Indian Journal of
Advances in Chemical Science

Article

Maximizing Profits in Bioreactors: Using Multi-Gene Genetic Programming and Genetic Algorithm Optimization for Explainable Models of Glucose to Gluconic Acid Conversion

Sucharita Pal*, Sandip Kumar Lahiri

Department of Chemical Engineering, University of Calgary, Calgary, Alberta, Canada

ABSTRACT

The present study focuses on building a model of a laboratory-scale bioreactor using genetic programming (GP) and optimizing it for profit maximization. The glucose to gluconic acid bioprocess was employed as a case study. It is challenging to create a reliable first principle-based model for the fermenter since it is a multiphase enzymatic bioreactor. On the other hand, data-driven models lack explicability. Consequently, a general methodology has been developed in this work, in which a data-driven approach, such as multi-gene GP, was used as a modeling tool, and the model was then post-processed to increase the explainability of the model. The model was used because it could effectively represent the underlying physics of the system. An acceptable model was constructed, and then it underwent optimization. This study looked at how to increase gluconic acid yield, which has a big influence on how profitable the process is. By applying an evolutionary algorithm to the produced model, an ideal solution was also discovered.

Key words: Multi-gene genetic programming, Bioprocess, Modeling, Optimization, Genetic algorithm.

1. INTRODUCTION

In the past decade, the corporate world has undergone significant changes due to heightened competition in a volatile market. As a result of globalization, chemical process firms are experiencing decreasing profit margins, necessitating technical innovation to increase process efficiency through yield maximization. To reduce costs and boost revenues, chemical companies worldwide are seeking innovative approaches. One such approach involves the utilization of data mining methods based on artificial intelligence to derive value from a substantial amount of experimental data. Knowledge discovery is a particularly promising cutting-edge method for process analysis. In this context, this paper explores the potential of knowledge discovery techniques to optimize chemical processes and improve profit margins. The study aims to contribute to the body of research on technical innovation in the chemical industry and offer insights into how data-driven methods can aid in addressing the challenges faced by chemical process firms in a highly competitive marketplace.

Chemical bioreactors have become an increasingly attractive avenue for researchers seeking new ways to generate revenue. These reactors are critical pieces of equipment that convert raw ingredients into finished goods and add significant value, making reactor optimization crucial for overall profitability. However, to fully utilize fed-batch bioreactors, intricate systems of bioprocesses must first be modeled [1]. Biochemical processes are characterized by intricate reaction kinetics and thermodynamics, making it challenging to create a workable phenomenological model for actual experimental reactors. The lack of knowledge of the kinetics of chemical reactions hinders reactor optimization, as most chemical bioreactors operate as "black boxes" due to reliability and safety concerns, leading to suboptimal operation and reduced profitability. Reactors are, therefore, unexplored territory in the world of chemical engineering. However, even a small improvement in catalyst selectivity and reaction yield in large-scale operations can have a significant impact on the number of raw materials used and profitability. To address these challenges, a straightforward option is to apply a data-driven, efficient computational technique to create approximate reactor models for complex reaction systems. These models can then be used to optimize the reactor and boost revenue. This paper presents a methodology that leverages data-driven techniques to optimize chemical bioreactors, using the example of fed-batch bioreactors for the production of finished goods. The study aims to contribute to the body of research on reactor optimization and offer insights into how data-driven models can aid in addressing the challenges faced by chemical engineering in optimizing bioreactors for maximum profitability [1]. The paper highlights the importance of understanding reaction kinetics and thermodynamics and the potential for data-driven approaches to generate significant value in the chemical industry.

Chemical experimentalists are accumulating vast quantities of reactor input and output experimental data every minute. The challenge is how to use this data to increase revenue. Data-driven modeling techniques, such as artificial neural networks (ANN) and support vector machines (SVM), have gained popularity due to their excellent prediction skills. However, these models are often rejected by engineers due

***Corresponding author:**
Sucharita Pal,
E-mail: palsucharita25@gmail.com

ISSN NO: 2320-0898 (p); 2320-0928 (e)
DOI: 10.22607/IJACS.2023.1102011

Received: 13th December, 2022;
Revised: 19th March, 2023;
Accepted: 22nd March, 2023.

to their lack of comprehensibility and "black-box" nature. While ANN produces equations with complex sigmoidal functions and tuning factors, SVM models are limited in their ability to explain the relationship between output variables and input characteristics. To better understand and profit from this relationship, process engineers seek intelligible equations in differential/algebraic form. Therefore, alternative computational techniques must be explored. This paper aims to investigate the effectiveness of multi-gene genetic programming (MGP), a data-driven approach, to develop a reliable model for the fed-batch glucose to gluconic acid bioreactor. The model will be optimized using a genetic algorithm (GA) to maximize profit while maintaining the explainability of the model. This study aims to contribute to the search for alternative computational techniques for reactor optimization, thus improving overall profitability in the chemical process industry.

The development of accurate models to predict the outcomes of complex systems is crucial in various fields of research. In chemical engineering, the use of data-driven modelling techniques such as ANN and SVM has proven successful in generating nonlinear models. However, the black-box nature of these models, combined with their lack of comprehensibility, has led to reluctance among engineers to adopt them for critical applications. This has spurred the exploration of alternative computational techniques to generate intelligible equations that can better explain the relationship between input variables and output characteristics of a system.

One such technique is GP, which has shown promise in overcoming the limitations of ANN and SVM models. GP is a subset of evolutionary modeling techniques that creates non-linear structured models as closed-form equations connecting the available data. The accuracy of each equation's numerous parameters is estimated to accurately forecast the outcome, and the application of "survival of the fittest" principles determines the likelihood of a model surviving to the next generation. GP allows the recombination of earlier models to generate new models, with the goal of improving predictability with each generation [2-4].

In the late 1980s, Koza's symbolic regression research demonstrated the breakthrough in GP, and subsequent studies have shown its effectiveness in various fields such as robotics, gaming, and control [3-5]. In chemical engineering, GP has been used for the dynamic and steady-state modeling of complex systems such as twin-screw frying extruders, binary distillation columns, non-linear vacuum distillation columns, and reaction test issues. In addition, GP has been combined with other techniques, such as principal component analysis (PCA), to develop nonlinear models for product design. GP has also been used to produce empirical models in a process system, and to automate the creation of nonlinear model predictive control. The promising results of GP in generating intelligible equations that accurately predict the outcome of complex systems make it a valuable technique for process engineers [5-9].

MGP has been proposed as a robust variation of GP for non-linear modeling and has been shown to produce more precise and computationally efficient models than normal GP. Despite its outstanding prediction skills, there are surprisingly few applications of MGP in fed-batch experimental bioreactors. In this study, we aim to fill this gap by applying MGP to model a biochemical reactor. MGP builds multi-gene mathematical models of predictor response data using low-order nonlinear combinations of input variables [10,11]. Unlike the conventional GP which evaluates a single tree expression, MGP combines many distinct genes using a multi-gene approach to make a single gene. The previous studies have shown that MGP performed better than other machine learning techniques such as ANN and SVM in terms of prediction and model simplicity. In this paper, we investigate the performance of MGP modeling in a biochemical reactor and compare it with other modeling techniques. Our results could provide new insights into the use of MGP for modeling fed-batch experimental bioreactors and could have significant implications for improving the profitability of these systems [11-15].

This study aims to develop a closed-form equation model for a fed-batch experimental bioreactor that is accurate, portable, and explainable for process engineers to understand the system better. The model will then be used to optimize the input process parameters using a nature-inspired metaheuristic optimization strategy to maximize the synthesis of gluconic acid. The multi-objective GA will be used to optimize the input space of the reactor, and Pareto optimal solutions will be obtained using the MGP model [16-20]. The purpose of this study is to demonstrate the applicability of the MGP technique in bioreactor modeling and optimization, as well as to provide a useful tool for process engineers in the bioprocessing industry [16-20]. Thus, a general methodology has been developed in this work, in which a data-driven approach, such as MGP, was used as a modeling tool, and the model was then post-processed to increase the explainability of the model. The model was used because it could effectively represent the underlying physics of the system. An acceptable model was constructed, and then it underwent optimization. This study looked at how to increase gluconic acid yield, which has a big influence on how profitable the process is. By applying an evolutionary algorithm to the produced model, an ideal solution was also proposed.

2. CASE STUDY OF GLUCONIC ACID BIOREACTOR

2.1. Background

Gluconic acid finds its applications in various industries, including medicine and textiles, as an acidulent, metal supplement, chelating agent, and more. However, due to the complexity of the multiphase enzymatic processes in fed-batch experimental bioreactors for gluconic acid production, it is challenging to obtain a reliable first-principle-based model. Therefore, data-driven modeling techniques are being explored as an alternative to overcome this challenge. The availability of a significant amount of process data from multiple bioreactor runs makes data-driven modeling an attractive option. The main objective of this project is to leverage this data to develop a framework that can translate the knowledge hidden in the data into increased profitability (Appendix Table 1). The project aims to create a reliable model of the fed-batch experimental bioreactor for gluconic acid production, which is accurate, portable, and easy to interpret. This model will be used to optimize the input parameters of the bioreactor using a multi-objective GA to maximize the yield of gluconic acid. By achieving these objectives, the project aims to offer insights into the optimization of gluconic acid production and its application in various industries.

2.1.1. Reactions

Commercial production of gluconic acid mainly involves two biological processes: Freecell fermentation and immobilized enzyme-based bioconversion of glucose. Glucose oxidase (GOD) of *Aspergillus niger* and *Gluconobacter* are commonly used in the latter method, which involves oxidation of glucose to glucono-d-lactone by GOD, followed by hydrolysis to gluconic acid by lactonase. However, immobilization and separation of enzymes pose significant challenges, leading to high cost and time consumption, as well as enzyme denaturization.

Free-cell fermentation, on the other hand, subjects mycelia to various mass and heat-transfer stressors. Although mechanical agitation helps overcome these limitations, it causes turbulence, which can result in cell disintegration, fracture, and pellet breaking, leading to reduced cellular activity. Immobilizing cells on a support matrix under submerged conditions is a more efficient and cost-effective approach for gluconic acid fermentation. Despite these methods, a reliable and efficient process for gluconic acid production is still required.

In this context, this paper aims to develop a data-driven modeling framework for a fed-batch experimental bioreactor to maximize gluconic acid yield. By leveraging the vast amount of process data available, the goal is to create a reliable and portable model that engineers can use to better understand the process and optimize input process parameters.

$$Gluconic\ Acid\ Yield = \frac{Moles\ of\ gluconic\,acid\ produced}{Moles\ of\ glucose\ consumed} \tag{1}$$

2.1.2. Process flow diagram

A novel batch fermentation process has been developed for the production of gluconic acid from glucose, which utilizes immobilized *A. niger* on a cellulosic fabric support matrix to achieve higher yields. The increased productivity is attributed to the better interaction between dissolved oxygen and fungal mycelia. This approach uses a continuous substrate dripping mechanism instead of mechanical agitation, as in free-cell fermentation, to enhance the reaction. Thus, the yield of the bioreactor is critical for overall profitability of gluconic acid production.

2.1.3. Production objectives

The aim of this study is to develop a mathematical model for the novel batch fermentation process of glucose to gluconic acid and to determine the optimal operating parameters for improved gluconic acid production. The fermenter model was created based on experimental data that considered the effects of substrate (glucose), biomass, and dissolved oxygen levels. Figure 1 shows that the conversion of glucose to gluconic acid using *A. niger* immobilized on cellulosic microfibrils involves complex mass transport and reaction processes. Creating a phenomenological or first-principles process model has become challenging due to the limited understanding of the physicochemical processes that drive bioconversion and the associated kinetic and transport mechanisms. In addition, the process dynamics have been reported to be non-linear. Therefore, the goal of this study is to develop a data-driven model that can predict the system's behavior and optimize the process parameters to enhance gluconic acid production.

3. GP: AT A GLANCE

GP is an optimization technique that utilizes a symbolic approach and is considered as a type of metaheuristic. It involves the creation of equations or computer programs to solve a given problem, inspired by the concept of natural selection's "survival of the fittest" in the Darwinian theory of evolution. GP starts with an initial population of randomly generated solutions, which are then evaluated and selected based on their fitness. The selected solutions then undergo genetic operations such as crossover and mutation to create new offspring, which are again evaluated and selected based on their fitness. This process is repeated for multiple generations until an optimal solution is found, based on the fitness function. GP has been shown to be effective in solving various optimization problems, including data analysis, image processing, and control system design, among others [9]. The following is the general form of the reactor model to be obtained (Equation 2).

$$y=f\,(X,\,\beta) \tag{2}$$

where y indicates the process output variable (Selectivity or catalyst temperature); X is the N-dimensional vector of input variables such as flow, pressure, and inlet concentration of various raw materials ($X = [x_1, x_2., x_n., x_N]^T$), and f denotes a non-linear function whose parameters are defined in terms of a P-dimensional vector, $\beta\ [\beta_1, \beta_2, \ldots, \beta_K]^P$. If experimental data of input and output variables are given, GP algorithm tries to best fit the data by changing its functional form and parameter vector β.

3.1. Executional Steps of GP

The algorithm of GP has been illustrated in Figure 2. The executional steps of GP are mentioned as below [1]:

Step 1 (Initialization): The GP algorithm generates random equations to fit data in Equation 4, creating a population of strings (chromosomes) that represent candidate solutions. These members consist of functions and terminals organized hierarchically in a tree-like structure. The

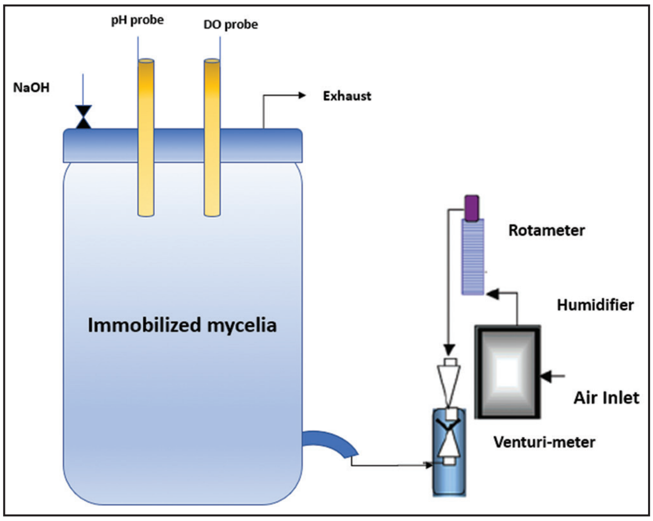

Figure 1: Experimental setup of the bioprocess [18].

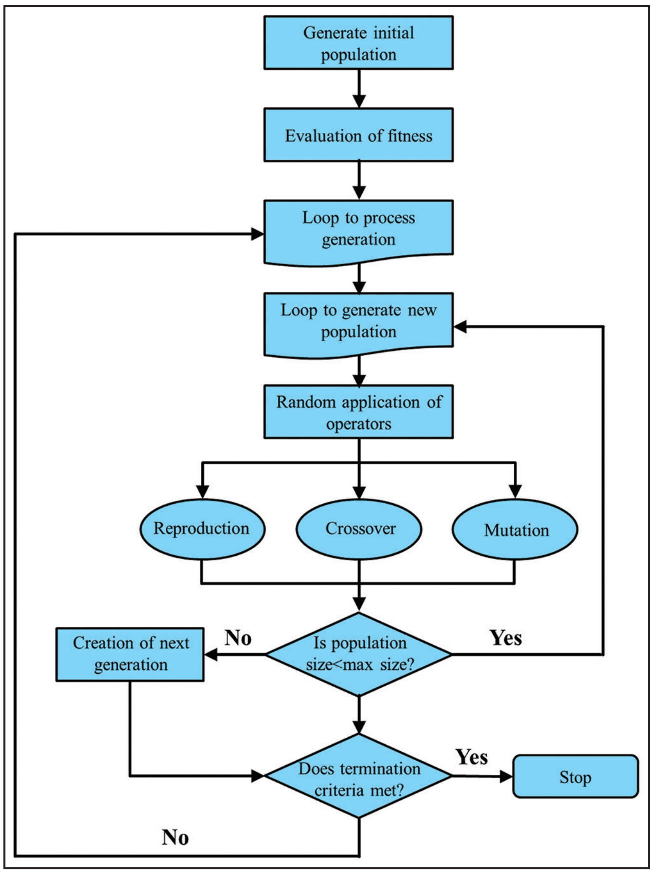
Generate initial population
Evaluation of fitness
Loop to process generation
Loop to generate new population
Random application of operators
Reproduction
Crossover
Mutation
Creation of next generation
No
Is population size<max size?
Yes
Does termination criteria met?
Yes
Stop
No

Figure 2: Algorithm of genetic programming.

function set includes algebraic and Boolean operators, while the terminal set consists of variables, numerical, and logical constants. Figure 3 provides an example of a typical tree structure.

Step 2 (Generation): This is an iterative procedure that aims to generate a population with a high fitness value. The steps involved are:

1. The fitness of each individual in the population is evaluated using a pre-specified fitness function. This function can be dependent on the coefficient of determination (R^2) or the error value. The higher the R^2 or the lower the error, the higher the fitness of an individual
2. Selection of individuals with high fitness is done through probabilistic determination
3. New individuals are created through genetic operators such as reproduction, crossover, and mutation. Reproduction involves copying the existing population without any changes, while crossover involves interchanging chromosomes of the parent generation to produce offspring. In mutation, existing elements in the offspring are replaced with other elements.

These steps are illustrated in Figure 3. The termination criteria for the algorithm are met when the best program is found as an approximate solution to the problem.

3.1.1. MGP

The standard GP is less accurate for symbolic regression of specified input output datasets, but the multi-gene strategy improves the model's accuracy to a greater extent. MGP creates a weighted linear combination of smaller GP trees or genes to enhance fitness. The anticipated output variable in MGP is the sum of the bias term and the weighted output of individual trees or genes. Equation 3 expresses the expected variables in MGP:

$$y_{pred} = b_0 + \sum_{i=1}^{N} w_i g_i \tag{3}$$

where y_{pred} is the predicted output, b_o is the bias term, g_i is the genes or trees, and w_i is the corresponding weightages. The weightages and bias are calculated using the least squares method, similar to linear regression. As a result, MGP utilizes tiny trees to capture nonlinear behavior while also employing traditional linear regression techniques [17,19].

In Figure 4, a typical MGP model is shown, which uses three input variables (x_1, x_2 and x_3) to predict the output variable. The MGP model is created by linearly combining two genes, as depicted in the figure. During training, the bias and weightages (b_0, w_1, and w_2) are determined using the least square approach. Users can set the maximum number of genes to be used and the maximum depth of a tree. It is important to note that increasing the depth and number of genes will result in higher accuracy, but it also increases the complexity of the model.

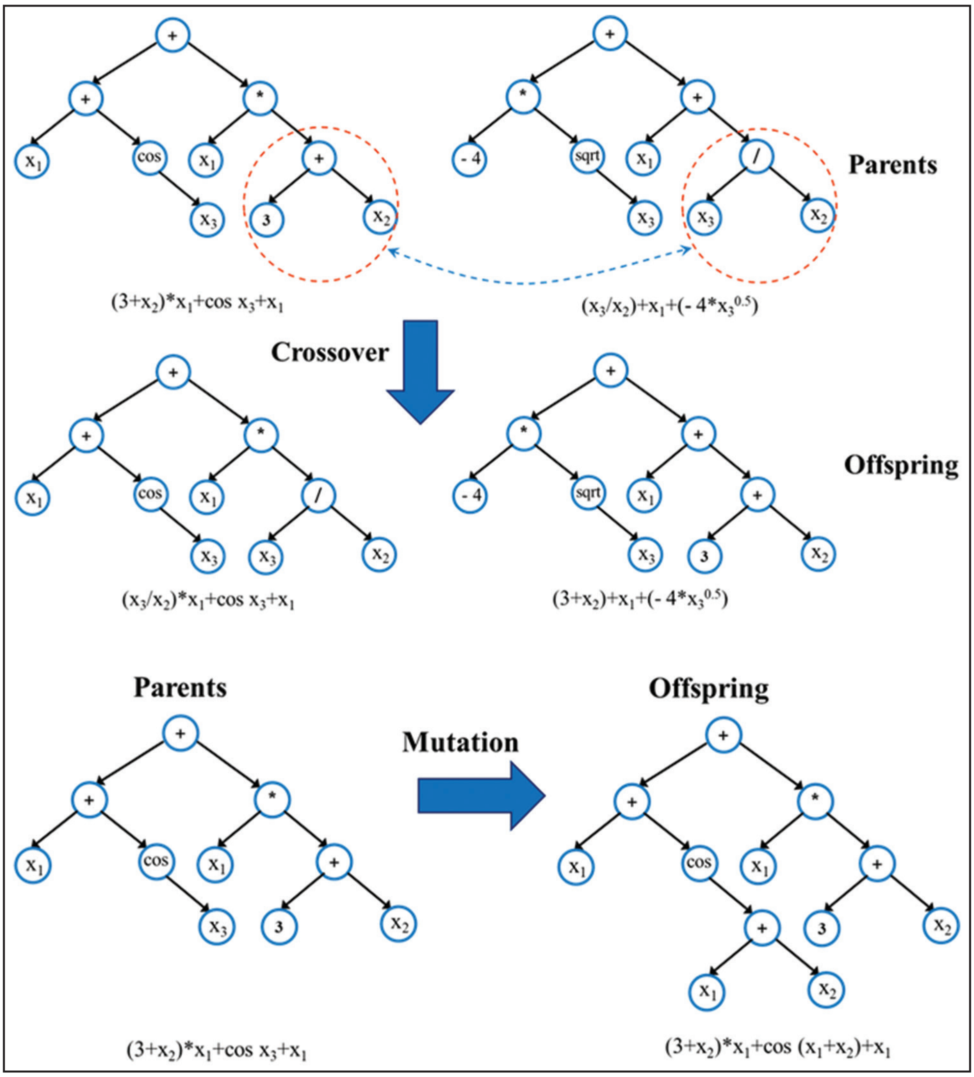
Parents
$(3+x_2)*x_1+\cos x_3+x_1$
$(x_3/x_2)+x_1+(-4*x_3^{0.5})$
Crossover
Offspring
$(x_3/x_2)*x_1+\cos x_3+x_1$
$(3+x_2)+x_1+(-4*x_3^{0.5})$
Parents
Mutation
Offspring
$(3+x_2)*x_1+\cos x_3+x_1$
$(3+x_2)*x_1+\cos (x_1+x_2)+x_1$

Figure 3: Crossover and mutation in GP.

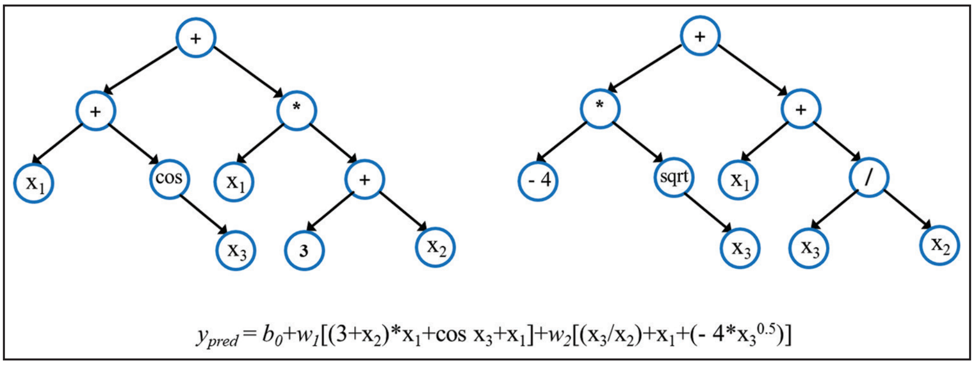
+
+
*
x1
cos
x1
+
x3
3
x2
+
*
+
- 4
sqrt
x1
/
x3
x3
x2
$y_{pred} = b_0+w_1[(3+x_2)*x_1+\cos x_3+x_1]+w_2[(x_3/x_2)+x_1+(- 4*x_3^{0.5})]$

Figure 4: A typical MGP model.

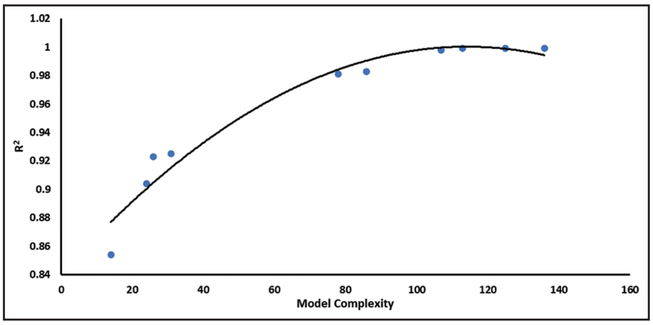

Figure 5: Pareto diagram of model complexity versus fitness for gluconic acid yield.

4. MATHEMATICAL MODELLING OF BIOPROCESS

4.1. Selection of Input and Output Variables for Modelling

Reactor yield is preserved as an output variable since it has such a substantial impact on total profitability. As a "wish list" of input variables, all reactor operational parameters that potentially affect yield are kept. All bioreactor experimental data were initially gathered. Then, after consulting with a technical specialist, all of the input variables that could affect the output variable were recorded. Following that, a cross-correlation study was performed. The correlation coefficients of each input variable with the output variable, as well as inter-input cross-correlation coefficients, were determined using this method [20,21].

The following criteria are used to shortlist the input variables:

1. For a particular input variable, there should be a high cross-correlation coefficient with the output variable
2. The values of cross-correlation coefficients of inter-input variables should be low
3. The input set of variables was kept as minimum as possible to avoid the complexity of the model.

Based on the above criteria, three input variables are finally shortlisted and tabulated in Table 1.

4.2. Data Collection, Data Cleaning and Removal of Outliers

The glucose to gluconic acid bioprocess GP-based model was constructed using experimental input-output data obtained from the fermenter, using the gluconic acid-producing *A. niger* NCIM 545 strain. The quality of data used in constructing data-driven models is crucial to the performance of the model. Thus, an automatic data cleaning technique was employed to ensure high-quality data. Due to the large amount of process data, an automated data cleaning method was developed to eliminate the need for manual cleaning. The data were pre-processed using multivariate PCA, and an automated MATLAB-based algorithm was developed to generate a multivariate statistical vector known as t-squared from the experimental operating dataset. Rows in the t-squared vector with values above the 95th percentile were regarded as outliers and were therefore removed from the dataset.

Table 1: Input output variables for model building and their range

Variables used in modeling	Data range
Input variables	
Glucose concentration, g/L (x_1)	100.0–180.0
Biomass concentration, g/L (x_2)	1.00–3.00
Dissolved oxygen concentration, mg/L (x_3)	10.0–60.0
Output variables	
Gluconic Acid Yield, % (Y_1)	5.9–94.58

It is important to note that noisy and faulty data can have a significant impact on the model's performance. Hence, data quality is a critical factor to consider in data-driven modeling. By utilizing an automated data cleaning technique, this study ensured that the data used to construct the GP-based model was of high quality.

4.3. Modeling through MGP

The study used MGP-based modeling with a dataset that contained three input variables and one output variable. The dataset was first cleansed using an automatic data-cleaning technique to eliminate outliers. The data were then randomly partitioned into a training set (80% of the total data) and a test set (20% of the total data) for the purpose of cross-validation. The MGP-based model was developed using the GPTIPS toolbox and MATLAB 2019a. The fitness function used in this study was the root-mean-squared error (RMSE), and the program was run to minimize the RMSE value. Since the MGP is stochastic in nature, the software was run 100 times to generate the model.

Cross-validation was employed to enhance the model's generalizability by testing its accuracy on the test data set. The goal of cross-validation is to assess how well the model can predict the outcomes of new data that it has not seen before. The model is trained on the training data set and then tested on the test data set to determine its predictive performance.

The use of MGP-based modeling allowed for the development of a highly accurate model for the glucose to gluconic acid bioprocess. The model's accuracy was evaluated using the RMSE, which measures the difference between the actual and predicted values. The study utilized a large dataset and employed an automatic data cleaning technique, which improved the quality of the data used to develop the model. Cross-validation was used to ensure that the model had good predictive performance on new data. Overall, the study demonstrated that MGP-based modeling is a highly effective technique for developing accurate models for complex bioprocesses.

4.4. Optimization through GA

After the development of a reliable and accurate bioreactor model, it is important to optimize the process parameters to achieve maximum profitability. This involves determining the optimal process conditions that will maximize reactor yield. To balance the two conflicting objectives, a multi-objective GA is used in the present study. GAs are a popular and effective optimization tool that have been widely applied in various fields, including engineering and medicine.

For implementation of GA algorithm, an objective function was developed which is as follows (Equation 4):

$$F_1(x) = 1/Y_1(x) \tag{4}$$

where $Y_1(x)$ is the function of the model corresponding to gluconic acid yield. Therefore, in MOGA $F_1(x)$ have to be minimized to maximize the yield of the bioprocess.

5. RESULTS AND DISCUSSION

5.1. Performance of GP Model

The primary objective of this study is to develop a straightforward, accurate, and transferable model equation for the gluconic acid fermenter. To create this model, MGP parameters were determined using a combination of trial-and-error methods and a literature review. The model was developed using basic arithmetic operators and functions, and the following parameters were set: A population size of 250, a maximum generation of 500, a maximum tree depth of 4, and a maximum number of genes of 6. It is important to note that while increasing these parameters may enhance the model's accuracy, it may also increase the complexity of the solutions and make the program computationally expensive. Using these parameters, the goal is to create a model that is both accurate and simple.

5.1.1. Developing closed form model equations

The primary objective of this study is to develop a straightforward, accurate, and transferable model equation for the gluconic acid fermenter. To create this model, MGP parameters were determined using a combination of trial-and-error methods and a literature review. The model was developed using basic arithmetic operators and functions, and the following parameters were set: A population size of 250, a maximum generation of 500, a maximum tree depth of 4, and a maximum number of genes of 6. It is important to note that while increasing these parameters may enhance the model's accuracy, it may also increase the complexity of the solutions and make the program computationally expensive. Using these parameters, the goal is to create a model that is both accurate and simple.

5.1.2. Controlling model complexity

Bloat, whether vertical or horizontal, is a common challenge when using multigene regression models. Vertical bloat refers to the tendency to evolve trees with terms that provide little or no performance benefit, which is associated with overfitting during model development. To address this issue, this study limited tree depth and used a Pareto tournament between expressional complexity and accuracy. Using the trial-and-error method, the tree depth was set to 4, and Pareto diagrams were formed to reduce vertical bloats [1].

Horizontal bloat occurs when multigene models acquire genes that are either performance neutral or offer extremely minor incremental performance improvements. This behavior is essentially the same as non-regularized models, where adding model terms results in a monotonically increasing R^2 on training data, even if the terms are not meaningful or do not allow the model to generalize well to testing or validation data sets. To avoid horizontal bloat in multigene regression, the simplest technique is to limit the number of genes allowed in the model. After using the trial-and-error procedure, the maximum number of genes in this study was kept at six.

The aim of the study is to create a gluconic acid fermenter closed-form model equation that is accurate, simple, portable, and easy to understand. The MGP parameters were determined using trial and error and a review of the literature. Basic arithmetic operators and functions were used, and the population size, maximum generation, maximum tree depth, and a maximum number of genes were set at 250, 500, 4, and 6, respectively. While increasing these parameters may improve the model's accuracy, it can also increase the complexity of the solutions and make the program computationally expensive. By balancing the accuracy and complexity of the model, this study developed a reliable and understandable model for optimizing the bioreactor process parameters to maximize reactor yield.

5.1.3. Shortlisting the models

The following criteria were used to pick a viable model from a pool of probable candidates or representative model equations (Table 2) with varied degrees of complexity and accuracy:

Simplicity: The model should have as little complexity as feasible.

Prediction accuracy: The constructed model should have a low RMSE and a high R^2 value.

The model equation should capture the process's underlying physics. In other words, model equations should contain a physical understanding of the system under study, not just a predictive association. This is a crucial factor to consider while creating realistic bioreactor models. To judge this capability, domain experts' qualitative knowledge about the bioreactor behavior is collected. Their theoretical knowledge, experience, and observations of bioreactor behavior are summarized in Table 3.

To evaluate the accuracy and effectiveness of the developed equations, a rigorous testing process was conducted. All ten equations listed in Table 4 were subject to scrutiny to ensure that they were consistent with experimental observations. The models were tested using ten separate datasets where all variables except glucose concentration were kept constant at their median values, while glucose concentration

Table 2: Rules to select best model from real experimental observations

Sl. no	Parameters changed keeping all other parameters constant	What happen to gluconic acid yield?
1	If glucose concentration increase	Increase
2	If biomass concentration increase	No change
3	If DO concentration increase	Increase

DO: Dissolved oxygen

Table 3: Performance of GP model

Model	Training			Testing		
	R^2	APE	RMSE	R^2	APE	RMSE
Gluconic acid yield model	0.998	0.169	0.232	0.984	0.483	0.186

GP: Genetic programming, R^2: Coefficient of determination, APE: Average percentage error, RMSE: Root-mean-squared error was varied between its minimum and maximum values in ten equal intervals. The resulting data sets were then input into each of the equations in Table 4 to generate ten sets of gluconic acid yield. The glucose concentration versus gluconic acid yield plots were then plotted, verifying observation 1 of Table 4.

This same process was repeated for all other observations in Table 4, with the resulting plots shown in Figure 6 for gluconic acid yield. This testing process was crucial to ensure that the developed models were accurate and effective in predicting the behavior of the bioreactor system. The models were also evaluated for simplicity, prediction accuracy, and their ability to capture the underlying physics of the system, as described in the previous section. Overall, the chosen model equation had the least complexity, the highest prediction accuracy (low RMSE and high R^2 value), and captured the physical understanding of the bioreactor system, as judged by domain experts' qualitative knowledge.

Any model equation that does not comply with the observations in Table 2 is rejected as it fails to capture the underlying physics of the bioreactor and is not consistent with general observations. These rejected models only represent a complex data fitting equation without any actual sense. Only one model equation for catalyst selectivity and one for reactor temperature were ultimately chosen from the shortlisted models, as mentioned in Equation 5. These two equations were considered as the representative models for selectivity and temperature, as they were highly accurate, adhered to the Table 2 observations, and captured the internal physics of the reactor.

$$Y_1 = 0.123\ (x_1^{3}\ x_3)^{1/2} - 0.00682\ x_1^{2}\ x_3^{1/2\,-}\ 12.6\ x_3 - 0.00288\ (x_3 - 1.0\ x_3^{1/2})^{3} + 12.6\ x_1^{1/2\,-}\ 171.0\ x_1^{1/4} + 12.6\ x_3^{1/2} + 0.286\ (x_3 - 1.0\ x_1^{1/2})\ (x_3 - 1.0\ x_3^{1/2}) + 334.0 \qquad (5)$$

The corresponding R^2 and average percentage error (APE) of the above model for training and test data are mentioned in Table 3.

The high values of R^2 and low values of APE obtained from the model for gluconic acid yield (Table 3) suggest that the predicted output values are in agreement with the actual output values, and the model

Table 4: Selectivity model equations: Expressional complexity/performance characteristics (on training data) of symbolic models on the pareto front

Model ID	Goodness of fit (R^2)	Model complexity	Model
2	0.999	113	$0.123\ (x_1^{3}\ x_3)^{1/2}$-$0.00682\ x_1^{2}\ x_3^{1/2}$-$12.6\ x_3$-$0.00288\ (x_3$-$1.0\ x_3^{1/2})^{3}$+$12.6\ x_1^{1/2}$-$171.0\ x_1^{1/4}$+$12.6\ x_3^{1/2}$+$0.286\ (x_3$-$1.0\ x_1^{1/2})\ (x_3$-$1.0\ x_3^{1/2}) + 334.0$
85	0.999	125	$0.125\ (x_1^{3}\ x_3)^{1/2}$-$0.00694\ x_1^{2}\ x_3^{1/2}$-$14.3\ x_3$+$14.3\ (2.0\ x_3$-$4.19)$ 1/2-0.0031 $(x_3$-$1.0\ x_3^{1/2})^{3}$+$14.3\ x_1^{1/2}$-$188.0\ x_1^{1/4}$+$0.313\ (x_3$-$1.0\ x_1^{1/2})\ (x_3$-$1.0\ x_3^{1/2}) + 370.0$
563	0.923	26	$1.22e$-$4\ x_1\ x_3^{2}$+$1.38\ (x_1\ x_3)^{1/2}$-$3.46e$-$4\ x_3^{3}$-38.2
581	0.983	86	$0.435\ x_1$+$0.0673\ (x_1\ x_3^{4})^{1/2}$-$0.00274\ x_1\ x_3^{2}$-$0.0635\ (x_3^{3}\ (x_3$+$9.0)\ (x_3$+$9.03))^{1/2}$+$0.00288\ x_3^{3}$-$46.0$
584	0.854	14	$6.84e$-$5\ x_1^{2}\ x_3$-$9.2e$-$5\ x_3^{3}$+11.9
627	0.904	24	$2.03\ x_3$+$0.00232\ (x_1^{4})^{1/2}$-$2.19e$-$4\ x_3^{3}$-44.9
873	0.998	107	$75.1\ x_3$-$0.0152\ x_{1+}0.0103\ x_1\ x_3$-$0.00381\ x_1\ x_3^{2}$-$8.7e$-$4\ x_1^{2}\ x_{3+}1.3\ x_1\ x_3^{1/2}$+$0.0712\ x_1\ x_3^{3/2}$-$36.1\ x_1^{1/4}\ x^{3}$-$0.00381\ x_1^{2}$-$86.0$
887	0.925	31	$5.18e$-$4\ x_1\ x_3$-$0.525\ x_3$-$4.36e$-$5\ x_1^{2}\ x_3$+$0.194\ x_1\ x_3^{1/2}$-53.8
894	0.981	78	$0.00197\ x_1\ x_3$-$32.4\ x_3$-$5.24e$-$4\ x_1$-$1.31e$-$4\ x_1\ x_3^{2}$-$1.66e$-$4\ x_1^{2}\ x_3$+$11.0\ x_1^{1/4}\ x_3$-$1.31e$-$4\ x_1^{2}$-2.99
1,131	0.999	136	$1.79\ x_3 + (40.3\ x_3^{3})/x_1^{2}$+$1.28\ (x_3^{2}\ (x_{1+}x_3$+$7.82))^{1/2}$-$43.5\ (x_3^{3}/x_1)^{1/2}$+$0.896\ x_3^{2}$-$0.0196\ x_3^{3/2}\ (x_1$+$x_{3+}20.4) - (0.326\ x_3^{2}\ (2.0\ x_3$+$7.91))/x_1$-$16.7$

R^2: Coefficient of determination

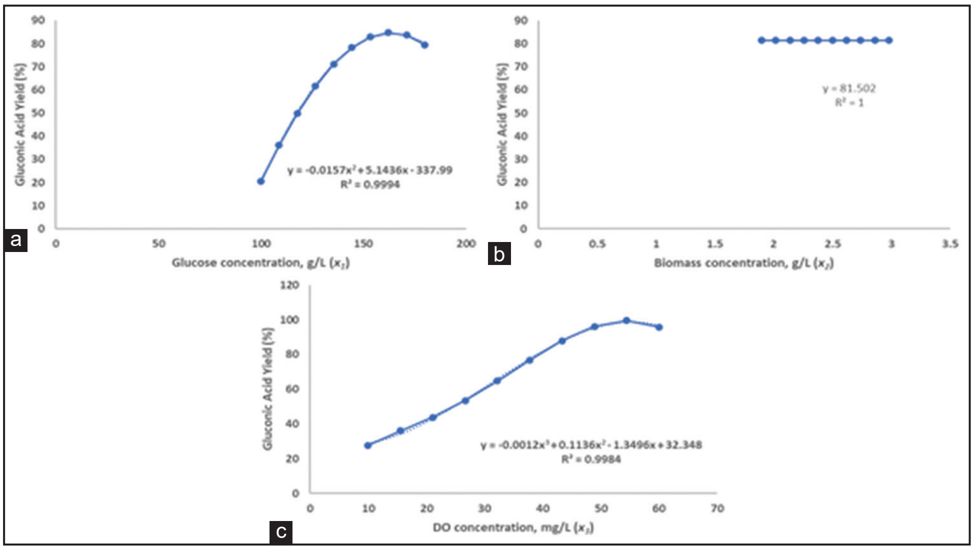

Figure 6: Influence of each variable on gluconic acid yield.

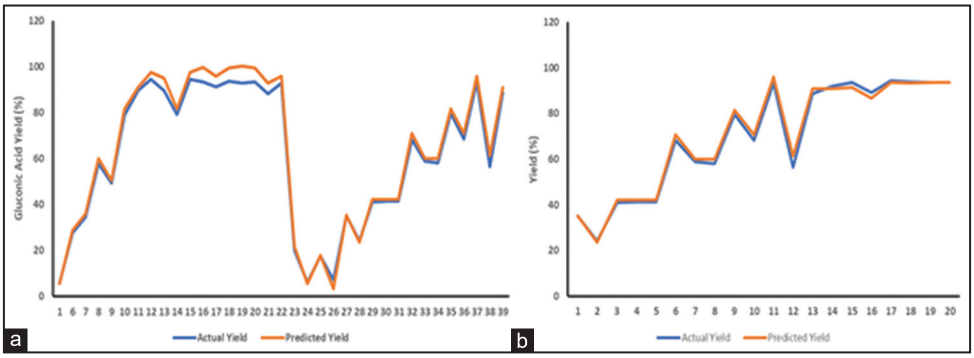

Figure 7: Actual versus predicted plots of (a) gluconic acid yield with training data (b) gluconic acid yield with testing data.

is reliable, reasonably accurate, and captures the underlying physics of the bioreactor. In addition, the high R^2 value on unseen test data and low APE indicate that the model can generalize well and accurately learn the non-linear input-output relationship. The performance of the models on both training and testing data is depicted in Figure 7, with the actual versus predicted curves being almost identical, indicating good prediction accuracy of the model.

From Table 3 and Figure 7, it is concluded that developed model is highly accurate and reliable as it also performs well with unseen test data.

5.1.4. Generation of explainable model equations

MGP modeling provides advantages over ANN and SVR approaches because it produces closed-form equations that are portable and implantable in distributed control systems. However, the created equations can be complex and difficult to interpret. To improve interpretability, a methodology was proposed in this study. The methodology involves adjusting one variable at a time from its lowest to highest value while keeping other input variables constant and using MGP's selectivity equations to predict selectivity values. A trendline is constructed based on the predicted values, and the equation and R^2 value are indicated in the figure. A trend line curve is chosen that closely matches the data, based on eye examination and R^2 value. The developed trend lines are decisive and match the actual experimental observations, following Table 2. The trend lines capture the nonlinear relationship between reactor selectivity and operating parameters, enabling experimental engineers to determine how input parameters impact gluconic acid yield. For example, increasing glucose and dissolved oxygen concentrations improve gluconic acid yield, while increasing biomass has no impact on yield. The relationship between yield and glucose and dissolved oxygen is represented by second and third-order polynomials, respectively. Finally, the trend line equations are used to develop the explainable Equation 6. The developed methodology enhances the interpretability of the MGP model, enabling experimental engineers to gain insight into how operating parameters impact gluconic acid yield.

$$Yield = -0.0157\left(x_1^2 - x_{avg}^2\right) + 5.1436\left(x_1 - x_{avg}\right) - 0.0012\left(x_3^3 - x_{avg}^3\right) + 0.1136\left(x_3^2 - x_{avg}^2\right) - 1.3496\left(x_3 - x_{avg}\right) + 81.50 \quad (6)$$

Where x_1, x_2, x_3 are the actual value of the 3 input variables and $x_{1,\ avg}$, $x_{2,\ avg}$, $x_{3,\ avg}$ are the average (50 percentile) value of input variables respectively.

Each term in the Equation 6 represents the change in selectivity if a particular parameter deviates from its average value. For example, the term, represents the deviation of glucose concentration from its average value and when it multiplied by coefficient 5.1436, represents the yield gain (or penalty) due to glucose. In this way, all three parameters contribution is calculated in Equation 6 and it is added with 81.5% (average yield) to get the actual yield.

Table 5: LB and UB for optimization

Bounds	x_1	x_2	x_3
LB	100	1	10
UB	180	3	60

LB: Lower bounds, UB: Upper bounds

Table 6: Optimal solution

x_1	x_2	x_3	Yield
162.89	1.07	60	99.6

The main advantage of this equation (Equation 6) over the GP equation (Equation 5) is that it may be easily understood by an experimental engineer. The equation is basic and incorporates terms or parametric coefficients that indicate the relative relevance of each parameter on overall yield if it differs from these base values. It also shows if each parameter's effect is linear or non-linear.

Equation 6 is then used to predict the selectivity of 3-year actual data and predicted, and actual yield is compared. The prediction error is 0.7% and R^2 is 0.98. This low value of prediction error and high value of R^2 signifies that the developed equation (Equation 10) is highly accurate and reliable.

5.2. Optimization through GA

After developing reliable models, the next step was to optimize the reactor operating parameters to achieve maximum yield. One of the crucial steps in this process is to fix the search space for finding the optimal process conditions. To achieve this, lower and upper bounds (UB) of the process variables were set in consultation with the experimental engineers. The lower bounds and UB for the process variables are shown in Table 5. These bounds are necessary to ensure that the optimization algorithm does not explore unrealistic or infeasible regions in the search space.

Using the GA tool in MATLAB, the optimal experimental conditions were determined which resulted in a 99.6% yield of gluconic acid, as shown in Table 6. The significant advantage of this study is that it offers experimental engineers a practical approach to operate the reactor at optimal conditions in real-time. Due to the absence of an explainable model, scientists had no idea about the optimal solution, and therefore, experimentalists had to rely on their experience and knowledge to optimize production heuristically. By running the GA with appropriate bounds in real-time, scientists can obtain a set of optimal operating conditions that they can set in the experiment, thereby minimizing the need for heuristic optimization (Figure 5).

6. CONCLUSION

In this study, the experimental data are utilized to build an accurate model for batch gluconic acid reactor through MGP. MGP provides a closed form model equation that is easily portable and useful in experimental analysis. The main contribution of this work is the development of an accurate and easily understandable model equation that offers valuable insights into the process. The created model equations are in line with scientific observations and adhere to the natural physics of the process. The use of these model equations enables the identification of the best solution to optimize yield, which ultimately ensures the maximization of profits.

7. REFERENCES

1. S. K. Lahiri, (2020) ***Profit Maximization Techniques for Operating Chemical Plants***, United States: John Wiley and Sons.
2. J. R. Koza. (1994) Genetic programming as a means for programming computers by natural selection. ***Statistics and Computing***, **4(2)**: 87-112.
3. J. R. Koza, J. P. Rice. (1992). ***Genetic Programming: The Movie***, Cambridge, MA: The MIT Press.
4. M. Willis, H. Hiden, M. Hinchliffe, B. Mckay, G. W. Barton, (1997) Systems modelling using genetic programming, ***Computers and Chemical Engineering***, **21(SUPPL 1)**: S1161-S1166.
5. E. G. Turitsyna, S. Webb, (2005) Simple design of FBG-based VSB filters for ultra-dense WDM transmission, ***Electronics Letters***, **41(2):** 40-41.
6. D. R. Lewin, (2005) Evolutionary algorithms in control system engineering. In: ***IFAC Proceedings Volumes (IFAC-Papers Online)***, **Vol. 16**. United States: IFAC.
7. S. Lakshminarayanan, H. Fujii, B. Grosman, E. Dassau, D. R. Lewin, (2000) New product design via analysis of historical databases, ***Computers and Chemical Engineering***, **24(2-7)**: 671-676.
8. B. Grosman, D. R. Lewin, (2004) Adaptive genetic programming for steady-state process modelling, ***Computers and Chemical Engineering***, **28(12)**: 2779-2790.
9. B. Grosman, D. R. Lewin, (2002) Automated nonlinear model predictive control using genetic programming. ***Computers and Chemical Engineering***, **26(4-5)**: 631-640.
10. D. Searson, M. Willis, G. Montague, (2007) Co-evolution of non-linear PLS model components, ***Journal of Chemometrics: A Journal of the Chemometrics Society***, **21(12):** 592-603.
11. D. P. Searson, D. E. Leahy, M. J. Willis, (2010) GPTIPS: An open source genetic programming toolbox for multigene symbolic regression. In: ***Proceedings of the International Multiconference of Engineers and Computer Scientists***, **Vol. 1**, Hong Kong: IMECS, p77-80.
12. I. Pan, D. S. Pandey, S. Das, (2013) Global solar irradiation prediction using a multi-gene genetic programming approach. ***Journal of Renewable and Sustainable Energy***, **5(6)**: 063129.
13. R. Barati, S. A. Neyshabouri, G. Ahmadi, (2014) Development of empirical models with high accuracy for estimation of drag coefficient of flow around a smooth sphere: An evolutionary approach, ***Powder Technology***, **257:** 11-19.
14. A. G. Floares, I. Luludachi, (2014) Inferring transcription networks from data. ***Springer Handbook of Bio-/Neuroinformatics***, Germany: Springer, p311-326.
15. A. H. Gandomi, A. H. Alavi, (2011) Multi-stage genetic programming: A new strategy to nonlinear system modeling, ***Information Sciences***, **181(23)**: 5227-5239.
16. S. K. Lahiri, S. Chowdhury, A. Hens, K. C. Ghanta, (2021) Modeling and multi-objective optimization of commercial ethylene oxide reactor to strike a delicate balance between profit and negative environmental impact, ***Environmental Science and Pollution Research International***, **29:** 20035-20047.
17. A. D. Mehr, V. Nourani, (2017) A Pareto-optimal moving average-multigene genetic programming model for rainfall-runoff modelling, ***Environmental Modelling and Software***, **92:** 239-251.
18. J. J. S. Cheema, N. V. Sankpal, S. S. Tambe, B. D. Kulkarni, (2002) Genetic programming assisted stochastic optimization strategies for optimization of glucose to gluconic acid fermentation. ***Biotechnology Progress***, **18(6)**: 1356-1365.
19. S. Pal, S. Chowdhury, A. Hens, S. K. Lahiri, (2022) Artificial intelligence based modelling and multi-objective optimization of vinyl chloride monomer (VCM) plant to strike a balance between profit, energy utilization and environmental degradation, ***Journal of the Indian Chemical Society***, **99(1)**: 100287.
20. S. Pal, S. K. Lahiri, (2022) Grey wolf optimizer trained ANN technique for development of explainable model of commercial ethylene oxide reactor and multi-objective optimization to maximize profit. ***International Research Journal of Engineering and Technology***, **9(11)**: 404-422.
21. S. Pal, S. K. Lahiri, (2023) Development of an explainable model for a gluconic acid bioreactor and profit maximization through grey wolf optimizer trained artificial neural network technique. ***International Research Journal of Engineering and Technology***, **10(3)**: 465-475.

**Bibliographical Sketch*

Sucharita Pal is a B.Tech graduate (2021) from NIT Durgapur from the Chemical Engineering Department. She is also the gold medalist for the year 2021. Currently, she is pursuing her Master of Science degree in Chemical and Petroleum Engineering from University of Calgary.

APPENDIX

Appendix Table 1: Experimental data utilized for building GP-based model taken from [18]

Batch no	Glucose concn (x1) (g/L)	Biomass concn (x2) (g/L)	DO (x3) (mg/L)	Gluconic acid yield (Y_1) (%)
1	100	1	10	5.9
2	150	2	10	29.42
3	120	2	15	20.76
4	150	2.5	15	35.51
5	150	3	15	35.16
6	120	2	25	27.77
7	120	2	30	34.48
8	150	2	30	57.86
9	150	3	25	49.32
10	150	2	40	78.99
11	150	2	45	89.48
12	150	2	50	94.5
13	180	2	50	89.63
14	150	3	40	79.05
15	150	2.5	50	94.58
16	150	2.5	55	93.41
17	150	2.5	60	91.26
18	160	2.5	60	93.67
19	175	3	55	92.69
20	160	3	60	93.3
21	180	3	60	88.13
22	150	3	60	92.7
23	100	3	60	20.04
24	100	2	10	6.13
25	120	2.5	10	17.58
26	100	2	15	7.2
27	150	2	15	35.09
28	120	2	20	24.12
29	150	2	20	40.99
30	150	2.5	20	41.33
31	150	3	20	41.25
32	150	2	35	68.22
33	150	2.5	30	58.82
34	150	3	30	58.03
35	150	2.5	40	79.61
36	150	3	35	68.38
37	150	2	60	93.4
38	120	2	60	56.3
39	150	3	45	88.63
40	180	2.5	55	91.94
41	150	3	50	93.68
42	180	2.5	60	89.09
43	150	3	55	94.5
44	166	3	60	93.82
45	165	3	60	93.53
46	162	3	60	93.54