PriMera Scientific Engineering
Volume 3 Issue 3 September 2023
ISSN: 2834-2550

PriMera Scientific Publications

Big Data Analytics in Weather Forecasting using Gradient Boosting classifiers Algorithm

Type: Review Article
Received: July 26, 2023
Published: August 18, 2023

Citation:
Kamel Maaloul., et al. "Big Data Analytics in Weather Forecasting using Gradient Boosting classifiers Algorithm". PriMera Scientific Engineering 3.3 (2023): 35-43.

Kamel Maaloul[1]* and Brahim Lejdel[2]

[1]*LIAP Laboratory, El-Oued University, Algeria*

[2]*Computer Science Department, University of El-Oued, Algeria*

***Corresponding Author:** Kamel Maaloul, LIAP Laboratory, El-Oued University, Algeria.

Abstract

Weather forecasting, a crucial and vital process in people's everyday lives, assesses the change taking place in the atmosphere's current state. Big data analytics is the practice of studying big data to uncover hidden patterns and useful information that might produce more beneficial outcomes. Big data is currently a topic of fascination for many facets of society, and the meteorological institute is no exception. Big data analytics will therefore produce better results for weather forecasting and assist forecasters in providing more accurate weather predictions. Several big data techniques and technologies have been proposed to manage and evaluate the enormous volume of weather data from various resources in order to accomplish this goal and to identify beneficial solutions. A smart city is a project that uses computers to process vast amounts of data gathered from sensors, cameras, and other devices in order to manage resources, provide services, and address problems that arise in daily life, such as the weather. Forecasting the weather is a crucial process in daily life because it assesses changes in the atmosphere's current state. A machine learning-based weather forecasting model was proposed in this paper, and it was implemented using 5 classifier algorithms, including the Random Forest classifier, the Decision Tree Algorithm, the Gaussian Naive Bayes model, the Gradient Boosting Classifier, and Artificial Neural Networks. These classifier algorithms were trained using a publicly available dataset. When the model's performance was assessed, the Gradient Boosting Classifier algorithm, which had a plus 98% predicted accuracy, came out on top.

Keywords: Weather forecasting; Big data; Machine Learning; smart city; Gradient Boosting Classifier

Introduction

The first weather predictions were made in the nineteenth century. Weather forecasting is defined as the process of analyzing atmospheric data, such as temperature, radiation, air pressure, wind speed, wind direction, humidity, and rainfall. There must be a large amount of data generated or collected in order to anticipate the weather. These data are also not well arranged. As a result, using weather data to predict the weather is a difficult task because it has too many variable variables [1].

By fusing probabilistic forecasts, for instance of snow, with data from the ground, smart cities may benefit from weather forecasting and choose where to place their snowplows. Given their known performance and available margin for demand response reasons, the smart city may also estimate the energy load of their own buildings by monitoring high local temperatures and humidity. Additionally, predicting the likelihood of traffic accidents by location and correlating precise fog forecasting to accident data. This will make it easier to set up variable speed limit signs and glare-reducing smart roadway lighting [2].

We should take into account the specific aspects of weather forecasting, such as continuity, data intensity, and multidimensional and chaotic behaviors, before we suggest an algorithm. Originally a labor-intensive human effort, weather forecasting has evolved into a computational process, necessitating high-tech machinery. The accuracy of forecasts can be impacted by a number of factors. Season, geographic location, data input accuracy, weather classifications, lead time, and validity time are some of these important variables [3].

Machine learning, a branch of data science, fundamentally manipulates data in a statistical manner. The system builds models from the data if there are input and output data, which are then used to generate predictions or finish tasks. Machine learning, on the other hand, is more resilient to disruptions and doesn't require a full understanding of the fundamental principles underlying the atmosphere. In order to forecast the weather, machine learning may be a useful alternative to physical models [4].

Big data is a term for massive, disorganized, and heterogeneous digital data. Big data processing cannot be accomplished simply or effectively using traditional data management techniques. We need look for a high-performance platform and a practical big data mining technique to obtain useful information in order to analyze this kind of data effectively. Huge data analytics is a big data search operation that aims to reveal hidden patterns, undiscovered relationships, and other relevant information to help people make better decisions. Big data has the potential to improve decision- making in the area of weather forecasting [5].

In theory, AI and machine learning will make it possible for human forecasters to work more productively, devoting less time to producing routine forecasts and more to explaining the implications and impacts of those forecasts to the general public, or, in the case of private forecasters, to their clients. We think the best way to accomplish these objectives and increase confidence in computer- generated weather forecasts is through rigorous collaboration between scientists, forecasters, and forecast users [6].

The remainder of the essay is structured as follows: The related study on weather forecasting that made use of different machine learning algorithms and techniques is described in chapter 2. The data and the methodology required to implement the artificial intelligence approach are described in Chapter 3. In Chapter 4, which is titled "Evaluation and Results," it is explained how the approach's outcome was assessed. At the end, it includes conclusions and suggestions for further work.

Related work

Since the modern world depends on data, smart cities heavily rely on artificial intelligence. Researchers have used a variety of metrological traits, features, and data produced from diverse sources to successfully apply a number of algorithms and models to forecast meteorological conditions. This section clarifies big data as the primary subject while talking about weather forecasting. Finally, the variables that affect the accuracy of weather forecasts are discussed.

Long-term memory (LSTM) technology was utilized by FENTE et al. [7] to gather a number of meteorological parameters from the National Climate Data Center. The neural network was trained using a variety of combinations. Weather variables like temperature, precipitation, wind speed, pressure, dew point visibility, and humidity were used to train the neural network used in LSTM weather forecasting. After utilizing these parameters to train the LSTM, the weather for the upcoming days is predicted.

Prior to discussing big data, Sahasrabuddhe and Jamsandekar [8] first discussed weather forecasting, covering fundamental procedures and several techniques. After that, they provided an overview of several publications and explained various data structures utilized in big data and weather forecasting. This study was not a systematic review because it lacked a classification, research methods, and discussion.

Weather forecasting's effects on the agriculture/food sector, tourism industry, sports industry, construction industry, transportation industry, disaster management industry, and energy industry were covered by Jain and Jain [9]. Then, they discussed the technical difficulties associated with weather forecasting, such as handling massive data sets, the accessibility of historical data, roadblocks associated with technology, the availability of forecast models, complexity, costly maintenance, and cost overrun. They omitted discussing the current big data methodologies and technologies. Instead of focusing on big data analytics, this study concentrated on weather forecasting topics. Because it lacked a research methodology, taxonomy, and discussion sections, it was not systematic.

Big data and climate change were first discussed in the context of agriculture by Rao [10]. Following that, the author went over big data technologies in agriculture, big data for climate smart agriculture, and a roadmap for using big data for climate smart agriculture in India. Because there were no sections devoted to the research methodology, taxonomy, or discussion, this paper was not an exhaustive review of the literature.

In order to increase forecasting efficiency and accuracy, Murugan et al. [11] suggested using a hybrid C5.0 decision tree algorithm with k-means clustering algorithm for short-range prediction. The k- means clustering algorithm was used to group together similar datasets. The prediction accuracy of the model was 90.18% when measured using mean absolute error and root mean square error.

The attributes, technology, and analytics of big data were introduced by Mittal and Sangwan [12]. Then, they offered a study of 11 publications' literature and noted difficulties in using traditional data mining algorithms to the large datasets. They compared these papers' various methods for predicting the weather, including the KNN classification algorithm, the MR-KNN and k-means algorithm on the MapReduce platform, the MapReduce and linear regression method, and the Bayesian model; they showed the outcomes in terms of the data set, the methods, the parameters, and the results of the experiments. They then composed the summary and future work. But none of the approaches we reviewed included any discussion of research methodology or taxonomy.

In order to comprehend, predict, and lower uncertainty in the WRF model when predicting precipitation as a result of the interaction of various physical processes included in the model, MOOSAVI et al. [13] proposed a novel use of machine learning techniques. To discover the connections between the configuration of the physical processes employed in the simulation and the observed model's prediction mistakes, they developed probabilistic techniques. The resolution of two significant issues involving model flaws and the identification of physical processes is then accomplished using these relationships.

Proposed Architecture

This section provides an outline for a thorough analysis of the research on big data analytical techniques for weather forecasting. As seen in picture 1, there are four layers that help us describe design architecture:

- ***Data Acquisition Layer:*** The input layer works on data gathering procedures; it gathers information from many sources that are managed by public and governmental entities.
- ***Cloud Layer:*** Large datasets are kept in the cloud by the cloud layer. All large datasets are stored in this layer and are processed in distributed parallel fashion.
- ***Data Processing Layer:*** Large quantities of financial and economic records are provided and used by the data layer, which uses distributed parallel computing.
- ***User Layer:*** The user layer serves as the user's interface for accessing large datasets and controls query management for analysis and report generation.

Our design architecture aims to produce accurate and efficient weather forecasts. To do this, we will carefully analyze the incoming data, identify the chosen features, and process the data. The processing performance of our suggested design depends on the following metrics.

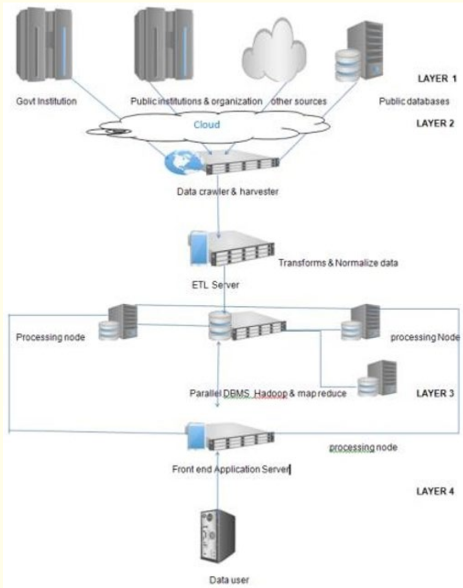

Figure 1: Proposed architecture [14].

Methodology

Dataset

This section includes a description of the dataset, its acquisition and pre- processing, as well as an analysis of the algorithm used, as well as the proposed model for weather forecasting using machine learning.

The dataset used for this analysis was acquired from an online data repository company, the data was for the town of Algeria for the period 11st Feb 2021 to 31st December 2021, the dataset is licensed under CC BY-NC- SA 4.0, the file format was a comma separated file (CSV) which contains 1461 rows and 6 columns, the columns which were identified as: Date, Precipitations, Temp_max, Temp_min, Wind and Weather. (See Table 1) The weather dataset was first stored as a CSV file, and Python 3.6 was utilized in Jupyter Notebook with the Anaconda installation for this work. Following preprocessing, the dataset was divided into training (75%) and testing (25%) datasets at random in order to train the algorithms and evaluate their performances.

N	*Date*	*Precipitations*	*Temp_max*	*Temp_min*	*Wind*	*Weather*
0	01-02-2021	0.0	12.5	4.0	2.4	CloudCover
1	02-02-2021	15.1	7.1	1.1	3.3	Precipitation
2	03-02-2021	14.8	8.2	1.3	1.8	Pressure
3	04-02-2021	0.0	5.0	-1.4	2.8	Solar Radiation
4	05-02-2021	0.2	12.1	3.8	4.4	Wind
5	06-02-2021	0.5	10.4	5.1	3.0	CloudCover
6	07-02-2021	1.1	9.8	4.2	2.4	Precipitation

Table 1: Data head overview.

The nine weather categories in the dataset are Temperature Humidity, Pressure, Wind, Precipitation, Snowfall, Cloud Cover, Solar Radiation, Probability of Precipitation and Probability of Snowfall.

To provide precise forecasts and ensure great algorithm performance, the data needs to be pre-processed. The data was pre-processed for this study by deleting the date column because it was judged unnecessary and the dataset was complete with no null values. Pre-processing involves transforming obtained data into a comprehensible format, eliminating redundant or null values, and eliminating undesired features [15].

Machine learning algorithms

Choosing the right learning The technique is difficult to use because the problem and the facts at hand are so crucial. Given the chaotic nature of atmospheric dynamics and the nonlinear nature of the fundamental phenomena that govern weather, we believe that the methodologies of these algorithms are ideally suited to capture related characteristics [16].

Our dataset has input and output labels, a supervised learning algorithm will be best for the problem. Supervised learning can be segregated further based on the type of output and if the data is sequential. Multiple parameters are needed to train the model [17], so we chose the Decision Tree Algorithm that works well in this direction. Random Forest was also used because it deals with classification and regression problems in ML, and it was implemented using classifier algorithms, including the Gaussian Naive Bayes model, the Gradient Boosting Classifier, and Artificial Neural Networks.

Random Forest classifier

Like its name suggests, a random forest classifier is an ensemble of several different decision trees that work together. The class with the most votes becomes the forecast of our model from each individual tree in the random forest. A randomly chosen portion of the training data is used by the Random forest classifier to generate a collection of decision trees. It simply consists of a collection of decision trees drawn from a randomly chosen subset of the training set, which are then used to decide the final prediction [18].

Decision Tree

Since decision trees are the foundation of the random forest model, let's quickly review them. They are thankfully fairly simple to understand. Most people have probably used a decision tree at some point in their life, whether consciously or unconsciously. The supervised machine learning family includes the Decision Trees technique, where data is continuously segmented by a certain parameter and represented by a tree structure. It is one of the machine learning algorithms that is most frequently employed, and it is used to handle classification and regression jobs [19].

Gaussian Naive Bayes

A probabilistic machine learning approach called Naive Bayes can be applied to a variety of categorization applications. Naive Bayes is frequently used for document classification, spam filtering, prediction, and other tasks. Thomas Bayes' discoveries served as the foundation for this method, hence its name [17]. The Naive Bayes classifier has these advantages: It can handle both continuous and discrete data, is straightforward and simple to implement, doesn't need a lot of training data, is quick, and can be used in real time [20].

Gradient Boosting Classifier

Gradient boosting refers to the process of improving the strength of a weak hypothesis or learning algorithm by a series of small adjustments. The concept of Probability Approximately Correct Learning serves as the foundation for this sort of hypothesis boosting. Gradient boosting classifiers update the classifiers and weighted inputs using the AdaBoosting algorithm in conjunction with weighted minimization. Reduce the loss, or the discrepancy between the actual class value of the training example and the predicted class value, is the goal of Gradient Boosting classifiers. Although it is not necessary to comprehend how the classifier's loss is reduced, it functions similarly to gradient descent in a neural network [21].

Gradient boosting classifiers minimize a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity in a regularized (L1 and L2) objective function (in other words, the regression tree functions). Adding new trees that forecast the residuals or errors of earlier trees, which are then integrated with earlier trees to produce the final prediction, is how the training process is carried out iteratively. Because it employs a gradient descent algorithm to reduce loss when introducing new models, the technique is known as gradient boosting [22].

Artificial Neural Networks

Artificial neurons, which are a set of interconnected units or nodes that loosely resemble the neurons in a biological brain, are the foundation of an ANN. Like the synapses in a human brain, each link has the ability to send a signal to neighboring neurons. An artificial neuron can signal neurons that are connected to it after processing signals that are sent to it. The output of each neuron is calculated by some non-linear function of the sum of its inputs, and the "signal" at a connection is a real number. Edges refer to the connections. The weight of neurons and edges often changes as learning progresses [23].

The ideal way to visualize an artificial neural network is as a weighted directed graph, where the nodes are the artificial neurons. The directed edges with weights represent the relationship between the neuron's inputs and outputs. The input signal for the artificial neural network comes from an outside source as a pattern and the input signal for the image as a vector [24].

Results and Discussion

Results

In our research, we used machine learning metrics for accuracy, precision, and recall to assess model performance. Five algorithms were used in this experiment with the aim of improving location prediction:

$$\text{Accuracy} = (TN + TP)/(TP + TN + FP + FN)$$

Where TP= True positive, TN= True Negative, FP= False positive and FN = false negative.

$$\text{Precision} = TP / (TP + FP)$$

Where TP= True positive, FP= False Positive.

$$\text{Recall} = TP / (TP + FN)$$

Where TP= True positive, FN= False Negative.

$$\text{F1-score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

According to the testing dataset, Table 1 shows the detection rate for each model:

Model Name	***Accuracy***	***Precision***	***Recall***	***F1-score***
Random Forest C	93.73%	90.00%	89.77%	2.85
Decision Tree	79.35%	72.42%	70.01%	1.98
Gaussian NaïveBayes	89.51%	87.39%	85.85%	2.59
Gradient BoostingC	98.12%	97.15%	97.00%	3.4
ANN	85.94%	84.66%	83.96%	2.41

Table 2: Displays Performance comparison of Model prediction.

Figure 2 depicts the count of activities for the training data set for additional clarity and analysis of the findings.

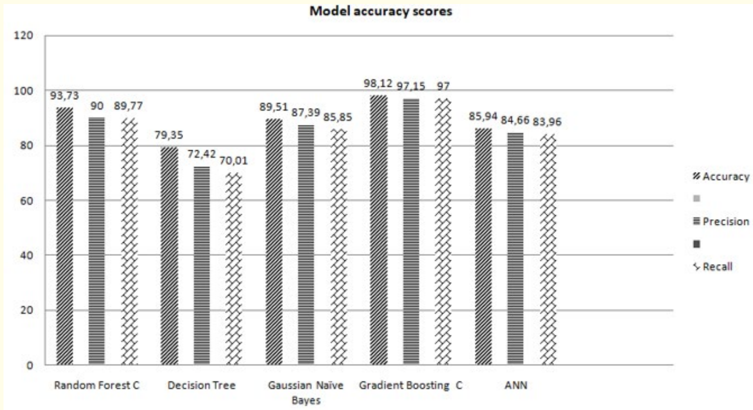

Figure 2: Model accuracy scores of Weather Forecasting using Algorithms.

Discussion

- The Gradient Boosting classifiers algorithm is the best for others, as seen in table 2 and figure 2 above. Results from the Gradient Boosting classification technique are favorable. It benefits from being the easiest of the five machine learning algorithms evaluated to understand and apply. We achieved a prediction accuracy of greater than 98% for weather forecasting. This algorithm works well. It does effectively with small, big, intricate, and data with subgroups. However, it struggles with sparse data and can also have difficulties with data that is widely dispersed. On those kinds of data problems, it typically performs better than the majority of supervised learning algorithms [25].
- The results were positive for the Gradient Boosting classifiers algorithm because the gradient boosting model adheres to ensemble learning, handling and interpreting the data is simpler. It is more accurate than many other algorithms, and accurate results can be obtained using cutting-edge techniques like bagging, random forest, and decision trees. Additionally, one of the best algorithms for processing bigger datasets and computing with weak learners at least loss is this one. This technique is effective at handling category data as well as numerical datasets. A robust technique for machine learning that can quickly detect over fitting training datasets is the gradient boosting algorithm [26].
- The results are less accurate for Artificial Neural Networks algorithm because the input data set is unbalanced, this may cause the neural network to learn patterns that are not representative of the real world. This may eventually lead to a less accurate model [27].
- Decision trees take less work to prepare the data during pre-processing than other methods do. Additionally, data standardization is not necessary for the decision tree. However, because a slight change in the data might result in a large change in the decision tree's structure, creating instability, the findings are weaker and less accurate. When compared to other algorithms, a decision tree's computation can occasionally be more challenging. They frequently include a tree. The decision causes the model training to take longer. Given the complexity and time required, training a decision tree is relatively expensive [28].
- The results were acceptable that the random forests can solve both types of problems namely classification and regression and do a decent estimate on both fronts. One of the advantages of Random Forest that stands most to me is its power to handle large, higher- dimensional datasets. It can handle thousands of input variables and select the most important ones so it is one of the dimension reduction methods. Furthermore, the model produces the significance of the variable, which can be a very useful feature. It has an efficient method for estimating missing data and maintains accuracy when a large percentage of data is missing [29].

- A small amount of training data is needed for Naive Bayes to estimate the test data. So, there is a shorter training period. Naive Bayes is also simple to implement. All of the attributes are implicitly assumed to be independent of one another by Naive Bayes. It is almost never possible to obtain a set of predictors that are entirely independent in real life [30].

Conclusion

In this study, we used Big Data Analytics in Weather Forecasting and machine learning algorithms to assess and compare the performance of five different predictors. We come to the conclusion that the five algorithms performed weather forecasting fairly accurately. This study predicts a variety of weather conditions, including mist, fog, rain, snow, and sun. Additionally, a model was given. The purpose of this project was to develop a weather forecasting model, evaluate its performance, and look into relevant literature. The Gradient Boosting Classifier model was found to be the most accurate of the used algorithms; the output of the model provided accurate prediction and useful guidance for meteorologists in their operational forecasting responsibilities.

Before the model can be used to predict real-time weather conditions and provide meteorologists with useful direction for their operational weather forecasting, a lot of work must be done. The model can be developed further as needed using a variety of techniques.

References

1. S Kumari and P Muthulakshmi. "A Wide Scale Survey on Weather Prediction Using Machine Learning Techniques". J. Inf. Knowl. Manag (2022): 2250093.
2. F Mehrpour. "Prediction of Bridge Fires Characteristics Using Machine Learning". Text, Carleton University (2022).
3. X Ren., et al. "Deep Learning-Based Weather Prediction: A Survey". Big Data Res 23 (2021): 100178.
4. K Maaloul and L Brahim. "Comparative Analysis of Machine Learning for Predicting Air Quality in Smart Cities". WSEAS Trans. Comput 21 (2022): 248-256.
5. KY Ngiam and IW Khor. "Big data and machine learning algorithms for health-care delivery". Lancet Oncol 20.5 (2019): e262-e273.
6. I Munoko, HL Brown-Liburd and M Vasarhelyi. "The Ethical Implications of Using Artificial Intelligence in Auditing". J. Bus. Ethics 167.2 (2020): 209-234.
7. TK Shivaprasad and J Shetty. "Sentiment analysis of product reviews: A review". 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (2017): 298-301.
8. DV Sahasrabuddhe and P Jamsandekar. "Data Structure for Representation of Big Data of Weather Forecasting: A Review". International Journal of Computer Science Trends and Technology 3.6 (2015): 10.
9. H Jain and R Jain. "Big data in weather forecasting: Applications and challenges". 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC) (2017): 138-142.
10. NH Rao. "Big Data and Climate Smart Agriculture - Review of Current Status and Implications for Agricultural Research and Innovation in India". Rochester, NY 25 (2017).
11. S Murugan Bhagavathi., et al. "Retracted: Weather forecasting and prediction using hybrid C5.0 machine learning algorithm". Int. J. Commun. Syst 34.10 (2021): e4805.
12. S Mittal and OP Sangwan. "Big Data Analytics Using Data Mining Techniques: A Survey". in Advanced Informatics for Computing Research, Singapore (2019): 264-273.
13. A Moosavi, V Rao and A Sandu. "Machine learning based algorithms for uncertainty quantification in numerical weather prediction models". J. Comput. Sci 50 (2021): 101295.
14. M Alam and M Amjad. "Weather forecasting using parallel and distributed analytics approaches on big data clouds". J. Stat. Manag. Syst 22.4 (2019): 791-799.
15. A Shehadeh., et al. "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression". Autom. Constr 129 (2021): 103827.
16. D Mukhin., et al. "Revealing recurrent regimes of mid-latitude atmospheric variability using novel machine learning method". Chaos Interdiscip. J. Nonlinear Sci 32.11 (2022): 113105.
17. IH Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions". SN Comput. Sci 2.3 (2021): 160.

18. F Kamalov, S Moussa and J Avante Reyes. "KDE-Based Ensemble Learning for Imbalanced Data". Electronics 11.17 (2022).
19. N Boonnam., et al. "Coral Reef Bleaching under Climate Change: Prediction Modeling and Machine Learning". Sustainability 14.10 (2022).
20. EK Ampomah., et al. "Stock Market Prediction with Gaussian Naïve Bayes Machine Learning Algorithm". Informatica 45.2 (2021).
21. Z Qinghe., et al. "Optimised extreme gradient boosting model for short term electric load demand forecasting of regional grid system". Sci. Rep 12.1 (2022): 19282.
22. R Sibindi, RW Mwangi and AG Waititu. "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices". Eng. Rep (2022).
23. C Jayakumar., et al. "Using the Artificial Neural Networks to Predict the Solubility Effects of Theophylline Drug in Hydrotropic Solutions". International Journal of Pharmaceutical Research (2021).
24. F Wang., et al. "Dynamic spatio-temporal correlation and hierarchical directed graph structure based ultra-short-term wind farm cluster power forecasting method". Appl. Energy 323 (2022): 119579.
25. K Maaloul, NM Abdelhamid and B Lejdel. "Machine Learning Based Indoor Localization Using Wi-Fi and Smartphone in a Shopping Malls". Artificial Intelligence and Its Applications, Cham (2022): 1-10.
26. M Madhavi and D Nethravathi. "Gradient Boosted Decision Tree (Gbdt) And Grey Wolf Optimization (Gwo)Based Intrusion Detection Model". 16 (2022): 15.
27. G Sambasivam and GD Opiyo. "A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks". Egypt. Inform. J 22.1 (2021): 27-34.
28. P Lipinski, E Brzychczy and R Zimroz. "Decision Tree-Based Classification for Planetary Gearboxes' Condition Monitoring with the Use of Vibration Data in Multidimensional Symptom Space". Sensors 20.21 (2020).
29. A Kanervo. "Random Forests an Application to Tumour Classification".
30. Maswadi., et al. "Human activity classification using Decision Tree and Naïve Bayes classifiers". Multimed. Tools Appl 80.14 (2021): 21709-21726.