

Predictive Modeling of Seismic Events: A Comparative Analysis Of K-Nearest Neighbors and Random Forest Algorithms

Dayang Sun

Harbin Engineering University, Harbin, China

2021029306@hrbeu.edu.cn

Abstract. This study processes and predicts seismic data using data visualization approaches, K-nearest neighbors (KNN) and the random forest (RF) algorithms. The analysis's dataset includes a number of variables connected to earthquakes. The primary goal is to devise a forecast algorithm capable of accurately categorizing seismic events, data visualization tools are utilized to gain insights into the dataset, producing informative charts that depict the distribution and correlations among different variables. This visual evaluation aids in pinpointing anomalies or trends, facilitating a deeper understanding of the data's characteristics and guiding decisions during the modeling phase. Subsequently, the KNN method classifies earthquake occurrences based on their attributes, predicting the class label by considering the characteristics of its nearest neighbors. Additionally, Accurate classification of seismic events is enhanced by using RF, an ensemble learning technique that combines many decision trees to produce predictions. To optimize outcomes, the study adjusts the random forest model's hyperparameters through cross-validation. The study compares the performance of KNN and RF using a confusion matrix. The confusion matrix shows a thorough insight of categorization performance, which provides a comprehensive view of categorization efficacy. This assessment underscores the models' precision and effectiveness in classifying seismic events.

Keywords: Seismic Data Prediction, Data Visualization, Machine Learning Algorithms (KNN and Random Forest).

1. Introduction

Earthquakes are natural disasters that can inflict significant damage to structures and infrastructure, resulting in casualties and economic losses. Predicting the degree of earthquake damage to homes can help with disaster preparation and management, allowing for more effective resource and aid distribution. Therefore, it is critical to develop accurate predictive models that can assess the extent of damage to houses based on various criteria. Machine learning, a branch of artificial intelligence, has shown significant potential in predicting outcomes and processing vast amounts of data.

Numerous studies have been conducted to far with the goal of estimating the level of damage resulting from earthquakes. These studies have employed a range of techniques and methodologies. Using data from several sources, such as hazard-causing causes, hazard-formative environments, and hazard-affected bodies, Jin Chen, Hong Tang, Jiayi Ge, et al. developed a unique approach for quickly assessing building damage following earthquakes. In order to determine the proportion of damaged structures with varying levels of damage, the approach combines pixel-level classification with microzone-level regression. It may be used to improve the efficiency of emergency rescue operations following earthquakes [1]. Ehsan Harirchian, Vandana Kumari, Kirti Jadhav et al. have examined the efficacy of five various Machine Learning (ML) algorithms in applications for vulnerability prediction. The damage data from four independent earthquakes that happened in Ecuador, Haiti, Nepal, and South Korea were used to train and evaluate the produced models. Eight performance-modifying elements have been added as variables with supervised machine learning. According to the research, the classifications of vulnerability indicated by ML methods and the actual degrees of damage detected in the structures were quite close [2]. Kuldeep Chaurasia's study uses machine learning to forecast the extent of building destruction brought on by the Gorkha earthquake in Nepal.

Eight tectonic markers that were mathematically analyzed together with historical vibration activity data were used to make the projections. The purpose of this project is to forecast earthquake damage using machine learning methods and an existing seismic data collection. In this study, K-nearest neighbors (KNN) and Random Forest (RF), two well-known machine learning algorithms, have been applied, and the best parameters for precise prediction are looked at. The data shows that for predicting building damage, the random forest method has superior to the technique using neural networks. The random forest classification yielded a score of 74.32% for F1 [3].

The results of these studies have enhanced our understanding of the relationship between earthquakes and building damage, providing valuable insights for disaster management agencies and policymakers. Despite the advancements in previous research, certain challenges persist in the field of earthquake damage prediction. Many of these studies have focused on certain types of structures or geographical regions, which limits the generalizability of their findings. Moreover, the choice of input features and the accuracy of the models differ among studies. These limitations underscore the necessity for a comprehensive and robust predictive model that can reliably estimate damage across a diverse range of houses in various geographic areas.

To evaluate the efficacy of two classic algorithms, KNN and RF, I constructed models on a comprehensive dataset to predict post-earthquake damage grades. The confusion matrix was utilized to compare accuracy and delve into misclassification specifics. The rest of this essay is structured as follows: The method is presented in Section 2, which includes data preparation, feature selection, and model construction. Section 3 shows the findings of models and discussions. Finally, Section 4 summarizes the implications and significance of the results obtained in this study.

2. Methodology

2.1. Dataset

The Kathmandu Valley is among the seismically active areas of the globe, and has experienced numerous earthquakes in the past. The soft sediments in the valley amplify the ground motion during earthquakes, making it a high-risk area for earthquake disasters. To enhance our knowledge of the seismic risk and hazard in the Kathmandu Valley and to offer important information for earthquake disaster mitigation. Michiko Shigefuji et al. collected the strong ground motion data [4]. The dataset includes the 7.8-magnitude earthquake that struck Gorkha, Nepal, in 2015. Nearly 9,000 people lost their lives and millions were immediately homeless in this damage, causing \$10 billion in losses – around 50 percent of Nepal's nominal GDP.

The dataset includes data from a network of 20 strong-motion stations, which recorded the mainshock and several large aftershocks. The dataset contains over 1,000 ground motion records, covering a wide range of frequencies and ground motion parameters, including peak ground acceleration, peak ground velocity, and spectral acceleration. The data can be used for numerous applications, including ground motion prediction, earthquake engineering, and seismic hazard assessment. Researchers, engineers, and decision-makers who are focusing on lowering earthquake risk and managing disasters in the area will benefit greatly from the dataset. The information can be used to evaluate the performance of buildings and other structures during earthquakes as well as to create more effective seismic design codes and standards. It can also be used to better understand the seismic risk and hazard in the Kathmandu Valley and to create formulas for predicting ground motion in the region. The dataset is available for download from the figshare data repository.

In this paper, I employed all 762,106 samples with 30 relevant features to building, including building_id, age_building, land_surface_condition, roof_type, etc., to predict the damage stage.

2.2. Prediction Algorithm

In this study, KNN and RF methods are used to calculate the damage data caused by the earthquake to the house.

2.2.1 KNN

KNN, or K-Nearest Neighbors, is a renowned supervised machine learning algorithm with roots dating back to the 1950s. It became more formalized in the 1960s through Thomas Cover's work. The fundamental idea behind KNN is to categorize a data item using the classification of its neighbors. KNN works by finding a predetermined number of training examples that are close together to a new data point and determining the label based on these. The number of samples can be a user-defined constant (K), or depending on the density of points nearby (radius-based neighbor learning). The distance can be any metric measure: the most typical option is the usual Euclidean distance.

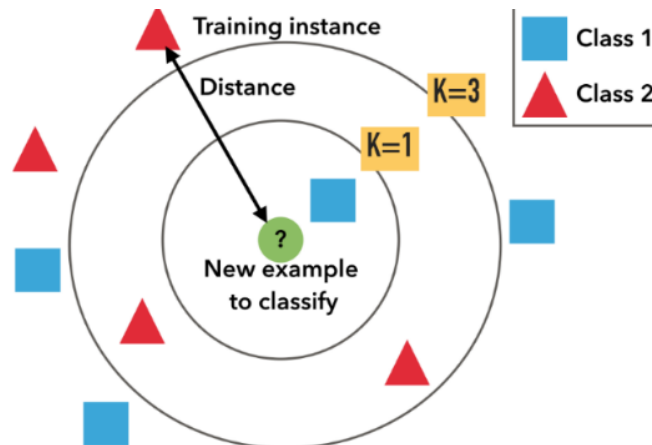


Figure 1. Mechanism of KNN

Fig 1 shows how KNN works in a more specific way: the green circle represents a data point that we want to classify. If we consider $k=3$ (3 nearest data points), it will be classified as "Red Class". If we consider $k=5$, it will be classified as "Blue Class". The choice of ' K ' is crucial. A larger value of k makes it computationally expensive, whereas a smaller value of k suggests that noise will have less of an impact on the outcome.

Over the years, KNN has undergone refinements to boost its efficiency. For instance, Chethana C's 2021 study utilized KNN to predict heart disease, achieving an accuracy of 69% with a prediction speed of 5600 obs/sec using an enhanced KNN model in MATLAB [5]. In 2019, N. H. Othman's research showcased KNN's classification prowess, where most classifier models, post-optimization, achieved a 100% performance rate in accuracy, precision, and ROC [6]. In essence, KNN's adaptability and continuous advancements make it a robust machine learning tool, with future enhancements expected to broaden its applications and elevate its performance.

2.2.2 Random Forest

Random Forest (RF) is an ensemble method that enhances prediction accuracy and minimizes overfitting by amalgamating multiple decision trees. Introduced by Leo Breiman in the early 2000s. Random Forest (RF) is a versatile machine learning algorithm that combines different decision trees to provide a prediction that is more accurate and reliable. Instead of relying on a single decision tree, RF takes the average or majority vote from a forest of trees to make its final decision. RF takes three steps: RF starts by selecting random samples from the dataset with replacement, creating multiple subsets first, known as "Bootstrap Sampling". Then for each subset, a decision tree is constructed. However, at each node, only a random subset of features is considered for splitting, adding another layer of randomness. And the last step is "Aggregation". For classification, the final prediction is taken to be the median (most common class) of the predictions from all trees. For regression, the average prediction is used. As shown in Fig 2, each colored box represents a decision tree, and the dataset is divided into multiple subsets. Then each tree makes its own prediction, and the final prediction is an aggregation of all tree predictions.

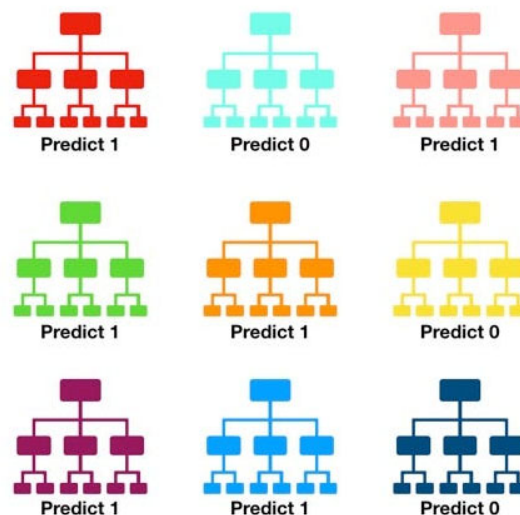


Figure 2. Construction of RF

High accuracy, handling huge datasets with increasing complexity, and handling missing values are all strengths of RF. It's widely used in various domains, from bioinformatics to finance. A.N.V.K Swarupa utilized RF for heart disease prediction, achieving a 95% accuracy using a dataset of 4920 patient records with 41 diseases. This RF-based system, developed in Python with the Tkinter Interface, facilitates early disease detection, enhancing healthcare outcomes [7]. In image recognition, RF excels. Suthapalli Uday Raj employed RF in designing a parking spot detection system. By capturing parking images and extracting edges using advanced algorithms, the system classifies spots as vacant or occupied. Tested on the PKlot dataset with 37,680 parking spots, this method achieved a remarkable 98.31% accuracy [8].

2.3. Metrics

In this research, a multi-classes metrics called confusion matrix is employed. A confusion matrix, often called an error matrix, is a particular table arrangement that illustrates how well an algorithm performs, usually a supervised learning method. It is especially useful for understanding the classification performance of models in contexts where the classes are imbalanced. The matrix itself is a two-dimensional table with actual classes represented by rows and predicted classes by columns [9].

The primary components of a confusion matrix for binary classification are:

Table 1 The primary components of a confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

True Positives (TP): Occurrences that were both positive and as expected. True Negatives (TN): Occurrences that were expected to be negative and actually turned out to be negative. False Positives (FP) are instances that were anticipated as positive even though they were actually negative. False Negatives (FN) are instances that were projected as negative but instead turned out to be positive.

The meaning of the confusion matrix representation is essentially the same for multi-classification issues. Like the binary classification confusion matrix, the addition of matrix row data is the number of real value categories, and the addition of column data is the number of categories after classification.

Table 2 The meaning of the confusion matrix representation

Confusion matrix	Prediction_category 1	Prediction_category 2	Prediction_category 3
True_category 1	a	b	c
True_category 2	d	e	f
True_category 3	g	h	i

3. Result and Discussion

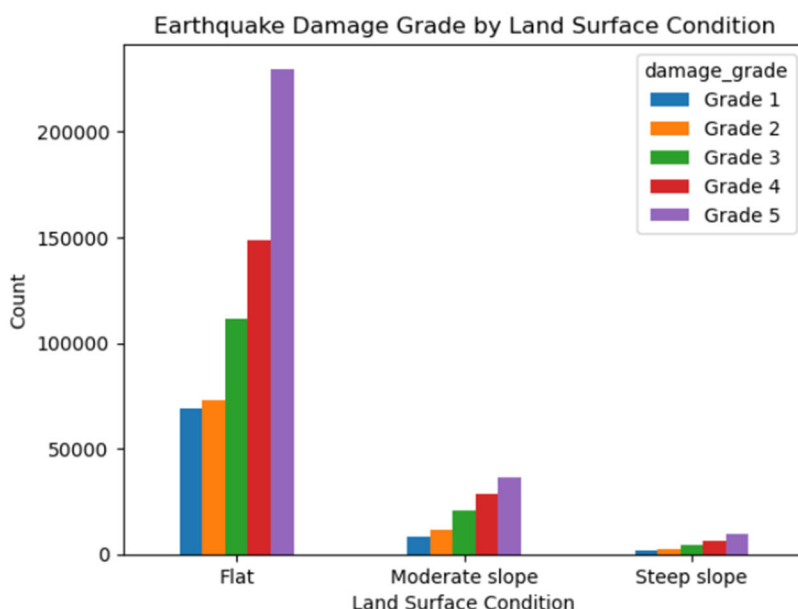


Figure 3. Data distribution, by Land Surface

This chart is about the relationship between the remaining amount of houses after the earthquake and different land surface conditions. From the chart, it can be seen that under the three land surface conditions of flat, moderate slope and slope, the trends of the damage degree of the five levels of earthquakes to houses are almost the same, and they all increase with the increase of the earthquake level. Houses are significantly more resistant to earthquakes on the flat surface, and there is a significant increase in resistance to magnitude five earthquakes. On the moderate slope and steep slope, the remaining amount of the house after the earthquake is much reduced. This shows that flat surface is the best choice for earthquake resistance, and architects should try to avoid building houses on slopes in the future.

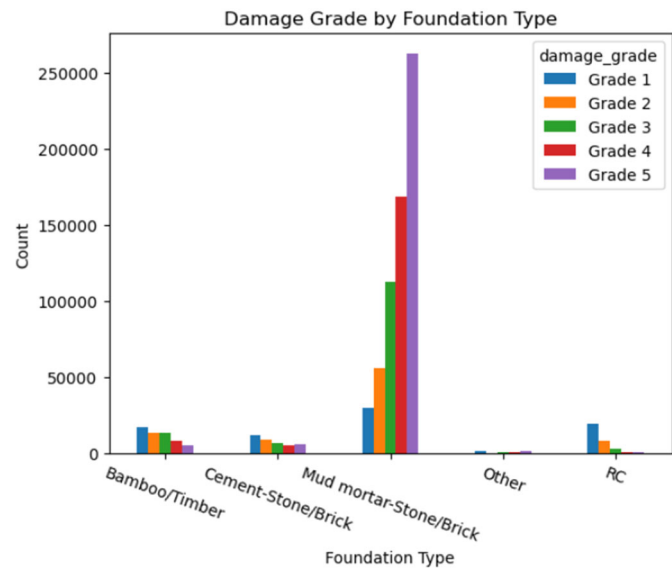


Figure 4. Data distribution, by Foundation Type

This bar chart is about trends in foundation types and post-earthquake housing conditions. Statistical foundation types include bamboo/timber, cement-stone/brick, mud mortar-stone/brick, RC and others. Under the influence of five earthquake levels, only mud mortar-stone/brick makes the house have the ability to resist earthquakes, and more are preserved after the disaster, and as the earthquake level increases, the better the earthquake resistance performance is. However, when the foundation houses of several other materials are faced with different levels of earthquakes, what they show is that as the level of earthquakes increases, the anti-seismic ability weakens. Although the seismic performance of each material is almost the same when the earthquake level is 1, there is a significant gap in their seismic capacity after the earthquake level increases. This shows that engineers and researchers should pay attention to the use of building materials, which should be able to withstand various levels of earthquakes. And if you want to consider the cost issue, you can consider mixing various materials to meet the two requirements of safety and cost.

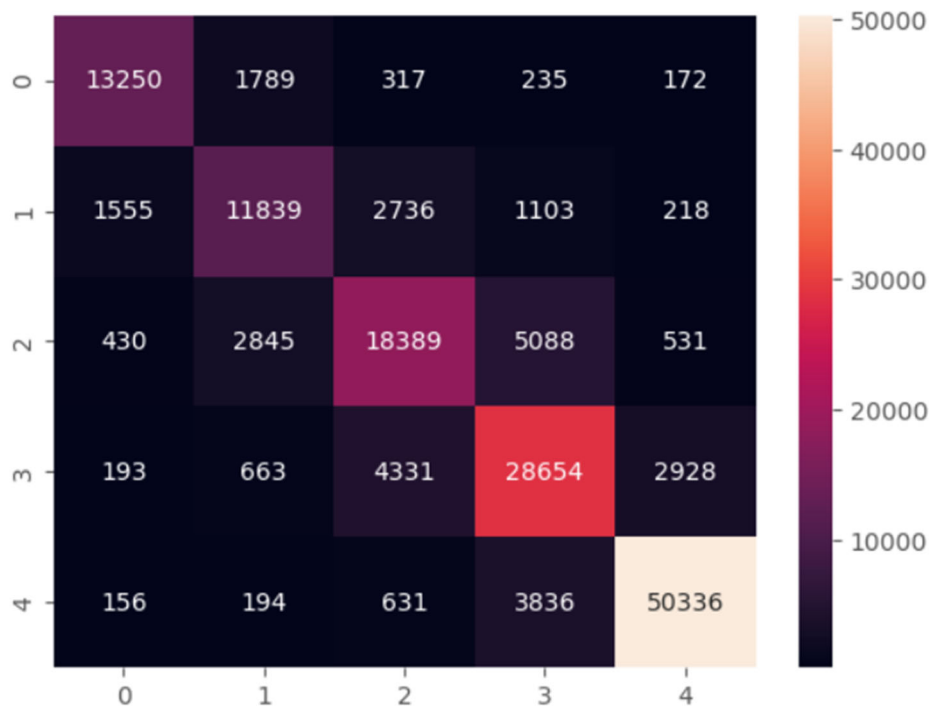


Figure 5. Confusion Matrix of KNN Prediction

This is the confusion matrix generated after using knn to predict the data set. According to the previous formula, we can calculate the relevant value of the confusion matrix. The Accuracy of garde 1 to 5 are 0.850, 0.683, 0.696, 0.736, 0.929, and the Recall are 0.841, 0.678, 0.674, 0.779, 0.913. It's easy to see that the model predicts poorly at classes two, three, and four.

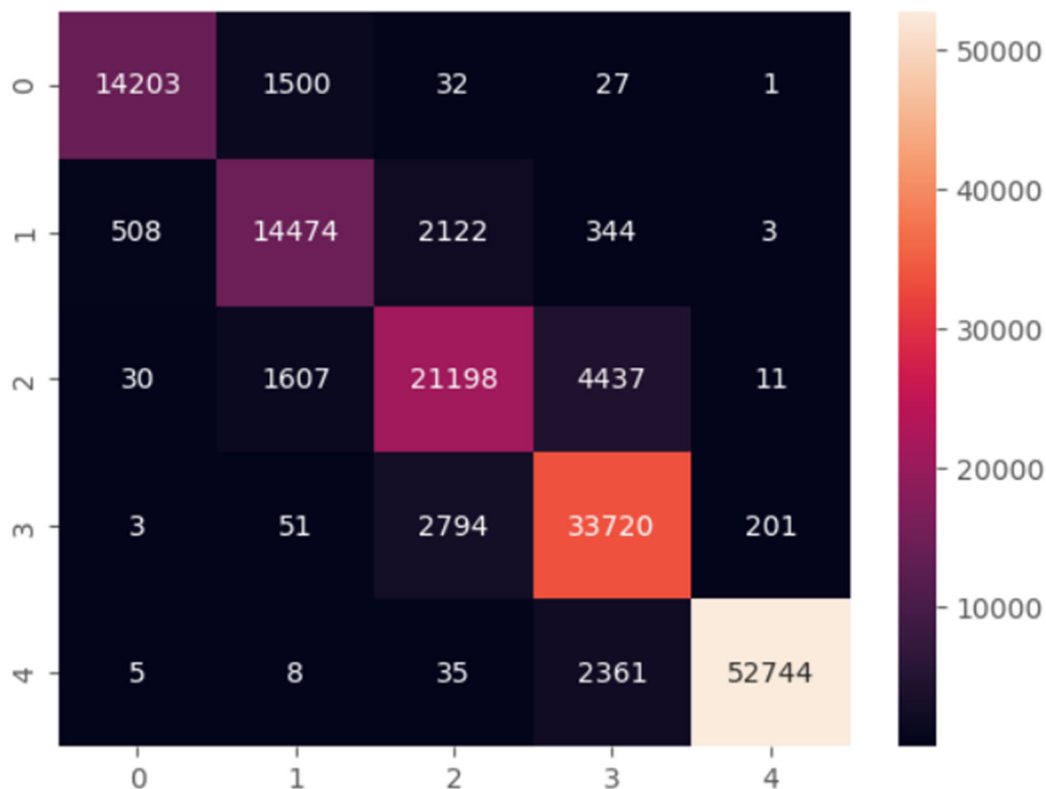


Figure 6. Confusion Matrix of RF Prediction

This is the confusion matrix generated after predicting the data set using random forest. Like the KNN confusion matrix in the previous picture, we can do the calculation. The Accuracy of garde 1 to 5 are 0.963, 0.821, 0.810, 0.825, 0.996, and the Recall are 0.901, 0.829, 0.777, 0.917, 0.956. The model performs well in almost every class, only the recall of class three is not very high. Compared with the KNN model, the prediction results of the RF model are significantly better.

4. Conclusion

In this study, data visualization methods, KNN and RF models were used to analyze the data sets of earthquakes that occurred in Nepal. Through this study, we found that: by analyzing the damage to houses caused by earthquakes, engineers should realize that houses should be built on flat ground as much as possible, and researchers should also pay attention to the use of engineering building materials. As shown in Figure 2, the anti-seismic performance of mud mortar-stone/brick is the best. When predicting the damage of earthquakes to houses, the prediction results of the two models performed well, and when the recall rate was analyzed by using the confusion matrix and compared, it is not difficult to see that the model of RF is better than that of KNN. The conclusion that can be drawn is that for this data set, the fitting effect of using random forest is stronger than that of using KNN.

References

- [1] Chen J, Tang H, Ge J, Pan Y. Rapid Assessment of Building Damage Using Multi-Source Data: A Case Study of April 2015 Nepal Earthquake. *Remote Sensing*. 2022; 14(6):1358.
- [2] Harirchian E, Kumari V, Jadhav K, Rasulzade S, Lahmer T, Raj Das R. A Synthesized Study Based on Machine Learning Approaches for Rapid Classifying Earthquake Damage Grades to RC Buildings. *Applied Sciences*. 2021, 11(16):7540.
- [3] K. Chaurasia, S. Kanse, A. Yewale, V. K. Singh, B. Sharma and B. R. Dattu, Predicting Damage to Buildings Caused by Earthquakes Using Machine Learning Techniques, 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India. 2019, pp.81-86.
- [4] Shigefuji, M., Takai, N., Bijukchhen, S. et al. Strong ground motion data of the 2015 Gorkha Nepal earthquake sequence in the Kathmandu Valley. *Sci Data*. 2022, 9, 513.
- [5] C. C. Prediction of Heart Disease using Different KNN Classifier, 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India. 2021, pp. 1186-1194.
- [6] N. H. Othman, K. Y. Lee, A. R. M. Radzol, W. Mansor and U. R. M. Rashid, Classification of Salivary Adulterated NS1 SERS Spectra Using PCA-Cosine-KNN, 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Shanghai, China. 2019, pp.159-163.
- [7] A. N. V. K. Swarupa, V. H. Sree, S. Nookambika, Y. K. S. Kishore and U. R. Teja, Disease Prediction: Smart Disease Prediction System using Random Forest Algorithm, 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India. 2021, pp.48-51.
- [8] S. U. Raj, M. Veera Manikanta, P. S. Sai Harsitha and M. Judith Leo, Vacant Parking Lot Detection System Using Random Forest Classification, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India. 2019, pp.454-458.
- [9] Kohavi, R., and Provost, F. Glossary of terms. *Machine Learning*. 1998, 30(2-3), 271-274.
- [10] Powers, D. M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*. 2011, 2(1), 37-63.