Electronic supplementary Information

Supplementary A. Machine Learning

Modelling huge volumes of data using any conventional approaches is time consuming and lacks accuracy as well. These approaches may be easily interpretable, but they are incapable of dealing with non-linear patterns and lack flexibility. Machine learning regression algorithms find application in prediction of continuous variable. The function of these algorithms is to find the best mapping function (f) from input variables (X) to continuous output variable (Y), given the time and resources. Reaction fraction is a continuous variable and thus, this study involves regression algorithms namely multivariate linear regression, decision trees and artificial neural networks.

SA.1 Multivariate Polynomial Linear Regression

Multivariate polynomial linear regression combines multivariate and polynomial regression to model non-linear relationships between a continuous output variable and multiple input variables. Equation a shows the relationship between input and output variables in multivariate polynomial regression.

$$y = b_0 + b_1X_1 + b_2X_{2+} \ldots + b_nX_{n+} c_{11}X_1^2 + c_{22}X_2^2 + \ldots + c_{nn}X_n^2 + c_{12}X_1X_2 + \ldots + c_{ij}X_iX_j + \ldots + c_{nm}X_n^m \quad \text{(a)}$$

where y is the output variable, b_0 is the intercept term, b_1 to b_n are the coefficients for the linear terms of the input variables X_1 to X_n respectively, c_{11} to c_{nn} are the coefficients for the quadratic terms of the input variables, c_{12} to c_{nm} are the coefficients for the interaction terms between pairs of input variables, and m is the degree of the polynomial.

The degree of polynomial used is a crucial hyperparameter, impacting the learning process. Choosing a low degree may result in underfitting, while a high degree may lead to overfitting and reduced accuracy on test data. Selecting the optimal degree is essential for achieving better accuracy.

SA.2 Decision Tree

A decision tree is a machine learning algorithm that can be used for both classification and regression tasks. It is a tree-like model where each internal node represents a feature or attribute, each branch represents a decision or rule based on the value of that feature, and each leaf node represents a class label or a numerical value [1]. Decision trees aim to split the dataset into subsets by choosing the feature that maximizes information gain, a measure of entropy

reduction. This process recurs until a stopping criterion, like reaching a maximum depth or minimum samples in a leaf node, is met.

SA.3 Artificial Neural Network

An Artificial Neural Network (ANN) is a machine learning algorithm inspired by the human brain [2]. It excels in solving nonlinear problems and handling defects and noise. ANNs can approximate any function but can become complex in multi-layered networks. They require a large initial dataset and are sensitive to data pre-processing. Addressing overfitting is crucial for minimizing testing errors. The network comprises input nodes, weights, biases, a transfer function, and output. Neuron outputs, influenced by weights, use activation functions to predict outputs for input variables.

Mathematical Equation:

$$y_j = \psi\left(\sum_{i=1}^{n} w_{ji} x_i + \theta_j\right) \quad \text{(b)}$$

where, θ_j : offset or bias

w_{ji} : weights

x_i : inputs

y_j : output

ψ : Transfer or Activation Function such as RelU, tanh etc.

Equation b shows the mathematical equation involving the input and output variables and all other parameters which are used to create artificial neural network model.

SA.4 Performance Metrics

Performance of regression models was evaluated using some standard evaluation measures. Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and coefficient of determination (R^2) are the most commonly used evaluation measures. Equation (c), (d) and (e) shows the formulae for calculation of MAPE, RMSE and R^2 respectively.

$$Mean\ Absolute\ Percentage\ Error,\ MAPE\ (\%) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{predict,i} - y_{data,i}}{y_{data,i}}\right| \times 10 \quad \text{(c)}$$

$$Root\ Mean\ Square\ Error,\ RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{predict,i} - y_{data,i})^2}{n}} \quad \text{(d)}$$

$$Coefficient\ of\ Determination,\ R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{predict,i} - y_{data,i})^2}{\sum_{i=1}^{n}(y_{predict,i} - \bar{y}_{data})^2} \quad \text{(e)}$$

where, $y_{predict,\,i}$ is the predicted output value for i^{th} observation,

$y_{data,\,i}$ is the actual value for i^{th} observation,

$\bar{y}_{data}$ is the mean of actual values and

n is the number of observations

Supplementary B. Optimization

An optimization algorithm is a process that is carried out repeatedly while evaluating numerous solutions in search of the best or most acceptable one. Optimization algorithms are mainly divided into two types: Gradient-based algorithms and Gradient-free algorithms. In addition to function evaluations, gradient-based algorithms need gradient or sensitivity information to decide on appropriate search directions for better designs throughout optimization iterations [3]. These algorithms can only be applied to continuous and differentiable functions as they use derivative information to guide the search process. On the other hand, Gradient free algorithms are independent of the problem and thus, can be applied to a wide range of problems. These algorithms are population-based algorithms which means that they begin with a random population of solutions and iteratively evolve the population towards better solutions. These algorithms are used in such optimization problems where the search space is large and complex. There are several population-based optimization algorithms such as Genetic Algorithm, Particle Swarm Algorithm, Simulated Annealing etc. Genetic Algorithm are good at exploring the search space and finding diverse set of solutions, but they take longer to converge to an optimal solution. On the other hand, Particle Swarm Algorithm finds optimal solution quickly

but are not as accurate as Genetic Algorithm in constraint satisfaction problems [1]. Simulated Annealing finds global optima but may take longer to converge to an optimal solution compared to other algorithms. Thus, Genetic Algorithm is used in the study to get accurate optimization results even with longer convergence time.

SB.1 Genetic Algorithm

Genetic Algorithms are inspired from Charles Darwin's theory of natural evolution where survival of the fittest is more likely to happen. Natural Selection works by choosing the fittest individuals for reproduction in order to produce the children of the next generation.

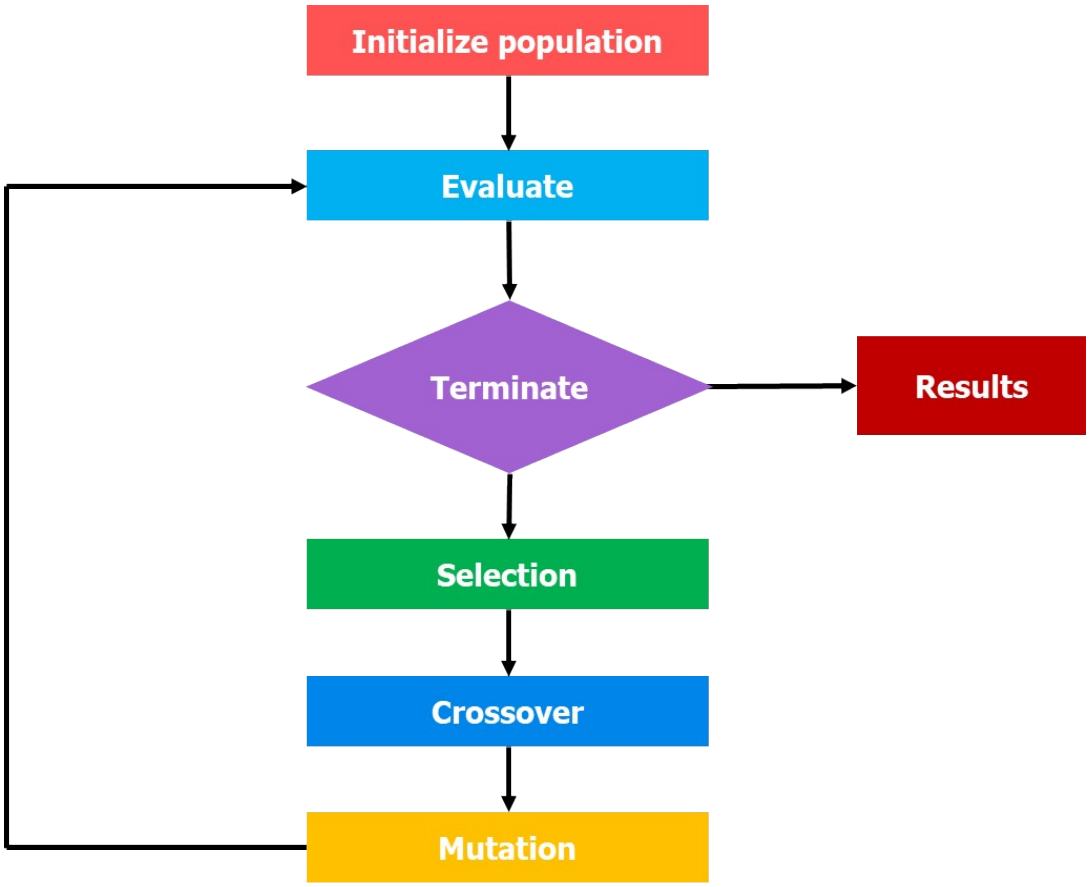

Figure a. Flowchart of Genetic Algorithm

The Genetic Algorithm (GA) operates in five stages, shown in Figure a. Firstly, the population is initialized by creating random potential solutions known as chromosomes. Each chromosome represents a solution to the optimization problem. The population is a matrix of chromosomes. Next, the chromosomes are evaluated using a fitness function to determine their suitability. The fittest solution is recorded, and the algorithm checks if the termination criteria are met. If the termination condition is not satisfied, the selection stage takes place. Individuals

are chosen from the population based on their fitness scores, with higher fitness increasing the probability of selection. Two selected individuals become parents. In the crossover stage, a random crossover point is chosen for each set of parents. This process combines genetic information from the parents to create offspring. Finally, in the mutation stage, a small percentage of genes in the offspring undergo random changes to maintain diversity in the population and avoid premature convergence. Overall, the GA progresses through population initialization, evaluation, selection, crossover, and mutation stages to find optimal solutions to the given optimization problem.

SB.2 Optimal Problem Formulation

The objective of formulation was to mathematically define the optimal design problem in a manner suitable for resolution by an optimization algorithm. Figure b outlines the typical steps involved in formulating an optimal design. Design variables were chosen and then, constraints were formulated for every design variable. Then, objective function was formulated to maximize or minimize certain parameters. There can be one or more objective functions. An optimization algorithm was chosen and solutions were obtained.

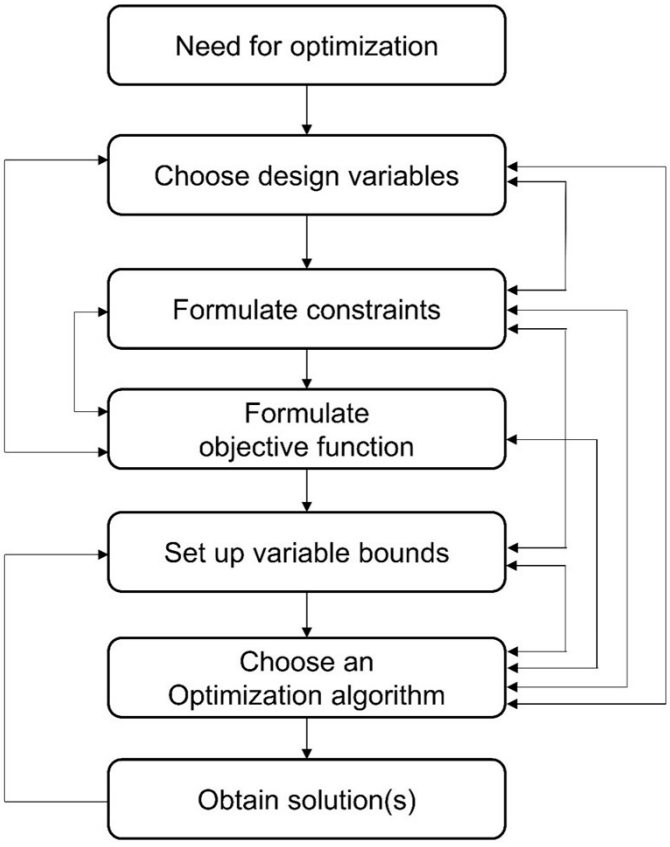

Figure b. Flowchart of optimal design procedure followed in this study

References

1 A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, ‘An introduction to decision tree modeling’, *J. Chemom.*, vol. 18, no. 6, pp. 275–285, 2004, doi: https://doi.org/10.1002/cem.873.

2 A. H. Neto and F. A. S. Fiorelli, ‘Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption’, *Energy Build.*, vol. 40, no. 12, pp. 2169–2176, 2008, doi: https://doi.org/10.1016/j.enbuild.2008.06.013.

3 J. Christensen and C. Bastien, ‘Chapter | four - Introduction to Structural Optimization and Its Potential for Development of Vehicle Safety Structures’, in *Nonlinear Optimization of Vehicle Safety Structures*, J. Christensen and C. Bastien, Eds., Oxford: Butterworth-Heinemann, 2016, pp. 169–207. doi: https://doi.org/10.1016/B978-0-12-417297-5.00004-3.

4 J. Bronson and K. Reed, ‘Particle Swarm Optimization Particle Swarm Optimization vs. Genetic Algorithms vs. Genetic Algorithms’, 2011. [Online]. Available: https://www.rose-hulman.edu/class/cs/csse453/archive/2011-12/presentations/PSOvsGA.pdf