

Adaptation issues of machine learning in safety digitization

Gyana Ranjana Panigrahi¹, Nalini Kanta Barpanda¹, Komma Anitha², Shanti Rathore³,
Preesat Biswas⁴, Prabira Kumar Sethy¹

¹Department of Electronics, Faculty of Engineering, Sambalpur University, Sambalpur, India

²Department of Electronics and Communication Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, India

³Department of Electronics and Telecommunication Engineering, Dr. C. V. Raman University Bilaspur Chhattisgarh, Bilaspur, India

⁴Department of Electronics and Telecommunication Engineering, Government Engineering College Jagdalpur, Jagdalpur, India

Article Info

Article history:

Received Oct 10, 2022

Revised Oct 26, 2022

Accepted Nov 18, 2022

Keywords:

Cyber security

Ethical contemplation

Internet community

Machine learning

ABSTRACT

The internet community is the only set of irreplaceable spaces in today's world and is used by millions for knowledge acquittance via the digital exchange between the landed gentry. The torrent of available e-contents in the Internet community attracts corporates and researchers to find the factual weightage of formed data. It is high time for digital diversification, which is the objective of using various learning-based machine learning (ML) systems for hands-on fortification. The main idea is to make stylistic communication more understandable. Here, the authors try to adapt the factual weightage procedure of formed data through the Internet community using machine learning schemes. Hence, the authors have chosen to emphasize cyber security, which is not well discussed and concerned with ethical contemplation from hackers' forums amidst internet communities. There are disparities in the continual growth of connotations, acronyms, spellings, and even technical jargon, which need periodic re-learning and their prototype implications through the proposed model.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Prabira Kumar Sethy

Department of Electronics, Faculty of Engineering, Sambalpur University

Burla, Odisha, India

Email: prabirsethy.05@gmail.com

1. INTRODUCTION

There must be a public place of interest for knowledge sharing and exchanging digital mediums that can perform by various Internet community tools like chronicles, the web of Internet societies, and different mediums [1], [2]. This interest could yield a torrent of unintellectual data utilizing amorphous stylistic communication commonly available to the public. The security investigators may be ethical white hat hackers assembling Internet communities to harmonize discovery schemes [3]–[5]. Here, in this scheme, specialists find vulnerabilities presented in the security applications and information to the manufacturer. There must be a period of agreement between the parties where the manufacturer can release patches for vulnerable packages. When the troubleshooting period ends, the internet communities will free the patches. The objective of this scheme can define in two different ways. One is to alarm the operators about the label of vulnerabilities that can harm their confidentiality [6]–[9]. The second focuses on organizations finding and releasing new patches that could eliminate the security fails from their packages. Gathering this valuable information in patches could help cyber experts prepare a balanced security framework for public implementation. The current discussion is to use machine-based automation to take out the visions from internet communities like chronicles of hackers' environments, the web of internet societies, and different mediums [10], [11]. These days the internet community seems to be the chief source of knowledge acquittance via digital exchange and social platforms between the landed gentry for different types of intimidations with cyberattacks and their challenges [12]–[14].

The importance of digital trade in online communities is increasing day-by-day. Preconceiving that the internet community is the core part of the programmer cyberspace for digital diversification for various incorporations and scholars. The internet community is one of the valued weaponries for making the digital exchange to understand the existing cyber threats for resolving different security issues [15]. This is open to those with a keen interest in the related area. Various internet mediums; are suitable tools for many experienced people to create money by selling mean products like confidential data, and bank card information. Package vulnerabilities are retailing as stated by regulatory fee body measuring its uniqueness and cruciality that could become available in bootleg emporium known as deep web [16], [17]. It is well known that security specialists are disposed to habit in internet communities by sharing various methodical examinations about package vulnerabilities to make them available in security patches. Further, the non-proprietary cyber security information in internet communities is an important aspect. The free spread of information posted on digital forums and broad reach is a treasured medium for electronic media exchange. Seeing the state of affairs, scholars are doing many studies. Correspondingly, communication on digital mediums may use to promote unlawful actions. One simple step is distributing illegal software and high-end programming facilities. Measures are typically associated with exploiting vulnerabilities in the system that allow intruders to enter the network and distribute personal data, rejection-service attacks with spying sort of works [18]. Machine automation in safety digitization is yet an innovative field, but they are impelling cyber forensic study into a new example of active defense. The objective is to forestall the aggressive efforts to distinguish intimidations before using any smart model and cogent. This innovative perception rests with existing security systems, which respond to known threats on a large scale [19].

The progression of digital diversification has proven to adapt to the new empirical milieu. However, it represents other benefits for various companies irrespective of the underlying area where it can operate or undertake the projects. Before and after digitization, the authors have examined the attached issues and the requirements of various industries and termed them IoTv4 [19]–[22]. The authors have labeled some paybacks of digital societies, the consequences of cyber security, and the definition of threats with different strategy types adopted to evade security jeopardies. The implication of machine automation and digital diversification as an essential part of the future solution for the self-propelled engineering sectors. Research happenings enable machine automation where they must meet many new requirements such as failed operations and cyber security measures of different industries [23]. Also, there are various drawbacks behind methods and approaches with their tools, which desire to quicken all results for legalization and authentication. The main challenge is correlating in better collaboration of prevailing growth methods, where authors have demanded and shown its requirement in their study process [24]. The process of detecting possible jeopardies presented the concept of evaluating the intensity of digital fortification in an electronic media forum. These limits under their study can discourse the idea by developing an outline to classify various jeopardies in cyber security for edifice future engineering [25]. The small offices and home office security issues for industrial corporations presented a cyber-automated system for small and medium-sized enterprises (SMEs). Significant features have been found in companies, but restoration schemes often lack overall safety awareness among employees. The development of digital wealth in Russia and its issues [26]. As it turns out, the country ranks high in the national cyber security rankings. These findings can apply to formulating a strategic plan for innovation development in Russia. The limitation is that the analysis only uses data from internet communities' organization for economic cooperation and development reports [27], [28].

2. METHOD

The critical attention of this section is to provide an outline regarding approaches and performances implemented in cyber security data research. Predictable methods have first been introduced to start the machine-based learning models as supervised, non-supervised, and semi-supervised methods. Supervision practice teaches that the input display output function is based on "I/O pairs, an example of working with machine learning. The result should contain data that matches our already mapped training model. Supervised learning solves two problems first is retrogression, and the second is cataloging.

In contrast, between two methods that have to do with the quality of the output, the earlier gives a series of outcomes, while the final gives a distinct one. Cyber security research has often found that security concerns involve a cataloging model with taxonomic output to generate the necessary models that classify transparent information that needs to be "trained" with a set of illustrative examples. Training is usually done for building or more for the model at this stage. There are widely available algorithms for learning them, each with its features and weights. Another traditional method is to teach without a teacher. It differs from the last because the categorized dataset model does not require training. In lieu, data-driven structural algorithms will distinguish between examples by identifying matches through the primary data structure. Not often are we semi-regulated as the first two methods discussed. The first gain of selecting this method is using uncategorized data to create a training model. Creating such data labels is difficult because it is time-killing and a costlier affair as a semi-regulated option. The model can train by uncategorized data with multiple brands, which helps

to get good results. This means that consuming both kinds of data during the training process improves the correctness of the resulting model, reducing markup time and costs. The correct plan for creating our model is as firstly, what kind of projected result that desire to use (retrogression or cataloging); secondly, resolving possible issues through a set of demonstrative data, thirdly what type of data that can input (special or incessant) and fourthly an evaluative round-trip assessment of the entire progression.

3. RESULTS AND DISCUSSIONS

To create an automated learning prototype, it is necessary to clear the digital content post for use in the selected machine learning (ML) algorithm's directory, which is the critical point in data-driven work. To read continuous or untrue input in the section, authors can follow natural language editor operators for text data conversations. The accomplishment can be done using denoise, transliteration, trivialization, originating, and Lemmatization which are recurrently performed in data-driven studies like cyber forensics. Figure 1. is a balanced framework for primary digital content processing.

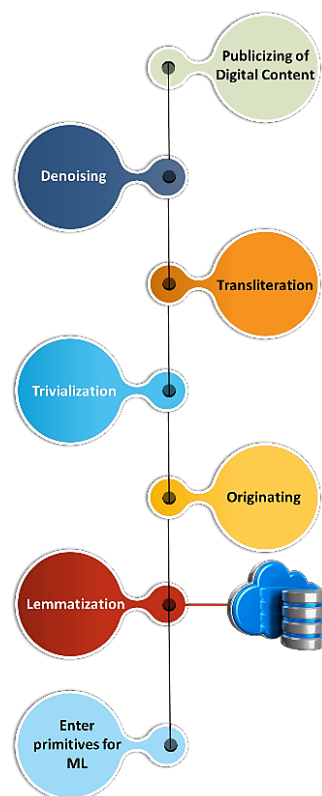


Figure 1. The framework of preliminary processing of digital content

3.1. Phase-1 (Denoising)

The first step after data publicizing to start with is denoising is the process of tag suppression from hypertext markup language (HTML) or extensible markup language (XML). Generally, the tool python can use to make this filtration successful. Also, the stylistic transferring of digital content may be done through HTML to JSON or vice versa. This experiment aims to create a model for the supervised categorization of HTML tags. This uses the same HTML document to annotate it, but some offspring of the <body> branch had labels manually added in a <class> property. Items that do not fall into these categories are considered noise.

3.2. Phase-2 (Transliteration)

Then, in the second phase, transliteration will start by gathering information from its domain area, which demarcates every notion in the transcript in the form of symbols or words. These marking notions will act as input to the ML procedures where it takes etymological study to resolve the issues in ML adaptations. Here more chances to face problems in defining the delimiters for various punctuations. These apps deal with

massive amounts of text for categorization or translation, which requires significant effort on the back end. It is a difficult procedure to convert text into something that an algorithm can understand.

3.3. Phase-3 (Trivialization)

After that, in the third phase, trivialization will be in place for the transfiguration of words and sentences into different ensigns, like the conversion from lower to upper. This step is done instead before the expressive investigation of the digital data. Cracks are a significant factor in monitoring and diagnosing the integrity of the structure of a concrete system. If cracks are discovered early, additional efforts can be made to improve structural health.

3.4. Phase-4 (Originating)

Then, originating helps to identify the source of the stemmed word by suppressing any before and after attachments in the fourth phase. The best reason for deploying this phase is to lessen the word counts, which saves storage space. Though transformed from one to another word, each word stemmed from its corresponding source. As it confirms that from trivialization of the conversion process, one note or comment is related to other because it roots from its source. Elimination can save the capacity of input primitives, which conserves the process of computational resources.

3.5. Phase-5 (Lemmatization)

Hence in the fifth phase, the well-known technique of lemmatization to eliminate these break words from the torrent of stylistic digital information is based on a database that can form by the predefined setlist. Thus, lemmatization aids in reducing terms such as studies and studying to a single base form or root word study. It should be noted that not all processes are required and depend on the application use case. It can use all of the proposed methods for spam filtering but not for language translation issues.

3.6. Phase-6 (Primitive conversion)

Then, at last, primitive conversion of stylistic digital data into input primitives requires some primary phrases. The first one is the signature of feature size, second is its depiction kind, which denotes whether the data are distinct or incessant by taking help from the space trajectory model. This proposed framework is a concise view of the adaptation issues of machine learning in safety digitization to provide defense against cyber security. It will give us comparative data among the number of approaches used in digital forums for resolution adoption, various procedures, preliminary processing techniques, and valuation stratagems.

Here the essential discussion on variances is collected data where the training model is identical to the social mediums. Now authors have assembled various adaptation issues where to:

- Spot vulnerability posts and leverage package products.
- Spot the service areas of various code runners and search the available versions in-store.
- Spot and track the malicious attackers.
- Track the tearfulness breakdown on attacker posts.

Experts in machine learning claim it is possible to automate ethical decisionmaking. Scholars use digital mediums as the usual way of gathering material from individuals across the globe. Otherwise, obtaining cyber security-related material through outmoded approaches such as research would take much time. The same is valid with cyber security research when scholars collect digital content from the dark web forum. Nonetheless, various legal ramifications of this information's usage are not adequately addressed across multiple cyber security studies. The following two questions will aid in our deliberations. An explicit contract in an internet community like Facebook denotes that user data is very open and can be used by any party of their interest. However, there is no such contract of information in hackers' forums. Therefore, permission through social media cannot be sufficient for any of the scholars to do their research against ethical contemplation. Obtaining up-to-date consent becomes more problematic, as it is almost implausible to search from the number of feature sets. In social media sites such as Facebook and Twitter, an explicit arrangement (commonly referred to as a contract agreement) informs users that third-party firms and academic organizations can use their data. However, there is no formal contract governing the usage of users' findings in hacker sites and chat rooms. However, in some instances, agreement via social media is an insufficient ethical justification for the researcher to continue the study. Researchers' judgment has to be improved before deciding whether to use this data; legal enforcement cannot be overlooked simply because the data seem public. Despite these concerns, cryptography testing and enterprise use data without prior authorization. In cyber security research, data are accessed and analyzed without the participants' informed consent, and they are frequently unaware of their involvement. Acquiring informed consent becomes more difficult when dealing with a dataset containing hundreds of data points.

4. CONCLUSION

It is the time for training, realizing, and assessing the prototypical features in a measured setting after using them in real time. To make it happen, there are various aspects to consider. The first thing this should notice is the vagaries in hacker terminology. There are differences in the constant evolution of connotations, acronyms, spellings, and even technical jargon, which requires periodic re-learning of the prototypical. The trained model in some digital forums does not certainly show the same in others because of terminology differences. Next can consider the lack of baseline truth feature sets for framework assessments. The future achievement of this study is to permit different cyber security scholars to distinguish their findings for further improvisation.

ACKNOWLEDGEMENTS

The authors are very thankful to Editor and Reviewers for their valuable suggestions towards improvement the article.




REFERENCES

- [1] A. El Kah, A. El Airej, and I. Zeroual, "Arabic authorship attribution on Twitter: what is really matters?," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1730–1737, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1730-1737.
- [2] R. Chingamtotattil and R. Gopikakumari, "Neural machine translation for Sanskrit to Malayalam using morphology and evolutionary word sense disambiguation," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1709–1719, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1709-1719.
- [3] A. K. Bitto *et al.*, "CryptoAR: scrutinizing the trend and market of cryptocurrency using machine learning approach on time series data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1684–1696, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1684-1696.
- [4] A. N. Tultul, R. Afroz, and M. A. Hossain, "Comparison of the efficiency of machine learning algorithms for phishing detection from uniform resource locator," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1640–1648, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1640-1648.
- [5] Z. I. A. Alrifae and T. Z. Ismaeel, "Cryptography based on retina information," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1697–1708, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1697-1708.
- [6] E. Rhee and J. Cho, "Security system using mobile image processing and color recognition for the visually impaired," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, p. 1363, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1363-1368.
- [7] Z. Zhou, Y. Matsubara, and H. Takada, "Resilience analysis and design for mobility-as-a-service based on enterprise architecture modeling," *Reliability Engineering and System Safety*, vol. 229, p. 108812, Jan. 2023, doi: 10.1016/j.res.2022.108812.
- [8] Ö. Sen, D. van der Velde, K. A. Wehrmeister, I. Hacker, M. Henze, and M. Andres, "On using contextual correlation to detect multi-stage cyber attacks in smart grids," *Sustainable Energy, Grids and Networks*, vol. 32, p. 100821, Dec. 2022, doi: 10.1016/j.segan.2022.100821.
- [9] M. Alawida, A. E. Omolara, O. I. Abiodun, and M. Al-Rajab, "A deeper look into cybersecurity issues in the wake of COVID-19: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8176–8206, Nov. 2022, doi: 10.1016/j.jksuci.2022.08.003.
- [10] M. M. Kamruzzaman and O. Alruwaili, "AI-based computer vision using deep learning in 6G wireless networks," *Computers and Electrical Engineering*, vol. 102, p. 108233, Sep. 2022, doi: 10.1016/j.compeleceng.2022.108233.
- [11] M. Ylönen *et al.*, "Integrated management of safety and security in Seveso sites - sociotechnical perspectives," *Safety Science*, vol. 151, p. 105741, Jul. 2022, doi: 10.1016/j.ssci.2022.105741.
- [12] J. Lee, I. Cameron, and M. Hassall, "Information needs and challenges in future process safety," *Digital Chemical Engineering*, vol. 3, p. 100017, Jun. 2022, doi: 10.1016/j.dche.2022.100017.
- [13] H. Pasman, H. Sun, M. Yang, and F. Khan, "Opportunities and threats to process safety in digitalized process systems—An overview," in *Methods in Chemical Process Safety*, 1st ed., sciencedirect, 2022, pp. 1–23.
- [14] G. Aiello, P. Catania, M. Vallone, and M. Venticinque, "Worker safety in agriculture 4.0: A new approach for mapping operator's vibration risk through Machine Learning activity recognition," *Computers and Electronics in Agriculture*, vol. 193, p. 106637, Feb. 2022, doi: 10.1016/j.compag.2021.106637.
- [15] S. K. Baduge *et al.*, "Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications," *Automation in Construction*, vol. 141, p. 104440, Sep. 2022, doi: 10.1016/j.autcon.2022.104440.
- [16] Y. Zhong, K. Guo, J. Su, and S. K. W. Chu, "The impact of esports participation on the development of 21st century skills in youth: A systematic review," *Computers and Education*, vol. 191, p. 104640, Dec. 2022, doi: 10.1016/j.compedu.2022.104640.
- [17] H. Ahmetoglu and R. Das, "A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions," *Internet of Things (Netherlands)*, vol. 20, p. 100615, Nov. 2022, doi: 10.1016/j.iot.2022.100615.
- [18] N. Agarwal, G. Sikka, and L. K. Awasthi, "A systematic literature review on web service clustering approaches to enhance service discovery, selection and recommendation," *Computer Science Review*, vol. 45, p. 100498, Aug. 2022, doi: 10.1016/j.cosrev.2022.100498.
- [19] Y. li Liu, L. Huang, W. Yan, X. Wang, and R. Zhang, "Privacy in AI and the IoT: The privacy concerns of smart speaker users and the personal information protection law in China," *Telecommunications Policy*, vol. 46, no. 7, p. 102334, Aug. 2022, doi: 10.1016/j.telpol.2022.102334.
- [20] J. Leng *et al.*, "Industry 5.0: Prospect and retrospect," *Journal of Manufacturing Systems*, vol. 65, pp. 279–295, Oct. 2022, doi: 10.1016/j.jmsy.2022.09.017.
- [21] R. Peres, M. Schreier, D. A. Schweidel, and A. Sorescu, "Blockchain meets marketing: Opportunities, threats, and avenues for future research," *International Journal of Research in Marketing*, Aug. 2022, doi: 10.1016/j.ijresmar.2022.08.001.



- [22] C. Virginia Anikwe *et al.*, "Mobile and wearable sensors for data-driven health monitoring system: State-of-the-art and future prospect," *Expert Systems with Applications*, vol. 202, p. 117362, Sep. 2022, doi: 10.1016/j.eswa.2022.117362.
- [23] M. H. Panahi Rizi and S. A. Hosseini Seno, "A systematic review of technologies and solutions to improve security and privacy protection of citizens in the smart city," *Internet of Things (Netherlands)*, vol. 20, p. 100584, Nov. 2022, doi: 10.1016/j.iot.2022.100584.
- [24] F. Aloraini, A. Javed, O. Rana, and P. Burnap, "Adversarial machine learning in IoT from an insider point of view," *Journal of Information Security and Applications*, vol. 70, p. 103341, Nov. 2022, doi: 10.1016/j.jisa.2022.103341.
- [25] J. J. Peralta Abadía, C. Walther, A. Osman, and K. Smarsly, "A systematic survey of internet of things frameworks for smart city applications," *Sustainable Cities and Society*, vol. 83, p. 103949, Aug. 2022, doi: 10.1016/j.scs.2022.103949.
- [26] J. Lichy and W. Ng, "Digital disruption in a state-controlled ecosystem: A sociomaterial perspective of public use of the internet under China's Social Credit System," *Technological Forecasting and Social Change*, vol. 183, p. 121948, Oct. 2022, doi: 10.1016/j.techfore.2022.121948.
- [27] O. Andriychuk, "Shifting the digital paradigm: Towards a sui generis competition policy," *Computer Law and Security Review*, vol. 46, p. 105733, Sep. 2022, doi: 10.1016/j.clsr.2022.105733.
- [28] Y. K. Dwivedi *et al.*, "Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 66, p. 102542, Oct. 2022, doi: 10.1016/j.ijinfomgt.2022.102542.

BIOGRAPHIES OF AUTHORS





Gyana Ranjana Panigrahi    is a Ph.D. scholar currently pursuing a Ph.D. from Sambalpur University in the department of Electronics, Sambalpur, Odisha, India. My research area includes cyber security, Digital forensics, Physical Cyber Systems, Communication, Wireless communication, Data communication & Networking, IoT, and Storage Area Networks (SAN). He can be contacted at email: gyana.ranjana.panigrahi@suiit.ac.in.





Dr. Nalini Kanta Barpanda   received his Ph.D. in Engineering from the Sambalpur University. He is working as Reader in Electronics, Sambalpur University, Odisha. He has published over 62 number of research articles in various areas of Performance Analysis of Communication Interconnection N/W, Wireless Sensor N/W, Image Processing, and Internet of Things. He can be contacted at email: nkbarpanda@suniv.ac.in.






Dr. Komma Anitha   is working as Associate professor in the Department of Electronics and Communication Engineering, PVP Siddhartha Institute of Technology, Vijayawada, A.P, India. She has published more than 40 articles in different reputed journals. Her area of interest is cybersecurity. She can be contacted at email: anithakomma108@gmail.com.






Dr. Shanti Rathore   is working as professor in the Department of Electronics and Telecommunication Engineering, Dr. C V Raman University, Chhatisgarh, India. She is working on networking, machine learning and IoT. She has published more than 50 articles in reputed journals. She produced 6 Ph.D. scholar and 4 are currently working in her guidance, Japan. She can be contacted at email: rathorepuja@gmail.com.



Dr. Preesat Biwas    is working as Assistant Professor in the Department of Electronics and Telecommunication Engineering, GEC Jagdalpur, CG, India. He has published more than 25 research articles in reputed journals and international conferences. His research area is communication system design and cyber security. He can be contacted at email: preesat.eipl@gmail.com.



Dr. Prabira Kumar Sethy    currently working as an Assistant Professor in the Department of Electronics at Sambalpur University since 2013. He has nine years of teaching, research & administrative experience and four years of Industry experience. Previously he worked as Engineer in Doordarshan, Prashar Bharati, from 2009 to 2013. He has received his Ph.D. and M. Tech degree from Sambalpur University and IIT (ISM) Dhanbad, respectively. His research area is image processing, machine learning, and deep learning. He has published 80 research papers in different reputed journals and conferences. In addition, he has two patents. He is an editorial board member of the International Journal of Electrical and Computer Engineering. He is also Editorial Board Member in Ingénierie des Systèmes d'Information. IIETA. He received the "InSc Young Achiever Award" for the research paper "Detection of coronavirus (COVID-19) based on Deep Features and Support Vector Machine, organized by the Institute of Scholars, Ministry of MSME, Government of India in the year 2020. He is a Senior Member of IEEE. He is a frequent reviewer of many journals and session chair of international conferences. He can be contacted at email: prabirsethy.05@gmail.com.