International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Identifying Fake News on ISOT Data using Stemming Method with a Subdomain of AI Algorithms

[1]Madhura Hemant Kulkarni, [2]Dr. Ravindra Sadashivrao Apare, [3]Gururaj L. Kulkarni, [4]Mukesh Singh, [5]Arun Pratap Srivastava, [6]Krishna Kant Dixit, [7]A. Deepak, [8]Dr. Anurag Shrivastava

Submitted: 03/02/2024 **Revised**: 11/03/2024 **Accepted**: 17/03/2024

Abstract: Nowadays, social media platforms have played a significant role in disseminating information throughout the world without any hindrance. Some people take this opportunity to propagate fake news in order to make money, by damaging the reputations of others. To tackle this issue, we proposed a methodology for detecting fake news on social media. This methodology extracts a feature of TF-IDF using N grams and Word2Vec in two ways i) with stemming method ii) without stemming method. Both of the process is performed and they fed an into supervised machine learning algorithms (ML) such as logistic regression (LR), random forest (RF), support vector machine (SVM), gradient boosting (Grad), adaptive boosting (Adaboost), and stochastic gradient descent (SGD) to detect a fake information. Evaluation shows that the unigram gives a better result with random forest when compared to the bigram and trigram. All classification algorithms were outperformed by Trigram. Unigram is more exact both with and without a stemming method. Word2vec has lower accuracy to detect fake information in the given dataset.

Keywords: *Fake news, Social media, TF-IDF, Word2Vec, Machine learning*

1. Introduction

In recent years, online platforms are mostly used to share the information from one person to others globally. Social media like Facebook, Twitter, Instagram and WhatsApp have a huge impact on people to make some decisions. Most of the decisions are taken based on the opinion of the person or the news but some people take this as an advantage to make profit. To identify fake news, First information needs to be verified before it spreads. Second, authentication and motivation of the news should be verified.

Fake or phony information/ news:

"Fake or dishonest information presented as a new article" is how fake news is defined. The news is based on some true news to either hurt or make people laugh, similar to memes. A variety of things can be considered "fake news," including satire, false connections, misleading content, false context, impostor content, manipulated content, and fabricated content.

1Assistant Professor, Department of Biotechnology, Willingdon College, Sangli
madhura09k@gmail.com
2Associate Professor, IT Department, Trinity College of Engineering and Research, SPPU Pune
ravi.apare@gmail.com
3Associate Professor, Department of Computer Science and Engineering, Vardhaman College of Engineering, Shamshabad, Hyderabad – 501218, Telangana
gururajlkulkarni@gmail.com
4Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun
mukeshsingh.cse@geu.ac.in
5Lloyd Institute of Engineering & Technology, Greater Noida
apsvgi@gmail.com
6 Department of Electrical Engineering, GLA University, Mathura
krishnakant.dixit@gla.ac.in
7Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha
University, Chennai, Tamilnadu
deepakarun@saveetha.com
8Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamilnadu
**anuragshri76@gmail.com*

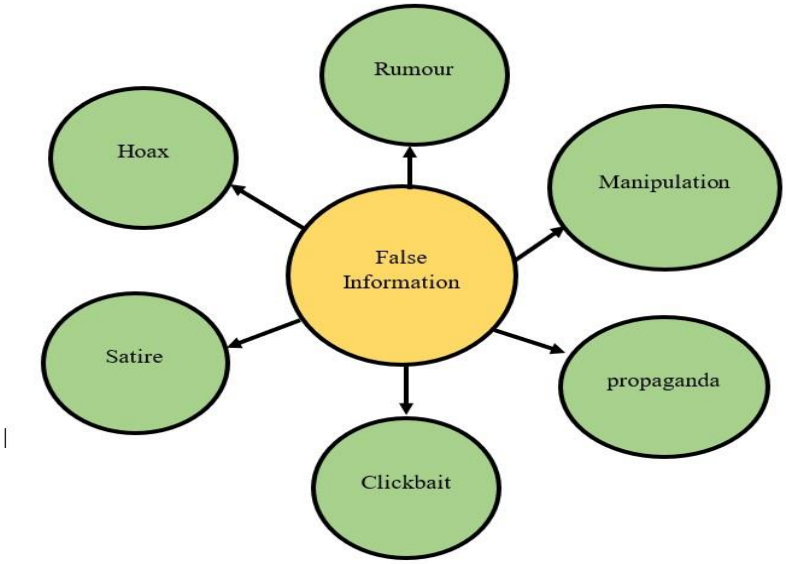

Fig 1. Types of False Information

Today, we can disseminate false information at an unprecedentedly high rate and scale due to low-cost online platforms like social media and websites [1]. A study found that fake news has a 70% higher chance of spreading than legitimate news. The public's reaction to fake news frequently reflects negative feelings of surprise, fear, and disgust. People's daily lives are impacted by fake news, which also influences their beliefs and may cause them to make poor decisions by manipulating their thoughts and feelings. The spread of false information on social media has a negative impact on society in a variety of areas, including politics, economics, social issues, health, technology, and sports [2]. The features are crucial for spotting fake news, but the majority of them are irrelevant. Therefore, the feature of the Term Frequency-Inverse Document Frequency with N-grams and Word2Vec will be the base of this work.

The model uses various machine learning algorithms with semantic features to analyze online articles and compares Naive Bayes, Recurrent Neural Network (RNN), and random forest algorithms with varied linguistic features. TF and TF-IDF numerical features are extracted using semantic features. The correlated TF and TF-IDF features are calculated using N-grams. The performance of the model is increased by bigrams and random forests [3]. To detect fake news, the model used text-based news features without any other related metadata. Stylometric features are divided into three groups in this methodology. For feature extraction in word vectors, the model used a bag of words count, BOW, TFIDF, CBOW, and Skip gram (SG). Two techniques are used after obtaining vectors for each token in the vocabulary. The performance of BOW count and BOW TF-IDF is evaluated using random forest, naive bayes, and logistic regression in four scenarios. The boosting method uses a gradient boosting algorithm, which provides better accuracy than the other two ensemble methods [4]. Hadeer Amed, Issa Traore, and Sherif Saad [5] extracted N-gram, TF, and TF-IDF characteristics. In order to represent the relevant documents, a feature extraction matrix is created after retrieving N-grams. TF and TF-IDF feature extraction are used with various N-gram sizes (1000, 5000, 10000, 50000) and classification algorithms. Marco L. Della Vedova, et al. [6] set the n size for n-gram features to seven, and every word in the documents was stemmed, with each text represented as a vector of TF-IDF frequencies on the stem's vocabulary.

2. Methodology

The proposed technique is the most effective at spotting fake news because it is based on an analysis of features. The TF-IDF and count vectorizer' features have already been studied [7].

This study incorporates Word2Vec and N-gram with TF-IDF features.

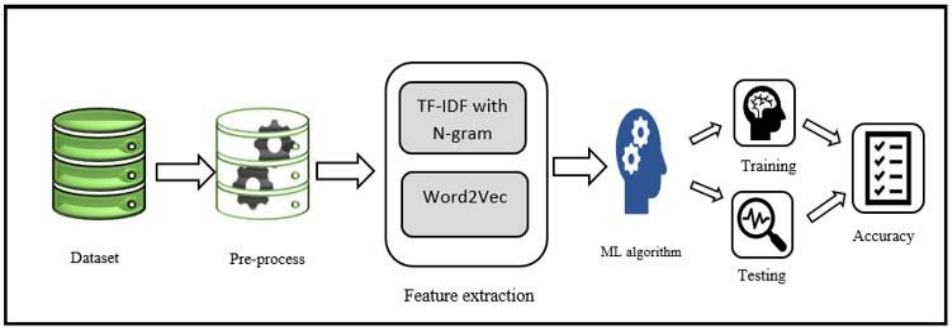

Fig. 2 Proposed Methodology

Proposed work algorithm:

Step 1: Import libraries

Step 2: Extract the dataset and read the csv data file.

Step 3: Label true as 0 and false as 1.

Step 4: Pre-process the data by removing punctuation, tokenization, stop-words, and applying porter stemmer method.

Step 5: Apply features TF-IDF and Word2Vec to convert vector form.

Step 6: Train the 80% of the data in an ML algorithm.

Step 7: Test the 20% data for evaluating the model.

Step 8: Accuracy score metrics are used to evaluate the model.

3. Materials & Methods

DATASET: The ISOT dataset, or fake and real news site Kaggle, is where the data was gathered. With 44898 rows, the dataset has four columns: title, text, subject, and date. Two CSV files, one fake and one true, are used to separate the dataset. The real CSV file has 21418 rows, but the fraudulent one has 23524 rows. There are 44942 rows total in the csv file, 44 of which are null values. Government news, Middle East news, US news, political news, and global news are all included in the article. The dataset is divided into two parts: training (80%) and testing (20%).

Table1. Types of Articles in ISOT Dataset

News	Number of articles	Subject	
		Type of News	Size
Fake	23481	Government	1570
		Middle east	778
		US	783
		Left	4459
		Politics	6841
		General News	9050
Real	21417	World	10145
		Politics	11272

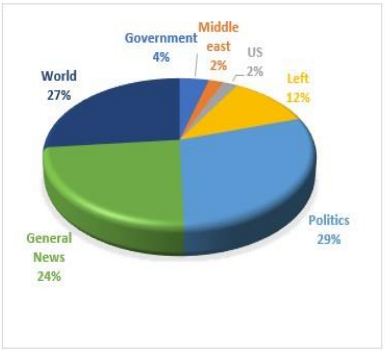

Table 2. The Description of fake and real dataset

Column	Description
Title	The title of the article
Text	The text of the article
Subject	The subject of the article
Date	The date at which the article was posted

Data Preprocessing: Social media data is largely unstructured, so it needs to be pre-processed. The majority of them involve casual language, including slang, typos, and poor grammar. It is essential to establish methods for resource utilization so that decisions may be made with knowledge in the quest for improved performance and dependability. Data needs to be cleaned before it can be used for predictive modeling in order to yield better insights. The news training data were subjected to some basic pre-processing for this purpose [8]. The components of this stage were

Fig 3. Preprocessing

Data Cleaning: Cleaning up the text data is crucial for identifying features that are going to be required by the machine learning system.

Remove punctuation: The punctuation removal approach will help treat all text equally.

Tokenization: Tokenization is the process of separating a statement into phrases, symbols, words, or other significant parts called tokens. Tokenization is typically used to identify meaningful keywords.

Remove stop-words: Many terms in papers recur quite frequently. However, the words are basically worthless as they're employed to link words together in a sentence.

Stemming: Stemming is the method of combining the different forms of a phrase into a common representation, the stem. The fundamental objective of stemming is to lower the frequency of derived terms. For example, terms such as hack, hacked, and hacking will be reduced to their lemma, which is the word hack. For this purpose, we employed the Porter stemmer method, which has become the most common stemming algorithm.

Feature Extraction:

TF-IDF with N grams are used to extract features for the ML models and generate feature matrix. To analyze the context of the text, we applied various forms of N-gram method, ranging from n1 to n2(i.e., Unigram and bi-gram). TF-ID assigns a weight to each word describing the relevance of the term in the document and corpus.

Word2Vec is used to extract features for the machine learning (ML) models and assign each word as a vector. The one bit in a vector is one. The main benefit of a word vector is that it captures both the position and meaning of the words in a text. Word2vec identifies a word in the context of the content, as well as lexical and synthetic matching and the link between other words. Word2vec is a commonly used neural network model for acquiring word embedding by utilizing text as an input. It produces a low-dimensional vector of the words that exist in the text corpus.

ML Algorithms: A supervised machine learning algorithm is implemented to train and evaluate the model for analyzing the attributes to identify fake news and increase its accuracy. A total of six ML algorithms is used to classify the news as bogus or authentic. They are logistic regression, random forest, support vector machine, ada boost, stochastic gradient descent classifier, and gradient boosting [7], [9].

4. Results & Discussion

The proposed work evaluated the features of TF-IDF with N grammes and W2V. Both aspects of TF-IDF and Word2Vec have been evaluated with or without stemming procedures.

Table 3. TF-IDF (With or without stemming process)

Algorithm	Without stemming			With Stemming		
	unigram	bigram	trigram	unigram	bigram	trigram
LR	98.85%	98.72%	96.61%	98.5%	98.61%	96.36%
RF	99.71%	98.52%	96.16%	99.62%	98.49%	96.01%

SVM	99.51%	99.16%	97.31%	99.28%	98.94%	96.85%
Grad	99.48%	93.14%	82.05%	99.47%	92.14%	81.41%
Ada	99.54%	94.34%	81.14%	99.45%	93.4%	80.77%
SG	99.19%	98.86%	96.49%	98.92%	98.71%	96.33%

Table 3 shows TF-IDF with N grams (unigram, bigram, trigram) evaluated with max feature of 15000.

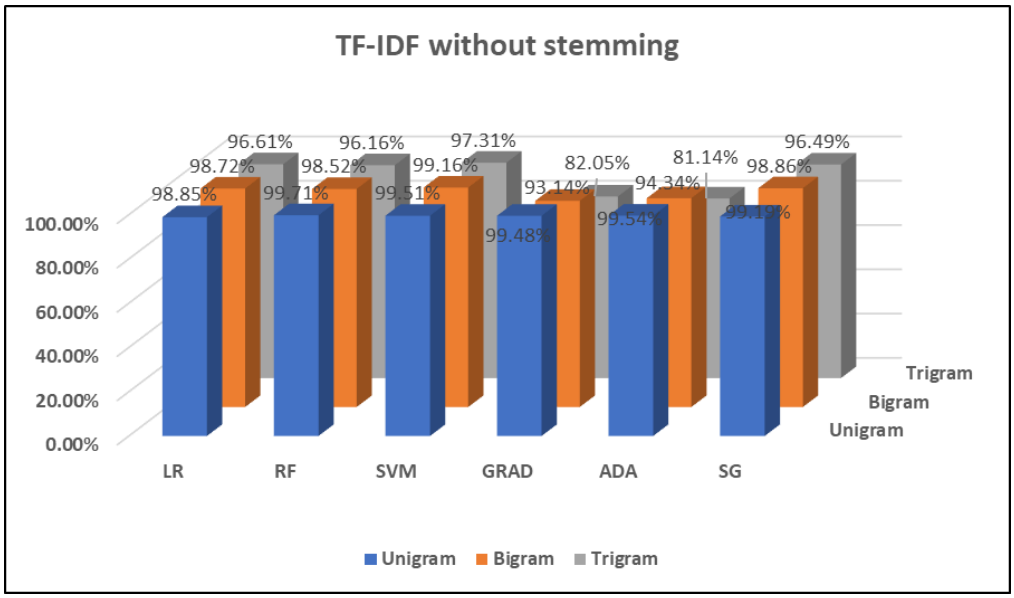

Fig 4. Result of TF-IDF without stemming

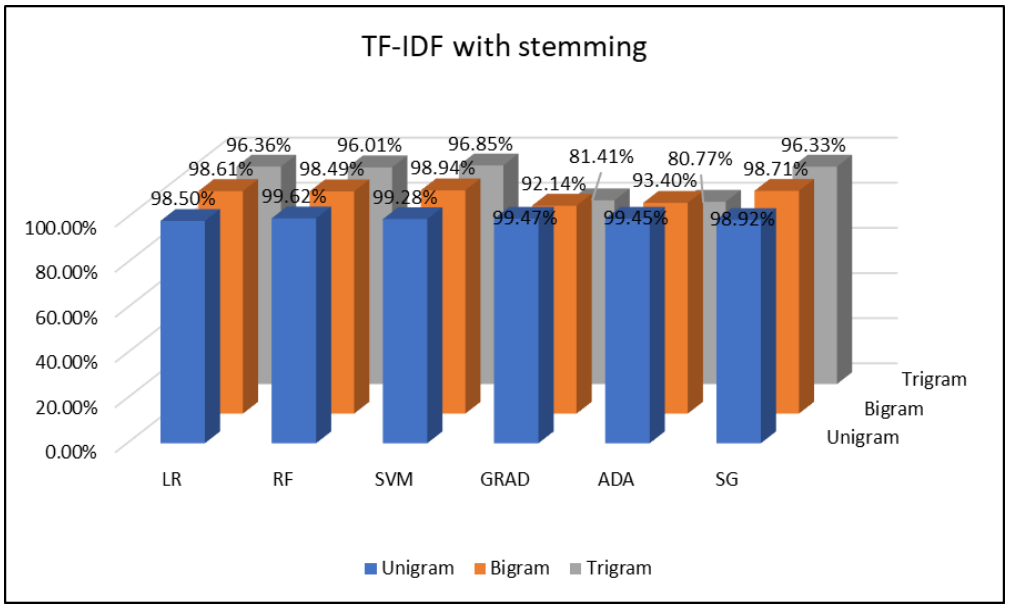

Fig 5. Result of TF-IDF with stemming

Both fig 4 & fig 5 shows that the TF-IDF (with or without stemming) with a unigram of Random Forest achieved better accuracy 99.71% and 99.62%.

Table 4. Word2Vec (With or without stemming process)

Algorithm	Word2Vec

	Without Stemming	With stemming
LR	59.5%	54.18%
RF	83.62%	88.66%
SVM	82.5%	85.68%
Grad	84.35%	87.09%
Ada	79.32%	81.55%
SGD	51.31%	54.23%

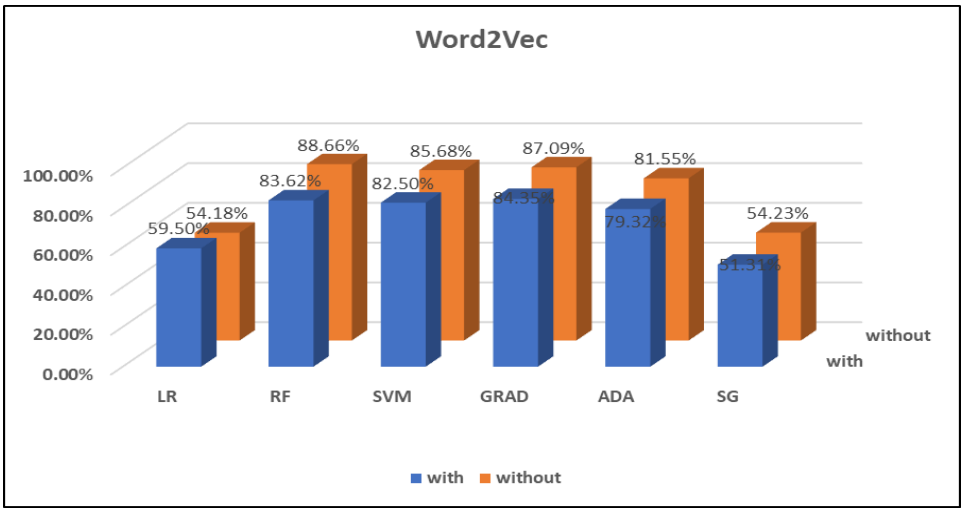

Fig 6. Result of Word2Vec

Word2Vec features analyze two phases: i) without stemming and ii) with stemming, which are shown in Table 3. In the first phase, without stemming, linear models achieved LR-59.5% and SGD-51.31%. Ensemble models of RF obtained 83.62%, Ada 99.5%, and Grad 79.32%. Support vector machine obtained 82.5%. In the second phase, with the stemming approach, linear models achieved LR-54.18% and SGD-54.23%. Ensemble models of RF reached 88.66%, Ada-81.55%, and Grad-87.09%. Support vector machine obtained 85.68%.

In N-grams, the unigram achieves a much better outcome when compared to the bigram and trigram. Trigram outperformed all classification algorithms. Unigram is more precise with and without the stemming process. Word2vec provides poorer accuracy. SVM takes a shorter time in TF-IDF for N-grams compared to TF-IDF and the count vectorizer [7], but in word2vec it consumes more. TF-IDF of a unigram with Random Forest generated a higher accuracy of 99.71%.

5. Conclusion & Future Work

This research investigates a machine-learning algorithm to focus on false news detection algorithms utilizing a supervised approach. We applied a classification algorithm for training and testing purposes that includes logistic regression, random forest, support vector machine, gradient boosting, ad boosting and stochastic gradient descent classifier. The evaluation of the model using ISOT datasets gave an accuracy of 99.71%. In the future, to identify the fake news update to date, we will build our own dataset that can update the latest news for our machine to train. All the live news and latest data will be kept in a database using a web crawler and an online database.

Reference

[1] Allein, Liesbeth, Marie-Francine Moens, and Domenico Perrotta. "Preventing profiling for ethical fake news detection." Information Processing & Management 60.2 (2023): 103206.

[2] Hamed, Suhaib Kh, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. "Fake News Detection

Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments." Sensors 23.4 (2023): 1748.

[3] Bharadwaj, Pranav, and Zongru Shao. "Fake news detection with semantic features and text mining." International Journal on Natural Language Computing (IJNLC) Vol 8 (2019).

[4] Reddy, Harita, et al. "Text-mining-based fake news detection using ensemble methods." International Journal of Automation and Computing 17.2 (2020): 210-221.

[5] Ahmed, Hadeer, Issa Traore, and Sherif Saad. "Detection of online fake news using n-gram analysis and machine learning techniques." Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1. Springer International Publishing, 2017.

[6] Della Vedova, Marco L., et al. "Automatic online fake news detection combining content and social signals." 2018 22nd conference of open innovations association (FRUCT). IEEE, 2018.

[7] Jayanthi, R., and Ms B. Jeevashri. "Fake News Detection using Supervised Machine Learning Algorithm with Feature Extraction Method." Telematique (2022): 5836-5843.

[8] Sharma, Uma, Sidarth Saran, and Shankar M. Patil. "Fake news detection using machine learning algorithms." International Journal of Creative Research Thoughts (IJCRT) 8.6 (2020): 509-518.

[9] Ahmad, Tahir, et al. "Efficient fake news detection mechanism using enhanced deep learning model." Applied Sciences 12.3 (2022): 1743.

[10] ELLOUMI, N., SLIM, B.C., SEDDIK, H. and NADRA, T., 2023. A 3D Processing Technique to Detect Lung Tumor. International Journal of Advanced Computer Science and Applications, 14(6),.

[11] . Bani Ahmad, A. Y. A. ., Kumari, D. K. ., Shukla, A. ., Deepak, A. ., Chandnani, M. ., Pundir, S. ., & Shrivastava, A. . (2023). Framework for Cloud Based Document Management System with Institutional Schema of Database. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 672–678.

[12] P. William, Anurag Shrivastava, Upendra Singh Aswal, Indradeep Kumar, Framework for Implementation of Android Automation Tool in Agro Business Sector, 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), 10.1109/ICIEM59379.2023.10167328

[13] P. William, Anurag Shrivastava, Venkata Narasimha Rao Inukollu, Viswanathan Ramasamy, Parul Madan, Implementation of Machine Learning Classification Techniques for Intrusion Detection System, 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), 10.1109/ICIEM59379.2023.10167390

[14] N Sharma, M Soni, S Kumar, R Kumar, N Deb, A Shrivastava, Supervised Machine Learning Method for Ontology-based Financial Decisions in the Stock Market, ACM Transactions on Asian and Low-Resource Language Information Processing.

[15] Ajay Reddy Yeruva, Esraa Saleh Alomari, S Rashmi, Anurag Shrivastava, Routing in Ad Hoc Networks for Classifying and Predicting Vulnerabilities, Cybernetics and Systems, Taylor & Francis, 2023

[16] P William, OJ Oyebode, G Ramu, M Gupta, D Bordoloi, A Shrivastava, Artificial intelligence based models to support water quality prediction using machine learning approach, 2023 International Conference on Circuit Power and Computing Technologie

[17] J Jose, A Shrivastava, PK Soni, N Hemalatha, S Alshahrani, CA Saleel, An analysis of the effects of nanofluid-based serpentine tube cooling enhancement in solar photovoltaic cells for green cities, Journal of Nanomaterials 2023

[18] K Murali Krishna, Amit Jain, Hardeep Singh Kang, Mithra Venkatesan, Anurag Shrivastava, Sitesh Kumar Singh, Muhammad Arif, Deelopment of the Broadband Multilayer Absorption Materials with Genetic Algorithm up to 8 GHz Frequency, Security and Communication Networks

[19] P Bagane, SG Joseph, A Singh, A Shrivastava, B Prabha, A Shrivastava, Classification of malware using Deep Learning Techniques, 2021 9th International Conference on Cyber and IT Service Management (CITSM).

[20] A Shrivastava, SK Sharma,Various arbitration algorithm for onchip (AMBA) shared bus multi-processor SoC, 2016 IEEE Students' Conference on Electrical, Electronics and Computer Science, SCEECS 509330

[21] A. Gandomi, M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management,* vol. 35, no. 2, pp. 137-144, 2015.

[22] Shrivastava, A., Chakkaravarthy, M., Shah, M.A..A Novel Approach Using Learning Algorithm for Parkinson's Disease Detection with Handwritten Sketches. In Cybernetics and Systems, 2022

[23] Shrivastava, A., Chakkaravarthy, M., Shah, M.A., A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics. In *Healthcare Analytics*, 2023, 4, 100219

[24] Shrivastava, A., Chakkaravarthy, M., Shah, M.A.,Health Monitoring based Cognitive IoT using Fast Machine Learning Technique. In *International Journal of Intelligent Systems and Applications in Engineering*, 2023, 11(6s), pp. 720–729

[25] Shrivastava, A., Rajput, N., Rajesh, P., Swarnalatha, S.R., IoT-Based Label Distribution Learning Mechanism for Autism Spectrum Disorder for Healthcare Application. In *Practical Artificial Intelligence for Internet of Medical Things: Emerging Trends, Issues, and Challenges*, 2023, pp. 305–321

[26] Boina, R., Ganage, D., Chincholkar, Y.D., .Chinthamu, N., Shrivastava, A., Enhancing Intelligence Diagnostic Accuracy Based on Machine Learning Disease Classification. In *International Journal of Intelligent Systems and Applications in Engineering*, 2023, 11(6s), pp. 765–774

[27] Shrivastava, A., Pundir, S., Sharma, A., ...Kumar, R., Khan, A.K. Control of A Virtual System with Hand Gestures. In *Proceedings - 2023 3rd International Conference on Pervasive Computing and Social Networking, ICPCSN 2023,* 2023, pp. 1716–1721

[28] A. P. Srivastava, P. Choudhary, S. A. Yadav, A. Singh and S. Sharma, A System for Remote Monitoring of Patient Body Parameters**,** International Conference on Technological Advancements and Innovations (ICTAI), 2021, pp. 238-243,