A Novel Hybrid Approach for Classification Problem Case Study: Heart Disease Classification

Ahmed Umer Khawaja(✉) and Yeh Ching Low

Department of Computing and Information Systems, School of Engineering and Technology, Sunway University, Petaling Jaya, Malaysia
khawajaahmedumer@gmail.com

Abstract. Heart disease is a major cause of death globally, with patients succumbing to death a few years of being diagnosed. This paper proposed a novel hybrid approach of Cuckoo Search Optimization – Extreme Learning Machine (CSO - ELM) to solve a classification problem. The approach was compared with established models proven in classifying heart disease. The CSO-ELM indicated significant predictive ability and outperformed the established and base models in machine learning.

Keywords: ELM-CSO · Extreme Learning Machine · Cuckoo Search Optimization · Heart Disease

1 Introduction

Heart disease is one of the leading causes of death in the present era and is expected to rise sharply in the future [1]. There are number of factors that have contributed to the significant rise, such as, unhealthy eating habits and lifestyle. It has been identified that early diagnosis of heart disease is crucial for reducing the mortality rate, as heart disease patient die within a few years of being diagnosed [2]. The diagnostic process of heart disease is known to be a challenging task, which includes multiple tests, generating abundance of data for the doctors to analyze. Due to this factor and doctors having to analyze pools of data daily, can lead to major error and fatigue [3]. To help reduce the mortality rate and increase the efficiency of data analytics, numerous methods have been employed, recently, the application of machine learning methods has gained popularity in various medical fields, such as, breast cancer prediction [4, 5], coronary heart disease prediction [2], Bupa, Diabetes and Hepatitis [5]. Researchers have also applied these machine learning methods in the different types of heart disease prediction and have acquired satisfactory results [3]. Although, there has been significant growth in this area of research, there is still room for improvement and application of newer and robust methods. Thus, in this paper, we present a novel Cuckoo Search Optimizer (CSO) – Extreme Learning Machine (ELM) hybrid model for heart disease identification.

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 413–423, 2022.
https://doi.org/10.2991/978-94-6463-094-7_32

2 Related Work

Several research studies have been carried out to predict various types of heart disease. Padmaja et al. [6] employed various Machine Learning (ML) algorithms, such as Random Forest (RF) Classifier, Logistic Regression (LR) Classifier, K-Nearest Neighbor (KNN) Classifier, Support Vector Machine (SVM) classifier, Decision Tree classifier, Naïve Bayes, and Gradient Boosting to accurately predict heart disease. Owusu et al. [7] proposed boosting Support Vector Machine classifier for heart disease risk prediction. The results highlighted that boosting SVM outperformed Naïve Bayes, Logistic Regression and Multilayer perceptron. Suresh et al. [8] also conducted a study by utilizing the SVM model with an iterative feature selection technique. It was discovered that the performance of SVM model is boosted when accompanied with the correct features and hyperparameters. Ansarullah et al. [9] used data mining techniques for cardiovascular disease detection. Dewangan et al. [4] introduced a novel Back Propagation Boosting Recurrent Wienmed model (BPBRW) with Hybrid Krill Herd African Buffalo Optimization (HKH-ABO) for detecting breast cancer in an earlier stage using breast MRI images. The results indicated that the proposed method acquired precision of 99.9% compared with other models. The authors attributed the success of the model to its dual hybridization approach. Ren et al. [10] conducted a novel study to predict the risk of kidney disease based on a patient having hypertension. Some research studies proposed cost effective ML methods to predict heart disease, the vision being to help reduce the cost incurred by the patients. Isinkaye et al. [11] proposed a mobile based neuro fuzzy model for cardiovascular disease diagnosis. It was noted that although the proposed method was cost effective, the results were inadequate. An ensemble heart disease prediction model was proposed by [1], the technique was a combination of Relief algorithm and ensemble classifiers, the results indicated that the method generated high accuracy performance. Dutta et al. [2] recommended a novel technique to improve heart disease prediction accuracy in imbalanced data by employing a using LASSO algorithm and shallow convolutional layers. Similarly, [12] also employed the LASSO algorithm for feature selection and combining the results with multiple machine learning and deep learning (DL) models. Kumar et al. [13] introduced a novel improved algorithm (S-CDF) for neural network weights optimization, novel method had significant improvement in predicting heart disease compared to established methods. Baccouche et al. [14] used ensemble deep learning methods for heart disease classification, the methodology was based on training two independent deep learning models on similar data and the results were a combination of them. Likewise, an incremental feed forward neural network method was introduced by [15]. Waghulde and Patil [16] gauged the performance of a genetic neural network (GNN) for heart disease classification, the results showed that an optimized GNN can achieve an accuracy of 98%. Sudha [17] developed an application-specific integrated circuit (ASIC) based on backpropagation neural network (BPNN).

3 Methodology

3.1 Data

The dataset used in this research study has been obtained from University of California Irvine (Kaggle) repository [18]. The dataset includes 918 observations, which are the combination of 5 different sources, namely, Cleveland: 303 observations, Hungarian: 294 observations, Switzerland 123 observations, Long Beach VA 200 observations and Stalong (heart) dataset: 270 observations. The total number of observations add up to 1190 as shown in Table 1. However, 272 observations were removed due to identical instances. 11 distinct features are included in the dataset as shown in the Table 2, the labels include 0- for no heart disease and 1- for heart disease.

Table 1. Heart disease data instance

Source	Observations
Cleveland	303
Hungarian	294
Switzerland	123
Long Beach VA	200
Stalong	270
Total	**1190**

Table 2. List of features and descriptions

Features	Description
Age	Age of patient
Sex	Male (M)/female (F)
ChestPainType	Chest pain type includes categories
RestingBP	Resting blood pressure, varies between 80 and 200
Cholesterol	Serum Cholesterol, varies between 0 and 425
FastingBS	Fasting blood Sugar, varies between 0.0 and 1.0
RestingECG	Resting electrocardiogram results includes 3 categories
MaxHR	Maximum heart rate achieved, varies between 60 and 205
ExerciseAngina	Exercise induced angina, true or false
Oldpeak	Exercise relative to rest, varies between −2.6 and 6.2
ST_Slope	Slope of the peak exercise ST segment, 3 categories

3.2 Data Pre-processing and Feature Selection

The dataset was analysed thoroughly to filter out any missing or duplicate values. Label encoding was implemented to the categorical features in the dataset. The method assigns a distinct number to each of the categorical feature, as some ML models are unable to interpret categorical values. The data was then split into 2 sections, 80% of the data was used for model training and 20% data was used for testing purposes. Furthermore, acknowledging that two benchmark models used in this study, originally had feature selection techniques applied to them. This study conducts the method in the same manner, three different data frames were created of the same data.

1. Data frame with no feature selection.
2. Data frame with recursive feature elimination (RFE) [8] applied.
3. Data frame with extra tree classifier algorithm [7] applied.

3.2.1 Feature Selection for SVM

Iterative feature selection was employed using recursive feature elimination (RFE) [8], the results showed that ST_Slope, Oldpeak, MaxHR, Cholesterol and ChestPainType, followed by Age and ExerciseAngina had the most influence on the outcome label. The featured that were less significant were removed and the data was recompiled to include only influential features.

3.2.2 Feature Selection for Boosted SVM

Feature selection was conducted using the extra tree classifier (ETC) algorithm [7]. The classifier uses the Gini impurity methodology to find the most significant features. The results are illustrated in Fig. 1. According to the results ST_Slope had the most significant impact on the outcome label, followed by ExerciseAngina, ChestPainType, Oldpeak, Cholesterol, MaxHR and Age. Compared with feature selection using RFE, results of ETC regards ExerciseAngina as the second most significant feature.

3.2.3 For Proposed Model (CSO-ELM) and ANN

Data for was pre-processed by the means of one hot encoding, this method turns all the categorical features into numeric arrays, which can be interpreted by neural networks. The data was further processed using the standard scaler, this can aid models that use gradient descent to converge faster.

3.3 ML Models

3.3.1 Extreme Learning Machine

Extreme learning machine is a relatively new learning algorithm developed by [19]. The algorithm aims to reduce the computational time incurred by gradient descent and back-propagation base algorithms. It has been proven by [19] to outperform SVM and back-propagation algorithms. Over the years, the algorithm has been optimized and applied

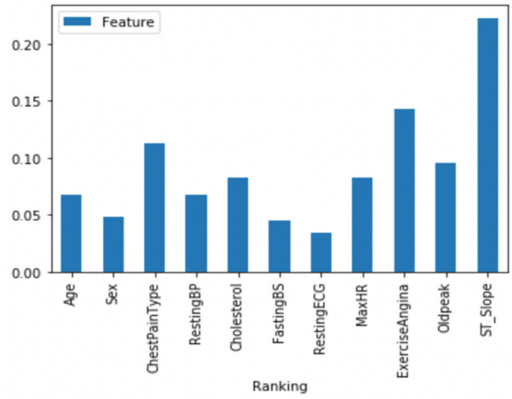

Fig. 1. Ranking of features using extra tree classifier

in various research area ranging from flood forecasting [20] to modulation classification [21]. The algorithm operates by assigning arbitrary values for input weights and biases and then calculates the output weights using the smallest norm least square method.

For N hidden node, the equation can be written as

$$f_N(X_i) = \sum_{i=1}^{N} \alpha_i g(w_i, b_i, x_i) = y_j \tag{1}$$

where g is the activation function.

In an ideal case the above equation as:

$$H\beta = T \tag{2}$$

Equation 2 can be represented by

$$H(w_i, w_2, \ldots w_N, b_1, b_2, \ldots b_N, x_1, x_2, \ldots x_N) = \left[g(w_i, x, b_i) \ldots g(w_N, x_N b_N)\right] \tag{3}$$

If all the values for w_i and b_i are set using arbitrary values, then an optimization problem for a non-ideal case can be represented by

$$||H\beta^* - T|| = \min_{\beta} ||H\beta^* - T|| \tag{4}$$

where β^* can be solved using the smallest norm least square equation

$$\beta^* = H^+ T \tag{5}$$

Here H^+ is the Moore-Penrose Generalized Inverse of H.

Considering the above equation, an optimization problem can be created that can determine the optimal number of neurons N.

3.3.2 Cuckoo Search Optimization

Cuckoo search optimization algorithm is a newly developed meta-heuristic approach [22]. The algorithm is inspired by the natural behavior of the cuckoo bird species. Cuckoo bird is known for its parasitic breeding behavior, the cuckoo bird does not make its own nest and chooses a host bird's nest to lay its egg. This approach has a risk of the host bird discovering the cuckoo's egg, upon discovery, the host bird can choose to throw the cuckoo's egg or build a new nest entirely. Several set of rules are to be followed by employing this algorithm:

1. Each cuckoo lays one egg at a time, and places it in a randomly chosen nest.
2. The best nests with the highest-quality eggs (solutions) will be carried over to the next generations.
3. The number of available host nests is fixed, and the cuckoo's egg is discovered by the host bird with the probability $p_a \in [0, 1]$. If the alien egg is discovered, the nest is abandoned, and a new nest is built in a new location.

To search for new solutions $x_i^{(t+1)}$, for cuckoo i, the levy flight method is utilized, such that:

$$x_i^{(t+1)} = x_i^t(\alpha \oplus Levy(\lambda)) \tag{6}$$

where $\alpha > 0$, is the step size and $\oplus$ denotes entry-wise multiplication. The levy flight is similar to the random walk method with its random steps being taken from a levy distribution of large steps.

3.3.3 CSO-ELM

Previous studies have used different approaches to optimize the input weights and biases of ELM by using different optimization algorithm. Liu et al. [23] used particle swarm optimization (PSO) to optimize the input-hidden weights layer and biases. Chen et al. [20] employed the backtracking search optimization to find the optimal input-hidden layer weights and biases. However, by optimizing the input-hidden layer wights and biases defeats the purpose of using ELM algorithm. As this tends to be a similar approach as using a backpropagation learning method [24]. Finding the optimal number of hidden neurons is still an unresolved issue [5]. Hence in this paper, we propose to utilize the CSO algorithm (Eq. 6) to find the optimal value of N in Eq. 3. The algorithm is repeated for a fixed number of terms until convergences to a global minimum,

Figure 2 illustrates the flowchart of proposed methodology.

3.4 Performance Metrics

To test the performance of the proposed model, several evaluation metrics are employed to measure the accuracy, recall, precision of the model. The values are then compared with benchmark models, namely, SVM, boosted SVM, ANN, a base ELM model and a hybrid ELM-PSO model.

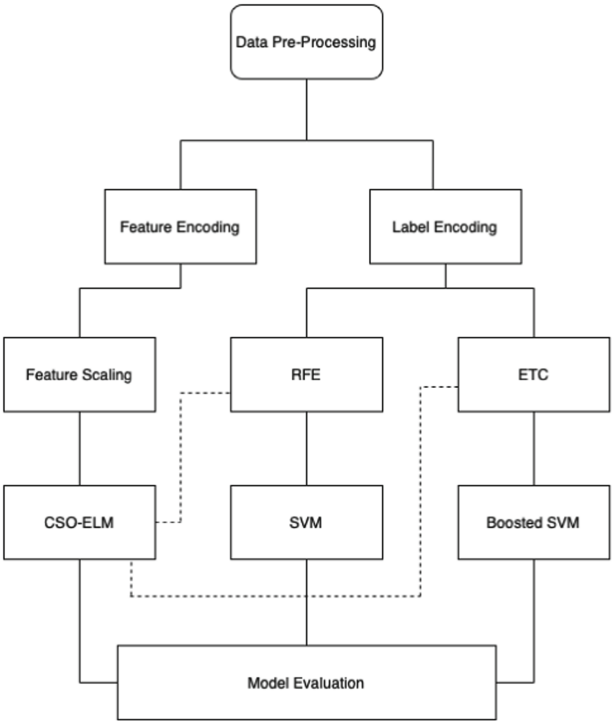

Fig. 2. Flowchart for methodology

3.4.1 Accuracy

Accuracy score represents the number of predictions a model was able to predict correctly over the total number of predictions.

$$\frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ predictions} \tag{7}$$

3.4.2 Recall

Recall is intuitively the ability of the classifier to find all the positive samples. It is the ratio of true positives divided by the combination of true positives and false negatives.

$$\frac{true\ positive}{(true\ positive\ +\ false\ negatives)} \tag{8}$$

3.4.3 Precision

Precision is intuitively the ability of the classifier not to label as positive a sample that is negative. It is the ratio of true positives divided by the number true positives and false positives.

$$\frac{true\ positives}{(true\ positives\ +\ false\ positives)} \tag{9}$$

4 Results and Discussion

Table 3 illustrates the accuracy, recall and precision score of proposed CSO-ELM model compared with base ELM, ANN, RFE-SVM, ETC – boosted SVM and PSO-ELM.

The results indicate that the proposed model outperforms all other models with an accuracy of 85.5% in classifying heart disease. The optimal number of neurons discovered by the CSO algorithm was 15. The number of neurons set for the base ELM was 11, following the standard convention of choosing the hidden layer neurons of a neural network. Base ELM attained the second highest accuracy of 83% followed by ANN. The result upholds [19] claim that ELM outperforms ANN in terms of generalization performance and learning speed. The ANN was trained with 50 epochs and Stochastic Gradient Descent (SGD) was used as the learning algorithm. SVM with tuned gamma and C value outperforms the boosted SVM by 1.7%, both methods were tested with their respective feature selection technique. These models were able to attain high accuracy in the original study, However, performed worse in terms of accuracy on this classification problem. In terms of recall, SVM and Boosted SVM obtain highest score, demonstrating that the models were able to predict all positive prediction correctly. Followed by the proposed model, ANN, and base ELM. The proposed model was also able to obtain the highest precision score. It should be noted that SVM and boosted SVM had the worst performance for precision score, indicating that the model had labelled most of the occurrences as a positive prediction. Hence, attaining a perfect score for Recall.

Table 4 illustrates the results of proposed CSO-ELM model compared with boosted SVM. The ETC method was used for feature selection and insignificant features namely, "Sex", "FastingBS" and "RestingECG" were discarded.

Proposed method was compared with the boosted SVM model using the same features extracted by the ETC technique, to acquire in-depth understanding of the model's performance. The results indicate that the proposed model outperforms the boosted SVM on the same data and amount of feature variables.

Table 3. Comparative results of various models for different metrics

Metric	ELM	ANN	RFE - SVM	ETC - boosted SVM	PSO-ELM	CSO-ELM
Accuracy (%)	83.0	81.5	59.9	58.2	84.9	85.8
Recall	0.86	0.87	1.0	1.0	86.8	0.88
Precision	0.85	0.82	0.58	0.58	86.8	0.86

Table 4. Comparative results of ETC - boosted SVM and ETC- CSO - ELM for different metrics

Metric	ELM-CSO	Boosted SVM
Accuracy (%)	84.0	58.2
Recall	0.84	1.0
Precision	0.87	0.58

Table 5. Comparative results of RFE - SVM and RFE - CSO-ELM for different metrics

Metric	ELM-CSO	SVM
Accuracy (%)	81.0	59.9
Recall	0.83	1.0
Precision	0.84	0.58

Table 5 illustrates the accuracy score of the proposed CSO-ELM model compared with SVM. The RFE method was used for feature selection and insignificant features namely, "Sex", "FastingBS", "RestingECG and "ExerciseAngina" were discarded.

Proposed method was compared with the SVM model using the same features extracted by the RFE algorithm, to acquire in-depth understanding of the model's performance. The results indicate that the proposed model outperforms the SVM on the same data and amount of feature variables.

5 Conclusion

The paper demonstrated a novel method to find the optimal number of hidden layer neurons for ELM classifier by using the CSO algorithm. The method was applied to classification problem and attained promising results on heart disease classification. One of the shortcomings encountered during this study was the ELM structure's random assignment of weights, lead to diverse range of weights, this had significant impact on the model's performance. Future studies may focus on developing a hybrid ELM model that can focus optimizing both weights and hidden layer neurons.

Acknowledgments. The authors acknowledge the dataset provided by UC Irvine Machine learning repository and [18] for providing the heart disease data used in this case study. The authors wish to thank the reviewers for their useful suggestions. The authors are supported by the Ministry of Higher Education grant FRGS/1/2020/STG06/SYUC/02/1.

Authors' Contributions. The first author conducted the literature review, design and implementation of the proposed model. The second author contributed to the model design and overall research framework.

References

1. Q. Zhenya & Z. Zhang, A hybrid cost-sensitive ensemble for heart disease prediction. BMC Medical Informatics and Decision Making, **21** (2021) 1–19. https://doi.org/10.1186/s12911-021-01436-7.
2. A. Dutta, T. Batabyal, M. Basu, & S. T. Acton, An efficient convolutional neural network for coronary heart disease prediction. Expert Systems with Applications, **159** (2020). https://doi.org/10.1016/j.eswa.2020.113408.
3. Y. Solanki, A Survey on Risk Assessments of Heart Attack Using Data Mining Approaches. International Journal of Information Engineering and Electronic Business, **11** (2019) 43–51. https://doi.org/10.5815/ijieeb.2019.04.05.
4. K. K. Dewangan, D. K. Dewangan, S. P. Sahu, & R. Janghel, Breast cancer diagnosis in an early stage using novel deep learning with hybrid optimization technique. Multimedia Tools and Applications, (2022). https://doi.org/10.1007/s11042-022-12385-2.
5. P. Mohapatra, S. Chakravarty, & P. K. Dash, An improved cuckoo search based extreme learning machine for medical data classification. Swarm and Evolutionary Computation, **24** (2015) 25–49. https://doi.org/10.1016/j.swevo.2015.05.003.
6. B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, & E. Krishna Rao Patro, Early and Accurate Prediction of Heart Disease Using Machine Learning Model. Turkish Journal of Computer and Mathematics Education 4516 Research Article, **12** (2021) 4516–4528.
7. E. Owusu, P. Boakye-Sekyerehene, J. K. Appati, & J. Y. Ludu, Computer-Aided Diagnostics of Heart Disease Risk Prediction Using Boosting Support Vector Machine. Computational Intelligence and Neuroscience, **2021** (2021). https://doi.org/10.1155/2021/3152618.
8. T. Suresh, T. A. Assegie, S. Rajkumar, & N. K. Kumar, A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model. International Journal of Electrical and Computer Engineering, **12** (2022) 1831–1838. https://doi.org/10.11591/ijece.v12i2.pp1831-1838.
9. S. I. Ansarullah, S. M. Saif, P. Kumar, & M. M. Kirmani, Significance of Visible Non-Invasive Risk Attributes for the Initial Prediction of Heart Disease Using Different Machine Learning Techniques. Computational Intelligence and Neuroscience, **2022** (2022). https://doi.org/10.1155/2022/9580896.
10. Y. Ren, H. Fei, X. Liang, D. Ji, & M. Cheng, A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. BMC Medical Informatics and Decision Making, **19** (2019). https://doi.org/10.1186/s12911-019-0765-4.
11. F. O. Isinkaye, J. Soyemi, & O. P. Oluwafemi, A Mobile-based Neuro-fuzzy System for Diagnosing and Treating Cardiovascular Diseases. International Journal of Information Engineering and Electronic Business, **9** (2017) 19–26. https://doi.org/10.5815/ijieeb.2017.06.03.
12. R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, & P. Singh, Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience, **2021** (2021). https://doi.org/10.1155/2021/8387680.
13. Kumar, P. R., Ravichandran, S., & Narayana, S. (2021). Ensemble classification technique for heart disease prediction with meta-heuristic-enabled training system. Bio-Algorithms and Med-Systems, 17(2), 119–136. https://doi.org/10.1515/bams-2020-0033
14. A. Baccouche, B. Garcia-Zapirain, C. C. Olea, & A. Elmaghraby, Ensemble deep learning models for heart disease classification: A case study from Mexico. Information (Switzerland), **11** (2020) 1–29. https://doi.org/10.3390/INFO11040207.
15. S. Elyassami & A. A. Kaddour, Implementation of an incremental deep learning model for survival prediction of cardiovascular patients. IAES International Journal of Artificial Intelligence, **10** (2021) 101–109. https://doi.org/10.11591/ijai.v10.i1.pp101-109.

16. N. P. Waghulde & N. P. Patil, Genetic Neural Approach for Heart Disease Prediction. International Journal of Advanced Computer Research, **4** (2014) 778–784.
17. M. Sudha, Evolutionary and Neural Computing Based Decision Support System for Disease Diagnosis from Clinical Data Sets in Medical Practice. Journal of Medical Systems, **41** (2017). https://doi.org/10.1007/s10916-017-0823-3.
18. Fedesoriano, Heart failure prediction dataset, Kaggle (2021). https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction (accessed April 10, 2022).
19. G. Bin Huang, Q. Y. Zhu, & C. K. Siew, Extreme learning machine: A new learning scheme of feedforward neural networks. IEEE Int. Conf. Neural Networks - Conf. Proc. (2004), pp. 985–990. https://doi.org/10.1109/IJCNN.2004.1380068.
20. L. Chen, N. Sun, C. Zhou, J. Zhou, Y. Zhou, J. Zhang, & Q. Zhou, Flood forecasting based on an improved extreme learning machine model combined with the backtracking search optimization algorithm. Water (Switzerland), **10** (2018). https://doi.org/10.3390/w10101362.
21. S. I. H. Shah, S. Alam, S. A. Ghauri, A. Hussain, & F. A. Ansari, A Novel Hybrid Cuckoo Search-Extreme Learning Machine Approach for Modulation Classification. IEEE Access, **7** (2019) 90525–90537. https://doi.org/10.1109/ACCESS.2019.2926615.
22. X.-S. Yang, S. Deb, Cuckoo search: Recent advances and applications, Neural Computing and Applications. 24 (2013) 169–174. https://doi.org/10.1007/s00521-013-1367-1.
23. T. Liu, Y. Ding, X. Cai, Y. Zhu and X. Zhang, Extreme learning machine based on particle swarm optimization for estimation of reference evapotranspiration, 36th Chinese Control Conference (CCC), 2017, pp. 4567–4572, https://doi.org/10.23919/ChiCC.2017.8028076.
24. S. Anupam & P. Pani, Flood forecasting using a hybrid extreme learning machine-particle swarm optimization algorithm (ELM-PSO) model. Modeling Earth Systems and Environment, **6** (2020) 341–347. https://doi.org/10.1007/s40808-019-00682-z.