International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Context Free Enciphering Technique for Protein Classification

Anju G S[1], Jyolsna Mary P[2]

[1]M. Tech Scholar, Department of Computer Science and Engineering, MBCCET, Idukki 695015, India

[2]Assistant Professor, Department of Computer Science and Engineering, MBCCET, Idukki 695015, India

Abstract: ***In bioinformatics, there is a profound scale of DNA and protein sequences available, but far from being fully utilized. Computational models can facilitate the analyses of large-scale data but require a numeric representation as input. Feature engineering aims at representing non-numeric data with numeric features and can help design features to cast the raw symbolic data effectively. Automated feature engineering, i.e., an encoding scheme preprogrammed the establishment of features, saves the redesigning process and allows the researchers to try different representations with minimal effort. Here an encoding scheme for protein sequences, which encodes the representative sequence dataset into a numeric matrix that can be fed into a downstream learning model. The method, Context-Free Enciphering Technique practice was preferred for a dataset with group of protein sequences. Compared with the traditional methods using task-specific designed features, this method improves the predicting accuracy and serve as an automated feature engineering method for protein sequences.***

Keywords: Enciphering technique, Data characterization, Feature engineering, Machine learning

1. Introduction

The main objective of feature engineering is to maintain the features of non-numeric data. For performing deep learning we have to make equal length input in order to make a non-numeric data in to numeric data. The learning methods can affect the standard and performance of data representation. Therefore, research in different areas (text mining, bioinformatics, speech recognition, etc) are helpful for designing an effective data representation carry and enhance succeeding learning technique regrettably features normally changes from problems. Particularly when master information is included in the plan, making them just helpful with regards to explicit assignments and models. In the period of big data, mechanized element building that is less delicate to the setting is progressively wanted. Studies have legitimized that information gained starting with one undertaking can be applied then onto the next by means of move learning. A decent portrayal that can profit largescale learning ought to keep up the inborn structure of information, be task vague yet keep the most applicable data about the job needing to be done.

In bioinformatics, the high-throughput sequencing methods have made scads of DNA and protein successions accessible, propelling the field into another time of enormous information. Be that as it may interpreting the groupings, for example to reveal the relationship among genotype and phenotype, stays a test. As a supplement to the expensive wet-lab tests, computational models give an elective point of view to unravel the shrouded designs in protein successions. Throwing the representative grouping dataset into a numeric portrayal generally serves as the upstream stage in a pipeline for examinations. Master information about the issue may encourage the investigations, yet this process is additionally dreary and restricted by human subjectivity. An encoding plan of protein groupings catching nonexclusive priors of amino acids are bound to be liberated from the unique situation (i.e., the errand, information, and model) so that the bioinformaticians can spare the exertion of structuring highlights for various issues.

Protein categorizing is a basic issue in examine the protein sequences, mention to multifaceted tasks. Proteins can be categorized in many ways and many levels of hierarchy [11]. Categorizing a protein series into a well-defined group is a introductory but non-insignificant examination, helpful for interpret its properties. For classifying many important functions or phenotype, usually exploratory evaluation is designed to check the properties of a selected entity [12]. Like, the hemagglutination inhibition (HI) evaluation is outlined to measure the antigenic sameness between the hemagglutinin proteins [13].

2. Related Works

Different methods of classification of proteins based on amino acid compositions were used timely. To anticipate enzyme subfamily classes, “amphiphilic pseudo amino acid compositions” were presented. In this strategy protein is represented by a discrete form, includes a lot of discrete numbers or a various measurement vector. The major lead behind using this form of representation is that, it is easy to be used in statistical prediction but it is difficult to include the sequence-order information. Usually amino acid composition of proteins does not comprise of its sequence-order information; hence the classification method can be improved by incorporating the sequence-order values into the predicator. Accordingly, the representation of protein samples by a set of discrete number, the so-called discrete form is required to frame an attainable predicator. The ‘amphiphilic pseudo amino acid composition’, viably ideal with both the perspective to represent the protein samples. It contains $20+2\lambda$ discrete numbers, where the initial twenty numbers are the part of regular amino acid composition and next 2λ are a lot of sequences correlation factors. The sequence relationship factor has different position of

coupling through the hydrophilicity and hydrophobicity of amino acid constituents along the sequence of proteins. The use of co-variant discriminant algorithm, depend on this representation scheme is astoundingly predominant by the achievement rates in recognizing the 16 subfamily classes of oxido reductase. With protein groupings going into databanks at a touchy pace, the early assurance of the family or subfamily class for a recently discovered chemical atom becomes significant on the grounds that this is legitimately identified with the itemized data about which explicit objective it follows up on, just as to its reactant procedure and organic capacity.

Tragically, it is both tedious and expensive to do as such by tests alone. In a past report, the covariant-discriminant calculation was acquainted with recognize the 16 subfamily classes of oxidoreductases. In spite of the fact that the outcomes were very reassuring, the whole forecast process depended on the amino corrosive structure alone without including any arrangement request data. Along these lines, it is deserving of further investigation.t the game plan of hydrophobicity and hydrophilicity of the amino corrosive deposits along a protein chain plays a important job in its collapsing, just as its association with different particles and synergist systems, and that various sorts of proteins will have diverse amphiphilic highlights, comparing to various hydrophobic and hydrophilic succession request designs [14].

The quantity of constituent particles in an amino acid sequence is its atomic composition. These atoms of amino acids are responsible for deciding the properties of amino acid and amino acid composition, which is a strong parameter for prediction of localization. The distribution of amino acid composition is disclosed in detail, by dividing it into three subregions, namely N-terminal, middle region and C-terminal. The length of every subregion is determined depend on the length of the amino acid sequences and every one of this length is equivalent. For every protein, the SVM module has a component vector of size 45. The physiochemical parameters from AAIndex database are utilized for feature distillation in physiochemical SVM module.

It considers the three components of amino acid composition distinctly. It brings out the compositional difference in each part of the sequence and discloses the distribution of amino acid composition. When the protein sequence p is split into three segments, Pn is the N-terminal segment, Pm is the middle segment, and Pc is the C-terminal segment whose length is the same and amino acid composition of each segment is calculated separately. There are four SVM modules to be specific, ATC-SVM, Phys-SVM, AAC-SVM, 3-AAC-SVM were intended for atomic composition, numerous physiochemical properties, amino acid composition, and three-section amino acid compositions. To make the last forecast from these discrete SVM modules, a voting framework has utilized. Cell is the essential unit of life and proteins are the work ponies in the cell. For a protein to play out its capacity, it ought to be situated in its focused-on cell area.

Data about a protein's area in the cell gives knowledge into the capacity of the protein and is valuable in, screening contender for medicate revelation and immunization structure, commenting on of quality items and, choosing significant proteins for further investigations. Computational subcellular restriction expectation strategies manage foreseeing the area of the protein from its amino corrosive groupings. The achievement of computational subcellular limitation expectation depends on two significant segments. First is the extraction of organic highlights which are applicable in the subcellular restriction and the second is the computational strategy utilized for making expectation. The higher exactness displayed by Smith–Waterman calculation can be a direct result of its capacity for discovering the neighbourhood arrangement inside the protein groupings, bringing out regular examples and areas inside the groupings. The location signals for every area share regular highlights and this calculation is equipped to identify the signs present. another boundary, nuclear organization for subcellular restriction forecast and viably coordinated it with different boundaries like amino corrosive piece, physiochemical boundaries and succession similitude. Our outcomes illustrated that the worldwide data of the grouping contributed more to the expectation precision. This is discovered valid on account of physiochemical properties and grouping arrangement modules. Another perception is that considering the full arrangement as a gathering of three sections, N-terminal, centre locale and C-terminal, will bring out hidden property dispersion to a more prominent detail to upgrade the forecast exactness. For grouping arrangement module, the Smith–Waterman calculation for entire succession performs better than Needleman–Wunsch calculation for entire succession. Our work unequivocally exhibits that nuclear arrangement can be effectively. used alongside other worldwide highlights of the succession to improve the precision of subcellular restriction expectation. [15].

Order of proteins successions into practical and basic families dependent on arrangement homology is a focal issue in computational science. Discriminative managed AI approaches give great execution, yet effortlessness and computational effectiveness of preparing and expectation are likewise significant concerns a class of string pieces, called confound parts, for use with help vector machines (SVMs) in a discriminative way to deal with the issue of protein arrangement and remote homology discovery. These pieces measure succession closeness dependent on shared events of fixed-length designs in the information, taking into account transformations between patterns. Thus, the bits give a naturally well motivated approach to look at protein successions without depending on family-based generative models, for example, shrouded Markov models. Register the pieces proficiently utilizing a bungle tree information structure, permitting us to figure the commitments of all examples happening in the information in one pass while navigating the tree.

At the point when utilized with an SVM, the pieces empower quick expectation on test groupings. Probes two benchmark SCOP datasets, where show that the befuddle portion utilized with a SVM classifier performs seriously with cutting edge strategies for homology discovery, especially at the point

when not many preparing models are accessible. Assessment of the most elevated weighted examples learned by the SVM classifier recuperates naturally significant themes in protein families and super families [41].

3. Methods

The block diagram of proposed frame work is shown in the Figure1.This frame work is a chain of process which begins from data recouping. The input data is fed to the enciphering module which convert the non-numeric data into numeric representation without lossing vital information. Then the data is modeled using different learning models. Finally it is fed to the evaluation model.

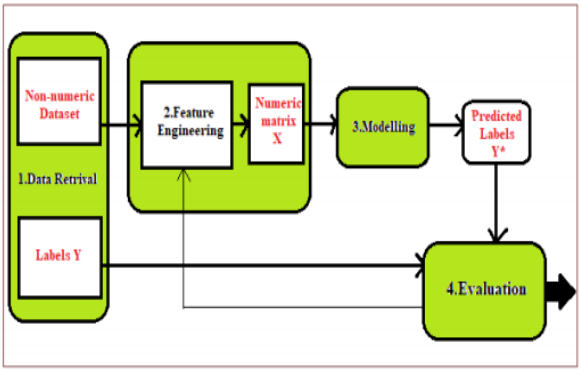

Figure 1: Block diagram of proposed frame work

The encoding plan can be applied to protein groupings with differing lengths which covers the most circumstance of grouping examinations in bioinformatics. The strategy ofenciphering techqniue, depends on the AAindex database [23], which is the assortment of amino corrosive lists and change lattices from distributed work, speaking to physiochemic and biochemical properties related to the explicitness and decent variety of protein structures and capacities. Right now, the AAindex contains 566 amino corrosive files in AAindex1. For encoding protein arrangements with differing length to generally gathering the proteins, the enciphering techqniue encodes the arrangements with AAindex1. In same manner, for describing increasingly unobtrusive differentiations between proteins, the replacement networks are used in enciphering techqniue.

When taking an arrangement cluster **S** of *m* successions with differing lengths as the information, enciphering techqniue encodes each succession $s_i$utilizing *k* amino corrosive records in AAindex1, which speak to conventional physicochemical and biochemical properties, α-helix, β-strand and turn penchants of amino acids. For the arrangement $s_i$encoded by record *j*, the yielding numeric vector is signified as s_i^j . The normal worth n_{ij} is determined, speaking to the estimation of $s_i$with the property *j*. After encoded by the *k* records, the arrangement $s_i$is spoken to by a vector n_i= [n_{i1}, n_{i2}, ...n_{i3}, ...n_{ik}]. Subsequent tostacking the vectors for *m* groupings, the emblematic dataset is encoded by the numeric network X with measurement *m*×*k*.The steps to perform enciphering technique is shown below:

1. **function** *Enciher(s, idL)* >Input: s is a protein sequence; idL is a list with k index IDs
2. **declare***n = { }* > n is a dictionary with keys as the IDs of amino acid indexes for enciphering
3: **for***id in idL***do**
4: n_s = [ ]
5: **for***j = 1 to len(s)***do**
6: n_s . *append(idx.get(s[j]))*>Get the score of each residue *s[j]* from the amino acid index id
7 *nid* = n_s . *mean()*
8: *n[id] = nid*
9: **return***n*

The set of data for protein classification are Identifying Antimicrobial Peptides (iAMP), TumorHPD (Tumor Homing Peptides), HemoPI (Hemolytic Peptide Identification), PVPred (Predicts the Phage Virion Proteins). Identifying antimicrobial peptides, incorporates antibacterial peptides, antiviral peptides, and antifungal peptides. The antimicrobial peptides are significant host guard atoms. TumorHPD is a web server for perceiving tumor homing peptides, which can perceive tumor cells. hemolytic peptide identification, is to screen hemolytic peptides from the non-hemolytic, where quantitative frameworks are produced for estimating the hemotoxicity. PVPred previses the phage virion proteins by investigating the fluctuation and improving the g-hole dipeptide

4. Result and Discussions

To test the viability of enciphering technique on throwing the protein successions to numeric portrayals, we directed protein characterization under various situations. The characterization results are contrasted and other conventional strategies utilizing high quality highlights exceptionally intended for each dataset. Likewise, we contrasted the enciphering technique and a best in class protein order technique named m-NGSG, which is propelled by normal language handling [21] in the natural resistant framework against microbes.

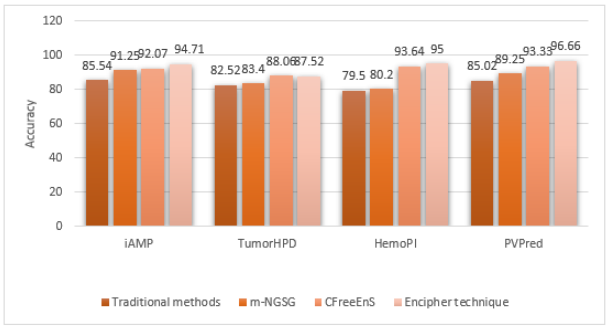

Figure 2: Comparison of accuracy different methods

An outline of the datasets for protein characterization using NN (Neural Network), Random Forest (RF), K-Nearest Neighbour (K-NN), SVM is shown in Table 1.

Table 1: Classification result of enciphering technique

Dataset	*K-NN*	*NN*	*SVM*	*RF*
PVPred	0.9	0.8	0.6333	**0.9666**
TumorHPD	0.7897	0.7761	0.7388	**0.8752**
iAMP	0.8348	0.8160	0.7368	**0.9471**
Hemo PI	0.9227	0.9363	0.8954	**0.95**

The four datasets are encoded by enciphering technique, utilizing all the accessible 566 amino corrosive files in the AA index database. Successions with fluctuating length are spoken to by vectors with length 566. Segments with high relationship are dropped before contributing into a downstream learning technique. To analyze the viability of information portrayal, we keep a similar downstream learning system as those conventional strategies utilizing planned highlights for each dataset. Plus, the m-NGSG, a cutting-edge technique rewarding protein successions as typical content and producing highlights from a book mining point of view, has been applied to the four datasets [21].

5. Conclusion

This paper gives a brief idea about various portrayals of protein successions may unravel or conceal various angles. An encoding plan catching the known nonexclusive properties of amino acids can help computerize the way toward developing highlights and encourage explaining protein capacities. For profiling different parts of proteins, for example foreseeing the protein folds, computational forecasts utilizing other novel portrayals may give more information. We done modelling over four datasets RF gives highest accuracy.

References

[1] Xinrui Zhou, Rui Yin, Jie Zheng and Chee-Keong Kwoh, "An Encoding Scheme Capturing Generic Priors and Properties of Amino Acids Improves Protein Classification, " IEEE, 2169-3536, 2018.
[2] Y. Bengio, A. Courville, and P. Vincent, ‘‘Representation learning: A review and new perspectives,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[3] S.-Y. Kung and M.-W. Mak, ‘‘Feature selection for genomic and proteomic data mining,’’ in Machine Learning in Bioinformatics. Hoboken, NJ, USA: Wiley,pp. 1–45, 2009.
[4] F. Seide, G. Li, X. Chen, and D. Yu, ‘‘Feature engineering in contextdependent deep neural networks for conversational speech transcription,’’ in Proc. IEEE Workshop Autom. Speech Recognit. Understand, pp. 24–29 (ASRU), Dec. 2011.
[5] A. S. Martinez-Vernon et al., ‘‘An improved machine learning pipeline for urinary volatiles disease detection: Diagnosing diabetes,’’ PLoS ONE, vol. 13, no. 9, p. e0204425, 2018.
[6] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, ‘‘Bilevel feature extraction-based text mining for fault diagnosis of railway systems,’’ IEEE Trans. Intell. Transp. Syst., vol. 18, no. 1, pp. 49–58, Jan. 2017.
[7] A. Zheng and A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Sebastopol, CA, USA: O’Reilly Media, 2018.
[8] S. J. Pan and Q. Yang, ‘‘A survey on transfer learning,’’ IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[9] N. Zou, Y. Zhu, J. Zhu, M. Baydogan, W. Wang, and J. Li, ‘‘A transfer learning approach for predictive modeling of degenerate biological systems,’’ Technometrics, vol. 57, no. 3, pp. 362–373, 2015.
[10] P. Domingos, ‘‘A few useful things to know about machine learning,’’ Commun. ACM, vol. 55, no. 10, pp. 78–87, 2012.
[11] D. Petrey and B. Honig, ‘‘Is protein classification necessary? Toward alternative approaches to function annotation,’’ Current Opinion Structural Biol., vol. 19, no. 3, pp. 363–368, 2009.
[12] C. B. Kennedy, ‘‘Multiplexed microfluidic devices and systems,’’ U.S. Patent 6 086 740 A, Jul. 11, 2000.
[13] G. K. Hirst, ‘‘The quantitative determination of influenza virus and antibodies by means of red cell agglutination,’’ J. Exp. Med., vol. 75, no. 1, pp. 49–64, 1942.
[14] M. H. Smith, ‘‘The amino acid composition of proteins,’’ J. Theor. Biol., vol. 13, pp. 261–282, Dec. 1966.
[15] K.-C. Chou, ‘‘Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes,’’ Bioinformatics, vol. 21, no. 1, pp. 10–19, 2004.
[16] B. S. Cherian and A. S. Nair, ‘‘Protein location prediction using atomic composition and global features of the amino acid sequence,’’ Biochem. Biophys. Res. Commun., vol. 391, no. 4, pp. 1670–1674, 2010.
[17] K. Blekas, D. I. Fotiadis, and A. Likas, ‘‘Motif-based protein sequence classification using neural networks,’’ J. Comput. Biol., vol. 12, no. 1, pp. 64–82, 2005.
[18] Ben-Hur and D. Brutlag, ‘‘Sequence motifs: highly predictive features of protein function,’’ in Feature Extraction. Berlin, Germany: Springer,pp. 625–645, 2006.
[19] I. Vujaklija, A. Bielen, T. Paradžik, S. Biin, P. Goldstein, and D. Vujaklija, ‘‘An effective approach for annotation of protein families with low sequence similarity and conserved motifs: Identifying gdsl hydrolases across the plant kingdom,’’ BMC Bioinf., vol. 17, no. 1, p. 91, 2016.
[20] E. Asgari and M. R. K. Mofrad, ‘‘Continuous distributed representation of biological sequences for deep proteomics and genomics,’’ PLoS ONE, vol. 10, no. 11, p. e0141287, 2015.
[21] S. M. A. Islam, B. J. Heil, C. M. Kearney, and E. J. Baker, ‘‘Protein classification using modified n-grams and skip-grams,’’ Bioinformatics, vol. 34, no. 9, pp. 1481–1487, 2017.
[22] X. Zhou, R. Yin, C.-K. Kwoh, and J. Zheng, ‘‘A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses,’’ BMC Genomics, vol. 19, no. 10, p. 936, 2018.
[23] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, ‘‘AAindex: Amino acid index database, progress report 2008,’’ Nucleic Acids Res., vol. 36, pp. D202–D205, Jan. 2007.
[24] K. Tomii and M. Kanehisa, ‘‘Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins,’’ Protein Eng., Des. Select., vol. 9, no. 1, pp. 27–36, 1996.
[25] P. Bhadra, J. Yan, J. Li, S. Fong, and S. W. I. Siu, ‘‘AmPEP: Sequencebased prediction of antimicrobial peptides using distribution patterns of amino acid

properties and random forest,'' Sci. Rep., vol. 8, no. 1, 2018, Art. no. 1697.
[26] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, ''iAMP-2L: A twolevel multi-label classifier for identifying antimicrobial peptides and their functional types,'' Anal. Biochem., vol. 436, no. 2, pp. 168–177, 2013.
[27] P. K. Meher, T. K. Sahu, V. Saini, and A. R. Rao, ''Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC,'' Sci. Rep., vol. 7, p. 42362, Feb. 2017.
[28] A. Sharma et al., ''Computational approach for designing tumor homing peptides,'' Sci. Rep., vol. 3, p. 1607, Apr. 2013.
[29] K. Chaudhary et al., ''A Web server and mobile app for computing hemolytic potency of peptides,'' Sci. Rep., vol. 6, Mar. 2016, Art. no. 22843.
[30] H. Ding, P.-M. Feng, W. Chen, and H. Lin, ''Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis,'' Mol. BioSyst., vol. 10, no. 8, pp. 2229–2235, 2014.
[31] R. B. Squires et al., ''Influenza research database: An integrated bioinformatics resource for influenza research and surveillance,'' Influenza Other Respiratory Viruses, vol. 6, no. 6, pp. 404–416, 2012.
[32] Y.-C. Liao, M.-S. Lee, C.-Y. Ko, and C. A. Hsiung, ''Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus,'' Bioinformatics, vol. 24, no. 4, pp. 505–512, 2008.
[33] X. Du et al., ''Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation,'' Nature Commun., vol. 3, Art. no. 709, Feb. 2012.
[34] J. Qiu, T. Qiu, Y. Yang, D. Wu, and Z. Cao, ''Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2,'' Sci. Rep., vol. 6, p. 31156, Aug. 2016.
[35] Y. Peng et al., ''A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures,'' Sci. Rep., vol. 7, Art. no. 42051, Feb. 2017.
[36] T. Frickey and A. Lupas, ''CLANS: A Java application for visualizing protein families based on pairwise similarity,'' Bioinformatics, vol. 20, no. 18, pp. 3702–3704, 2004.
[37] J. Wen and Y. Zhang, ''A 2D graphical representation of protein sequence and its numerical characterization,'' Chem. Phys. Lett., vol. 476, nos. 4–6, pp. 281–286, 2009.
[38] M. A. Remita, A. Halioui, A. A. M. Diouara, B. Daigle, G. Kiani, and A. B. Diall**o**, ``A machine learning approach for viral genome classification,"*BMC Bioinf.*, vol. 18, no. 1, p.d208, Apr. 2017
[39] D. Struck, G. Lawyer, A.-M. Ternes, J.-C. Schmit, and D. P. Bercoff, "COMET: Adaptive context-based modeling for ultrafast HIV-1 subtypeidenti_cation," *Nucleic Acids Res.*, vol. 42, no. 18, p. e144, Oct. 2014.
[40] D. Loewenstern, H. Hirsh, P. Yianilos, and M. Noordewier, "DNAsequence classi_cation using compression-based induction," Center DiscreteMath. Theor. Comput. Sci., Rutgers Univ., Piscataway, NJ, USA,Tech. Rep. LCSR-TR-240, 1995.
[41] Christina S. Leslie, Eleazar Eskin, Adiel Cohen1, Jason Weston and William Stafford Noble,"*Mismatch string kernels for discriminative protein classification,"*,2004.

Author Profile

Anju G S received the B. Tech degree in Computer Science and Engineering from Cochin University of Science and Technology in 2016. During 2016-2017, she worked as web development trainee at Tektide Innovations. Currently doing her Master degree in Computer Science and Engineering from APJ Abdul Kalam Technical University at MBC College of Engineering and Technology.