University of South Wales
Prifysgol De Cymru

Assignment on

Breast Cancer Prediction with Machine Learning Algorithms

Module Title: Machine Learning

Module Code: CS4S771

Student Id: 30091254

Submission Date: 11-01-2024

Table of Contents

1. Introduction

Machine learning provides the ability of automatically learning process and improve the form of experiences without being explicitly programmed. The learning process is being provide by the data observation, direct experiences or instruction for identifying pattern in data for better decision making. Machine learning's basic proposition is to build algorithms for predicting results within an acceptable range from the input data set. Machine learning system deals with different types of algorithms which are supervised or unsupervised. In supervised learning both input and output data need to be given to learn the machines in addition to furnishing feedback about the accuracy of predictions during training. After training the algorithm is being processed with the new data set to find the performance of the machines. In unsupervised learning the desired output need not to be given. It works with more complicated data than supervised learning. A technique named Mammography is used for breast cancer identification. But this technique fails because of choosing features and class imbalance problem.

Breast cancer is one of the most dangerous diseases in the recent times. It can be affected both men and women. Though it is more common in the women. It makes one physically and psychologically weaker. It happens for the abnormal grows of cells that can affect other body part. It is not easy to identify breast cancer in the starting stage.

In our study our aim is to predict Malignant or Benign from a given dataset using different types of machine learning algorithms also find out which algorithm gives the better result.

2. Algorithms

2.1 Logistic Regression

Logistic Regression is used for binary classification which is a supervised machine learning algorithm. Logistic regression predicts the output of a categorical conditional variable. The result must be a categorical or discrete value. It gives a value within 0 and 1. It is used for classifying any solutions. This algorithm is easier to implement as it predicts between binary value 0 and 1. It can work with categorical data as well as continuous data.

2.2 K-Nearest

(K-NN) algorithm is a supervised machine learning algorithm that is used for its clarity and mitigation of implementation. It does not demand any assumptions about the underlying data distribution. It can work with numerical and categorical data. The K-NN algorithm works by finding the Euclidean distance with K nearest neighbors to a given data point. This algorithm works with the local data structure for making predictions.

2.3 SVM

Support Vector Machine is a supervised machine learning text classification algorithm. Here each data item is being plot as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiate the two classes very well. SVM works well with clear margin of separation, and effective in high dimensional spaces. It is effective in cases where number of

dimensions is greater than the number of samples. It is memory efficient as it uses a subset of training points in the decision function.

In SVM, a discriminative classifier formally defined by a separating hyper plane with given labelled training data. The algorithm outputs an optimal hyper plane which categorizes new examples. In two-dimensional space this hyper plane is a line dividing a plane in two parts where in each class lay in either side

2.4 Decision Tree Model

Decision Tree Model works with a tree structure of nodes where the nodes are presented the features. Decision Tree Model handle high dimensional data to fin out accuracy instead of prediction. Decision Tree has a root node that partitions the attributes. It helps to make choice according to the priority. This technique can be used for classification and regression.

2.5 Random Forest

Random Forest is a supervised learning algorithm. To get a more accurate and stable prediction this algorithm builds multiple decision trees and merges them together. It can be used for classification problems which form most current machine learning systems. The random-forest algorithm brings extra randomness into the model, when it is growing the trees. To create wide diversity for a better model, random-forest algorithm instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features.

For applications in classification problems, Random Forest algorithm will avoid the over-fitting problem. It can be used for identifying the most important features from the training data set, in other words, feature engineering.

3. Training

We implement a model to classify Malignant or Benign from the given dataset.

3.1 Data Preprocessing: From the dataset we need to preprocess them. We need to make ready the dataset by cleaning noise, outliers, categorical data to numeric, normalize and standardize for the model. We must handle if any null value is present in the data set.

- **Null Value:** In the data set there are no null values.
- **Duplicate Value:** In the data set there are no duplicate values.
- **Label Encoding:** Label Encoding is a technique for encoding categorical values to numerical. Values are replaced by some numeric number. We use Label encoding for our categorical values Malignant (M) for 0 and Benign (B) for 1

3.2 Feature selection: There are 30 features in our data set. From the we select 10 features first and train the model and after that we add three more feature to find out if there any changes occur or not. We use correlation matrix for feature selection.

```
selected_features = ['concave points_worst', 'perimeter_worst', 'concave
points_mean', 'radius_worst','perimeter_mean', 'area_worst', 'radius_mean',
'area_mean', 'concavity_mean', 'compactness_mean']
```

Final_selected_features = ['concave points_worst', 'perimeter_worst', 'concave points_mean', 'radius_worst', 'perimeter_mean', 'area_worst', 'radius_mean', 'area_mean', 'concavity_mean', 'compactness_mean','radius_se',’area_se','perimeter_se']

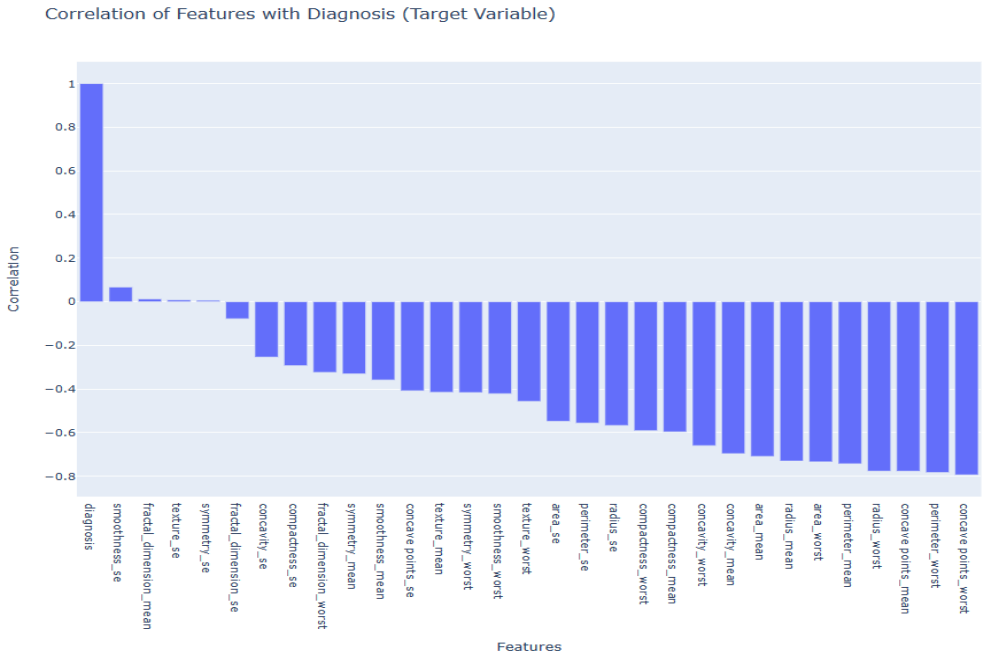


Fig1: Correlation Matrix

3.4 Model Selection: We select different algorithms to train our model.

1. Logistic Regression
2. K-Nearest
3. SVM
4. Decision Tree Model
5. Random Forest

3.5 Model Training: We train our model with the selected algorithms and get the accuracy to understand how the model is working.

3.6 Model Evaluation: Finally, we get classification report with confusion matrix, f-score, recall, precision and training and testing accuracy.

4. Result

We train our model with machine learning algorithms and get the accuracy. We make classification report for showing how the model works with different models.

4.1 Accuracy

Logistic Regression	Support Vector Machine	Decision Tree	Random Forest	k-Nearest Neighbors
Accuracy (Train): 0.9451 Accuracy (Test): 0.9825	Accuracy (Train): 0.9121 Accuracy (Test): 0.9474	Accuracy (Train): 1.0000 Accuracy (Test): 0.9298	Accuracy (Train): 1.0000 Accuracy (Test): 0.9561	Accuracy (Train): 0.9385 Accuracy (Test): 0.9474

Fig2: Accuracies of Algorithms

Here, we find that Logistic regression has the highest accuracy in both training and testing. In Decision Tree model it gives more accuracy in training but less accuracy in testing that means this algorithm overfits the model. On the other hand, Random Forest also gives more accuracy in training and 2nd highest in testing. K-Nearest Neighbor also gives more accuracy in both training and testing.

4.2 Classification Report

Evaluating model performance, we generate classification report. In classification report we generate precision, recall, f1-score and support.

- **Precision:** Precision value deals with the correct positive prediction value and total positive predicted value. It shows the accurate prediction of positive outcomes.
- **Recall:** Recall deals with the correct positive prediction value and actual total positive value. It shows the sensitivity of the prediction of positive outcomes.
- **F1-score:** F1-score is calculated from the precision and recall value which is the harmonic mean of them. It deals with both accuracy and sensitivity. It is useful to identify imbalance result.
- **Support:** It shows the representation of the prediction of the actual distribution of the dataset classes.

```
Classification Report:
              precision    recall  f1-score   support

           B       0.93      0.99      0.96        71
           M       0.97      0.88      0.93        43

    accuracy                           0.95       114
   macro avg       0.95      0.93      0.94       114
weighted avg       0.95      0.95      0.95       114

--------------------------------------------------
```

Fig3: Classification Report for Logistic Regression

```
Classification Report:
              precision    recall  f1-score   support

           B       0.93      0.96      0.94        71
           M       0.93      0.88      0.90        43

    accuracy                           0.93       114
   macro avg       0.93      0.92      0.92       114
weighted avg       0.93      0.93      0.93       114

--------------------------------------------------
```

Fig4: Classification Report for Decision Tree Model

```
Classification Report:
              precision    recall  f1-score   support

           B       0.93      0.99      0.96        71
           M       0.97      0.88      0.93        43

    accuracy                           0.95       114
   macro avg       0.95      0.93      0.94       114
weighted avg       0.95      0.95      0.95       114

--------------------------------------------------
```

Fig5: Classification Report for K-Nearest Neigbors

```
Classification Report:
              precision    recall  f1-score   support

           B       0.92      1.00      0.96        71
           M       1.00      0.86      0.92        43

    accuracy                           0.95       114
   macro avg       0.96      0.93      0.94       114
weighted avg       0.95      0.95      0.95       114

--------------------------------------------------
```

Fig6: Classification Report for SVM

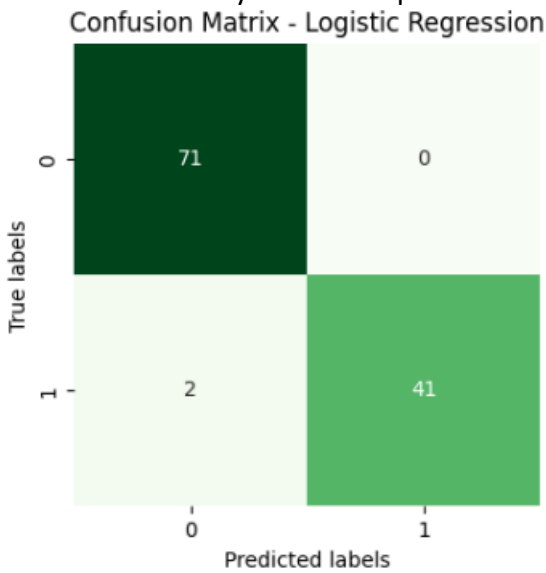
```
Classification Report:
              precision    recall  f1-score   support

           B       0.96      0.97      0.97        71
           M       0.95      0.93      0.94        43

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114

--------------------------------------------------
```

Fig7: Classification Report for Random Forest

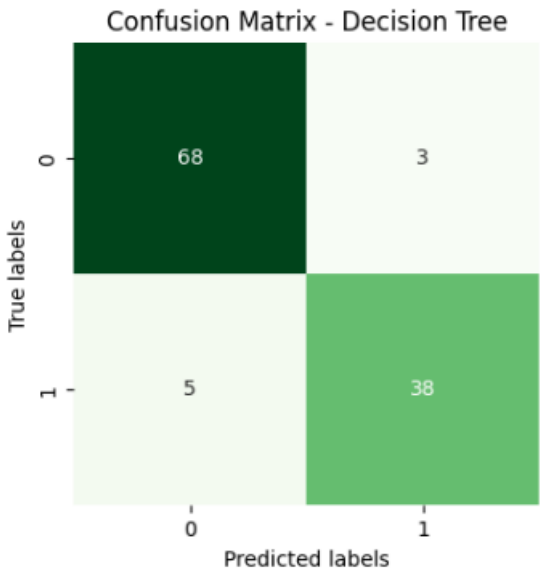
4.3 Confusion Matrix:

Confusion matrix shows the accurate and inaccurate prediction values of test data. It helps to calculate the model's predictions accuracy.

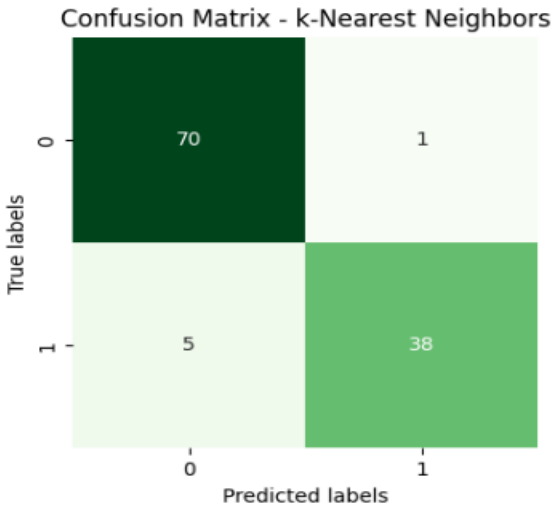
- **True Positive:** It shows how correctly the model predict the true values.
- **True Negative:** It shows how correctly the model predict the false values.
- **False Positive:** It shows how incorrectly the model predict the true values.
- **False Negative:** It shows how incorrectly the model predict the false values.



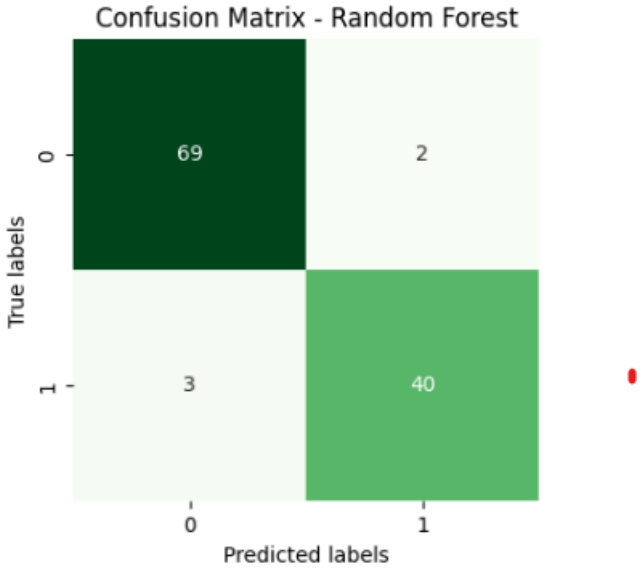
Logistic Regression gives 71 True positive and 41 True Negative and 0 False positive and 2 False negative.



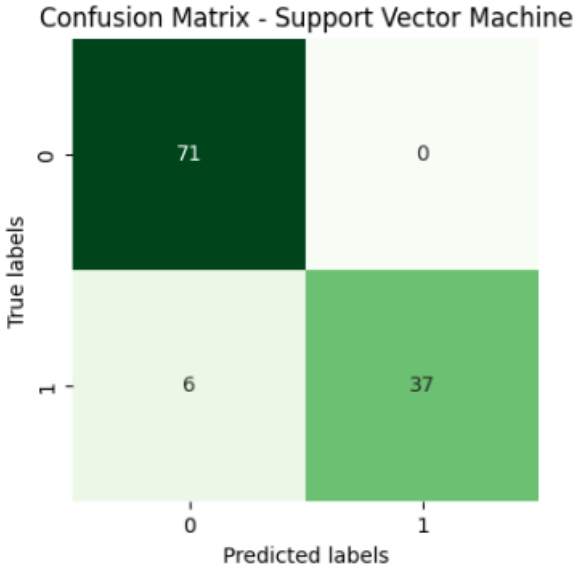
Decision Tree gives 68 True positive and 38 True Negative and 3 False positive and 5 False negative.



K-Nearest Neighbor gives 70 True positive and 38 True Negative and 1 False positive and 5 False negative.



Random Forest gives 69 True positive and 40 True Negative and 2 False positive and 3 False negative.



SVM gives 71 True positive and 37 True Negative and 0 False positive and 6 False negative.

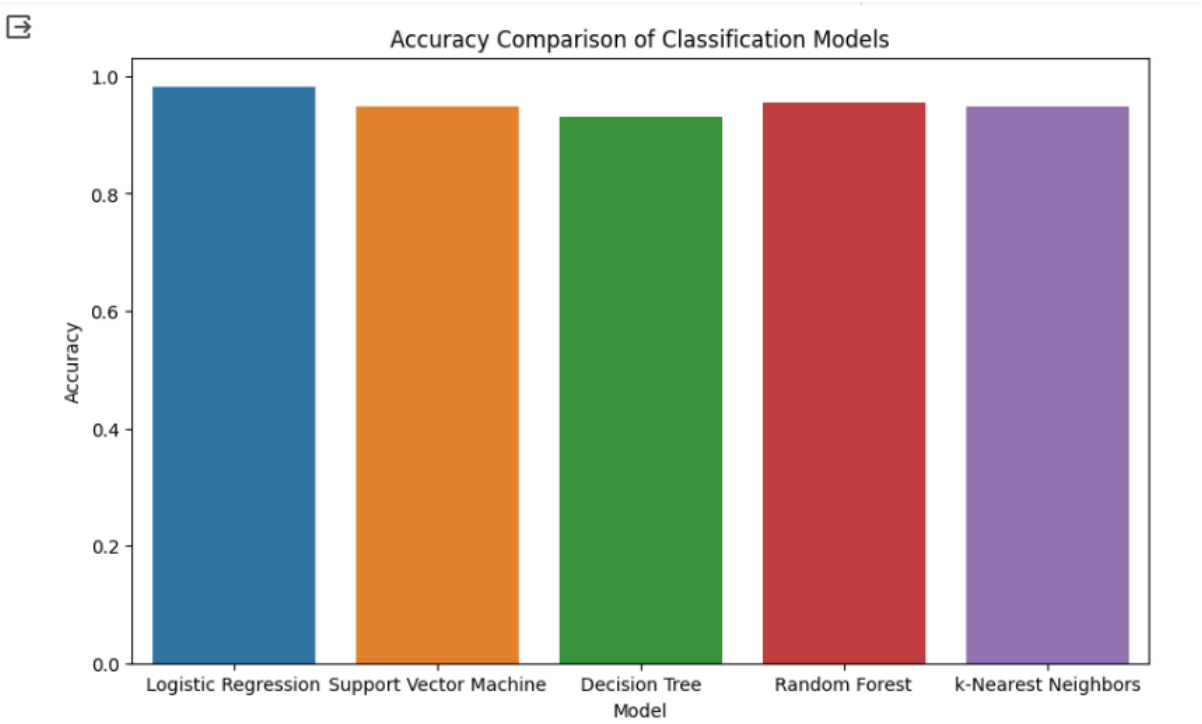


Fig8. Accuracy Comparison Chart

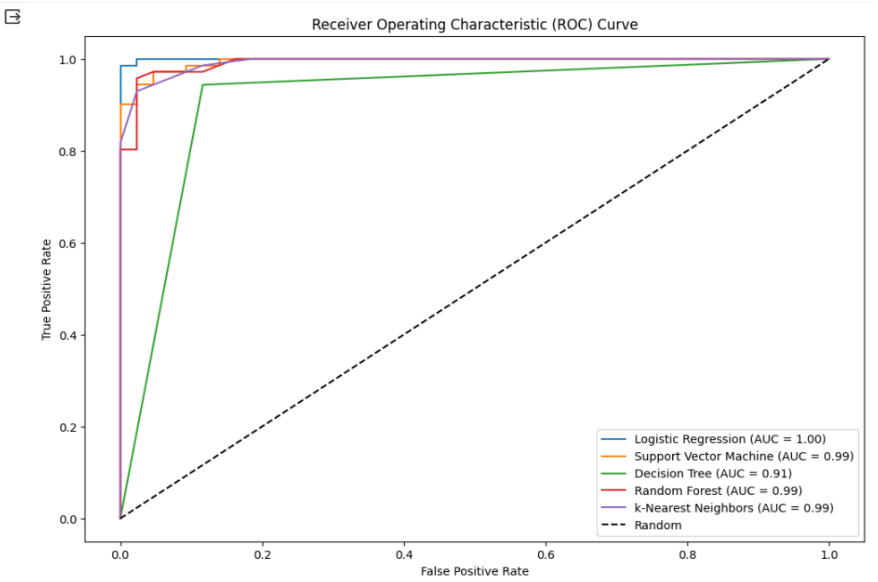


Fig9: Accuracy Line plot with 10 Features

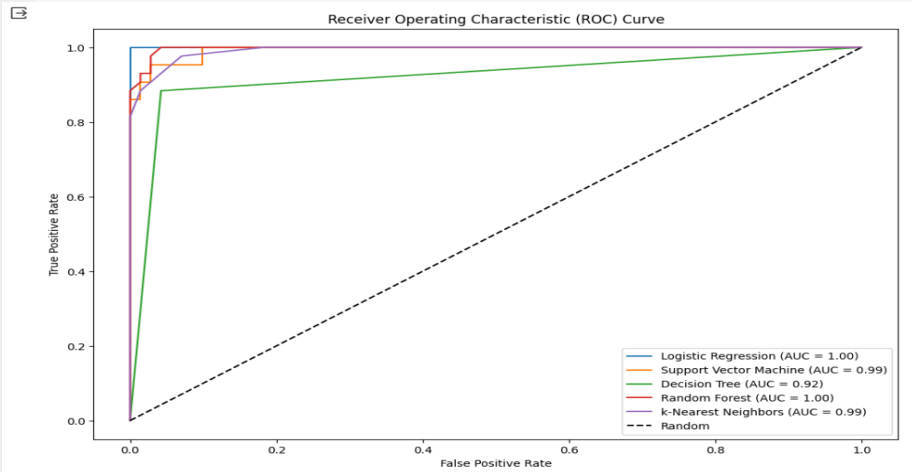


Fig10: Accuracy Line plot with 13 Features

From the accuracy graph we find that the Logistic Regression is giving the highest accuracy in prediction and the random forest gives the 2nd highest accuracy, but we also find the random forest gives more accuracy in training set than testing set. The SVM gives more accuracy in both training and testing data set.

5. Limitations

The limitation of our study is the dataset is not enough to train a model that can work in the real world. We also can not implement any software application with we can test it in real. We can work in this model to launch it in a software application for the medical sector using purpose in future.

6. Conclusion

Machine Learning algorithms are being used to train a model for predicting any solutions. In our study we use different types of machine learning algorithms to understand how they works in predicting cancer from the given dataset. In the medical field machine learning algorithms and the prediction model are being used for analysing huge data to find out a pattern that can easily identify the diseases. Breast Cancer becomes very common in women in the world. It is necessary to detect breast cancer in the early stage so people can get treatment before it goes in serious issue. Our model can make a system that will help to identify the breast cancer.

7. Reference

Das, A.K., Biswas, S.Kr., Mandal, A., Bhattacharya, A. and Sanyal, S. (2024) 'Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP)'. *Expert Systems with Applications* 242, p. 122673. doi: 10.1016/j.eswa.2023.122673.

Duan, H. et al. (2024) 'Machine learning-based prediction model for distant metastasis of breast cancer'. *Computers in Biology and Medicine* 169, p. 107943. doi: 10.1016/j.compbiomed.2024.107943.

Hassan, Md.M. et al. (2023) 'A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction'. *Decision Analytics Journal* 7, p. 100245. doi: 10.1016/j.dajour.2023.100245.

'Logistic Regression in Machine Learning'. (2017) *GeeksforGeeks* 9 May. Available at: https://www.geeksforgeeks.org/understanding-logistic-regression/ (Accessed: 12 January 2024).

Ray, S. (2017) 'Learn How to Use Support Vector Machines (SVM) for Data Science'. *Analytics Vidhya* 12 September. Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (Accessed: 12 January 2024).

Shaikh, F.J. and Rao, D.S. (2022) 'Prediction of Cancer Disease using Machine learning Approach'. *Materials Today: Proceedings* 50, pp. 40–47. doi: 10.1016/j.matpr.2021.03.625.