Original Article

An Improving Accuracy in Predicting the spread of Covid over Online Social Networks based on Geographical Location using Novel Autoregressive Algorithm comparing Support Vector Clustering Algorithm

Arani Girish[1], A. Shri Vindhya[2*]

[1]Research Scholar,Department of Computer Science and Engineering, Saveetha School of Engineering,Saveetha Institute of Medical and Technical Sciences,Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105
[2]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,Saveetha University, Chennai, Tamil Nadu,India. Pincode:602105.

Abstract

Aim: The purpose of this research work is to heighten the efficiency percent of geographical location identification to relieve the effect of covid using device studying classifiers by evaluating novel Logistic Regression algorithm and Random Forest algorithm.
Materials and Methods: Logistic Regression algorithm with sample size = 10, G-power (value=0.8)and Random Forest algorithm with sample size = 10 were predicted many times to evaluate the efficiency percentage. Logistic Regression is evaluated by using its weights and configurations.
Results and Discussion: Logistic Regression algorithm has better accuracy (92%) when compared to Random Forest Algorithm accuracy(21%). The results achieved with significance value p=0.680 (p>0.05) shows that two groups are statistically insignificant.Conclusion: Logistic Regression algorithm performed significantly better than the Random Forest algorithm.
Keywords: Novel Autoregressive, Support Vector Clustering algorithm, Efficiency, Covid, Geographical Location Identification, Covid Hotspot**.**

DOI: 10.47750/pnr.2022.13.S04.034

INTRODUCTION

The purpose of this research is to predict the pandemic Spread of Covid the use of Geographical Location Identification using the novel Autoregressive Algorithm and to evaluate proposed algorithms with Support Vector Clustering Algorithm. In this case the experiment aims to improve the rate of efficiency in detecting social distancing violations between locations (Guo, Li, and Temkin-Greener 2022). With the recent wave and rapid transmission of the COVID-19 pandemic (Wong et al. 2022), spread of covid is simply increasing. The proposed approachchanged intopreviously used to end upaware about the hotspots of covid infected areas. Geographical location identification model proposed in this research to identify the covid hotspots by using Novel Random Forest Algorithm (Tuan 1986). A relatedresearch concludes that the Autoregressive Algorithm has betterefficiency and quicker detection time. The applications of this researchapproachenables in identifying the covid hotspots with geographical location(Palser, Lazerwitz, and Fotopoulou 2022).
A second strand of research, which has acquired muchinterestduring the COVID-19 pandemic, makes use of geolocation information to trace covid hotspots. Around 126 articles had beenpublished in IEEE xplore and 182 articles had beenposted in google scholar on identifying hotspots on detecting physical distancing.These recordshave been used to find out the determinants and outcomes of social distancing behavior(Alkhatib 2020). A methodology for spread of covid using geographical location identification using hotspots with social distancing during the covid-19 crisis (Matthew Seah 2022). The geographic shape of social networks is usuallytough to measure on a national or international scale. In this paper, we triumph over this challengebyusing aggregated data from twitter to calculate social connections among places. A close

research discusses the challenges within hotspot identifying the covid hotspot areas (Erber et al. 2022). We then show that these connectedness measures can help predict the geographic growth of communicable diseases such as COVID-19 (Burrough, Lloyd, and Mcdonnell 2015). In every othermethodthe writer proposed a technique that overcomes detection disasters and efficiently detects the covid hotspots whilesurprisingalternatepasses off in geographical location(Lewis 2020).Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021)

The drawback of the present Geographical LocationIdentification is lessperformance in detecting objects in hotspots specificallywhilethey're in movement and detecting much less frames in step with second. The main aim of our proposed system is to improve efficiency in detecting social geographical location identifications for identifying covid 19 using the novel Autoregressive algorithm.

MATERIALS AND METHODS

This research work was performed at Cyber Forensic Laboratory, Saveetha School of Engineering, SIMATS (Saveetha Institute of Medical and Technical Sciences). The proposed work contains two groups. Group1 is taken as Autoregressive and Group 2 as Support Vector Clustering (Davis et al. 1982). The Autoregressive algorithm and Support Vector Clustering algorithm were evaluated a different number of times with a sample size of 10(Gupta et al. 2019)with confidence interval of 95%, and with pretest power of 80% and maximum accepted error is fixed as 0.05.

After dataset collection, the null values and unimportant content in the datasets were removed by preprocessing and data cleaning steps. After cleaning and preprocessing the data, an ideal input for the detection model is produced, which are processed into the detection model using opencv library and efficiency of both novel Autoregressive algorithm and Support Vector Clusteringalgorithm is calculated. The learning process of Autoregressive and Support Vector Clustering algorithms are given

Autoregressive Algorithm

An Autoregressive is a machine learning technique used to solve regression and classification problems. It is used to predict when there will be a correlation between the values in a time series and the values that precede and follow them (Davis et al. 1982).Autoregressive Algorithm The process is essentially a linear regression of data in the current series against one or more values passed in the same series.Table. 5 shows the Pseudocode for Random Forest from dataset processing to output generation.

Support Vector Clustering Algorithm

Support vector clustering Algorithm data points are mapped from the data space to a high-dimensional feature space using a Gaussian kernel (Hacoupian 2013). These contours are disturbed as cluster boundaries. The points enclosed by each dividing contour are assigned to the same group.The result is the effect of each variable on the exposure odds ratio of the observed event of interest. Table 6 predicts the Support Vector Clustering Algorithm.

The detection model gives the following procedure. Table 1 gives the source of the covid affected hotspots. The hotspotsare processed using the open csv library and each frame is detected at once. It is represented in a graph that can show the tweets count at the location selected from the frame and then transformed into the top-down view. The location for every hotspot can be estimated based on the top-down view. The distance between each hotspot can be measured and scaled. According to the preset minimum distance, any distance less than the acceptable distance between any two locations as a warning.Python programming language was used to implement this work.

Hardware configuration references the details and system resource settings allotted for specific devices, the following are minimum hardware requirements to implement this model processor: intel i5, RAM 8GB, 500 GB HDD storage.

Software specifications are concerned with the resources that must be installed in the target system in order to get an application to work. The minimal software specifications for this model to work are windows operating system version 7/8/10 python programming language version 3 or above, Google collab.

Statistical Analysis

IBM SPSS is used for statistical analysis.The independent variable is user id and the dependent variable is hotspot location and country (George and Mallery 2019; Aldrich 2018). The independent T-Test analysis is performed.

RESULTS

Table 1 shows the dataset for several users and their locations. Table 2 represents the simulated efficiency analysis of novel AutoregressiveSupport Vector Clustering algorithms. Table 3 represents group statistical analysis with the mean value of

91.80 and 21.40, standard deviation of 4.085 and 6.085 for novel Autoregressive and Support Vector Clustering algorithms respectively. Table 4 represents the independent T-test analysis of both the groups with significance value p= 0.680 (p>0.05) states that both groups are statistically insignificant. Figure 1 shows Architecture for Geographic location identification using novel Auto regression from dataset processing to output of each location.

Figure2 shows the bar graph analysis based on efficiencies of two algorithms. The mean efficiencies of novel Autoregressive and Support Vector Clustering are 92% and 21% respectively. From the results obtained it is inferred that the novel Autoregressive algorithm is more efficient than the Support Vector Cluster algorithm.

DISCUSSION

In this research work, Autoregressive Algorithm and Support Vector Clustering Algorithm were evaluated for predicting the efficiency of geographical location identification in covid hotspots using twitter. After validating the two models using the same datasets it was observed that the Autoregressive algorithm has better performance than the Support Vector Clustering algorithm. The novel Autoregressive detection model for finding covid hotspots and geographical location identification between hotspots was developed, which makes use of opencv library to process the identifying the location. The proposed model detects the hotspots and their distances using Autoregressive, and displays a user's tweets by location detected. The datasets from different ranges of location helped in improving the efficiency percentage.

The research resulted in less development of efficiency in detecting geographical location identification between covid hotspots (Bouffanais and Lim 2020). A similar work in identification of covid hotspots (Charandabi and Gholami 2021) and areas distance calculation using Support Vector Clustering algorithm (Bisen and Dubey 2018). The results achieved after all iterations on each dataset showed a constant 92% efficiency. The model proposed resulted in achieving more than 71 % increase of efficiency compared to the existing model. The similar research carried out is about covid hotspots identification which is best for future researchers who are interested in covid hotspots detection (Nasir 2003). There are no such opposite findings with regardance of existing item detection for geolocation identification of covid hotspots.

The proposed system of this paper is to get the information about the Geographical Location Identification on predicting the hotspots of covid 19 pandemic (Charandabi and Gholami 2021). Although our proposed system is faster than Support Vector Clustering Algorithm Algorithm in detecting covid hotspot areas, it is generally extracting only limited features from the hotspots and it is limited to testing only hotspots data. Further this research work can be improved by deploying a model that identifies the covid hotspots so that wait will be less and it can be embedded in geographic locations to identify spread of covid as in this research.

CONCLUSION

In this research work, prediction of efficiency percentage for Geographic Location Identification using Auto regressive Algorithm appears to have enhanced efficiency (92%) when compared to Support Vector Clustering algorithm (21%). Location identification has been successfully employed for Geographical location identification for tweet counts by users. The results reveal the maximum number of true positives compared to true negatives from all the observations.

DECLARATIONS

Conflict of Interest

The author declares no conflict of interest.

Authors Contribution

Author AG was involved in data collection, data analysis, and manuscript writing. Author SVA was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The Authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the research.

1.Indigitech, Chennai, Tamil Nadu.

2.Saveetha University.

3.Saveetha Institute of Medical and Technical Sciences.

4.Saveetha School of Engineering.

REFERENCES

1. Adhinarayanan, Rajesh, Aravindh Ramakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and Damodharan Dillikannan.

2020. "Comparative Analysis on the Effect of 1-Decanol and Di-N-Butyl Ether as Additive with diesel/LDPE Blends in Compression Ignition Engine." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
2. Aldrich, James O. 2018. *Using IBM SPSS Statistics: An Interactive Hands-On Approach*. SAGE Publications.
3. Alkhatib, Ahed J. 2020. "Social Marketing and Social Distancing:" *Proceedings of DIALOGO-CONF 2020*. https://doi.org/10.18638/dialogo.2020.6.2.2.
4. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." *Biomass Conversion and Biorefinery*, August. https://doi.org/10.1007/s13399-020-00938-0.
5. Aurtherson, P. Babu, Bhanu Teja Nalla, Karthikeyan Srinivasan, Kulmani Mehar, and Yuvarajan Devarajan. 2021. "Biofuel Production from Novel Prunus Domestica Kernel Oil: Process Optimization Technique."*Biomass Conversion and Biorefinery*, May. https://doi.org/10.1007/s13399-021-01551-5.
6. Bhansali, Karan J., Kamlesh R. Balinge, Subodh U. Raut, Shubham A. Deshmukh, M. Senthil Kumar, C. Ramesh Kumar, and Pundlik R. Bhagat. 2021. "Visible Light Assisted Sulfonic Acid-Functionalized Porphyrin Comprising Benzimidazolium Moiety for Photocatalytic Transesterification of Castor Oil." *Fuel* 304 (November): 121490.
7. Bisen, Minakshi, and Amit Dubey. 2018. "An Intrusion Detection System Based on Support Vector Machine Using Hierarchical Clustering and Genetic Algorithm." *The SIJ Transactions on Computer Science Engineering & Its Applications (CSEA)*. https://doi.org/10.9756/sijcsea/v6i1/03010040101.
8. Bouffanais, Roland, and Sun Sun Lim. 2020. "Cities — Try to Predict Superspreading Hotspots for COVID-19." *Nature*. https://doi.org/10.1038/d41586-020-02072-3.
9. Burrough, Peter A., Christopher D. Lloyd, and Rachel A. Mcdonnell. 2015. *Principles of Geographical Information Systems*.
10. Charandabi, Neda Kaffash, and Amir Gholami. 2021. "COVID-19 Spatiotemporal Hotspots and Prediction Based on Wavelet and Neural Network." *COVID-19 Pandemic, Geospatial Information, and Community Resilience*. https://doi.org/10.1201/9781003181590-19.
11. Davis, Herbert T., H. Joseph Newton, Marcello Pagano, and TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS. 1982. *A Toeplitz Gram-Schmidt Algorithm for Autoregressive Modeling*.
12. Deepanraj, B., N. Senthilkumar, D. Mala, and A. Sathiamourthy. 2021. "Cashew Nut Shell Liquid as Alternate Fuel for CI Engine—optimization Approach for Performance Improvement." *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01312-4.
13. Erber, Johanna, Verena Kappler, Bernhard Haller, Hrvoje Mijočević, Ana Galhoz, Clarissa Prazeres da Costa, Friedemann Gebhardt, et al. 2022. "Infection Control Measures and Prevalence of SARS-CoV-2 IgG among 4,554 University Hospital Employees, Munich, Germany." *Emerging Infectious Diseases* 28 (3): 572–81.
14. George, Darren, and Paul Mallery. 2019. *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference*. Routledge.
15. Guo, Wenhan, Yue Li, and Helena Temkin-Greener. 2022. "COVID-19 in Assisted Living Communities: Neighborhood Deprivation and State Social Distancing Policies Matter." *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America*, February, 1–18.
16. Gupta, Savyasachi, Rudraksh Kapil, Goutham Kanahasabai, Shreyas Srinivas Joshi, and Aniruddha Srinivas Joshi. 2019. "SD-Measure: A Social Distancing Detector." 2019. https://ieeexplore.ieee.org/document/9242628.
17. Hacoupian, Yourik. 2013. *Mining Aspects Through Cluster Analysis Using Support Vector Machines and Genetic Algorithms*.
18. Jayanth, Bellappu Venkat, Melvin Victor Depoures, Gopal Kaliyaperumal, Damodharan Dillikannan, Dilipsingh Jawahar, Kumaran Palani, and Ganesha Prasad Meravanigee Shivappa. 2021. "A Comprehensive Study on the Effects of Multiple Injection Strategies and Exhaust Gas Recirculation on Diesel Engine Characteristics That Utilize Waste High Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
19. Kamath, Manjunath, Subha Krishna Rao, Jaison, Sridhar, Kasthuri, Gopinath, Sivaperumal, and Shantanu Patil. 2020. "Melatonin Delivery from PCL Scaffold Enhances Glycosaminoglycans Deposition in Human Chondrocytes – Bioactive Scaffold Model for Cartilage Regeneration." *Process Biochemistry* 99 (December): 36–47.
20. Lewis, Dyani. 2020. "Why Schools Probably Aren't COVID Hotspots." *Nature*. https://doi.org/10.1038/d41586-020-02973-3.
21. Matthew Seah, K. T. 2022. "Covid Recovery and Digital Technologies - Correspondence." *International Journal of Surgery* , February, 106579.
22. Nasir, J. 2003. "Hotspots." *Clinical Genetics*. https://doi.org/10.1034/j.1399-0004.2002.610502.x-i2.
23. Palser, Eleanor R., Maia Lazerwitz, and Aikaterini Fotopoulou. 2022. "Gender and Geographical Disparity in Editorial Boards of Journals in Psychology and Neuroscience." *Nature Neuroscience*, February. https://doi.org/10.1038/s41593-022-01012-w.
24. Rajasekaran, S., D. Damodharan, K. Gopal, B. Rajesh Kumar, and Melvin Victor De Poures. 2020. "Collective Influence of 1-Decanol Addition, Injection Pressure and EGR on Diesel Engine Characteristics Fueled with diesel/LDPE Oil Blends." *Fuel* 277 (October): 118166.
25. Rajesh, A., K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A. P. Sathiyagnanam, and D. Damodharan. 2020. "Effect of Anisole Addition to Waste Cooking Oil Methyl Ester on Combustion, Emission and Performance Characteristics of a DI Diesel Engine without Any Modifications." *Fuel* 278 (October): 118315.
26. Raju, P., K. Raja, K. Lingadurai, T. Maridurai, and S. C. Prasanna. 2021. "Glass/Caryota Urens Hybridized Fibre-Reinforced nanoclay/SiC Toughened Epoxy Hybrid Composite: Mechanical, Drop Load Impact, Hydrophobicity and Fatigue Behaviour." *Biomass Conversion and Biorefinery*, March. https://doi.org/10.1007/s13399-021-01427-8.
27. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
28. Shanmugam, Rajasekaran, Damodharan Dillikannan, Gopal Kaliyaperumal, Melvin Victor De Poures, and Rajesh Kumar Babu. 2021. "A Comprehensive Study on the Effects of 1-Decanol, Compression Ratio and Exhaust Gas Recirculation on Diesel Engine Characteristics Powered with Low Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 43 (23): 3064–81.
29. Sudhakar, M. P., Merlyn Ravel, and K. Perumal. 2021. "Pretreatment and Process Optimization of Bioethanol Production from Spent Biomass of Ganoderma Lucidum Using Saccharomyces Cerevisiae." *Fuel* 306 (December): 121680.
30. Tuan, Pham Dinh. 1986. *An Efficient Algorithm for Maximum Likelihood Estimation for Autoregressive Moving Average Model*.
31. Wong, Shuk-Ching, Jonathan Hon-Kwan Chen, Lithia Lai-Ha Yuen, Veronica Wing-Man Chan, Christine Ho-Yan AuYeung, Sally Sau-Man Leung, Simon Yung-Chun So, et al. 2022. "Air Dispersal of Meticillin-Resistant Staphylococcus Aureus in Residential Care Homes for the Elderly: Implication in Transmission during COVID-19 Pandemic." *The Journal of Hospital Infection*, February. https://doi.org/10.1016/j.jhin.2022.02.012.

TABLES AND FIGURES

Table 1. Dataset name- Name, Extension and Source.

S.NO	DATASET NAME	DATASET EXTENSION	DATASET SOURCE
1	Covid19_tweets	CSV	Kaggle.com
2	world cities	CSV	Kaggle.com

Table 2. Efficiency of Autoregressive and Support Vector Clustering. The Autoregressive algorithm is 71% more efficient than the Support Vector Clustering algorithm.

ITERATION NO.	Autoregressive (%)	Support Vector Clustering (%)
1	99	28
2	96	26
3	94	24
4	93	23
5	92	22
6	91	21
7	89	20
8	88	19
9	87	18
10	86	17

Table 3. Group Statistics of Autoregressive and Support Vector Clustering algorithm with the mean valueof 91.80% and 21.40%

GROUPS	N	Mean(%)	Std.Deviation	Std.Error Mean
AR	10	91.80	4.290	1.356

SVC	10	21.40	4.061	1.284

Table 4. Independent sample T-test is performed for the two groups for significance and standard error determination. The significance value p=0.680 (p>0.05) shows that two groups are statistically insignificant.

		Levene's Test for Equality of Variance		**T-test for Equality of Means**						
									95% Confidence Interval of the Difference	
	Equal Variance	**F**	**Sig**	**t**	**df**	**Sig (2-tailed)**	**Mean Difference**	**Std. Error Differ-ence**	**Lower**	**Upper**
Accuracy	**Assumed**	0.176	.680	37.315	18	.001	70.400	1.868	66.475	74.324
	Not Assumed			37.315	17.946	.001	70.400	1.868	66.475	74.324

Table 5. Pseudocode for novel Auto Regressive algorithm.

1. Start the program
2.Load the data(train set, test set).
3. Remove unwanted variables from the dataset
4. Visualization of data set
5. Import Autoregressive Model from sklearn.linear_model
6. Fit train set and test set to Autoregressive model.
7. Do K-fold cross validation on our train set

8. Predict the geographical location identification of train set
9. Measure the accuracy of the model on test set
10.Compile model
11.End the program

Table 6. Pseudocode for Support Vector Clusteringalgorithm.

1. Start the program
2.Load the data(train set, test set).
3. Remove unwanted variables from the dataset
4. Visualization of data set
5. Import Support Vector Clustering model from sklearn.linear_model
6. Fit train set and test set to Support Vector Clustering model.
7. Do K-fold cross validation on our train set
8. Predict the geographical location identification of train set
9. Measure the accuracy of the model on test set
10.Compile model
11.End the program

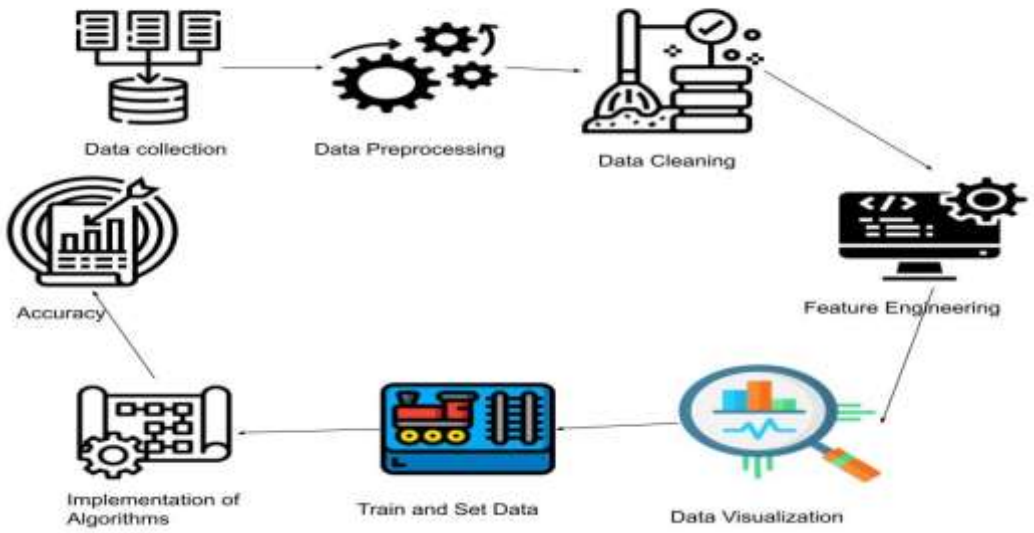

Fig. 1. Architecture for Geographic location identification using novel Auto regressive from dataset collection to output of Accuracy location.

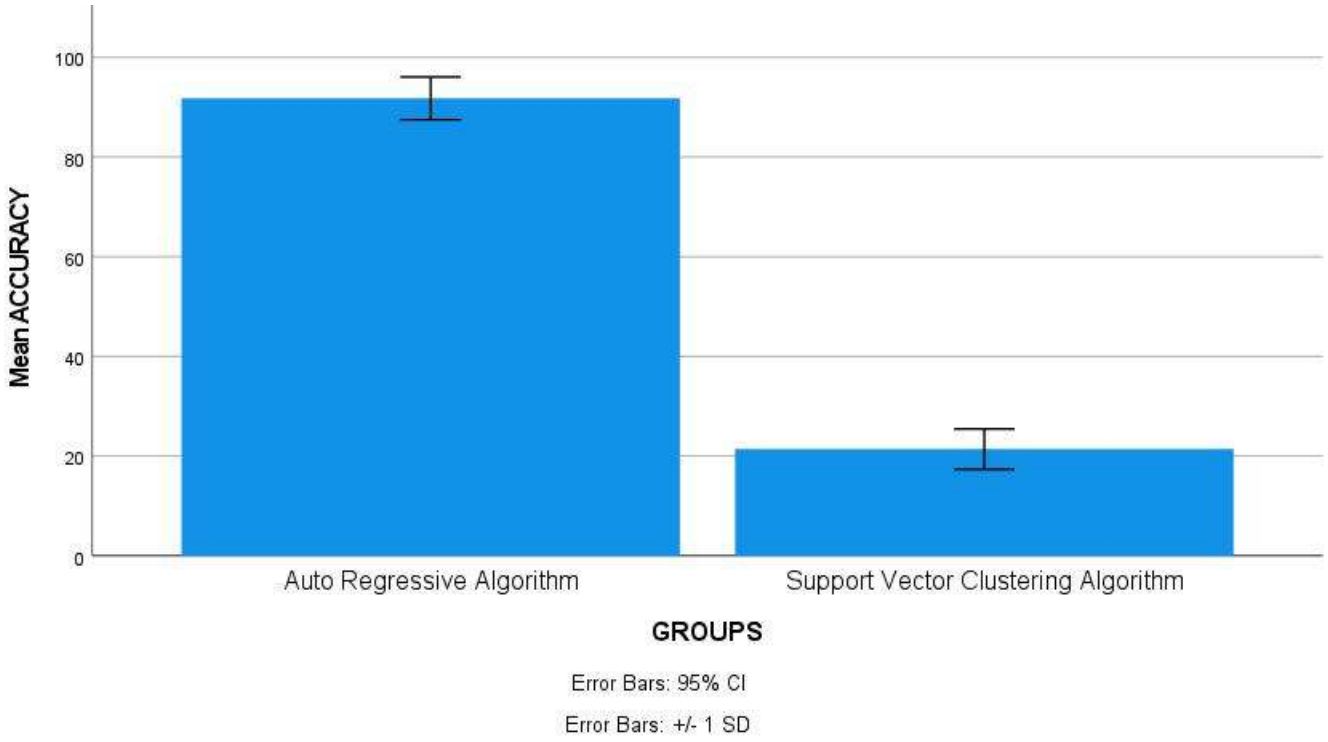

Fig. 2. Bar graph analysis of Novel Autoregressive algorithm and Support Vector Clustering algorithm. Graphical representation shows the mean efficiency of 92% and 21% for the proposed algorithm Autoregressive and Support Vector Clustering respectively. X-axis : Autoregressive vs Support Vector Clustering, Y-axis : Mean precision ± 1 SD.