JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2822 - 2828
https://publishoa.com
ISSN: 1309-3452

Landslide Detection Using Machine Learning Algorithms

Devi Naveen [1], Dr.Nirmala M [2], Dummala Roopesh [3], J. Kiran Reddy [4], P. Karthik Raju [5]

[1] Sr.Assistant Professor, CSE Department, New Horizon College of Engineering, Bangalore, Karnataka, India.
[2] Associate Professor, CSE Department, New Horizon College of Engineering, Bangalore, Karnataka, India.
[3, 4, 5] Student, CSE Department, New Horizon College of Engineering, Bangalore, Karnataka, India.

Email : *[1] devinaveen74@gmail.com, [2] drmnirmala15@gmail.com, [3] 1nh18cs062.dummalaroopesh@gmail.com, [2] 1nh18cs078.kiran@gmail.com, [2] 1nh18cs131.karthikkumar@gmail.com*

ABSTRACT

Landslides are among the most destructive natural disasters that may occur in hilly terrain such as the Himalaya. The study of landslides has gotten a lot of interest lately, mostly because people are becoming more conscious of the socio-economic consequences of landslides. Remote sensing pictures give a wealth of important land use information that may be combined in a GIS setting with other spatial characteristics that influence the incidence of landslides to get a more complete picture of the landscape. The creation of a landslide inventory is an essential step in conducting a landslide hazard analysis using geographic information systems (GIS)[1].
The use of geographic information systems (GIS) enabled the rapid analysis of a large amount of data, and the artificial neural network proved to be an excellent tool for landslide hazard estimates. In order to perform a risk analysis, the DEM, the distance from the danger zone, the land cover map, and the damageable items that were at risk were all considered. Demarcating catchments and masking risky zones in the landslide area were accomplished via the use of digital elevation models (DEMs). The hazard map was generated via the use of geographic information system (GIS) map overlaying technology. This information might be used to calculate the danger to people, property, and existing infrastructure, such as transportation.
As part of the effort to develop real-time weather forecasting and image processing methodologies, this study may benefit from the addition of concepts and technologies such as embedded systems, the Internet of Things, and digital image processing to its repertoire.

Keywords— ANN, SMOTE, Landsliding.

I. INTRODUCTION

In the case of weathered rock that has been broken and decomposed as a consequence of the weathering process, weathered debris that has been wet with rainwater may fall as a result of gravity. For a rapid downhill slide movement of rock debris, the word "landsliding" is used to characterise the phenomenon. Any terrain may be capable of supporting them provided the proper conditions exist in terms of soil, moisture content, and slope aspect, among other things. In the planet's surface geology, landslides are an essential element of the natural process that redistributes soil and sediments in a process that may take the shape of quick collapses or slow mudflows, debris flows, earth failures, slope failures, and other forms of failure. Landslides are extremely common in geodynamic sensitive belts, which are zones and territories that have been repeatedly shaken by earthquakes and impacted by other neotectonic activities which are zones and territories that have been repeatedly shaken by earthquakes and impacted by other neotectonic activities[2].
Gravitational force is the primary driving factor behind every landslide, and the movement of this mass will be proportional to the slope angle of the hill below the landslide. They are related to the mass's friction angle and to the same angle of the hill slope as the mass's resistive forces. They are also proportional to the friction angle of the material that is employed to make that mass slide down that slope. If there is a prolonged period of rain or seismic shocks, it is also likely that the resistive forces will be greatly decreased. Varying types of landslides occur at different rates, and there is a significant difference in the rates of occurrence among them (Fig. 1). In this image, it can be shown that the failure rate of rockfall is much more than the failure rate of slumps or soil creeping, as can be noticed [1].

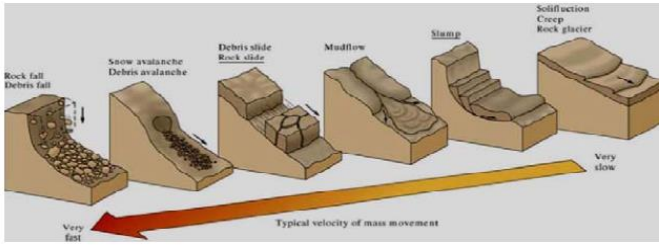


Figure 1 Types of Landslides.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2822 - 2828
https://publishoa.com
ISSN: 1309-3452

The identification of landslide-prone areas or the zonation of landslide hazards is possible with the help of remote sensing data, which can be gathered from a variety of sources, including drainage maps, contour maps, digital elevation models, slope angle maps, land use/cover maps, relative relief maps, thrust (buffer) maps, photo-lineament (buffer) maps, and geological maps, among others. The digital elevation model (DEM) depicts regional variations in height and is used to create a slope map and a relative elevation map, both of which are seen in the image above. When evaluating stability, the slope is an important factor to take into consideration. Each pixel in the picture indicates the angle of slope at a given location, and the image is the first derivative of elevation; the image is the first derivative of elevation. A natural result of increasing slope angle is that shear stress in soil or other unconsolidated materials increases, which is a natural consequence of this phenomenon. A slope's stability improves as the quantity of vegetation on it grows, and ratings were provided in line with this pattern. According to the definition, relative relief is defined as the difference in elevation between the highest and lowest points in a certain area or aspect, and it is established geographically. Higher relief levels have a bigger influence on the chance of landslides happening than lower relief values[1].

A. **Aim:**

It is the primary purpose of this project to develop a landslide detection system, which will be accomplished via the usage of Artificial Neural Networks (ANNs) and the Synthetic Minority Oversampling Technique (SMOT) (SMOTE Algorithm). The Python programming language will be used to create the actual system that will be used to implement it. Particular attention is being paid to identifying areas at risk of landslides, with the ultimate objective being the establishment of a safe and effective early warning system in the event of a potentially catastrophic catastrophe.

B. **Objective:**

This research article's primary goal is to develop the capability of identifying the presence of landslides, as well as the development of a model that can do so successfully, allowing us to issue warnings and prevent landslides from occurring in the first place. This will allow us to issue warnings and prevent landslides from occurring.

II. LITERATURE SURVEY

According to the findings of the study paper [1,] landslides, ground settlement, and avalanches cause major disruptions to long-term gatherings and activities. The collapse of a hillside or valley side as a result of a combination of geological, climatic, and biotic variables is a key component of the process. These disasters have destroyed towns and cities, communication networks, and critical structures such as dams and bridges, among other things. The slope is crucial in the formation of gravitational force in wasting processes such as landslides, soil creeping, and slumping. As a result of a flurry of construction initiatives, landslides have become far more prevalent in the Himalayan area. The only method to entirely avoid or mitigate landslide-related tragedies is to have a comprehensive understanding of the projected frequency, kind, and scale of mass movements in a given location, which can only be achieved by extensive research. Understanding the multiple elements that contribute to landslide formation and spread is crucial for forecasting the danger of future landslides in a particular place. Understanding the many factors that contribute to the genesis and spread of landslides Geospatial technologies are now being used to solve a wide variety of environmental challenges because of their broad range of applications. This strategy, which is also beneficial in natural hazard assessments, may be used to a wide range of applications, including environmental planning. Using remote sensing data, it is possible to identify landslide-prone areas and categorise landslide hazards. Remote sensing data sources include drainage maps, contour maps (including photolineament), digital elevation models (including slope angle maps), land use/land cover mapping data, relative relief mapping data, thrust (buffer) maps, photolineament (buffer) maps, and geological maps. The Geographic Information System (GIS), a computer-based system capable of data capture and input, manipulation and transformation, visualisation and combination with other information, querying and analysing data, modelling and producing output, has received a lot of attention in the field of natural disaster assessment due to its superior spatial data processing capacity. GIS stands for Geographic Information System (GIS)

The frequency ratio model is crucial, according to the authors of the research [3.] Inputting data, creating findings, and presenting those findings are all straightforward. A geographic information system (GIS) allows massive amounts of data to be processed quickly and easily. When using the logistic regression model, data must be translated to ASCII or other formats for statistical analysis before being returned to the GIS database format. In addition, the statistical software creates a large amount of data that is difficult to manage. In a similar statistical model (discriminant analysis), the components must have a normal distribution, but in multi regression analysis, they must have a numerical distribution. Logistic regression is a useful technique for landslide analysis since the dependent variable must be presented as a number between zero and one in order for the model to be valid.

Landslide hazard mapping is increasingly being used in urban planning applications. When it comes to slope management and land use planning, the results of this study [3] may be valuable to developers, planners, and engineers,

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2822 - 2828
https://publishoa.com
ISSN: 1309-3452

among other professionals. When using the models to develop unique websites, it is recommended that you approach them with care. Other slope components must be examined due to the extensive scope of the investigation. As a consequence, the models developed in this study are suitable for long-term planning and assessment.
In their investigation, they state that the factors involved in landsliding include aspects, slopes, and curvature as derived from the topography database; soil texture, material, drainage, thickness, and topography as derived from the soil database; timber type and diameter as derived from the forest database; lithology as derived from the geology database; lineament as derived from the IRS image; and land cover classification as derived from the LCI database. Furthermore, according to the study, the probability–likelihood ratio approach was adopted in order to analyse the link between landslides and other components in order to determine the chance that a landslide would occur. A landslide risk map was constructed based on the ratios acquired from this link, with the resultant landslide hazard index serving as a guide for the construction process. Specifically, in order to check and validate the risk map, it was computed the association between the site of landslide occurrence and the data obtained from the study and then compared to the risk map, which was constructed based on the research's data. In accordance with the findings of the verification, the recommendations have been implemented.

III. METHODOLOGY

This section will describe the application of the algorithms selected for the research, as well as the rationale for selecting the specific method for the investigation.

SMOTE(Synthetic Minority Over-sampling Technique):

Because geographic information systems (GIS) combine many different kinds of data layers depending on geographic location, it is argued that the data from geographic information systems (GIS) is unbalanced, therefore justifying the adoption of SMOTE (Synthetic Minority Over-sampling Technique). It was discovered throughout the investigation that the vast bulk of the data included a geographic component. Terrain visualisations, topographic features, and base maps are all included in the data collected by geographic information systems (GIS), which are then linked to spreadsheets and tables for further analysis and display.
Despite the fact that SMOTE is widely regarded as a successful approach for dealing with unbalanced data, it is likely that a significant amount of critical information will be lost as a result of under- or even over-sampling, and that a large number of duplicate data will be generated as a result of under- or even over-sampling.
An easy solution to such problems is data augmentation, and it is a straightforward process that needs no technical knowledge or skill.
The addition of data points that have been regenerated as a result of the SMOTE process improves the content of a dataset.
Most crucially, by making slightly different versions of the same information, SMOTE minimises the need for duplicate data and so saves time and money. Long-term, this leads in more efficient data storage and retrieval because of the increased efficiency.
By using SMOTE, it is feasible to fine-tune the model such that it generates fewer false positives while also enhancing recall, which is the intended outcome.
It is feasible that we may observe a decrease in accuracy, but that we will experience an increase in recall as a consequence of producing more predictions for a minority class, which would result in a greater rise in correct forecasts than we would otherwise see.
SMOTE is meant to minimise the quantity of data created as a remedy for imbalanced data, rather than increase it.

1. ANN (Artificial Neural Networks) :

In order to process information, a vast number of linked processing units that work together to analyse information and make meaningful conclusions from that information are used in this research. In order to handle information efficiently, ANNs (artificial neural networks) are being employed in conjunction with a high number of linked processing units.
It is crucial to highlight that the employment of ANN has the advantage of learning from the data that has been analysed and so does not need reprogramming afterwards, which is advantageous. In contrast, they are referred regarded as "black box" models since they provide very little information about how these models really operate in real-world circumstances. Users just need to submit input, sit back and watch the system train, and then wait for the system to choose an output for them to be able to utilise the system successfully.
A class of mathematical models known as artificial neural networks (ANNs) are widely considered to be vital in mathematics, and they have the potential to enhance the performance of existing data processing technologies. Despite the fact that it does not have the same processing power as the human brain, some experts feel that it is the most significant building block in the creation of artificial intelligence.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2822 - 2828
https://publishoa.com
ISSN: 1309-3452

The approach flowchart is shown in (Fig. 2)

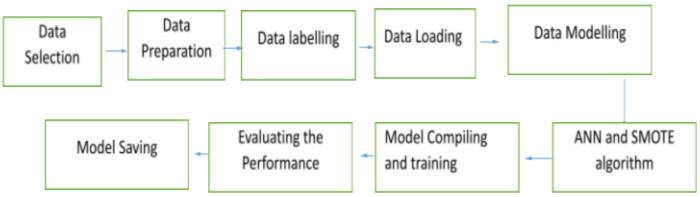


Figure 2 Approach flowchart.

Data Selection:

As defined by the American Society for Information Management, it is the process of selecting the most relevant data type and source, as well as the most appropriate instruments to use, in order to obtain the necessary information in order to make a decision in order to collect the necessary information. Following the completion of the data selection process, the real practise of data collection may begin. This process continues indefinitely.
Specifically, when it comes to our predicament, we have chosen geographic information system (GIS) data because, as opposed to other types of data, it is a computer system that collect and store data that is related to real locations on the Earth's surface, validate and display that data, and does so in a more visually appealing manner than other types of data. In order to get a better knowledge of geographical patterns and linkages, it is feasible for people and organisations to benefit from the usage of geographic information systems (GIS).

Data Preparation:

Generally speaking, data preparation is the pre-processing procedure that happens prior to the usage of data from one or more sources in a machine learning model in the area of artificial intelligence. It is vital to clean and update data in order to enhance the overall quality of the data and make it more useable in the future.

Data labelling:

Gathering raw data and associating meaningful labels to it in order to provide context for the information that has been gathered is known as data labelling.
As a result of this procedure, the computer is able to recognise the most significant aspects of the data and train itself to do so even more efficiently in the future.

Model training and testing:

Following the completion of the data tagging, we divided the information into two unique sets of information. When the fit strategy of SMOTE is used, a subset of the data is used to train the machine learning mode, resulting in a more accurate machine learning model than when utilising the whole set of data.
Following the construction of the machine learning model, the data from the second batch is used to test the model and determine if it is successful or unsuccessful.
As a consequence of the findings, it becomes possible to construct a classification matrix for the machine learning model, which will allow us to evaluate how successful the model is at precession and recall in terms of classification accuracy.

Data modeling (ANN and SMOTE):

Make classification() of the scikit-learn library may be used to construct a synthetic binary classification dataset with 10,000 instances and a 1:100 class distribution by combining the make classification() tool with the make classification() function and a 1:100 class distribution.
With the use of this object, we will be able to summarise the number of samples in each class, which will aid us in establishing whether or not the dataset was constructed appropriately.
As a starting point for our study, we will use the binary classification dataset from the previous section. We will next fit and assess a decision tree technique to classification using the data from this section.
In the next step, the method is generated with any necessary hyperparameters (we will leave the defaults in place), and the model is tested using repeated stratified kfold cross-validation, which is a statistical procedure used to assess models. In order to reach our goal of fitting and evaluating 30 models on the dataset in total, we will conduct three rounds of 10-fold cross-validation on the data. This indicates that a 10-fold cross-validation procedure will be carried out three times in total.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2822 - 2828
https://publishoa.com
ISSN: 1309-3452

Model evaluation:

It is possible to get a genuine positive result if you anticipate that an observation belongs to a class, and the observation does in fact belong to the class in which you anticipated it to belong.
True negatives arise when you predict that an observation does not belong to a class, and the observation does not really belong to that class in the manner that you anticipated it to belong to that class. True negatives are also known as false negatives.
When you predict that an observation belongs to a class, but the observation does not belong to any of the classes that you anticipated it to be a member of, you are said to have generated a "false positive."
In the case of expecting erroneously, when you anticipate an observation to belong to a class but it turns out that the observation does belong to the class in question, this is referred to as inaccurate anticipation. A "false negative" is what is referred to as this.
They are often shown on a confusion matrix in order to draw attention to the link between these four possibilities. The confusion matrix depicted below is intended to serve as an instance of a circumstance in which binary classification is utilised, and it is intended to serve as an example of such a situation. It would be necessary to build this matrix based on your test results, which would then be used to categorise each prediction into one of the four most likely alternatives listed above. In order to create this matrix, you would follow the steps outlined above.

IV. RESULTS AND DISCUSSION

This section explains the results obtained during the whole dissertation. The results are obtained based on the data set provided. The accuracy of the model is evaluated and plotted in the following graphs.

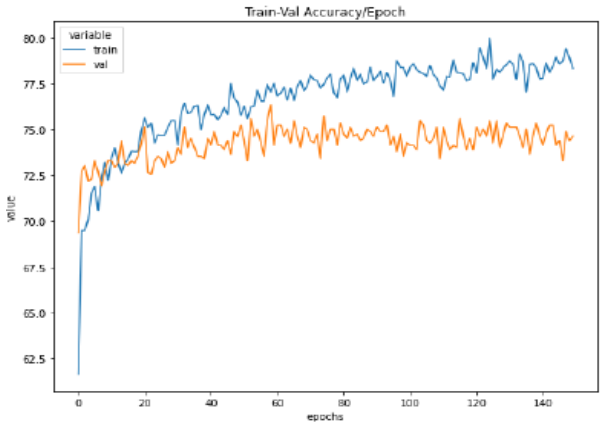


Fig 17: Train- Validation Accuracy

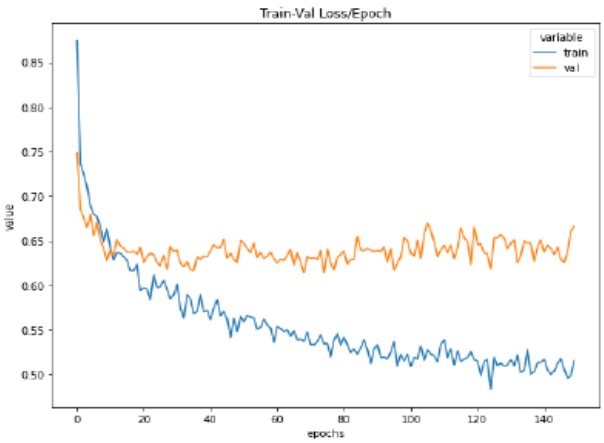


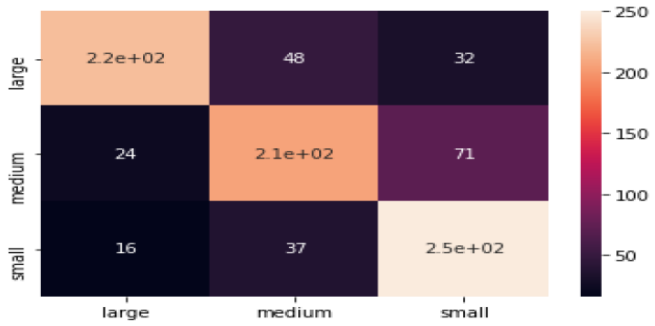
Fig 18: Train- Validation loss

Loss is defined as the difference between the problem's real values and the model's projected values. The greater the loss, the more significant the data mistakes you committed.
The number of errors you made on the data is referred to as accuracy.
That is to say: A low accuracy and large loss indicate that you made large errors on a large amount of data, whereas a low accuracy but small loss indicates that you made little errors on a large amount of data.
A high accuracy combined with minimal loss indicates that you made few errors on a small set of data (best case)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2822 - 2828
https://publishoa.com
ISSN: 1309-3452

Confusion Matrix:

A confusion matrix is a visual depiction of the results of a classification problem prediction.
The total number of correct and incorrect predictions is added together and divided by class using count values. This is the key to the confusion matrix.
The confusion matrix shows how your classification model becomes bewildered while making predictions.
It notifies you not just of your classifier's flaws, but also of the kind of errors that are being made.
The disadvantage of relying only on categorization accuracy is addressed in this study.



	precision	recall	f1-score	support
0	0.85	0.74	0.79	302
1	0.71	0.69	0.70	302
2	0.71	0.83	0.76	303
accuracy			0.75	907
macro avg	0.75	0.75	0.75	907
weighted avg	0.75	0.75	0.75	907

Fig 19: Classification report

Support - Support is the number of real events of the class in the predefined dataset. Imbalanced support in the preparation of information might demonstrate underlying shortcomings in the announced scores of the classifier and could show the requirement for separated inspecting or rebalancing. Support doesn't change between models yet rather analyzes the assessment interaction.
The accuracy of the developed model is 75%.

V. CONCLUSION

By producing slightly distinct copies of the same information, SMOTE reduces the need for duplicate data and the amount of data that must be stored. Using SMOTE as a result, time and resources are saved as a result of the process. As technology progresses, data storage and retrieval will become more efficient, resulting in more efficient data storage and retrieval in the long run. It is the pre-processing method that happens prior to the use of data from one or more sources in a machine learning model that is known as data preparation in the machine learning field. As an example, consider the following definition of data preparation: Data that is up to date and clean must be maintained if we are to improve the overall quality of the data and make it more relevant for future research and development. Geographic information systems (GIS) are computer systems that collect and store data that is linked to real-world geographic locations on the Earth's surface, validate and display that data, and do so in an aesthetically pleasing manner. Geographic information systems (GIS) are used in the context of the Earth's surface.
With regard to data structure, we will specifically examine how a synthetic binary classification dataset with 10,000 occurrences and a 1:100 class distribution is organised in the next section. The create classification() tool will be used to generate a model that will be used to represent each class. It is necessary to test the model using repeated stratified kfold cross-validation, which is an iterative procedure, once it has been generated with any appropriate hyperparameters (we will leave the defaults in place). This is known to as "false positive" generation when you forecast that an observation belongs to a class, but the observation does not belong to any of the classes to which you expected it to belong when you made the prediction. The confusion matrix presented below is meant to be used in combination with the scenario described above in order to serve as an illustration of such a situation. This matrix would be needed once you have received the results of your tests because it would be essential to categorise each prediction into one of the four most probable choices indicated above after you have received the results of your tests. Once you have gotten the results of your tests, it will be essential to develop this matrix using the information.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2822 - 2828
https://publishoa.com
ISSN: 1309-3452

REFERENCES

1. Rai, P.K., Mohan, K. and Kumra, V.K., 2014. Landslide hazard and its mapping using remote sensing and GIS. Journal of Scientific Research, 58, pp.1-13.
2. Bolt, B.A., (1975). Landslide Hazard, Geological Hazard, Springer Verlog, New York, 150.
3. Pradhan, B. and Youssef, A.M., 2010. Manifestation of remote sensing data and GIS on landslide hazard analysis using spatial-based statistical models. Arabian Journal of Geosciences, 3(3), pp.319-326.
4. Lee, S., Choi, J. and Min, K., 2004. Probabilistic landslide hazard mapping using GIS and remote sensing data at Boun, Korea. International Journal of Remote Sensing, 25(11), pp.2037-2052.
5. Dr. Vijayalakshmi A. Lepakshi1, Dr. Prashanth C S R2," Efficient Resource Allocation with Score for Reliable Task Scheduling in Cloud Computing Systems", 978-1-7281-4167-1/20/$31.00 ©2020 IEEE
6. A. P. Nirmala, Prasenjit Kumar Das," IoT based Automatic Light Control with Temperature Monitoring and Alert mechanism",DOI:10.35940/ijeat.E7701.088619
7. S. SHANMUGA PRIYA," Define – use Testing An Example", Retrieval Number: E1947017519 & Sciences 19©BEIESP
8. Gopal M.K., Amirthavalli M. "Applying machine learning techniques to predict the maintainability of open source software", International Journal of Engineering and Advanced Technology, Vol. 8, 2019.
9. Karthikayini T., Srinath N.K. "Comparative Polarity Analysis on Amazon Product Reviews Using Existing Machine Learning Algorithms", 2nd International Conference on Computational Systems and Information Technology for Sustainable, 2018.
10. 10. Nithya B., Ilango V. "Predictive analytics in health care using machine learning tools and techniques", Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems, ICICCS 2017.
11. A. P. Nirmala, Prasenjit Kumar Das, " IoT based Automatic Light Control with Temperature Monitoring and Alert mechanism", DOI: 10.35940/ijeat.E7701.088619.
12. Godwin J.J., Krishna B.V.S., Rajeshwari R., Sushmitha P., Yamini M. "IoT Based Intelligent Ambulance Monitoring and Traffic Control System", Vol. 193, Intelligent Systems Reference Library, 2021.
13. Shanmugaraj G., Santhosh Krishna B.V., SriSahithya S., Sandhya M., Monikca T.H. "Unhindered Safety Monitoring System for Underground Workers", Vol. 665, Lecture Notes in Electrical Engineering, 2020.
14. Suji Prasad S.J., Thangatamilan M., Suresh M., Panchal H., Rajan C.A., Sagana C., Gunapriya B., Sharma A., Panchal T., Sadasivuni K.K. "An efficient LoRa-based smart agriculture management and monitoring system using wireless sensor networks", International Journal of Ambient Energy, 2021.
15. Mannan J M., Kanimozhi Suguna S., Dhivya M., Parameswaran T. "Smart scheduling on cloud for IoT-based sprinkler irrigation", Vol. 17, International Journal of Pervasive Computing and Communications, 2021.
16. Manojkumar P., Suresh M., Ayub Ahmed A.A., Panchal H., Rajan C.A., Dheepanchakkravarthy A., Geetha A., Gunapriya B., Mann S., Sadasivuni K.K. "A novel home automation distributed server management system using Internet of Things", International Journal of Ambient Energy, 2021.
17. Santhosh Krishna B.V., Jijin Godwin J., Tharanee Shree S., Sreenidhi B., Abinaya T. "Detection of Leukemia and Its Types Using Combination of Support Vector Machine and K-Nearest Neighbors Algorithm", Vol. 201, Lecture Notes in Networks and Systems, 2021.