Volume: 5, Issue: 2
Page: 128-142
2021

International Journal of
Science and Business
Journal homepage: ijsab.com/ijsb

IJSAB International

Machine Learning Classification Based on Radom Forest Algorithm: A Review

Nasiba Mahdi Abdulkareem & Adnan Mohsin Abdulazeez

Abstract:

Machine Learning is a significant technique to realize Artificial Intelligence. The Random Forest Algorithm can be considered as one of the Machine Learning's representative algorithm, which is known for its simplicity and effectiveness. It is also can be defined as a Decision Tree-Based Classifier that chooses the best classification tree as the final classifier's classification of the algorithm via voting. Random Forest is the most accepted group classification technique because of having excellent features such as Variable Importance Measure, Out-of-bag error, Proximities, etc. Currently, it is in the new classification, intrusion detection, content information filtering, and sentiment analysis that is why there is an extensive range of applications in image processing. In this paper, the construction process of Random Forests and the study status of Random Forests would primarily be introduced in terms of capacity enhancement and performance indicators. The use of Random Forest in different fields such as Medicine, Agriculture, Astronomy, etc. is often mentioned.

IJSB
Literature review
Accepted 24 January 2021
Published 27 January 2021
DOI: 10.5281/zenodo.4471118

Keywords: *Machine Learning, Random Forest, Ensembles of Decision Tree, Classification.*

About Author (s)

Nasiba Mahdi Abdulkareem (corresponding author), Information Technology Department, Akre Technical College of Informatics, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq. E-mail: nasiba.mahdi@dpu.edu.krd
Professor Adnan Mohsin Abdulazeez, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq. E-mail: adnan.mohsin@dpu.edu.krd

1. Introduction

Machine learning (ML) is the analysis of computational algorithms that are used as a branch of artificial intelligence that progresses automatically through experience (Abdulqader et al., 2020; Adeen et al., 2020). In order to make predictions or choices without being explicitly programmed to do so, machine learning algorithms create a model based on sample data, defined as 'training data' (Zeebaree et al., 2019b). Classification is one of the supervised machine learning models with the objective of predicting the categorical class labels of new instances (discrete, unordered values, group membership) based on previous observations (Sadiq et al., 2020).A type of classification algorithm is a decision tree that constructs a model in the form of a tree. By splitting information into smaller subsets, a related decision tree is incrementally constructed (Abdulqader et al., 2020; Zeebaree et al., 2019a). A decision tree is a map or graph that is used for individuals to evaluate a course of action to display a predictive probability (Zebari et al., 2020a). A future choice, consequence, or response is defined by a branch of the decision tree. The end results are represented by the farthest nodes of the tree. The Decision Tree does not need any understanding of the domain and is simple to understand (Sadeeq & Abdulazeez, 2018; Najat & Abdulazeez, 2017) . To identify a huge number Inaccuracy of prediction can result from data set in the respective class labels using the decision tree and the results are uncertain. Instead of using a single classifier, we used many classifiers identified as ensembles by decision forest (Zebari et al., 2019a). Inaccuracy of estimation will arise from the data set in the respective class labels utilizing the decision tree to classify a huge amount, and the effects are unknown. We used several classifiers known as ensembles by decision forest (Mienye et al., 2019) instead of using a single classifier. Random forest is a scalable, easy-to-use machine learning algorithm that delivers, much of the time, a great result even without hyper-parameter tuning. Because of its flexibility and diversity, it is now one of the most used algorithms (it can be used for both classification and regression tasks). In the ensemble methodology, several classifiers are built for a particular function. Then, the separate classifiers are combined to create a new classifier (Das et al., 2007), with that said; random forests are a powerful and far more stable modeling methodology than a single decision tree. Many decision trees are aggregated to limit over fitting as well as error due to prejudice and therefore produce valuable outcomes. The Rest of this report is structured as Follows: We have the background theory in Section 2, The Related Work is in Section 3, The Results and Discussion is in Section 4, Finally Conclusion in section 5.

2. Theoretical Background

In modern times, there are a number of classification difficulties because of the huge volume of data. Many widely employed algorithms have not performed correctly in some instances (Zebari et al., 2020b). Random Forest classification system is the best approach for classifying big results. Random Forest is essentially a group of Decision Trees whose outcomes are merged into one final outcome .(Schonlau & Zou, 2020) Their ability to limit over fitting without significantly increasing error due to bias is why they are models that are so powerful. One way for Random Forests to minimize variance is by training on multiple data samples (Han et al., 2019; Zhou et al., 2020).

2.1 Decision Tree

Decision Trees use a variety of algorithms to decide to divide one node into two or more sub-nodes. The way of making sub-nodes enlarges the homogeneity of consequent sub-nodes (Kumar et al., 2016). As well as the Decision Tree divides the nodes on all available variables and then selects the split which results in the most homogeneous sub-nodes (Li et al., 2019). Random Forests are basically composed of multiple Decision Trees, which are the basic classifiers that makeup Random Forests. Common Decision Tree algorithms include ID3, C4.5,

CART (Classification and Regression Tree), etc. as one of the earliest Decision Tree algorithms, ID3 algorithm constructs Decision Tree by selecting the attribute with the largest information achievement and the critical value for node splitting. It can only support discrete data processing, and the training model is prone to over-fitting phenomenon(Singh & Giri, 2014). C4.5 algorithm is an enhancement of ID3 algorithm. In order to prevent the overfitting phenomenon, it introduces the pruning step on basis of ID3. The implementation process is to specify a threshold; the numbers of samples are smaller than the given threshold. The collection can be seen as a leaf node, which reduce the over-fitting phenomenon, but the threshold selection needs to depend on experience and lack necessary theoretical support. CART algorithm (Band et al., 2020)performs two-ways recursive segmentation on the training samples according to the Gini impurity minimum criterion divides the current sample set into two sub-sample sets, so each non-leaf node of the generated Decision Tree has two branches which are forming a binary tree and Formal Decision Tree classifier (Sarker et al., 2020) .It is important to say that CART tree is a binary tree, while ID3 and C4.5 can be multi-fork trees.

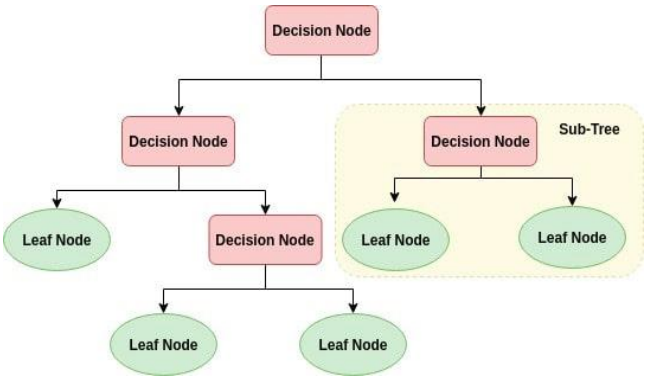

Fig. 1 Decision Tree training flow chart

2.2 Random Forest

Random Forest is a classifier consisting of a set of tree-structured classifiers with identically distributed independent random vectors and each tree casting a unit vote at input x for the most popular class (Reis et al., 2018). A random vector that is independent of the previous random vectors of the same distribution is generated and a tree is generated using the training test , an upper bound is extracted for Random Forests to get the generalization error in terms of two parameters Exactitude and interdependence of individual classifiers (Ozgode Yigin et al., 2020). Figure 2 shows the flow chart of Random Forest.

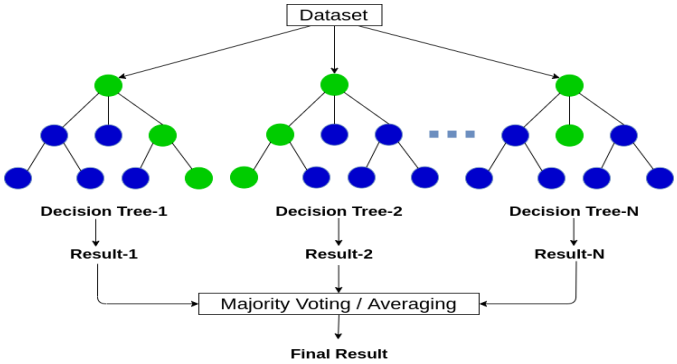

Fig. 2 Flow Chart of Random Forest

Table1: Random Forest Algorithm Advantages and Limitations

Advantages	Disadvantages
There is greater accuracy. Effective in working with large databases. It manages thousands of input variables quickly and effectively. Provides information on variables that are important and are not in the Classifying. Provides techniques to estimate incomplete data. Deals with lost details without losing accuracy. Prototypes are used to provide data or meta data on the relationship between different factors. Permits the analysis of variable relationships **(Bhattacharyya et al., 2019)**	One of the main problems found is over-fitting a single data set, especially in the tasks of regression. Random Forests have trouble dealing with multi-valued and multi-value attributes Multi-dimensionally. They prefer multi-level categorical variables (Shaik & Srinivasan, 2019)

2.3 Random Forest Algorithm

An integrated learning model suggested by Breiman in 2001, with the Decision Tree as the basic classifier is Random Forest. To get multiple subsets of samples, it implements the bootstrap method (Denisko & Hoffman, 2018), creates a Decision Tree utilizing each subset of samples, and combines several Decision Trees into a Random Forest. When the sample to be classified is reached, the final outcome of the classification is decided by a vote on the Decision Tree (Utkin et al., 2020). Generally, scholars increase the precision of the classifier starting from the classifier and reduce the association between classifiers (Demidova & Ivkina, 2019). Random Forest algorithm in the classification process, where the effects of the classification of each base classifier have a common distribution of errors, the final reduction of the classification effect is accomplished (Abdulazeez et al., 2020). Takes the test characteristics and uses the rules of each randomly generated Decision Tree to forecast the result and store the expected result (target). Determine the votes for each predicted goal. Consider the predicted high-voted goal as the final prediction from the Random Forest algorithm (Kolhe et al., 2020; Gajowniczek et al., 2020). Figure 3 illustrates the training process of the Random Forest.

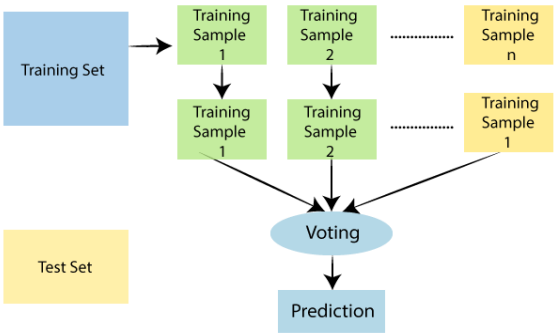

Fig. 3 Random Forest training flow chart

The Random Forest's basic algorithm steps are as follows:

As one of the main classifier applications Random Forest incorporation is made of many individual forests (Abdel-Hamid et al., 2012) .Classifiers of the Decision Tree and deciding on research samples according to such laws. The following are the steps (Bingzhen et al., 2020).

In the Random Forest algorithm, there are two steps, one is Random Forest formation, and the other is to make a guess from the first step of the Random Forest classifier (Sun et al., 2020; Kulkarni & Sinha, 2012). Here, the author first reveals the pseudo-code for the development of the Random Forest (Computer Science & Engineering &GZSCCET Bhatinda, Punjab, India et al., 2017):

1. Select "K" features at random from the complete "m" features, where k << m.
2. Calculate the node "d" among the "K" features using the best split point.
3. Using the best division to divide the network into daughter nodes.
4. Repeat measures from 1 to 3 until the number of nodes 'n' has been reached.
5. Develop a forest to build the "n" number of trees by repeating steps 1 to 4 for "n" number of times.

With the Random Forest classifier generated in the next step, we will make the forecast. The random pseudo-code for forest prediction is seen below:
Takes the test features and uses the rules of each Decision Tree generated at random to forecast the result and store the anticipated outcome (target).
For each forecast goal, measure the votes. As the final prediction from the Random Forest algorithm, consider the strongly voted predicted objective (Paul et al., 2018);(Xu, n.d.).

The Decision Formula [(Das et al., 2007)] by using shown in equation 1.

$$H(x) = \arg\max_{y} \sum_{i=1}^{k} I(h_i(x) = Y) \quad (1)$$

Where:
x = test sample
h_i = single Decision Tree
Y = output variable (i.e. classification label)
I= indicator function
H = Random Forest model
That is, the classification outcome of each test tree for the test sample is summarized and the final classification result is the class with the maximum number of votes. In addition, several Random Forest promotion algorithms have appeared and their pairs are seen in Table 2 with the Random Forest algorithm(Imaizumi et al., 2020).

Table2. Random Forest promotion algorithm

Algorithm name	Different from Random Forest
Extra trees *(Darbanian et al., 2020; Motamedidehkordi et al., 2017)*	The key distinction between Random Forests and Extra Trees (usually referred to as extreme Random Forests) lies in the fact that with each feature under consideration, a random value is chosen for the split instead of determining the locally optimal feature/split combination (for the Random Forest) (for the extra trees).
Isolation Forest (IForest) *("Isolation Forest and Local Outlier Factor for Credit Card Fraud Detection System," 2020) (Chaudhary, n.d.)*	Forest of Solitude, The Forest of Solitude functions a little better than the Forest of Random. It, therefore, produces a lot of Decision Trees, so then the path length required to extract an observation in the tree is determined.
Random Survival Forest (RSF) *(Chen et al., 2019)*	The tree-building decision is identical to RF. Each Decision Tree in the RSF is a two-class survival tree to process survival data. It is superior to other approaches in survival analysis for evidence on high-dimensional survival.

3. Related Work

Random Forest algorithm is a Decision Tree-based classifier. It selects the best classification tree as the classification algorithm of the final classifier by voting (Zebari et al., 2019b). It can be used for both classifications and regression tasks. It provides higher accuracy through cross-validation (Bargarai et al., 2020). Currently, it is in news classification, intrusion detection, content information filtering, sentiment analysis, there is a wide range of applications in the field of image processing. In terms of continuous enhancement and performance metrics, this literature review would primarily present the Decision Tree, the development method of Random Forests, and the study status of Random Forests (Jahwar & Abdulazeez, 2021) .

Iwendi & Jo (2020) proposed a model proposed to apply the Random Forest algorithm With an F1 Score of 0.866, improved by the AdaBoost algorithm On the patient dataset for COVID-19. Also, pointed out that the Boosted Random Forest algorithm gives detailed forecasts on imbalanced datasets too. The knowledge reviewed in this analysis has It indicated that among the Wuhan natives, death rates were higher. Non-natives as opposed. Male patients have had a higher percentage of Compared with female patients, the mortality risk. The largest of those impacted patients are aged 20 and 70 years of age. Zhang & Yang (2020) owed to massive human migrations, land conversions, and global environmental change, coastal regions ace a great deal of tension. Because of their geographical, heterogeneous, spectral sophistication, mapping urbanized coastal areas can be very difficult. Seven variable rating methods focused on Random Forests were tested in this research. To choose the best classification form, feature exclusion techniques were implemented using both CART and CIT models. CPVIM has been proven to be more reliable in providing stable and reasonable feature rankings from correlated remotely sensed data. The optimal model was found through the NRFE process based on the CART tree using CPVIM. It achieved an overall accuracy of 89.03% with ten features only, i.e., Green, NIR, SWIR1 and SWIR2, Greenness, MSAVI, NDII, ED, SVVI, and DEM.

Moreover, Saenz-Cogollo & Agelli (2020) presented time-domain characteristics derived from the single-lead ECG was critically chosen by their data quality, and the efficiency of the heartbeat classification using RF was reasonably assessed by adopting the (AAMI) and the inter-patient paradigm principles. The most discriminative features for the classification task were considered to be normalized features relative to R-R intervals and to the width of the main wave of the QRS complex. With the top six most insightful features and a 40-tree RF classifier, the best results were produced. The MIT-BIH Arrhythmia Database measurement culminated in an average precision of 96.14 percent for the NB, SVEB, and VEB groups, with individual F1 ratings of 97.97 percent, 73.06 percent, and 90.85 percent, respectively. Results are one of the best performances recorded to date in accordance with state-of-the-art methods tested in comparable conditions. The findings not only indicate that RF is an outstanding heartbeat classification method, but also that relatively few features are necessary to achieve state-of-the-art efficiency.

Additionally, Chai & Zhao (n.d.) Presented a modern ObRF learning method are made up of ObRF-BM and ObRF-DIL. The planned one the system analytically measures the oblique hyperplanes the costly hunt for the right function and split threshold is stopped. In addition, the decision node characteristics are projected into a random higher dimensional space, which injects further Randomness to the model of the ensemble and boosts output From ObRF. In comparison, the creation of gradual approaches for situations with sample increments and class increments, In order for the predefined model to be efficient, without laborious retraining, revised. Empirical findings suggest that the superior efficiency of the

ObRF suggested. International Conference on Artificial Intelligence and Computer Vision (2020) mentioned the Random Forests, which are variations of Decision Trees equipped with data sub-samples, are created using under-sampling and over-sampling. The author contrasts fit metrics derived from the different requirements of the models evaluated and assesses their results within and outside the study. The findings revealed that Random Forest strategies utilizing imbalanced sub-samples smaller than the initial study showed the highest efficiency and change of the Random Forests used relative to the medical dataset. Shiroyama et al. (2020)discussed the influence of sample size on habitats was explored in this review. Plots for suitability utilizing RF .The outcome revealed that the predictive one was the efficiency of the approximate RF models is positive, the sample sizes were associated. Next, it was determined that the Plots for habitat suitability are often impacted by the sample size and Output as prediction. In the case of minimal accessible sample evidence, to find a realistic approach for delineating habitat suitability plots, the "average plot of habitat suitability" was suggested. This demonstrates that the typical habitat suitability plot can theoretically be Improves also in a habitat suitability plot calculation in a Small quantity of samples. Wang & Zhu (2020)proposed a framework for soil mapping by the integration of a methodology focused on similarities and Random Forests. To check its electiveness, the approach proposed was extended to the Heshan study field. The following conclusions can be taken from this study: (1) The SB-RF system achieved better accuracy output than either the RF or SB alone, demonstrating the integrated method's e-efficacy and superiority; (2) The similarity covariates provided by the similarity-based approach embedded useful data that can e-effectively boost the accuracy of the mapping. The precision of the SB-RF system is influenced by the sampling technique.

There is four aspects in which (Cervantes et al., 2020) noted to the current literature: first, the socio-economic determinants of wellbeing, with an emphasis on European countries. Second, from the viewpoint of the economy as a whole, perceive greenhouse gas pollution. Third, there is scarce data to date indicating the degree to which government environmental spending leads to the enhancement of the welfare of a community. Fourth, employ a technique for Spontaneous Trees. As a plus, this approach offers a classification of social variables to describe life expectancy at birth according to their relative value. Moreover, based on the findings of this study of the Random Forest, it has been concluded in this paper that some roles of public spending are more relevant in understanding life expectancy at birth in European countries. Consequently, the emphasis on public services, such as environmental and social security investments, of better relevance to the health condition of the community, would be more de efficient. Thonfeld et al. (2020) illustrated an effective approach to the measurement of land use and land cover (LULC) utilizing Multi-temporal Landsat data metrics. As LULC conversion and slight within-class shifts could be detected, the combined method of spectral shift detection and PCC showed complementarity. The minimization of spurious improvements was rendered possible by utilizing the least linked multitemporal metrics of TC items. The findings have also shown that there is a rising shortage of land per capita for agricultural production in the Kilombero catchment. The Kilombero wetland is now a precious habitat situated alongside a largely unchecked river channel. In the current circumstance of steady economic and demographic development and the pace of the LULC transition in recent decades, anthropogenic land-use expansion is expected to accelerate further. Dolejš (2020)suggested a model with the objective of predicting EMS coverage with a minimum deviation from real-time was developed and detailed algorithms and model construction methods were described. The results indicate that the model based on real ambulance tracking using Random Forest learning results in a better (more accurate) travel time prediction than those obtained from an empirical model. At the same time, it was shown

that the results can be improved by using additional constraints that are potentially relevant to the accuracy of travel time prediction or training the data separately for different situations. The use of traffic-flow models complemented with EMS-specific components is another possible research route. Dikshit et al. (2020) believed Droughts will inflict significant harm to farmland and water supplies, resulting in severe economic losses and loss of life. The research may be used for other uses dependent on drought, such as urban heat, irrigation and preparedness for fire emergencies. The research results are as follows:
(i) The relative significance of the hydro-meteorological variables used to predict the drought index suggests that PET is the most critical component, apart from rainfall, accompanied by vapor pressure and mean temperature. (ii) For SPEI 1 and SPEI 3 examples, the model indicates strong forecasting capability, with the R2 value being 0.73 and 0.76, respectively. However, SPEI 3 revealed a larger number of related classes in comparison with SPEI 1 when evaluating the difference according to drought classes, thus having a marginally improved predictive power of the model for the former scenario. (iii) During the validation phase, the grouping component of the model into various drought groups was evaluated. The findings reveal that for SPEI 1 and SPEI 3 time ranges, the model was able to identify 82% and 84% correctly. The outcome demonstrates that the usage of the Random Forest model has the potential to work well for the NSW region at short-term time scales for both regression and classification problems about drought.

Lan et al. (2020a) reported the composition of ionospheric anomalies has a detrimental effect on the propagation in the ionosphere of electromagnetic waves. The automatic identification of ionospheric spread-F is of considerable importance. Three automatic spread-F identification approaches focused on machine learning are identified and implemented: Decision Tree, Random Forest, and convolutional neural network (CNN). Using a wide collection of test results, the accuracy of these automated recognition methods was checked. Results indicate that the precision of all three approaches surpassed 90% in recognizing ionograms with spread-F. Noticed that the Decision Tree approach was the simplest and easier to explain with the framework, after analyzing the findings of the three approaches, and it took the shortest interpretation period. The Random Forest approach produced stronger results than the Decision Tree system in terms of the detection results, and the CNN method was the right one to correctly classify ionograms using spread-F. Yeap (2020) used computer vision methods and machine learning To explain how they can be used to detect peaks and to conduct Random Forests binary classification. Grayscale imaging, RIP reduction, hat top Filtering and thresholding is used to minimize background and threshold noise. The picture is transformed to a binary image. Segmentation in the Watershed helped diagnose each peak compound ion is labeled and classified as a separate compound. A Full Alignment a peak table summarizing the compounds identified was created by an algorithm using compensation voltage and retention time limits. Random Forests, a model of machine learning, demonstrated strong precision, suggesting that Random Forests are a stable model for predicting binary classification on GC/DMS samples. Zong et al. (2020) proposed in particular with the use of global positioning system (GPS) data, a variety of algorithms have been proposed to identify travel modes, although most algorithms seldom recognize traffic conditions. This paper distinguishes two symbolic transport types, i.e. bus and automobile, by utilizing the random-forest approach to fill the void, which explores the related characteristic variables under varying traffic conditions. In order to minimize uncertainties between the bus and the vehicle, local congestion variables are specified. The findings show that the average detection performance of the not-in-congestion trips is as high as 94.0 percent, and that of in-congestion trips is 91.1 percent, showing that identifying traffic patterns utilizing Random Forests will reliably boost travel modes detection accuracy. Distinguishing local traffic

patterns will further boost precision, it is found. Sharma et al. (2020) think in order Argued in order to uninstall junk mails, unused storage space, and network capacity, spam emails require too much time. The filter is therefore important for the filtering of unsolicited emails with great precision. Here, the collection of features is carried out using MapReduce Minimal Redundancy Maximum Significance (mRMR) to process spam emails to pick the right classification features. By using the Random Forests algorithm, the chosen characteristics are categorized. It classifies the junk mails and ham emails in the Random Forests by way of voting methods. The distinction shows that in the distributed setting, the suggested email spam classification offers greater precision than classification using Random Forests.

Fourth International Congress on Information and Communication Technology (2019) reported that the Land cover classification based on the use of remote sensing images is one of the common remote sensing applications, and several remote sensing image classification techniques have been enhanced and implemented. In the remote sensing region, RF and SVMs, which are supervised classification techniques, have recently been employed. The author's purpose is to present findings obtained with the RF classifier and Decision Tree and to equate their utility with the methodology of the SVMs. Results shows that the efficiency of the Random Forest classifier outperforms the performance of the Decision Tree and SVM techniques regarding the amount of misclassification instances and the precision of the classification with an overall accuracy of 86 percent, whereas the accuracy of the Decision Tree is 67 percent, and the accuracy of the SVMs is 56 percent, respectively.

4. Comparison and Discussion

The above review of the recent study shows the assessment of the Random Forest algorithm in different fields of life sciences. The successful point of this algorithm as one of the powerful machine learning techniques has been approved in both classification and regression-based problems. This has been argued as the main advantage of this algorithm in the majority of these researches. A summary of the results of these studies that have been conducted during the last year (The year 2020) is given in Comparison Table3.

Table3. Summary of Literature Review Related of Random Forest Algorithm

Reference	*Year*	*Objectives*	*Data Sets*	*Results and Accuracy*	*Used Techniques*
(Iwendi & Jo, 2020)	2020	Health Forecast COVID-19 Patient	Of Kaggle as "The Novel" Dataset Corona Virus 2019 (26). From multiple outlets, such as the International Health Organisation, and the University of John Hopkins.	0.86%	RF, AdaBoost , BRFC
(Zhang & Yang, 2020)	2020	Selecting the best function domain to maximize the classification of land cover in a dynamic urbanized coastal region.	NRFE-CPVIM	89.03%,71.79%	DT,RF, CPVIM , CART-RF,CIT-RF
(Saenz-Cogollo & Agelli, 2020)	2020	Investigating Feature Selection and Random Forests for Inter-	ECG signals used in from the MIT-BIH online library. Database Arrhythmia	96.14% ,97.97%, 73.06%, 90.85%	RFC, AAMI, ECG, GMM, Bagging Trees (BT).

		Patient Heartbeat Classification.			
(Chai & Zhao, n.d.)	2020	Providing dual incremental learning (DIL) ability for Oblique Random Forests(ObRF) to conduct on-the-fly classification	ObRF-BM, ObRF-DIL	86.49% , 85.39%, 85.88%, 79.18%, 71.66%, 73.98%,	SVM,DIL,RF,DT
(International Conference on Artificial Intelligence and Computer Vision et al., 2020)	2020	To address the imbalance dilemma in medical datasets.	Eight medical datasets selected.	89%,	DT,RFC,CART,RF
(Shiroyama et al., 2020)	2020	Knowing the effects of sample size on the habitat and the bluegill (Lepomis macrochirus), the largest exotic fish species in the rivers of Japan, was chosen as the target.	Japan's State Censuses on River Ecosystems (NCRE).	0.75%	RF, (PCC), p-value SDM (RF) combined with partial dependence function.
(Wang & Zhu, 2020)	2020	To present a soil mapping method	mapping soil subgroups	66.61%, 57.39%, 59.62%,66.67%	SB, RF, Integration of SB and RF(SB-RF).
(Cervantes et al., 2020)	2020	Identify and classify the relative significance of many socioeconomic variables in the European Union that clarify life expectancy at birth The 2008–2017 era.	main greenhouse gases (CO2, N2O, and CH4)	98.98%	RF, RFC, DT
(Thonfeld et al., 2020)	2020	Assess land use/land cover (LULC) adjustments in Tanzania's Kilombero catchment.	two major regions, one in the Kilombero floodplain, one in the West of the catchment	50%,58.97%	PCC,RF, RCVA
(Dolejš, 2020)	2020	Generating a regional-level geographical coverage strategy for emergency	Real data from a single region of Czechia (Central Europe)	76%	RFE,DT, EMS

		medical services			
(Dikshit et al., 2020)	2020	Spatio-Temporal Drought Forecasting in New South Wales, Australia for the short term.	Climate Research Unit Time Series (CRU-TS) dataset.	82%, 84%	DT, RF ,ERT
(Lan et al., 2020b)	2020	A comparative analysis of the spread-F identity Decision Tree, Random Forest, and convolutional neural network	Wuhan Ionospheric Sounding System (WISS)	90%	DT, RF, and deep learning method of CNN

Classification of the COVID-19 patient dataset and their status has been successfully predicted by Iwendi & Jo (2020) and using the new improved version of the RF algorithm. (Saenz-Cogollo & Agelli, 2020) reported the great performance of the RF in heartbeat classification that was based on cardiac dataset obtained from different patients. Similarly, the RF tool possesses a great potential to deal even with the medical imbalanced dataset and with a good performance in the prediction of the outputs (International Conference on Artificial Intelligence and Computer Vision, 2020). Dolejš (2020) adopted the RF technique for tracking the ambulance movements. It has been claimed that the model helps the ambulances to reach the patient at any location and in the shortest possible time. During the validation of their results, it has been concluded that the travel time predicated by the RF-based model is generally improved compared with other locally used empirical models. Zhang & Yang (2020) highlighted the computational efficiency of RF framework, which is based on the variable selection process, to predict the optimal domain area to improve the land cover classification in urbanized coastal areas. With the same concept, the RF has been used by (Thonfeld et al., 2020) for classification of the land use (land cover) information that will help decision-makers to plan for sustainable land management in the future. An integrated RF model was used by (Wang & Zhu, 2020)to predict the spatial variation of soil information such as type and/or properties (soil mapping). In this framework, the RF is integrated with similarity-based methods. The results clearly show the higher accuracy of the developed model (71.79%) than applications of both approaches RF (66.67%) and similarity-based method (58.97%) separately or alone. Furthermore, the RF approach itself still performed with better accuracy than the similarity-based method. In addition, Dikshit et al. (2020)examined the tool for both classification and regression problems for short-term drought forecasting of the New South Wales region in Australia confirming its good performance. The relative importance of several socioeconomic factors in the European Union (EU) countries has been identified and classified using the RF algorithm(Cervantes et al., 2020). Their study identified the main sources of greenhouse gas emissions in these countries that affect the future public expenditures to pay for protection of the society and environment. Moreover, the RF has been utilized by Shiroyama et al. (2020) to study the effect of sample size on habitat suitability for a type of exotic fish in the Japan rivers called Bluegill. Their results show the good accuracy of RF even under a small sample size. Finally, in an interesting study, Lan et al. (2020a)compared the performance of three machine learning techniques to identify ionospheric spread-F (electron density perturbation in the F-layer of ionization). Decision Tree, Random Forest, and convolutional neural network (CNN) are the three-algorithms used in the study. They concluded that the RF has a better potential than the Decision Tree method but less than CNN to identify the problem.

5. Conclusion

This paper has shown an overview of Random Forest and its performance in Classification Model. Random Forest is an ensemble classifier that includes multiple classifiers to predict class label values with past data set. Random Forests are fast to build and even faster to predict. They don't require any cross-validation or fully parallelizable. Random Forest algorithms are often more accurate than a single classifier. It can handle the data without preprocessing, which means data doesn´t need be rescaled or transformed. However, as a widely used algorithm, it is worthy of additional study on improving classification accuracy.

References

Abdel-Hamid, O., Mohamed, A., Jiang, H., & Penn, G. (2012). Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4277–4280. https://doi.org/10.1109/ICASSP.2012.6288864

Abdulazeez, A. M., Sulaiman, M. A., & Qader, D. (2020). *Evaluating Data Mining Classification Methods Performance in Internet of Things Applications*. *1*(2), 15.

Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). *Machine Learning Supervised Algorithms of Gene Selection: A Review*. *62*(03), 13.

Adeen, I. M. N., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). *Systematic Review of Unsupervised Genomic Clustering Algorithms Techniques for High Dimensional Datasets*. *62*(03), 21.

Band, S. S., Janizadeh, S., Saha, S., Mukherjee, K., Bozchaloei, S. K., Cerdà, A., Shokri, M., & Mosavi, A. (2020). Evaluating the Efficiency of Different Regression, Decision Tree, and Bayesian Machine Learning Algorithms in Spatial Piping Erosion Susceptibility Using ALOS/PALSAR Data. *Land*, *9*(10), 346. https://doi.org/10.3390/land9100346

Bargarai, F. A. M., Abdulazeez, A. M., Tiryaki, V. M., & Zeebaree, D. Q. (2020). Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio. *International Journal of Interactive Mobile Technologies (IJIM)*, *14*(13), 107. https://doi.org/10.3991/ijim.v14i13.14211

Bhattacharyya, S., Hassanien, A. E., Gupta, D., Khanna, A., & Pan, I. (Eds.). (2019). *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2* (Vol. 56). Springer Singapore. https://doi.org/10.1007/978-981-13-2354-6

Bingzhen, Z., Xiaoming, Q., Hemeng, Y., & Zhubo, Z. (2020). A Random Forest Classification Model for Transmission Line Image Processing. *2020 15th International Conference on Computer Science & Education (ICCSE)*, 613–617. https://doi.org/10.1109/ICCSE49874.2020.9201900

Cervantes, P. A. M., López, N. R., & Rambaud, S. C. (2020). *Life Expectancy at Birth in Europe: An Econometric Approach Based on Random Forests Methodology*. 17.

Chai, Z., & Zhao, C. (n.d.). Multiclass Oblique Random Forests With Dual-Incremental Learning Capacity. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, 12.

Chaudhary, A. (n.d.). *An improved Random Forest Classifier for multi-class classification*. 25.

Chen, S., Mulder, V. L., Martin, M. P., Walter, C., Lacoste, M., Richer-de-Forges, A. C., Saby, N. P. A., Loiseau, T., Hu, B., & Arrouays, D. (2019). Probability mapping of soil thickness by random survival forest at a national scale. *Geoderma*, *344*, 184–194. https://doi.org/10.1016/j.geoderma.2019.03.016

Computer Science & Engineering &GZSCCET Bhatinda, Punjab, India, Goel, E., Abhilasha, Er., & Computer Science & Engineering &GZSCCET Bhatinda, Punjab, India. (2017). Random Forest: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(1), 251–257. https://doi.org/10.23956/ijarcsse/V7I1/01113

Darbanian, E., Rahbari, D., Ghanizadeh, R., & Nickray, M. (2020). IMPROVING RESPONSE TIME OF TASK OFFLOADING BY RANDOM FOREST, EXTRA-TREES AND ADABOOST CLASSIFIERS IN MOBILE FOG COMPUTING. *Jordanian Journal of Computers and Information Technology*, *0*, 1. https://doi.org/10.5455/jjcit.71-1590557276

Das, K., Behera, R. N., & Tech, B. (2007). *A Survey on Machine Learning: Concept, Algorithms and Applications*. *5*(2), 10.

Demidova, L., & Ivkina, M. (2019). Defining the Ranges Boundaries of the Optimal Parameters Values for the Random Forest Classifier. *2019 1st International Conference on Control Systems, Mathematical*

Modelling, Automation and Energy Efficiency (SUMMA), 518–522. https://doi.org/10.1109/SUMMA48161.2019.8947569

Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, *115*(8), 1690–1692. https://doi.org/10.1073/pnas.1800256115

Dikshit, A., Pradhan, B., & Alamri, A. M. (2020). Short-Term Spatio-Temporal Drought Forecasting Using Random Forests Model at New South Wales, Australia. *Applied Sciences*, *10*(12), 4254. https://doi.org/10.3390/app10124254

Dolejš, M. (2020). Generating a spatial coverage plan for the emergency medical service on a regional scale: Empirical versus random forest modelling approach. *Journal of Transport Geography*, 10.

Fourth international congress on information and communication technology. (2019). SPRINGER. https://link.springer.com/book/10.1007/978-981-15-0637-6

Gajowniczek, K., Grzegorczyk, I., Ząbkowski, T., & Bajaj, C. (2020). Weighted Random Forests to Improve Arrhythmia Classification. *Electronics*, *9*(1), 99. https://doi.org/10.3390/electronics9010099

Han, J., Fang, M., Ye, S., Chen, C., Wan, Q., & Qian, X. (2019). Using Decision Tree to Predict Response Rates of Consumer Satisfaction, Attitude, and Loyalty Surveys. *Sustainability*, *11*(8), 2306. https://doi.org/10.3390/su11082306

Imaizumi, T., Okada, A., Miyamoto, S., Sakaori, F., Yamamoto, Y., & Vichi, M. (Eds.). (2020). *Advanced Studies in Classification and Data Science*. Springer Singapore. https://doi.org/10.1007/978-981-15-3311-2

International Conference on Artificial Intelligence and Computer Vision, Azar, A. T., Gaber, T., Oliva, D., Ṭulbah, M. F., & Hassanien, A. E. (2020). *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*. Springer. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=6144671

Isolation Forest and Local Outlier Factor for Credit Card Fraud Detection System. (2020). *International Journal of Engineering and Advanced Technology*, *9*(4), 261–265. https://doi.org/10.35940/ijeat.D6815.049420

Iwendi, C., & Jo, O. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health*, *8*, 9.

Jahwar, A. F., & Abdulazeez, A. M. (2021). *META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING: A REVIEW*. 20.

Kolhe, M. L., Tiwari, S., Trivedi, M. C., & Mishra, K. K. (Eds.). (2020). *Advances in Data and Information Sciences: Proceedings of ICDIS 2019* (Vol. 94). Springer Singapore. https://doi.org/10.1007/978-981-15-0694-9

Kulkarni, V. Y., & Sinha, P. K. (2012). Pruning of Random Forest classifiers: A survey and future directions. *2012 International Conference on Data Science & Engineering (ICDSE)*, 64–68. https://doi.org/10.1109/ICDSE.2012.6282329

Kumar, G. K., Viswanath, P., & Rao, A. A. (2016). Ensemble of randomized soft decision trees for robust classification. *Sadhana*. https://doi.org/10.1007/s12046-016-0465-z

Lan, T., Hu, H., Jiang, C., Yang, G., & Zhao, Z. (2020a). A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. *Advances in Space Research*, *65*(8), 2052–2061. https://doi.org/10.1016/j.asr.2020.01.036

Lan, T., Hu, H., Jiang, C., Yang, G., & Zhao, Z. (2020b). A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. *Advances in Space Research*, *65*(8), 2052–2061. https://doi.org/10.1016/j.asr.2020.01.036

Li, Y., Jiang, Z. L., Yao, L., Wang, X., Yiu, S. M., & Huang, Z. (2019). Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties. *Cluster Computing*, *22*(S1), 1581–1593. https://doi.org/10.1007/s10586-017-1019-9

Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: A review. *Procedia Manufacturing*, *35*, 698–703. https://doi.org/10.1016/j.promfg.2019.06.011

Motamedidehkordi, N., Amini, S., Hoffmann, S., Busch, F., & Fitriyanti, M. R. (2017). Modeling tactical lane-change behavior for automated vehicles: A supervised machine learning approach. *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 268–273. https://doi.org/10.1109/MTITS.2017.8005678

Najat, N., & Abdulazeez, A. M. (2017). Gene clustering with partition around mediods algorithm based on weighted and normalized mahalanobis distance. *2017 International Conference on Intelligent*

Informatics and Biomedical Sciences (ICIIBMS), 140–145. https://doi.org/10.1109/ICIIBMS.2017.8279707

Ozgode Yigin, B., Algin, O., & Saygili, G. (2020). Comparison of morphometric parameters in prediction of hydrocephalus using random forests. *Computers in Biology and Medicine*, *116*, 103547. https://doi.org/10.1016/j.compbiomed.2019.103547

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved Random Forest for Classification. *IEEE Transactions on Image Processing*, *27*(8), 4012–4024. https://doi.org/10.1109/TIP.2018.2834830

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. *The Astronomical Journal*, *157*(1), 16. https://doi.org/10.3847/1538-3881/aaf101

Sadeeq, H., & Abdulazeez, A. M. (2018). Hardware Implementation of Firefly Optimization Algorithm Using FPGAs. *2018 International Conference on Advanced Science and Engineering (ICOASE)*, 30–35. https://doi.org/10.1109/ICOASE.2018.8548822

Sadiq, S. S., Abdulazeez, A. M., & Haron, H. (2020). Solving multi-objective master production schedule problem using memetic algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(2), 938. https://doi.org/10.11591/ijeecs.v18.i2.pp938-945

Saenz-Cogollo, J. F., & Agelli, M. (2020). Investigating Feature Selection and Random Forests for Inter-Patient Heartbeat Classification. *Algorithms*, *13*(4), 75. https://doi.org/10.3390/a13040075

Sarker, I. H., Colman, A., Han, J., Khan, A. I., Abushark, Y. B., & Salah, K. (2020). BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model. *Mobile Networks and Applications*, *25*(3), 1151–1161. https://doi.org/10.1007/s11036-019-01443-z

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Shaik, A. B., & Srinivasan, S. (2019). A Brief Survey on Random Forest Ensembles in Classification Model. In S. Bhattacharyya, A. E. Hassanien, D. Gupta, A. Khanna, & I. Pan (Eds.), *International Conference on Innovative Computing and Communications* (Vol. 56, pp. 253–260). Springer Singapore. https://doi.org/10.1007/978-981-13-2354-6_27

Sharma, N., Chakrabarti, A., & Balas, V. E. (Eds.). (2020). *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 1* (Vol. 1042). Springer Singapore. https://doi.org/10.1007/978-981-32-9949-8

Shiroyama, R., Wang, M., & Yoshimura, C. (2020). Effect of sample size on habitat suitability estimation using random forests: A case of bluegill, *Lepomis macrochirus*. *Annales de Limnologie - International Journal of Limnology*, *56*, 13. https://doi.org/10.1051/limn/2020010

Singh, S., & Giri, M. (2014). Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm: A Survey. *International Journal of Advanced Information Science and Technology*, 6.

Sun, Y., Li, Y., Zeng, Q., & Bian, Y. (2020). Application Research of Text Classification Based on Random Forest Algorithm. *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 370–374. https://doi.org/10.1109/AEMCSE50948.2020.00086

Thonfeld, F., Steinbach, S., Muro, J., & Kirimi, F. (2020). *Long-Term Land Use/Land Cover Change Assessment of the Kilombero Catchment in Tanzania Using Random Forest Classification and Robust Change Vector Analysis*. 25.

Utkin, L. V., Kovalev, M. S., & Coolen, F. P. A. (2020). Imprecise weighted extensions of random forests for classification and regression. *Applied Soft Computing*, *92*, 106324. https://doi.org/10.1016/j.asoc.2020.106324

Wang, D., & Zhu, A.-X. (2020). *Soil Mapping Based on the Integration of the Similarity-Based Approach and Random Forests*. 16.

Xu, R. (n.d.). *Improvements to random forest methodology*. 88.

Yeap, D. (2020). Peak detection and random forests classification software for gas chromatography/differential mobility spectrometry (GC/DMS) data. *Chemometrics and Intelligent Laboratory Systems*, 9.

Zebari, D. A., Haron, H., Zeebaree, D. Q., & Zain, A. M. (2019a). A Simultaneous Approach for Compression and Encryption Techniques Using Deoxyribonucleic Acid. *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 1–6. https://doi.org/10.1109/SKIMA47702.2019.8982392

Zebari, D. A., Haron, H., Zeebaree, D. Q., & Zain, A. M. (2019b). A Simultaneous Approach for Compression and Encryption Techniques Using Deoxyribonucleic Acid. *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 1–6. https://doi.org/10.1109/SKIMA47702.2019.8982392

Zebari, D. A., Zeebaree, D. Q., Saeed, J. N., Zebari, N. A., & AL-Zebari, A. (2020a). *Image Steganography Based on Swarm Intelligence Algorithms: A Survey*. 14.

Zebari, D. A., Zeebaree, D. Q., Saeed, J. N., Zebari, N. A., & AL-Zebari, A. (2020b). *Image Steganography Based on Swarm Intelligence Algorithms: A Survey*. 14.

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019a). Machine learning and Region Growing for Breast Cancer Segmentation. *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 88–93. https://doi.org/10.1109/ICOASE.2019.8723832

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019b). Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 106–111. https://doi.org/10.1109/ICOASE.2019.8723827

Zhang, F., & Yang, X. (2020). Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection. *Remote Sensing of Environment*, *251*, 112105. https://doi.org/10.1016/j.rse.2020.112105

Zhou, Z., Wang, Y., He, X., & Zhang, X. (2020). Optimization of Random Forests Algorithm Based on ReliefF-SA. *IOP Conference Series: Materials Science and Engineering*, *768*, 072065. https://doi.org/10.1088/1757-899X/768/7/072065

Zong, F., Zeng, M., He, Z., & Yuan, Y. (2020). Bus-Car Mode Identification: Traffic Condition–Based Random-Forests Method. *Journal of Transportation Engineering, Part A: Systems*, *146*(10), 04020113. https://doi.org/10.1061/JTEPBS.0000442

Cite this article:

Nasiba Mahdi Abdulkareem & Adnan Mohsin Abdulazeez (2021). **Machine Learning Classification Based on Radom Forest Algorithm: A Review**. *International Journal of Science and Business, 5*(2), 128-142. doi: https://doi.org/10.5281/zenodo.4471118

Retrieved from http://ijsab.com/wp-content/uploads/676.pdf

Published by

IJSAB International