

Null-values imputation using different modification random forest algorithm

Maad M. Mijwil¹, Alaa Wagih Abdulqader¹, Sura Mazin Ali², Ahmed T. Sadiq³

¹Department of Computer Techniques Engineering, Baghdad College of Economic Sciences University, Baghdad, Iraq

²College of Political Science, Mustansiriyah University, Baghdad, Iraq

³Department of Computer Science, University of Technology, Baghdad, Iraq

Article Info

Article history:

Received Jan 28, 2022

Revised Jul 20, 2022

Accepted Aug 18, 2022

Keywords:

COVID-19

Datasets

Decision making

Machine learning

Null values

Random forest

University of California Irvine

ABSTRACT

Today, the world lives in the era of information and data. Therefore, it has become vital to collect and keep them in a database to perform a set of processes and obtain essential details. The null value problem will appear through these processes, which significantly influences the behaviour of processes such as analysis and prediction and gives inaccurate outcomes. In this concern, the authors decide to utilise the random forest technique by modifying it to calculate the null values from datasets got from the University of California Irvine (UCL) machine learning repository. The database of this scenario consists of connectionist bench, phishing websites, breast cancer, ionosphere, and COVID-19. The modified random forest algorithm is based on three matters and three number of null values. The samples chosen are founded on the proposed less redundancy bootstrap. Each tree has distinctive features depending on hybrid features selection. The final effect is considered based on ranked voting for classification. This scenario found that the modified random forest algorithm executed more suitable accuracy results than the traditional algorithm as it relied on four parameters and got sufficient accuracy in imputing the null value, which is grown by 9.5%, 6.5%, and 5.25% of one, two and three null values in the same row of datasets, respectively.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Maad M. Mijwil

Computer Engineering Techniques Department, Baghdad College of Economics Sciences University

Baghdad Province, Yarmouk, Nafaq Al-Shurta, Iraq

Email: mr.maad.alnaimiy@baghdadcollege.edu.iq

1. INTRODUCTION

Machine learning [1]–[4] is the most exciting science today in the research community, which is characterised by its ability to design and develop algorithms that allow machines to learn [5], [6]. It is a sub-field of artificial intelligence where the learning process consists of automatically extracting rules and patterns from a data file [7], [8]. Machine learning is closely related to fields such as data mining, statistics, pattern recognition, other things [9]–[11]. Supervised machine learning algorithms are illustrated by using new practices to predict future events and using what has been learned from past practices to recent data [12]–[15]. In addition, these algorithms analyse well-known scaling data through which they produce a function to make predictions about the output values, whereby the system can provide targets for any new input after adequate training [16]–[18]. Furthermore, machine learning algorithms can compare their calculated and accurate outputs to find errors in which the model can be modified accordingly [19]–[22]. One of the most classical machine learning techniques utilised for prediction is the random forest [23]–[25]. This technique is marked by being more flexible and straightforward to predict [26], as the forest consists of trees, and it is said that the more trees, the more influential the forest. In other words, the random forest generates

decision trees based on randomly selected data samples [27], [28]. Then the predictions are got from each tree, and the best accurate result is chosen through voting, which is a good indication of the significance of this technique [29]. In general, this technique is employed for both classification and regression [30], [31].

The most critical issue that databases face is the existence of null value [32], as organisations rely heavily on the collection, storage, and analysis of this value for decision-making purposes. In short, a null value can be described as an empty field and means that the values are missing or unknown. Databases are a set of columns and rows that include data [33], but some of them will consist of a null or missing value [34]. Moreover, dealing with or knowing this value is not effortless as it may take a great time to realise it and understand its whereabouts [35]. As a result, databases suffer significantly from the problem of empty data that leads to inaccurate records and incorrect calculations, which leads to a return to the traditional manual method of data entry and therefore there will be a great effort and time in managing the database and consequently unreliable data will be obtained.

The foremost contribution of this scenario is to make different modifications to the random forest algorithm to impute the null value from five datasets gathered from the University of California Irvine (UCI) machine learning repository. The modification process depends on three main things (bootstrap with less redundancy, add features selection, and modified ranking stage) that are improved within the algorithm. Also, this scenario compares the modified algorithm with the algorithm without modification to know the performance of the two approaches in estimating null values and reaching convincing effects.

2. LITERATURE SURVEY

This section will address a bunch of literature involved in the random forest technique in solving a null values or missing values in large datasets. In a study executed by Sadiq *et al.* [36], they proposed using swarm intelligence and iterative dichotomiser 3 (ID3) techniques to solve the problem of null values in a large set of data. The intelligent swarm algorithm is used to feature selection that represents the bee's algorithm, while ID3 is used to find the statistics effects. This study makes a comparison between these two approaches for estimating null values; the outcomes indicate that the best performance is for ID3 in finding results without affecting the accuracy of the null value and no matter how much these values improved. Sadiq and Chawishly [37] executed the growth and improvement of the performance of the ID3 algorithm to solve the problem of null values in a large dataset. This investigation concluded, in the event of the happening of null values one and two with a row, the proposed system has the ability to estimate 99% of the null values, as well as if three null values appear within the row, the approximation is 97%, which are efficient and sound effects. In a study conducted by Ramosaj and Pauly [38], they suggested involving several techniques (stochastic gradient tree boosting, C5.0 algorithm, and random forest) in predicting missing values from credit information and Facebook data. The authors are able to develop these techniques to work more efficiently, as they are able to analyse the performance of obtaining continuous categorical and mixed data. It is concluded that the best performance was for the random forest as it gave high effects in finding the missing values in less time.

According to Salman *et al.* [39], they presented developing a random forest algorithm to increase its performance via meerkat clan algorithm to impute the missing value. After 100 iterations, the performance and accuracy of the random forest are good in calculating these values, but at 200 and 300 iterations, the execution becomes more complex. Increasing the block size in the modified algorithm improves the accuracy of null-value computation. This paper is characterized by the use of types of null values (categorical and numeric), which makes this piece more efficient. In a study by Jackins *et al.* [40], suggested that artificial intelligence techniques (naive bayes and random forest) be applied to predict diabetes, heart disease, and breast cancer. The database for this investigation is taken from National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), and all patients' data are from 21-year-old females. After running several experiments, it is found that a missing value is replaced with null values. The results of this study prove the ability of the techniques to remove the missing value and the efficiency of data classification. Another study executed by Gök and Olgun [41] collected blood samples from patients from Einstein Hospital in Brazil. They used them to predict the level of severity of COVID-19 utilising machine learning algorithms (decision tree, random forest, k-nearest neighbour, support vector machine classifier, gradient boosting, Gaussian naive bayes, multi-layer perceptron, Gaussian process). A set of missing data appeared during the work that affected the work, but they can use several approaches to fill the missing values, which are replaced with the most common value. This study got an accuracy of 0.98 from the random forest classifier.

3. METHOD

Random forest is a supervised algorithm [42]–[45]; by its name, its work is understood, and it makes a random forest that is its goal. It relies in its work on creating multiple decision trees and combining them to obtain

it is less inclined to overfitting than decision tree and other algorithms, and it's essential to demonstrate the significance of features. The overfitting phenomenon is more insignificant in the tree if the dataset increases, as a sufficient amount of data assists machine learning models in finding new patterns efficiently.

4. THE PROPOSED WORK

This scenario concentrates on modifying three critical points in the random forest algorithm: bootstrap with less redundancy, add features selection method, and modified ranking stage. Besides, bootstrap is a crucial stage in the random forest algorithm. In the modification step, a specific bootstrap strategy is based on decreasing the redundant of samples. Reducing the redundancy will increase the diversity of samples. Algorithm 1 illustrate the essential idea (steps) of bootstrap with less redundancy. This idea will guarantee a fair diversity of bootstrap samples that leads to different trees in the random forest.

Algorithm 1. Bootstrap with less redundancy

For $i = 1$ to No. of Samples **Do**

Repeat

Select sample k ;

Until

the similarity between sample k and other is greater than threshold

End For

Moreover, to increase the performance of the random forest algorithm in the null-value estimation problem, the proposed modification of this algorithm concentrates on several steps. Features selection step plays a significant role to increase the accuracy of the random forest algorithm. Thus, the proposed modification will be making this step hybrid, it depends on the hybrid feature selection method. This method indicates that the selected features will be depending on more than one feature. Also, this method can be calculated with (1). From this equation, the random forest will be selecting the features depending on two feature selection methods. Thus, the selected features will be more powerful and relevant to the target.

$$\text{Hybrid Feature Selection} = w * \text{Feature Selection1} + (w - 1) * \text{Feature Selection2} \quad (1)$$

Another modification is based on the ranking strategy of trees. In fact, the random forest algorithm before it is modified builds a set n of tree classifications to assume the assumed outcome from the predictors. In addition, each tree is trained on a different specific sample of N subjects with a random subset of m tries predictors believed in every node from the tree. The primary purpose of random forest is to aggregate tree-level effects evenly across trees. In general, the traditional random forest algorithm is enforced for structuring forest trees, but the ranking is based on the undertaking of tree aggregation. Notably, every tree in the forest's ranking class 'votes' is believed. Thus, the superior-performing trees are ranked extra accurate. In other words, the ranking depends directly on the performance; its execution on another data set that is matching and differs in size will lead to calculating the bias prediction error rating. The data diverges originally into training and testing sets during the traditional performance of this algorithm in order to avert the bias while making trees on the bootstrap samples. By utilising the individuals of out-of-bag error, the predictive rating ability for each tree is calculated. In this scenario, the training data of ranking random forests included three quarters of the actual sample. Thus, approximately one half of the completed sample is in-bag in every tree, is employed to construct the tree, and one quarter is out-of-bag. Likewise, it is used to estimate tree implementation to calculate tree accuracy. Subsequent, the tree accuracy is calculated in the training data. Also, n trees are operated to gain votes for one quarter by observing independent test groups, where the votes (predicted classifications) over trees using ranking. Algorithm 2 illustrates the stages of the modified random forest algorithm with the ranking prediction for the class, which is based on every tree in it. The principal stages of this scenario are:

- Stage I: n no. of random records is accepted from the dataset having k no. of records. The samples selected are founded on the proposed less redundancy bootstrap.
- Stage II: Unique decision trees are created for each sample. Each tree has distinctive features depending on hybrid features selection.
- Step III: Each decision tree will generate an effect.
- Step IV: The Final effect is evaluated based on ranked voting for classification.

Algorithm 2. Random forest modification

Begin

For each tree in the random forest

if $(0.6 \leq \text{Tree's Accuracy} < 0.80)$ **Then**

add tree to predict list within 1 point;

```

else if ( $0.8 \leq \text{Tree's Accuracy} < 0.95$ ) Then
  add tree to predict list within 2 points;
else if ( $\text{Tree's Accuracy} > 0.95$ ) Then
  add tree to predict list within 3 points;
end if
end for
predict = obtain more frequency tree in predict
appropriate = match between predict all and actual
RankVote = correct ÷ no. of class in test data
End

```

5. EXPERIMENTAL RESULTS

5.1. Dataset description and parameters

In this scenario, the proposed algorithm is executed on five datasets shown in Table 1. The first dataset is connectionist Bench include sonar, mines vs. rocks dataset [63]. The assignment is to train a network to determine sonar signals reflected off a metal cylinder and those reflected off a roughly cylindrical rock. This dataset includes files; the first is "sonar. mines" consists of 111 patterns achieved by bouncing sonar signals off a metal cylinder at different angles and under other circumstances. The second is "sonar. rocks" with 97 patterns earned of rocks under the equal status. The transmitted sonar signal is a frequency-modulated chirp, growing in frequency. Moreover, this dataset is characterised as the signals from the collection of different part angles, travelling 90 degrees for the cylinder and 180 degrees for the rock. In addition, every pattern in this dataset consists of a set of 60 numbers between the scopes of 0.0 to 1.0. Also, every number symbolises the energy within a characteristic frequency band, integrated over an express length of time. The integration aperture for heightened frequencies materialises later since these frequencies are subsequently transmitted during the chirp. The label connected with every record includes (*R*) if the object is a rock while (*M*) if it is a mine (metal cylinder). On the other hand, the labels' numbers are in growing order of factor angle, but the angle is not encoded directly. The second dataset [64] is data collected from phishing sites, namely phish tank archive, Google searching operators, miller smiles archive while the third dataset is breast cancer Wisconsin [65]. The fourth dataset is Ionosphere dataset classification of radar returns from the ionosphere [66]. Finally, the fifth dataset is COVID-19 pandemic. There are several parameters in the proposed modified random forest algorithm for null-values imputation. Table 2 includes each parameter's ranges value. In this scenario, four feature selection methods have been utilised in the experiments: Information Gain, Gini Index, Chi-Squared and Correlation.

Table 1. Dataset description

	Datasets				
	Connectionist bench	Phishing websites	Breast cancer wisconsin	Ionosphere	COVID-19
Data set characteristics	Multivariate	Multivariate	Multivariate	Multivariate	Multivariate
Attribute characteristics	Real	Integer	Integer	Integer, Real	N/A
Associated tasks	Classification	Classification	Classification	Classification	Classification
Number of instances	208	1353	699	351	14
Number of attributes	60	10	10	34	7
Null values?	N/A	N/A	Yes	No	N/A
Area	Physical	Computer	Life	Physical	Computer
Date donated	N/A	2016-11-02	1992-07-15	1989-01-01	2020-04-24
Number of web hits	116625	33498	389445	166509	47691

Table 2. Parameters ranges

Parameter	Values	Means
No. of trees	10–30	Number of trees in the random forest.
Weight (W)	0.4–0.6	Weight for hybrid features selection.
Threshold	0.6–0.8	Threshold value for bootstrap sampling.
No. of null values	1–3	Number of null values in each dataset row.

5.2. The effects and discussion

Several experimental results have been conducted to test the proposed algorithm within ranges of parameters in Table 2. Typically, three matters of null-values imputation, which are loss 1, 2 and 3 values in each row, have been taken, respectively. The proposed algorithm has experimented with 10, 20 and 30 trees in the random forest in each matter. Also, applied different values of threshold (0.6, 0.7, 0.8) and different weight of hybrid feature selection ($W=0.4, 0.5$ and 0.6). The accuracy of null-values estimation is computed using (2):

$$Null\ Value\ Accuracy = \frac{Actual\ Correct\ Null\ Values}{Desired\ Null\ Values} \tag{2}$$

Several experiments selected two important feature selection methods (Information Gain and Gini Index) within different weight values. Directly, the effects of this scenario will be given. Matter I: loss 1 value in each row, the experimental results performance is exhibited in Tables 3-5. Matter II: loss 2 values in each row, the experimental results performance is exhibited in Tables 6-8. Matter III: loss 3 values in each Row, the experimental results performance is exhibited in Tables 9-11. Also, the original random forest algorithm runs on the same dataset matters. Table 12 displays the most acceptable results of the proposed work compared with the original random forest.

Table 3. One null-value accuracy within 10 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	15%	13%	14%	19%	21%	18%	16%	17%	15%
Phishing websites	67%	66%	66%	68%	69%	67%	68%	68%	66%
Breast cancer	85%	86%	85%	85%	87%	86%	84%	86%	85%
Ionosphere	32%	33%	33%	32%	34%	33%	32%	34%	33%
COVID-19	9%	9%	8%	9%	11%	12%	11%	13%	13%

Table 4. One null-value accuracy within 20 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	16%	18%	18%	21%	23%	19%	17%	16%	16%
Phishing websites	68%	67%	66%	69%	71%	70%	68%	69%	67%
Breast cancer	85%	87%	86%	89%	93%	87%	85%	86%	85%
Ionosphere	34%	34%	33%	35%	36%	34%	33%	34%	33%
COVID-19	11%	11%	10%	12%	11%	12%	10%	11%	10%

Table 5. One null-value accuracy within 30 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	16%	18%	18%	20%	20%	19%	17%	16%	16%
Phishing websites	68%	68%	66%	68%	69%	70%	67%	65%	66%
Breast cancer	86%	88%	85%	86%	89%	86%	84%	87%	84%
Ionosphere	36%	34%	35%	36%	39%	35%	34%	35%	33%
COVID-19	12%	11%	11%	12%	13%	11%	12%	11%	10%

Table 6. Two null-values accuracy within 10 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	13%	15%	14%	14%	16%	15%	13%	13%	14%
Phishing Websites	2%	2%	1%	2%	3%	1%	2%	1%	1%
Breast Cancer	77%	74%	77%	79%	78%	77%	78%	77%	78%
Ionosphere	15%	15%	14%	15%	16%	14%	14%	13%	13%
COVID-19	1%	1%	1%	1%	1%	1%	2%	1%	1%

Table 7. Two null-value accuracy within 20 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	15%	16%	15%	17%	17%	15%	14%	13%	14%
Phishing websites	2%	3%	1%	3%	4%	1%	3%	1%	1%
Breast cancer	83%	82%	83%	84%	86%	83%	84%	83%	83%
Ionosphere	15%	16%	15%	16%	17%	13%	14%	12%	11%
COVID-19	2%	2%	1%	2%	2%	3%	2%	1%	1%

Table 8. Two null-values accuracy within 30 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	15%	14%	15%	17%	15%	15%	15%	13%	14%
Phishing websites	2%	3%	1%	3%	4%	1%	3%	1%	1%
Breast cancer	82%	83%	83%	82%	85%	83%	83%	82%	82%
Ionosphere	15%	16%	15%	15%	18%	16%	15%	13%	11%
COVID-19	1%	1%	2%	2%	2%	2%	2%	1%	2%

Undoubtedly, the problem of null values is one more complex problem for several reasons such as: i) Weakness of datasets because of no real associations among the attributes or features of these datasets; ii) Weakness of some null values associated with the target or other completed attributes/features; iii) Little completed data compared with the size of null values; and iv) The nature of the dataset, for instance, hasn't a strong association or relevance between the features and target, even among the features.

Table 9. Three null value accuracy within 10 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	10%	11%	10%	10%	12%	12%	11%	10%	11%
Phishing websites	1%	1%	1%	2%	2%	1%	2%	1%	1%
Breast cancer	77%	78%	78%	77%	78%	78%	75%	76%	76%
Ionosphere	5%	4%	5%	6%	5%	3%	2%	2%	2%
COVID-19	2%	2%	2%	1%	1%	2%	1%	1%	1%

Table 10. Three null-value accuracy within 20 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	10%	12%	11%	11%	14%	12%	12%	10%	11%
Phishing websites	2%	3%	1%	3%	4%	1%	3%	1%	1%
Breast cancer	78%	78%	78%	77%	79%	78%	76%	77%	76%
Ionosphere	5%	5%	4%	6%	6%	4%	2%	3%	2%
COVID-19	2%	1%	1%	2%	3%	1%	2%	1%	1%

Table 11. Three null-value accuracy within 30 trees

Dataset	Threshold =0.6			Threshold =0.7			Threshold =0.8		
	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6	W=0.4	W=0.5	W=0.6
Connectionist bench	10%	12%	11%	11%	13%	12%	12%	10%	11%
Phishing websites	2%	3%	1%	3%	4%	1%	3%	1%	1%
Breast cancer	76%	78%	77%	76%	77%	78%	76%	76%	76%
Ionosphere	5%	6%	5%	6%	7%	3%	3%	2%	2%
COVID-19	2%	1%	1%	2%	3%	1%	3%	1%	1%

Table 12. Null-value accuracy using random forest and modified random forest

Dataset	No. of null values	Random forest	Modified random forest
Connectionist bench	1	18%	23%
	2	11%	17%
	3	2%	4%
Phishing websites	1	67%	71%
	2	1%	4%
	3	1%	4%
Breast cancer	1	83%	93%
	2	77%	86%
	3	67%	79%
Ionosphere	1	18%	36%
	2	9%	17%
	3	2%	6%
COVID-19	1	15%	11%
	2	4%	2%
	3	1%	3%

No. of tree =20, W=0.5, Threshold=0.7 and (information gain and gini index)

Through the above reasons, some results are unsuitable or don't meet ambition in predicting effects. In this scenario, the most profitable results have been obtained through the number of trees =20, W =0.5, threshold =0.7 and the two feature selection methods (information gain and gini index). The performance of the modified random forest results increased by 9.5%, 6.5% and 5.25% of 1, 2 and 3 null values, respectively. The results depended on average values for the five datasets. Besides, the nature of the dataset plays a significant role in increasing the accuracy of null-values estimation. In addition, one null value imputation gave a good result for all the five datasets, two null values gave less than one null value, and three null values showed minor effects. The breast cancer dataset gave the best results compared with the four others. Connectionist bench, Phishing websites, and Ionosphere datasets gave inadequate effects within two and three null values. While the performance with COVID-19 is not satisfactory.

6. CONCLUDING REMARKS AND FUTURE DIRECTION

The modified random forest algorithm focuses on three modifications to increase the performance of the original one, less redundancy bootstrap, hybrid features selection and ranked voting. These three modifications made the random forest algorithm more efficient by selecting diverse samples using less redundancy bootstrap and more than one feature selection method to enhance the selected features more relevant to the target. Lastly, the voting strategy is based on ranking the trees. Also, these three modifications on the random forest algorithm gave enhanced results compared to the original one. The experimental results for the five datasets showed significant improvement in outcomes by 9.5%, 6.5% and 5.25% for one, two, and three null values, respectively. In the null values imputation problem, increasing the number of missing values decreases the imputation accuracy. Also, the nature of the dataset plays a significant role in the imputation; some dataset does not contain relational relevance in their attributes, which causes poor extracted learned rules. Unfortunately, these inadequate, learned rules don't enough to estimate the missing values. In the future, other machine learning techniques will be applied to solve the situation of null values in the same datasets.

REFERENCES




- [1] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mechanical Systems and Signal Processing*, vol. 138, p. 106587, Apr. 2020, doi: 10.1016/j.ymsp.2019.106587.
- [2] O. A. von Lilienfeld and K. Burke, "Retrospective on a decade of machine learning for chemical discovery," *Nature Communications*, vol. 11, no. 1, 2020, doi: 10.1038/s41467-020-18556-9.
- [3] A. N and A.-T. I, "Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: A review.," *Informatics in medicine unlocked*, vol. 24, p. 100564, 2021, [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/33842685/>.
- [4] A. Khaleel Faieq and M. M. Mijwil, "Prediction of heart diseases utilising support vector machine and artificial neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 374–380, 2022, doi: 10.11591/ijeecs.v26.i1.pp374-380.
- [5] H. Mohammad-Rahimi, M. Nadimi, A. Ghalyanchi-Langeroudi, M. Taheri, and S. Ghafouri-Fard, "Application of machine learning in diagnosis of covid-19 through x-ray and ct images: A scoping review," *Frontiers in Cardiovascular Medicine*, vol. 8, 2021, doi: 10.3389/fcvm.2021.638011.
- [6] K. Aggarwal *et al.*, "Has the future started? the current growth of artificial intelligence, machine learning, and deep learning," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, no. 1, pp. 115–123, Jan. 2022, doi: 10.52866/ijcsm.2022.01.01.013.
- [7] M. van der Schaar *et al.*, "How artificial intelligence and machine learning can help healthcare systems respond to covid-19," *Machine Learning*, vol. 110, no. 1, pp. 1–14, 2021, doi: 10.1007/s10994-020-05928-x.
- [8] A. Al-Janabi, E. A. Al-Zubaidi, and R. H. Abdulzhray Al-Sagheer, "Encapsulation of semantic description with syntactic components for the Arabic language," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 961–967, 2020, doi: 10.11591/ijeecs.v22.i2.pp961-967.
- [9] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Systems with Applications*, vol. 166, p. 114060, Mar. 2021, doi: 10.1016/j.eswa.2020.114060.
- [10] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, 2021, doi: 10.1007/s42979-021-00592-x.
- [11] A. Le Glaz *et al.*, "Machine learning and natural language processing in mental health: Systematic review," *Journal of Medical Internet Research*, vol. 23, no. 5, 2021, doi: 10.2196/15708.
- [12] G. Y. H. Lip, A. Genaidy, G. Tran, P. Marroquin, and C. Estes, "Incident atrial fibrillation and its risk prediction in patients developing covid-19: A machine learning based algorithm approach," *European Journal of Internal Medicine*, vol. 91, pp. 53–58, 2021, doi: 10.1016/j.ejim.2021.04.023.
- [13] S. Wang *et al.*, "Application of machine learning to predict the occurrence of arrhythmia after acute myocardial infarction," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, 2021, doi: 10.1186/s12911-021-01667-8.
- [14] W. L. L. ZQ, and W. A., "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, 2020.
- [15] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of covid-19 in chest X-ray images using DeTraC deep convolutional neural network," *Applied Intelligence*, vol. 51, no. 2, pp. 854–864, 2021, doi: 10.1007/s10489-020-01829-7.
- [16] H. Zhang *et al.*, "A tool to early predict severe corona virus disease 2019 (covid-19): A multicenter study using the risk nomogram in wuhan and guangdong, china.," *Cancer*, vol. 46, no. May, pp. 1–17, 2020, [http://dx.doi.org/10.1016/S1470-2045\(20\)30310-7](http://dx.doi.org/10.1016/S1470-2045(20)30310-7).
- [17] M. M. Mijwil and E. A. Al-Zubaidi, "Medical image classification for coronavirus disease (covid-19) using convolutional neural networks," *Iraqi Journal of Science*, vol. 62, no. 8, pp. 2740–2747, 2021, doi: 10.24996/ijcs.2021.62.8.27.
- [18] Ö. Yıldırım, P. Pławiak, R. S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Computers in Biology and Medicine*, vol. 102, pp. 411–420, 2018, doi: 10.1016/j.combiomed.2018.09.009.
- [19] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction*, vol. 129, p. 103827, Sep. 2021, doi: 10.1016/j.autcon.2021.103827.
- [20] T. Lan, X. Liu, S. Wang, K. Jermsittiparsert, S. T. Alrashood, *et al.*, "An advanced machine learning based energy management of renewable microgrids considering hybrid electric vehicles' charging demand," *Energies*, vol. 14, no. 3, p. 569, 2021.
- [21] M. M. Mijwil, "Iraqi food image detection using convolutional neural network classification method," 2021, pp. 249–257.
- [22] B. Farsi, M. Amayri, N. Bouguila, and U. Eicker, "On short-term load forecasting using machine learning techniques and a novel parallel deep LSTM-CNN approach," *IEEE Access*, vol. 9, pp. 31191–31212, 2021, doi: 10.1109/ACCESS.2021.3060290.
- [23] M. M. Mijwil, "Implementation of machine learning techniques for the classification of lung x-ray images used to detect covid-19 in humans," *Iraqi Journal of Science*, vol. 62, no. 6, pp. 2099–2109, 2021, doi: 10.24996/ijcs.2021.62.6.35.
- [24] K. K. Singh, S. Kumar, P. Dixit, and M. K. Bajpai, "Kalman filter based short term prediction model for covid-19 spread," *Applied Intelligence*, vol. 51, no. 5, pp. 2714–2726, 2021, doi: 10.1007/s10489-020-01948-1.
- [25] S. Hartini, Z. Rustam, G. S. Saragih, and M. J. S. Vargas, "Estimating probability of banking crises using random forest," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, pp. 407–413, 2021, doi: 10.11591/IJALV10.I2.PP407-413.
- [26] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of covid-19 infection from routine blood exams

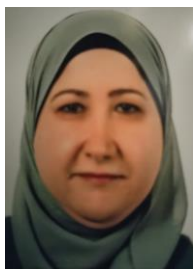
- with machine learning: A feasibility study,” *Journal of Medical Systems*, vol. 44, no. 8, 2020, doi: 10.1007/s10916-020-01597-4.
- [27] D. Liu *et al.*, “Optimisation and evaluation of the random forest model in the efficacy prediction of chemoradiotherapy for advanced cervical cancer based on radiomics signature from high-resolution T2 weighted images,” *Archives of Gynecology and Obstetrics*, vol. 303, no. 3, pp. 811–820, 2021, doi: 10.1007/s00404-020-05908-5.
- [28] L. Xue, Y. Liu, Y. Xiong, Y. Liu, X. Cui, and G. Lei, “A data-driven shale gas production forecasting method based on the multi-objective random forest regression,” *Journal of Petroleum Science and Engineering*, vol. 196, p. 107801, Jan. 2021, doi: 10.1016/j.petrol.2020.107801.
- [29] R. Zhu, Y. Wang, J. X. Liu, and L. Y. Dai, “IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier,” *BMC Bioinformatics*, vol. 22, no. 1, 2021, doi: 10.1186/s12859-021-04104-9.
- [30] X. Xie *et al.*, “Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land,” *Ecological Indicators*, vol. 120, p. 106925, Jan. 2021, doi: 10.1016/j.ecolind.2020.106925.
- [31] G.-F. Fan, M. Yu, S.-Q. Dong, Y.-H. Yeh, and W.-C. Hong, “Forecasting short-term electricity load using hybrid support vector regression with grey catastrophe and random forest modeling,” *Utilities Policy*, vol. 73, p. 101294, Dec. 2021, doi: 10.1016/j.jup.2021.101294.
- [32] L. Jin *et al.*, “A comparative study of evaluating missing value imputation methods in label-free proteomics,” *Scientific Reports*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-81279-4.
- [33] C. Ribeiro and A. A. Freitas, “A data-driven missing value imputation approach for longitudinal datasets,” *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6277–6307, 2021, doi: 10.1007/s10462-021-09963-5.
- [34] R. Wei *et al.*, “Missing value imputation approach for mass spectrometry-based metabolomics data,” *Scientific Reports*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-017-19120-0.
- [35] H. Demirhan and Z. Renwick, “Missing value imputation for short to mid-term horizontal solar irradiance data,” *Applied Energy*, vol. 225, pp. 998–1012, 2018, doi: 10.1016/j.apenergy.2018.05.054.
- [36] A. T. Sadiq, M. G. Duaimi, and S. A. Shaker, “Data missing solution using rough set theory and swarm intelligence,” in *Proceedings - 2012 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2012*, 2012, pp. 173–180, doi: 10.1109/ACSAT.2012.29.
- [37] A. T. Sadiq and S. A. Chawishly, “Intelligent methods to solve null values problem in databases intelligent,” *Journal of Advanced Computer Science and Technology Research*, vol. 2, no. 2, pp. 91–103, 2012.
- [38] B. Ramosaj and M. Pauly, “Predicting missing values: a comparative study on non-parametric approaches for imputation,” *Computational Statistics*, vol. 34, no. 4, pp. 1741–1764, 2019, doi: 10.1007/s00180-019-00900-3.
- [39] Z. W. Salman, K. A. Hussein, and A. T. Sadiq, “Solving null values problem using modified random forest algorithm via meerkat clan algorithm,” *Solid State Technology*, pp. 1118–1130, 2020.
- [40] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, “AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes,” *Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.
- [41] E. C. Gök and M. O. Olgun, “SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples,” *Neural Computing and Applications*, vol. 33, no. 22, pp. 15693–15707, 2021, doi: 10.1007/s00521-021-06189-y.
- [42] Z. Chai and C. Zhao, “Enhanced random forest with concurrent analysis of static and dynamic nodes for industrial fault classification,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 54–66, 2020, doi: 10.1109/TII.2019.2915559.
- [43] T. Chen, X. Yin, L. Peng, J. Rong, J. Yang, and G. Cong, “Monitoring and recognizing enterprise public opinion from high-risk users based on user portrait and random forest algorithm,” *Axioms*, vol. 10, no. 2, 2021, doi: 10.3390/axioms10020106.
- [44] S. J. Buckley, R. J. Harvey, and Z. Shan, “Application of the random forest algorithm to *Streptococcus pyogenes* response regulator allele variation: from machine learning to evolutionary models,” *Scientific Reports*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-91941-6.
- [45] R. V. K. Reddy, S. Subhani, B. S. Rao, and N. L. Anantha, “Machine learning based outlier detection for medical data,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 1, pp. 564–569, 2021, doi: 10.11591/ijeecs.v24.i1.pp564-569.
- [46] B. Xiao and B. Xiao, “A novel approach for internal short circuit prediction of lithium-ion batteries by random forest,” *International Journal of Electrochemical Science*, vol. 16, pp. 1–19, 2021, doi: 10.20964/2021.04.21.
- [47] G. Stefanos *et al.*, “Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling,” *Geocarto International*, vol. 0, no. 0, pp. 1–16, 2019, [Online]. Available: <https://doi.org/10.1080/10106049.2019.1595177>.
- [48] Shamis N. Abd. Mohammad Alsjari, and Hind Raad Ibraheem, “Rao-SVM machine learning algorithm for intrusion detection system,” *Iraqi Journal for Computer Science and Mathematics*, vol. 1, no. 1, pp. 23–27, Jan. 2020, doi: 10.52866/ijcsm.2019.01.01.004.
- [49] Y. E. Almallki *et al.*, “A novel method for covid-19 diagnosis using artificial intelligence in chest x-ray images,” *Healthcare (Switzerland)*, vol. 9, no. 5, 2021, doi: 10.3390/healthcare9050522.
- [50] N. A. Hakimi, M. A. Mohd Razman, and A. P. P. Abdul Majeed, “The classification of covid-19 cases through the employment of transfer learning on x-ray images,” *Mekatronika*, vol. 3, no. 1, pp. 44–51, 2021, doi: 10.15282/mekatronika.v3i1.7151.
- [51] D. Singh, V. Kumar, V. Yadav, and M. Kaur, “Deep neural network-based screening model for covid-19-infected patients using chest x-ray images,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 3, 2021, doi: 10.1142/S0218001421510046.
- [52] S. H. Kassania, P. H. Kassanib, M. J. Wesolowski, K. A. Schneidera, and R. Detersa, “Automatic detection of coronavirus disease (covid-19) in x-ray and ct images: A machine learning based approach,” *Biocybernetics and Biomedical Engineering*, vol. 41, no. 3, pp. 867–879, 2021, doi: 10.1016/j.bbe.2021.05.013.
- [53] L. Peng, L. Wang, X. Y. Ai, and Y. R. Zeng, “Forecasting tourist arrivals via random forest and long short-term memory,” *Cognitive Computation*, vol. 13, no. 1, pp. 125–138, 2021, doi: 10.1007/s12559-020-09747-z.
- [54] R. P. E. C. A. Subasini, A. V. Katharine, V. Kumaresan, S. G. Kumar, and T. M. Nithya, “Cardiovascular disease prediction using machine learning algorithms,” *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 2, pp. 904–912, 2021, doi: 10.17762/turcomat.v12i6.2426.
- [55] M. K. Yusof, W. M. A. F. W. Hamzah, and N. S. M. Rusli, “Efficiency of hybrid algorithm for covid-19 online screening test based on its symptoms,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 440–449, 2022, doi: 10.11591/ijeecs.v25.i1.pp440-449.
- [56] C.-F. Yen, H.-Y. Hsieh, K.-W. Su, M.-C. Yu, and J.-S. Leu, “Solar power prediction via support vector machine and random forest,” *E3S Web of Conferences*, vol. 69, p. 01004, Nov. 2018, doi: 10.1051/e3sconf/20186901004.
- [57] E. Antoniou *et al.*, “EEG-based eye movement recognition using the brain-computer interface and random forests,” *Sensors*, vol. 21, no. 7, 2021, doi: 10.3390/s21072339.




- [58] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for random forests," *Machine Learning with Applications*, vol. 6, p. 100094, 2021, doi: 10.1016/j.mlwa.2021.100094.
- [59] A. T. Sadiq and K. S. Musawi, "Modify random forest algorithm using hybrid feature selection method," *International Journal on Perceptive and Cognitive Computing*, vol. 4, no. 2, pp. 1–6, Dec. 2018, doi: 10.31436/ijpc.v4i2.59.
- [60] A. Fornaser, M. De Cecco, P. Bosetti, T. Mizumoto, and K. Yasumoto, "Sigma-z random forest, classification and confidence," *Measurement Science and Technology*, vol. 30, no. 2, 2019, doi: 10.1088/1361-6501/aaf466.
- [61] K. S. Mohsen and A. T. Sadiq, "Random forest algorithm using accuracy-based ranking," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 3, pp. 1039–1045, 2019, doi: 10.1166/jctn.2019.7995.
- [62] A. Y. Hussein, P. Falcarin, and A. T. Sadiq, "Enhancement performance of random forest algorithm via one hot encoding for IoT IDS," *Periodicals of Engineering and Natural Sciences*, vol. 9, no. 3, pp. 579–591, 2021, doi: 10.21533/pen.v9i3.2204.
- [63] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks*, vol. 1, no. 1, pp. 75–89, 1988, doi: 10.1016/0893-6080(88)90023-8.
- [64] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Computing and Applications*, vol. 25, no. 2, pp. 443–458, 2014, doi: 10.1007/s00521-013-1490-z.
- [65] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computer-derived nuclear features distinguish malignant from benign breast cytology," *Human Pathology*, vol. 26, no. 7, pp. 792–796, 1995, doi: 10.1016/0046-8177(95)90229-5.
- [66] Y. Jiang and Z. Zhou, "NeC4.5: neural ensemble based C4.5," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 770–773, 2004.

BIOGRAPHIES OF AUTHORS






Maad M. Mijwil    is an Iraqi Academician; he is born in Baghdad, Iraq, in 1987. He received his B.Sc. degree in software engineering from Baghdad college of economic sciences university, Iraq, in 2009. He received a M.Sc. degree in 2015 from the computer science department, university of Baghdad in the field of wireless sensor networks, Iraq. Currently, he is working as a lecturer and an academic member of staff in the computer techniques engineering department at Baghdad College of Economic Sciences University, Iraq. He can be contacted at email: mr.maad.alnaimiy@baghdadcollege.edu.iq.






Alaa Wagih Abdulqader    is an Iraqi Academician; she was born in Basrah, Iraq, in 1969. She received her B.Sc. degree in computer sciences from Baghdad College of Economic Sciences University, Iraq, in 2009. She received a M.Sc. degree in 2012 from the *Iraqi Commission for Computers and Informatics / Informatics Institute* for Post Graduate Studies in the field of security and ANN, Iraq. Currently, she is working as a Lecturer and an academic member of staff in the computer techniques engineering department at Baghdad College of Economic Sciences University, Iraq. She has over ten years of experience in teaching and guiding projects for undergraduates. She can be contacted at email: alaa_wagih@baghdadcollege.edu.iq.



Sura Mazin Ali    received her B.Sc. degree in software Engineering in 2001 from the AlRafidein college University in Baghdad, Iraq. Her M.Sc. degree in Computer science from ICCI in 2014. Her current research interests include data security and artificial intelligence applications. She can be contacted at email: Suraaz2007@uomustansiriyah.edu.iq.



Ahmed T. Sadiq    received a B.Sc., M.Sc. & Ph. D. degree in Computer Science from the University of Technology, Computer Science Department, Iraq, 1993, 1996 & 2000 respectively. He is Professor in A.I. since 2014. His research interests in artificial intelligence, data security, patterns recognition and data mining. He can be contacted at email: Ahmed.t.sadiq@uotechnology.edu.iq.