International Journal of Computer and Communication Technology

Volume 8 | Issue 4 Article 3

November 2022

Empirical Analysis of Machine Learning algorithms in Fake News detection

Bhagyalaxmi Devi
CIME, Bhubaneswar, bhagyalaxmidevi87@gmail.com

Sudhir Senapati
Asst. Prof. College of IT and Management Education, BBSR, sudhir.aricent@gmail.com

Recommended Citation

Devi, Bhagyalaxmi and Senapati, Sudhir (2022) "Empirical Analysis of Machine Learning algorithms in Fake News detection," *International Journal of Computer and Communication Technology*: Vol. 8: Iss. 4, Article 3.
DOI: 10.47893/IJCCT.2022.1432
Available at: https://www.interscience.in/ijcct/vol8/iss4/3

Empirical Analysis of Machine Learning algorithms in Fake News detection

Cover Page Footnote

Dear Editor, We have modified the manuscript as per the reviewers comments.

Empirical Analysis of Machine Learning algorithms in Fake News detection

Bhagyalaxmi Devi, Sudhir Kumar Senapati*
CIME, Bhubaneswar
bhagyalaxmidevi87@gmail.com, sudhir.aricent@gmail.com

Abstract

Social media is the finest venue for thinking and expressing in the modern world. And this is the best place to share information about your identity, culture, religion, and customs. It entails an immediate information interchange that covers news from every industry. These days, social media has a big impact on how we live and how society functions. Currently, social media is the best medium for expressing your thoughts. Social media has also evolved into a channel for disseminating information about nearby events. how the locals in the other place are made aware of what is going on there. People benefit from this through learning about various cultures. However, some evil people use social media to spread their lies, which affects society and our everyday lives. Furthermore, fake news spreads like a forest fire if it is not dealt with promptly. And this bogus news offends certain individuals and occasionally sparks riots in public places. We need instruments in the modern day that can confirm any news, whether it is real or fraudulent. The current work considers a variety of machine-learning techniques for detecting false news, including Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM). The performance evaluation was then conducted using several criteria, including F-1 score, recall, accuracy, and precision. The empirical investigation shows DT has the greatest accuracy level at 100%.

Keywords: Fake News, Machine Learning, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM)

1. Introduction

A strange type of media is online media. The consequence is the creation of a virtual environment that can be visited online. Global connectivity is made possible by the enormous online media sector. It is not a trustworthy form of communication. Each field's information is present and engages in quick data exchange. Today's culture and our daily lives are greatly impacted by internet-based media. Online media is also the best technique to communicate your opinions in the current world. Online media now provides a platform to share what is happening in your immediate environment. how the people of the other place learn about what is going on in the other place.

People also learn about different cultures' ways of living in addition to this. However, certain wicked elements use internet media to disseminate their erroneous ideas, which impacts our way of life and society. Furthermore, this fake news is nothing like a forest fire if it isn't put out at the right time. Additionally, some people are offended by this fake news, and it occasionally even sparks riots among the populace. Today's new media format, social media, gives everyone the chance to freely voice their opinions. Distance is no longer an obstacle since social media makes it simple for individuals who live far apart to interact. Nowadays, users use social media platforms to exchange papers and send text messages. Through social media, you may also establish friends with people from other societies, cultures, lifestyles, or nations who

remain your friends despite these differences. You can also inform people of the crime against you and ask for a new one through social media.

But since every coin has two sides—one positive and the other—as we all know, social media is no exception to the norm. Some people aim to deceive others by disseminating false information through social media. Because of this, people's feelings are wounded, and occasionally this leads to riots in the community. The propagation of fake news on social media must be stopped because it costs the nation and its citizens both lives and property. Today, social media is a world unto itself where everyone can freely express their opinions and speak publicly on any topic. Social networking has made it feasible for people to reconnect with old-school pals and relive their memories with them in ways that weren't before conceivable. Social media has dissolved international borders and drawn even remote individuals closer together.

Social media allows you to network with new people and learn about different cultures, religions, and lifestyles. People may now discuss news among themselves via social media, which up until now hasn't even been reported by newspapers or television networks. To resist the injustice being done to us and inform society and the rest of the world of what is happening to us, social media has given us a weapon. The term "deception," which is frequently used to refer to "fake news" in modern times, is defined as false or inaccurate information that may be given to mislead the readers.

Data or opinions that you can't help but disagree with might not be untrue. Although the term "fake news" is frequently used as a derogatory in news reporting today, this is an unreliable use of the phrase. To be sure, labeling reality-based reporting "deception" because it contradicts your political views might perhaps be considered a lie in and of itself. Consider a few modern models to help you understand the concept of deceit:

- A video that appeared to show Nancy Pelosi stuttering and slurring her speech appeared online in May 2019, which led many experts to speculate about her psychological makeup. This was a doctored film, as revealed by The New York Times.
- Midway through the year 2020, several stories about the supposed treatment of COVID-19, often known as Covid, began to stand out as extremely noteworthy. Many fake examples were presented as facts, including the idea that treating the condition by consuming more red meat or coconut oil.

1.1 Objective

- To implement machine learning-based algorithms for fake detection.
- To evaluate the performance of the machine learning algorithms in terms of some influencing parameters such as accuracy, precision, recall, and F-1 score.

2. Literature Survey

Ethar Qawasmeh and others and others [1] The author of this research study made extensive use of a large dataset. In this research, the dataset was essentially cleaned first by taking the dataset, and then the vectorization process was used. However, the source of the information is not discussed in this work. If we don't know where the news is coming from, how will we know if it's fake? This is a vital consideration when deciding on any news. The author of this article has focused more on the topic of news, writing about political news, sports news, etc. But to identify bogus news, we must carefully consider each fact. The accuracy of the system algorithm utilized by the author in this work is 85.3%.

Wang William Yang and others [2] A challenging problem in misdirection discovery, automatic fake news identification has significant real-world political and societal repercussions. However, the lack of designated benchmark datasets has severely constrained the measurement of approaches to combating fake news. In this article, we introduce LIAR, the publicly available dataset for detecting fake news. The site provides a detailed examination report and connects to source documents for each case. We collected long-term, 12.8K physically marked brief explanations in various settings from the site. Additionally, studies on certainty-checking can make use of this dataset. This new dataset outpaces the largest publicly available fake news datasets of the same kind in size.

Costin BUSIOC and others, etc. [3] Fighting fake news is a difficult and challenging endeavor. False news has a historically spectacular impact on people's lives and is gaining power in the social arena of politics. Drives aboutcomputerized fake news findings have become more popular as a result of this miracle, creating unavoidable inspection curiosity. However, while trying to come up with such arrangements, the majority of approaches that concentrate on English and low-asset languages run into problems. This analysis highlights current plans, issues, and viewpoints held by many research groups while focusing on the progress of such analyses. In addition, considering the limited scope of computational analyses conducted on Romanian fake news, we assess the relevance of the available approaches in the Romanian context while simultaneously identifying potential future research directions.

Alim Al Ayub Ahmed et al. [4] In this survey paper user has solved enter news using the important algorithm. In this, the author has divided the query into pieces and has searched the related news through free web scraping. Then based on the manipulated data, it tried to tell whether the news is fake or not. Its accuracy is 82.

Razan Masood et al. [5] This paper was published in early 2018 and this paper's machine learning algorithm has been used to predict fake news. In this paper, the author first selects the features from the dataset and then uses the SVM algorithm for classification. In this research paper, linear regression has been used for prediction.

Sohan De Sarkar and coauthors [6] To stop the propagation of misinformation on the Internet, satirical news identification is important. Current methods for handling fake news parodies combine hand-drawn highlights with AI models like SVM and other levels of neural networks, but they do not take into account the distinction between sentences and archives. In this study, a robust, progressive, profound brain organization strategy for parody recognition is proposed. This approach may be used to detect parody both at the sentence and report levels. Pluggable non-exclusive neural organizations like CNN, GRU, and LSTM are combined in engineering. The genuine news parody dataset test results indicate considerable performance improvements, demonstrating the suitability of our suggested strategy. A review of the learned models reveals the existence of crucial phrases that regulate the occurrence of parody in news.

Abd-All-Tanvire and others [7] This research paper offers a forecast for AIM Fake News for the year 2019. In this study, the user's input is gathered first, and the Python API is subsequently employed by that information. Then, using Twipy, a collection of data was prepared from Twitter. The author then purifies the dataset and categorizes it using a support vector machine. The Trained and Test datasets are prepared after the categorization. The trained-to-test datasets ratio is assumed to be 6:4. The decision tree algorithm then forecasts whether the news is phony or real.

3. Methodology

This section discusses various machine-learning approaches [8-25].

3.1 Decision Tree

A Decision Tree is a very popular machine learning algorithm that is being used to classify problems within supervised learning. Decision Tree can be used in both Regression and Classification.Itclassifiestheinputdatawithinaparticularclass.WhilepreparingtheDecision Tree model, it is trained in such a way that whenever it is given any unknown input data, it can find out which class it belongs to. For example, take an insurance company and suppose that company has to sell its insurance policies, then with the help of a decision tree, they can find out how many people can buy insurance according to their age through a decision tree. If they are classified.

3.2 Random Forest

Random forest tree is an algorithm of machine learning and is an advanced decision tree form. Therearedifferentdecisiontreesintherandomforesttreealgorithmandallthesetreesgenerate different results. Then the prediction of the results of all these decision trees is combined to form a new decision tree. And the prediction result from that is the final one.

3.3 Support Vector Machine

Classification and regression problems are resolved using a Support Vector Machine, or SVM, one of the most used supervised learning techniques. It is mostly used, nevertheless, in Machine Learning Classification problems. In order to swiftly categorize new data points in the future, the SVM algorithm aims to define the best line or decision boundary that can split n-dimensional space into classes. The name for this ideal decision boundary is a hyperplane. The extreme vectors and points that help create the hyperplane are chosen via SVM. These extreme cases are referred to as support vectors, and the technique is called a support vector machine.

3.4 Naïve Bayes

The two terms Nave and Bayes that make up the Nave Bayes algorithm may be expressed as:

- Naive: It is so called because it assumes that the occurrence of one feature is unrelated to the occurrence of other features. A red, spherical, sweet fruit, for instance, is recognized as an apple if the fruit is identified based on its color, form, and flavor. Because of this, each characteristic may be used independently of the others to help determine that something is an apple.
- Bayes: It is known as Bayes because it relies on the Bayes' Theorem.

4. Implementation

The dataset has been taken from the Kaggle open-source data repository. For implanting the algorithms python with an Anaconda environment has been considered. Figure 1 and Figure 2 show the dataset’s head and tail respectively. Figure 3-5.13 are the classification report with a confusion matrix for every algorithm. For calculating the efficiency accuracy, F1-score, precision, and recall has been considered which can be calculated from the confusion matrix [26- 32].

	title	text	subject	date	target
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake

Figure 1 Dataset head description

	title	text	subject	date	target
44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	true
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

Figure 2 Dataset tail description

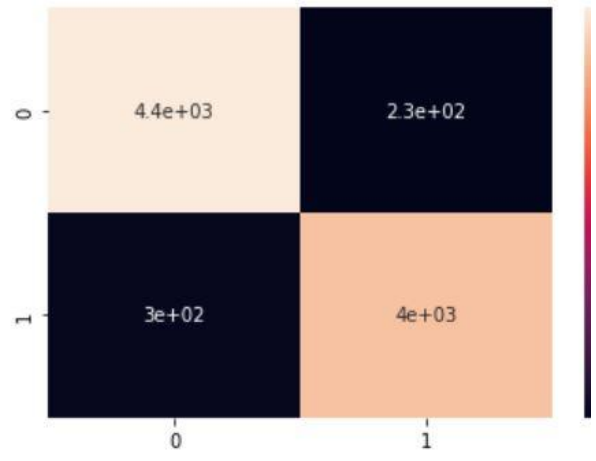

Figure 3 Confusion Matrix for NB

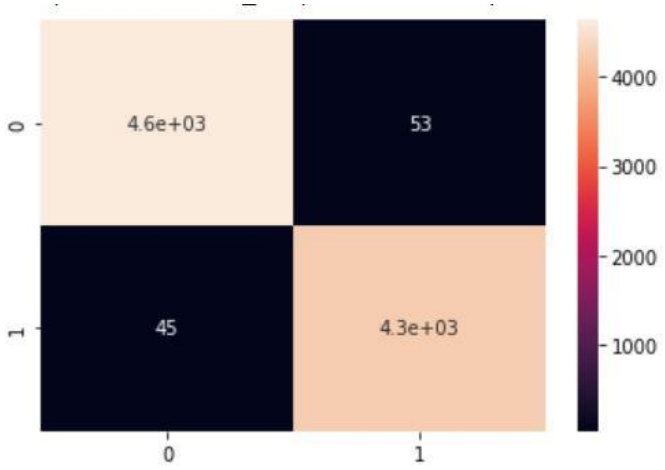

Figure 5 Confusion Matrix for RF

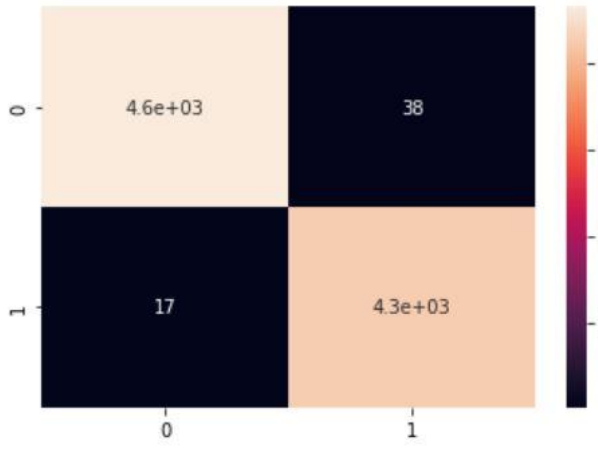

Figure 4 Confusion Matrix for SVM

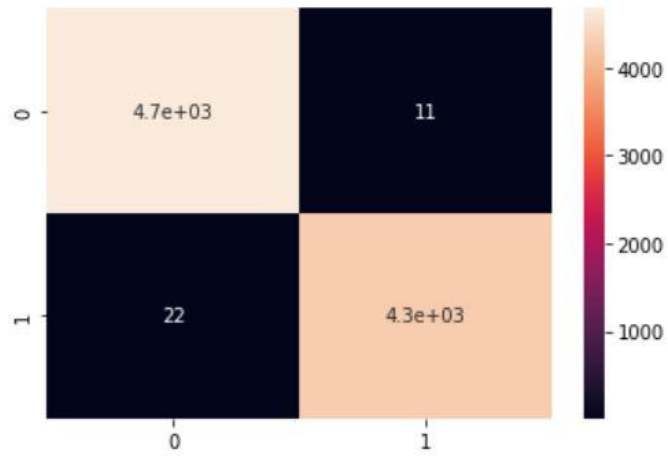

Figure 6 Confusion Matrix for DT

```
[[4450  232]
 [ 300 3998]]

Classification Report of Naive Bayes Classifier:

              precision    recall  f1-score   support

        fake       0.94      0.95      0.94      4682
        true       0.95      0.93      0.94      4298

    accuracy                           0.94      8980
   macro avg       0.94      0.94      0.94      8980
weighted avg       0.94      0.94      0.94      8980
```

Figure 7 Classification Report for NB

```
[[4644   38]
 [  17 4281]]

Classification Report of SVM Classifier:

              precision    recall  f1-score   support

        fake       1.00      0.99      0.99      4682
        true       0.99      1.00      0.99      4298

    accuracy                           0.99      8980
   macro avg       0.99      0.99      0.99      8980
weighted avg       0.99      0.99      0.99      8980
```

Figure 8 Classification Report for SVM

```
Confusion Matrix of Naive Bayes Classifier:

[[4629   53]
 [  45 4253]]

Classification Report of Naive Bayes Classifier:

              precision    recall  f1-score   support

        fake       0.99      0.99      0.99      4682
        true       0.99      0.99      0.99      4298

    accuracy                           0.99      8980
   macro avg       0.99      0.99      0.99      8980
weighted avg       0.99      0.99      0.99      8980
```

Figure 9 Classification Report for RF

```
Confusion Matrix of Decision Tree Classifier:

[[4671   11]
 [  22 4276]]

Classification Report of Decision Tree Classifier:

              precision    recall  f1-score   support

        fake       1.00      1.00      1.00      4682
        true       1.00      0.99      1.00      4298

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980
```

Figure 10 Classification Report for DT

5. Conclusion

According to our study, everything has a good side and a bad side, and our decision is based on which side we select. The same thing applies to social media platforms, which some individuals are now utilizing more carelessly. By spreading false information, some bad people are really harming society. My findings may be extremely useful in putting a stop to fake news. Any bogus news may be prevented from reaching society if this mechanism is deployed on the social media platform. In order to make the best choice, people may utilize it to confirm any news they have seen on any social media site. You can also stop anything bad from happening. The primary goal of this algorithm was accomplished, and the DT algorithm's accuracy was 100%, which highly satisfied the goal.

In the future, I would like to work on this research again because the research I have done was based only on textual data. But some smart wrong people also try to spread fake news through pictures, audio, and videos. They spread the wrong thing by tampering with the video, audio, or picture.

Reference

[1] EtharQawasmeh,MaisTawalbeh,MalakAbdullah,“AutomaticIdentificationofFakeNews Using Deep Learning”, 2019 Sixth International Conference on Social NetworksAnalysis,Management and Security (SNAMS), 978-1-7281-2946-4/19/$31.00 ©2019 IEEE

[2] WilliamYangWang,“Liar,LiarPantsonFire”:ANewBenchmarkDatasetforFakeNews Detection ”, arXiv:1705.00648v1 [cs.CL] 1 May2017.

[3] Costin BUSIOC, Stefan RUSETI, Mihai DASCALU, “ A Literature Review of NLP Approaches to Fake News Detection and Their Applicability to Romanian- Language News Analysis”, 2020, *Romanian Ministry of Education and Research, CNCS - UEFISCDI, project number PN-III-P1-1.1-TE-2019-1794, within PNCDIIII.*

[4] Alim Al Ayub Ahmed, Ayman

Aljarbouh, Praveen Kumar Donepudi," Detecting Fake News using Machine Learning: A Systematic Literature Review",2020, IEEEconference.

[5] Razan Masood and Ahmet Aker," The Fake News Challenge: Stance Detection using Traditional Machine Learning Approaches", In Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KMIS 2018), pages128-135

[6] Sohan De Sarkar, Fan Yang," Attending Sentences to detect Satirical Fake News", Proceedings of the 27th International Conference on Computational Linguistics, pages 3371– 3380 Santa Fe, New Mexico, USA, August 20-26,2018.

[7] Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter" Detecting Fake News using Machine Learning and Deep Learning Algorithms", 2019 7th International Conference on Smart Computing & Communications(ICSCC).

[8] Pati, A., Parhi, M., & Pattanayak, B. K. (2022). IHDPM: an integrated heart disease prediction model for heart disease prediction. International Journal of Medical Engineering and Informatics, 14(6), 564-577.

[9] Roul, A., Pati, A., & Parhi, M. (2022). COVIHunt: An Intelligent CNN-Based COVID-19 Detection Using CXR Imaging. In Electronic Systems and Intelligent Computing (pp. 313-327). Springer, Singapore.

[10] Rout, S. K., Sahu, B., Panigrahi, A., Nayak, B., & Pati, A. (2023). Early Detection of Sepsis Using LSTM Neural Network with Electronic Health Record. In Ambient Intelligence in Health Care (pp. 201-207). Springer, Singapore.

[11] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, pages 15–19. ACM,2015.Mathieu Cliche. The sarcasm detector,2014.

[12] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, page 82. American Society for Information Science,2015.

[13] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals inlarge-scaletext.InProceedingsofthe2016CHIConferenceonHumanFactorsinComputing Systems, pages 4647–4657. ACM,2016.

[14] SongFeng,RitwikBanerjee,andYejinChoi. Syntacticstylometryfordeceptiondetection.

In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pages 171–175. Association for Computational Linguistics,

2012.

[15] Johannes F¨urnkranz. A study using n-gram features for text categorization. Austrian Research Institute for Artificial Intelligence, 3(1998):1–10, 1998.

[16] Sahu, B., Panigrahi, A., Rout, S. K., & Pati, A. (2022, July). Hybrid Multiple Filter Embedded Political Optimizer for Feature Selection. In 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) (pp. 1-6). IEEE.

[17] MykhailoGranik and Volodymyr Mesyura. Fake news detection using naive Bayes classifier. In Electrical and Computer Engineering (UKRCON), 2017 IEEE First Ukraine Conference on, pages 900–903. IEEE, 2017.

[18] Pati, A., Parhi, M., Pattanayak, B. K., Singh, D., Samanta, D., Banerjee, A., ... &Dalapati, G. K. (2022). Diagnose Diabetic Mellitus Illness Based on IoT Smart Architecture. Wireless Communications and Mobile Computing, 2022.

[19] Johan Hovold. Naive Bayes spam filtering using word-position-based attributes. In CEAS, pages 41–48,2005.

[20] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759,2016.

[21] Pati, A., Parhi, M., & Pattanayak, B. K. (2022). HeartFog: Fog Computing Enabled Ensemble Deep Learning Framework for Automatic Heart Disease Diagnosis. In Intelligent and Cloud Computing (pp. 39-53). Springer, Singapore.

[22] Hadeer Ahmed, Issa Traore, Sherif Saad, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques“,International Conferenceon Intelligent, Secure, and Dependable Systemsin Distributedand Cloud Environments ISDDC 2017: Intelligent, Secure, and Dependable Systems in Distributedand Cloud Environments pp 127-13.

[23] Parhi, M., Roul, A., Ghosh, B., & Pati, A. (2022). IOATS: an Intelligent Online Attendance Tracking System based on Facial Recognition and Edge Computing. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 252-259.

[24] Sahu, B., Panigrahi, A., Mohanty, S., &Sobhan, S. (2020). A hybrid cancer classification based on SVM optimized by PSO and reverse firefly algorithm. International Journal of Control and Automation, 13(4), 506-517.

[25] Sahu, B., & Panigrahi, A. (2020, February). Efficient role of machine learning classifiers in the prediction and detection of breast cancer. In 5th International Conference on Next Generation Computing Technologies (NGCT-2019).

[26] Sahu, B., Badajena, J. C., Panigrahi, A., Rout, C., & Sethi, S. (2020). An intelligence-based health biomarker identification system using microarray analysis. In Applied intelligent decision making in machine learning (pp. 137-161). CRC Press.

[27] Pati, A., Parhi, M., & Pattanayak, B. K. (2021, January). IDMS: an integrated decision-making system for heart disease prediction. In 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON) (pp. 1-6). IEEE.

[28] Pati, A., Parhi, M., & Pattanayak, B. K. (2022). IADP: An Integrated Approach for Diabetes Prediction Using Classification Techniques. In Advances in Distributed Computing and Machine Learning (pp. 287-298). Springer, Singapore.

[29] Sahu, B., Panigrahi, A., & Rout, S. K. (2020). DCNN-SVM: A new approach for lung cancer detection. In Recent Advances in Computer Based Systems, Processes, and Applications (pp. 97-105). CRC Press.

[30] Pani, S., Sahu, B., Mishra, J., Mohanty, S. N., & Panigrahi, A. (2022). Pragmatic Analysis of Social Web Components on Semantic Web Mining. Social Network Analysis: Theory and Applications, 83-108.

[31] Panigrahi, A., Sahu, B., Panigrahi, S. S., Khan, M. S., & Jena, A. K. (2021). Application of Blockchain as a solution to the real-world issues in health care system. In Blockchain Technology: Applications and Challenges (pp. 135-149). Springer, Cham.

[32] Sahu, B., Panigrahi, A., Pani, S., Swagatika, S., Singh, D., & Kumar, S. (2020, July). A crow particle swarm optimization algorithm with deep neural network (CPSO-DNN) for high dimensional data analysis. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 0357-0362). IEEE.