Design a new Method for Prediction of DNA-binding Protein using Deep Learning

Kangkan Ray, Sufal Das

School of Technology, North Eastern Hill University, Shillong

Abstract

A protein that can bind with DNA is called as DNA binding protein. It can also interact with DNA. It has an important and critical role in gene expression and transcription. Therefore, using DNA binding protein we can develop some important drugs which can be used to treat cancers and genetic diseases. So in the domain of molecular biology, it has become a challenging and essential problem for researchers to develop highly accurate and efficient methods for identifying DNA binding protein. Also, the experimental methods are more time consuming and very expensive, hence there is a need for a method based on machine learning. Here, various experiments are demonstrated and analyzed. Finally, we have proposed a method to predict DNA-binding protein using Convolutional Neural Networks. This proposed method takes a 2D PSSM (Position Specific Scoring Matrix) of a protein sequence as input with the dataset that is available in Protein Data Bank. We have attained an accuracy of 97.67% when we have performed our method on the PDB1075 dataset, and it has attained an accuracy of 89.32% on the PDB186 dataset.

Keywords: *Protein, DNA-Biding, PSSM, CNN, Deep Learninig*

1. Introduction

A protein that can bind with DNA is called as DNA binding protein. It consists of different DNA binding domains. The process of transcription is regulated by transcription factors, DNA molecules are split by nucleases and chromosome packing in the cell nuclei are done by histones. DNA-BPs performs two important functions: at first, it arranges and close-packed the DNA then it controls and also influences different cellular processes. DNA binding protein can be used to develop some important drugs which can be used to treat cancers and genetic diseases. So in the domain of molecular biology, it has become a challenging and essential problem for researchers to develop highly accurate and efficient methods for identifying DNA binding protein. Traditionally, this DNA binding protein has been recognized by using various techniques experimentally. These include filter binding assays [2], genetic analysis [3], X-ray crystallography [4], chromatin im-munoprecipitation on microarrays [5], etc. Nowadays computational techniques have been used by researchers for identifying DNA binding protein [6] because these experimental methods are more time consuming and very costly.

2.Related Works

2.1 Prediction of DNA-binding proteins based on revolutionary profiles using SVM. [1]

Introduction : Manish Kumar et al.[1] have put forward a method for identifying DNA-binding protein based on PSSM profiles using SVM and a web server named DNAbinder has been designed[9].

Dataset Description : On DNAaset, which consists of 1153 DNA- binding and 1153 non-DNA-binding proteins an SVM model is designed.

ISSN: 2005-4238 IJAST

Results : Accuracy of 71.59% with Sensitivity = 72.59% , Specificity = 70.59%, MCC = 0.43 are obtained using dipeptide compositions and Accuracy of 72.42% with Sensitivity = 72.51% , Specificity = 72.33%, MCC = 0.45 are obtained using amino acids. When PSSM profiles are used the execution of the SVM model is increased to Accuracy of 74.22% with Sensitivity = 73.53% , Specificity = 74.92%, MCC = 0.49. Also, on DNAset an SVM model has been developed, which comprises 146 numbers of DNA-binding and 250 numbers of non-DNA binding and it attained an Accuracy of 86.62% with Sensitivity = 86.32% , Specificity = 86.80%, MCC = 0.72 using PSSM profiles and Accuracy of 79.80% with Sensitivity = 78.11% , Specificity = 80.80%, MCC = 0.58 using amino acid composition [8].

Advantage : Better performance is obtained by using PSSM profiles.
Disadvantage : Dipeptide gives very poor performance.

2.2 A Framework to identify DNA Binding Protein [9]

Introduction : Xiu-Juan Liu et al.[9] have proposed a model stacking framework to combine and analyze freely-coupled models to identify DNA-binding proteins by MSFBinder. This framework combines ACStruc, 188D, PSSM DWT, and Local DPP feature extraction techniques which are put as an input into random forest and SVM. After that, a logistic regression model was enforced to do the latter prediction.

Dataset Description : Two standard datasets namely PDB186 and PDB1075 are used. The dataset PDB1075 comprises of 525 number of DNA-binding proteins and 550 number of non-DNA-binding proteins, whereas the dataset PDB186 comprises of 93 number of DNA-binding proteins and 93 number of non-DNA-binding proteins.

Results : The above method attained an Accuracy of 83.53% with Sensitivity = 83.81% , Specificity = 83.27%, MCC = 0.6707 when performed on the PDB1075 dataset, and it attained an Accuracy of 81.72% with Sensitivity = 89.25% , Specificity = 74.19%, MCC = 0.6417 on the independent dataset PDB186.

Advantage : Local DPP beats all the other three features with a better performance.
Disadvantage : The attainment of AC struct is poorer than the others.

2.3 Prediction of DNA-binding proteins based on PSSM information. [10]

Introduction : Yubo Wang et al. [10] have used three feature extraction tech-niques NMBAC, PSSM-DCT, and PSSM-DWT to extract the features from the protein sequence. After feature selection, the selected features are put as input to train the SVM (support vector machine).

Dataset Description : Here, PDB186, PDB1075, and PDB594 datasets are used. The PDB1075 dataset comprises of 525 number of DNA-binding proteins and 550 number of non-DNA binding proteins. The PDB594 dataset comprises 297 number of DNA-binding proteins and 297 number of DNA- non-binding proteins. The PDB186 dataset comprises of 93 number of DNA-binding proteins and 93 number of non-DNA binding proteins. Jackknife test is done by using the datasets PDB1075 and PDB594 and an independent test are done by using the dataset PDB186.

Results : The above method attained an Accuracy of 86.23% with Sensitivity = 87.43% , Specificity = 85.09%, MCC = 0.73 when performed on the PDB1075 dataset, and it attained an Accuracy of 76.3% with Sensitivity = 92.5% , Specificity = 60.2%, MCC = 0.557 on the independent dataset PDB186.

Advantage : Better performance is obtained when all the feature extraction techniques are merged.

Disadvantage : PSSM-DCT features are not as effective as the other two features.

2.4 DNA-Binding Protein recognition using Evolutionary detail. [11]

Introduction : Xiangzheng FU et al.[11] put forward a feature extraction method named K-PSSM-Composition. During the process of evolution, these features can preserve the evolutionary detail of a protein. RFE methods are used to find the best features which are put as an input to the support vector machine.

Dataset Description : PDB1075 dataset comprises of 525 number of DNA binding proteins and 550 number of non-DNA binding proteins and PDB186 comprises of 93 number of DNA binding proteins and 93 number of non-DNA binding proteins.

Results : The above method attained an Accuracy of 89.77% with Sensitivity = 90.29% , Specificity = 89.27%, MCC = 0.80 when performed on the PDB1075 dataset, and it attained an Accuracy of 88.71% with Sensitivity = 95.70% , Specificity = 81.72%, MCC = 0.782 on the independent dataset PDB186.

Advantage : It takes less time and space.

Disadvantage : Attainment is not so good.

Feature selection is essential to get good performance.

2.5 PseAAC, a model to recognize DNA-binding protein. [12]

Introduction : M.Saifur Rahman et al.[12] have put forward a method that can pull out relevant information right away from the protein string. After that, the RFE [13] method is used to find the best features which are put as an input to the support vector machine[14].

Dataset Description : Dataset PDB1075 comprises of 525 number of DNA-binding proteins and 550 number of non-DNA binding proteins. For indepen-dent testing, the PDB186 dataset comprises 93 number of DNA binding and 93 number of non-DNA binding proteins is used.

Results : The above method attained an Accuracy of 95.91% with Sensitivity = 94.10% , Specificity = 97.64%, MCC = 0.92 when performed on the PDB1075 dataset, and it attained an Accuracy of 77.42% with Sensitivity = 83.87% , Specificity = 70.97%, MCC = 0.553 on the independent dataset PDB186.

Advantage : Good performance.

Disadvantage : Complex due to the number of feature dimension.

It requires more time and space due to the ensemble classifier.

3. Proposed Method

Through techniques like Random Forest, Support Vector Machine (SVM), and Logistic Regression, the execution of the prediction of DNA binding protein rely on the feature extraction technique. So, we need to perform the feature extraction separately which is time-consuming. Also, we have seen that all these methods do not give better accuracy. Here, we have proposed a technique to predict a DNA binding protein using Convolutional Neural Network (CNN) [15] that will consider all the above discussed issues. The proposed method is shown in Figure 2.

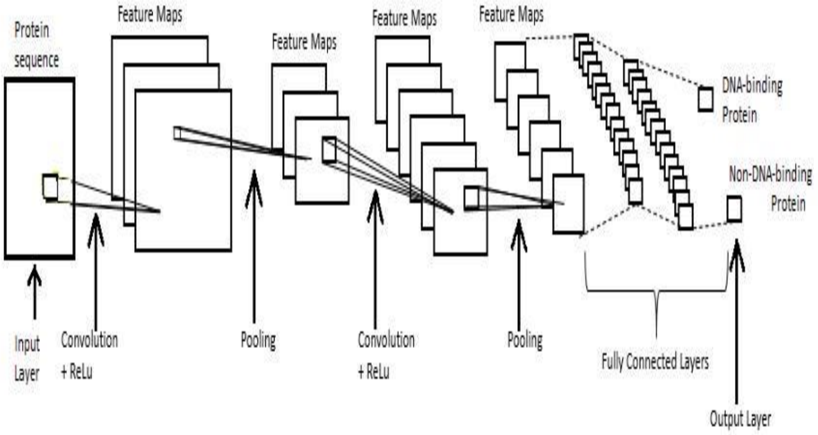


Figure 2. Prediction of DNA-binding protein using CNN

4.Convolutional Neural Network(CNN)

Prediction of DNA-binding protein has been done using different types of classifiers like Support vector machine, [18] Naive Bayes Algorithm, Random Forests, Artificial Neural Networks [17]. In this paper, we have discussed the Convolutional Neural Networks(CNN) [19] approach. CNN is like the Feed Forward Artificial Neural Network. In the adjacent layers, CNN [20] has a local connectivity pattern between the neurons.

4.1 Input Layer

In this layer, a 2D PSSM matrix is given as input.

PSSM PSSM stands for Position Specific Scoring Matrix. This PSSM matrix is generated by PSI-BLAST that preserves the revolutionary detail of a protein. The size of the PSSM of a protein string is M×20 formulated as follows:

$$\text{PSSM} = \begin{bmatrix} f1{,}1 & \cdots & f1{,}20 \\ \vdots & \ddots & \vdots \\ fM, 1 & \cdots & fM, 20 \end{bmatrix}$$

where M is the length of the protein string and f_{xy} is the score of an amino acid that becomes different during the process of evolution from x^{th} location to y^{th} location in the sequence of protein. x = 1,2,. . . ,M and y = 1,2,. . . ,20.

4.2 Convolution Layer

Convolution layer is used for the purpose of the feature extraction. Kernels or filters are used in this layer to produce the output called a feature map. The size of this feature map depends on the number of filters used, stride and padding.

4.3 ReLU

The purpose of ReLU stands for Rectified Linear Unit is to maintain non-linearity in CNN. The output of ReLU is f(y) = max(0,y), where y is the input to the neuron.

4.4 Pooling Layer

This pooling layer is used to reduce the number of parameters or dimensionality of each feature map. It is also called downsampling. There may be max pooling, sum pooling and average pooling.

4.5 Fully Connected Layer

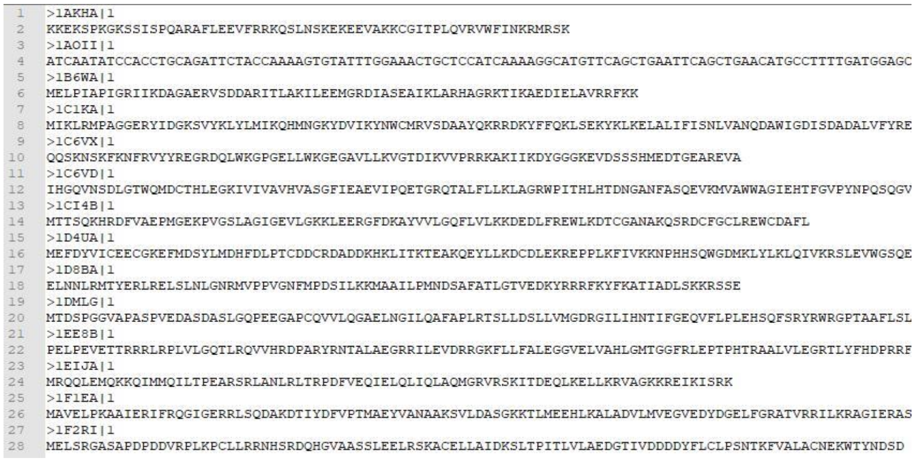
After the Convolution and Pooling layer, the Fully Connecter layer is added to build the CNN model. This FC layer takes the input as a vector.

4.6 Output layer

Finally, an activation function such as softmax or sigmoid is used to classify the output as DNA-binding or non-DNA-binding protein.

5. Implementation and Result Analysis

Input Dataset Here, we have considered two datasets, PDB1075 and PDB186. Dataset PDB1075 comprises of 525 number of DNA-binding proteins and 550 number of non-DNA binding proteins. Whereas the PDB186 dataset comprises 93 number of DNA binding and 93 number of non-DNA binding proteins. A sample of the dataset PDB1075 has been shown in the Figure 3.



```
>1AKHA|1
KKEKSPKGKSSISPQARAFLEEVFRRKQSLNSKEKEEVAKKCGITPLQVRVWFINKRMRSK
>1AOII|1
ATCAATATCCACCTGCAGATTCTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGCATGTTCAGCTGAATTCAGCTGAACATGCCTTTTGATGGAGC
>1B6WA|1
MELPIAPIGRIIKDAGAERVSDDARITLAKILEEMGRDIASEAIKLARHAGRKTIKAEDIELAVRRFKK
>1C1KA|1
MIKLRMPAGGERYIDGKSVYKLYLMIKQHMNGKYDVIKYNWCMRVSDAAYQKRRDKYFFQKLSEKYKLKELALIFISNLVANQDAWIGDISDADALVFYRE
>1C6VX|1
QQSKNSKFKNFRVYYREGRDQLWKGPGELLWKGEGAVLLKVGTDIKVVPRRKAKIIKDYGGGKEVDSSSHMEDTGEAREVA
>1C6VD|1
IHGQVNSDLGTWQMDCTHLEGKIVIVAVHVASGFIEAEVIPQETGRQTALFLLKLAGRWPITHLHTDNGANFASQEVKMVAWWAGIEHTFGVPYNPQSQGV
>1CI4B|1
MTTSQKHRDFVAEPMGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDTCGANAKQSRDCFGCLREWCDAFL
>1D4UA|1
MEFDYVICEECGKEFMDSYLMDHFDLPTCDDCRDADDKHKLITKTEAKQEYLLKDCDLEKREPPLKFIVKKNPHHSQWGDMKLYLKLQIVKRSLEVWGSQE
>1D8BA|1
ELNNLRMTYERLRELSLNLGNRMVPPVGNFMPDSILKKMAAILPMNDSAFATLGTVEDKYRRRFKYFKATIADLSKKRSSE
>1DMLG|1
MTDSPGGVAPASPVEDASDASLGQPEEGAPCQVVLQGAELNGILQAFAPLRTSLLDSLLVMGDRGILIHNTIFGEQVFLPLEHSQFSRYRWRGPTAAFLSL
>1EE8B|1
PELPEVETTRRRLRPLVLGQTLRQVVHRDPARYRNTALAEGRRILEVDRRGKFLLFALEGGVELVAHLGMTGGFRLEPTPHTRAALVLEGRTLYFHDPRRF
>1EIJA|1
MRQQLEMQKKQIMMQILTPEARSRLANLRLTRPDFVEQIELQLIQLAQMGRVRSKITDEQLKELLKRVAGKKREIKISRK
>1F1EA|1
MAVELPKAAIERIFRQGIGERRLSQDAKDTIYDFVPTMAEYVANAAKSVLDASGKKTLMEEHLKALADVLMVEGVEDYDGELFGRATVRRILKRAGIERAS
>1F2RI|1
MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPITLVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWTYNDSD
```

Figure 3. A sample of the dataset PDB1075

Platform

1. Python 3.6
2. Tensorflow 1.8

Result Analysis : Our method has attained an accuracy of 97.67% when performed on the PDB1075 dataset, and it attained an accuracy of 89.32% on the PDB186 dataset.

Comparison with Other Methods : For better comparison a table 1 and two graphs (graph 4 and graph 5) have shown to compare our method with all other methods.

Table 1. Comparison of our method with other methods

SL No.	Methods	Accuracy(%) on PDB1075	Accuracy(%) on PDB186
1	Xiu-Juan Liu	83.53	81.72
2	Yubo Wang	86.23	76.30
3	K-PSSM	89.77	88.71
4	PseAAC	95.91	77.42
5	Our method	97.67	89.32

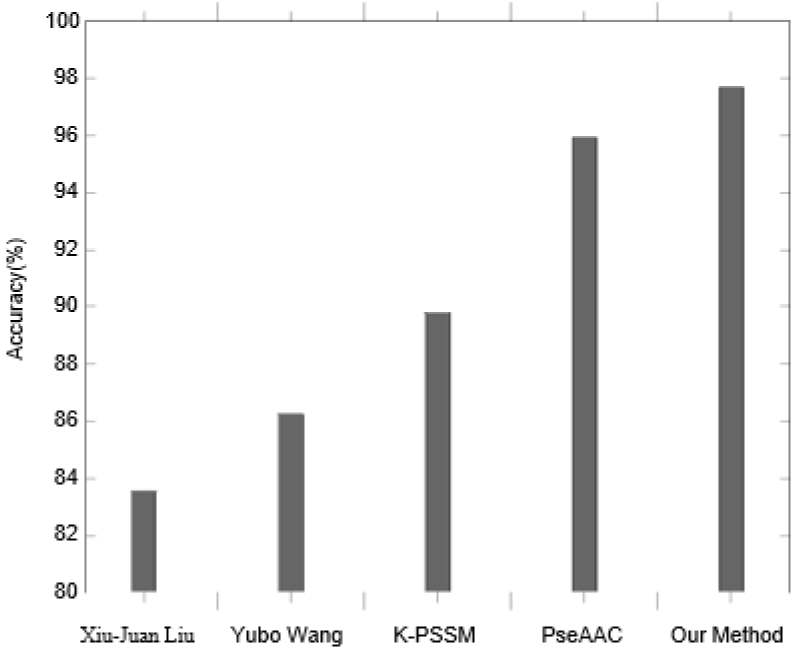


Figure 4. Comparison of Accuracy Value using PDB1075 Dataset

ISSN: 2005-4238 IJAST

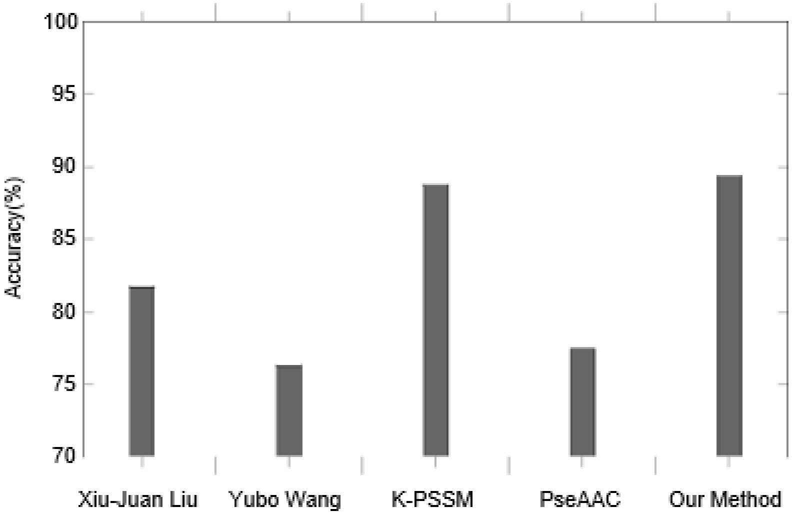


Figure 5. Comparison of Accuracy Value using PDB186 Dataset

References

[1] Kumar, Manish, Michael M. Gromiha, and Gajendra PS Raghava. "Identification of DNA-binding proteins using support vector machines and evolutionary profiles." BMC bioinformatics 8.1 (2007): 463.

[2] Helwa, Reham, and Jorg D. Hoheisel. "Analysis of DNA{protein interactions: from nitrocellulose filter binding assays to microarray studies." Analytical and bioana-lytical chemistry 398.6 (2010): 2551-2561.

[3] Freeman, Katie, Marc Gwadz, and David Shore. "Molecular and genetic analysis of the toxic effect of RAP1 overexpression in yeast." Genetics 141.4 (1995): 1253-1262.

[4] Chou, Chia-Cheng, et al. "Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso10b2 at a resolution of 1.85 Angstroms." Journal of bac-teriology 185.14 (2003): 4066-4073.

[5] Buck, Michael J., and Jason D. Lieb. "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experi-ments." Genomics 83.3 (2004): 349-360.

[6] Zhao, Huiying, Yuedong Yang, and Yaoqi Zhou. "Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural ge-nomics targets." Nucleic acids research 39.8 (2011): 3017-3025.

[7] Liu, Bin, et al. "iDNA-Prot| dis: identifying DNA-binding proteins by incorporat-ing amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition." PloS one 9.9 (2014).

[8] Yu, Bin, et al. "Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising." Journal of Molecular Graphics and Modelling 76 (2017): 260-273.

[9] Liu, Xiu-Juan, et al. "A Model Stacking Framework for Identifying DNA Binding Proteins by Orchestrating Multi-View Features and Classifiers." Genes 9.8 (2018): 394.

[10] Wang, Yubo, et al. "Improved detection of DNA-binding proteins via compression technology on PSSM information." PloS one 12.9 (2017).

[11] Fu, Xiangzheng, et al. "Improved DNA-binding protein identification by incorpo-rating evolutionary information into the Chou's PseAAC." IEEE Access 6 (2018): 66545-66556.

[12] Rahman, M. Saifur, et al. "Dpp-pseaac: A dna-binding protein prediction model using chou's general pseaac." Journal of theoretical biology 452 (2018): 22-34.

[13] Qiu, Wang-Ren, et al. "iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC." Onco-target 7.28 (2016): 44310.

[14] Zou, Chuanxin, Jiayu Gong, and Honglin Li. "An improved sequence based pre-diction protocol for DNA-binding proteins using SVM and comprehensive feature analysis." BMC bioinformatics 14.1 (2013): 90.

[15] Pan, Xiaoyong, et al. "Prediction of RNA-protein sequence and structure bind-ing preferences using deep convolutional and recurrent neural networks." BMC genomics 19.1 (2018): 511.

[16] Leung, Michael KK, et al. "Machine learning in genomic medicine: a review of computational problems and data sets." Proceedings of the IEEE 104.1 (2015): 176-197.

[17] Chen, Chin-Fu, Xin Feng, and Jack Szeto. "Identification of critical genes in mi-croarray experiments by a Neuro-Fuzzy approach." Computational biology and chemistry 30.5 (2006): 372-381.

[18] Liu, Bingqiang, et al. "Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies." IEEE/ACM transactions on computational biology and bioinformatics 16.4 (2018): 1211-1218.

[19] Siar, Masoumeh, and Mohammad Teshnehlab. "Brain Tumor Detection Using Deep Neural Network and Machine Learning Algorithm." 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2019.

[20] Li, Qing, et al. "Medical image classification with convolutional neural network." 2014 13th international conference on control automation robotics and vision (ICARCV). IEEE, 2014.

Authors

Kangkan Roy is doing Master in Technology from North-Eastern Hill University. He obtained bachelor degree in technology from the same university. His area of research work is Bio-informatics, Data Mining.

Dr. Sufal Das is currently working as Assistant Professor in North-Eastern Hill University since 2010. He obtained bachelor degree in Engineering in 2005 and Master degree in Engineering in 2008. He received Ph.D degree from the NEHU in 2018. His area of research work is Big Data Analysis, Data Mining, Machine Learning. He has published several research articles in journals and conferences.