Breast Cancer Classification using Big Data Tools

Kotha Venkata Naga Harischandra Prasad
Department of Computer Science and Engineering
Vignan's Foundation For Science,
Technology and Research Guntur, India

Gnanadeep Settykara
Department of Computer Science and Engineering
Vignan's Foundation for Science,
Technology and Research Guntur, India

Abstract— **Breast cancer is the second most cause of death of women worldwide and is one of the most deadly diseases being faced by women. Breast cancer is difficult to detect, and women tend to hide the problems and early stages of breast cancer which leads to the increase, and thus a lot of death occurs due to this disease. Therefore, many breast cancer awareness programs are being organized across the world so that women would open up on their problems of breast cancer detection. Therefore, this experiment has been based on detecting breast cancer with the help of big data tools like transfer learning and employing the machine learning algorithm, Convolutional neural networks to aid in this process. A practical model has been created using these two technologies, discussed here in this research paper.**

Keywords—CNN, breast cancer, Benign, Malignant, Tumor, Transfer learning.

I. INTRODUCTION

Cancer is a disease that is considered one of the most life-threatening and most dangerous illnesses among various deadly diseases. Cancer can be originated in several parts of the body, such as throat cancer, neck cancer, breast cancer in women, blood cancer, and there are several other types of cancer. In this research paper, we are going to focus our attention mostly on breast cancer identification and detection techniques, so we will discuss breast cancer a bit [1]. Then we will also discuss our approach on how we are moving on to approach the techniques for detection of cancer at a very early stage so that we can help the patients suffering from breast cancer with a model that has an increased accuracy for prediction. The most crucial point to note is that there is no cure for treating cancer patients if the cancer is detected at a later stage. It is also important to note that breast cancer is the second most leading cause responsible for deaths of women all across the world.

Breast cancer is a form of cancer that develops from breast tissues and can be detected by looking at the size and development of breasts. The signs for the detection of breast cancer include a change in the size of the breasts, appearance of a lump in the breasts fluid coming out from nipples, presence of red or scaly patch on the skin of the patient, dimpling of patients skin in the breast area, and appearance of inverted nipple [2]. Therefore, it is complicated to detect breast cancer because it appears in the private parts of women's bodies that almost every woman tends to hide from the world, and therefore it becomes more deadly. However, if breast cancer is detected successfully at an early stage, then the patient's chances of death reduce in a very significant manner. Therefore, it is necessary to see this deadly disease using essential statistical and mathematical tools such as automation or machine learning algorithms to detect this disease.

There are various risk factors or reasons for cancer development in breasts. The most significant risk factor of breast cancer is being a female due to breasts in females. Other risk factors include the lack of involvement of physical exercise in daily lives. Obesity is also a significant risk factor for breast cancer, including alcohol intake. Moreover, diagnosis with hormone replacement therapy during menopause, first menstruation at an early age, having no children or having them at a later stage of life, or a family history involving breast cancer [3]. There are about 5 to 10% of cases passed down genetically from parents to their children in case of breast cancer.

The origin of breast cancer or its development has been most commonly observed in the cells lining the milk ducts and the lobules used to supply these ducts with milk. The type of breast cancer which develops from the milk ducts is also termed ductal carcinomas, whereas the breast cancer originating from the lobules of the milk ducts is termed lobular carcinomas. Various other 18 types of breast cancer are based on their origin and effects. Breast cancer can be diagnosed with the help of a biopsy of the concerned tissue. Following the breast cancer diagnosis, further tests are performed, which helps determine the area in which cancer has been affected and helps identify if the breast cancer or the cancer cells have been spreading beyond the breast section of the body [4]. These tests also help in knowing the kind of treatment that needs to be provided to the patients to cure it as early as possible.

Breast cancer can be treated if it is detected at an early stage with the help of several treatment plans which has been deduced in these previous years. Some of the primary treatments developed and famous for application are surgery, chemotherapy, targeted therapy, radiation therapy, and hormonal therapy. The process of surgery and chemotherapy is the most used method for treating breast cancer. The survival rate of the patients suffering breast cancers from the time cancer starts to develop is five years as per the 85% of patients of US and UK [5]. Death frequency as per the data of 2020 due to breast cancer is 685000 according to Wikipedia, 2.2 million patients have been diagnosed by 2020 with breast cancer been recorded. Thus, it becomes necessary that we do our best to find a cure for breast cancer.

As per recent times, we have a lot of scientific data related to various diseases, which helps us solve and detect or predict several conditions, which is very difficult to process manually [6]. Therefore, with the help of this research experiment, we try to deduce a method for predicting and detecting breast cancer at an early stage in women. We will use the technologies and the advantages of the huge amount of data stored for research all these years and given the name of Big Data along with the help of a deep learning algorithm to detect the symptoms or the tumor at an early stage. This massive amount of data can also be used to keep an eye on or monitor the patients' health. The cancer of breast cancer can be classified as benign or malignant. The benign tumor means that there is no danger from breast cancer; however, a malignant tumor proves to be the presence of breast cancer in the patient's body. Big data and machine learning have been used to research patients' data to detect breast cancer and several other diseases, which is difficult and very time-consuming when done manually by doctors or an expert. Thus, it provides us with an advantage in this field.

In this research, we will build a model for the detection of breast cancer using the tools of big data and machine learning algorithms or the deep learning algorithm to predict the presence of a tumor and detect the presence of breast cancer in the concerned patients. We will use the transfer learning technology of big data and combine it with the convolutional neural networks to build this model for the detection of breast cancer. In this research experiment, three machine learning algorithms are employed: logistic regression, random forest classifier, and gradient boosting classification algorithms. This experiment is performed to provide an added advantage to the patients who have breast cancer and is very helpful for the early detection of breast cancer. Thus, this model will be developed, and the application and results will be discussed later. The following section in the research shows some of the previous research that has effectively solved the prediction problem of breast cancer.

II. LITERATURE REVIEW

Since breast cancer is a very deadly disease that must be worked upon, a cure must be found to resolve this disease so that the women facing this problem might lead a successful life. Therefore, the medical research team has worked on a cure for this disease. However, it is not enough, and various researchers in machine learning, data science, big data, IoT, etc., have been working to provide aid with these technologies. Big data is an emerging field that uses the data from previous patients and helps in the detection and recovery of breast cancer using a massive resource of data. As we are going to implement this research with the help of machine learning and big data, we must discuss some of the research done in this field which will help us in a practical approach.

To move forward with the prediction of breast cancer, big data is an incredible field that will help us effectively. Therefore, extensive data mining is performed to predict breast cancer in this research. The prediction model has been built with the help of a hybrid model using the decision tree and SVM model [7]. The strategy to make a prediction model in this research consists of two parts. The first step is the extraction of option and information treatment, and the second step is deploying the hybrid model using decision trees and support vector machines for prediction. The dataset used is the data available for breast cancer from Wisconsin dataset from the UCI machine learning dataset library. Three classification techniques or models are being compared using the Weka software. The results show that the decision tree–support vector machine hybrid learning algorithm has greater prediction accuracy than minimal sequential optimization, instance-based learning, and naïve Bayes classifier algorithm. This model classifies malignant tumors from benign cancer tumors. There are 458 benign and 241 cancer tumors in the dataset. The accuracy obtained by this model is 91%, the error rate obtained is 2.58%, and the correctly classified instances are 459 with several incorrect ones like 240.

This research presents a model based on the scalability of machine learning algorithms to solve the problem of prediction of breast cancer with context to extensive data analysis. In this analysis, two varieties of data are being analyzed, the gene expression and the methylation of DNA. In this research, classification is performed on the datasets individually with the help of scaling the machine learning algorithms [8]. To aid this research, the Apache Spark platform is being used, and the machine learning algorithms used for this purpose are decision trees, random forest, and support vector machines. These three algorithms are used to create nine models, with the help of these two datasets used individually and then after combining both of these datasets. This research helps in knowing which type of model is best to be employed for this problem in terms of error rate and accuracy. The two big data platforms, Spark and Weka, are also being compared here. The results obtained through this comparison show that the scaled support vector machine employed in the Apache Spark platform performs better than the other classifiers and is observed to attain the highest accuracy with lost error rate when used in the gene expression dataset.

This research paper presents a model for predicting breast cancer disease as being Benign or Malignant. The framework in which the model has been created is the Apache Spark framework. Various machine learning algorithms have been used to perform this research, including support vector machines for classification, logistic regression classifier algorithms, and the random forest classification algorithm [9]. In this research, classification is performed on the datasets individually with the help of scaling the machine learning algorithms. To aid this research, the Apache Spark platform is being used, and the machine learning algorithms used for this purpose are decision trees, random forest, and support vector machines. The dataset used is the data available for breast cancer from Wisconsin dataset from the UCI machine learning dataset library. The experiment has been performed and executed successfully in the Scala environment. As per the results of this experiment, the support vector machine serves better than the rest of the models created and obtains greater accuracy and a lower error rate with less consumption of time when compared to the other two algorithms. To perform the experiment, single-node and multi-node Spark clusters are being created.

Besides classification and prediction or detection of breast cancer, it is also essential that correct drugs or medicine are also being provided to the patients suffering from this disease to cure the disease and lead a peaceful life once again. Therefore, this research is based on discovering drugs for breast cancer with the help of big data analytics. As per recent times, we have a lot of scientific data related to various diseases, which helps us solve and detect or predict several conditions, which is very difficult to process manually [10]. This massive amount of data can also be used to keep an eye on or monitor the patients' health. This research focuses on discovering drugs with the help of high-performance computational techniques and big data analysis and processing systems, such as using statistics in the form of machine learning algorithms or data science. This kind of processing is also termed virtual screening and is one of the most intensive processes in terms of computational power. This research is performed using the machine learning algorithms designed for working with extensive data analysis on MapReduce and the Mahout, which helps pre-filter the massive set of ligands for virtual screening of the patients suffering from protein receptors related to breast cancer disease.

This research paper presents an efficient model for improving the accuracy of the breast cancer classification or prediction models. The dataset used is the data available for breast cancer from Wisconsin dataset from the UCI machine learning dataset library. The dataset used is the data available for breast cancer from Wisconsin dataset from the UCI machine learning dataset library. It is also important to note that breast cancer is the second most leading cause responsible for deaths of women all across the world. However, if breast cancer is detected successfully at an early stage, then the chances of death of the patient reduce in a very significant manner [11]. Therefore, it is necessary to detect this deadly disease using necessary statistical and mathematical tools such as automation or machine learning algorithms to detect this disease. To obtain much higher accuracy, several machine learning algorithms have been considered: decision trees, random forest, convolutional neural networks logistic regression, naïve Bayes, k nearest neighbors, MLP, and support vector machine used to increase the accuracy of classification. It is observed that the deep neural networks get harder to optimize with the increase in the time of configuration, and the best result is obtained when we perform the experiment using three hundred feature maps. As a result, this experiment brings a higher accuracy of up to 98% to 99%.

In this research paper, an optimized model of artificial neural networks is being used, based on the big data environment, to perform the classification to predict breast cancer disease. The data used for this processing is the unprocessed data for breast cancer. This research has been conducted to outperform the accuracies obtained by the previous research algorithms and methods for the detection of breast cancer [12]. Breast cancer is being classified as a benign or malignant tumor based on the model created in this research experiment. Hadoop MapReduce is used for the elimination of repeated or duplicate data. The data is then processed with the help of the RMA method or the replacement of the missing attributes method. Then various normalization techniques are being employed to experiment. The features for the model building have been chosen using the modified dragonfly algorithm or MDF, and then the selected features are taken as input for classification. Finally, optimization of the model is carried out, with the help of a gray wolf optimization algorithm where the experiential outcomes are being considered and contrasted with the prevailing IWDT or improved weight decision tree and are regarded in terms of precision, recall, accuracy, and ROC curve.

In this research paper, the model has been created to identify breast cancer to help the patients recover at an early stage with the help of the tweets streaming from patients. This experiment has been performed on the Spark platform with the support of four different machine learning algorithms [13]. The two feature selection algorithms employed in this research model are the univariate feature selection algorithm and the recursive feature elimination algorithms, which are then used to the features after the correlation is done to select essential elements. To perform this experiment, four machine learning algorithms are employed: the decision trees, random forests, logistic regression, and the support vector machine (SVM) for performing the classification of the patients suffering from breast cancer and identifying the tumor as benign or malignant. Additionally, to optimize the models further, cross-validation techniques and hyperparameter tuning are applied along with the machine learning algorithms to enhance the accuracy of the models created. To develop the second component of the research, tools such as Apache Kafka, Apache Spark, and Twitter streaming API are used. The first component obtains the highest accuracy to predict breast cancer disease in real-time from the streaming of the tweets. According to the results of the models, the random forest classification model achieves the best accuracy.

III. METHODOLOGY USED IN THIS RESEARCH

Dataset used for this research

To perform this research experiment, we took the dataset from Kaggle, the breast cancer Wisconsin diagnostic data set. This dataset consists of about 32 columns and is very effective for working on experiments related to breast cancer. Many investigations concerning machine learning algorithms have been done relative to this dataset. This dataset presents the data in the form of benign and malignant tumors and is very useful for the classification of breast cancer.

To continue with modeling the dataset, we will use several machine learning algorithms, which will help us efficiently model this research. We will use logistic regression, random forest classifier, and gradient boosted tree classifier algorithms in this research. This section provides a brief detail regarding these algorithms and enlists why we are applying these algorithms.

Logistic regression: The logistic regression algorithm is a regression algorithm that can also be implemented for classification purposes, and in this model, we are using it as a classification algorithm. It is a statistical model that is often used for predictive analysis. As this research focuses on the prediction of breast cancer in patients, this approach can help us gain a positive difference as per the other models as they predict the outcomes to help us understand the relationships and predict the outcomes. The logistic regression is very beneficial when applying the algorithm for a binary result system, as in this case, we are trying to understand and differentiate whether the breast cancer tumor is malignant or benign. It can also be used to classify multinomial datasets, but the binomial is being applied in this case. The binomial logistic regression is most helpful in modeling the event probability for a categorical response variable with two different outcomes. The most crucial benefit of using logistic regression is that it is easier to implement and train efficiently. Moreover, it does not make any assumptions about the distribution of the classes in feature space. The equation for logistic regression is:

$$h\theta(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

Random forest classifier: The random forest classifier is a machine learning algorithm that can be applied for classification and regression problems. The random forest uses the ensemble learning method, and through the bagging process of various decision trees, the random forest is formed. After that, results of the decision trees are combined altogether, and a voting process is involved, which will be used for the prediction results. A decision tree is a form of machine learning algorithm which forms a tree through the help of the binary decision-making process and answers the questions to decide in the form of binary such as true or false, yes or no, and similar method. First, the root node is set up with this decision-making process, and then several nodes which make the body of the tree are being set, and the final node or the leaf node is the one that produces the result or the output. The decision tree provides the development with a lower variance; therefore, an ensemble of many decision trees is formed using random forest, which builds a model that will nullify the disadvantages of the decision tree, and the best possible result is obtained through this method. However, the complexity increases due to the number of trees to be trained. The random forest can be demonstrated with the help of this mathematical equation:

$$\hat{Y}(\boldsymbol{X}) = \sum_{i=1}^{5} Y_i \times I_i(\boldsymbol{X})$$

Gradient-boosted tree classifier: The gradient boosted tree classifier is one of the most critical and influential techniques used to build predictive models. The gradient boosting tree classifier is a boosting algorithm and therefore helps the weak learner to become better and produce better results with the help of modification. Adaboost was the first boosting algorithm, then generalized as gradient boosting for gradient tree growing algorithm. The gradient boosting algorithm works based on three key steps, which involve optimizing a loss function. A weak learner is used for making predictions, and the third step is the additive model for the addition of vulnerable learners to minimize the loss function. In this research experiment, weak learner concept of the decision tree is being applied in which the decision tree is being used. The decision tree is an inefficient model, and as it is weak and has several flaws, gradient boosting is being applied to strengthen the models. It helps in maintaining the models based on their purpose and use. Finally, we obtain a model with higher accuracy than the original one. The equation for gradient boosted tree classifier is:

$$y_i^p = y_i^p + \alpha * \delta \sum (y_i - y_i^p)^2 / \delta y_i^p$$

which becomes, $y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p)$

where, α is learning rate and $\sum (y_i - y_i^p)$ is sum of residuals

The discussions related to models being trained through this research are carried out, and new models have been introduced. In the next section, the discussions related to the application of these algorithms and their results are being provided.

The experiment procedure can be overviewed with the help of this picture provided below:

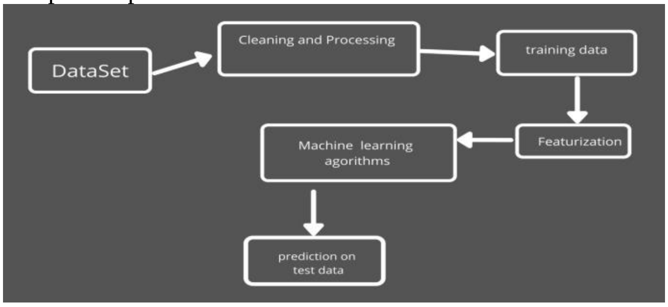

Figure 1: Methodology steps.

IV. RESULTS AND DISCUSSIONS

In this section, discussion on how experiment is conducted is produced, and results are being attained throughout the research. This experiment has been carried out with the help of Scala language, and machine learning techniques are being employed to carry out this research. First, we load the dataset in the CSV format and present a dataset schema to get an insight into it. Then the "NA" values are dropped, and the indexer is employed, which performs a procedure to merge the diagnosis column with the label, and the model is encoded. After this, columns that are not of much importance to this research are being dropped off, and the columns dropped are diagnosis, _c32, and id.

Once the final dataset is received, we will use the vector assembler, which will help us convert the dataset into vectors. Now the models can be trained through this. The logistic regression model is applied by setting the maxiter parameter as 10, regpapam as 0.3, elasticnetparam as 0.8, and the labelcol is set as the label. The traintestsplit function is used for the random split of the dataset. Then the run function is applied to fit the model in the data frame.

The pipeline is being utilized for encoding the model's accuracy and producing the results. Similarly, the random forest classifier and the gradient boosting tree classifier are being employed in this architecture. This research has been conducted to outperform the accuracies obtained by the previous research algorithms and methods for the detection of breast cancer. The results are being tested through two parameters which are accuracy and recall. With the help of this research experiment, we try to deduce a method for predicting and detecting breast cancer at an early stage in women. We will use the technologies and the advantages of the huge amount of data stored for research all these years and given the name of Big Data along with the help of a deep learning algorithm to detect the symptoms or the tumor at an early stage. The table below shows results obtained by each algorithm in the form of precision and recall parameters.

Algorithm Used	Accuracy	Recall
Logistic regression	1.0	1.0
Random forest classifier	0.987	0.964
Gradient boosting tree classifier	1.0	1.0

Table 1: Results obtained through the algorithms.

The table of results provided above shows that the logistic regression and the gradient boosting tree classifiers offer the best, and the classification results show that the algorithms have provided accurate classification with no errors. Thus, we can rely on this model. This model will help reduce the work of the doctors by saving a lot of time by classifying the cancer cells and the presence of breast cancers. However, even though an automated classification of breast cancer is helpful for society, the patients must get their cancer diagnosed. If the classification is done at an early stage, it can be treated efficiently, which is helpful in the early detection and treatment of breast cancer. In this research, classification is performed on the datasets individually with the help of scaling the machine learning algorithms.

Besides classification and prediction or detection of breast cancer, it is also essential that correct drugs or medicine are also being provided to the patients suffering from this disease to cure the disease and lead a peaceful life once again. The awareness among women is thus essential, which will motivate them to visit the diagnosis center at an early stage, which can help save the lives of many women around the world. As this research focuses on the prediction of breast cancer in patients, this approach can help us gain a positive difference as per the other models as they predict the outcomes to help us understand the relationships and predict the outcomes. The death rate due to breast cancer is relatively high, and due to breast cancer, a patient only hopes to live for five to six years which is less. Although some methods can help the patients get rid of the infections of breast cancer, if it infects the cells of other body parts, it will become hazardous, and this is why it must be stopped in its early stages.

V. CONCLUSION

In this research experiment, a machine learning model is being trained to classify breast cancer which is one of the most dangerous diseases concerning women. There are several factors responsible for this. Therefore, it is necessary to reduce the dangers of this disease so that a quick classification of the tumor can be classified, which will help society to fight this type of cancer efficiently. Due to the complexity of cancer and why women are not open about this disease is an essential factor to consider that breast cancer is dominant and is one of the most dangerous diseases. However, several programs and awareness campaigns are being organized to motivate women to speak up and save themselves at an early stage of this disease. This research performs a classification of the breast cancer tumor to identify if the cancer is deadly or not. Therefore, this research will help classify breast cancer with the help of three machine learning algorithms: logistic regression, random forest classifier, and gradient boosting classification algorithms. The accuracy obtained through logistic regression with the pipeline is 1.0, random forest classifier is 0.98, and gradient boosting tree classifier is 1.0. Thus, logistic regression and gradient boosting tree classifiers are excellent for solving this problem. Through this research, breast cancer can be classified more quickly and efficiently.

REFERENCES

[1] R. A. Ibrahem Alhayali, M. A. Ahmed, Y. M. Mohialden, and A. H. Ali, "Efficient method for breast cancer classification based on ensemble hoffeding tree and naïve Bayes," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 2, pp. 1074–1080, 2020, doi: 10.11591/ijeecs.v18.i2.pp1074-1080.

[2] P. D. Hung, T. D. Hanh, and V. T. Diep, "Breast cancer prediction using spark MLlib and ML packages," *ACM Int. Conf. Proceeding Ser.*, pp. 52–59, 2018, doi: 10.1145/3309129.3309133.

[3] T. Daghistani, H. AlGhamdi, R. Alshammari, and R. H. AlHazme, "Predictors of outpatients' no-show: big data analytics using apache spark," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00384-9.

[4] A. Eleyan, "Breast cancer classification using moments," *2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet.*, pp. 1–4, 2012, doi: 10.1109/siu.2012.6204778.

[5] M. J. Rasool, A. S. Brar, and H. S. Kang, "Risk Prediction of Breast Cancer From Real Time Streaming Health Data Using Machine Learning," no. 11, pp. 409–418, 2020, doi: 10.5281/zenodo.4284315.

[6] K. Men *et al.*, "Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning," *Phys. Medica*, vol. 50, no. December 2017, pp. 13–19, 2018, doi: 10.1016/j.ejmp.2018.05.006.

[7] K. Sivakami, "Mining Big Data : Breast Cancer Prediction using DT - SVM Hybrid Model," *Int. J. Sci. Eng. Appl. Sci.*, no. 5, pp. 418–429, 2015.

[8] S. Alghunaim and H. H. Al-Baity, "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context," *IEEE Access*, vol. 7, pp. 91535–91546, 2019, doi: 10.1109/ACCESS.2019.2927080.

[9] W. S. Albaldawi and R. M. Almuttairi, "Prediction Breast Cancer as Benign or Malignant in Apache Spark Framework," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 928, no. 3, 2020, doi: 10.1088/1757-899X/928/3/032046.

[10] R. Mennour and M. Batouche, "Drug discovery for breast cancer based on big data analytics techniques," *2015 5th Int. Conf. Inf. Commun. Technol. Access. ICTA 2015*, 2016, doi: 10.1109/ICTA.2015.7426894.

[11] C. Shahnaz, J. Hossain, S. A. Fattah, S. Ghosh, and A. I. Khan, "Efficient approaches for accuracy improvement of breast cancer classification using Wisconsin database," *5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017*, vol. 2018-January, pp. 792–797, 2018, doi: 10.1109/R10-HTC.2017.8289075.

[12] M. Supriya and A. J. Deepa, "A novel approach for breast cancer prediction using optimized ANN classifier based on big data environment," *Health Care Manag. Sci.*, vol. 23, no. 3, pp. 414–426, 2020, doi: 10.1007/s10729-019-09498-w.

[13] N. F. Omran, S. F. Abd-El Ghany, H. Saleh, and A. Nabil, "Breast Cancer Identification from Patients' Tweet Streaming Using Machine Learning Solution on Spark," *complexity*, vol. 2021, 2021, doi: 10.1155/2021/6653508.