IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

A Deep Learning Model for Dimension Reduction and Multi-Class Classification of Gene Expression Data

Aradhita Mukherjee[1], Dibyendu Bikash Seal[2*]

[1]Dept. of Computer Science and Engineering, University of Calcutta
[2]A. K. Choudhury School of Information Technology, University of Calcutta

[*]*Corresponding Author: dibs.adi@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 14/Aug/2018, Published: 31/Aug/2018

Abstract — Gene expression analysis has been vital in cancer detection across the world. Genes regulating cell growth in cancer, suffer altered expressions. This leads to various phenotypic traits. Gene expression profiling has been extensively used by researchers to accurately identify tumours and has thus enabled better understanding of tumour biology. However, feature extraction and classification of gene expression datasets is challenging due to the high dimension of gene expression datasets and the non-linear relationships among the data. In this article, we have developed a deep learning-based dimension reduction and multi-class classification model using deep auto-encoder and multi-layer perceptron (MLP). We have trained the auto-encoder to extract meaningful features from the RNA-Seq data. These features are then used for supervised classification of tumour samples using a multilayer perceptron. Our (deepAE-MLP) model showed better feature extraction and disease classification capabilities when compared to benchmark methods.

Keywords —Gene expression, Deep Learning, Auto-encoder, Multi-layer perceptron, Dimension Reduction, Multi-class Classification

I. INTRODUCTION

Gene expression analysis helps us to characterize cellular states of various diseases. Changes on gene expressions may lead to phenotypic variation. Manifestation on gene expression depends on genetic variants at DNA level. Gene expression analysis can lead to significant discoveries in biological fields. Identification of genes that are differentially expressed or critical for disease pathways, finding regulatory targets and drug development are the main focus of such research [1-7].

However, there are certain challenges associated with these tasks. The "dimensionality curse" (high number of features against very small number of samples) and noise are the two most worrying factors. A lot of research is still required in this direction to improve results obtained from gene expression analysis.

Soft computing and machine learning techniques for interpreting gene expression patterns, selecting critical genes or reducing dimension have achieved some success [8, 9]. K-Means clustering, Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are some techniques used for dimension reduction [10]. Hybrid methods involving genetic algorithms and SVM or neural networks were applied for gene selection and classification [11, 12].

Auto extraction of knowledge and summarization is one of current research need. Many supervised learning algorithms and unsupervised clustering algorithms have been applied for extraction of knowledge from biological data, but their performance mostly depends on the knowledge of biology or identifying the most significant signals in the data. A deep neural network is preferred in this context. Deep learning models claim to automatically learn maximum information from the data in an unsupervised manner, without the need of domain knowledge. Feature extraction architectures like Auto-encoders and Restricted Boltzmann machines have become the de-facto tools for knowledge extraction. The most striking feature of an auto-encoder network is that unlike the other learning techniques, it does not depend on known biology.

In recent years, deep neural network architectures have been used profoundly for supervised and unsupervised learning tasks. Deep multi-layer perceptron models have been trained using auto-encoders for learning a low-dimensional representation of the data. Deep Belief Networks (DBNs) have also been used for this purpose [13, 14]. Variations of auto-encoder network, namely, stacked auto-encoders, de-noising auto-encoders, sparse auto-encoders and variational auto-encoders (VAEs) have been employed to learn better representations of the data [15, 16]. Deep architectures like

© 2018, IJCSE All Rights Reserved

auto-encoder extract features in non-linear space unlike standard dimension reduction techniques like PCA.

In this article, we have introduced a deep learning framework for dimension reduction and classification of gene expression dataset. The deep auto-encoder (deepAE) model has been used to reduce the high dimensional gene expression data to a lower dimensional, more meaningful representation. The new extracted features are then used to classify the tumor samples into one of the five classes. Various other standard machine learning algorithms for dimension reduction and classification are then studied and compared, to test the usefulness of our proposed model.

Rest of the article is organized as follows, Section II contain the related work on deep learning techniques for computational biology, Section III explains the data and the methods used, Section IV discusses the results, and section V concludes the research work with future directions.

II. RELATED WORK

There are multiple approaches for classification and clustering of microarray gene expression data. Support vector machines (SVM) have been used to classify between leukemia, ovarian and colon cancers [17] and breast cancer tissues [18].

To reduce the high dimension of data and extract meaningful features from it, PCA has been the standard tool. It uses an orthogonal transformation to map a set of high-dimensional correlated observations to a set of uncorrelated low-dimensional components [19]. The transformation is such that maximum variability in the data is explained by the first principal component, the next highest variability by the next principal component as so on and so forth.

Deep architectures overcome the limitation of PCA in extracting non-linear relationships from the data [20]. Auto-encoders have been used extensively in recent years to extract meaningful features from data. Gupta et. al. in [21] used de-noising auto-encoders to pre-train deep architectures., which were further used to regenerate gene expression data. De-noising auto-encoders were further used in [22] on breast cancer gene expression dataset to identify features associated with molecular subtypes and estrogen receptor (ER) status. Danaee et. al. employed stacked de-noising auto-encoders on breast cancer dataset and analyzed the auto-encoder matrices to identify highly interactive genes [23].

To analyze the impact of genetic factors on gene expression, Rui Xie et. al. constructed a predictive regression model using stacked de-noising auto-encoders along with multi-layer perceptrons [24]. Variational auto-encoders were used in [25] to extract latent variables from leukemia data. These latent variables were accessed for estimating drug response. Auto-encoders have not only been instrumental in gene expression analysis, but also in multi-omics integration approaches that study the effect of integrating multiple omics data like DNA methylation, RNA-Seq, microRNA-Seq and clinical information [26, 27].

Deep learning architectures have been reviewed in [28] where they focus on two important research areas deep learning and Big Data. They also discuss different challenges of deep learning architectures when working with big data and scope to work on in future like handling high dimensional data, analyzing streaming data, distributed computing, data tagging, information retrieval, selecting criteria for extracting good data representations, etc.
In another review on deep learning in bioinformatics [29], the authors discuss how deep learning breaks the barrier of convention machine learning approaches for problems in the bioinformatics domain. They focus on hyper-parameter tuning of deep architectures for various applications.
Christof Angermueller et. al. in [30] reviewed deep learning architectures for computational biology, regulatory genomics and image analysis. Other review articles on deep learning applications for computational biology, health informatics, biomedicine and big data processing can be found in [31-33], where the authors discuss how the accuracy of decision support systems can be increased using deep learning. They also focus on building robust techniques to integrate massive semi-structured biological data.

III. METHODOLOGY

Data Acquisition and Pre-processing

We have used the RNA-Seq (HiSeq) PANCAN dataset https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq from UCI Machine Learning Repository for building our model. It contains gene expression values of patients suffering from Breast Cancer (BRCA), Kidney renal papillary cell carcinoma (KIRC), Colon adenocarcinoma (COAD), Lung adenocarcinoma (LUAD) and Prostate adenocarcinoma (PRAD). There are 801 samples and 20531 features or attributes in the dataset. It does not contain any missing value. However, some attributes had 0 expression value across all samples.

Attributes having 0 expression values across all samples were removed and the number of features was reduced to 20264 features. To further reduce the impact of erroneous readings, we have removed features (genes) having less than 80% non-zero values. The final reduced gene expression data contain 801 samples and 16479 genes. To normalize and scale the data within [0-1] range, we have used the sklearn.preprocessing.MinMaxScalar (Sckit-learn) package [34].

© 2018, IJCSE All Rights Reserved

Dimension Reduction and Feature Extraction using Deep Auto-encoders

To reduce the dimension of the dataset and also extract good features, we used a deep auto-encoder network. An auto-encoder network is a feed-forward, non-recurrent neural network that employs an encoder function and a decoder function. The aim of an auto-encoder is to learn a lower dimensional representation of a given dataset in an unsupervised fashion. The encoder function $e = enc(x)_i$ and the decoder function $d = dec(e)$ is used to encode the data and reconstruct the original data from the encoded representation respectively. It typically consists of an input layer, one or more hidden layers and an output layer. The output layer produces a reconstruction of the original input. To reduce the input dimension, the number of neurons in the hidden layer is usually kept much lower than that in the input layer.

In our model, the gene expression dataset with 16479 genes forms the input to the auto-encoder. As illustrated in figure 1, the auto-encoder model consists of an input layer, two auto-encoders in between with dimension 500 and 150 respectively and an output layer. We have used the ReLu (Rectified Linear Units) activation function in each layer. The ReLu function solves the vanishing gradient problem [35] that other functions suffer from. Thus, for each layer i, $y = f(x_{tot}) = \max(0, x_{tot})$, where x_{tot} is the weighted sum of the inputs given by: $x_{tot} =$ and b_i is the bias. We have trained the auto-encoder for 30 epochs using 'mse' as the loss function, i.e., the squared difference between the original and the reconstructed input is taken as the reconstruction error and the auto-encoder is trained to minimize this error. For an input x and its reconstruction x, the loss function thus becomes:

$$\text{loss}(x, x') = \|x - x'\|^2 = \left\|x - \sigma'(W'(\sigma(Wx + b)) + b)'\right\|^2 \quad (1)$$

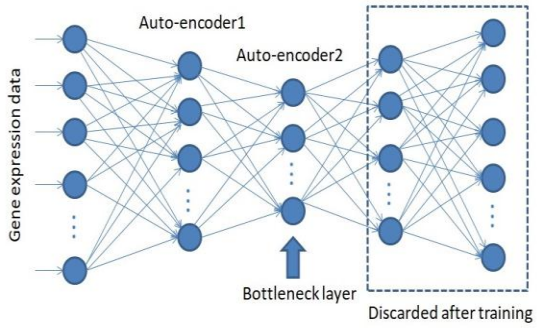

Figure 1. Our deep auto-encoder model for feature extraction and dimension reduction

For our multi-class classification problem, the output layer has been designed to have the same number of neurons as the number of classes in the data. The ReLu activation function has been used in the intermediate layers and a softmax function at the output layer. The MLP has been trained for 30 epochs using the 'categorical crossentropy' as the loss function and a 'adam' optimizer. The overall architecture of our deepAE-MLP model is shown in figure 2.

Multi-class classification using Multi-layer Perceptron

After training, we have extracted the reduced feature set from the bottleneck layer of the auto-encoder and used it as the input to the Multi-layer perceptron (MLP). An MLP is a feed-forward artificial neural network used for supervised learning. It consists of an input layer, one or more hidden layers and an output layer. Neurons in each layer are connected to all neurons in the next layer. Given, a set of features $X = \{x_1, x_2, \ldots, x_n\}$ and a target Y, an MLP can be trained to learn a non-linear estimator for a classification problem.

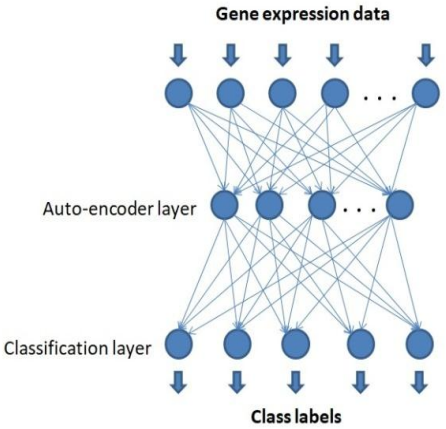

Figure 2. Our deepAE-MLP model

Other methods for comparison

To compare our results, we have used a few benchmark methods for feature extraction and classification. We have the performed the same experiments using the Principal Component Analysis algorithm (PCA) and Kernel PCA for dimension reduction. PCA has been the benchmark tool for dimension reduction or feature extraction since it allows transforming data to a much low-dimensional space, with the first few principal components explaining the overall variance of the data. However, PCA is unable to exploit the non-linear relationships between the data [21]. This is

© 2018, IJCSE All Rights Reserved

overcome by Kernel PCA that uses the kernel functions to find new directions of variance [36, 37].

For classification, we have evaluated our results by comparing our deepAE-MLP model with other state-of-the-art classifiers like SVM, Decision Tree and Naïve Bayesian classifier. A support vector machine can be used for a classification or regression task. It separates data from different classes by constructing a hyperplane or a set of hyperplanes in a very high-dimensional space. The best separation is achieved by the hyperplane which has the largest distance to the nearest training-data point of any class (margin).

A Decision tree classifier, on the other hand, performs classification by posing a set of questions to the data. Based on the question related to its attributes, a split is made at the root and each of the internal nodes, until a pure partition is obtained. Each internal node of the tree represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. A Naive Bayesian classifier works on the principle of Bayes' theorem. It is called 'Naive' since it assumes that every pair of feature is independent of each other.

IV. Results and Discussion

Our proposed method deepAE-MLP is first compared with PCA-MLP and KPCA-MLP to evaluate the effectiveness of our dimension reduction method. As can be seen in table 1, deepAE-MLP outperformed both PCA-MLP and KPCA-MLP by a large margin. This establishes the strength of our auto-encoder based dimension reduction model.

Table 1. Comparison with other standard methods

Method	Accuracy%
PCA + MLP	84.90
KPCA + MLP	85.63
deepAE + SVM	88.76
deepAE + DecisionTree	87.65
deepAE + NB	89.18
deepAE + MLP	**99.37**

We have also tried to find out the minimum number of features that produce the optimal results. The plot shown in figure 3 shows that using 150 features was sufficient enough for our deepAE-MLP model to classify all cancer samples accurately.

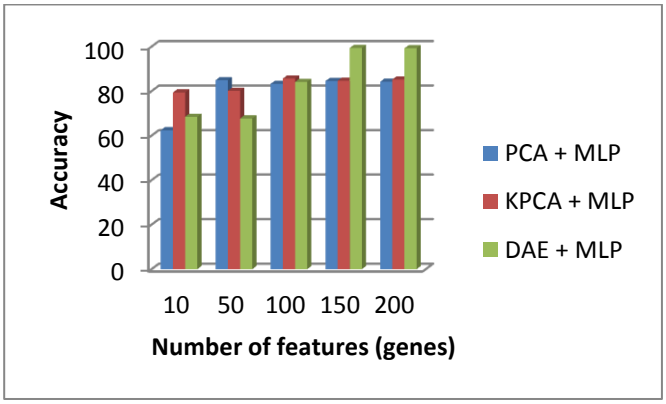

Figure 3. Finding the minimum number of features necessary for classification

To evaluate the potency of our overall deep learning model, we have compared our results with those from three other methods that use SVM, Decision Tree and Naïve Bayes' for classification. Results from table 1 once more establish that our deep learning based model outperformed all other methods, the optimal accuracy being 99.37 % as shown in table 2.

Table 2. The optimal results for our deepAE-MLP model

Accuracy	**0.9937**			
Class	**Precision**	**Recall**	**F1-score**	**support**
0	1.00	1.00	1.00	59
1	0.95	1.00	0.97	18
2	1.00	1.00	1.00	18
3	1.00	0.96	0.98	26
4	1.00	1.00	1.00	40
avg/total	**0.99**	**0.99**	**0.99**	**161**

V. Conclusion and Future Scope

In this article, we have proposed a deep auto-encoder architecture for feature extraction and dimension reduction. We then used a multi-layer perceptron for the classification of tumour samples. The auto-encoder reduced features are found to be useful enough to correctly classify cancer samples into one of the five classes. Results show that our deep learning-based model produces an accuracy of 99.37 % which is more than several state-of-the-art methods used for dimension reduction and classification.

Future work may consider identifying genes that are potential biomarkers for a particular type of cancer. This may be done by analyzing the weight matrices in each layer of the auto-encoder network and ranking genes according to their level of relevance.

© 2018, IJCSE All Rights Reserved

REFERENCES

[1] Creighton CJ et al., "Comprehensive molecular characterization of clear cell renal cell carcinoma", Nature 499(7456):43-9 (2013)

[2] H. Li, B. Yu, J. Li, L. Su, M. Yan, J. Zhang, C. Li, Z. Zhu and B. Liu, "Characterization of differentially expressed genes involved in pathways associated with gastric cancer", PloS one 10, p. e0125013

[3] (2015).

[4] T. Zhou, Y. Du and T. Wei, "Transcriptomic analysis of human breast cancer cells reveals differentially expressed genes and related cellular functions and pathways in response to gold nanorods", Biophysics Reports 1, 106 (2015)

[5] J. S. Myers, A. K. von Lersner, C. J. Robbins and Q.-X. A. Sang, "Differentially Expressed Genes and Signature Pathways of Human Prostate Cancer", PloS one 10, p. e0145322 (2015)

[6] M. Maienschein-Cline, J. Zhou, K. P. White, R. Sciammas and A. R. Dinner, "Discovering transcription factor regulatory targets using gene expression and binding data", Bioinformatics 28, 206 (2012)

[7] K. Shabana, K. A. Nazeer, M. Pradhan and M. Palakal, "A computational method for drug repositioning using publicly available gene expression data", BMC bioinformatics 16, p. 1 (2015)

[8] Yoo, C.K., Leeb, I. and Vanrolleghema, P.A. (2005) "Interpreting patterns and analysis of acute leukemia gene expression data by multivariate fuzzy statistical analysis", Computers & Chemical Engineering, Vol. 29, No. 6, pp.1345–1356.

[9] Liao, C., Li, S. and Luo, Z. (2006) "Gene selection using Wilcoxon rank sum test and support vector machine for cancer classification", Proceedings of the International Conference on Computational Intelligence and Security, 3–6 November, Guangzhou, China, pp.57–66.

[10] Peterson, L.E. and Coleman, M.A. (2005) "Comparison of gene identification based on artificial neural network pre-processing with k-means cluster and principal component analysis", Proceedings of the 6th Conference Workshop on Fuzzy Logic and Applications, 15–17 September, Crema, Italy, 267–276.

[11] Huerta, E.B., Duval, B. and Hao, J.K. (2006) "A hybrid GA/SVM approach for gene selection and classification of microarray data", Proceedings of the EvoWorkshops 2006: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoINTERACTION, EvoMUSART, and EvoSTOC, 10–12 April, Budapest, Hungary, pp.34–44.

[12] Baena, R.M.L., Urda, D., Subirats, J.L., Franco, L. and Jerez, J.M. (2013) "Analysis of cancer microarray data using constructive neural networks and genetic algorithms", Proceedings of the 1st International Work-Conference on Bioinformatics and Biomedical Engineering, 18–20 March, Granada, Spain, pp.55–63.

[13] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets.". Neural computation 18, no. 7 (2006): 1527-1554.

[14] Wang, Haohan, and Bhiksha Raj. "A Survey: Time Travel in Deep Learning Space: An Introduction to Deep Learning Models and How Deep Learning Models Evolved from the Initial Ideas." arXiv preprint arXiv:1510.04781 (2015).

[15] Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and composing robust features with de-noising autoencoders." In Proceedin7,s of the 25th international conference on Machine learning, pp. 1096-1103. ACM, 2008.

[16] Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Y oshua Bengio, and Pierre-Antoine Manzagol. "Stacked de-noising autoencoders: Learning useful representations in a deep network with a local de-noising criterion." The Journal of Machine Learning Research II (20 I 0): 3371- 3408.

[17] T. S. Furey, N. Cristianini, N. Du y, D. W. Bednarski, M. Schummer and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data", Bioinformatics 16, 906 (2000).

[18] S. Reddy, K. T. Reddy, V. V. Kumari and K. V. Varma, "An SVM Based Approach to Breast Cancer Classification using RBF and Polynomial Kernel Functions with Varying Arguments", International Journal of Computer Science and Information Technologies 5, 5901 (2014).

[19] S. Wold, K. Esbensen and P. Geladi, "Chemometrics and intelligent laboratory systems", 2, 37 (1987).

[20] Fakoor R, Ladhak F, Nazi A, Huber M, editors, "Using deep learning to enhance cancer diagnosis and classification", Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare Atlanta, Georgia: JMLR: W&CP; 2013.

[21] A. Gupta, H. Wang and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering", in Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, 2015.

[22] Jie Tan, MAtthew Ung, Chao Cheng and Casey S Greene, "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with de-noising autoencoders", Pac Symp Biocomput. 2015; 20: 132–143.

[23] Padideh Danaee, Reza Ghaeini, and David A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification", Pac Symp Biocomput. 2016; 22: 219–229.

[24] Rui Xie, JiaWen, Andrew Quitadamo, Jianlin Cheng and Xinghua Shi, "A deep auto-encoder model for gene expression prediction", BMC Genomics 2017, 18(Suppl 9):845 DOI 10.1186/s12864-017-4226-0

[25] Ayse Dincer, Safiye Celik, Naozumi Hiranuma, and Su-In Lee, "DeepProfile: Deep learning of patient molecular profiles for precision medicine in acute myeloid leukemia", bioRxiv preprint first posted online Mar. 8, 2018; doi: http://dx.doi.org/10.1101/278739.

[26] Kumardeep Chaudhary, Olivier B Poirion, Liangqun Lu, Lana X Garmire, "Deep Learning based multi-omics integration robustly predicts survival in liver cancer", Clinical Cancer Research, Pages clincanres. 0853.2017

[27] Fadhl M. Alkawaa, Kumardeep Chaudhary, and Lana X. Garmire, "Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data", J. Proteome Res., DOI: 10.1021/acs.jproteome.7b00595, Publication Date (Web): 07 Nov 2017

[28] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall WaldEmail author and Edin Muharemagic, "Deep learning applications and challenges in big data analytics", Journal of Big Data20152:1 https://doi.org/10.1186/s40537-014-0007-7

[29] S Min, B Lee, S Yoon, "Deep learning in bioinformatics", Briefings in bioinformatics 18 (5), 851-869

[30] Angermueller C, Pärnamaa T, Parts L, Stegle O, "Deep learning for computational biology", Mol Syst Biol. 2016 Jul 29;12(7):878. doi: 10.15252/msb.20156651.

[31] Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ, "Deep Learning for Health Informatics", IEEE J Biomed Health Inform. 2017 Jan;21(1):4-21. doi: 10.1109/JBHI.2016.2636665. Epub 2016 Dec 29.

[32] Mamoshina P, Vieira A, Putin E, Zhavoronkov A, "Applications of Deep Learning in Biomedicine", Mol Pharm. 2016 May

© 2018, IJCSE All Rights Reserved

2;13(5):1445-54. doi: 10.1021/acs.molpharmaceut.5b00982. Epub 2016 Mar 29.

[33] Imad, Hafidi & Rochd, Yassir. (2018). An Enhanced Apriori Algorithm Using Hybrid Data Layout Based on Hadoop for Big Data Processing. International Journal of Network Security. 18. 161.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825{2830, 2011.

[35] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol. 6, pp. 107{116, Apr. 1998.

[36] B. Scholkopf, A. Smola and K.-R. M□uller, "Kernel principal component analysis", in International Conference on Articial Neural Networks, 1997.

[37] T. SenthilSelvi1 and R. Parimala, "Improving Clustering Accuracy using Feature Extraction Method", International Journal of Scientific Research in Computer Science and Engineering, Vol. 6, Issue. 2, pp.15-19, 2018.

Authors Profile

Mrs. Aradhita Mukherjee pursued Bachelor of Science (Computer Science Honours) from University of Calcutta in 2009, Master of Computer Technology from Burdwan University in 2011 and Master of Computer Science and Engineering from University of Calcutta in 2013. She is currently a research scholar in the department of Computer Science and Engineering in University of Calcutta, under the Visvesvaraya PhD Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India. Her research interest includes Big Data Technologies and Machine Learning.

Mr. Dibyendu Bikash Seal pursued Bachelor of Science (Computer Science Honours) in 2008, Master of Computer Application from in 2011 and Master of Computer Science and Engineering in 2015, all from University of Calcutta. He is currently working as an Assistant Professor in A. K. Choudhury School of Information Technology in the University of Calcutta. Prior to that, he has worked as a Systems Engineer in TATA Consultancy Services Limited and as a faculty of Computer Science in various colleges and universities. His research interest includes Computational Biology and Machine Learning.

© 2018, IJCSE All Rights Reserved