International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Machine Learning for The Diagnosis of Covid-19

Vaibhavi Sujit Dhumal, Afsha Akkalkot, Arunadevi Khaple
Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India

Article Info
Volume 8, Issue 2
Page Number : 390-396

Publication Issue :
March-April-2022

Article History
Accepted: 10 March 2022
Published: 22 March 2022

ABSTRACT

A singular coronavirus (SARS-CoV-2) is an unusual viral pneumonia in sufferers, first determined in overdue December 2019, latter it declared a virus via world health corporations because of its deadly consequences on public health. In this present, cases of COVID-19 pandemic are exponentially increasing every day within the entire international. here, we're detecting the COVID-19 cases, i.e., showed, demise, and cured cases in India handiest. We are performing this evaluation based totally at the cases taking place in the world in chronological dates. Our dataset incorporates a couple of instructions so we're performing multi-elegance classification. in this dataset, first, we completed statistics cleaning and feature choice, then done forecasting of all lessons the usage of random forest, linear model, assist vector machine, selection tree, and neural network, in which random wooded area model outperformed the others, therefore, the random forest is used for prediction and analysis of all the outcomes. The okay-fold move-validation is carried out to measure the consistency of the version.

Key words: Coronavirus; COVID-19; Respiratory System; Classification Techniques; Random Forest

I. INTRODUCTION

In overdue Dec. 2019, the sector came to recognise about a lethal coronavirus disease in Wuhan, China. soon, this sickness started to unfold in specific international locations. This turned into initially named the 2019 novel coronavirus through the world fitness company (WHO). Later, in Feb. 11, 2020, the WHO formally named the disease coronavirus disease 2019 (COVID-19). The Coronavirus study group of the global Committee termed the virus that brought on COVID-19 an intense acute respiration syndrome coronavirus 2 (SARS-CoV-2) on Feb. 11, 2020. The Chinese scientists swiftly remoted a SARS-CoV-2 from a patient inside a short time on Jan. 7, 2020 and genome sequenced the SARS-CoV-2. As of Apr. 22, 2020, there have been 24, seventy-one,136 confirmed cases and 169,006 deaths globally.

In India, the first case of coronavirus sickness 2019 (COVID-19) was announced on thirtieth January 2020. This virus extends to the entire of India (of their extraordinary districts) until April 2020 end. In India, the whole cases announced were 5734 in which 472 were recovered and 166 humans have been lifeless till ninth April 2020. In India, the entire cases announced were 236 184 in which 113 233 were recovered and 6649 human beings had been lifeless till sixth June

2020. After this date, sparkling cases are still entering mild every day that is around 10 000.

The increase within the wide variety of cases in COVID-19 is threatening to overwhelm health systems around the sector with the call for in depth care unit beds a way above the present potential. consistent with WHO, more than 50 COVID-19 vaccine applicants are present in trial degree. therefore, preventive measures want to be taken to avoid becoming infected via this virus. normally, the virus receives into human frame through the eyes, throat, and nose, carried by way of the hand. The WHO has asked people to preserve personal hygiene through washing palms with cleaning soap and water frequently, and covering the mouth and nostril with an elbow or a tissue while coughing or sneezing. furthermore, cleansing surfaces with disinfectants and keeping social distancing are crucial to save you this ailment. In many nations, the spread of SARS-CoV-2 has resulted in the loss of trying out kits to diagnose the virus. There are not enough trying out kits or trained personnel to test the virus in suspected patients. In some instances, medical doctors, nurses, and medical aid team participants are also becoming inflamed, or they have to quarantine, which makes the situation extra difficult for sufferers. therefore, there may be a want for the statistics-pushed diagnosis of COVID-19 patients. This chapter makes a speciality of the automated detection of COVID-19 using a dataset to be had in https://www.kaggle.com/einsteindata4u/covid19. First, the circumstance of COVID-19 in the world is visualized the usage of some of tables and graphs. This shows showed deaths and recovered cases in distinct nations and continents. 2d, gadget gaining knowledge of strategies are used on a dataset of COVID-19 sufferers. each characteristic choice and class algorithms are carried out, and it's miles shown that for the given dataset, the COVID-19 sickness can be predicted reliably.

II. VISUALIZATION OF THE SPREAD OF CORONAVIRUS DISEASE 2019

In this phase, the spread of COVID-19 is provided via statistics visualization. for visualization functions, the dataset is accrued from an internet site. Fig. 9.1 illustrates parent 9.1 Continentally mentioned cases (as much as April 2020).

176 statistics science for COVID-19 continentally mentioned cases for which showed deaths and recovered or active instances are discussed. The mortality fee consistent with 100 humans is also defined. Fig. nine.1 suggests that the variety of showed instances is the very best on the European continent (11, 65, and 661, respectively). The mortality charge is likewise the highest in Europe (9.55%). Fig. 9.2 depicts the list of pinnacle 10 international locations of confirmed COVID-19 instances, recovered cases, death instances and energetic instances.

continent	Confirmed	Deaths	Recovered	Active	Incident_Rate	Mortality Rate (per 100)
Africa	25934	1240	6964	17730	249.022188	4.780000
Asia	422714	15862	199153	207699	1449.915987	3.750000
Australia	8024	83	5197	2744	57.902293	1.030000
Europe	1165661	111264	364944	689453	8528.897067	9.550000
North America	908771	50390	95369	763012	746.250242	5.540000
Others	1870	36	826	1008	1152.929553	1.930000
South America	97031	4595	41315	51121	252.352363	4.740000

FIGURE 9.1 Continentally reported cases (up to Apr. 23, 2020).

III. METHODOLOGY

This dataset became generated from patients at the clinic Israelita Albert Einstein in Sao Paulo, Brazil. The samples are collected anonymously acting the SARS-CoV-2 contrary transcriptase polymerase chain response and extra laboratory exams. The statistics were standardized by way of using changing the samples just so the mean fee of the samples is 0 at the same time as the standard deviation is concord. The dataset has 5644 rows and 111 columns and it is imbalanced.

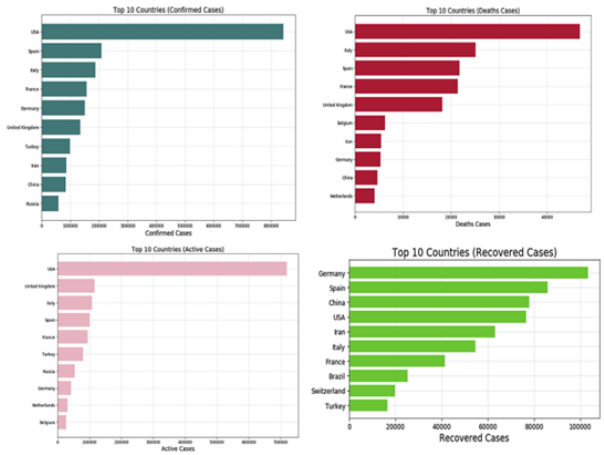

FIGURE 9.2 Top 10 countries.

There are a number of lacking values within the records samples. For this, features with greater than 99.88% of parent 9.2 top 10 countries. Fig. 9.3 indicates that the dataset became in the beginning imbalanced with 90.1% samples representing poor instances. After getting rid of attributes with at least 99.8% null values, the dataset became balanced with 51.1% representing terrible cases. table nine.1 suggests the list of dropped functions which have at the least 99.8% null values. The info of function filtering is illustrated in Fig. 9.4 below. The figures show the percentage of poor and effective cases after below sampling. thus the dataset can be considered balanced. the new dataset has 1091 rows and sixty-one columns; it will best have numerical capabilities. The goal feature is converted to 0 or 1, wherein 1 way superb and 0 approach bad. on this section, experiments are finished to categorise regular and COVID-19 sufferers the use of samples inside the dataset. This studies paintings are carried out using the Scikit-examine library of Python programming language. Steps observed in this implementation are proven in Fig. 9.5.

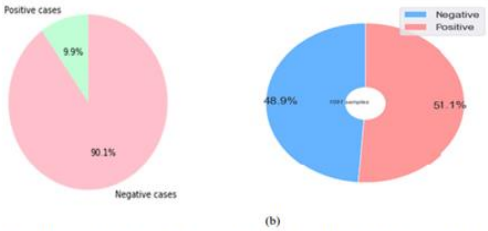

FIGURE 9.3 Positive and negative cases of data samples: (A) imbalanced case in original dataset; (B) balanced case after processing.

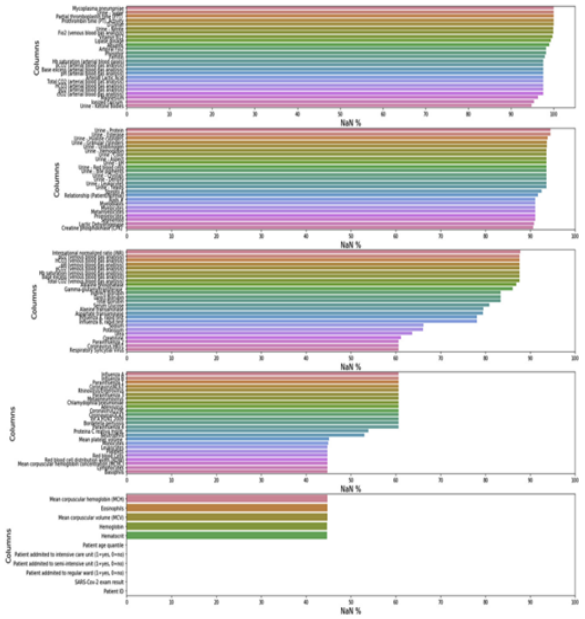

FIGURE 9.4 Feature filtering.

Some of strategies are performed, including records labelling and records filtering, which can be part of pre-processing. subsequent, essential capabilities are decided on. Statistics technological know-how for COVID-19 class algorithms are then implemented on the selected features. Some of popular category algorithms which includes random forest (RF), logistic regression (LR), choice tree (DT), and XGBoost are considered. both pass-validation (cv) and holdout techniques are taken into consideration. For, cv, the KFold () function, and for holdout, the train_test_split() feature from scikit-analyze library are used to break up the dataset. Next, the type models are geared up with the training facts and the models are then used to are expecting COVID19 samples.

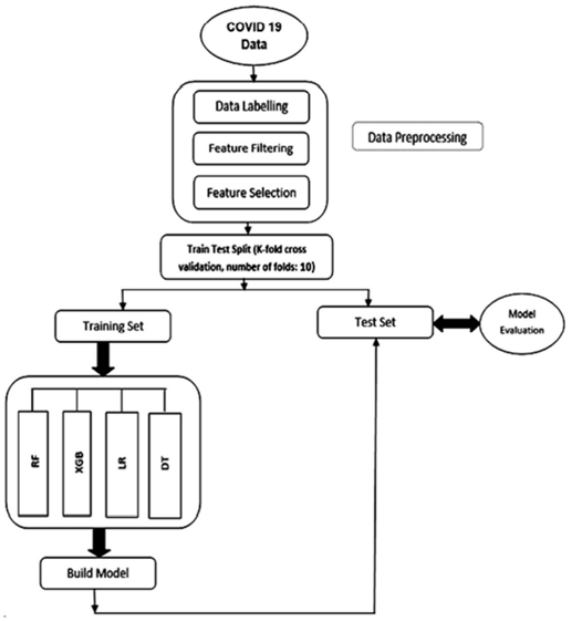

FIGURE 9.5 Workflow diagram. *DT*, decision tree; *LR*, logistic regression; *RF*, random forest; *XGB*, XGBoost.

IV. FEATURE IMPORTANCE AND FEATURE SCORING

There are some of characteristic choice algorithms. In this case, a univariate function selection approach is taken into consideration. For this, the SelectKBest() feature of the scikit-examine library is used. desk nine.2 shows the pinnacle 25 capabilities with their corresponding ratings obtained the usage of SelectKBest(). desk 9. three indicates the ranking of those top 25 capabilities. For the dataset considered, serum glucose is the fine-ranked characteristic, or the maximum influential feature in predicting a COVID-19 affected person

Table 9.2 Top 25 features using univariate selection method.

Name of feature	Score
Protein C-reactive, mg/dL	20,857.493593
Leukocytes	5800.140368
Lymphocytes	2888.050022
Neutrophils	2774.036881
Alanine transaminase	2237.563824
pO_2 (venous blood gas analysis)	1835.797530
γ-Glutamyltransferase	1584.636027
Platelets	1505.770002
Monocytes	1462.219391
Eosinophils	1441.501397
Red blood cells	1243.403849
Mean corpuscular volume	1128.205213
Aspartate transaminase	1030.129529
Indirect bilirubin	949.508049
Hematocrit	934.136514
pCO_2 (partial pressure of carbon dioxide within venous blood)	870.907532
Red blood cell distribution width	807.758742
Creatinine	733.620615
Serum glucose	729.902624
pH (venous blood gas analysis)	664.791565
Total bilirubin	587.205740
Parainfluenza 3	518.255556
Urea	450.905016
Hb saturation (venous blood gas analysis)	439.066476
Parainfluenza 4	362.666667

V. CLASSIFICATION USING MACHINE LEARNING

After deciding on pinnacle capabilities with the aid of the function selection approach, the function subset is then taken into the classifier education stage. Within the education degree, XGBoost, RF, LR and DT are employed. Fig. 9.5 above illustrates the stages of this implementation. Fig. 9.5 above suggests that the dataset is initially pre-processed, followed with the aid of the characteristic choice technique. subsequent, the statistics samples are break up into training and testing samples. Then, the training records is used to match a classifier model. The checking out records are then implemented to the model to expect the target: in this example, COVID-19. Sooner or later, the testing goal value that is the real value is compared with the anticipated cost.

5.1 XGBoost

XGBoost is a famous shape of gradient boosting set of rules designed for top-rated hardware use. It's miles an implementation of gradient-boosted DTs. XGBoost can penalize a version for complexity using L1 and L2 regularization in which regularization prevents overfitting of the XGBoost model. Regularization helps prevent overfitting.

Algorithm 1 shows how XGBoost is used to classify COVID-19 patients.

Algorithm 1. Detection of positive COVID-19 patients using XGBoost

Input: List of features

Output: classification file, confusion matrix, receiver running feature (ROC) curve

Process:

1. Standardize the chosen functions the use of StandardScaler() characteristic
2. Follow XGBoost classifier the usage of XGBClassifier (base_score¼zero.five, booster¼'gbtree', gamma¼zero, learning_rate¼zero.1, max_depth¼3,

n_estimators¼100, objective¼'reg:- linear', random_state¼zero) feature on the chosen functions three.
3. Train the version using selected functions
4. Predict end result using take a look at dataset
5. Evaluate the accuracy of the classifier using accuracy_score() function
6. Use confusion_matrix() feature to evaluate proper bad (TN), fake superb (FP), fake poor (FN), and proper high quality (TP).
7. Use classification_report() feature to calculate precision, bear in mind, and F1 score

5.2 Random forest

RF is an aggregate of more than one DTs. two crucial concepts make this set of rules random: the randomness inside the sampling of the training portion of the records and the randomness within the selection of features for the splitting nodes. The RF algorithm maintains the reliability of a large part of the dataset by means of handling any missing pattern values. algorithm 2 describes the levels of RF in classifying COVID-19 sufferers.

Algorithm 2. Detection of positive COVID-19 patients using RF

Input: A list of features based on rank

Output: Classification report, confusion matrix, accuracy

Process:

1. Standardize the selected functions using StandardScaler() feature
2. Apply RF the usage of RFClassifier (n_estimators¼100, criterion¼'gini') characteristic with some parameter on the chosen features
3. three. Train the version using selected capabilities
4. k-fold parameters for k-fold cv: thresh ¼ 0.5, k_fold_seed ¼ 13, n_folds ¼ 10
5. Predict the end result the use of test dataset
6. Evaluate the accuracy of the classifier characteristic
7. Use confusion_matrix() feature to evaluate TN, FP, FN, and TP
8. Use classification_report() feature to calculate precision, take into account, and F1 score

5.3 Decision tree and logistic regression

Other popular class algorithms are DT and LR.

Algorithm 3 describes crucial steps of DT and LR classifiers in predicting patients low with COVID-19.

Algorithm 3. Detection of positive COVID-19 patients using DT and LR

Input: List of features according to rank

Output: Classification report, confusion matrix, accuracy

Process:

1. Standardize the chosen capabilities using StandardScaler() characteristic
2. Practice DT the use of DTClassifier (criterion¼'entropy', max_depth¼five, random_state¼zero) characteristic with a few parameter on the selected functions Or follow LR the usage of LogisticRegression() feature three.
3. Teach the version the usage of selected features .
4. Parameters for k-fold cv are: thresh ¼ 0.5, n_folds ¼ 10, k_fold_seed ¼ 13
5. Predict the result the use of take a look at dataset
6. Compare the accuracy of the classifier feature
7. Use confusion_matrix() feature to assess TN, FP, FN, and TP
8. Use classification_report() function to calculate precision, don't forget, and F1 score

Table 9.5 Performance results of random forest using cross-validation.

Fold (cross-validation)	Precision (%)	Recall (%)	F1 score (%)	Testing accuracy (%)
3	91	90	90	89.9083
4	92	91	91	90.8257
5	96	95	95	95.4128
6	92	92	92	91.7431
7	94	94	94	93.5779
8	88	86	86	86.2385
9	91	91	91	90.8257
10	93	93	93	92.6606

Table 9.6 Performance results of logistic regression using cross-validation.

Fold (cross-validation)	Precision (%)	Recall (%)	F1 score (%)	Testing accuracy (%)
3	91	91	91	90.8257
4	92	92	92	91.7431
5	98	98	98	98.1651
6	93	93	93	92.6606
7	96	96	96	96.3303
8	90	90	90	89.9083
9	88	88	88	88.0734
10	93	93	93	92.6606

Table 9.7 Performance results of decision tree using cross-validation.

Fold (cross-validation)	Precision (%)	Recall (%)	F1 score (%)	Testing accuracy (%)
3	89	89	89	88.9908
4	85	85	85	85.3211
5	96	95	95	95.4128
6	89	89	89	88.9908
7	96	95	95	95.4128
8	84	83	83	83.4862
9	87	87	87	87.1559
10	89	88	88	88.0734

Table 9.8 Overall performance comparison of classifiers using cross-validation.

Classifier	Precision (%)	Recall (%)	Miss rate or false-negative rate (%)	Specificity (%)	F1 score (%)	Accuracy (%)	Confusion matrix
XGBoost	93	93	7	94.183	93	92.67	Predicted 0.0 1.0 Actual 0.0 502 49 1.0 31 509
Random forest	92	92	8	97.373	92	91.84	Predicted 0.0 1.0 Actual 0.0 519 75 1.0 14 483
Logistic regression	93	93	7	93.058	93	92.58	Predicted 0.0 1.0 Actual 0.0 496 44 1.0 37 514
Decision tree	90	90	10	92.308	90	89.73	Predicted 0.0 1.0 Actual 0.0 492 71 1.0 41 487

Table 9.4 compute statistics for average cv. further, this is completed for different classifiers the usage of Tables 9.5e9.7. table 9.8 compares the classifiers in terms of precision, recall, F1 rating, testing accuracy, leave out fee, specificity, and confusion matrix using a mean cost of cv. XGBoost has the highest accuracy price of ninety two.67%, and LR has the second one maximum accuracy of 92.58%. both XGBoost and LR have the highest values of precision, keep in mind, and F1 rating, all of which might be ninety three%. as a result, XGBoost and LR are accurate picks for classifying this particular dataset and as a result are expecting COVID-19 sufferers reliably.

Next, the overall overall performance of the classifiers is shown for the holdout technique, wherein the dataset is break up into distinctive portions of checking out and education samples. The results vary with the difference inside the splitting. In this case, we cut up the dataset so that eighty% of the data samples are used for training and the ultimate 20% are used for checking out. desk nine.nine shows the overall performance effects for distinctive classifiers whilst 20% facts samples are used for trying out. LR has the nice trying out accuracy of 94.06%. moreover, LR outperforms different classifiers in phrases of precision, don't forget, miss price, F1 rating, and AUC. table nine.eight shows that for the case of cv, XGBoost and LR have excessive testing accuracies of 92.sixty seven% and ninety two.58%, respectively, while for the case of holdout (20% checking out), LR has the very best accuracy price of 94.06%

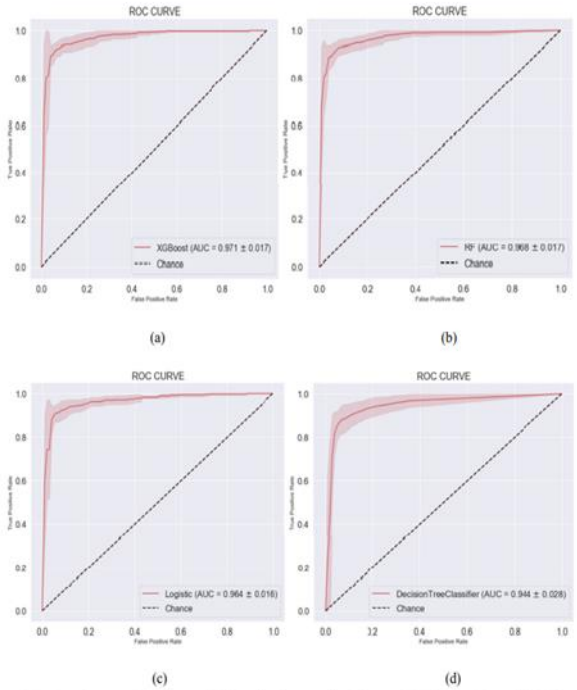

FIGURE 9.7 Receiver operating characteristic (ROC) and area under the curve (AUC) comparison of (A) XGBoost, (B) random forest, (C) logistic regression, and (D) decision tree classifiers.

Table 9.9 Performance comparison of the classifiers using holdout method.

Classifier	Precision (%)	Recall (%)	Miss rate or false negative rate (%)	F1 score (%)	Accuracy (%)	Area under the curve (%)
XGBoost	92	92	8	92	92.8374	92
Random forest	91	91	9	91	91.7574	90
Logistic regression	94	94	6	94	94.0639	94
Decision tree	89	89	11	89	88.5845	89

VI. CONCLUSION

This is an overview of the spread of COVID-19. The US and some ECU international locations which includes Italy, Spain, the United Kingdom, and Germany are closely tormented by the sickness. This chapter makes use of gadget mastering algorithms to are expecting COVID-19 for a given dataset. For this particular dataset, our experimental outcomes suggest that serum glucose is the most influential attribute in predicting COVID-19. Our results additionally show that for the case of cv, XGBoost has the best accuracy cost of ninety-two.67% and LR has the second one highest accuracy of 92.58%, whereas each XGBoost and LR have the identical ninety-three% value for precision, don't forget, and F1 score. For the case of the holdout method with 20% trying out data samples, LR well-known shows the very best trying out accuracy of 94.06%. for this reason, XGBoost and LR may be used to are expecting COVID-19.

The reliability of the prognosis results offered on this bankruptcy relies upon on the reliability of the dataset used. In destiny, with the provision of greater reliable datasets, machine gaining knowledge of algorithms should be carried out to the ones new datasets to validate the effectiveness of the classifiers. Hybrid deep getting to know algorithms can also be correctly implemented in numerous chest X-ray or computed tomography photograph datasets to detect COVID-19 patients.

VII. REFERENCES

[1]. Mondal MRH, Bharati S, Podder P. Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review of Curr Med Imaging. 2021;17(12):1403-1418. doi: 10.2174/1573405617666210713113439. PMID: 34259149.

[2]. V. K. Gupta, A. Gupta, D. Kumar and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," in Big Data Mining and Analytics, vol. 4, pp. 116-123, June 2021, doi: 10.26599/BDMA.2020.9020016.

[3]. Mainak Adhikari, M. Ambigavathi, Varun G Menon, Mohammad Hammoudeh, "Random Forest for Data Aggregation to Monitor and Predict COVID-19 Using Edge Networks", Internet of Things Magazine IEEE, vol. 4, pp. 40-44, 2021.

[4]. X. Liao, D. Zheng and X. Cao, "Coronavirus pandemic analysis through tripartite graph clustering in online social networks," in Big Data Mining and Analytics, vol. 4, pp. 242-251, Dec. 2021, doi: 10.26599/BDMA.2021.9020010.

[5]. X. Yu, M. D. Ferreira and F. V. Paulovich, "Senti-COVID19: An Interactive Visual Analytics System for Detecting Public Sentiment and Insights Regarding COVID-19 From Social Media," in IEEE Access, vol. 9, pp. 126684-126697, 2021, doi: 10.1109/ACCESS.2021.3111833.