step is to describe and extract the protein target sequence using PsePSSM. Then represent the drug using FP2 molecular fingerprint and accession numbers of the drugs such as DB0002, DB00005 etc. are used for these studies. Then reduce the dimensionality of the dataset. But the sample has some imbalance so use a cost sensitive ensemble method is used. Many machine learning methods has been proposed for the prediction of drug-target interaction. Then it compared the prediction performance with the other prediction method using the same datasets. Based on the comparison, paper use XGBoost classifier algorithm for the drug-target interaction prediction. To evaluate the performance evaluation of the proposed model, to identify the Accuracy (ACC), Specificity (SP), Sensitivity (SE). And also determine the confusion matrix, TP (True positive), FP (False Positive), TN (True Negative), FN (False Negative). The figure 3 shows the comparison of various method with proposed methods on the dataset. From the evaluation, the proposed method classifier has higher accuracy rate than other classifier.

V. CONCLUSION

Drug target interaction is extrusive area in the drug discovery. Drug discovery is a method which identify new drug and their new target. The experimental method which discover the drug target pairs is very time consuming and expensive. Use machine learning technique provides an effective and efficient method. So, a novel proposed method is used to predict drug- target interaction. The feature-based method which predict the interaction between these drug target pairs by discovering features. In the proposed method, firstly represent the protein target by the using PsePSSM and then represent the drug by using the FP2 Molecular fingerprint. Then construct the bipartite graph of known drug target pairs. Then Lasso method is used to extract feature from the original dataset and reduce the dimension.

But the dataset has some imbalanced to handle it and cost sensitive ensemble method is used to deal with the imbalance of positive and negative samples, and finally, the processed data is input into the extreme gradient boosting classification algorithm for drug-target interactions prediction. The cost- sensitive set method will effectively prevent over-fitting of the model. Extreme gradient-enhancing classification algorithms can handle multiple data types with faster learning speeds, effectively manage noise data, and create highly accurate classifiers.

REFERENCES

1. J.T. Dudley, T. Deshpande, A.J Butte, "Exploring drug disease relationships for computational drug repositioning", Briefings Bioinf. 12(4) (2011) 303-311
2. A. Ezzat, et.al., "Computational prediction of drug- target interactions using chemogenomic approaches: an empirical survey", Briefings Bioinf. (2018)
3. L. C. Wang, Z. X. Yang, Y. Wang, N.Y. Deng, "Computationally probing drug-protein interactions via support vector machine", Lett. Drug Des Discov. 7 (2010) 370–378
4. J. Wang, Z.H. You, X. Chen, S. X. Xia, F. Liu, X. Yan, Y. Zhou, K. J. Song, "A computational-based method for predicting drug–target interactions by using stacked auto encoder deep neural network", J.Comput.Biol.25(2017)361–373.
5. H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang, Y. Wang, "A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data," PLoS One7(5)(2012)e37608.
6. Chih. -Fong Tsaia, Wei-Chao Lin, Ya-Han Hue, Guan-Ting Yao, "Under sampling class imbalanced datasets by combining clustering analysis and instance selection" (2019) 47-54.
7. K. Bleakley, Y. Yamanishi, "Supervised prediction of drug- target interactions using bipartite local models", 25(18) (2009)2397–2403.
8. F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference", PLoS Comput.Biol.8(5)(2012)e1002503.
9. Z Mousavian, S Khakabimamaghani, K Kavousi, A Masoudi-Nejad "Drug–target interaction prediction from PSSM based evolutionary information" J. Pharn. Toxicol. Methods 78(2016)42-51.
10. F. R. Meng, Z. H. You, X. Chen, Y. Zhou, J.Y, "An prediction of drug-target interaction networks from the integration of protein sequences and drug chemical structures", Molecules 22(2017)1119.
11. R. S. Olayan, H. Ashoor, V. B. Bajic, "DDR:efficient computational method to predict drug-target interactions using graph mining and machine learning approaches", Bioinformatics 34(7)(2018)1164–1173
12. D. T. Jones, "Protein secondary structure prediction based on position specific scoring matrices",J.Mol.Biol.292(1999)195–202.
13. S. F.Altschul, T.L.Madden, A.A.Schäffer, J. Zhang, W. Miller, D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids".25(1997)3389–3402.
14. Y. Yamanishi, E. Pauwels, H. Saigo, V. Stoven, "Extracting sets of chemical substructures and protein domains governing drug-target interactions" J.Chem. Inform.Model.51(2011)1183–1194.
15. M. O. Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, "Open Babel: an open chemical tool box interactions" form.3(2011)33.
16. R. Tibshirani, "Regression shrinkage and selection via the LASSO:"J.R Statics. Soc. B73(2011)273-283
17. Xinmin Tao, Qing Li, Wenjie Guo, Chao ren, Chenxi Li, Rui Liu, Junrong "Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification", Information Sciences 487(2019),31-56.
18. J. Zhong, Y. Sun, W. Peng, M. Xie, J. Yang, and X. Tang, "XGBFEMF: An XGBoost-based framework for essential protein prediction,", IEEE Trans. Nanobiosci., vol. 17, no. 3, pp. 243–250, Jul. 2018.

AUTHORS PROFILE

Ansa Baiju received Bachelor of Technology in Computer Science and Engineering from Younus College of Engineering in 2018 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in machine learning and data mining

Prof. Linda Sara Mathew is currently working as assistant professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. She received her B-Tech Degree in 2002 and M-Tech in 2011 in Computer Science and Engineering from Mahatma Gandhi university and Anna University. She has around 15 years of teaching and industrial experience. She is interested in the areas of Data Mining, Neural Network, Image Processing, Soft Computing

Prof. Neethu Subash is currently working as assistant professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. She received her B-Tech Degree in 2008 and M-Tech in 2013 in Computer Science and Engineering from Mahatma Gandhi university Kerala. She has around 6 years of teaching and 2 years of industrial experience. Her research interest is in Cryptography, Image security, Blockchain and Machine Learning.