SELECTED NONLINEAR MODELING AND STOCHASTIC OPTIMIZATION APPLICATIONS OF ARTIFICIAL INTELLIGENCE FORMALISMS IN CHEMICAL ENGINEERING AND TECHNOLOGY

A THESIS SUBMITED TO

SAVITRIBAI PHULE PUNE UNIVERSITY (SPPU)

FOR AWARD OF DEGREE OF

DOCTOR OF PHILOSOPHY (Ph.D.)

IN THE FACULTY OF **CHEMICAL ENGINEERING**

SUBMITTED BY

Ms. Vinadevi V. Patil

UNDER THE GUIDANCE OF

Dr. Sanjeev S. Tambe

CHEMICAL ENGINEERING AND PROCESS DEVELOPMENT DIVISION
CSIR-NATIONAL CHEMICAL LABORATORY
Dr. HOMI BHABHA ROAD
PUNE – 411 008, INDIA

December 2016

Certificate of the Guide

This is to certify that the work incorporated in the thesis titled **"Selected Nonlinear Modeling and Stochastic Optimization Applications of Artificial Intelligence Formalisms in Chemical Engineering and Technology"** submitted by **Ms. Vinadevi V. Patil**, for the degree of Doctor of Philosophy, in Chemical Engineering at Savitribai Phule Pune University (SPPU), Pune, was carried out by her under my supervision at Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune – 411 008, India. Such material has been obtained from other sources has been duly acknowledged in the thesis.

Date: December 14, 2016

Dr. Sanjeev S. Tambe
(Research Guide)

Declaration by the Candidate

I, hereby declare that the thesis entitled **"Selected Nonlinear Modeling and Stochastic Optimization Applications of Artificial Intelligence Formalisms in Chemical Engineering and Technology"** submitted by me for the degree of Doctor of Philosophy, in Chemical Engineering at Savitribai Phule Pune University (SPPU), Pune, is the record of work carried out by me during the period from December-2011 to December-2016 at Chemical Engineering and Process Development Division of CSIR-National Chemical Laboratory, Pune- 411 008, India, under the guidance of Dr. Sanjeev S. Tambe. The thesis has not formed the basis for the award of any degree, diploma, associateship, and fellowship, titles in this or any other University or other institutions of Higher Learning. I further declare that the material obtained from other sources has been duly acknowledged in the thesis.

Date: December 14, 2016 **Ms. Vinadevi V. Patil**

ACKNOWLEDGEMENT

'You may not realize it, but Artificial Intelligence is all around us'

This statement of Judy Woodruff, an eminent news anchor sums up the outcome of the deep association with my eminent guide **Dr. Sanjeev S. Tambe** during my research tenure in the unexplored area of artificial intelligence (AI). I am honored to have him as a mentor for my thesis. His accurate observations, critical analysis of a problem and clarity in content always drove me to achieve the goals with perfection. He taught me all the useful skills to carry out the research work in the area of AI, and also trained me in the technical presentation skills during the course of the study. His time management abilities, strong determination, exemplary commitment and positive attitude inspired me to attempt dizzying heights during this entire period. His creative thinking and administrative skills continue to inspire me a lot. Intense sessions of discussions converging to valuable suggestions and constant encouragement in stressful times during my Ph. D. work will be an integral part of my professional life. ***Working with him was a truly enriching and once in a lifetime experience***. ***Thank you Sir.***

I feel privileged to express my heartfelt thanks to **Dr. Ashwinikumar Nangia**, Director, and **Dr. Sourav Pal**, ex-Director, CSIR-National Chemical Laboratory for allowing me to carry out research work and extending all the required infrastructural facilities during my research stay in National Chemical Laboratory. I also thank **Dr. B. D. Kulkarni**, Distinguished Scientist, and CSIR-National Chemical Laboratory for his help and support. The entire library staff is gratefully acknowledged for providing unlimited access to the excellent facilities.

I owe my sincere thanks to my **father** for inspiring me to pursue higher studies and encouraging me to face all challenges with patience and zeal.

My special gratitude goes out to my late **mother**, who instilled in me a desire to learn, strong work ethics, and the value of hard work to achieve my goals.

I would like to highly acknowledge the help rendered to me at various stages by my loving and caring daughter **Anushka**. Her adapting disposition, sense of responsibility, and understanding nature assisted me to confront the distressful periods of this journey with courage. Her timely counseling helped me a great deal at various stages of this endeavor to boost my moral in difficult times.

I would like to express my gratefulness towards my husband **Anand**, whose love, patience and appreciation helped me to enjoy the journey towards this destination.

I express my heartfelt appreciation towards my sister **Vidula** for her caring attitude and for extending help in all possible ways in times of need. Special thanks to my sisters **Dr. Vaishali and Vijayamala**, brother **Virendra** for their moral support.

I would like to express my gratitude towards my parent organization, **Management Bharati Vidyapeeth Pune**, for permitting me to pursue my higher studies, and entire staff of **Department of Chemical Engineering**, for their support.

Last but not the least, I am thankful to my lab colleagues **Rahul Kulkarni, Purva Goel, Devendra Verma, Shishir Tiwari and Suhas Ghugare** for proffering their help at various stages of this endearing journey.

I would like to sum up my journey till now with a few words of Stephen Hawking.

'Success in creating Artificial Intelligence would be the biggest event in human history'

Thank you all once again.

Ms. Vinadevi V. Patil

Dedicated to my Parents (Akka and Dada)

and

my Guide

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

Mathematical reaction/process models are needed for a variety of tasks in chemical engineering and technology. These tasks include but are not limited to, equipment design, operation, and scale-up, prediction of steady-state and dynamic behavior, monitoring, control, fault detection and diagnosis and optimization. Conventionally, two approaches, namely, *phenomenological* (also termed "mechanistic" or "first principles"), and *empirical* (which comprise of regression methods), are used in chemical reaction/process modeling. Both these approaches suffer from several disadvantages especially when underlying reaction/process behavior is nonlinear, which is often the case in real practice. Another important chemical engineering task, namely, process optimization is traditionally conducted using deterministic gradient based methods. These methods also suffer from drawbacks such as entrapment into a local minimum.

The difficulties involved in the *phenomenological* and *regression-based* modeling and *deterministic* optimization techniques necessitated exploration of alternative nonlinear modeling and optimization strategies. In recent years, *Artificial Intelligence* (AI) based nonlinear modeling and stochastic optimization techniques owing to their several advantages have provided an attractive avenue for modeling highly nonlinear, complex multivariable systems as also optimization of chemical reactions and processes. Similar to AI, machine learning (ML) based modeling methods also possess certain attractive characteristics. Accordingly, in the present thesis, artificial intelligence and machine learning formalisms have been extensively employed to build exclusively data-driven models for tasks such as steady state and dynamic reaction/process modeling. The specific AI- and ML-based methodologies used in process modeling are *artificial neural networks* (ANNs), *genetic programming* (GP), and *support vector regression* (SVR). Additionally, an AI-based stochastic method, namely, *genetic algorithms* (GA) has been used for optimizing a chemical process. Apart from the stated AI-based methods, conventional mathematical methods such as *principal component analysis* (PCA) and *sensitivity analysis* have been used for conducting dimensionality reduction and identifying influential causal (input/independent) variables, respectively.

Notable features of the studies presented in the thesis are:

- Artificial intelligence and machine learning methods have been comprehensively used for modeling coal gasification pilot plant process; the coals used in gasification are high ash Indian coals.
- It has been clearly demonstrated that the genetic programming technique while searching and optimizing the form and associated parameters of an appropriate linear/nonlinear data-fitting function, also identifies those inputs which significantly influence the model output.
- An entirely AI-based hybrid methodology integrating GP and GA formalisms has been employed for modeling and optimization of resin-based adsorptive removal of toxic metal ions from contaminated water.
- The GP strategy has been utilized for an accurate prediction of API gravity values of crude oils. The nonlinear model developed uses SARA composition of crude oils to predict the API gravity magnitudes.
- In a first of its kind of study, genetic programming has been employed to develop models for VLE prediction, where it has been shown that a single GP model under certain conditions can predict VLE of multiple binary systems.

This thesis is divided into nine chapters. A brief description of these chapters is provided below.

Chapter 1 gives a bird's eye-view of the significance of the work reported in the thesis. It also presents information about the conventional modeling and optimization techniques, and difficulties encountered thereof. The chapter next presents salient features of the AI-based modeling and optimization strategies and their generic application areas in chemical engineering and technology.

Chapter 2, first describes in detail the various AI-based formalisms utilized in the various studies reported in the thesis, such as *multilayer perceptron* (MLP) *neural network*, *genetic programming* (GP), *support vector regression* (SVR) and *genetic algorithms* (GAs). This chapter also provides a description the conventional mathematical techniques, namely, *principal component analysis* (PCA) and *sensitivity analysis*, which have been used for performing dimensionality reduction, and identifying influential causal (input/independent) reaction/process variables, respectively. Additionally, statistical measures, namely, *coefficient of correlation*

(*CC*), *root mean squared error* (*RMSE*), and Steiger's test that has been used for the evaluation and comparison of the prediction and generalization performance of the AI and ML-based models are explained in the chapter.

Chapter 3 reports study a wherein data were collected from extensive gasification experiments conducted in a pilot-plant scale fluidized-bed coal gasifier (FBCG)—located at CSIR-Central Institute of Mining and Fuel Research (CIMFR), Dhanbad, India—using high-ash Indian coals. Specifically, the effects of eight coal and gasifier process related parameters on the four gasification performance variables, namely $CO+H_2$ *generation rate*, *syngas production rate*, *carbon conversion*, and *heating value of the syngas*, were rigorously studied. The data collected from these experiments were used in the FBCG modeling, which was conducted by utilizing two artificial intelligence (AI) strategies namely *genetic programming* (GP) and *artificial neural networks* (ANNs). The original eight-dimensional input space of the FBCG models was reduced to three-dimensional space using principal component analysis (PCA), and the PCA-transformed three variables were used in the AI-based FBCG modeling. A comparison of the GP and ANN-based models reveals that their output prediction accuracies and the generalization performance vary from good to excellent as indicated by the high training and test set correlation coefficient magnitudes. This study also presents results of the sensitivity analysis performed to identify those coal and process related parameters, which significantly affect the FBCG process performance.

Chapter 4 reports development of the data-driven models for the gasification of chars derived from the high ash coals. Specifically, the models predict two important gasification performance parameters, viz. *gasification rate constant* and *reactivity index*. These models have been constructed using three computational intelligence (CI) methods, namely *genetic programming* (GP), *multilayer perceptron* (MLP) neural network (NN), and *support vector regression* (SVR). The inputs to the CI-based models consist of seven parameters representing the gasification reaction conditions and properties of high ash coals and chars. The data used in the modeling were collected from the extensive gasification experiments. These were performed in the CO_2 atmosphere in a thermo-gravimetric analyzer (TGA) using char samples derived from the Indian coals containing high ash content. Values of the two gasification performance parameters were obtained by fitting the experimental data to

the shrinking unreacted core (SUC) model. It has been observed that all the CI-based models possess an excellent prediction accuracy and generalization capability. Accordingly, these models can be gainfully employed in the design and operation of the fixed and fluidized bed gasifiers using high ash coals.

In Chapter 5, a GP-based strategy has been suggested for (a) simultaneously identifying the important predictor (independent/causal/input) variables that significantly influences the output (dependent variable) of an input-output model, and (b) searching and optimizing an optimal data fitting function and its parameters. The said strategy has been illustrated by conducting two process identification case studies wherein the GP formalism has been shown to (i) identify the influential time-delayed inputs and outputs, and (ii) simultaneously perform system identification using these influential predictors. The two chemical engineering systems chosen in the case studies are nonlinear height control system for a conical tank, and nonlinear adiabatic CSTR concentration control system. It is noticed from the GP-based models obtained in these case studies that although the data supplied to the GP algorithm contained six predictor variables, it searched and optimized models with only four predictor variables; noticeably, these predictors were identified by the sensitivity analysis to be having most influence on the model output. The GP-based system identification strategy suggested here—being computationally economical and much less tedious—has the potential to become an effective alternative to the conventionally used linear/nonlinear identification strategies. Having identified a process using the GP strategy the corresponding model can be gainfully utilized to implement the model predictive control (MPC) strategy.

Chapter 6 presents, the API gravity (^{o}API) is an important physicochemical characteristic of crude oils and often used in determining their properties and quality. There exist models— predominantly linear ones —for predicting the ^{o}API magnitude from the molecular composition of crude oil. This approach is tedious and time-consuming since it requires quantitative determination of numerous crude oil components. Usually, the hydrocarbons present in the crude oils are grouped according to their molecular average structures into *Saturates*, *Aromatics*, *Resins* and *Asphaltenes* (SARA) fractions. An ^{o}API prediction model based on these four fractions is relatively easier to develop although this approach has been rarely utilized. A rigorous scrutiny suggests that some of the dependencies between the

individual SARA fractions and the corresponding ^{o}API magnitude could be nonlinear. Accordingly, in this study, SARA fractions based nonlinear models have been developed for the prediction of ^{o}API magnitudes using three computational intelligence (CI) formalisms, namely, *genetic programming*, *artificial neural networks* and *support vector regression*. The SARA analyses and API-gravity values of 403 crude oil samples covering wide ranges have been utilized in developing these models. A comparison of the CI-based models with an existing linear model indicates that all the former class of models possesses a significantly better ^{o}API prediction and generalization performance than that exhibited by the linear model. Also, the SVR-based model has been found to be the most accurate API gravity predictor. Owing to their better prediction accuracy, CI-based models can be gainfully used to predict ^{o}API values of crude oils.

In Chapter 7, a computational intelligence (CI) based hybrid strategy was employed to model and optimize, tannin-formaldehyde (TFA) and tannin-aniline-formaldehyde (TAFA) resin-based adsorption of arsenite [As(III)] and arsenate [As(V)] ions for securing optimal reaction conditions. This strategy first uses an exclusively reaction data driven modeling strategy, namely, *genetic programming* (GP), to predict the extent (%) of As(III)/As(V) adsorbed on the TFA and TAFA resins. Next, the input space of the GP-based models consisting of reaction condition variables was optimized using *genetic algorithm* (GA), which is an artificial intelligence based stochastic nonlinear optimization method; the objective of this optimization was to maximize the adsorption of As(III) and As(V) ions on the two resins. Finally, the sets of the optimal reaction condition variables provided by the GP-GA hybrid method were verified experimentally. The verification results indicate that the optimized conditions have lead to 0.3% and 1.3% increase in the adsorption of the As(III) and As(V) ions respectively on the TFA resin. More significantly, the optimized conditions resulted in an improvement of 3.02 % in the adsorption of As(III), and 12.77% in the adsorption of As(V) on the TAFA resin. The GP-GA hybrid strategy employed in this study can be gainfully utilized for modeling and optimization of similar type of contaminant-removal processes.

Chapter 8 presents, a study wherein *genetic programming* (GP) has been introduced for the prediction of VLE. Specifically, four case studies have been performed wherein seven GP-based VLE models have been developed using

experimental data for predicting the *vapor phase composition*, (y_i) of a ternary and groups of non–ideal binary systems. The input space of these models consists of three attributes of pure components (*acentric factor*, *critical temperature*, and *critical pressure*), and three intensive thermodynamic parameters (*liquid phase composition*, *pressure*, and *temperature*). The prediction and generalization performance of the GP-based models was rigorously compared with that of the corresponding conventionally employed Van Laar, NRTL, and UNIQUAC models. The results obtained thereby indicate superior prediction accuracy and generalization performance of the GP-based models vis-a-vis that of the conventional thermodynamic models. The GP-based modeling method proposed in this study can be gainfully utilized in the prediction of VLE as also designing corresponding experiments at different pressure and temperature ranges.

Chapter 9 gives an overview of the important results presented in this thesis and the conclusions drawn thereof. Directions for future research are also presented in this chapter.

Chapter 1

Introduction

ABSTRACT

Modeling and optimization of chemical reactions and processes is an important activity in chemical engineering/technology. It assists in the prediction of reaction/process behavior, equipment design and scale-up, process operation and monitoring, control, etc. There exist conventional methods for conducting modeling and optimization of chemical reactions and processes. These have certain deficiencies. Accordingly, in the present thesis, artificial intelligence and machine learning based formalisms have been used for modeling and optimization of a number of reactions and processes. This chapter outlines the currently used principle reaction and process modeling and optimization methods, and the need for newer approaches thereof. Additionally, the chapter presents an overview of the contents of the subsequent chapters.

1.1 MOTIVATION FOR THE THESIS

For designing and operating a chemical process and carrying out related tasks, it is necessary to understand its behavior completely. Conducting experiments for getting an insight in to the process behavior is often an expensive, complicated, tedious, and a time-consuming proposition. These difficulties can be overcome if a representative process model is available. The objective of mathematical modeling has been stated as (Constantinides, 1987) — "to construct, from theoretical and empirical knowledge of the process, a mathematical description, which can be used to predict the process behavior." The mathematical model of a chemical process provides—over specific ranges of operating variables and parameters—quantitative information on the process behavior; it describes at least the major features of the chemical and physical mechanisms underlying the process. The process behavior described by mathematical models mainly includes steady-state, dynamic, and spatiotemporal phenomena. A properly constructed model of a process (physical, chemical, biological, and biochemical) can be used to predict its behavior under different operating scenarios. In chemical engineering practice, an accurate, robust, and reliable mathematical process model assists in the preliminary process design, complex simulation, prediction of the steady-state and dynamic behavior, startup, shutdown, scaling up, process monitoring, model based control, fault detection and diagnosis, and process optimization.

Chemical processes comprise a set of unit operations, and reactors that convert raw materials into desirable products through physicochemical conversions. Modern day chemical processes are highly complex with a multitude of interconnected equipments, sensors, and control systems. Consequently, a large number of variables and parameters associated with these systems interact with each other thereby making design, operation, control, and analysis of processes a difficult task. Notwithstanding these difficulties, it is at most necessary that operation of chemical processes is safe, robust, efficient, commercially viable, and environment friendly. In the modern times of advanced software and hardware technologies, the stated goal is achieved via computer-aided chemical process design, operation, control, simulation, and optimization. It helps in (a) reducing the time lag between process innovation and its commercial implementation and exploitation, (b) ensuring efficiency, safety,

competitiveness, and flexibility of new chemical plants, and (c) improving operational efficiencies of existing plants.

There exist various types of models, namely, phenomenological, empirical, black-box, stochastic, statistical, Monte-Carlo, cellular automata, etc. Each one of these possesses advantages and drawbacks. In the last three decades, a new class of modeling paradigm that uses various artificial intelligence (AI) and machine learning (ML) formalisms is being increasingly utilized in chemical process modeling. Models belonging to this class have several attractive properties. Accordingly, the principal motivation of this thesis is to explore selective AI and machine learning formalisms for modeling chemical reactions and processes. Additionally, an AI-based method is employed for conducting chemical process optimization. The present chapter provides (a) a bird's eye-view of the AI and ML based modeling and optimization formalisms used in conducting the studies described in the subsequent chapters, and (b) an overview of the contents of chapters 2 to 9 of this thesis.

1.2 PROCESS ENGINEERING TASKS

The principal tasks encountered in chemical engineering and technology that involves development of models are described below in brief.

(a) Prediction of steady-state and dynamic process behavior

A reaction or a process essentially displays two types of behavior, namely *steady-state* and *dynamic*. The corresponding mathematical models are categorized similarly. In the case of former, after an initial transient behavior, reaction proceeds at a constant (steady) rate. In a dynamic state, however, reaction behavior is not steady and varies with time. When a reaction/process reaches a *steady* (static) *state*, its operating variables reach a constant value and do not vary unless an external force is applied. Steady-state modeling is particularly useful in design calculations. When a process is in a dynamic state, its operating variables exhibit time-dependent variations. Dynamic process models are crucial in getting a comprehensive view of the reaction/process/plant behavior.

(b) Process optimization

Process optimization aims at determining optimal values of operating variables/parameters for securing a desirable performance, such as better product

quality, higher conversion, improved (lower) selectivity for the desired (undesired) product, minimum operating cost, and profit maximization. It can also help in ensuring a safe, cost-efficient, and environment friendly process operation. Availability of a process model is a pre-requisite for process optimization.

(c) Model based process control

For any chemical process, influential parameters and variables need to be manipulated (controlled) for achieving the desired process performance. This is achieved by implementing a process control mechanism. It is a critical engineering activity in any chemical process since it ensures a safe, economical, and environ-friendly process operation. Using the knowledge of the process's steady-state and dynamical behavior, process control maintains the magnitude of a specific process variable within a desired range. For instance, the temperature of a chemical reactor may be controlled to maintain a consistent product output or conversion. The conventional proportional-integral-derivative (PID) control strategy does not explicitly take into account the process model. In general, model-based control is found to yield better performance than the PID scheme especially for controlling nonlinear systems.

(d) Process monitoring

In order to deliver quality products, critical process operating variables should precisely follow their specified trajectories. Process variability can be reduced by employing an efficient monitoring strategy. Such a system, based on a dynamic process model and functioning online is capable of quickly identifying any abnormal process behavior so that corrective measures can be taken swiftly. It also helps in diagnosing process faults.

(e) Process identification

Process identification is necessary in implementing model-based process control. It involves development of empirical input-output models and thereby identifying dynamic process behavior. Given past and present values of the manipulated and controlled variables, this type of model typically predicts the single/multistep ahead magnitude of the manipulated variable.

(f) Quantitative Structure—Activity/Property Relationships (QSAR/QSPR)

QSAR/QSPR represents a relationship (model) between the structural parameters of a molecule and its activity/property. These relationships are unquestionably of great importance in modern chemistry and its sub-disciplines. Once a correlation between structure and activity/property is developed, any number of compounds, including those not yet synthesized, can be readily screened; it helps in screening specific structures (molecules) possessing the desired activity/property. In the next step, the screened compounds are synthesized and tested in the laboratory. Thus, the QSAR/QSPR approach conserves resources and accelerates development of new molecules for use as drugs, materials, additives, or for any other purpose (Karelson et al., 1996).

(g) Fault detection and diagnosis (FDD)

A properly and timely detection and diagnosis of occurrence of faults in chemical plants of all sizes, assumes greatest importance from the viewpoint of personnel and equipment safety. Of concern is also the monetary loss incurred during the short- and long–term plant shut-downs owing to an equipment malfunction or a failure. Since it directly helps to prevent any impending hazardous situation, process fault detection and diagnosis (FDD) has become an integral part of the process design and operation activity. The task of FDD is greatly simplified if a process model is available (or can be developed) since various equipment fault and malfunction scenarios can be simulated using models.

(h) Soft-sensors

In the absence of hardware sensors, product analysis is conducted in the quality control laboratory using instrumental and chemical methods. Some analyses are tedious, and time-consuming. Consequently, the plant continues to produce off-spec product during the time taken for the chemical analysis. This difficulty is overcome by developing soft-sensors. These are software based sensors (mathematical models), which given the information about the current values of process variables, can predict the values of the quality control variables. The soft-sensor models are developed using historical data of process variables and parameters, and the corresponding values of quality control variables determined via instrumental and/or chemical methods. When operated in the prediction mode, soft-sensor models are capable of predicting the values of quality control variables almost instantaneously.

(i) Data mining

Process operation over time generates huge amounts of data regarding, for example, process operating variables and parameters and the corresponding conversions, yields, selectivities, and quality control variables. These data contain wealth of information and knowledge hidden in them. Data mining is a non-trivial task of identifying valid, novel, potentially useful, and ultimately understandable patterns within process data. It is the process of extracting previously unknown comprehensible and actionable information from large databases and using it to make crucial process/business decisions (Provost and Fawcett, 2013). One of the important elements of data mining is to develop models unraveling hidden relationships between different process variables and/or parameters and the corresponding product specific and process performance attributes.

1.3 CONVENTIONAL PROCESS MODELING TECHNIQUES

A great deal of effort has been spent over the last several decades towards mathematical modeling of chemical processes. Availability of an accurate process model is essential for predicting the process behavior under wide-ranging input conditions. Commonly, two approaches, namely, *phenomenological* and *empirical* are employed for modeling a chemical process.

1.3.1 Phenomenological Modeling

The phenomenological (also termed *first principles* or *mechanistic*) models rigorously account for the reaction mechanism, mass and heat transport phenomena, and thermodynamics associated with the chemical process under consideration.

Principal advantages of the phenomenological modeling approach

- Provides valuable insight into the process behavior.
- Model can be used in extrapolated regions of the input space.
- Can be used in process scale-up.
- Since these represent physico-chemical phenomena underlying a process, first principles models provide an insight into the intrinsic phenomena responsible for process behavior.

Difficulties encountered in phenomenological modeling

Being inherently complex and nonlinear, many chemical processes are difficult to model phenomenologically. Specific difficulties encountered in the phenomenological modeling of the chemical processes are:

- Most chemical processes witness existence of multiple nonlinear interactive relationships between process variables and parameters.
- Cost-intensive and exhaustive experimentation is required for studying the effects of influential process operating variables and parameters on the process behavior.
- Often, there exists insufficient knowledge of the physicochemical phenomena (e.g., reaction kinetics, heat and mass transport mechanisms and thermodynamics) underlying a process and, thus, extensive effort is needed to arrive at a reasonable model**.**

In view of the difficulties associated with the phenomenological modeling, it becomes necessary to explore alternative modeling approaches. One such practical option is the development of exclusively data-driven models. Commonly, data-driven models are developed using *empirical* (regression) methods.

1.3.2 Empirical Modeling

Empirical models, sometimes also termed as “black-box” models provide a convenient alternative to *first-principles* models. In mathematical modeling, when the primary goal is the most accurate replication of data, regardless of the mathematical model structure, a black-box modeling approach is useful (Sjöberg et al., 1995). In conventional empirical modeling, process behavior is modeled using appropriately chosen empirical equations, for example, polynomial or multivariable linear/nonlinear expressions. This procedure termed *regression* uses a heuristic procedure wherein an appropriate functional form that possibly fits the process data is selected in advance following which the unknown function parameters are estimated using a suitable parameter estimation method. Since several efficient linear/nonlinear parameter estimation methods are available, the real difficult part in empirical modeling is specification of the model structure. For linear systems model specification is easy; however for a nonlinear systems it poses significant difficulties

since it involves selecting an appropriate model structure from the numerous competing ones (Verma et al., 2016).

Principal advantages of the empirical modeling approach

- Appropriate linear or nonlinear models are fitted exclusively from the process data containing values of dependent and independent variables and parameters.
- The detailed knowledge of physico-chemical phenomena underlying the process is not needed.

Difficulties encountered in empirical modeling

- The exact form of the data-fitting function needs to be specified before parameters associated with it can be estimated. This is a difficult task that requires a "trial-and-error" approach since very often a number of variables nonlinearly influence the process behavior and the precise interactions between them are not known.
- Mostly provide correct predictions over a limited range of the process data used in developing the model.
- In general can not be used for extrapolation.
- Large amounts of statistically well distributed data are needed to develop an empirical model possessing good prediction accuracy and generalization capability.

1.4 ARTIFICIAL INTELLIGENCE (AI)-BASED PROCESS MODELING TECHNIQUES

Artificial intelligence (AI) is a branch of computational science, which develops mathematical algorithms mimicking various kinds of intelligent behavior exhibited by the biologically evolving species with the aim of providing novel and efficient solutions to complex modeling, classification and optimization problems (Fogel, 2006). Stated differently, AI is essentially concerned with the development of algorithms and techniques, which allow computers to "learn" and utilize this knowledge to solve problems such as function approximation, classification, image and speech recognition and clustering. The AI, however, does not have to confine itself to methods that are observed only in the nature. Accordingly, often *machine*

learning (ML) algorithms are also considered to be part of AI. Unlike AI, the ML algorithms are not based on the intelligent behavior observed in nature although their working can be termed "intelligent." Both AI and ML-based modeling formalisms are exclusively data-driven and their performance critical depends upon the quality and quantity of the data. In the following, major AI and ML-based modeling methods are described in brief.

(a) Artificial Neural Networks (ANNs)

Artificial neural networks (ANNs) (Freeman and Skapura, 1996) are an information-processing paradigm founded on the mechanisms followed by the highly interconnected cellular structure of the human brain. They basically simulate the brain's lower level mechanisms, such as, learning, pattern recognition, pattern association, generalization, and self-organization (Tambe et al., 1996). ANNs are a black-box empirical modeling paradigm where process modeling is possible solely based on the historic process input-output data. The commonly employed ANNs for modeling purposes, such as multilayer perceptron (MLP) and radial basis function (RBF) neural network utilize a generic nonlinear function as a building block of the function to be approximated and, thus, the troublesome task of specifying the form of the fitting function gets completely eliminated. ANNs possess certain added advantages such as amenability to parallel processing, due to which process modeling becomes easier, less cumbersome, and faster compared to the phenomenological modeling approach.

(b) Genetic Programming (GP)

There exists a novel member of the evolutionary algorithms family, namely, *genetic programming* (GP) (Koza, 1992) that in its original form provided a method for automatically creating computer programs that perform pre-specified tasks simply from a high-level statement of the problem. Genetic programming follows Darwin's theory of biological evolution comprising "survival of the fittest" and "genetic propagation of characteristics" principles. It addresses the goal of automatic generation of computer programs by: (i) genetically breeding a random population of computer programs, and (ii) iteratively transforming the population into a new generation of computer programs, by applying analogs of nature-inspired genetic operations, namely, *selection*, *crossover,* and *mutation* (Vyas et al., 2015). Another

important application of GP termed “symbolic regression” is in data-driven modeling, which has been extensively explored in this thesis. The novel aspect of GP when used in modeling is that given an example input-output data set, the method is capable of automatically obtaining an appropriate linear/nonlinear data-fitting function and its parameters.

(c) Fuzzy Logic (FL)

Fuzzy Logic is a systematic mathematical formulation for investigating and characterizing different types of uncertainties (Tootoonchy and Hashemi, 2013). It is best suited when a mathematical model of the process either does not exist, or exists but is too complex to be evaluated fast enough for a real time operation, or is too difficult to encode, when data are imprecise and noisy. It was popularized by Lotfi Zadeh in the sixties (Zadeh, 1965). It is based on the premise that Boolean logic, represented by 0 and 1, does not adequately represent imprecise or fuzzy information. Fuzzy logic uses membership functions having values between 0 and 1. The degree of membership allows an object in a set to be anywhere in the range of 0 (completely not in the set) to 1 (completely in the set), thus permitting to deal with uncertain situations naturally (Bose, 1994). The values of fuzzy variables are expressed with English words such as *cold*, *warm*, *hot* or *weak*, *medium*, and *high*; each of these is defined by a suitable (e.g., Guassian, triangular, or trapezoidal) membership function. In contrast to the abrupt changes from 0 to 1 in Boolean logic, the membership functions allow gradual variations in the variables. FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy or missing input information. This unique ability of FL has been utilized to model complex nonlinear processes where development of a suitable phenomenological mathematical expression becomes difficult (Mendel, 1995).

1.5 MACHINE LEARNING (ML)-BASED PROCESS MODELING TECHNIQUE: Support Vector Regression (SVR)

The SVR (Vapnik, 1995; Burges, 1998) is a regression analog of the statistical machine learning theory based classification paradigm, namely, *support vector machines* (Vapnik, 1995). It is a linear method in a high-dimensional feature space that is nonlinearly related to the input space. SVR formalism possesses some desirable characteristics, such as good generalization ability of the regression function, robustness of the solution, sparseness of the regression, and an automatic control of

the solution complexity. It also provides an explicit knowledge of the data points that define the regression function. This feature assists in interpreting an SVR-based model in terms of the training data.

1.6 CONVENTIONAL PROCESS OPTIMIZATION TECHNIQUES

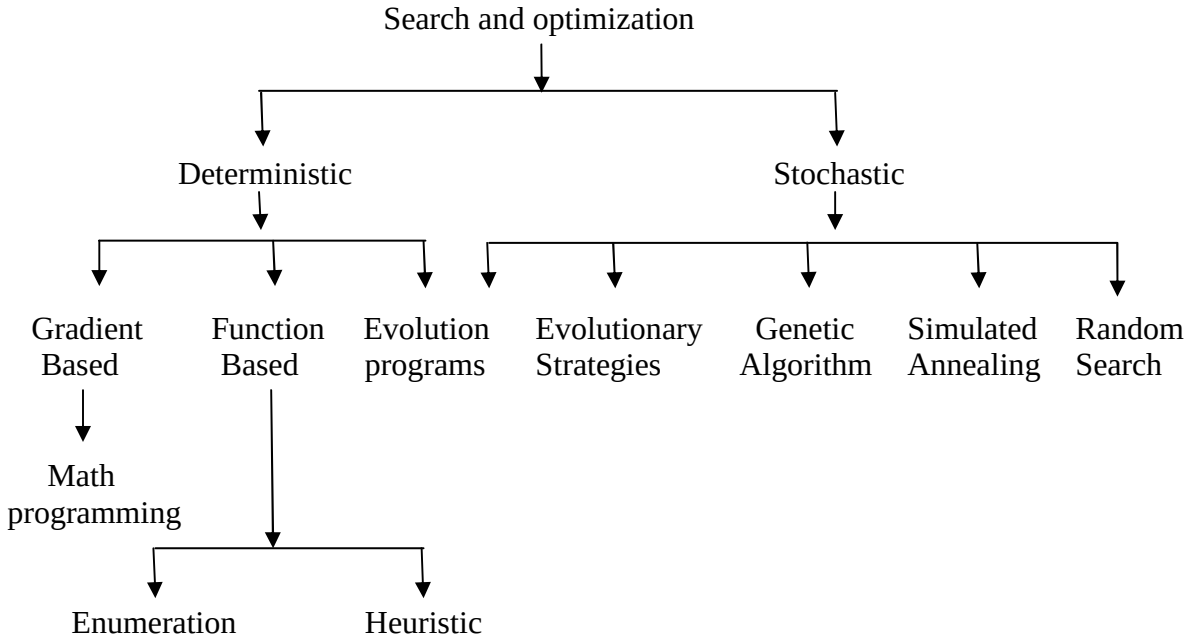
Apart from predicting the process performance under varying operating conditions, a process model can also be used to secure optimum process conditions that would maximize process performance. The objective of process optimization could involve maximization of conversion, product yield, product selectivity, process profit, etc., and/or minimization of operating loss, production cost, selectivities of the undesirable products, etc. A generalized optimization problem statement is given as:

$$\textit{maximize/minimize } f = (\mathbf{x}, \beta); \text{ subject to constraints } C_1, C_2, \ldots \qquad (1.1)$$

where, $\mathbf{x}$ = set of decision variables (to be optimized), f = objective function, and β = set of objective function parameters

Depending upon the type of process model (phenomenological/empirical/AI-based) a suitable formalism should be selected for the model optimization. There exist two principle methods of optimization, namely, *deterministic*, and *stochastic*. Figure 1.1 adapted from Devillers (1996) shows an overview of search and optimization methods.

Figure 1.1: Classification of search and optimization methods.



1.6.1 Deterministic Optimization Methods

These classical methods emboding algorithms which rely heavily on linear algebra since they are commonly based on the computation of the gradient, and in some cases also Hessian, of the response variables (Cavazzuti, 2013). They aim to arrive at the optimum by approximating the local neighborhood of a given solution in the search space and moving to a better solution whenever possible. All gradient based methods and some line search methods fall under this class.

***Gradient based methods*:** As suggested by their name, these methods evaluate gradient of the dependent variable (e.g., prediction error) with respect to the decision variable, and move (update) the decision variable in the negative direction of the gradient. There exist several gradient-based optimization methods such as *delta rule* and *conjugate gradient*.

Advantage of gradient based optimization method

- They converge to an optimium solution speedily, meaning as compared to the stochastic optimization methods, they need smaller number of objective function evaluations to reach the optimal solution.

Disadvantages of gradient-based optimization methods

- Most of these methods require the objective function (to be maximized/minimized) to be continuous, smooth and differentiable. In many real-life systems, the objective function could be noisy, non-smooth and discontinuous and, thus, not amenable to gradient-based methods.
- Invariably get stuck in a local optimum leading to a sub-optimal solution.

1.6.2 Stochastic Optimization Methods

These methods are mostly used in nonlinear optimization. They randomly generate candidate solutions, which are subsequently manipulated according to a specific algorithm. Here, the emphasis is on sampling the search space as widely as possible while trying to locate the promising regions for further search. In the stochastic techniques, randomly generated initial population of candidate solutions is constantly refined so as to find better solutions. In contrast to the traditional deterministic optimization techniques, which invariably operate on a single candidate solution, the stochastic methods operate on a population of candidate solutions. This makes it possible for the stochastic techniques to search several areas of the solution

space. The size of the candidate solution population is user-defined and depends on the size of decision variable space under consideration.

1.7 AI-BASED STOCHASTIC OPTIMIZATION TECHNIQUES

In recent years, several AI-based nonlinear search and optimization techniques such as *genetic algorithms*, *particle swarm*, *ant colony* and *artificial immune systems*, have been proposed. All these have a random component in their implementation.

Advantages of stochastic methods

- Unlike deterministic gradient based optimization methods, stochastic ones do not require the objective function to be smooth, continuous, and differentiable.
- Since they operate on a population of candidate solutions, they scan a wider solution Space. Invariably, they converge to a solution that is global or the deepest local minimum.

***Genetic algorithms* (*GAs*):** Genetic algorithms (Holland, 1975) are the most widely used stochastic optimization formalism. They belong to the AI-based class of search and optimization methods namely *evolutionary algorithms*. GAs enforces the survival of the fittest paradigm of evolution along with the genetic propagation of characteristics. This brings to bear a balanced tradeoff between exploitation and exploration (Michalewicz, 1996) during search for an optimum solution. Beginning from a randomly generated population of candidate solutions to the optimization problem at hand, GA produces offspring population from *parent* candidates that are fitter in some respect. The mechanisms used in offspring production are *selection*, *crossover* and *mutation*. Unlike deterministic optimization methods, which move from point to point, in GA procedure an initial population of solutions is constantly refined in a manner imitating selection and adaption in the biological evolution, while discovering expectedly better solutions.

1.8 OUTLINE OF THE THESIS

The principal aim of this thesis is to employ AI-based modeling formalisms such as artificial neural networks, genetic programming and support vector regression for developing data-driven models of a number of chemical engineering systems including reactions and processes. Additionally, genetic algorithms are used for optimization of reaction conditions. The remainder of this thesis is divided in eight

chapters and references are listed alphabetically at the end of each chapter. In what follows, a brief overview of chapters 2 to 8 is provided.

Chapter 2: Modeling and Optimization Methodologies

In this chapter, first the various AI-based modeling and optimization formalisms utilized in the studies reported in this thesis, are described in detail. The chapter also provides details of the conventional mathematical techniques, namely, *principal component analysis* and s*ensitivity analysis*. These methods have been used in reducing the dimensionality of the input space of the models and identifying influential causal (predictor) variables in the example data sets used in modeling. Additionally, the statistical measures, namely, *coefficient of correlation* (*CC*) and *root mean squared error* (*RMSE*), and the Steiger's z-test used in the evaluation and comparison of the prediction and generalization performance of the data-driven models, are described.

Chapter 3: Modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier

The quality of coal—especially its high ash content—significantly affects the performance of coal-based processes. Accordingly, in this study, data were collected from extensive gasification experiments conducted in a pilot-plant scale fluidized-bed coal gasifier (FBCG)—located at CIMFR, Dhanbad—using high-ash Indian coals. Specifically, the effects of eight coal and gasifier process related parameters on the four gasification performance variables, namely $CO+H_2$ *generation rate*, *syngas production rate*, *carbon conversion*, and *heating value of the syngas*, were rigorously studied. The data collected from extensive gasification experiments were used in the FBCG modeling, which was conducted by utilizing two artificial intelligence (AI) strategies namely *genetic programming* (GP) and *artificial neural networks* (ANNs). The original eight-dimensional input space of the FBCG models was reduced to three-dimensional space using principal component analysis (PCA). This study also presents results of the sensitivity analysis performed to identify those coal and process related parameters, which significantly affect the FBCG process performance.

Chapter 4: High ash char gasification in thermo-gravimetric analyzer and prediction of gasification performance parameters using computational intelligence formalisms

This chapter reports development of the data-driven models for the gasification of chars in the CO_2 atmosphere in a thermo-gravimetric analyzer (TGA); these chars were derived from the high ash Indian coals. Specifically, the models predict two important gasification performance parameters, viz. *gasification rate constant* and *reactivity index*. These models were constructed using three computational intelligence (CI) methods, namely *genetic programming* (GP), *multilayer perceptron* (MLP) neural network, and *support vector regression* (SVR).

Chapter 5: Genetic programming methodology for selecting predictor variables and modeling in process identification

In this chapter, a GP-based strategy has been suggested for (a) simultaneously identifying the important predictor (independent/causal/input) variables that significantly influences the output (dependent variable) of an input-output model, and (b) searching and optimizing an optimal data fitting function and its parameters. The said strategy has been illustrated by conducting two process identification case studies wherein the GP formalism has been shown to (i) identify the influential time-delayed inputs and outputs, and (ii) simultaneously perform system identification using the identified influential predictors.

Chapter 6: Prediction of API gravity of crude oils using SARA analysis: Computational intelligence based models

This chapter presents results of SARA (*Saturates*, *Aromatics*, *Resins* and *Asphaltenes*) fractions based development of nonlinear models predicting ^{o}API magnitudes of crude oils using three computational intelligence (CI) formalisms, namely, *genetic programming*, *artificial neural networks* and *support vector regression*. The SARA analyses and API-gravity values of 403 crude oil samples covering wide ranges have been utilized in developing these models. The CI-based models are found to possess an excellent ^{o}API prediction accuracy and generalization performance.

Chapter 7: Removal of arsenic ions from wastewater using TFA and TAFA resins: Computational intelligence based reaction modeling and optimization

In this study, tannin-formaldehyde (TFA) and tannin-aniline-formaldehyde (TAFA) resins were synthesized and employed successfully for an adsorptive removal of arsenite [As(III)] and arsenate [As(V)] ions from the contaminated water. Further, a computational intelligence (CI) based hybrid modeling-optimization strategy integrating genetic programming and genetic algorithm has been employed to model and optimize, tannin-formaldehyde (TFA) and tannin-aniline-formaldehyde (TAFA) resin-based adsorption of arsenite [As(III)] and arsenate [As(V)] ions for securing optimal reaction conditions**.**

Chapter 8: Genetic programming formalism for prediction of vapor-liquid equilibrium (VLE)

This chapter presents a study wherein *genetic programming* (GP) has been introduced for the prediction of vapor-liquid-equilibria (VLE). Specifically, three case studies have been performed wherein four GP-based VLE models have been developed using experimental data for predicting the *vapor phase composition*, (y_i) of a ternary, and a group of non–ideal binary systems.

Chapter 9: Conclusions

An overview of the important results presented in this thesis and the conclusions drawn thereof are presented in this chapter.

REFERENCES

Bose, B. K. (1994). Expert system, fuzzy logic, and neural network applications in power electronics and motion control. *Proceedings of the IEEE*, *82*(8), 1303-1323.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

Cavazzuri, M. (2013). General guidelines: How to proceed in an optimization exercise. In *Optimization Methods*: *From Theory to Design*, (pp. 147-152), Springer-Verlag Berlin Heidelberg. DOI: 10.1007/978-3-642-31187.

Constantinides, A. (1987). *Applied Numerical Methods with Personal Computers*. Schowalter, W. P., Carberry, J. J., and Fair, J. R. (Eds.), McGraw-Hill, Inc. New York, NY, USA, ISBN: 0070796904.

Devillers, J. (1996). Genetic algorithms in computer-aided molecular design. In *Genetic Algorithms in Molecular Modeling*. Devillers, J. (Ed.), Academic Press. London, pp.1-21.

Fogel, D. B. (2006). *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. 3rd ed., (Vol. 1). John Wiley & Sons, Inc.

Freeman, J. A., and Skapura, D. M. (1991). *Neural Networks Algorithms, Applications, and Programming Techniques*. Addison-Wesley Publishing Company, Reading, M.A, USA.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.

Karelson, M., Lobanov, V. S., and Katritzky, A. R. (1996). Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical reviews*, *96* (3), 1027-1044.

Koza, J. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA.

Mendel, J. M. (1995). Fuzzy logic systems for engineering: A tutorial. *Proceedings of the IEEE*, *83*(3), 345-377.

Michalewicz, Z. (1994). GAs: What are they?. In *Genetic Algorithms+ Data Structures= Evolution Programs*. 3rd ed., Springer-Verlag Berlin Heidelberg, pp. 13-30.

Provost, F., and Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P. Y., Hjalmarsson, H., and Juditsky, A. (1995). Nonlinear black-box modeling in system identification: A unified overview. *Automatica*, *31*(12), 1691-1724.

Tambe, S. S., Kulkarni, B. D., and Deshpande, P. B. (1996). *Elements of Artificial Neural Networks with Selected Applications in Chemical Engineering, and*

Chemical & Biological Sciences. Simulation & Advanced Controls Inc., Louisville, K.Y.

Tootoonchy, H., and Hashemi, H. H. (2013). Fuzzy logic modeling and controller design for a fluidized catalytic cracking unit. In *Proceedings of the World Congress on Engineering and Computer Science*, Vol II WCECS 2013, (pp. 982-987), San Francisco, USA

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. 2nd ed., Springer Verlag, New York. ISBN 978-1-4419-3160-3

Vapnik, V. (1998). *Statistic Learning Theory*. Willey, New York.

Verma, D., Goel, P., Patil-Shinde, V., and Tambe, S. S. (2016, January). Use genetic programming for selecting predictor variables and modeling in process identification. In *IEEE explore, 2016 Indian Control Conference* (*ICC*) (pp. 230-237). IEEE. (ISBN: 978-1-4673-7992-2), doi: 10.1109/INDIANCC.2016.7441133.

Vyas, R., Goel, P., and Tambe, S. S. (2015). Genetic programming applications in chemical sciences and engineering. In *Handbook of Genetic Programming Applications*; Gandomi, A.H., Amir H., Alavi, Ryan, C. (Eds.), Springer International Publishing, Switzerland, pp.99–140.doi:http://dx.doi.org/10.1007/978-3-319-20883-1.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338-353.

Chapter 2

Modeling and Optimization Methodologies

ABSTRACT

In the present thesis, a number of artificial intelligence (*AI*) *and machine learning* (*ML*) *based formalisms have been employed to build exclusively data-driven models for a variety of chemical reactions and processes. These methods are artificial neural networks, genetic programming, and support vector machines. Additionally, an AI-based method namely genetic algorithm has been employed for optimizing reaction conditions of a chemical reaction. Apart from the AI-based methods, conventional methods such as principal component analysis and sensitivity analysis have been employed for dimensionality reduction of the input space and ranking predictor variables in the order of their influence on the response variables, respectively. This chapter describes all the stated methods in sufficient details and lays a strong foundation for the subsequent chapters.*

2.1 INTRODUCTION

Chapter 1 (sections 1.3.1 and 1.3.2) has explained the complexities associated with the *phenomenological* and *empirical* (regression-based) reaction/process modeling, and section 1.6.1 has presented the drawbacks of deterministic optimization techniques. The principal observations that can be made from the stated complexities and drawbacks are:

- Deficiencies of phenomenological modeling necessitate (a) investigation of optional nonlinear modeling strategies, which do not need full details of the physicochemical phenomena underlying the system/process, and (b) it should be possible to model a system/process simply from its relevant data consisting of independent (causal) and dependent (response) variables/parameters.
- The drawbacks of the regression-based modeling techniques require modeling approaches that do not need an explicit specification of the structure (form) of the model. That is, it should be possible to perform modeling without making assumptions regarding the data fitting function and associated parameters.
- Deficiencies of the conventional deterministic optimization formalisms necessitate exploration of methods that do not need the objective function (to be maximized or minimized) to be smooth, differentiable and continuous.

In recent years, *Artificial Intelligence* (AI) and *Machine Learning* (ML) based modeling techniques owing to their several advantages, have provided an attractive avenue for modeling nonlinear and complex multivariable systems. There also exist AI-based efficient stochastic methods that overcome the drawbacks of the deterministic optimization formalisms.

In the present thesis, AI- and ML-based formalisms have been employed to build exclusively data-driven models for tasks such as (a) steady-state modeling of coal-gasifier pilot plant, and char gasification in thermo-gravimetric analyzer, (b) batch reaction modeling of resin-based adsorptive removal of arsenic ions from contaminated water, (c) process identification of a conical tank and adiabatic

continuous stirred tank reactor (CSTR) systems, (d) prediction of API gravity values of crude oils, and (e) vapor-liquid equilibria (VLE) prediction for non-ideal systems. The specific AI-based methodologies used in these modeling studies are *genetic programming* (GP), *artificial neural networks* (ANNs), and *support vector regression* (SVR). The AI-based stochastic optimization strategy used in the thesis for the optimization of process conditions of resin-based adsorptive removal of arsenic ions from contaminated water is *genetic algorithm* (GA).

Often, variables in the data pertaining to a process operation are linearly correlated. This poses problems, such as redundancy and excessive computational load, during process modeling. There exists a method, namely, *principal component analysis* (PCA), which assists in removing linearly correlated variables; thereby, the dimensionality of a data set can be reduced. In the present thesis, PCA has been used for reducing the dimensionality of the input space of several AI-based reaction/process models.

In a chemical process, multiple operating condition variables/parameters (model inputs) affect the process outputs (such as conversion, yield, and selectivity) to different degrees. Some variables and/or parameters are simply more influential than others. A method known as *sensitivity analysis* (SA) is capable of ranking the process operating condition variables/parameters according to their influence on a specific output variable. The SA method has been used in this thesis for ranking the process input variables/parameters according to their influence on the model outputs.

In the studies presented in chapters 3 to 8, it was often necessary to compare the prediction accuracy and generalization capability of competing models. This comparison was performed using two statistical measures, namely, *coefficient of correlation* (*CC*) and *root mean squared error* (*RMSE*). Additionally, Steiger's z-test was employed to compare equivalence of correlation coefficients of competing models, and thereby determining the better performing models.

This chapter presents the essential details of the various modeling and optimization formalisms as also the statistical measures and the test, used in the

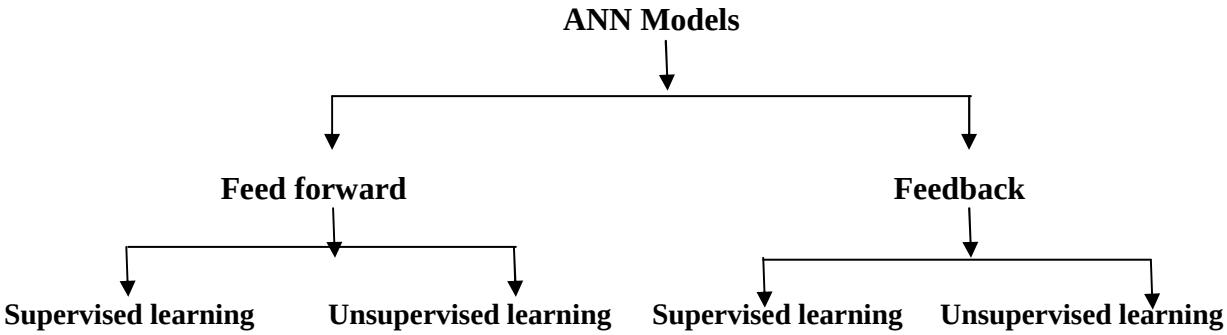
studies presented in the thesis. The remainder of this chapter is structured as follows. Section 2.2 describes the AI-based modeling techniques, namely *artificial neural networks* (ANNs), *multilayer perceptron* (MLP) neural networks (section 2.2.1), and *genetic programming* (GP) (section 2.2.2); the machine learning based *support vector regression* (SVR) modeling method is explained in section 2.3. Next, section 2.4 presents an overview of the AI-based stochastic optimization formalisms; the widely used *genetic algorithm* (GA) method is detailed in section 2.4.1. The description of *principal component analysis* (PCA) and *sensitivity analysis* (SA) is provided in sections 2.5 and 2.6, respectively. Finally, essentials of *Steiger's z-test* are presented in section 2.7.

2.2 ARTIFICIAL INTELLIGENCE (AI)-BASED MODELING TECHNIQUES

2.2.1 Artificial Neural Networks (ANNs)

Artificial neural networks are over-simplified systems that simulate the intelligent performance displayed by human beings; they imitate the types of physical neurological connections occurring in the human brain. ANNs are founded on the conception that a highly interconnected system of simple processing nodes (also called ''processing elements'' or ''artificial neurons'') can learn the complex nonlinear relationships that may exist between variables of a data-set (Tambe et al., 1996). There exist several types of ANNs as presented in Table 2.1. These essentially belong to two categories namely *feed-forward* and *feed-back* ANNs. In the first type, information flow is in the forward direction only, whereas in feedback neural networks information is fed back to the nodes in the same layer and/or to those in the preceding layer(s). Feed forward neural networks (FFNs) are the most frequently used class of ANNs. Among FFNs, an architecture termed *multilayer perceptron* (MLP) has found maximum number of applications in almost every science and engineering discipline. Another widely used FFN is *radial basis function* (RBF) neural network.

Table 2.1: Commonly used artificial neural network architectures (Tambeet al., 1996)



Supervised learning	Unsupervised learning	Supervised learning	Unsupervised learning
▪ Associative-Reward-Penalty(ARP)	▪ Kohonen Self-organizing Feature Map (SOFM)	▪ Brain-State in-a-Box (BSB)	▪ Discrete Bidirectional Associative Memory(BAM)
▪ Adaptive Heuristic Critic (AHC)	▪ Fuzzy Associative Memory (FAM)	▪ Fuzzy Cognitive Map (FCM)	▪ Additive Grossberg (AG)
▪ Adaline/Madaline	▪ Drive-Reinforcement Learning (DR)		▪ Analog Adaptive Resonance Theory (ART2)
▪ Counterpropogation[1] (CP)	▪ Linear Associative Memory (LAM)		▪ Binary Adaptive Resonance Theory (ART1)
▪ Cauchy Machine(CM)	▪ Learning Vector Quantizer (LVQ)[1]		▪ Shunting Grossberg (SG)
▪ Boltzmann Machine(BM)	▪ Learning Matrix (LM)		▪ Adaptive Bidirectional Associative Memory (ABAM)
▪ Multilayer Perceptron (MLP)			▪ Discrete Hopfield (DH)
▪ Perceptron			▪ Continuous Hopfield (CH)
▪ Radial Basis Function[1] (RBF)			

1 *model that utilizes hybrid* (*competitive + supervised*) *learning schemes*

Multilayer Perceptron Neural Network (MLPNN)

Given a representative data set (example set) consisting of *independent* (causal/input/predictor), and the corresponding *dependent* (output/response) variables of a system/process, an MLP neural network possesses an ability of learning and generalizing the nonlinear relationships that exist between the inputs, and outputs to an arbitrary degree of accuracy. An MLP has been found to be an attractive ANN architecture to conduct exclusively data-driven nonlinear process modeling especially in situations wherein development of the first principles (phenomenological) or classical empirical (regression-based) modeling becomes impractical, tedious, and/or costly. The principle features of MLP-based models are (Tambe et al., 1996; Patel et al., 2007):

- Used in approximating complex and nonlinear input-output relationships and performing supervised classification.
- The detailed knowledge of the causative mechanistic phenomena that underlies a reaction or process is unnecessary for the model development.
- A well-trained MLP-based model possesses "generalization" ability due to which it can exactly predict outputs for a fresh set of inputs, which do not belong the example set.
- Even *multiple input- multiple output* (MIMO) nonlinear relationships can be approximated effortlessly and simultaneously.
- It uses a generic nonlinear function for fitting the example set data, and thus it is unnecessary to pre-specify the form of the data-fitting function explicitly.

In Figure 2.1, a commonly used MLPNN is depicted. It comprises four layers of processing nodes—an *input layer*, two intermediate layers called *hidden layers*, and an *output layer*; these layers house *I, J, K* and *L* number of processing nodes, respectively. Very often, MLPNN consists of just a single hidden layer.

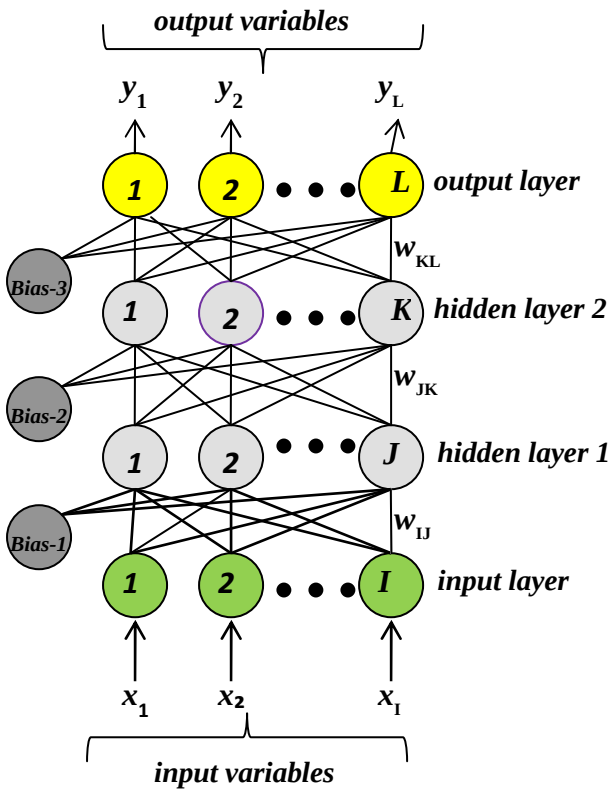


Figure 2.1: Schematic of two hidden layers multiple input−multiple output (MIMO) MLP network.

The processing nodes, alternatively referred to as *artificial neurons*, *nodes*, *processing units/elements*, are the fundamental constituents of an MLPNN. They perform simple mathematical manipulations on the numerical information (data) received from their input connections with the processing elements (PEs) in the previous layer and pass on the computed outputs to the PEs located in the next layer. Each connection of an MLPNN has a parameter termed "weight" associated with it. Although a PE may have multiple output connections, an output signal of the same strength is transmitted across each one of them. In MLP, there exist inter-layer connections, which are classified as *excitatory* or *inhibitory*, according to their resultant actions. The excitatory connection carries a positive signal and enhances the activation level of the destination node. An inhibitory connection has a negative sign, and it reduces the destination PE's activation level.

The MLPNN's input layer houses a number of nodes (I) equal to the number of predictor variables in the example data; the number of nodes in the output layer equals the number of outputs (L) in the system being modeled. It may however be noted, that the number of hidden layers and the number of nodes each one of them houses, are

determined heuristically based on the desired output prediction accuracy and generalization capability of the MLPNN-based model. The weights $\{w_{ij}\}$ on the MLPNN's connections represent the parameters of the model that it approximates during training. As shown in Figure 2.1, an MLPNN contains a bias node with its output fixed at +1, in its input and hidden layers; these nodes are connected to all the nodes in the next layer. The significance of bias nodes is that these help the MLPNN-fitted function to be positioned anywhere in the *I*-dimensional input space.

Table 2.2: Commonly used transfer functions in MLP neural networks (Simpsons, 1990; Hunt et al., 1992)

	Function	Equation	Properties
a.	Linear	$f(x) = ax$	Differentiable, scales-up or scales-down the $f(x)$ values in proportion to real valued constant a, used at output layer.
b.	logistic sigmoid	$f(x) = \frac{1}{(1+e^{-x})}$	Positive, differentiable, monotonic, step-like, symmetric around 0.5, output range [0, 1].
c.	hyperbolic tangent	$f(x) = \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	Differentiable, monotonic, symmetric, step-like, zero-mean, output range [-1, 1].
d.	Gaussian	$f(x) = \exp(\frac{-x^2}{\sigma^2})$	Differentiable, pulse-like.

A processing unit in the network's active i.e., hidden and output layers essentially performs three numerical operations. First, it combines all the input signals to compute the net input. It is then transformed into the node's activation level (*net activation* or simply *activation*) using an activation function. Lastly, the net activation is operated upon using a transfer function to yield the node's output. The transfer function can transform PE's activation in a linear or nonlinear manner. Different types of transfer functions are used for performing such transformations. Table 2.2 lists the formulae of some commonly used transfer functions and their properties. The nonlinear Gaussian function listed in Table 2.2 possesses some special properties and

is used mostly in radial basis function (RBF) networks. On the other hand, the *logistic* and *tanh sigmoid* functions are the common choices for *multilayer perceptron* (MLP) networks.

The input layer neurons of an MLP perform no computations. They simply pass their outputs to the neurons to the next layer (hidden layer). Following application of an input vector, $\mathbf{x}_n$, from the example set to the input layer, each hidden layer neuron first calculates its activation according to the weighted sum of inputs using the following equation:

$$net_{ij}^{h} = \mathbf{w}_j^h \, \mathbf{x}_p + \theta_j^h = \sum_{i=1}^{I} \mathrm{w}_{ij}^h \; x_{pi} + \theta_j^h$$
$$= w_{1j}^h \; x_{p1} + w_{2j}^h \; x_{p2} + \ldots.. + w_{Ij}^h \, x_{kI} + \theta_j^h; \; j = 1,\ldots,J \qquad (2.1)$$

where, net_{ij}^{h} represent activation of j^{th} hidden layer neuron when p^{th} input (p = 1,2, …, N_{P}) pattern/vector in the example set is applied to the input nodes. The vector $\mathbf{w}_j^h$ denotes the weights of the connections linking the input layer nodes to the j^{th} hidden node, and θ_j^h represents the strength of the link between the bias and j^{th} hidden node. The subscripts "h" and "o" designates the quantities associated with hidden and output layers, respectively. The hidden layer outputs are computed by nonlinearly transforming their activations using a transfer function. Outputs of the first hidden layer neurons are either passed to the neurons of the next hidden layer or the output layer. The hidden neurons' outputs are computed using a nonlinear activation function that nonlinearly transforms the net activation level of a hidden neuron. The outputs of the processing nodes in first hidden layer form inputs to the nodes in the subsequent layer; this layer could be another hidden layer, or an output layer. The outputs of these nodes are computed similarly as shown in Eq. (2.1). It may, however, be noted that output layer neurons can use either linear or a nonlinear transfer function to compute their outputs.

***MLPNN training*:** Towards performing a nonlinear function approximation, an MLPNN is trained in a manner such that a pre-specified error function is minimized. The training (learning) process for MLP essentially aims at obtaining an optimal set of network's connection weights that would minimize an error function. There are essentially two methods of training an MLP neural network namely *batch* and *continuous* mode. In the batch mode, network outputs are evaluated using all input

patterns in the example data set following which all network weights are updated once. In the continuous mode, network weights are adjusted immediately after computing the network output pertaining to a single input pattern in the example set.

In the present thesis, continuous mode has been used in training the MLP-based models. It consists of two passes through the network architecture; these are termed *forward* and *reverse* passes. In the forward pass, outputs of all the output nodes (network output) are evaluated using input patterns/vectors of example set. In the reverse pass, the magnitude of the error function specific to the input pattern is calculated using the desired (target) network output, and it is used in updating the network weights. A single training iteration is completed when weight-updation procedure is carried out for all the patterns in the example set. Typically, MLP training needs to be conducted over several iterations till convergence is achieved.

Commonly, "*root mean squared error*" (*RMSE*) is used as the error function in MLP training; the widely employed error function minimization technique is known as the "*error back-propagation*" (EBP) algorithm (Rumelhart et al., 1986). The *RMSE* is calculated as:

$$RMSE = \sqrt{\frac{\sum_{p=1}^{N_p}(y_p^{exp} - y_p^{mdl})^2}{N_p}} \tag{2.2}$$

where N_p represents the number of patterns in the example data set; p is the pattern/vector index, and y_p^{exp} and y_p^{mdl} respectively, denote the experimental (target/desired), and MLP-predicted outputs pertaining to the p^{th} input pattern.

The EBP algorithm uses a gradient-descent technique known as *generalized delta rule* (GDR) for iteratively updating the network weights. Irrespective whether the destination neuron *j* belongs to the hidden or an output layer, the basic weight updation rule for training MLP follows the same basic principle given as the delta rule:

$$\left\{\begin{array}{c}\text{Magnitude of weight}\\ \text{correction at training}\\ \text{iteration, } t, (\Delta w_{ij}\ (t))\end{array}\right\} = \left\{\text{learning rate, } \eta\right\} \times \left\{\begin{array}{c}\text{scaled- error}\\ \text{with respect}\\ \text{to } j^{th} \text{ node}\end{array}\right\} \times \left\{\text{output of } i^{th} \text{ node}\right\} \tag{2.3}$$

The EBP algorithm for MLP training uses two free parameters, namely *learning rate*, η ($0 < \eta < 1$), and *momentum coefficient*, μ_{ebp} ($0 < \mu_{\text{ebp}} < 1$), in its formulation; both these parameters are tuned heuristically.

***Over-fitting of MLP weights and how to avoid it*:** For building an optimal MLP model with good output prediction accuracy, and generalization capability, it is necessary to avoid what is known as ‘‘model over-fitting.’’ An over-fitted MLP model has captured even the micro details such as noise in the data at the cost of learning the smooth trends therein. Such a model is practically useless since it makes poor predictions for a new set of inputs (poor generalization). Over-fitting occurs when (a) an MLP model—with an aim of reducing the prediction error to minimum possible—is trained over a very large number of training iterations (known as ‘‘over-training’’), and (b) MLP’s architecture houses more hidden layers and neurons than are necessary (known as ‘‘over-parameterization’’). Hence, it is absolutely critical to take suitable precautions to avoid over-fitting of an ANN model.

To avoid over-fitting, the example input-output data set is divided into two subsets, namely, *training* and *test* sets. While the first set is used in training the network weights, the test set is used for evaluating the generalization ability of the network undergoing training. Specifically, after each training step, *RMSE* is computed for both training ($RMSE_{\text{trn}}$) and test ($RMSE_{\text{tst}}$) sets; While $RMSE_{\text{trn}}$ indicates the data-fitting ability (also termed “recall ability”) of the network undergoing training, $RMSE_{\text{tst}}$ measures how well the network is generalizing. After training the network over a large number of iterations, the set of network weights resulting in the smallest $RMSE_{\text{tst}}$ magnitude for the test set data is accepted to be an optimal weight set. It may, however, be noted that this weight set pertains to the specific number of hidden units considered in the network architecture.

The complete procedure for constructing an optimal architecture and the related weight matrix of an MLP neural network using the GDR strategy is summarized in the following steps (Bishop, 1994):

1. Choose a small magnitude, for example, one or two for the number of hidden units, *J*, and randomly initialize the network weight matrix.
2. Minimize the test set $RMSE_{\text{tst}}$ using error back propagation algorithm. Repeat training multiple times using each time a different random number sequence for

initializing the network weights. This helps in exploring MLP's weight space widely and, consequently, locating the deepest local or the global minimum on the error surface. Store the network weights that produced the smallest $RMSE_{tst}$.

3. Repeat steps 1 and 2 by systematically increasing the number of hidden units until $RMSE_{tst}$ attains its smallest possible magnitude.

***Issues related to MLP training*:** To construct an optimal MLP model, the effects of variation in its structural attributes, namely, number of hidden layers, number of nodes in each hidden layer, and the type of transfer function, and the two EBP algorithm-specific parameters, namely, learning rate (η) and momentum coefficient (μ_{ebp}), need to be rigorously investigated. The details of the heuristic procedure involved in obtaining an optimal MLP network model possessing good prediction accuracy and generalization capability can be found in, for example, Freeman and Skapura (1991); Zurada (1992); Bishop (1994); and Tambe et al. (1996).

Applications of MLP neural networks in chemical sciences and engineering/ technology

Artificial Neural Networks (ANNs) have been used in chemical science with a great success for providing potential solutions to a variety of data-driven problems. There are some notable generic reviews of applications of artificial neural networks in chemical science and engineering/technology. These are Burns and Whitesides, 1993; Bishop, 1994; Himmelblau, 2000; and Zhang and Friedrich, 2003; for books see, Tambe et al., 1996; and Bulsari, 1995.

Table 2.3: Representative recent applications of MLP neural networks in chemical engineering/technology

Sr. No.	Application	Specific study	Reference
1.	Process modeling	Vapor–liquid equilibrium predictions.	Sharma et al. (1999)
		Modeling of an industrial fluid catalytic cracking unit	Michalopoulos et al. (2001)
		Prediction of vapor-liquid equilibria for binary systems.	Mohanty (2005)
		Prediction of nonlinear viscoelastic behavior of polymeric composites	Al-Haik et al. (2006)

Table 2.3 continued…

Sr. No.	Application	Specific study	Reference
1.	Process modeling	Modeling of anaerobic tapered fluidized bed reactor for starch wastewater treatment	Rangasamy et al. (2007)
		Modeling of the activated sludge process	Moral et al. (2008)
		Thermal conductivity prediction of aqueous electrolyte solutions	Eslamloueyan et al. (2011)
		Estimation of thermal conductivity of ionic liquids	Hezave et al. (2012)
		Modeling of biomass gasification process in fluidized bed reactors.	Puig-Arnavat et al. (2013)
		Modeling of ultrasound-assisted trans-esterification process	Badday et al. (2014)
		Modeling of photocatalyatic process on synthesized ZnO nanoparticles	Amani-Ghadim and Dorraji (2015)
2.	Data analysis	Gas mixture analysis	Moore et al. (1993)
		A neural network methodology for heat transfer data analysis.	Thibault and Grandjean (1991)
3.	Process fault detection/diagnosis	Fault diagnosis in complex chemical plants	Hoskins et al. (1991)
		Framework for enhancing fault diagnosis capabilities	Farell and Roat (1994)
4.	Soft sensor development	Soft sensors development for on-line bioreactor state estimation	de Assis and Maciel (2000)
		Real-time process monitoring and control of an industrial polymerization.	Gonzaga et al. (2009)
5.	Process identification	Identification of dynamic process model	Pollard et al. (1992)
		Robust model predictive control architecture for a neutralization process	Tsai et al. (2003)
6.	Model based process control	Dynamic prediction and control of heat exchangers	Díaz et al. (2001)
		System identification and model predictive control for a flotation column	Mohanty (2009)
7.	Quantitative Structure-Activity/Property Relationships (QSAR/QSPR).	Prediction of fluid properties	Lee and Chen (1993)
		Prediction of polymer properties	Sumpter and Noid (1996)
		Developing Quantitative Structure-Activity Relationships	Dudek et al. (2006)

ANN software packages

There are numerous open source and commercial software packages for training MLP neural networks. Two software packages namely, *IBMSPSS* (2011) and *RapidMiner* (2011) have been used to develop MLP-based models in this thesis. Their details are as given below.

- *IBM® SPSS®* statistics is a comprehensive system for analyzing data. The Advanced Statistics optional add-on module offers the additional analytic skills. This module has been used with the SPSS Statistics Core system and is entirely integrated into that system.
- *RapidMiner* is a software platform that offers an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It supports all steps of a typical data-mining exercise including data preparation, validation and optimization, and results visualization. The RapidMiner (free) Basic Edition is restricted to a single logical processor, and 10,000 data rows are available under the AGPL license.

2.2.2 Genetic Programming (GP)

The principal features of the GP formalism (Koza, 1992; Kinnear, 1994) are conceptually similar to the genetic algorithms (GAs); GA (Goldberg, 1989) is a stochastic search and optimization technique. Both GP and GA are based on the principles of *natural selection* ("survival of the fittest") and *genetics* followed by the biologically evolving species. Given an objective function, the GA is capable of efficiently searching and obtaining the optimal values of the decision variables that would maximize or minimize the function. Although the GP method utilizes same principles as employed by GA, it conducts what is termed *symbolic regression* (SR). It is a methodology of searching both the structural form of a data-fitting function and all of its parameters. Thus, GP is capable of automatically attaining the mathematical model that fits a given set of process data comprising dependent (also termed "response" variables) and independent (also termed "predictor" or "causal") variables. Although intellectually novel and appealing, the GP formalism has not been applied as extensively as other AI-based modeling formalisms such as artificial neural networks and fuzzy logic.

***GP implementation*:** The general form of the model to be obtained using GP-based symbolic regression is given as

$$y = f(\mathbf{x}, \alpha) \tag{2.4}$$

where y denotes the dependent variable; $\mathbf{x} = [x_1, x_2, ..., x_i, ..., x_I]^T$ refers to the I-dimensional vector of independent variables , f represents a linear/nonlinear function whose parameters are defined in terms of a K-dimensional vector, $\alpha = [\alpha_1, \alpha_2, ..., \alpha_k, ..., \alpha_K]^T$. Given data consisting of values of the operating variables, $\mathbf{x}$ (*model inputs*), and the corresponding values of the reaction/process output, y (*model output*), the task of GP is to secure the best fitting functional form, f, and its parameter vector, α.

The GP procedure initiates by creating a random initial population of mathematical expressions, $\{f\}$, representing candidate solutions to the data fitting problem defined in Eq. (2.4); each candidate solution represents a different mathematical data-fitting function, and it is coded symbolically in the form of a tree-like structure. This structure comprises two types of building units namely *functions* (*operators*) and *terminals* (*operands*) (see Figure 2.2). While functions are nodes with branches, terminals are leaves (nodes without branches) of the tree. The function nodes represent operators of a candidate solution. The set of operators that can be used to form a mathematical expression is given below.

- Arithmetic operators: *addition, subtraction, multiplication, division*
- Trigonometric and other mathematical operators: *sine*, *cosine*, *tan*, *cot*, *logarithm*, *exponentiation, etc.*
- Conditional operators: *IF-THEN-ELSE*
- Boolean operators: *AND, OR,* etc.

The terminal nodes define "operands," which are arguments of the mathematical model represented by a candidate solution or entities upon which an operator acts. The terminal set comprises variables, constants (elements of the parameter vector,α), and zero-arity functions (i.e., functions with no arguments) such as *rand* (random number). When arranged properly, the operators and operands appearing in a tree form a complete mathematical expression. The elements of function and terminal sets are the building units of a candidate solution. An illustrative tree structure defining a mathematical expression "$(x_1 + 6) \times (x_4 - 3)$" is shown in Figure 2.2 (*a*).

A typical implementation of the GP-based symbolic regression is shown in the form of a flowchart in Figure 2.3; it consists of following major steps.

1. *Initialization*: Randomly create an initial population of candidate solutions in the symbolic form using tree structures.
2. REPEAT
    a. *Ranking*: Evaluate fitness scores of candidate solutions and rank them according to their scores.
    b. *Selection:* Choose candidates possessing high fitness scores to form a mating pool to undergo crossover and mutation operations.
    c. *Crossover*: Generate offspring candidate solutions by implementing crossover operation.
    d. *Mutation*: Create a new generation of candidate solutions by performing mutation operation on offspring candidate solutions.
3. UNTIL TERMINATION

Each of the above steps can be implemented a number of ways. In what follows, steps corresponding to a generic GP implementation are explained in sufficient details.

Step 1 (*Initialization*): Set the generation index (N_{gen}) to zero (Figure 2.3) and randomly form an initial population of a pre-specified number of candidate solutions/expressions using symbolically coded tree structures as described above.

Step 2a (*Ranking*): Using a pre-specified fitness function evaluates fitness value of the each candidate solution. Fitness function measures the data-fitting ability of a candidate solution. Typically, the mathematical expression represented by the tree-structure is used to compute the model predicted value of the output variable, *y,* and thereby that solution's fitness value. One of the several possible fitness functions is as follows:

$$R_q = \frac{1}{1+\Delta_q^2}; \quad q = 1, 2, \ldots, N_q \tag{2.5}$$

where R_q refers the fitness value (score) of q^{th} candidate solution, N_q refers to the number of candidate solutions in the population and Δ_q^2 refers to the

mean-squared-error (*MSE*) between the desired (target) and model predicted outputs; it is computed as:

$$\Delta_q^2 = \frac{\sum_{p=1}^{N_p}(y_p^{exp} - y_p^{mdl})^2}{N_p} \tag{2.6}$$

where N_p denotes the number of patterns in the data set; p is the pattern index, and y_p^{exp} and y_p^{mdl}, respectively represent the desired (experimental) and the model-predicted outputs to the p^{th} input pattern. Following computation of the fitness values, candidate solutions are ranked in the decreasing order of their fitness.

Step 2b (*Selection*): From the ranked population, this step selects fitter solutions to form a mating pool of parent candidate solutions possessing high fitness scores to undergo crossover and mutation operations (See Figure 2.2 (*b*)). There exist several methods such as *Roulette-wheel selection* (Lipowski and Lipowska, 2012), *Tournament selection* (Miller and Goldberg, 1995), *elitist mating* (Thierens and Goldberg, 1994) etc., for carrying out the stated selection.

Step 2c (*Crossover*): In this key step, a pair of offspring candidate solutions is generated from each of the randomly selected pairs of parent trees in the mating pool. The crossover operation can be performed in a number of ways. For example, in a crossover scheme termed "*single-point*", a point is chosen randomly along the length of each parent tree (see Figure 2.2(*b*)), and both the parent trees are sliced at the respective points. Next, two offspring candidate solutions are created by mutually exchanging and combining the sliced portions of the two parents (see Figure 2.2(*c*)). In another crossover scheme, termed "*two-point crossover*," a pair of nodes is selected randomly from each parent tree, and the contents lying between them are exchanged mutually among the parents to form a pair of offspring. Commonly, crossover is performed with a higher probability than the mutation operation.

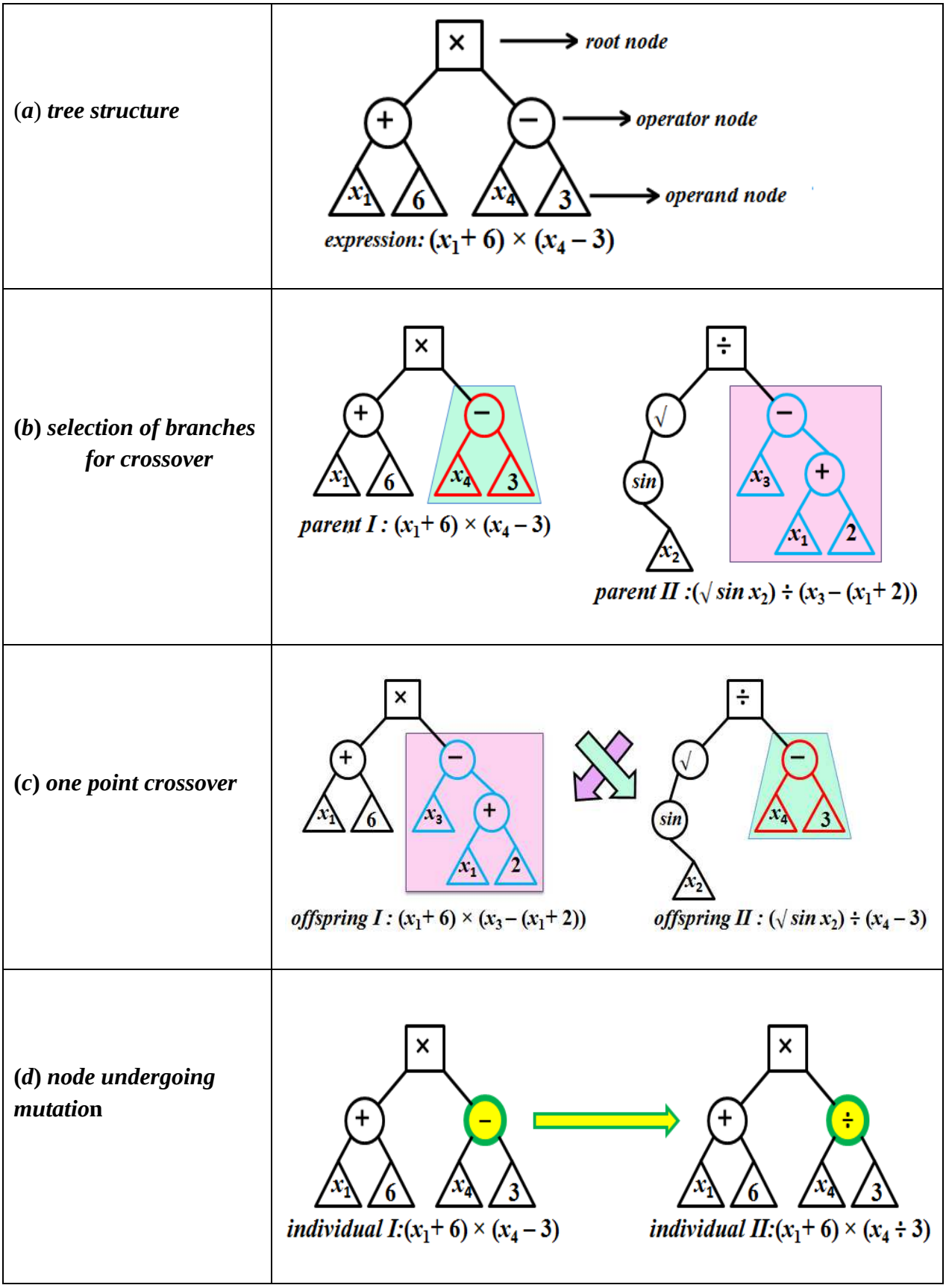


Figure 2.2: Schematic of genetic programming: (*a*) basic tree structure, (*b*) random selection of branches for reproduction, (*c*) crossover operation, and (*d*) mutation operation.

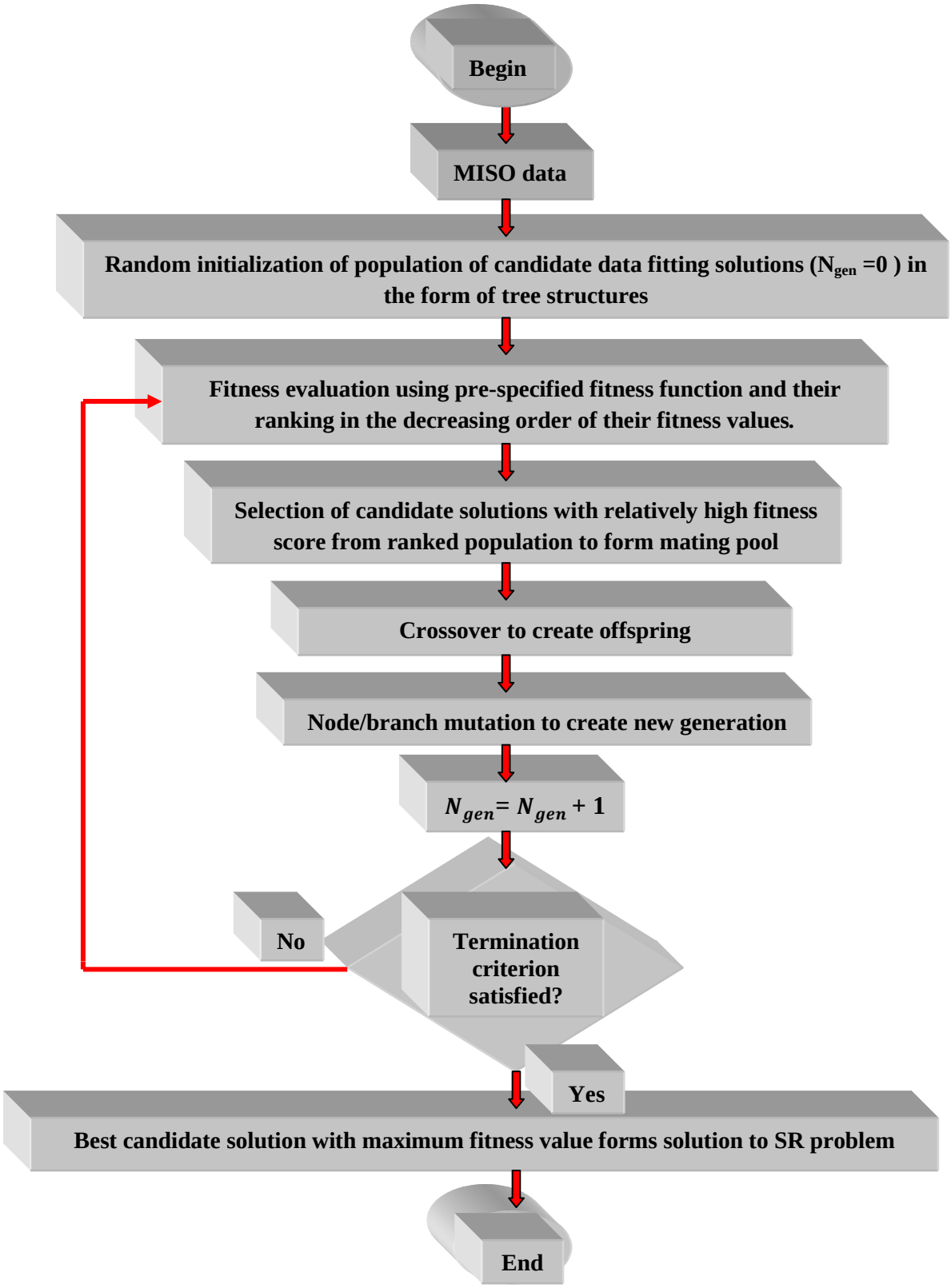


Figure 2.3: Flow-chart of generic GP implementation

Step 2d (*Mutation*): It modifies (mutates) contents of randomly chosen function and/or terminal node(s) of offspring solution trees produced by the crossover operation; it is albeit conducted with a small probability. This operation maintains population diversity and broadens the search for good data-fitting models. Mutation can be conducted two ways namely "branch" and "node" mutation. In node mutation (Figure 2.2(*d*)) an operator (operand) of a randomly chosen function (terminal) node is replaced by another operator (operand), whereas in the branch mutation a randomly chosen branch is replaced by a randomly generated another branch. The population of candidate solutions resulting upon mutated offspring forms a new generation of candidate solutions (i.e. $N_{gen} = N_{gen} + 1$).

Step 3 (*Termination*): Repeat step 2 iteratively till one of the two following termination criteria gets satisfied: (i) the GP has evolved over a pre-specified number of generations, and (ii) the fitness value of the best candidate solution no longer increases significantly or remains constant over a sufficiently large number of generations.

***Over-fitting of GP-based model and how to avoid it*:** Similar to MLP neural networks, a GP-based model is prone to "over-fitting." An over-fitted GP-based model learns even the micro-details in the data at the cost of learning the smooth trends therein. Such a model is useless since it yields sub-optimal predictions for a new set of inputs (poor generalization). In GP training procedure, over-fitting occurs when the model—in an attempt to reduce the prediction error—contains more terms and parameters than necessary. In short, complexity of the model becomes high owing to the more-than-necessary number of terms and parameters in the data-fitting function. It is well-known that a model with high complexity performs poorly at generalization. Thus, it is important to take an appropriate precaution to avoid an

over-fitted GP-based model. This is commonly achieved (as in MLP training) by partitioning the entire input-output example data set available for model building into two sets, namely, *training* and *test* sets. While the GP steps are implemented using the training set, upon convergence the top ranking solution is evaluated using the test set and the solution is accepted as the "best-so-far" only if its data-fitting performance in respect of the test set is closely comparable with that of the training set.

In another method to avoid over fitting, the fitness value of an over-fitted model is appropriately penalized so that it does not enter the mating pool. To obtain an overall optimal data fitting model (f^*) a number of runs may be required by varying the GP-algorithmic parameters systematically. A model is accepted as an overall optimal one only if (i) the correlation coefficients in respect of the training and test set outputs are highest and comparable, and (ii) the *MSE*/*RMSE* magnitudes in respect of training and test set outputs are lowest and comparable. Once an appropriately validated optimal model is secured, its parameters, α, can be fine-tuned further by utilizing a standard nonlinear regression technique, for instance, Marquardt's algorithm (Marquardt, 1963). Such a refinement, if indeed feasible, does improve the prediction accuracy and generalization performance of the GP-based model.

Applications of GP in chemical sciences and engineering

The applications of GP in chemical sciences have focused mainly on data mining, which can be further categorized into *rule-based classification,* and *symbolic regression* based model development. It is the second GP application that has been exploited in this thesis. Earlier the GP technique has been successfully exploited in various fields of chemistry and chemical engineering. Comprehensive reviews of GP applications in chemistry and chemical engineering are provided by Willis et al., 1997 and Vyas et al., 2015. A few selected applications of GP in chemical sciences and engineering are listed in Table 2.4.

Table 2.4: Representative applications of genetic programming in chemical engineering/technology

Sr. No.	Application	Specific study	Reference
1.	Process modeling	Steady-state modeling of chemical process systems.	McKay et al. (1997)
		GP-assisted stochastic optimization strategies for the optimization of glucose to gluconic acid.	Cheema et al. (2002)
		Optimization of a controlled release pharmaceutical formulation.	Barmpalexis et al. (2011)
		K-value program for crude oil components at high pressures based on PVT laboratory data.	Fattah (2012)
		Prediction of permeation flux decline during MF of oily wastewater.	Shokrkar et al. (2012)
		Estimation of the magnitude of *minimum spouting velocity* (U_{ms}) in spouted beds with a conical base.	Hosseini et al. (2014)
		Prediction of char gasification performance parameters derived from high ash coals.	Patil-Shinde et al. (2016)
2.	Process synthesis	Synthesis of heat-integrated complex distillation systems	Wang et al. (2008)
3.	Process monitoring	Bioprocess monitoring: application to continuous production of gluconic acid by immobilized Aspergillus niger.	Sankpal, et al. (2001)
4.	Process fault detection/diagnosis	Process identification and fault diagnosis of non-linear dynamic systems.	Witczak et al. (2002)
		Fault classification using genetic programming.	Zhang and Nandi (2007)
5.	Soft sensor development	Data-driven Soft Sensors in the process industry.	Kadlec et al. (2009)
		The development of soft-sensors for biochemical processes.	Sharma and Tambe (2014)
6.	Process /system identification	System identification of a fluidized catalytic cracking (FCC) unit, for an exothermic reaction.	Nandi et al. (2000)

Table 2.4 continued…

Sr. No.	Application	Specific study	Reference
6.	Process /system identification	System identification of Tennessee Eastman chemical process reactor.	Faris and Sheta (2013)
		To identify the influential time-delayed inputs and outputs, and simultaneously perform system identification using these influential predictors.	Verma et al. (2016)
7.	Model based process control	Generation of empirical dynamic GP models to implement the nonlinear model predictive control (NMPC) strategy.	Grosman and Lewin (2002)
		Development of steady-state and dynamic temperature control models.	Dassau et al. (2006)
8.	Quantitative Structure-Activity/property Relationships (QSAR/QSPR).	Building quantitative structure—property relationship (QSPR) models	Barmpalexis et al. (2011)
		Development of a linear genetic programming (LGP) based quantitative structure-property relationship (QSPR) model for the prediction of standard state real gas entropy of pure materials	Bagheri et al. (2014)

Software packages for GP implementation

Following are the details of two user-friendly software packages that are available for implementing GP algorithm.

- *Eureqa Formulize* (Schmidt and Lipson, 2009; 2014) makes use of symbolic regression technique to capture the intrinsic relationships existing in a given data set, and explain them in a simple mathematical form (structure). It uses GP heavily in its framework, and is optimized to provide "parsimonious" solutions meaning of low complexity.
- *SyMod* software uses machine learning to build symbolic models of the relationship existing between one or more discrete and/or continuous attributes (i.e. independent/causal variables), and a discrete or continuous dependent (response/output) variable. It allows the user to specify a set of mathematical functions, and operators; these are subsequently used to construct predictive

models using genetic programming algorithm. More information of SyMod package can be obtained at the following URL: http://www.symbolicmodeler.org/.

2.3 MACHINE LEARNING BASED MODELING METHOD: Support Vector Regression (SVR)

Support vector regression (SVR) (Vapnik, 1995; Burges, 1998) is an adaptation of the statistical/machine learning theory based classification paradigm, namely *support vector machine*s (Vapnik, 1998). This formalism possesses some desirable characteristics, such as good generalization ability of the regression function, the robustness of the solution, sparseness of the regression, and an automatic control of the solution complexity. Moreover, it provides an explicit knowledge of the data points that define the regression function. This feature allows an interpretation of an SVR-approximated model in terms of the training data.

Given an example data set, $D = \{(\mathbf{x}_p, y_p)\}$, $i = 1, 2, ..., p, \ldots, N_p$, where $\mathbf{x}_p$ is a I-dimensional vector of input variables, and y_i the corresponding scalar output (target), the objective of the SVR algorithm is to fit a regression function, $y = f(\mathbf{x})$, such that it accurately predicts the outputs $\{y_i\}$ corresponding to a new set of input examples $\{\mathbf{x}_i\}$ (Sharma and Tambe, 2014). In SVR, the inputs are first nonlinearly mapped into a high dimensional feature space (Φ) wherein they are correlated linearly with the outputs. The SVR algorithm attempts to place a tube around the regression function as shown in Figure 2.4. The region enclosed by the tube is called an *ε-insensitive* zone, where ε represents the radius of the tube. The optimization criterion in SVR penalizes those data points, the y values of which lie more than ε distance away from the regression function $f(\mathbf{x})$. A detailed description of the SVR and its implementation is found in, for example, Vapnik (1995), Nandi et al. (2004), and Desai et al. (2005). The SVR-based regression function has the following form:

$$f(\mathbf{x}, \mathbf{w}) = f(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \sum_{p=1}^{N_p} (\alpha_p^* - \alpha_p) K(\mathbf{x}_p, \mathbf{x}) + b \qquad (2.7)$$

where, α_p and α_p^* (≥ 0) are the coefficients (Lagrange multipliers) satisfying $\alpha_p \alpha_p^* = 0$; $p = 1, 2, \ldots, P$, and $K(\mathbf{x}_p, \mathbf{x})$ denotes the kernel function describing the dot product in the feature space. The vector $\mathbf{w}$ is described in terms of the Lagrange multipliers α

and α^*. In Eq. (2.7), only some of the coefficients,$(\alpha_p - \alpha_p^*)$, are non-zero and the corresponding input vectors, $\mathbf{x}_i$, are called "support vectors (SVs)." The SVs can be thought of as the most informative data points, which compress the information content of the training set. A number of guidelines for the judicious selection of SVR parameters are provided by Cherkassky and Ma (2004). In the present study, SVR based models have been developed using the ε-SVR module of the data-mining package known as *Rapid Miner* (2014).

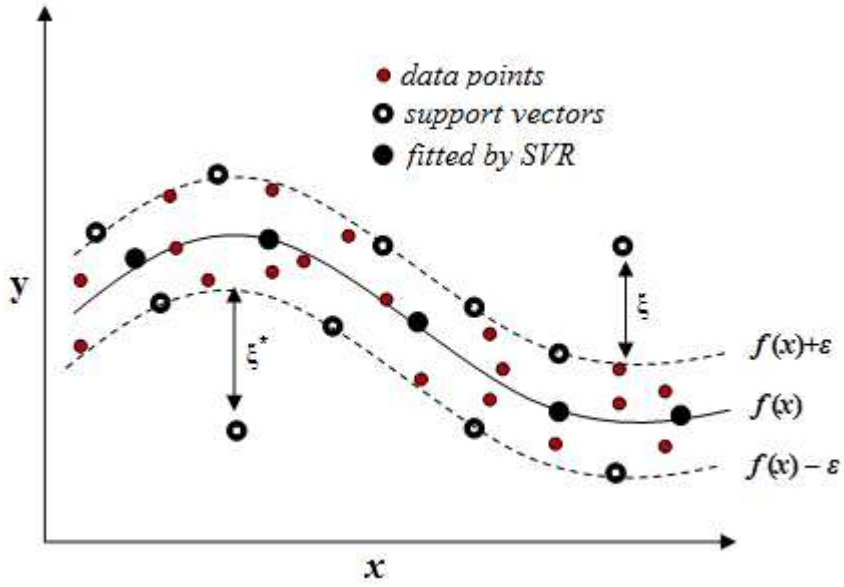


Figure 2.4. A schematic of support vector regression using ε-insensitive loss function

Applications of SVR in chemical engineering and technology

In a short time, SVM/SVR have found plenty applications in chemistry, such as drug design (discriminating between ligands and nonligands, inhibitors and non-inhibitors, etc.), development of quantitative structure-activity relationships, where SVM formalism is used to predict various physical, chemical, or biological properties, chemometrics (optimization of chromatographic separation or compound concentration prediction from spectral data as examples), text mining (automatic recognition of scientific information), and sensor technology (for qualitative and quantitative prediction from sensor data). A comprehensive review of SVR applications in chemistry is provided by Ivanciuc (2007). In chemical engineering too SVR has found a number of applications. A representative list of a few such applications is provided in Table 2.5.

Table 2.5: Representative applications of support vector regression in chemical engineering/technology

Sr. No.	Application	Specific study	Reference
1.	Process modeling	Prediction of pressure drops of slurry flow in pipeline.	Lahiri and Ghanta (2008)
		Predicting the point gas hold-up for bubble column reactor through recurrence quantification analysis of LDA time-series	Gandhi et al. (2008)
2.	Process fault detection/diagnosis	Fault diagnosis based on Fisher discriminant analysis.	Chiang et al. (2004)
		SVR method has been applied for the fault diagnosis in sheet metal stamping processes.	Ge et al. (2004)
3.	Soft sensor development	Soft sensing modeling based on SVM and Bayesian model selection.	Yan et al. (2004)
		Soft-sensor development for bioprocesses in fed-batch bioreactors	Desai et al. (2006)
4.	Model based process control	Predictive functional control design for output temperature of coking furnace	Zhang and Wang (2008)
		Modeling and predictive control of a neutralization reactor	Ławryńczuk (2016)
5.	Quantitative Structure-Activity/property Relationships (QSAR/QSPR).	Development of a QSAR model for the prediction of toxicities of 153 phenols.	Yao et al. (2004)
		Support vector machines QSAR for the toxicity of organic chemicals	Yi and Qin (2007)
		Predictions of chromatographic retention indices of alkyl phenols	Fatemi et al. (2009)

Advantages of SVR

- It uses the structural risk minimization principle by penalizing the model complexity while minimizing the training data error. This results in a model with a better generalization capability.
- Solves a quadratic objective function endowed with a single minimum and, thus, SVR provides globally optimal minimal solutions.
- It permits computations in the input space itself and, hence, reduces the computational load significantly.
- SVR defines a robust regression function and allows sparseness of regression function.

Software packages for SVR implementation

In the present thesis, SVR-based models have been developed using *Rapidminer* package (*2014*).

2.4. ARTIFICIAL INTELLIGENCE (AI) BASED STOCHASTIC OPTIMIZATION FORMALISMS

There exist a number of AI-based optimization methods such as *particle swarm, ant colony, artificial immune systems,* and *genetic algorithms*; these belong to the class termed as "stochastic search and optimization" algorithms and possess some unique advantages as explained in section 1.7 over commonly employed deterministic gradient based algorithms. Among various AI-based optimization methods, genetic algorithms are used most widely. In what follows, an overview of particle swarm, ant colony, and artificial immune system methods is provided followed a detailed description of GA.

(a) Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is an evolutionary computation technique developed by Kennedy and Eberhart in 1995 (Kennedy and Eberhart, 1995; Eberhart and Kennedy, 1995; Eberhart et al., 1996). In PSO a number of simple entities—the particles—are placed in the search space of some problem or function, and each estimates the objective function at its current location. Each particle then determines its movement through the search space by combining some aspect of the history of its

own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next iteration takes place after all particles have been moved. Eventually, the swarm as a whole, like a flock of birds jointly foraging for food, is probably to move close to an optimum of the fitness function.

(b) Ant Colony Optimization (ACO)

The *ant colony optimization algorithm* (ACO), proposed by Dorigo et al. (1996, 1999), is a probabilistic methodology used in solution of those computational problems, which can be reduced to searching good paths through graphs. The ACO method is inspired by the behavior of ants while discovering paths from the colony to food source. In the real world, to begin ants wander randomly, and upon locating a food source return to their colony; while going back to the nest, they lay pheromone trails. When other ants find such a trail, they are less likely to keep travelling at random; instead, they are most likely to follow the trail, and if they indeed find food reinforce the path by depositing pheromones while returning to the colony. Over time, however, the pheromone trail begins to evaporate and as a result its attractive strength decreases. As the time taken to and fro the food increases, the pheromones have more time to evaporate. A trail short in length, gets marched over faster and, therefore, the pheromone concentration remains at high levels as it is laid on the trail as fast as it can evaporate. Pheromone evaporation also has an advantage of getting entrapped in to a locally optimal solution. If no evaporation was to take place, the paths chosen by the first few ants would tend to be highly attractive to the following ones. In such a case, rigorous exploration of the solution space would be severely limited. Thus, it is important that one ant finds a short (good, in other words) path from the nest to a food source in which case, other ants are more likely to follow that path. The positive feedback that gets created eventually leaves all the ants following a single “good” path. The principle of the ACO is to mimic the stated behavior of actual ants with "simulated ants" walking around the graph representing the optimization problem under consideration. ACO algorithms have been used to generate near optimal solutions to the *Travelling Salesman Problem* (Dorigo and Gambardella, 1997). The

ant colony algorithm has advantages such as these can be run continuously and adapt to changes in real time.

(c) Artificial Immune Systems (AIS)

The vertebrate immune system, which defends our body from foreign substances, is one of the most complex and elaborate bodily systems. Its complexity is, in fact, comparable with that of the brain. With advances in the technology, the curiosity about how the immune system functions increased very rapidly (de Castro and Timmis, 2002). This led to its study including the development of mathematical models based on several of its main operative mechanisms. Similar to the study of the nervous system that led to the emergence of ANNs, the study of the immune system has lately inspired the development of AIS as a novel computational/artificial intelligence (CI/AI) paradigm (de Castro and Timmis, 2002). The tremendous information-processing capabilities of the immune system, such as feature extraction, pattern recognition, learning, memory, and its distributive nature provide rich metaphors for its artificial counterpart, i.e. AIS (Aickelin and Dasgupta, 2005; Dasgupta and Nino, 2009). A number of computational methods performing above tasks have been derived from the functioning of the immune system and applied for the solution of much complex real world mathematics, science and engineering problems.

Applications of particle swarm, ant colony, and artificial immune system methods in chemical science, engineering, and technology

A large number of studies have been performed by employing the stochastic optimization methodologies. A wide variety of research papers and reviews on particle swarm, ant colony, and artificial immune systems, and their applications in various fields are available in the literature. Some notable studies and reviews on these methods are Dasgupta and Stephanie (1999), Shi (2001), Maniezzo and Carbonaro (2002), Dasgupta et al. (2003), Martens et al. (2007) and García and Fernández (2012). A severely curtailed representative sample of these studies in chemical engineering is given in the following table.

Table 2.6: Representative applications of particle swarm, ant colony, and artificial immune systems in chemical engineering/technology

Optimization method	**Specific study**	**Reference**
Particle swarm method	Optimization of the hydrolysis of lingo cellulosic residues.	Giordano et al. (2013)
	Prediction of phase equilibrium of binary systems containing ionic liquids.	Lazzús (2013)
	Techno-economic optimization of a shell and tube heat exchanger	Sadeghzadeh et al. (2015)
	Model re-parameterization and parameters estimation for solid-state fermentation process.	da Silveira et al. (2016)
	Bi-level heat exchanger network synthesis with evolution method for optimization.	Wang et al. (2016)
Ant colony method	Introduction to ant colony optimization and survey its most notable applications.	Dorigo et al. (2006)
	Reduce NO_x emissions in coal-fired utility boilers	Zheng et al. (2008)
	Optimization of significant process variables in the biogas production process.	Beltramo et al. (2016)
	Design and scheduling of batch plants	Jayaraman et al. (2000)
Artificial immune system	Fault Diagnosis of batch chemical processes using a dynamic time warping (DTW)-based artificial immune system	Dai and Zhao (2011)
	Removal of heavy metals from residual waters	Dragoi et al. (2012)
	Optimization of process parameters for biodegradable iron chelate for H_2S abatement.	Hamid et al. (2014)

2.4.1. Genetic Algorithm (GA)

Genetic algorithms developed by Holland (1975) (also see Goldberg, 1989; Davis, 1991; Deb, 1995) is a nonlinear search and optimization technique based on the mechanisms of natural selection and genetics. It is the most widely used AI-based stochastic nonlinear optimization formalism, which enforces “the survival of the fittest” paradigm of evolution along with the “genetic propagation of characteristics” followed by the biologically evolving species. It is a robust nonlinear search and optimization technique for conducting function maximization/ minimization. Being a stochastic technique, it differs substantially from the widely used deterministic gradient-based optimization methods (such as conjugate gradient) in that it involves a random component in some stages in its implementation. The advantages of the GA technique are as follows:

- Random initialization of the candidate solution population assists GA in escaping from a locally optimum solution (termed ‘local minimum’) and reaching the globally optimum solution or at least the deepest local minimum.
- It is a zeroth order optimization technique meaning it does not use derivative information of the objective function. GA requires only the measurements of the objective function and not the measurements (or direct calculation) of the gradient (or the higher order derivatives) of the said function (Deb, 1995; Nandi et al., 2001).
- It searches the solution space heuristically and, hence, unlike most deterministic gradient-based methods it is un-affected by the properties (e.g., smoothness, differentiability, continuity, etc.) of the objective function (Goldberg, 1989).
- It has a remarkable capability of handling nonlinear and noisy objective functions.

In the present thesis, GA has been used for process optimization wherein an MLP/GP/SVR based process model is available. This model relates the process operating conditions (inputs) to its output that defines process performance. Accordingly, in what follows, the procedure for the GA-based optimization of the input space of a data-based model (ANN/SVR/GP) is given.

The process optimization objective under consideration is stated as;

Given process data comprising values of process operating (*input*) *variables and the corresponding values of process output* (*response*) *variables, find the optimal values of input variables such that the pre-specified measures of process performance are simultaneously minimized/maximized.*

Having specified an objective function, *f*, GA searches and optimizes its *I*-dimensional decision variable space (**x**) such that the function is minimized or maximized. The function maximization/minimization problem can be defined as:

$$\textit{Maximize/Minimize}\ (y) = f\ (\mathbf{x}, \beta); \quad x_i^L < x_i^{opt} < x_i^U \tag{2.8}$$

Where, *y* denotes the output variable; the *I*-dimensional vector, $\mathbf{x} = [x_1, x_2, ..., x_i, \ldots, x_I]^T$, represents the set of process operating (decision) variables and/or parameters to be optimized; x_i^L and x_i^U are the lower and upper bounds on x_i, and x_i^{opt} denotes the i^{th} optimized decision variable, f refers to the function correlating the output variable with the inputs, and β represents the parameter vector of function f.

In the current thesis, GA has been used to find reaction optimal operating conditions that would maximize the extent of arsenic adsorption (output) on resins (see Chapter 7). In this study, the developed data-driven model itself acts as an objective function to be maximized. This data-driven model predicts the extent of adsorption on a resin.

Overview of GA implementation:

A simple GA-implementation procedure comprises following components:

In each cell of an organism, there is the same set of chromosomes. Chromosomes are strings of DNA and serve as a model for the whole organism. A chromosome consists of genes, which are blocks of DNA. Each gene has its own position in the chromosome called its *locus*. Complete set of genetic material (all chromosomes) is called a *genome*. GA encodes all candidate solutions to the optimization problem in the form of a genetic code. Commonly, these candidate solutions (also termed *strings* or *chromosomes*) are represented using binary digits (binary coding), i.e., in terms of *zero* and *one*.

***Binary Encoding*:** Binary encoding is the most common. In binary encoding, every chromosome is a string of bits, 0 or 1.

Chromosome A	10110010110010101110010 1
Chromosome B	11111110000011000001111 1

Drawbacks of binary coding: Number of bits used scales with the number of variables and the precision of each variable.

***Real Value Coding*:** Direct value encoding can be used in problems, where some complicated value, such as real numbers, is used. Use of binary encoding for this type of problems would be very difficult. In real value encoding, every chromosome is a string of some real numbers. These can be anything connected to the problem, for instance, catalyst concentration, reactant concentration, temperature etc.

Chromosome A	1.2324	5.3243	0.4556	2.3293	2.4545

In the GA procedure, a random population of N_q number of strings is created. either using binary digits or real numbers. In binary coding, each string containing l_{chr}^{A} number of bits is divided into *I* segments where an i^{th} (*i* = 1, 2, ..., *I*) segment of length l_i^A represents the binary representation of the i^{th} decision variable. The decimal equivalent, x_i, of the i^{th} binary segment is evaluated as

$$x_i = x_i^L + \frac{(x_i^U - x_i^L)S_i}{2^{l_i^A} - 1}\,; \quad i = 1, 2, \ldots\ldots, I\,; \qquad \sum_{i=1}^{I} l_i^A \;=\; l_{chr}^{A} \tag{2.9}$$

where S_i represents the decimal value of the i^{th} binary segment comprising l_i^A bits. Upon decoding all N_q strings in the current population in this manner, their fitness values, $\{R^A\}$, are evaluated using a pre-specified fitness function. Next, the string population is subjected to the actions of four genetic operators, namely, *selection*, *reproduction*, *crossover*, and *mutation*, to obtain a new generation of candidate solutions. The actions of these GA operators are repeated with successive generations of solutions till convergence is achieved. The entire GA-implementation can now be summarized as follows:

Step 1 (*Initialization*): Set generation index (N_{gen}^{A}) to zero and generate a population of N_q binary strings (chromosomes) randomly. Each string consisting of l_{chr}^{A} bits is divided into I segments equal to the number of decision (input) variables to be optimized.

Step 2 (*Fitness computation*): Decode q^{th} (q= 1, 2,..., N_q) binary string to obtain the corresponding decimal values of the decision variables, x_{qi}, i = 1, 2, ..., I (see eq. 2.9), and evaluate the fitness (R_q^A) of the q^{th} string as given by

$$R_q^A = H\,(\,y_q) = H\,[f^*(X_q\,,\,\alpha\,)\,] \tag{2.10}$$

where X_q refers to the real-valued decision variable vector, X_q= [x_{q1}, x_{q2}, ..., x_{qI}]^T. After computing fitness values of all the N_q strings in the current population, the strings are ranked in the decreasing order of their fitness values.

Step 3 (*Selection of parents*): From the current population, choose N_q number of parent strings to form the mating pool. The members of this pool, which are used to produce offspring population, possess relatively high fitness scores. The commonly used parent selection techniques are *Roulette-wheel selection* (Lipowski and Lipowska, 2012), *Tournament selection* (Miller and Goldberg, 1995), and *elitist mating* (Thierens and Goldberg, 1994).

- *Roulette-wheel selection:* Selection of the candidate solutions in the mating pool is done, such that candidates with higher fitness scores contribute higher number of copies to the mating pool. It is conducted by creating copies of the candidates in proportion to their fitness scores. This ensures that the mating pool has more number of candidates with higher fitness as compared to those with lower fitness scores.
- *Tournament selection:* This is a static selection scheme where the probability of selection of a candidate remains fairly constant across generations. In this scheme, a specified number, called the “tournament size”, of members are chosen from the parent population and these enter competition for selection. The winner is decided based on the

best fitness and allowed to enter the reproductive phase. This process is repeated sufficiently, along with recombination and mutation, to produce the offspring population.

- *Elitist Selection*: *Elitism* is sometimes the case that a good solution is found early on in the GA run but gets deleted from the population as the GA progresses. One solution is to "memorize" the best solution found so far. A technique called *elitism* has been used to ensure that the best members of the population are carried forward from one generation to the next.

Step 4 (*Crossover*): From the mating pool, select (N_q/2) number of parent pairs randomly; the crossover operation is performed on each pair using a high value for the crossover probability, P_{cr}^{A} (range $0.9-1.0$). A random number is drawn, and whenever it falls below the crossover probability, two individuals (selected using one of the selection schemes described in the following section) are allowed to undergo crossover. If the random number test fails, the chosen individuals are duplicated and placed in the offspring population. This crossover operation, when repeated on the (N_q/2) number of parent pairs, produces N_q number of offspring strings.

- *One point crossover*: Here, a random cut-point is chosen along the length of the coded solution and the two parent chromosomes are split at this point. The tail portion (i.e., the entire bit positions following the cut-point) of the two parents are exchanged to create two offspring chromosomes.
- *Two-point crossover*: Two random cut-points are chosen and the portions of the encoded representations of the parents between these cut-points are mutually exchanged.
- *Uniform crossover*: It is the generalized form of crossover where chromosomal exchanges happen between parents, across multiple (the number is chosen randomly) cut-points. The recombination operator has a probability associated with it which dictates how often it is used.

Step 5 (*Mutation*): Mutate the bits of offspring strings wherein the probability that a randomly selected bit undergoes mutation is P_{mut}^{A} (range $0.01-0.05$). In mutation, a randomly selected bit has been flipped from zero to one and vice versa. The population emerging after the mutation operation represents a new generation of solutions and thus, ($N_{gen}^{A}= N_{gen}^{A} + 1$).

Step 6 (*Termination*): Repeat steps 2 to 5 on the new generation of strings till it is observed that the fitness of the best solution shows no increase over a large number (say ≈ 1000) of successive generations or the GA has evolved over a specified number (N_{max}^{A}) of generations. Finally, the *N* binary segments in the string possessing maximum fitness score are decoded (see eq. 2.9), and the optimal values of the decision variables obtained thereby represent the optimized solution, $X^* = [x_1^*, x_2^*, ..., x_I^*]^T$. Analogous to the GP procedure, it is necessary in the GA procedure also that the entire GA implementation is repeated several times using different seed values for the random number generator. The optimal solutions obtained thereby are compared, and the one satisfying the optimization objective of function maximization or minimization in a best possible manner is selected as an overall optimal solution.

Applications of GA in chemical sciences and engineering

In chemical engineering, GAs are primarily used for the steady-state/dynamic process optimization, nonlinear process identification and control, fault detection and diagnosis, QSAR (quantitative structure-activity relationships) and QSPR (quantitative structure property relationships) tasks. A number of short and comprehensive reviews of GA applications in chemistry and chemical engineering are provided by, for example, Lucasius and Kateman (1993), Lucasius and Kateman (1994), and Venkatasubramanian and Sundaram (1998). Some of the important chemical engineering applications of GAs are listed in Table 2.7.

Table 2.7: Representative applications of genetic algorithm in chemical engineering/technology

Sr. No.	Specific study	Reference
1.	Forecasting chaotic time series	Szpiro (1997)
2.	Optimization study of benzene isopropylation on Hbeta catalyst to maximize the process performance.	Nandi et al. (2004)
3.	Development of correlations for the overall gas hold-up, volumetric mass transfer coefficient, and effective interfacial area in bubble column reactors.	Gupta et al. (2009)
4.	Experimental optimization of supercritical extraction of β-carotene from Aloe barbadensis Miller.	Bashipour and Ghoreishi (2012)
5.	Determination of interaction parameters in multicomponent systems of liquid–liquid equilibria	Khansary and Sani(2014)
6.	Chemometrics tools in QSAR/QSPR studies	Yousefinejad and Hemmateenejad (2015)
7.	Mathematical modeling of continuous ethanol fermentation in a membrane bioreactor by pervaporation	Esfahanian et al. (2016)
8.	Study of change in particle size distribution in a gas-solid fluidized bed due to particle attrition.	Farizhandi et al. (2016)
9.	Optimization of chemical reactors network.	Leong et al. (2016)
10.	Modeling and optimization of a pharmaceutical crystallization process.	Velásco-Mejía et al. (2016)
11.	Modeling and optimization of toluene oxidation over perovskite-type nanocatalysts	Zonouz et al. (2016)

2.5 DIMENSIONALITY REDUCTION METHOD: Principal Component Analysis (PCA)

PCA was first introduced in statistics by Pearson (1901), who formulated the analysis as finding "lines and planes of closest fit to systems of points in space". PCA was briefly mentioned by Fisher and MacKenzie (1923) as more suitable than analysis of variance (ANOVA) for the modeling of response data. Fisher and MacKenzie (1923) also outlined the nonlinear iterative partial least squares (NIPALS) algorithm, which was later rediscovered by Wold (1966) and Hotelling (1933) that have further developed PCA to its present stage (also see Geladi and Kowalski, 1986).

When large multivariate datasets are analyzed, it is often desirable to reduce their dimensionality. Principal component analysis (PCA) is one technique for achieving the stated task; it is a multivariate statistical technique that analyzes a data set in which original variables are described by several inter-correlated quantitative dependent (derived) variables. Multivariate techniques can consider a number of factors, which control data variability simultaneously and therefore offer significant advantages over univariate techniques, where errors associated with repeated statistical testing can occur. In simple terms PCA, in essence, computes new orthogonal variables (*principal components* or *factors*) from linear combinations of the original variables to display the pattern of similarity of the observations, and of the original variables. Principal components are a transformed variable set defining the eigenvectors of the covariance of the data and the associated parameters. The first principal component, or factor, accounts for the greatest variability in the data; or the first few variables retain most of the variation present in all of the original data (Nomikos and MacGregor, 1994), and there can be an infinite number of new factors with each accounting for less data variability than the previous (Dong and McAvoy, 1996).

To illustrate the PCA method, consider a two dimensional matrix, $X(N_p, I)$, defining N_p measurements of I variables. The PCA decomposes, X, into matrices of

latent variables and the corresponding parameters (known also as "loadings") as given by:

$$X = TP' + E \tag{2.11}$$

where, matrix X is assumed to be mean-centered (mean = 0) and variance-scaled (i.e. the standard deviation of elements of each column is unity); T (N_p, I) denotes the matrix of I principal component (PC) scores (each column of matrix T signifies a principal component); P' refers to the transpose of the loading matrix, $P(I, I)$, and E denotes the residuals. In the event of linearly correlated variables, first R principle component scores capture a large amount of variance in the data, and thus Eq. (2.11) can be rewritten as

$$X = \sum_{r=1}^{R} t_r (p_r)' + E' \tag{2.12}$$

where, t_r denotes the N_p-dimensional r^{th} score vector; p_r refers to the transpose of the r^{th} I-dimensional loading vector, p_r, and E'denotes the residual matrix. It can be seen from Eq. (2.12) that the original ($N_p \times I$) dimensional data matrix, X, can now be represented in terms of N_p number of R-dimensional score vectors. Since R is smaller than I, the original data can be represented in terms of a smaller matrix. The sum of squares of elements of a score vector (t_r) is related to the eigenvalue (also known as "*trace*") of that vector and it serves as a measure of the variance captured by the r^{th} principle component. It thus follows that larger the magnitude of a trace, more significant is the respective principal component.

Application areas of Principal Component Analysis (PCA)

A wide variety of research papers and reviews are available in the technical literature wherein PCA has been used in various studies. Some selective studies and reviews on this method are, Kruger et al. (2008), Abdi and Williams (2010), and Bro and Smilde (2014).

Table 2.8: Representative applications of principal component analysis in chemical engineering/technology

Sr. No.	Specific study	Reference
1.	Principal component analysis in linear systems: Controllability, observability, and model reduction	Moore(1981)
2.	Detection and diagnosis of abnormal batch operations.	Nomikos(1996)
3.	Non-linear principal components analysis using genetic programming	Hiden et al. (1997)
4.	The application of principal component analysis and kernel density estimation to enhance process monitoring	Chen et al. (2000)
5.	Fault detection behavior and performance analysis of principal component analysis based process monitoring methods	Wang et al. (2002)
6.	Fault identification for process monitoring using kernel principal component analysis.	Cho et al. (2005)
7.	A review of principal component analysis and its applications to color technology.	Tzeng and Berns (2005)
8.	Sensor-fault detection, diagnosis and estimation for centrifugal chiller systems.	Wang and Cui (2005)
9.	Coal gasification in a pilot plant scale fluidized bed gasifier	Patil-Shinde et al. (2014)
10.	Prediction of high ash char gasification performance parameters	Patil-Shinde et al. (2016)

2.6 SENSITIVITY ANALYSIS

In this thesis, sensitivity analysis (also termed "*importance*" analysis) has been performed for the example input-output data used in the development of various models. It is the analysis of the importance of imprecision or uncertainty in the model inputs in a decision-making or modeling exercise. It is conducted to ascertain the extent of influence exerted by each input (independent/causal) variable on the output

(dependent/response) variable. The "importance" of a predictor/independent variable is a measure of how much the model-predicted output value changes when the predictor magnitude is changed. The related quantity termed "*normalized importance*" is simply the importance value divided by the largest importance value and expressed it as a percentage (IBM SPSS, 2011). The importance analysis is conducted using the entire set of data which have been used in developing the models. The *importance chart* is purely a bar chart of the values in the importance table, sorted by the descending values of importance.

Methods of sensitivity analysis

Numerous methods have been developed to determine how sensitive model outputs are to changes in model inputs. Most methodologies examine the effects of changes in a single parameter value or input variable assuming no variations in all the other inputs. (UNESCO Report, 2005).

***Analytical methods*:** Analytical approaches for sensitivity analysis do not exist for complex simulation models. However, procedures based on simplifying assumptions and guesstimates can be used to yield useful sensitivity information.

Difficulties faced with analytical methods

- Obtaining the derivatives for many models.
- Needing to assume mathematical (usually linear) relationships when obtaining estimates of derivatives by making small changes of input data values near their nominal or most likely values.
- Having large variances associated with most process models.

Above stated difficulties have motivated the replacement of analytical method by numerical and statistical approach for sensitivity analysis.

***Numerical and statistical methods*:** There exist a number of *numerical* and *statistical* methods for sensitivity analysis. A few prominent ones are: *deterministic sensitivity analysis, first-order sensitivity analysis,* and *Monte Carlo sampling methods*. A detailed description of these methods can be found, for example, in UNESCO Report (2005). In the present thesis sensitivity analysis was conducted for MLP based models. An overview of MLP-based sensitivity analysis is given below.

2.6.1 Artificial neural network based sensitivity analysis

For feed-forward network *multilayer perceptron* (MLP), sensitivity is analyzed through the hyper-rectangle model. In this method, the sensitivity measure is defined as the mathematical expectation of output deviation due to expected input deviation with respect to overall input patterns in a continuous interval. Based on the structural characteristics of the MLP, a bottom-up approach is adopted. A single neuron is considered first, and algorithms with approximately derived analytical expressions that are functions of expected input deviation are given for the computation of its sensitivity. Then another algorithm is given to compute the sensitivity of the entire MLP network. In the present thesis, the sensitivity analysis was performed using (IBM SPSS, 2011) package.

Application areas of sensitivity analysis in chemical engineering/technology

Table 2.9: Representative applications of sensitivity analysis in chemical engineering/technology

Sr. No.	Specific Study	Reference
1.	To study mechanics of artificial neural networks for the relative influence of the independent variables in the prediction process.	Olden and Jackson(2002)
2.	Importance of input variables on the output of a feed forward neural network have been proposed	Montano and Palmer(2003)
3.	Quantifying variable importance in artificial neural networks	Olden et al. (2004)
4.	Identifying, quantifying and communicating the uncertainties in model outputs.	Loucks et al. (2005)
5.	To rank the impact of object oriented metrics in fault prediction modeling.	Kaur et al. (2006)
6.	To predict and simulate the behavior of the Fenton process.	Elmolla et al. (2010)

2.7 STEIGER'S Z-TEST

In a variety of situations in research, it is desirable to be able to make statistical comparisons between correlation coefficients measured on the same individuals. For example, an experimenter may wish to assess whether two predictions correlate equally with a criterion variable. In another situation, the experimenter may wish to test the hypothesis that an entire matrix of correlations has remained stable over time. A statistical test known as Steiger's z-test (Steiger, 1980) is performed for comparing the performance of a pair of models. Specifically, this test is used to examine whether the two correlation coefficients corresponding to the predictions of two competitive models are significantly different. It tests the *null hypothesis* (H_0) that statistically two correlation coefficient magnitudes are not different, i.e. $CC_{AB} = CC_{AC}$. Subscripts *A*, *B*, and *C*, respectively denote the experimental values and those predicted by the models *B* and *C*, where $CC_{AB}(CC_{AC})$ refers to the correlation coefficient pertaining to the model *B* (model *C*) predicted outputs and their corresponding experimental counterparts. If the obtained *p-values* are less than 0.05, this indicates a uniform rejection of the null hypothesis (at 95% confidence level) regarding the statistical equivalence of the *CC* magnitudes pertaining to the respective model pairs. It can thus be concluded that the differences in the *CC* magnitudes of the stated model pairs are statistically significant. From the *CC* magnitudes and the results of Steiger's z-test for a pair of models, it is possible to determine which model possesses higher prediction accuracy and generalization capability.

The formula for computing Steiger's Z-statistic is given below.

$$Z = [Z_{12} - Z_{13}] \times \frac{\sqrt{[N-3]}}{\sqrt{2\times[1-r_{23}]\times h}} \tag{2.13}$$

where, Z_{12} and Z_{13} are the Fisher's Z transformations of r_{12} and r_{13} , respectively.

$$h = \frac{1-[f\times rm^2]}{1-rm^2}; \quad \text{where} \quad rm^2 = \frac{r_{12}^2 + r_{13}^2}{2} \tag{2.14}$$

If Z > 1.96, $p < .05$; Z > 2.58, $p < .01$

2.8 CONCLUSION

The AI-based modeling methods such as artificial neural networks, genetic programming, and support vector regression have some attractive features. As a

result, they have found numerous modeling applications in chemical engineering and technology. In this chapter, these methods, which have been extensively employed to conduct various modeling studies in chapters 3 to 8, have been described in sufficient details. This chapter also presents the commonly utilized AI-based stochastic optimization method, namely, genetic algorithm, which has been employed in Chapter 7 for obtaining optimal conditions for a resin-based waste-water treatment reaction.

In addition to the AI-based modeling, a number of studies reported in this thesis have utilized dimensionality reduction and sensitivity analysis methods (sections 2.5 and 2.6); these are respectively used for reducing the dimensionality of the input space of the models and identifying influential input variables.

In each of the modeling studies presented in the thesis, there was a need to rigorously compare the prediction and generalization performance of the competing AI-based and other models. This comparison was performed mostly using the Steiger's z-test described in section 2.7. In summary, this chapter lays a strong foundation for the subsequent chapters by presenting in detail the various AI and machine learning-based modeling and optimization methods, as also conventional mathematical methods used in data pre-processing.

NOMENCLATURE

E'	residual matrix in PCA
f	linear/nonlinear function whose parameters are defined in terms of a *K*-dimensional vector, α
I	Number of input nodes in MLPNN, and pattern index in SVR formulation
K	kernel function in SVR , and number of input nodes in 2nd hidden layer of MLPNN
l_{chr}^{A}	length of a chromosome or a string in GA simulation
l_{i}^{A}	number of bits to represent i^{th} decision variable
net_{ij}^{h}	activation of j^{th} hidden layer
N_{gen}	generation index
N_{gen}^{A}	generation index in GA simulation

N_{max}^{A} maximum number of generations for GA evolution

N_p number of patterns in the data set

N_q number of binary strings (population size) in GA simulation; number of candidate solutions in the GP population

P' transpose of the loading matrix

P_{cr}^{A} crossover probability in GA procedure

P_{mut}^{A} mutation probability in GA procedure

p_r transpose of the r^{th} J- dimensional loading vector in PCA

R^A string fitness in GA procedure

R_q fitness score of q^{th} candidate solution in GP

R_q^A fitness score of q^{th} candidate solution in GA

S_i decoded decimal value of i^{th} binary segment

t_r N_p-dimensional r^{th} score vector in PCA

$\mathbf{w}_j^h$ weights of the connections between input layer nodes and j^{th} hidden node

$\mathbf{x}_p$ = $[x_1, x_2, ..., x_i, ..., x_I,]^T$ refers to the I-dimensional vector of independent/input variables

x_i^{opt} i^{th} optimized decision variable

y Dependent/output variable

y_p^{exp} Experimental (target) outputs pertaining to the p^{th} input pattern.

y_p^{mdl} Model-predicted outputs pertaining to the p^{th} input pattern.

Greek letters

α = $[\alpha_1, \alpha_2, ..., \alpha_k, ..., \alpha_K]^{\text{T}}$, parameter vector in GP

α_p, α_p^* Lagrange multipliers in SVR

β Parameter vector of function f

Δ_q^2 mean-squared-error between the target and model predicted outputs for the entire solution population

η learning rate in the EBP algorithm

θ_j^h Strength of the connection that the bias neuron makes with j^{th} hidden node.

μ_{ebp} momentum coefficient in the EBP algorithm

REFERENCES

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459.

Aickelin, U., Dasgupta, D. (2005). Artificial immune systems. In *Search Methodologies*: *Introductory Tutorials in Optimization and Decision Support Techniques*; Burke, E.K., Kendall, G. (Eds.), Chapter 13, 1st Ed., Springer, NewYork.

Al-Haik, M. S., Hussaini, M. Y., and Garmestani, H. (2006). Prediction of nonlinear viscoelastic behavior of polymeric composites using an artificial neural network. *International Journal of Plasticity*, *22*(7), 1367-1392.

Amani-Ghadim, A. R., and Dorraji, M. S. (2015). Modeling of photocatalyatic process on synthesized ZnO nanoparticles: Kinetic model development and artificial neural networks. *Applied Catalysis B: Environmental*, *163*, 539-546.

Badday, A. S., Abdullah, A. Z., and Lee, K. T. (2014). Artificial neural network approach for modeling of ultrasound-assisted transesterification process of crude Jatropha oil catalyzed by heteropolyacid based catalyst. *Chemical Engineering and Processing: Process Intensification*, *75*, 31-37.

Bagheri, M., Borhani, T. N. G., Gandomi, A. H., and Manan, Z. A. (2014). A simple modelling approach for prediction of standard state real gas entropy of pure materials. *SAR and QSAR in Environmental Research*, *25*(9), 695-710.

Barmpalexis, P., Kachrimanis, K., Tsakonas, A., and Georgarakis, E. (2011). Symbolic regression via genetic programming in the optimization of a controlled

release pharmaceutical formulation. *Chemometrics and Intelligent Laboratory Systems*, *107*(1), 75-82.

Bashipour, F., and Ghoreishi, S. M. (2012). Experimental optimization of supercritical extraction of β-carotene from Aloe barbadensis Miller via genetic algorithm. *The Journal of Supercritical Fluids*, *72*, 312-319.

Beltramo, T., Ranzan, C., Hinrichs, J., and Hitzmann, B. (2016). Artificial neural network prediction of the biogas flow rate optimized with an ant colony algorithm. *Biosystems Engineering*, *143*, 68-78.

Bishop, C. M. (1994). Neural networks and their applications. *Review of Scientific Instruments*, *65*(6), 1803-1832.

Bro, R., and Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812-2831.

Bulsari, A. B. (1995). *Neural Networks for Chemical Engineers*. Elsevier Science Inc. New York. ISBN:0444820973.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121-167.

Burns, J. A., and Whitesides, G. M. (1993). Feed-forward neural networks in chemistry: mathematical systems for classification and pattern recognition. *Chemical Reviews*, *93*(8), 2583-2601.

Cheema, J. J. S., Sankpal, N. V., Tambe, S. S., and Kulkarni, B. D. (2002). Genetic programming assisted stochastic optimization strategies for optimization of glucose to gluconic acid fermentation. *Biotechnology Progress*, *18*(6), 1356-1365.

Chen, Q., Wynne, R. J., Goulding, P., and Sandoz, D. (2000). The application of principal component analysis and kernel density estimation to enhance process monitoring. *Control Engineering Practice*, *8*(5), 531-543.

Cherkassky, V., and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, *17*(1), 113-126.

Chiang, L. H., Kotanchek, M. E., and Kordon, A. K. (2004). Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Computers & Chemical Engineering*, *28*(8), 1389-1401.

Cho, J. H., Lee, J. M., Choi, S. W., Lee, D., and Lee, I. B. (2005). Fault identification for process monitoring using kernel principal component analysis. *Chemical Engineering Science*, *60*(1), 279-288.

da Silveira, C. L., Mazutti, M. A., and Salau, N. P. (2016). Solid-state fermentation process model reparametrization procedure for parameters estimation using particle swarm optimization. *Journal of Chemical Technology and Biotechnology*, *91*(3), 762-768.

Dai, Y., and Zhao, J. (2011). Fault diagnosis of batch chemical processes using a dynamic time warping (DTW)-based artificial immune system. *Industrial & Engineering Chemistry Research*, *50*(8), 4534-4544.

Dasgupta, D., and Stephanie, F. (1999). Artificial immune systems in industrial applications. *Intelligent Processing and Manufacturing of Materials, 1999. IPMM'99. Proceedings of the Second International Conference on*. *Vol. 1*. IEEE.

Dasgupta, D., Ji, Z., and González, F. A. (2003, December). Artificial immune system (AIS) research in the last five years. In *IEEE Congress on Evolutionary Computation* (*1*) (pp. 123-130).

Dasgupta, D., Nino, L.F. (2009). *Immunological Computation*. 1st ed., Auerbach Publications, Taylor & Francis Group, Boca Raton.

Dassau, E., Grosman, B., and Lewin, D. R. (2006). Modeling and temperature control of rapid thermal processing. *Computers & chemical engineering*, *30*(4), 686-697.

Davis, L. (1991). *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York.

de Assis, A. J., and Maciel Filho, R. (2000). Soft sensors development for on-line bioreactor state estimation. *Computers & Chemical Engineering*, *24*(2), 1099-1103.

de Castro, L.N., Timmis, J. (2002). Artificial Immune Systems: A New Computational Intelligence Approach, 1st ed., Springer-Verlag, London, pp. 67–84.

Deb, K. (1995). *Optimization for Engineering Design: Algorithms and Examples*. Prentice-Hall, New Delhi.

Desai, K., Badhe, Y., Tambe, S. S., and Kulkarni, B. D. (2005). Soft-sensor development for fed-batch bioreactors using support vector regression. *Biochemical Engineering Journal*, *27*(3), 225-239.

Díaz, G., Sen, M., Yang, K. T., and McClain, R. L. (2001). Dynamic prediction and control of heat exchangers using artificial neural networks. *International Journal of Heat and Mass Transfer*, *44*(9), 1671-1679.

Dong, D., and McAvoy, T. J. (1996). Nonlinear principal component analysis—based on principal curves and neural networks. *Computers & Chemical Engineering*, *20*(1), 65-78.

Dorigo, M., Caro, G. Di., and Gambardella, L. M. (1999). Ant Algorithms for Discrete Optimization. *Artificial Life*, *5* (2), 137–172.

Dorigo, M., and Gambardella, L. M. (1997). Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. *IEEE Transactions on Evolutionary Computation, 1* (1), 53–66.

Dorigo, M., Birattari, M., and Stutzle, T. (2006). Ant colony optimization. *IEEE Computational Intelligence Magazine*, *1*(4), 28-39.

Dorigo, M., Maniezzo V., and Colorni, A. (1996). Ant System: Optimization by a Colony of Cooperating Agents. *IEEE Transactions on Systems, Man, and* Cybernetics–*Part B, 26* (1), 29–41.

Dragoi, E. N., Suditu, G. D., and Curteanu, S. (2012). Modeling methodology based on artificial immune system algorithm and neural networks applied to removal of heavy metals from residual waters. *Environmental Engineering and Management Journal*, *11*(11), 1907-1914.

Dudek, A. Z., Arodz, T., and Galvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial Chemistry & High Throughput Screening*, *9*(3), 213-228.

Eberhart, R. C., and Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Proceedings of the sixth international symposium on micro machine and human science* (pp. 39–43), Nagoya, Japan. Piscataway: IEEE.

Eberhart, R. C., Simpson, P. K., and Dobbins, R. W. (1996). *Computational Intelligence PC Tools*. Boston:Academic Press.

Elmolla, E. S., Chaudhuri, M., and Eltoukhy, M. M. (2010). The use of artificial neural network (ANN) for modeling of COD removal from antibiotic aqueous solution by the Fenton process. *Journal of Hazardous Materials*, *179*(1), 127-134.

Esfahanian, M., Rad, A. S., Khoshhal, S., Najafpour, G., and Asghari, B. (2016). Mathematical modeling of continuous ethanol fermentation in a membrane bioreactor by pervaporation compared to conventional system: Genetic algorithm. *Bioresource Technology*, *212*, 62-71.

Eslamloueyan, R., Khademi, M. H., and Mazinani, S. (2011). Using a multilayer perceptron network for thermal conductivity prediction of aqueous electrolyte solutions. *Industrial & Engineering Chemistry Research*, *50*(7), 4050-4056.

Farell, A. E., and Roat, S. D. (1994). Framework for enhancing fault diagnosis capabilities of artificial neural networks. *Computers & Chemical Engineering*, *18*(7), 613-635.

Faris, H., and Sheta, A. (2013). Identification of the tennessee eastman chemical process reactor using genetic programming. *International Journal of Advanced Science and Technology*, *50*, 121-140.

Farizhandi, A.A.K., Zhao, H., and Lau, R. (2016). Modeling the change in particle size distribution in a gas-solid fluidized bed due to particle attrition using a hybrid artificial neural network-genetic algorithm approach. *Chemical Engineering Science*, *155*, 210–220.

Fatemi, M. H., Baher, E., and Ghorbanzade'h, M. (2009). Predictions of chromatographic retention indices of alkylphenols with support vector machines and multiple linear regression. *Journal of Separation Science*, *32*(23-24), 4133-4142.

Fattah, K. A. (2012). K-value program for crude oil components at high pressures based on PVT laboratory data and genetic programming. *Journal of King Saud University-Engineering Sciences*, *24*(2), 141-149.

Fisher, R. A., and Mackenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *The Journal of Agricultural Science*, *13*(03), 311-320.

Freeman, J. A., and Skapura, D. M. (1991). *Neural Networks Algorithms, Applications, and Programming Techniques*. Addison-Wesley Publishing Company, Reading, M.A, USA.

Gandhi, A. B., Joshi, J. B., Kulkarni, A. A., Jayaraman, V. K., and Kulkarni, B. D. (2008). SVR-based prediction of point gas hold-up for bubble column reactor through recurrence quantification analysis of LDA time-series. *International Journal of Multiphase Flow*, *34*(12), 1099-1107.

García-Gonzalo, E., and Fernández-Martínez, J. L. (2012). A brief historical review of particle swarm optimization (PSO). *Journal of Bioinformatics and Intelligent Control*, *1*(1), 3-16.

Ge, M., Du, R., Zhang, G., and Xu, Y. (2004). Fault diagnosis using support vector machine with an application in sheet metal stamping operations. *Mechanical Systems and Signal Processing*, *18*(1), 143-159.

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*, 1-17.

Giordano, P. C., Beccaria, A. J., Goicoechea, H. C., and Olivieri, A. C. (2013). Optimization of the hydrolysis of lignocellulosic residues by using radial basis functions modeling and particle swarm optimization. *Biochemical Engineering Journal*, *80*, 1-9.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA.

Gonzaga, J. C. B., Meleiro, L. A. C., Kiang, C., and Maciel Filho, R. (2009). ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process. *Computers & Chemical Engineering*, *33*(1), 43-49.

Grosman, B., adn Lewin, D. R. (2002). Automated nonlinear model predictive control using genetic programming. *Computers & Chemical Engineering*, *26*(4), 631-640.

Gupta, P. P., Merchant, S. S., Bhat, A. U., Gandhi, A. B., Bhagwat, S. S., Joshi, J. B., Jayaraman, V.K., and Kulkarni, B. D. (2009). Development of correlations for overall gas hold-up, volumetric mass transfer coefficient, and effective interfacial area in bubble column reactors using hybrid genetic algorithm-support vector regression technique: viscous Newtonian and non-Newtonian liquids. *Industrial & Engineering Chemistry Research, 48*(21), 9631-9654.

Hamid, A., Deshpande, A. S., Badhe, Y. P., Barve, P. P., Tambe, S. S., and Kulkarni, B. D. (2014). Biodegradable iron chelate for H_2S abatement: Modeling and optimization using artificial intelligence strategies. *Chemical Engineering Research* and *Design*, *92*(6), 1119-1132.

Hezave, A. Z., Raeissi, S., and Lashkarbolooki, M. (2012). Estimation of thermal conductivity of ionic liquids using a perceptron neural network. *Industrial & Engineering Chemistry Research, 51*(29), 9886-9893.

Hiden, H. G., Willis, M. J., Tham, M. T., Turner, P., and Montague, G. A. (1997, September). Non-linear principal components analysis using genetic programming. In *Genetic Algorithms in Engineering Systems: Innovations and Applications, 1997. GALESIA 97. Second International Conference* (*Conf. Publ. No. 446*) (pp. 302-307). IET.

Himmelblau, D. M. (2000). Applications of artificial neural networks in chemical engineering. *Korean Journal of Chemical Engineering*, *17*(4), 373-392.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Univ. of Michigan Press, Ann Arbor.

Hoskins, J. C., Kaliyur, K. M., and Himmelblau, D. M. (1991). Fault diagnosis in complex chemical plants using artificial neural networks. *AIChE Journal*, *37*(1), 137-141.

Hosseini, S. H., Karami, M., Olazar, M., Safabakhsh, R., and Rahmati, M. (2014). Prediction of the minimum spouting velocity by genetic programming approach. *Industrial & Engineering Chemistry Research*, *53*(32), 12639-12643.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417.

Hunt, K. J., Sbarbaro, D., Żbikowski, R., and Gawthrop, P. J. (1992). Neural networks for control systems—A survey. *Automatica*, *28*(6), 1083-1112.

IBM SPSS. (2011).Neural Networks 20 manual, IBM: Chicago.

Ivanciuc, O. (2007). Applications of support vector machines in chemistry. *Reviews in Computational Chemistry*, *23*, 291-400.

Jayaraman, V. K., Kulkarni, B.D., Karale, S., and Shelokar, P. (2000) Ant colony framework for optimal design and scheduling of batch plants, *Computers & Chemical Engineering 24* (*8*), 1901-1912.

Kadlec, P., Gabrys, B., and Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, *33*(4), 795-814.

Kaur, K., Kaur, A., and Malhotra, R. (2006). Alternative methods to rank the impact of object oriented metrics in fault prediction modeling using neural networks. *Transactions on Engineering, Computing and Technology*, *13*, 207-212.

Kennedy, J., and Eberhart, R. C. (1995). Particle swarm optimization. In *Proceedings of the IEEE international Conference on Neural Networks IV* (pp. 1942–1948). Piscataway: IEEE.

Khansary, M. A., and Sani, A. H. (2014). Using genetic algorithm (GA) and particle swarm optimization (PSO) methods for determination of interaction parameters in multicomponent systems of liquid–liquid equilibria. *Fluid Phase Equilibria*, *365*, 141-145.

Kinnear, K. E. (1994). *Advances in Genetic Programming* (Vol. 1). MIT press, Cambridge, MA.

Koza, J. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA.

Kruger, U., Zhang, J., and Xie, L. (2008). Developments and applications of nonlinear principal component analysis–A review. In *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer, Berlin, Heidelberg. pp.1-43.

Lahiri, S. K., and Ghanta, K. C. (2008). Prediction of pressure drop of slurry flow in pipeline by hybrid support vector regression and genetic algorithm model. *Chinese Journal of Chemical Engineering*, *16*(6), 841-848.

Ławryńczuk, M. (2016). Modelling and Predictive Control of a neutralisation reactor using sparse support vector machine wiener models. *Neurocomputing*. *205*, 311-328.

Lazzús, J. A. (2013). Thermodynamic modeling based on particle swarm optimization to predict phase equilibrium of binary systems containing ionic liquids. *Journal of Molecular Liquids*, *186*, 44-51.

Lee, M. J., and Chen, J. T. (1993). Fluid property predictions with the aid of neural networks. *Industrial & Engineering Chemistry Research*, *32*(5), 995-997.

Leong, C. C., Blakey, S., and Wilson, C. W. (2016). Genetic Algorithm optimised Chemical Reactors network: A novel technique for alternative fuels emission prediction. *Swarm and Evolutionary Computation*, *27*, 180-187.

Lipowski, A., and Lipowska, D. (2012). Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications*, *391*(6), 2193-2196.

Loucks, D. P., Van Beek, E., Stedinger, J. R., Dijkman, J. P., and Villars, M. T. (2005). *Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications*. Paris: UNESCO.

Lucasius, C. B., and Kateman, G. (1993). Understanding and using genetic algorithms Part 1. Concepts, properties and context. *Chemometrics and Intelligent Laboratory Systems*, *19*(1), 1-33.

Lucasius, C. B., and Kateman, G. (1994). Understanding and using genetic algorithms Part 2. Representation, configuration and hybridization. *Chemometrics and Intelligent Laboratory Systems*, *25*(2), 99-145.

Maniezzo, V., and Carbonaro, A. (2002). Ant colony optimization: An overview. In *Essays and Surveys in Metaheuristics* (pp. 469-492). Springer US.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431-441. http://dx.doi.org/10.1137/0111030.

Martens, D., De Backer, M., Haesen, R., Vanthienen, J., Snoeck, M., and Baesens, B. (2007). Classification with ant colony optimization. *IEEE Transactions on Evolutionary Computation*, *11*(5), 651-665.

McKay, B., Willis, M., and Barton, G. (1997). Steady-state modelling of chemical process systems using genetic programming. *Computers & Chemical Engineering*, *21*(9), 981-996.

Michalopoulos, J., Papadokonstadakis, S., Arampatzis, G., and Lygeros, A. (2001). Modelling of an industrial fluid catalytic cracking unit using neural networks. *Chemical Engineering Research and Design*, *79*(2), 137-142.

Miller, B. L., and Goldberg, D. E. (1995). Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, *9*(3), 193-212.

Mohanty, S. (2005). Estimation of vapour liquid equilibria of binary systems, carbon dioxide–ethyl caproate, ethyl caprylate and ethyl caprate using artificial neural networks. *Fluid Phase Equilibria*, *235*(1), 92-98.

Mohanty, S. (2009). Artificial neural network based system identification and model predictive control of a flotation column. *Journal of Process Control*, *19*(6), 991-999.

Montano, J. J., and Palmer, A. (2003). Numeric sensitivity analysis applied to feedforward neural networks. *Neural Computing & Applications*, *12*(2), 119-125.

Moore, B. (1981). Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, *26*(1), 17-32.

Moore, S. W., Gardner, J. W., Hines, E. L., Göpel, W., and Weimar, U. (1993). A modified multilayer perceptron model for gas mixture analysis. *Sensors and Actuators B: Chemical*, *16*(1-3), 344-348.

Moral, H., Aksoy, A., and Gokcay, C. F. (2008). Modeling of the activated sludge process by using artificial neural networks with automated architecture screening. *Computers & Chemical Engineering*, *32*(10), 2471-2478.

Nandi, S., Badhe, Y., Lonari, J., Sridevi, U., Rao, B. S., Tambe, S. S., and Kulkarni, B. D. (2004). Hybrid process modeling and optimization strategies integrating neural networks/support vector regression and genetic algorithms: study of benzene isopropylation on Hbeta catalyst. *Chemical Engineering Journal*, *97*(2), 115-129.

Nandi, S., Ghosh, S., Tambe, S. S., and Kulkarni, B. D. (2001). Artificial neural-network-assisted stochastic process optimization strategies. *AIChE Journal*, *47*(1), 126-141.

Nandi, S., Rahman, I., Tambe, S.S., Sonolikar, R.L., and Kulkarni, B.D. (2000). Process identification using genetic programming: A case study involving Fluidized catalytic cracking (FCC) unit. In *Petroleum Refining and Petrochemicals Based Industries in Eastern India*; Saha, R.K., Maity, B.R., Bhattacharyya. D., Ray. S., Ganguly. S., and Chakraborty, S.L. (Eds.), Allied Publishing Ltd., New Delhi, pp. 195-201.

Nomikos, P. (1996). Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis World Batch Forum, Toronto, May 1996. *ISA Transactions*, *35*(3), 259-266.

Nomikos, P., and MacGregor, J. F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, *40*(8), 1361-1375.

Olden, J. D., and Jackson, D. A. (2002). Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, *154*(1), 135-150.

Olden, J. D., Joy, M. K., and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, *178*(3), 389-397.

Patel, S. U., Kumar, B. J., Badhe, Y. P., Sharma, B. K., Saha, S., Biswas, S., Chaudhury, A., Tambe, S.S., and Kulkarni, B. D. (2007). Estimation of gross calorific value of coals using artificial neural networks. *Fuel*, *86*(3), 334-344.

Patil-Shinde, V., Kulkarni, T., Kulkarni, R., Chavan, P. D., Sharma, T., Sharma, B. K., Tambe, S. S., and Kulkarni, B. D. (2014). Artificial intelligence-based modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier. *Industrial & Engineering Chemistry Research*, *53*(49), 18678-18689.

Patil-Shinde, V., Saha, S., Sharma, B. K., Tambe, S. S., and Kulkarni, B. D. (2016). High ash char gasification in Thermo-gravimetric analyzer and prediction of gasification performance parameters using computational intelligence formalisms. *Chemical Engineering Communications*, *203*(8), 1029-1044.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2* (6), 559–572.

Pollard, J. F., Broussard, M. R., Garrison, D. B., and San, K. Y. (1992). Process identification using neural networks. *Computers & Chemical Engineering*, *16*(4), 253-270.

Puig-Arnavat, M., Hernández, J. A., Bruno, J. C., and Coronas, A. (2013). Artificial neural network models for biomass gasification in fluidized bed gasifiers. *Biomass and Bioenergy*, *49*, 279-289.

Rangasamy, P., Iyer, P. V. R., and Ganesan, S. (2007). Anaerobic tapered fluidized bed reactor for starch wastewater treatment and modeling using multilayer perceptron neural network. *Journal of Environmental Sciences*, *19*(12), 1416-1423.

RapidMiner. (2011). http://rapid-i.com/content/view/181/190/lang.en/

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating error. *Nature, 323*(9), 533-536.

Sadeghzadeh, H., Ehyaei, M. A., and Rosen, M. A. (2015). Techno-economic optimization of a shell and tube heat exchanger by genetic and particle swarm algorithms. *Energy Conversion and Management, 93*, 84-91.

Sankpal, N. V., Cheema, J. J. S., Tambe, S. S., and Kulkarni, B. D. (2001). An artificial intelligence tool for bioprocess monitoring: application to continuous production of gluconic acid by immobilized Aspergillus niger. *Biotechnology Letters, 23*(11), 911-916.

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science, 324*(5923), 81-85.

Sharma, R., Singhal, D., Ghosh, R., and Dwivedi, A. (1999). Potential applications of artificial neural networks to thermodynamics: Vapor–liquid equilibrium predictions. Computers *&* *Chemical Engineering*, *23*(3), 385-390.

Sharma, S., and Tambe, S. S. (2014). Soft-sensor development for biochemical systems using genetic programming. *Biochemical Engineering Journal, 85*, 89-100.

Shi, Y. (2001). Particle swarm optimization: developments, applications and resources. In *evolutionary computation, 2001. Proceedings of the 2001 Congress on* (Vol. 1, pp. 81-86). IEEE.

Shokrkar, H., Salahi, A., Kasiri, N., and Mohammadi, T. (2012). Prediction of permeation flux decline during MF of oily wastewater using genetic programming. *Chemical Engineering Research and Design, 90*(6), 846-853.

Simpsons, P. (1990). *Artificial Neural Systems: Foundations, Paradigms, Applications, and Implimentations*. Pergamon Press, New York.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245.

Sumpter, B. G., and Noid, D. W. (1996). On the use of computational neural networks for the prediction of polymer properties. *Journal of Thermal Analysis and Calorimetry*, *46*(3-4), 833-851.

Szpiro, G. G. (1997). Forecasting chaotic time series with genetic algorithms. *Physical Review E*, *55*(3), 2557-2568.

Tambe, S. S., Kulkarni, B. D., and Deshpande, P. B. (1996). *Elements of Artificial Neural Networks with Selected Applications in Chemical Engineering, and Chemical & Biological Sciences*. Simulation & Advanced Controls Inc., Louisville, K.Y.

Thibault, J., and Grandjean, B. P. (1991). A neural network methodology for heat transfer data analysis. *International Journal of Heat and Mass Transfer*, *34*(8), 2063-2070.

Thierens, D., and Goldberg, D. (1994, June). Elitist recombination: An integrated selection recombination GA. In *Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference*, pp. 508-512, IEEE.

Tsai, P. F., Chu, J. Z., Jang, S. S., and Shieh, S. S. (2003). Developing a robust model predictive control architecture through regional knowledge analysis of artificial neural networks. *Journal of Process Control*, *13*(5), 423-435.

Tzeng, D. Y., and Berns, R. S. (2005). A review of principal component analysis and its applications to color technology. *Color Research & Application*, *30*(2), 84-98.

UNESCO Report. (2005). Water Resources Systems Planning and Management – ISBN 92-3-103998-9, chapter 9.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. 2nd ed., Springer Verlag, New York. ISBN 978-1-4419-3160-3.

Vapnik, V. (1998). *Statistic Learning Theory*. Willey, New York.

Velásco-Mejía, A., Vallejo-Becerra, V., Chávez-Ramírez, A. U., Torres-González, J., Reyes-Vidal, Y., and Castañeda-Zaldivar, F. (2016). Modeling and optimization

of a pharmaceutical crystallization process by using neural networks and genetic algorithms. *Powder Technology*, *292*, 122-128.

Venkatasubramanian, V., and Sundaram, A. (1998). Genetic algorithms: Introduction and Applications. *Encyclopedia of Computational Chemistry*.

Verma, D., Goel, P., Patil-Shinde, V., and Tambe, S. S. (2016, January). Use genetic programming for selecting predictor variables and modeling in process identification. In *IEEE explore, 2016 Indian Control Conference* (*ICC*) (pp. 230-237). IEEE. (ISBN: 978-1-4673-7992-2), doi: 10.1109/INDIANCC.2016.7441133.

Vyas, R., Goel, P., and Tambe, S. S. (2015). Genetic programming applications in chemical sciences and engineering. In *Handbook of Genetic Programming Applications*; Gandomi, A.H., Amir H., Alavi, Ryan, C. (Eds.), Springer International Publishing, Switzerland, pp 99–140. doi:http://dx.doi.org/10.1007/978-3-319-20883-1

Wang, H., Song, Z., and Li, P. (2002). Fault detection behavior and performance analysis of principal component analysis based process monitoring methods. *Industrial & Engineering Chemistry Research*, *41*(10), 2455-2464.

Wang, J., Cui, G., Xiao, Y., Luo, X., and Kabelac, S. (2016). Bi-level heat exchanger network synthesis with evolution method for structure optimization and memetic particle swarm optimization for parameter optimization. *Engineering Optimization*, 1-16.

Wang, S., and Cui, J. (2005). Sensor-fault detection, diagnosis and estimation for centrifugal chiller systems using principal-component analysis method. *Applied Energy*, *82*(3), 197-213.

Wang, X. H., Li, Y. G., Hu, Y. D., and Wang, Y. L. (2008). Synthesis of heat-integrated complex distillation systems via Genetic Programming. *Computers & Chemical Engineering*, *32*(8), 1908-1917.

Willis, M., Hiden, H., Hinchliffe, M., McKay, B., and Barton, G. W. (1997). Systems modelling using genetic programming. *Computers & Chemical Engineering*, *21*, S1161-S1166.

Witczak, M., Obuchowicz, A., and Korbicz, J. (2002). Genetic programming based approaches to identification and fault diagnosis of non-linear dynamic systems. *International Journal of Control*, *75*(13), 1012-1031.

Wold, H. (1966), Nonlinear estimation by iterative least squares procedures. In *Research Papers in Statistics*; F. David (Ed.), pp. 411–444, Wiley, New York,

Yan, W., Shao, H., and Wang, X. (2004). Soft sensing modeling based on support vector machine and Bayesian model selection. *Computers & Chemical Engineering*, *28*(8), 1489-1498.

Yao, X. J., Panaye, A., Doucet, J. P., Zhang, R. S., Chen, H. F., Liu, M. C., Hu, Z. D., and Fan, B. T. (2004). Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *Journal of Chemical Information and Computer Sciences*, *44*(4), 1257-1266.

Yi, Z. S., and Qin, L. T. (2007). Support Vector Machines QSAR for the toxicity of organic chemicals to Chlorella Vulgaris with SVM parameters optimized with simplex. *Internet Electronic Journal of Molecular Design*, *6*(8), 229-236.

Yousefinejad, S., and Hemmateenejad, B. (2015). Chemometrics tools in QSAR/QSPR studies: a historical perspective. *Chemometrics and Intelligent Laboratory Systems*, *149*, 177-204.

Zhang, L., and Nandi, A. K. (2007). Fault classification using genetic programming. *Mechanical Systems and Signal Processing*, *21*(3), 1273-1284.

Zhang, R., and Wang, S. (2008). Support vector machine based predictive functional control design for output temperature of coking furnace. *Journal of Process Control*, *18*(5), 439-448.

Zhang, Z., and Friedrich, K. (2003). Artificial neural networks applied to polymer composites: A review. *Composites Science and Technology*, *63*(14), 2029-2044.

Zheng, L., Zhou, H., Wang, C., and Cen, K. (2008). Combining support vector regression and ant colony optimization to reduce NOx emissions in coal-fired utility boilers. *Energy & Fuels*, *22*(2), 1034-1040.

Zonouz, P. R., Niaei, A., and Tarjomannejad, A. (2016). Modeling and optimization of toluene oxidation over perovskite-type nanocatalysts using a hybrid artificial neural network-genetic algorithm method. *Journal of the Taiwan Institute of Chemical Engineers, 65*, 276-285.

Zurada, J.M. (1992). *Introduction to Artificial Neural Network.* West Publ Co., St. Paul.

Chapter 3

Artificial Intelligence-based Modeling of High Ash Coal Gasification in a Pilot Plant Scale Fluidized Bed Gasifier

ABSTRACT

The quality of coal—especially its high ash content—significantly affects the performance of coal-based processes. Coal gasification is a cleaner and an efficient alternative to the coal combustion for producing the syngas. The high-ash coals are found in a number of countries, and they form an important source for the gasification. Accordingly, in this study, extensive gasification experiments were conducted in a pilot-plant scale fluidized-bed coal gasifier (*FBCG*) *using high-ash coals from India. Specifically, the effects of eight coal and gasifier process related parameters on the four gasification performance variables, namely* $CO+H_2$ *generation rate, syngas production rate, carbon conversion, and heating value of the syngas, were rigorously studied. The data collected from these experiments were used in the FBCG modeling, which was conducted by utilizing two artificial intelligence* (*AI*) *strategies namely genetic programming* (*GP*) *and artificial neural networks* (*ANNs*). *The novelty of the GP formalism is that it searches and optimizes both the form and parameters of an appropriate linear/nonlinear function that best fits the given process data. The original eight-dimensional input space of the FBCG models was reduced to three-dimensional space using the principal component analysis* (*PCA*) *and the PCA-transformed three variables were used in the AI-based FBCG modeling. A comparison of the GP and ANN-based models reveals that their output prediction accuracies and the generalization performance vary from good to excellent as indicated by the high training and test set correlation coefficient magnitudes lying between 0.92 and 0.996. This study also presents results of the sensitivity analysis performed to identify those coal and process related parameters, which significantly affect the FBCG process performance.*

3.1 INTRODUCTION

The widely employed coal-based combustion technologies for the power generation suffer from a number of drawbacks, such as lower efficiency, and significant emissions of CO_2, SO_x, and NO_x gases. These emissions lead to the climate change, and air pollution. An important factor that influences the performance of a thermal power station is the quality of the coal used in the combustion. Specifically, usage of the high ash (>20%) coal produces following adverse effects: (a) emission of more particulate matter into the atmosphere, (b) reduced power station boiler efficiency leading to consumption of higher volumes of coal to achieve the targeted power output, which results in higher coal transportation costs and, consequently, costlier power, and (c) higher levels of impurities from the coal (e.g., ash and moisture) that do not contribute to the combustion process; these also lead to severe waste disposal problems.

There are a number of countries such as India, China, Australia, and Turkey, where high ash coal deposits are found. In India, coal-based energy meets nearly 70% of the country's energy needs. The Indian thermal power stations invariably receive high ash coals and, therefore, CO_2 emission control has become a major concern. For achieving the stringent pollution control targets, changes in the coal utilization practices and the development of clean coal technologies have become essential. These measures are expected to result in a high coal conversion efficiency and lower environmental impact (Takematsu and Maude, 1991). The gasification of coal is such a promising clean coal technology (Miller, 2011). The typical thermal efficiencies of the conventional pulverized-fuel (PF)-fired power stations are approaching 37%, whereas supercritical PF units can achieve net efficiencies of 47% (Clean coal technology, 2000). In comparison, power generation using an *Integrated Gasification Combined Cycle* (IGCC) system has achieved thermal efficiencies of approximately 47% (Heaven, 1996) and it is believed that the efficiencies exceeding 50% are possible in the near future (Clean coal technology, 2000; Davidson, 1983). The newer gas turbine concepts and increased process temperatures are targeting efficiencies up to 65% (Davidson, 1983).

The gasification technology, being environment-friendly is a potential alternative to the conventional coal combustion-based power generation. The

conventional thermal power plants with steam cycles alone cannot achieve the high efficiency targets, and hydrogen production from the combustion plants is not feasible. These limitations are not applicable to the gasification technologies and they possess several other advantages as well due to their flexibility in the syngas applications. There exist three major coal gasification technologies, namely moving (fixed) bed, fluidized bed, and entrained bed gasifiers. Among these, the fluidized bed coal gasifier (FBCG) possesses following advantages (Chavan, 2012):

- Process conditions in an FBCG are more uniform leading to better heat and mass transfer in the bed and steady product composition.
- Provides better contact between the solid and gaseous reactants, which is favorable for maximizing carbon conversion.
- It has a high solids residence time, can use bigger particle sizes, and is capable of handling the high-sulfur coals without a need for the flue-gas desulfurization systems, which incur high capital and operational costs.
- It operates at lower temperatures and thus emits lower amounts of nitrogen oxides (NO_x). The low temperature process, besides improving the system's reliability, is also inherently more energy efficient since it consumes nearly all the heat generated in the gasifier in supporting the gasification.
- The FBCG can handle a wide variety of coals ranging from the high quality bituminous coal to the lignite. For more reactive fuels, such as the sub-bituminous coal and lignite, the fluidized bed gasifiers can achieve good gasification yields and carbon conversion at relatively mild conditions.
- The tar and phenol formation is low or negligible.
- The large fuel inventory provides safety, reliability, and stability.
- Potential for in situ sulfur capture.
- Better turn-down ratio.

Due to the above-stated several attractive characteristics, the fluidized bed coal gasifiers are better suited—in comparison with the other types of gasifiers—for handling high ash coals.

A large number of studies have been performed to investigate the fluidized bed coal gasification (see, for example, Gutierrez and Watkinson, 1982; Ocampo et al., 2003; and Ju et al., 2010). However, a detailed experimental investigation and

analysis of the fluidized bed gasification with a focus on the high ash coals cannot found in the open scientific literature. Such a study is important since in countries like India a large portion of the electrical energy comes from the coal-fired thermal power stations and there exists a dire need to switch over to more efficient clean coal technologies, such as gasification. Another reason for studying the FBCG in depth is the following: the gasifier operates in the dry ash removal mode owing to which the operating temperature needs to be lower. This may result in an unconverted carbon in the fly and bottom ashes. To address the issues arising from the low temperature FBCG operation, an in-depth investigation of the various factors affecting the FBCG performance has become necessary. Accordingly, this study first reports the results of the fluidized bed gasification in a pilot-plant scale gasifier using high ash coals of Indian origin.

Availability of an accurate, robust, and reliable mathematical process model of an FBCG assists in the preliminary process design, complex simulation, prediction of the steady-state and dynamic behavior, startup, shutdown, change of fuel and load, scaling up, control, fault detection and diagnosis, and process optimization. Such models are also helpful in fixing the right magnitude of the bed temperature so as to (i) avoid bed agglomeration and incomplete char conversion (due to lower temperatures), (ii) avert tar formation (owing to high temperatures), and (iii) ensure high gasification efficiency. Conducting experiments, especially at a large scale, is often expensive, complicated and time-consuming task; modeling can save time and money (Gómez-Barea and Leckner, 2010). Owing to these advantages, a great deal of effort has been spent over the last five decades toward mathematically modeling different types of gasifiers.

3.1.1 Phenomenological Modeling of Fluidized Bed Coal Gasification

Commonly, *phenomenological* ("*first principles*" or "*mechanistic*") models of an FBCG are developed for gaining design and performance related information on the process operating under a variety of reaction conditions. These models incorporate complex and nonlinear reaction and mass and heat transport phenomena. Different types of models can be developed for the FBCG—from the simple black-box or zero-dimensional models, where mass and heat balances are made over the entire gasifier to predict the exit gas composition, to the complex non-isothermal, three-dimensional

ones taking into account the fluid dynamics and thermal behavior. Owing to the sheer complexity of the underlying physicochemical phenomena, the phenomenological modeling of the coal gasification/gasifier is a challenging task. This task is often simplified by making a number of assumptions regarding the numerous mechanisms underlying the gasification process.

Broadly, two types of *phenomenological* models, namely thermodynamic (equilibrium) and kinetic (rate), are developed for the FBCG (Lee, 2007). The models belonging to the first category are independent of the gasifier type and assume complete oxygen consumption. Being independent of the gasifier type, these models are not useful for examining the effects of operating parameters on the gasifier behavior. The other type, i.e., *kinetic* models, comprises an appropriate hydrodynamic model of the fluidized bed coupled with the kinetics of various reactions occurring in the gasifier. Given a set of operating conditions of a specific type of a gasifier, its kinetic model is capable of predicting the process behavior in terms of, for instance, product composition, and temperature profiles. In the phenomenological modeling of an FBCG, once the model structure is fixed, then a large number of kinetic, thermodynamic, and heat and mass transport related parameters appearing in the model need to be determined either by conducting experiments and/or by simulation. Some notable representative studies as also reviews pertaining to the phenomenological modeling (including computational fluid dynamics modeling) of the fluidized bed gasification can be found in Rhinehart et al. (1987), Sett and Bhattacharya (1988), de Souza-Santos (1989), Goyal et al. (1989), Gururajan et al. (1992), Lim et al. (1995), Witt and Perry (1996), Witt et al. (1997), Donne et al. (1998), Moorea-Taha (2000), Villanueva et al. (2008), Mazumder (2010), Armstrong et al. (2011), Irfan et al. (2011), Yang et al. (2012), Xiangdong et al. (2013), and Singh et al. (2014).

The specific difficulties encountered in the phenomenological modeling of gasification processes are: (i) nonlinear interplay of a number of process variables, (ii) lengthy throughput dependent process dynamics, (iii) cost-intensive and exhaustive experimentation required for studying the effects of influential operating variables and parameters, and (iv) unavailability of an in-depth knowledge of the physicochemical phenomena (e.g., kinetics, heat and mass transport mechanisms) underlying the coal gasification phenomena.

3.1.2 Alternate FBCG Modeling Strategies

An alternative approach to the phenomenological modeling of the gasification process is to utilize regression methods to formulate *empirical* models. However, in this approach the exact form of the data fitting function (model) needs to be specified before the function parameters could be estimated. This is a difficult task since in the gasification process multiple factors influence the nonlinear gasification phenomena and the precise interactions between them are not fully known. The complexities involved in the phenomenological and regression-based modeling of FBCG necessitate exploration of alternative nonlinear modeling strategies that do not require full details of the underlying physicochemical phenomena. The AI-based modeling strategies, for example, *artificial neural networks* (ANNs), and the statistical *machine learning* (ML) theory-based formalism, namely *support vector regression* (SVR), are exclusively data-driven strategies and thus these can be used for modeling FBCG. There exist a number of studies wherein ANNs have been employed in the energy-related science and engineering (Mjalli and Al- Mfargi, 2008; Nougues et al., 2000; Liukkonen et al., 2012; Puig-Arnavat et al., 2013; Behera, 2014). In an exhaustive data driven modeling study of FBCG, Chavan et al. (2012) developed two ANN based models for the prediction of gas production rate and heating value of the product gas, using process data from 18 globally located coal gasifiers. These models use six inputs namely, *fixed carbon*, *volatile matter*, *mineral matter*, *air feed per kilogram of coal*, *steam feed per kilogram of coal*, and *temperature*. Despite their potential, however, the AI and ML based strategies have been only rarely employed in the modeling of fluidized bed gasifiers.

Apart from the ANNs, the AI comprises a novel exclusively data-driven modeling formalism, namely *genetic programming* (GP). The uniqueness of the GP methodology is that given an example input-output data set, it is capable of searching and optimizing both, the specific form (structure) and the parameters, of an appropriate linear/nonlinear data-fitting function and unlike ANNs, the GP does this without making any assumptions regarding the structure and parameters of the data-fitting function. A detailed description of GP formalism is given in Chapter 2 (section 2.2.2). Despite its novelty, the GP has not been used widely for the data-driven modeling applications in chemical engineering/technology to the same extent as the ANNs and SVR. Accordingly, the principal objectives of this paper are (i) to

rigorously study the gasification of the high ash Indian coals in the pilot-plant scale FBCG and (ii) to develop GP-based models for the prediction of four FBCG performance variables, namely CO+ H_2 *generation rate* (y_1) (kg/kg coal), *syngas production rate* (y_2) (kg/kg coal), *carbon conversion* (y_3) (%), and *heating value of the syngas* (y_4) (kcal/Nm^3). In the FBCG modeling, following eight process variables and parameters have been used as the model inputs: *fuel ratio* (fixed carbon/volatile matter) (x_1), *ash content of coal* (x_2) (wt %), *specific surface area of coal* (x_3) (m^2/g), *activation energy of gasification* (x_4) (kJ/mol), *coal feed rate* (x_5) (kg/h), *gasifier bed temperature* (x_6) (°C), *ash discharge rate* (x_7) (kg/h) and *air/coal ratio* (x_8) (kg/kg coal). The novel features of the present study are as follows.

a) Extensive experimentation has been conducted for studying the high ash coal gasification under steady state conditions in a pilot-plant scale FBCG located at the *Central Institute of Mining and Fuel Research* (CIMFR), Dhanbad, India.
b) A rigorous literature search shows that this is the first study wherein the GP strategy has been employed for the data-driven modeling in the coal-related energy science and engineering.
c) The *principal component analysis* (PCA) has been performed for reducing the dimensionality of the models' eight-dimensional input space representing the various coal and gasifier parameters.
d) The *sensitivity analysis* of the eight model inputs has been performed to gauge their influence on the four process performance variables.

Conventionally, phenomenological models for the coal gasification/gasifier use four types of inputs: (i) coal properties (proximate and/or ultimate analysis), (ii) process operating variables and parameters (reactor temperature, pressure, feed rate, etc.), (iii) physicochemical parameters (e.g., surface area) of the coal, and (iv) reaction kinetics parameters. In an earlier study on the FBCG modeling, Chavan et al. (2012) used the first two types of inputs. The present FBCG modeling study, in comparison, utilizes all the four types of inputs owing to which the input space now contains additional information pertaining to the physicochemical phenomena occurring in the gasifier. The importance of the selected eight model inputs ($x_1 - x_8$) is described below.

- The extents of fixed carbon and volatile matter reflect the effect of hydrogen and oxygen containing functional groups in the coal. This effect is represented in the form of fuel ratio (FC/VM) (x_1) as a model input.
- Ash (wt %) (x_2) is an indicator of coal's mineral matter content. Several formulas have been proposed for converting the ash content in the coal to its mineral matter content. One of the oldest yet still widely used correlation is given by Parr (1932):

$$mineral\ matter\ (wt\ \%) = (1.08 \times ash\ (wt\ \%)) + (0.55 \times S\ (wt\ \%)) \quad (3.1)$$

  where S refers to the sulfur content in the coal. Since the sulfur (as also carbonate) content of the Indian coals is low, a simplified version of the Parr correlation (Eq. 3.1) is used for these coals as given by Choudhury (2013):

$$mineral\ matter\ (wt\ \%) = 1.1 \times ash\ (wt\ \%) \quad (3.2)$$

  The ash in the coal is responsible for lowering the carbonaceous material in the coal matrix thereby negatively affecting the quality of the product gas.
- The rate of gasification depends on the accessibility of the reactant gases to the internal surface of the porous coal where active sites reside. Accordingly, specific surface area (m^2/g) (x_3), has been considered as a model input.
- In the coal gasification process, the char$-CO_2$ gasification is one of the rate controlling steps and hence the activation energy (kJ/mol) (x_4) of the char$-CO_2$ gasification reaction forms an input to the model.
- The coal feed rate (kg/h) (x_5) has been chosen because it defines the flow rate of the basic carbonaceous raw material.
- The significance of the model input, namely gasifier bed temperature (°C) (x_6), is that, as its magnitude increases, the product gas generation per kilogram of the coal increases. Also, higher gasification temperature results in the faster pyrolysis generating an increased amount of the CO_2, which in turn gets converted to the CO via the Boudouard reaction.
- The ash discharge rate (kg/h) (x_7) has been considered as a model input (Satonsaowapak et al., 2011) because together with the coal feed rate (x_5), it significantly influences the residence time of the coal particles in the gasifier bed. For instance, the residence time decreases with an increase in the coal

feed rate (x_5) or the ash withdrawal rate (x_7). This results in the lower product gas generation per kilogram of the coal feed. Accordingly, in the low temperature gasification, it is necessary to allow a sufficient residence time for the coal particles to achieve maximum carbon conversion and product gas generation per kg of the coal (Chavan et al., 2012; Kim et al., 1997).

- The air/coal ratio (kg/kg of coal) (x_8) is an important gasification process parameter since (Ocampo et al., 2003; Kim et al., 1997): (a) the air-assisted oxidation of the carbon is one of the key reactions for attaining the desired temperature for the gasification, (b) an increase in the air/coal ratio increases the carbon conversion, and (c) an excessive air/coal ratio decreases the heating value of the syngas thereby negatively affecting the performance of the gasification process.

In this study, the prediction accuracies and generalization capabilities of the four GP-based models have been compared with those of the corresponding four multilayer perceptron neural network (MLPNN) based models. This comparison indicates that both types of models possess an excellent ability to predict the magnitudes of the four gasifier performance variables.

The structure of the chapter is as follows. The details of the FBCG and the experiments conducted thereof are given in "experimental section" (section 3.2), Section 3.3 titled ''Results and Discussion'' first presents the results pertaining to (a) the sensitivity analysis of the eight model inputs (section 3.3.1), (b) artificial intelligence (AI)-based FBCG Modeling (section 3.3.2), which includes the details of development of the GP and MLP-based FBCG modeling followed by a comparison of prediction and generalization performance of GP and MLP-based models. Finally, section 3.4 summarizes the principal findings of the study.

3.2 EXPERIMENTAL SECTION

In this study, four types of Indian coals with ash content varying between 27% and 48% have been used. Their basic properties were evaluated by the proximate and ultimate analyses (see Table 3. 1) carried out according to the Indian standards, viz. IS: 1350 (Part-I) 1984, IS: 1350 (Part-III) 1969, IS: 1350 (Part-IV/Sec-1) 1974, IS: 1350 (Part-IV/Sec-2) 1975. The specific surface area of the coal samples was measured using Tristar 3000 surface area analyzer (Micromeritics, U.S.A.).

Coal gasification comprises two steps: *pyrolysis* and *char gasification* (Ollero et al., 2003). The kinetics of the char gasification was investigated in the reaction rate controlling regime. Here, char−CO_2 gasification kinetic parameters, such as the gasification rate constant and the gasification activation energy thereof, were evaluated using the laboratory scale thermo-gravimetric analyzer (TGA) by following the procedure given in Shaw et al. (1997), Beamish et al. (1998), and Çakal et al. (2007).

3.2.1 FBCG Pilot Plant

Gasification experiments were conducted in an air-blown FBCG pilot plant (Figure 3.1). The gasifier with a capacity to handle 10−20 kg coal/h and operating temperature <1000 °C consists of the following subsystems: coal feeding system, gasifying agent feeding system, fluidized bed gasifier, ash extraction system, cyclone separator, syngas cooling and cleaning system, and flare stack. Gasification experiments were conducted using four types of coals by varying the process operating conditions. The gasifier temperature was raised by an external electric heating system. The preheated air (200−250 °C) and the superheated steam (200−250 °C) were mixed using an air/steam mixer and fed to the gasifier through a conical distributor. Ash in the bed was extracted at a controlled rate and cooled to ≈ 40 °C prior to discharging in the ash bin. The hot dusty raw fuel gases leave the gasifier from the freeboard section and enter the cyclone where most of the elutriated particles are captured.

The fuel gas from the cyclone enters the quench column. Following the softening treatment, water from the settler tank was directly sprayed by spray nozzles on the hot syngas to reduce its temperature. The dust laden water gets collected in the seal pot situated at the bottom of the quench pipe. The cooled gas exits from the top side in the seal pot and passes through a venturi scrubber wherein any left-over acidic contents of the cooled gas are cleaned further. Prior to flaring, the clean gas from the knockout drum is transferred through a system pressure control valve and water-sealed flare stack. By using the water displacement method, the syngas samples were extracted in the glass pipettes through a sample collection port located at the downstream of the knockout drum. These gas samples were analyzed using an offline gas chromatograph (Model GC 1000; Chemito, India). The ranges of the various

constituents of the syngas given by the GC analysis were: CO (%) 10−22; H_2 (%) 10−22; CO_2 (%) 10−25; N_2 (%) 45−60; and CH_4 (%) < 2.

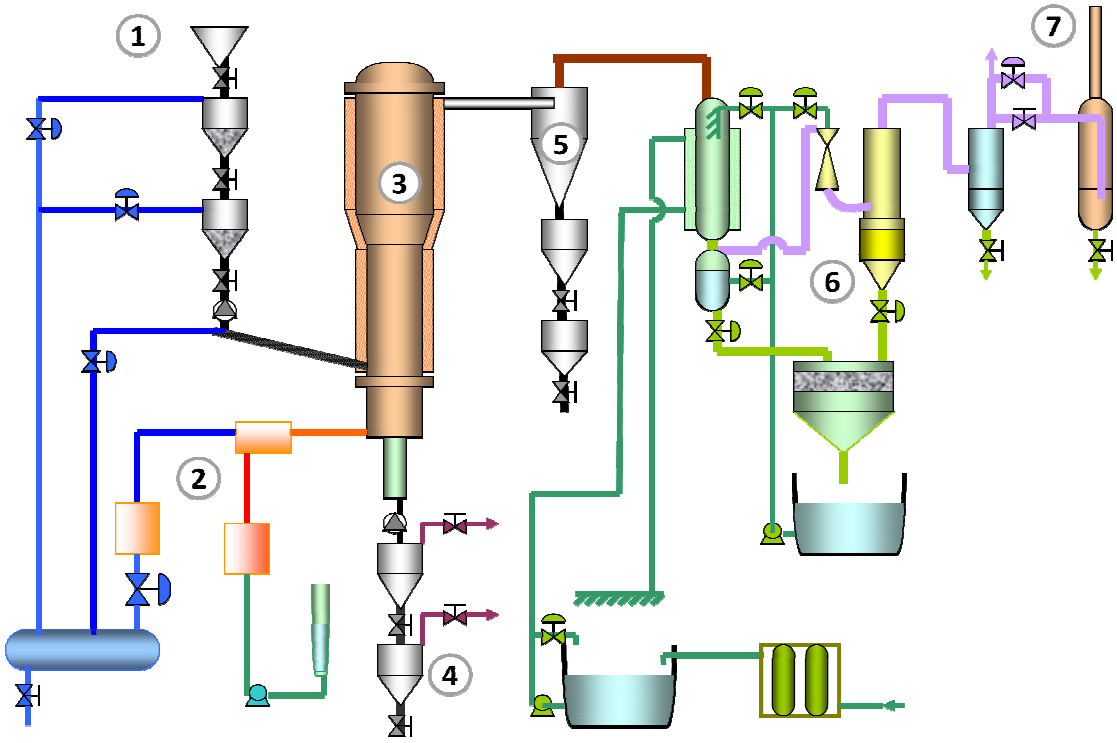


Figure 3.1: Fluidized bed gasification pilot plant consisting of process elements: (1) Coal feeding system, (2) Gasifying agent feeding system, (3) Fluidized bed gasifier, (4) Ash extraction system, (5) Cyclone separator, (6) Syngas cooling and cleaning system, and (7) Flare stack.

Table 3.1: Analysis of Coal Samples (Air Dried Basis)

Coal	Proximate analysis				Ultimate analysis				
	Ash (wt %) (x_2)	Moisture (wt %)	Volatile Matter (wt %)	Fixed Carbon (wt %)	*C* (%)	*H* (%)	*N* (%)	*S* (%)	*O* (%)*
C1	41.3	6.5	24.5	27.7	37.15	2.83	0.86	0.55	6.68
C2	48.9	7.1	20.4	23.6	30.82	1.90	0.60	0.24	5.55
C3	27.0	9.7	25.7	37.6	48.46	3.44	1.03	0.60	7.07
C4	36.0	8.1	20.7	35.2	43.51	3.03	0.98	0.51	4.27

*By difference

Table 3.2: FBG Experimental data

Expt. no.	Coal type	FC/VM (x_1)	Sp. surface area (m^2/g) (x_3)	Activation energy (kJ/mol) (x_4)	Coal feed rate (kg/h) (x_5)	Gasifier temp. (^oC) (x_6)	Ash discharge rate (kg/h) (x_7)	Air/coal ratio (kg/kg coal) (x_8)	$CO+H_2$ (kg/kg coal) (y_1)	Syngas production rate (kg/kg coal) (y_2)	Carbon conversion (%) (y_3)	Syngas heat value ($kcal/Nm^3$) (y_4)
1	C1	1.13	103.60	120.60	11.5	852	6.0	1.13	0.27	1.61	66.34	1104.0
2*	C1	1.13	103.60	120.60	11.0	896	5.0	1.20	0.44	1.78	80.64	1278.0
3*	C1	1.13	103.60	120.60	11.0	905	5.0	1.23	0.43	1.83	81.13	1260.0
4	C1	1.13	103.60	120.60	10.3	912	4.5	1.26	0.46	1.89	84.19	1260.0
5*	C1	1.13	103.60	120.60	11.0	918	4.5	1.29	0.49	1.92	85.42	1275.3
6	C1	1.13	103.60	120.60	10.5	925	4.5	1.34	0.49	1.95	84.82	1245.0
7	C2	1.16	115.15	117.15	16.0	815	10.3	0.87	0.14	1.19	52.78	0962.4
8	C2	1.16	115.15	117.15	16.0	825	9.5	0.87	0.18	1.27	61.76	1023.6
9	C2	1.16	115.15	117.15	15.0	834	9.5	0.91	0.18	1.29	61.27	1025.4
10*	C2	1.16	115.15	117.15	14.0	839	7.0	0.92	0.22	1.33	64.89	1102.2
11*	C2	1.16	115.15	117.15	14.0	839	7.0	0.94	0.24	1.36	68.13	1077.6
12	C2	1.16	115.15	117.15	14.0	845	8.0	0.94	0.23	1.29	62.64	1122.6
13	C2	1.16	115.15	117.15	15.0	855	8.5	0.88	0.20	1.29	63.57	1095.6
14	C2	1.16	115.15	117.15	15.0	859	8.0	0.89	0.24	1.36	71.01	1161.3
15	C2	1.16	115.15	117.15	15.0	859	8.5	0.91	0.22	1.34	67.43	1137.0
16	C2	1.16	115.15	117.15	13.0	872	5.5	0.94	0.32	1.42	77.56	1282.2
17*	C2	1.16	115.15	117.15	14.3	879	8.0	0.96	0.31	1.40	72.86	1233.3
18	C2	1.16	115.15	117.15	13.0	880	6.0	0.95	0.32	1.44	75.79	1113.6

Table 3.2 continued…

Expt. no.	Coal type	FC/VM (x_1)	Sp. surface area (m^2/g) (x_3)	Activation energy (kJ/mol) (x_4)	Coal feed rate (kg/h) (x_5)	Gasifier temp. (^oC) (x_6)	Ash discharge rate (kg/h) (x_7)	Air/coal ratio (kg/kg coal) (x_8)	$CO+H_2$ (kg/kg coal) (y_1)	Syngas production rate (kg/kg coal) (y_2)	Carbon conversion (%) (y_3)	Syngas heat value ($kcal/Nm^3$) (y_4)
19	C2	1.16	115.15	117.15	14.0	880	5.8	0.91	0.32	1.41	77.96	1295.4
20	C2	1.16	115.15	117.15	13.0	887	5.5	0.98	0.33	1.46	78.39	1261.5
21	C2	1.16	115.15	117.15	13.4	890	6.0	0.98	0.33	1.44	76.66	1281.9
22	C2	1.16	115.15	117.15	13.3	892	7.0	1.04	0.33	1.48	76.02	1211.7
23*	C2	1.16	115.15	117.15	13.0	893	5.8	1.00	0.34	1.48	78.73	1254.3
24	C2	1.16	115.15	117.15	13.5	896	6.0	1.02	0.35	1.48	79.43	1291.5
25	C2	1.16	115.15	117.15	12.5	900	6.0	1.10	0.35	1.53	77.94	1238.1
26	C2	1.16	115.15	117.15	13.0	901	5.8	0.97	0.34	1.46	79.88	1298.4
27	C2	1.16	115.15	117.15	13.0	903	6.0	0.95	0.33	1.43	78.04	1266.3
28	C2	1.16	115.15	117.15	13.0	904	5.8	0.98	0.36	1.47	81.19	1320.0
29	C3	1.46	86.25	133.37	11.5	833	6.5	1.27	0.34	1.87	60.51	1110.0
30*	C3	1.46	86.25	133.37	11.0	889	5.5	1.30	0.48	2.00	70.06	1257.0
31	C3	1.46	86.25	133.37	10.5	911	5.0	1.34	0.52	2.09	74.33	1269.0
32	C3	1.46	86.25	133.37	10.0	966	4.0	1.43	0.57	2.34	86.22	1215.6
33	C4	1.70	94.52	125.87	12.0	841	7.0	1.12	0.32	1.64	60.12	1161.0
34*	C4	1.70	94.52	125.87	11.5	875	6.7	1.24	0.39	1.83	67.49	1164.0
35	C4	1.70	94.52	125.87	11.5	890	6.0	1.22	0.44	1.82	68.74	1242.0
36	C4	1.70	94.52	125.87	11.0	900	6.0	1.34	0.47	1.96	73.59	1221.0

During gasification, the major controlled parameters were coal feed rate, bed temperature, air flow rate, steam flow rate, ash withdrawal rate, and bed height. The bed temperature was major feedback for the control loops to control various other operating parameters. It was controlled via manipulating the coal and air feed rates. The bed height was controlled by adjusting the rate of ash extraction from the bottom of the gasifier. Upon following the stated experimental and control procedures, a number of experiments (=36) were conducted by varying FBCG operating conditions in the following ranges: (i) coal feed rate = 10−16 (kg/h), (ii) air/ coal ratio = 0.8−1.5 (kg/kg of coal), (iii) steam feed rate ≈ 0.2 (kg/kg of coal), (iv) gasifier bed temperature = 800−960 (°C), (v) ash withdrawal rate = 4−10 (kg/h), and (vi) bed height ≈ 10 cm. All the experiments were conducted with the minimum fluidization velocity of 0.625 m/s. The proximate and ultimate analyses of the four types (C1−C4) of coals are given in Table 3.1, and the values of the seven inputs (x_1 to x_8) and the corresponding four outputs (y_1−y_4) pertaining to the 36 gasification experiments are listed in Table 3.2; the values of the second input, namely percentage of ash (x_2), are listed in Table 3.1.

3.3 RESULTS AND DISCUSSION

3.3.1 Sensitivity Analysis of Model Inputs

In this study, the sensitivity analysis (also termed “importance” analysis) of the predictor/input/independent variables in the gasifier data has been performed using the IBM-SPSS (2011) package to ascertain the extent of influence exerted by the eight input variables (x_1−x_8) on the four process performance variables (y_1−y_4); the details of the sensitivity analysis are given in Chapter 2 (section 2.6). The importance analysis was conducted using the entire set of experimental data listed in Tables 3.1 and 3.2. The four panels (*a*−*d*) of Figure 3.2 exhibit the importance and normalized importance charts, which indicate the extent of influence exerted individually by the eight input variables (x_1−x_8) on the four performance variables (y_1−y_4). The importance chart is simply a bar chart of the values in the importance table, sorted in the descending values of importance. From Figure 3.2, it is observed that the *air/coal*

ratio (x_8), *gasifier bed temperature* (x_6), *ash discharge rate* (x_7), and *coal feed rate* (x_5) influence the gasifier performance variables most significantly. These sensitivity results are in conformity with those observed in the studies by Pinto et al. (2003), Lee et al. (2002), and Ponzio et al. (2006). It is also noticed that the basic coal properties viz. FC/VM ratio (x_1), ash content (x_2), specific surface area (x_3), and gasification activation energy (x_4) impart relatively lower influence on the gasification performance. Chavan et al. (2012) and Kim et al. (1997) have also made similar observations during their gasification studies.

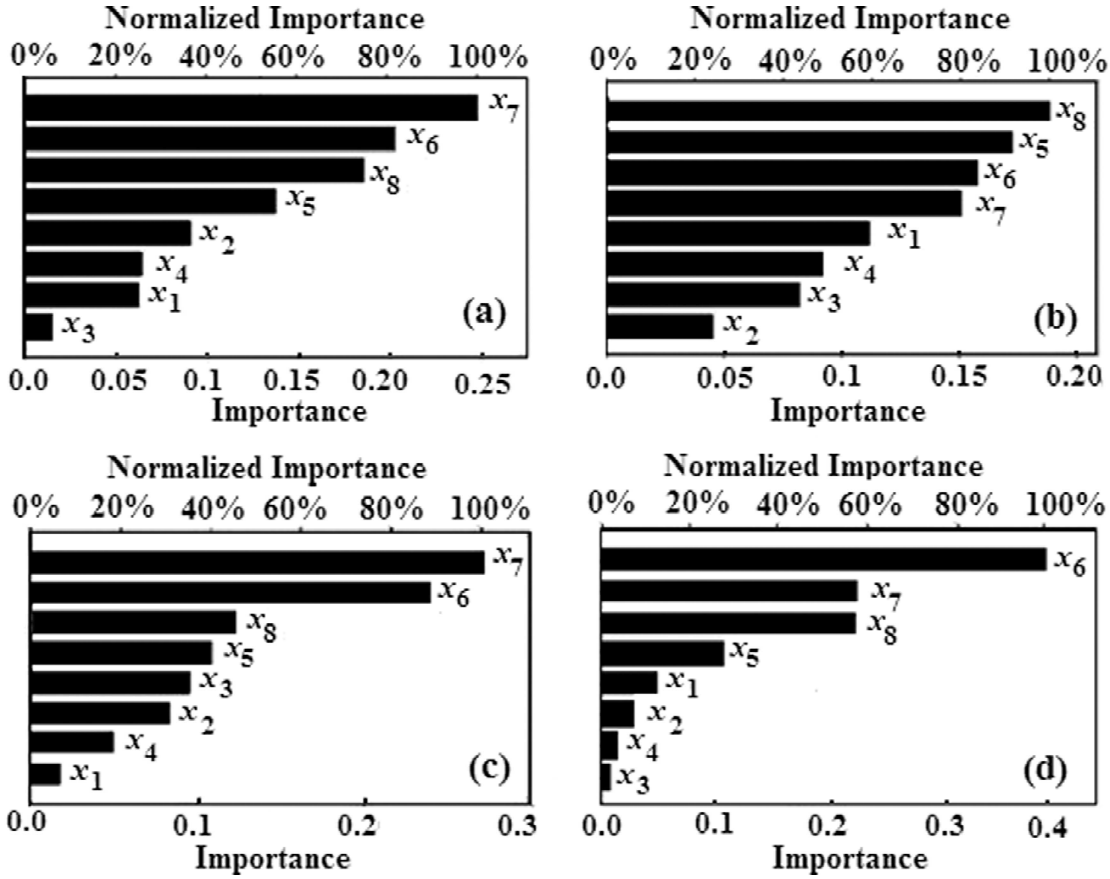


Figure 3.2: Normalized importance of eight model inputs (x_1− x_8) on four model outputs namely $CO+H_2$ generation rate (panel *a*), syngas generation rate (panel *b*), carbon conversion (panel *c*), heating value of syngas (panel *d*).

3.3.2 Artificial Intelligence (AI)-based FBCG Modeling

The principal component analysis (PCA) (Geladi and Kowalski, 1986) described in Chapter 2 (section 2.5) was performed on the eight-dimensional input space of the

GP and MLP-based models with a view to reduce the dimensionality of the input space and thereby the complexity of the models. A low dimensional input space also lowers the computational load during the GP/MLP-based modeling. The results of the PCA provided the following magnitudes of the variance in the experimental data captured by the eight principal components (PCs): PC_1, 70.2%; PC_2, 21.5%; PC_3, 4%; PC_4, 2.4%; PC_5, 1.6%; PC_6, 0.2%; PC_7, 0.1%; and PC_8, 0.0%. It is thus seen that the first three PCs have captured a large percentage (≈ 95.7%) of the data variance. Thus, it was possible to reduce the dimensionality of the input space of the GP and MLP-based models from eight to three by considering the elements of the first three PCs in place of the original eight inputs. The three PCA-transformed inputs (v_1, v_2, v_3) are defined as

$$v_1 = 0.288\,\hat{x}_1 - 0.396\,\hat{x}_2 - 0.399\,\hat{x}_3 + 0.382\,\hat{x}_4 - 0.387\,\hat{x}_5 + 0.242\,\hat{x}_6 - 0.289\,\hat{x}_7 + 0.405\,\hat{x}_8 \quad (3.3)$$

$$v_2 = 0.424\,\hat{x}_1 - 0.228\,\hat{x}_2 - 0.233\hat{x}_3 + 0.273\,\hat{x}_4 + 0.243\,\hat{x}_5 - 0.556\,\hat{x}_6 + 0.514\,\hat{x}_7 - 0.08\,\hat{x}_8 \quad (3.4)$$

$$v_3 = -0.811\hat{x}_1 - 0.25\,\hat{x}_2 - 0.162\,\hat{x}_3 + 0.174\,\hat{x}_4 - 0.115\hat{x}_5 - 0.431\,\hat{x}_6 + 0.005\,\hat{x}_7 + 0.159\,\hat{x}_8 \quad (3.5)$$

where $\hat{x}_i$; i = 1, 2, ..., 8, denote the *normal scores* (standardized variables) of the eight input variable values (x_i) listed in Tables 3.1 and 3.2. For developing the models possessing good prediction accuracy and generalization ability, the experimental data were split randomly wherein 75% data (27 patterns) were used as the training set for developing the models while 25% data (9 patterns) were used as the test set for assessing the generalization ability of the models. In Supporting Information Table 3.2, the test set data are marked using the asterisk (“ ***** ”) symbol.

Table 3.3: Details of GP-based FBCG Models

Model No	GP-based models*	CC_{trn}	MSE_{trn}	CC_{tst}	MSE_{tst}
I	$y_1 = 0.1058\,[0.8374\,z_1 + 0.1965\,z_1 z_3 - 0.3645\,z_2 - 0.09381\,z_3^3] + 0.3414$	0.993	1.41×10^{-4}	0.981	3.5×10^{-4}
II	$y_2 = 0.2806\,[1.118\,z_1 - 0.103\,z_2 - 0.2117\,z_1 z_2] + 0.3414$	0.993	1.28×10^{-3}	0.997	5.37×10^{-4}
III	$y_3 = 8.4187\,[0.3003\,z_1^2 + 0.1524\,z_1 z_2^2 - 0.1977\,z_3 - 1.089\,z_2 - 0.276] + 72.9869$	0.980	3.905	0.980	2.288
IV	$y_4 = 91.441\,[0.8156\,z_1 z_3 + 0.2545\,z_1 z_2 + 0.07754\,z_2^3 - 0.1946\,z_3^3 - 0.2459\,z_1^2 z_3 - 0.657\,z_2 - 0.02941] + 1197.72$	0.925	1296.95	0.969	621.62
*$z_1 = (v_1)/(2.3708)$, $z_2 = (v_2)/(1.3099)$ and $z_3 = (v_3)/(0.5676)$; v_i denotes ith PCA-transformed variable					

Table 3.4: Details of MLP-based FBCG Models

Model No	Output variable	Input nodes	No. of hidden layers	Hidden nodes in each hidden layer	Transfer function for hidden layer	Transfer function for output layer	μ_{ebp}	η	CC_{trn}	MSE_{trn}	CC_{tst}	MSE_{tst}
I	y_1	3	1	2	tanh	Identity	0.05	0.1	0.993	1.4×10^{-4}	0.978	5.3×10^{-4}
II	y_2	3	1	2	tanh	Identity	0.005	0.1	0.994	9.0×10^{-4}	0.996	9.8×10^{-4}
III	y_3	3	1	2	tanh	Identity	0.05	0.1	0.977	3.370	0.982	1.822
IV	y_4	3	1	2	tanh	Identity	0.05	0.2	0.920	1371.81	0.960	617.99

GP-based modeling

The four GP-based models predicting as many (y_1− y_4) gasifier performance variables were developed using the *Eureqa Formulize* software package (Schmidt and Lipson, 2009). The detailed procedure for GP (Koza, 1992; Kinnear, 1994) implementation has been explained in Chapter 2 (section 2.2.2). This package has been optimized to construct parsimonious models possessing good generalization ability. In the GP-based modeling, the *mean squared error* (*MSE*) dependent fitness function was used. The effects of the GP procedural parameters such as the size of the training and test sets as also the various input normalization schemes were studied rigorously. The prediction accuracy and the generalization performance of each model were evaluated by computing the *coefficient of correlation* (*CC*) and the *MSE* between the experimental (target) and the corresponding model-predicted values of the four process performance variables. These quantities were evaluated separately for the training and test data sets. The overall best models were selected on the basis of their high CC and low *MSE* magnitudes in respect of both training and test set data. The four GP-based models, respectively predicting CO + H_2 *generation rate* (y_1), *syngas production rate* (y_2), *carbon conversion* (y_3), and *heating value of the syngas* (y_4), are listed in Table 3.3 along with the corresponding magnitudes of the training and test set coefficients of correlation and mean squared errors.

The four panels (*a*−*d*) of Figure 3.3 respectively show the parity plots of the experimental versus GP model-predicted values of the four process performance variables (y_1− y_4) in respect of both training and test set data. As can be noticed from panels (*a*−*c*), the model predicted values of the performance variables y_1, y_2 and y_3 exhibit a close match with their experimental counterparts. The prediction accuracy of the GP-based model for the performance variable, y_4 (heating value of the syngas), though high is marginally inferior to that possessed by the GP-models for y_1, y_2, and y_3.

MLP-based modeling

The MLP-based analogs of the four GP-based models were developed using the same training and test sets as used in the development of the GP-based models. To construct an optimal MLP-based (Zurada, 1992; Bishop,1994) model the detailed

procedure explained in Chapter 2 (section 2.2.1) has been followed. The effects of network's structural parameters (i.e., the number of hidden layers, number of nodes in each hidden layer and type of transfer function and the two EBP algorithm parameters, namely *learning rate* (η) and *momentum coefficient* (μ_{ebp}), on the model's prediction accuracy and generalization capability were systematically examined. Also, the effect of random weight initialization was studied to obtain an MLP model that corresponds to the global or the deepest local minimum on the model's nonlinear error surface. The details of the model architecture along with the values of the training and test set *CC* and *MSE* for the four MLP models are listed in Table 3.4. The four panels (*a*−*d*) of Figure 3.4 respectively show the parity plots of the experimental versus MLP model-predicted values of the performance variables y_1 to y_4. Similar to Figure 3.3, it can be observed in Figure 3.4 that the MLP predicted values of the performance variables y_1, y_2, and y_3 exhibit a close match with their experimental counterparts. From the *CC* and *MSE* values listed in Tables 3.3 and 3.4 following observations can be made:

- All the four GP-based models are nonlinear.
- The high (≥0.925) and comparable *CC* magnitudes are observed in respect of the training and test set outputs for all the GP and MLP-based models.
- Among the four models constructed separately using the GP and MLP methods, the prediction accuracy and generalization performance of the first three models, respectively predicting the magnitudes of CO+H_2 *generation rate* (y_1), *syngas generation rate* (y_2), and *carbon conversion* (y_3) are excellent (CC_{trn} and $CC_{\mathrm{tst}} \geq 0.98$); the fourth one predicting the *heating value of the syngas* (y_4), however, possesses relatively lower prediction accuracy and generalization performance ($\mathrm{CC_{trn}} \sim 0.92$ and $\mathrm{CC_{tst}} \sim 0.96$).
- The *CC* and *MSE* magnitudes for the training and test set data obtained using the GP and MLP models indicate that both types of models possess comparable prediction and generalization performance.

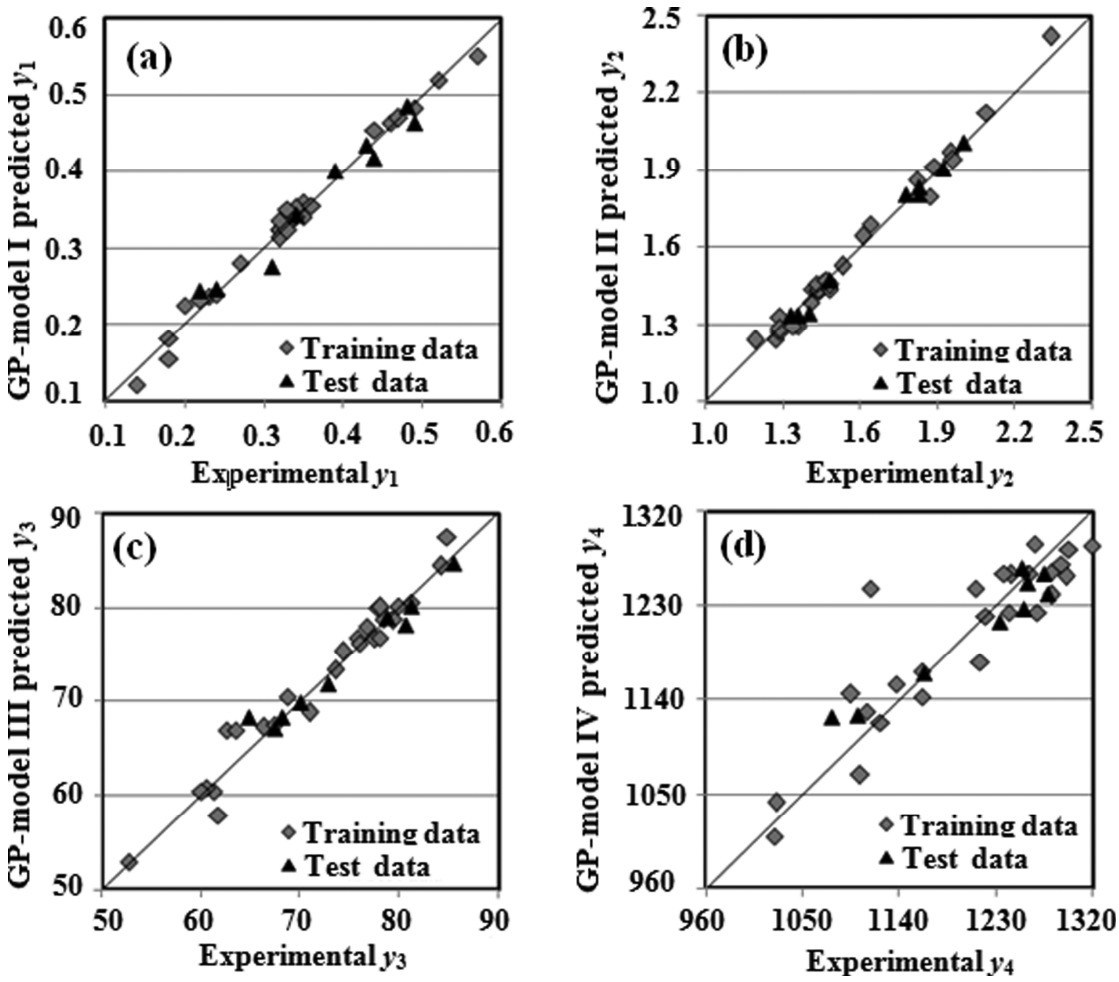


Figure 3.3: Plots of experimental versus GP model-predicted values of performance variables, namely CO+H_2 generation rate (y_1, kg/kg coal) (panel *a*), syngas production rate (y_2, kg/kg coal) (panel *b*), carbon conversion (y_3, %) (panel *c*), and heating value of syngas (y_4, kcal/Nm^3) (panel *d*).

An explanation is in order for the relatively lower prediction accuracy and generalization performance in predicting the *heating value of the syngas* (y_4) by both GP and MLP-based models. In this study, the overall heating value of the generated syngas is computed by adding the heating values of CO (3014 kcal/nm^3), H_2 (3050 kcal/nm^3), and CH_4 (9530 kcal/nm^3) in their respective proportions in the syngas. In the experiments conducted in the FBCG, the individual percentages of the generated CO and H_2 varied between 10 and 22, while the percentage of the generated CH_4 varied between 0.5 and 2.0. The quantitative analysis of the composition of the syngas was made using the gas chromatography, and it is quite plausible that the accuracy of the measurement of concentration of CH_4 whose proportion in the syngas is much lower than that of the CO or H_2 was not as good as that of the stated major

components. The effect of even a marginal inaccuracy in the measurement of the methane gets amplified in the computation of the heating value of syngas, due to methane's nearly 3.15 times higher heating value when compared with that of the CO or H_2. In essence, slight inaccuracies in the measurements of the minor syngas component (i.e., CH_4) together with its much higher heating value could have led to the small deviations in the actual magnitudes of the overall heating values of the syngas. These deviations are possibly responsible for the lower (albeit marginally) prediction accuracies of both the GP and MLP-based models predicting the overall heating value of the generated syngas. It is thus clear that accurate measurements of the low concentrations of methane in the syngas product should assist in improving the y_4 prediction accuracy of the GP- and MLP-based models.

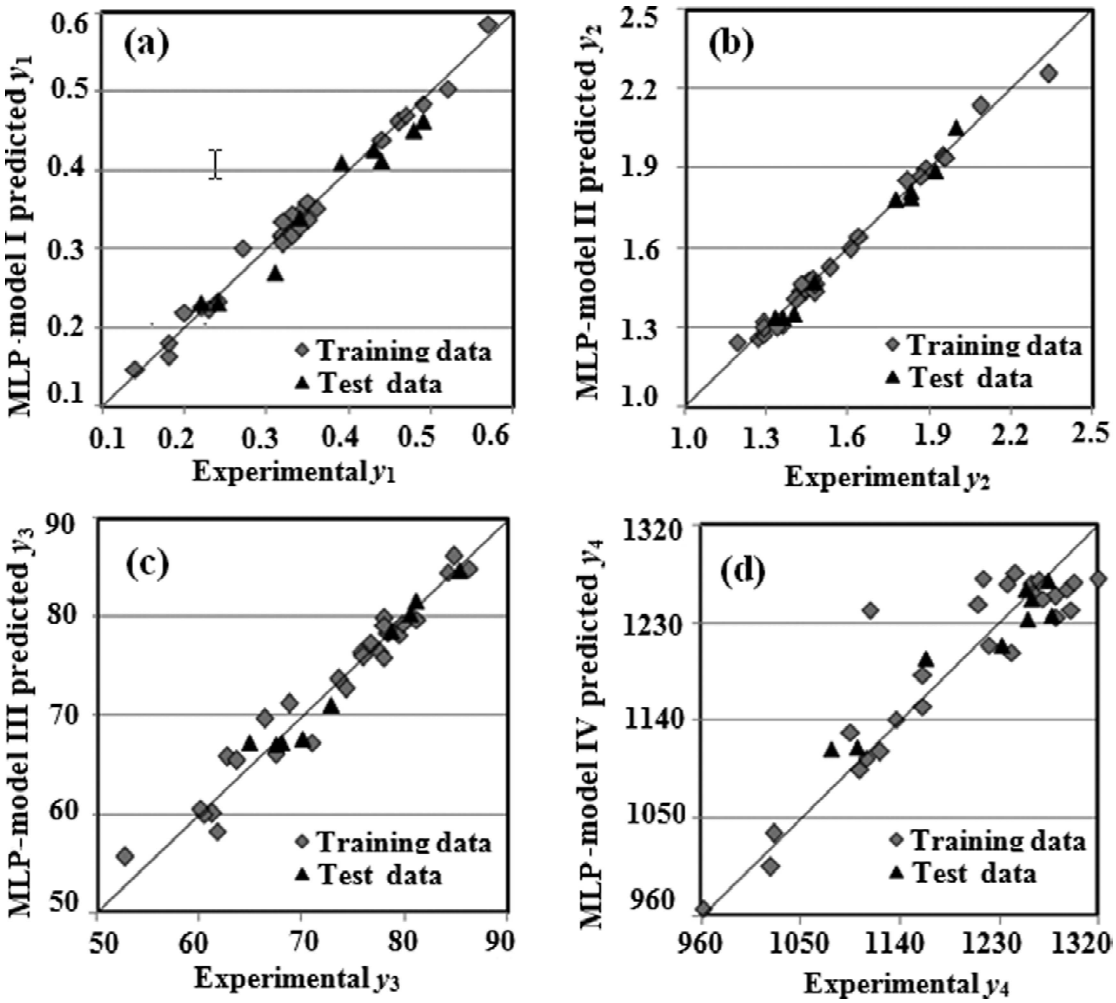


Figure 3.4: Plots of experimental versus MLP model-predicted values of performance variables, namely CO+H_2 generation rate (y_1, kg/kg coal) (panel *a*), syngas production rate (y_2, kg/kg coal) (panel *b*), carbon conversion (y_3, %) (panel *c*), and heating value of syngas (y_4, kcal/Nm^3) (panel *d*).

3.4 CONCLUSION

Coal gasification is a cleaner and an efficient alternative to the coal combustion for producing the syngas. The high-ash coals are found in a number of countries, and they form an important source for the gasification. Accordingly, in this study, extensive gasification experiments were conducted in a pilot-plant scale fluidized-bed coal gasifier (FBCG) using high-ash coals from India. The FBCG is a complex nonlinear process and the complete details of the underlying physicochemical phenomena are not available; thus, development of the phenomenological model for an FBCG process is a cumbersome, time-consuming, and costly task. To overcome the difficulties associated with the phenomenological models, in this study the knowledge of the proximate analysis, char−CO_2 gasification activation energy, surface area of the coal, and influential process parameters has been utilized for developing exclusively data driven FBCG models predicting four important gasification performance variables. For modeling, novel *artificial intelligence* (AI) formalism, namely *genetic programming* (GP) has been used and the performance of the GP-based models was compared with the corresponding MLP neural network-based ones. Both types of models have been found to possess output prediction accuracies and the generalization performance that vary from good to excellent as indicated by the high training and test set correlation coefficient magnitudes lying between 0.920 to 0.996. A rigorous literature search shows that this is the first study wherein the GP strategy has been employed for the data-driven modeling in the coal sciences and engineering. The models developed in this study can be gainfully used in designing and control of the FBCG, and in selecting process operating conditions leading to an optimal gasifier operation. These models can also be used in predicting the gasification performance of similar types of coals in the bubbling FBCG of pilot scale capacity.

NOMENCLATURE

v_i ith PCA-transformed variable

$\hat{x}_i$ the *normal scores* (standardized variables) of the eight input variable values (x_i)

y_i ith output (dependent) variable

REFERENCES

Armstrong, L. M., Gu, S., and Luo, K. H. (2011). Parametric study of gasification processes in a BFB coal gasifier. *Industrial & Engineering Chemistry Research*, *50*(10), 5959-5974.

Beamish, B. B., Shaw, K. J., Rodgers, K. A., and Newman, J. (1998). Thermogravimetric determination of the carbon dioxide reactivity of char from some New Zealand coals and its association with the inorganic geochemistry of the parent coal. *Fuel processing technology*, *53*(3), 243-253.

Behera, S. K., Rene, E. R., Kim, M. C., and Park, H. S. (2014). Performance prediction of a RPF-fired boiler using artificial neural networks. *International Journal of Energy Research*, *38*(8), 995-1007.

Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments*, *65*(6), 1803-1832.

Çakal, G. Ö., Yücel, H., and Gürüz, A. G. (2007). Physical and chemical properties of selected Turkish lignites and their pyrolysis and gasification rates determined by thermogravimetric analysis. *Journal of analytical and applied pyrolysis*, *80*(1), 262-268.

Chavan, P. D. (2012). Studies on effect of coal properties and process parameters on gasification kinetics. Ph.D. Thesis, Indian School of Mines: Dhanbad, India.

Chavan, P. D., Sharma, T., Mall, B. K., Rajurkar, B. D., Tambe, S. S., Sharma, B. K., and Kulkarni, B. D. (2012). Development of data-driven models for fluidized-bed coal gasification process. *Fuel*, *93*, 44-51.

Chavan, P., Datta, S., Saha, S., Sahu, G., and Sharma, T. (2012). Influence of high ash Indian coals in fluidized bed gasification under different operating conditions. *Solid Fuel Chemistry*, *46*(2), 108-113.

Choudhury, S. (2013). Studies on demineralization of coal: Fractional factorial design. *Int. J. Innovative Technol. Res*, *1*(1), 2320-5547.

Clean Coal Technology Demonstration Program: Program Update 2000; U.S. Department of Energy (DOE): Washington, DC, 2001; http:// www.netl.doe.gov/

technologies/coalpower/cctc/resources/pdfsprog/cm3tupdat/ ct_pgm_2000_all.pdf (accessed Jan. 19, 2014).

Davidson, R. M. (1983). *Mineral effects in coal conversion* , Report ICTIS/TR22 , (Vol. 22). IEA Coal Research, London.

de Souza-Santos, M. L. (1989). Comprehensive modelling and simulation of fluidized bed boilers and gasifiers. *Fuel*, *68*(12), 1507-1521.

Donne, M. S., Dixon, R., Pike, A. W., Odeku, A. J. L., and Ricketts, B. E. (1998). *Dynamic modelling of the ABGC prototype integrated plant*. COAL R143; Coal R & D Programme, Energy Technology Support Unit: Harwell Laboratory, U.K.

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, *185*, 1-17.

Gómez-Barea, A., and Leckner, B. (2010). Modeling of biomass gasification in fluidized bed. *Progress in Energy and Combustion Science*, *36*(4), 444-509.

Goyal, A., Zabransky, R. F., and Rehmat, A. (1989). Gasification kinetics of Western Kentucky bituminous coal char. *Industrial & engineering chemistry research*, *28*(12), 1767-1778.

Gururajan, V. S., Agarwal, P. K., and Agnew, J. B. (1992). Mathematical modelling of fluidized bed coal gasifiers: Chemical reaction engineering. *Chemical engineering research & design*, *70*(A3), 211-238.

Gutierrez, L. A., and Watkinson, A. P. (1982). Fluidized-bed gasification of some Western Canadian coals. *Fuel*, *61*(2), 133-138.

Heaven, D. L., Daniel, F., and Calif, I. (1996). Gasification converts a variety of problem feed-stocks and wastes. *Oil and Gas Journal*, *94* (22), 49−54.

IBM SPSS Neural Networks 20 manual, IBM: Chicago, 2011.

Irfan, M. F., Usman, M. R., and Kusakabe, K. (2011). Coal gasification in CO_2 atmosphere and its kinetics since 1948: a brief review. *Energy*, *36*(1), 12-40.

Ju, F., Chen, H., Yang, H., Wang, X., Zhang, S., and Liu, D. (2010). Experimental study of a commercial circulated fluidized bed coal gasifier. *Fuel Processing Technology*, *91*(8), 818-822.

Kim, Y. J., Lee, J. M., and Kim, S. D. (1997). Coal gasification characteristics in an internally circulating fluidized bed with draught tube. *Fuel*, *76*(11), 1067-1073.

Kinnear, K. E. (1994). *Advances in Genetic Programming* (Vol. 1). MIT press, Cambridge, MA.

Koza, J. (1992). *Genetic Programming: On the Programming of Computers by* Means *of Natural Selection*. MIT Press, Cambridge, MA.

Lee, S. (2007). Gasification of coal. In *Handbook of Alternative Fuel Technologies*; Lee, S., Speight, J. G., Loyalka, S. K., Eds.; CRC Press: Boca Raton, pp 25−80.

Lee, W. J., Kim, S. D., and Song, B. H. (2002). Steam gasification of an Australian bituminous coal in a fluidized bed. *Korean Journal of Chemical Engineering*, *19*(6), 1091-1096.

Lim, K. S., Zhu, J. X., and Grace, J. R. (1995). Hydrodynamics of gas-solid fluidization. *International journal of multiphase flow*, *21*, 141-193.

Liukkonen, M., Hälikkä, E., Hiltunen, T., and Hiltunen, Y. (2012). Dynamic soft sensors for NOx emissions in a circulating fluidized bed boiler. *Applied energy*, *97*, 483-490.

Mazumder, A. (2010). Development of a simulation model for fluidized bed mild gasifier. Thesis, Paper 101, University of New Orleans, New Orleans; http://scholarworks.uno.edu/td/101/ (accessed Jan. 20, 2014).

Miller, B. G. (2011). Clean coal technology for advanced power generation; in: *Clean Coal Engineering Technology*; Chapter 7; Elsevier: Burlington, MA 01803, USA; pp 251−296.

Mjalli, F. S., and Al-Mfargi, A. (2008). Artificial neural approach for modeling the heat and mass transfer characteristics in three-phase fluidized beds. *Industrial & Engineering Chemistry Research*, *47*(13), 4542-4552.

Moorea-Taha, R. (2000). Modeling and Simulation for Coal Gasification; IEA Coal Research 2000; IEA Clean Coal: London; ISBN 92- 9029-354-3, pp 1−50.

Nougues, J. M., Pan, Y. G., Velo, E., and Puigjaner, L. (2000). Identification of a pilot scale fluidised-bed coal gasification unit by using neural networks. *Applied thermal engineering*, *20*(15), 1561-1575.

Ocampo, A., Arenas, E., Chejne, F., Espinel, J., Londono, C., Aguirre, J., and Perez, J. D. (2003). An experimental study on gasification of Colombian coal in fluidised bed. *Fuel*, *82*(2), 161-164.

Ollero, P., Serrera, A., Arjona, R., and Alcantarilla, S. (2003). The CO_2 gasification kinetics of olive residue. *Biomass and Bioenergy*, *24*(2), 151-161.

Pinto, F., Franco, C., Andre, R. N., Tavares, C., Dias, M., Gulyurtlu, I., and Cabrita, I. (2003). Effect of experimental conditions on co-gasification of coal, biomass and plastics wastes with air/steam mixtures in a fluidized bed system. *Fuel*, *82*(15), 1967-1976.

Ponzio, A., Kalisz, S., and Blasiak, W. (2006). Effect of operating conditions on tar and gas composition in high temperature air/steam gasification (HTAG) of plastic containing waste. *Fuel Processing Technology*, *87*(3), 223-233.

Puig-Arnavat, M., Hernández, J. A., Bruno, J. C., and Coronas, A. (2013). Artificial neural network models for biomass gasification in fluidized bed gasifiers. *biomass and bioenergy*, *49*, 279-289.

Rhinehart, R. R., Felder, R. M., and Ferrell, J. K. (1987). Coal gasification in a pilot-scale fluidized bed reactor. 3. Gasification of a Texas lignite. *Industrial & engineering chemistry research*, *26*(10), 2048-2057.

Satonsaowapak, J., Kulworawanichpong, T., Oonsivilai, R., and Oonsivilai, A. (2011). Gasifier system identification for biomass power plants using neural network. *International Journal of Chemical, Molecular, Nuclear, Materials and Metallurgical Engineering*, *5*(12), 1079-1084.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, *324*(5923), 81-85.

Sett, A., and Bhattacharya, S. C. (1988). Mathematical modelling of a fluidised-bed charcoal gasifier. *Applied energy*, *30*(3), 161-186.

Shaw, K. J., Beamish, B. B., and Rodgers, K. A. (1997). Thermogravimetric analytical procedures for determining reactivities of chars from New Zealand coals. *Thermochimica Acta*, *302*(1), 181-187.

Singh, N., Raghavan, V., and Sundararajan, T. (2014). Mathematical modeling of gasification of high-ash Indian coals in moving bed gasification system. *International Journal of Energy Research*, *38*(6), 737-754.

Takematsu, T., and Maude, C. (1991). *Coal gasification for IGCC power generation*. International Energy Agency, Coal Research: London.

Villanueva, A., Gómez-Barea, A., Revuelta, E., Campoy, M., Ollero, P. (2008). Guidelines for selection of gasifiers modeling strategies. In *Proceedings of 16th European Biomass Conference and exhibition*; Valencia, Spain, pp 980−986.

Witt, P. J., Perry, H., and Schwartz, M. P. (1997). Application of CFD to fluidised bed systems. In *Proceedings of International conference on CFD in Minerals and Metals Processing and Power Generation*; CSIRO Minerals: Melbourne, Australia, pp 353−360.

Witt, P. J.; Perry, J. H. (1996). A study in multiphase modeling of fluidized bed. In *Proceedings of the 7th Biennial Conference on Computational Techniques and Applications*: CTAC95, Melbourne, Australia, 3−5 Jul. 1995; World Scientific: Singapore, 1996; pp 787− 794.

Xiangdong, K., Zhong, W., Wenli, D. U., and Feng, Q. I. A. N. (2013). Three stage equilibrium model for coal gasification in entrained flow gasifiers based on Aspen Plus. *Chinese Journal of Chemical Engineering*, *21*(1), 79-84.

Yang, S., Yang, Q., Li, H., Jin, X., Li, X., and Qian, Y. (2012). An integrated framework for modeling, synthesis, analysis, and optimization of coal gasification-based energy and chemical processes. *Industrial & Engineering Chemistry Research*, *51*(48), 15763-15777.

Zurada, J.M. (1992). *Introduction to Artificial Neural Network*. West Publ Co., St. Paul.

Chapter-4

High Ash Char Gasification in Thermo-gravimetric Analyzer and Prediction of Gasification Performance Parameters Using Computational Intelligence formalisms

ABSTRACT

Coal gasification is a cleaner and more efficient process than coal combustion. Although high ash coals are routinely used in the energy generation, systematic gasification kinetic studies using chars derived from these coals are scarce. Accordingly, this chapter reports the development of the data-driven models for the gasification of chars derived from the high ash coals. Specifically, the models predict two significant gasification performance parameters, viz. gasification rate constant, and reactivity index. These models have been constructed using three computational intelligence (*CI*) *methods, namely genetic programming* (*GP*), *multilayer perceptron* (*MLP*) *neural network* (*NN*), *and support vector regression* (*SVR*). *The inputs to the CI-based models consist of seven parameters representing the gasification reaction conditions and properties of high ash coals and chars. The data used in the modeling were collected by performing extensive gasification experiments in the* CO_2 *atmosphere in a thermo-gravimetric analyzer* (*TGA*), *using char samples derived from Indian coals with high ash content. Values of the above-state two gasification performance parameters were obtained by fitting the experimental data to the shrinking un-reacted core* (*SUC*) *model. It has been observed that all the CI-based models developed in this study possess an excellent prediction accuracy and generalization capability. Accordingly, these models can be gainfully employed in the design and operation of the fixed and fluidized bed gasifiers using high ash coals.*

4.1 INTRODUCTION

The commonly used coal combustion technologies in the power generation industry produce significant amounts of emissions of greenhouse (CO_2) and polluting gases, such as SO_x and NO_x. Thus, the development of clean coal technologies has received a global attention for overcoming the adverse effects of coal combustion. For mitigating the undesirable impact of coal combustion on the environment, stringent pollution control norms have been prescribed by the regulatory agencies of countries generating coal-based power. The suggested measures are expected to result in higher coal conversion efficiencies and a lower environmental impact (Takematsu and Maude, 1991). One of the important norms that have been prescribed includes changing coal utilization practices.

Gasification is a cleaner and more efficient process than the combustion for converting carbonaceous materials into energy (Miller, 2011). In a gasification reaction, solid fuel is converted at high temperatures into a gaseous fuel (syngas) that burns relatively cleanly. There exist three major coal gasification technologies, namely *moving* (fixed), *fluidized* and *entrained bed* gasifiers. The gasification of coals and chars has been studied extensively for understanding the specific underlying reactions and developing the corresponding kinetic models (see, for example, Ballal and Zygourakis, 1986; Ye et al., 1998; Ochoa et al., 2001; Zhang et al., 2006; Irfan et al., 2011).

The gasification of coal occurs in two steps. The first step is pyrolysis, which produces volatiles and char. Normally, char (pyrolysis residue) represents 55–70% of the original coal. In the second step, solid char is converted to gaseous products (char gasification). The principal reactions occurring during the gasification of char are as follows:

$$C + CO_2 \rightarrow 2\,CO; \qquad \Delta H = +\,159.7\ \text{kJ mol}^{-1} \tag{4.1}$$

$$C + H_2O \rightarrow CO + H_2; \qquad \Delta H = +\,118.9\ \text{kJ mol}^{-1} \tag{4.2}$$

$$C + O_2 \rightarrow CO_2; \qquad \Delta H = -\,405.9\ \text{kJ mol}^{-1} \tag{4.3}$$

$$CO + H_2O \rightarrow CO_2 + H_2; \qquad \Delta H = -\,40.9\ \text{kJ mol}^{-1} \tag{4.4}$$

$$C + 2\,H_2 \rightarrow CH_4; \qquad \Delta H = -\,87.4\ \text{kJ mol}^{-1} \tag{4.5}$$

The reaction enthalpies of the above reactions are given at standard conditions, i.e., at 25^{o}C and 0.1013 MPa pressure (Kristiansen, 1996).

Due to its several attractive characteristics and higher efficiency, coal gasification is gaining importance for producing electrical energy. It is, however, a complex nonlinear process and, therefore, several issues need to be addressed while designing and operating a coal-based gasifier.

- Commonly, in industry an air-steam or oxygen-steam mixture is used as a gasifying agent. The endothermic gasification reactions, such as the Boudouard (reaction 1) and water-gas (reaction 2), are driven by the heat generated during the air- or oxygen-assisted partial combustion of a fraction of the coal (reaction 3). Ideally, gasification needs to be carried out with a minimum amount of air or oxygen to avoid generation of undesired products, such as CO_2 and H_2O, in high quantities. However, an inadequate quantity of air or oxygen results in an incomplete coal conversion and, consequently, insufficient amount of heat generation to drive the endothermic gasification reactions.
- The char–CO_2 reaction is the slowest compared to the other heterogeneous gasification reactions with oxygen and steam. It is, therefore, the rate-determining reaction. The mechanisms of the char–CO_2 and char–steam reactions are considered to be identical (Kristiansen, 1996; Jayaraman et al., 2015).
- Being a well-known gasifying agent, CO_2 in the flue gas can be utilized in the fuel system of the gasifier, which helps in reducing the CO_2 emissions to the atmosphere as also increasing the gasifier efficiency.

It is thus clear that a thorough study of the reactivity and kinetics of the char-gasification (Boudouard) reaction is necessary for: (i) determining the quantity of heat required to drive the reaction, (ii) fixing the amount of air or oxygen required in the exothermic oxidation reaction 3, so that just enough amount of heat is generated for driving the char-gasification reaction, and (iii) using CO_2 in the flue gas as a gasifying agent.

Owing to their importance, the reactivity and kinetics of coal-char gasification have been studied widely in the CO_2 atmosphere (Adschiri et al., 1986; Ahn et al., 2001; Ochoa et al., 2001; Kim et al., 2011; Saha et al., 2011, 2013; Silbermann et al.,

2013; Jayaraman et al., 2015). Presently, coals mined in countries such as India, Australia, China, and Turkey contain ash in high percentages and these constitute a major raw material for thermal power stations. However, in the literature, systematic experimental and modeling studies addressing the reactivity and kinetics of gasification of high ash coals are limited (see, for example, Saha et al., 2011, 2013). In the past, an attempt was made by Adschiri et al. (1986) in which a first principles model was proposed to predict the change in the rate during char gasification. This model utilized the gasification temperature, CO_2 partial pressure, and characteristics of only the parent coal as inputs—i.e., the properties of char produced were not considered. For deriving the aforementioned model, Adschiri et al. (1986) utilized char gasification data collected using a thermo-gravimetric analyzer (TGA). Chars produced from 14 different parent coals in a fluidized bed were employed in the gasification experiments. A significant limitation of this model is that its gasification rate prediction accuracy is suboptimal and a majority of the coals used in the TGA-based experiments contained low amounts of ash. From a rigorous literature survey, it is noticed that although necessary, a model based on the coal and char properties, and gasification conditions, is not available for predicting the char gasification rate constants and char reactivity.

The phenomenological (first principles) modeling of a coal-char gasification process is a difficult task. The specific difficulties encountered in this modeling are (Patil-Shinde et al., 2014): (i) a widely differing gasification behavior due to the variation in the coal-char characteristics, (ii) nonlinear interplay of multiple process variables, (iii) cost-intensive, tedious, and exhaustive experimentation required for studying the effects of influential process operating variables and parameters, and (iv) unavailability of the detailed knowledge regarding physicochemical phenomena (e.g., kinetics and heat and mass transport mechanisms) underlying the gasification process. Some notable representative studies and reviews on the modeling of coal gasification are by Gururajan et al. (1992), Moreea-Taha (2000), Chejne et al. (2011), and Zhao et al. (2012).

In view of the difficulties encountered in the phenomenological modeling of coal gasification process, it becomes necessary to explore alternative modeling approaches. One such practical option is development of exclusively data-driven models. The advantage of these models is that they can be utilized in predicting the gasification behavior under a variety of process operating conditions for a number of

coals and chars produced from them. Consequently, the efforts involved in conducting the time-consuming, costly, and tedious experiments, are reduced drastically. The said data-driven models can also be useful in selecting a suitable coal for an efficient and optimal gasifier operation.

Commonly, data-driven gasification/gasifier models are developed using regression methods. In this approach, the exact structure of the data-fitting function needs to be specified before the parameters associated with it can be estimated. This is a difficult task since, in coal/char gasification, a number of variables nonlinearly influence the process behavior, and the precise interactions between them are not known. These difficulties associated with the development of the standard regression-based modeling however can be overcome by constructing computational intelligence (CI) based models. Accordingly, in the present chapter, data-driven CI-based generalized models have been developed for the prediction of the char gasification rate constant and reactivity index from the knowledge of the properties of coals containing high ash content, and the corresponding chars, as also the gasification conditions. The CI-based modeling formalisms used are *genetic programming* (GP), *multilayer perceptron* (MLP) *neural network* (NN), and *support vector regression* (SVR). The details of all these three data-driven modeling formalisms are provided in Chapter 2 (sections 2.2.2, 2.2.1 and 2.3), respectively.

Due to its simplicity of operation, and high accuracy of measurement, TGA has been widely used in the determination of gasification reactivity and related kinetic studies (Irfan et al., 2011). The experimental data for the present investigation were collected by conducting gasification in the CO_2 atmosphere in a TGA. A total of 108 gasification experiments were conducted using the sub-bituminous high ash Indian coals. The performance of the gasification reaction was monitored in terms of *char gasification rate constant* (k_s) (min^{-1}) and *reactivity index* (r_I) (min^{-1}). The general forms of the CI-based models developed in this study are given as

$$k_s = f_1\,(T_G, C_{CO_2}, T_P\,, S_{N_2}, C_A\,, S_{CO_2}, \phi_P, \alpha\,) \tag{4.6}$$

$$r_I = f_1\,(T_G, C_{CO_2}, T_P\,, S_{N_2}, C_A\,, S_{CO_2}, \phi_P, \beta\,) \tag{4.7}$$

where $\alpha = [\,\alpha_1, \alpha_2, \ldots, \alpha_M]^T$ and $\beta = [\,\beta_1, \beta_2, \ldots, \beta_N]^T$, respectively, represent the parameter vectors; the seven process variables and parameters that form the input space of the models are defined as: (i) T_G : *char gasification temperature* (^oC), (ii)

C_{CO_2}: CO_2 *fraction*, (iii) T_P : *char preparation temperature* (^oC), (iv) S_{N_2} : *surface area-BET method* (m^2/g), (v) C_A: *coal-ash* (%), (vi) S_{CO_2} : *surface area*-CO_2 *method* (m^2/g), and (vii) ϕ_P: *porosity* (%).

The data-driven models presented in this study are expected to be significantly useful in predicting the values of the rate constant and reactivity index for chemically controlled gasification reactions utilizing high ash coals, and chars derived from them. Applicability of the models, however, may be limited for the gasification of the low ash coals since the reaction behavior exhibited by these coals differs from that displayed by their counterparts containing high ash content.

The remainder of this chapter is structured as follows. The details of the char preparation and characterization, as also TGA-based gasification experiments, are provided in the ‘‘Experimental’’ (section 4.2). Section 4.3 titled ‘‘Results and Discussion’’ first presents the results of the phenomenological modeling of the gasification reaction using the shrinking un-reacted core (SUC) approach, followed by the development of the CI-based generalized models for the prediction of the char gasification rate constant and reactivity index. This section also presents results of (a) the principal component analysis (PCA) conducted to perform dimensionality reduction of the input space of the models, and (b) a comparison of the prediction and generalization performance of the three types of CI-based models. Finally, in ‘‘Concluding Remarks’’ (section 4.4), the principal findings of this study are summarized.

4.2 EXPERIMENTAL

4.2.1 Selection of Coal Samples

Three sub-bituminous high ash coals with varying ash content (27 – 48.9% ash on air-dried basis) were selected from three Indian coal mines. These samples are the true representatives of Indian coals. A major portion (around 70%) of the coals being mined currently in India has an average ash percentage of 45% (Patel et al., 2007).

4.2.2 Char Preparation

Char samples were prepared in the Argon (A_r) atmosphere (Naredi and Pisupati, 2007; Jayaraman et al., 2015) at 800, 900, and 1000^{o}C using a TGA (Model: STA449F3 Jupiter of Netzsch, Germany). The A_r flow rate was kept constant at 50 ml/min throughout the duration of char preparation. Approximately, 500 mg of the

air-dried coal sample was taken in a flat alumina sample container and temperature was raised at the rate of 10^{o}C/min until it reached the desired value. After attaining the targeted temperature, the sample was kept in the TGA for an additional 30 min; this ensured that the sample is free from the volatile matter (Saha, 2013).

4.2.3 Characterization of Coal and Char

***Proximate and Ultimate Analyses*:** The basic properties of coal samples were evaluated by conducting proximate and ultimate analyses (see Table 4.1) performed according to the Indian standards, viz. IS: 1350 (Part-I) 1984, IS: 1350 (Part-III) 1969, IS: 1350 (Part-IV/Sec-1) 1974, and IS: 1350 (Part-IV/Sec-2) 1975 (Saha et al., 2007).

Table 4.1: Analysis of three types of high ash coal samples used in the experimentation

Coal	Proximate analysis (air-dried basis)				Ultimate analysis (dry ash-free basis)				
	Moisture (*M*) (wt %)	Ash (*B*) (wt %)	Volatile Matter (*VM*) (wt %)	Fixed Carbon (*FC*) (wt %)	*C* (%)	*H* (%)	*N* (%)	*S* (%)	*O* (%)*
coal 1	6.5	41.3	24.5	27.7	71.17	5.42	1.65	1.05	20.71
coal 2	7.1	48.9	20.4	23.6	70.05	4.32	1.36	0.55	23.72
coal 3	9.7	27.0	25.7	37.6	76.56	5.43	1.63	0.95	15.43

*By difference

***Porosity Determination*:** Porosity of the coal samples was calculated using the true (ρ_t) and particle densities (ρ_P) as follows (Parkash and Chakrabartty, 1986; Saha, 2013):

$$Porosity\ (\%) = \frac{(\rho_t - \rho_P)}{\rho_t} \times 100 \qquad (4.8)$$

***Surface Area Measurements*:** The BET and CO_2 surface areas of the char samples were measured using Tristar 3000 surface area analyzer (Micromeritics, U.S.A.); BET surface area was determined using nitrogen as an adsorbate (99.999% purity). When CO_2 was used as an adsorbate, the respective surface areas were determined with the help of Dubinin–Radushkevich (D–R) equation. The adsorption isotherms for the BET- and CO_2-based surface areas were measured at -196 and 0^{o}C, respectively.

4.2.4 Gasification Experiments

The gasification experiments were conducted in the isothermally operated TGA (STA 449 F3 Jupiter, Netzsch, Germany) in a CO_2 atmosphere at 900, 950, 1000, and 1050^{o}C. The ultrapure dry nitrogen (N_2) was chosen as an inert gas; CO_2 (purity 99.999%) in concentrations of 30, 70, and 100% (balanced with N_2) was used as the gasifying agent. A char sample weighing 50 mg was spread uniformly on a flat alumina (Al_2O_3) container, which was placed on the TGA's sample carrier. The sample was heated at the rate of 10^{o}C/min up to the desired temperature with the inert gas (N_2) flow rate of 50 ml/min. For conducting a gasification experiment in the CO_2 atmosphere, the nitrogen flow was replaced—post attainment of the desired temperature—by the CO_2 flow (50 ml/min) to maintain the CO_2 atmosphere of the desired concentration. The TGA instrument used in the gasification experiments was calibrated and the repeatability of its measurements was tested by performing several experiments by employing calcium oxalate as a reference sample. For minimizing the buoyancy effect, each gasification experiment was corrected by a blank run, which was conducted under conditions identical to the gasification experiment. The TGA instrument has an ''S''-type thermocouple integrated with the furnace. It is positioned just below the sample holder and has the ability to measure the temperature accurately within $\pm$1.5^{o}C.

The char gasification reaction was conducted in a manner such that diffusional resistance is avoided. The particle size of the char not only influences the gasification reaction rate but also plays a crucial role in determining the rate controlling step (i.e., whether gasification is a chemical reaction or diffusion controlled). The absence of diffusional resistance is confirmed if no change in the reaction rate is observed for different sizes of the particles while all other reaction conditions remain unchanged. In this study, the char particle size was kept within -0.21 to $+$ 0.15 mm range for all gasification experiments. Earlier, Saha (2013) had conducted experiments with similar char samples to examine the presence/absence of the diffusional resistances using smaller particle sizes (-0.15 to $+$ 0.10 mm) at 1050^{o}C, and CO_2 partial pressure varying between 0.1 and 0.03 MPa. No change in the gasification reactivity was observed in these experiments. Thus, it is safe to infer that the char gasification reactions reported in this study, which are conducted in the temperature, CO_2 partial

pressure, and char particle size ranges of [900, 1050^{o}C], [0.1, 0.03 MPa], and [−0.21 mm, +0.15 mm], respectively, are kinetically (chemically) controlled.

During the char gasification experiments, the data consisting of the following seven attributes representing TGA operating conditions, and coal and char properties (these form the input space of the gasification models) were recorded (see Appendix 4.A). The basis of selection of the seven model inputs is given below.

- Char gasification temperature (^{o}C) (T_{G}) is a significant attribute since according to the Arrhenius law, the rate of the endothermic CO_2 gasification reaction increases with increasing T_G (Ahn et al., 2001; Liu et al., 2009). This can be easily verified in Appendix 4.A wherein it is noticed that the magnitudes of the reactivity index (r_{I}) and rate constant (k_{s}) of the char- CO_2 gasification reaction increase with increasing T_{G}.
- The magnitudes of r_{I} and k_{s} also increase with increasing CO_2 fraction (C_{CO_2}). Such an increase in the gasification rate is attributed to an increase in the number of reactant molecules diffusing to and getting adsorbed on the active sites of the char surface (Ahn et al., 2001; Zhang et al., 2006).
- Char preparation temperature (^{o}C) (T_{P}) is one of the multiple factors influencing the pyrolysis phenomenon during gasification. It is considered as a model input since both the model outputs, namely k_{s} and r_{I}, decrease with increasing T_{P} (Van Heek and Muhlen, 1987; Fermoso et al., 2010). It has been also reported (Wu et al., 2009) that increasing pyrolysis temperature adversely affects the gasification reaction, which is attributed to the decrease in the char's surface area as the pyrolysis temperature increases.
- Ash (wt %) (C_{A}) is an indicator of the coal's mineral matter content [*mineral matter* (wt %) =1.1 × *ash* (wt %)]. It lowers the extent of the carbonaceous material in the coal matrix and, thereby, negatively influences the quality and quantity of the gas produced. During combustion and gasification, mineral $matter$ in coal is converted into ash by chemical reactions. A typical sample of an Indian coal ash contains 90% or more SiO2, Al_2O_3, Fe_2O_3, and CaO. The balance 10% or less consists of MgO, Na_2O, K_2O, and TiO_2 as the basic constituents, and SO_3 and P_2O_5 as the acidic constituents. Details of the elemental analysis of the three coal-ash samples are given in Saha (2013). Some of these inorganic components act as catalysts in the coal conversion

processes. It is known that alkali and alkaline earth metals act as catalysts for carbon gasification (Miura et al., 1989; Takarada et al., 1986; Saha et al., 2011). The active metal ions (e.g., sodium, potassium, and calcium) must be connected to the carboxylic and phenolic groups to form active sites on the coal surface for exhibiting the catalytic activity. In the present investigation, it is found that the char gasification reaction rate increases with the increasing ash content. This result can be attributed to a higher catalytic activity of the inorganic elements with increasing ash content.

- During gasification, the macro- and meso-pores present in the char provide channels for the reacting gas to reach the active sites in the micropores where reaction takes place (Ng et al., 1984). The BET surface area (S_{N_2}) specifically measures the area of the meso- and macro-pores, whereas the CO_2 surface area (S_{CO_2}) indicates the micropore area. In the present investigation, Appendix 4.A clearly shows that both the rate constant of the char-CO_2 gasification reaction and reactivity index decrease with decreasing S_{N_2}, S_{CO_2}; and porosity (ϕ_P) values of the char. Accordingly, these three influential factors have been considered as model inputs (Bhatia and Gupta, 1992; Chi and Perlmutter, 1989; Feng and Bhatia, 2003).

4.3 RESULTS AND DISCUSSION

4.3.1 Determination of Reactivity Index Values

The reactivity index (r_I) is commonly used in determining and comparing the gasification reactivities of different chars under varying reaction conditions. It is defined as $r_I = \frac{0.5}{\tau_{0.5}}$, where $\tau_{0.5}$ refers to the time required to achieve 50% conversion (Takarada et al., 1985). This definition has been used in the present study for determining the gasification reactivity values of char samples. The r_I magnitudes were computed from $\tau_{0.5}$ values derived from the fractional conversion (x) versus time (t) relationship monitored in each experiment.

4.3.2 Determination of Rate Constant (k_s) Values Using Shrinking Un-Reacted Core Model

A number of kinetic models have been used to characterize coal gasification reactions. Among these, the most widely employed are the homogeneous, shrinking

unreacted core (SUC), and random pore models. Considering its simplicity and efficient representation of underlying phenomenon, SUC model has been utilized in the present investigation for representing the CO_2-char gasification kinetics (Molina and Mondragon, 1998; Irfan et al., 2011). This model assumes that the reaction occurs only on the surface of the progressively shrinking carbon core. In the beginning, the particle is surrounded by the gas. As conversion progresses, an increasing ash layer surrounds the continuously shrinking internal core of the unconverted material. This also indicates that the reaction front moves from the surface toward the particle's interior. The external radius of the particle remains unchanged during the entire reaction. In the char gasification reaction, it is reasonable to consider the porous inert solid product layer to be the ash layer. The SUC model also assumes that the un-reacted solid is impervious to the gas since it is densely packed. On the other hand, the ash layer is porous so that the reactant gas can diffuse inside and the product gas can diffuse out. The SUC model considers three reaction scenarios, namely diffusion through the gas film controlling, ash layer diffusion controlling, and chemical reaction controlling (Kim et al., 2011). As described earlier, the gasification reaction conditions used in this study correspond to the chemically controlled region and the corresponding SUC model is represented as

$$\frac{dx_{con}}{dt} = k\,(1 - x_{con})^{2/3} \tag{4.9}$$

Its solution is given by

$$3\,[1 - (1 - x_{con})^{1/3}] = kt \tag{4.10}$$

Or

$$1 - (1 - x_{con})^{1/3} = k_s\, t \tag{4.11}$$

where x_{con} represents the char conversion, t refers to the time (min), and k_s denotes the rate constant ($k_s = k/3$). In this study, the magnitude of k_s was determined from the slope of the $[1 - (1 - x_{con})^{1/3}]$ versus t plot. The k_s and r_I values corresponding to a total of 108 gasification experiments along with the activation energies computed using the Arrhenius equation are listed in Appendix 4.A.

4.3.3 Principal Component Analysis

While developing the data-driven models, it is necessary to avoid correlated inputs since these cause redundancy and unnecessarily increase the computational

load involved in the model construction. Accordingly, the seven inputs of the CI-based models were subjected to the principal component analysis (PCA) (Geladi and Kowalski, 1986). It helps in removing the linear correlations existing (if any) between the variables and, thereby, reducing the dimensionality of the input space of the model. In this study, seven principal components (PCs) were extracted from the gasification related input data listed in Appendix 4.A. The PCA yielded following values of the variance in the experimental data captured by the seven PCs—PC_1: 49.6%; PC_2: 19.8%; PC_3:14.3%; PC_4: 14.3%; PC_5: 4%; PC_6: 0.5%; and PC_7: 0.2% (PC_i denotes the i^{th} PC). It is thus seen, that the first four PCs have captured a large percentage (≈ 95%) of the variance in the seven inputs. This result indicates that it is possible to consider only the first four PCs (v_1– v_4) as defined below, in place of the original seven inputs for developing the gasification models.

$$\upsilon_1 = -0.376\ \hat{x}_3^j + 0.465\ \hat{x}_4^j + 0.379\ \hat{x}_5^j + 0.518\ \hat{x}_6^j + 0.48\ \hat{x}_7^j \tag{4.12}$$

$$\upsilon_2 = 0.599\ \hat{x}_3^j - 0.401\ \hat{x}_4^j + 0.597\ \hat{x}_5^j + 0.033\ \hat{x}_6^j + 0.35\ \hat{x}_7^j \tag{4.13}$$

$$\upsilon_3 = -\ \hat{x}_1^j \tag{4.14}$$

$$\upsilon_4 = \ \hat{x}_2^j \tag{4.15}$$

where $\hat{x}_q^j$ (q = 1, 2,. . ., Q; Q = 7) denote the normal scores (standardized variables) pertaining to the values of the seven inputs listed in Appendix 4.A. The normalized variables were obtained as follows:

$$\hat{x}_q^j = \frac{x_q^j - \overline{x}_q}{\sigma_q}\ ; \quad j = 1, 2, \ldots,\ N_{\text{pat}} \tag{4.16}$$

where x_q^j represents j^{th} value of q^{th} un-normalized input variable, x_q; $\overline{x}_q$ refers to the mean of x_q, and σ_q represents standard deviation of x_q. The mean and standard deviation values used in the normalization procedure are given below where $\overline{x}_1$, $\overline{x}_2$, $\overline{x}_3$, $\overline{x}_4$, $\overline{x}_5$, $\overline{x}_6$ and $\overline{x}_7$ respectively represent the mean values of T_{G}, C_{CO_2}, T_{P}, S_{N_2}, C_{A}, S_{CO_2} and ϕ_P.

$\overline{x}_1$ = 975 (^oC); $\overline{x}_2 = 0.667$; $\overline{x}_3 = 900$ (^oC); $\overline{x}_4 = 28.81$ (m^2/g);

$\overline{x}_5 = 56.387(\%)$; $\overline{x}_6 = 239.24\ (m^2/g)$; $\overline{x}_7 = 19.40\ (\%)$. (4.17)

The corresponding standard deviation values are as given below.

$\sigma_1 = 56.162\ (^oC)$; $\sigma_2 = 0.288$; $\sigma_3 = 82.03\ (^oC)$; $\sigma_4 = 12.84\ (m^2/g)$;

$\sigma_5 = 10.826\ (\%)$; $\sigma_6 = 40.167\ (m^2/g)$; $\sigma_7 = 4.172\ (\%)$. (4.18)

Similar to the model inputs, the two outputs, namely k_s and r_I, were also normalized as follows:

$$\hat{k}_\mathrm{S}^j = \frac{k_\mathrm{S}^j - \overline{k}_\mathrm{S}}{\sigma_k};\ j = 1, 2, \ldots, N_\mathrm{pat} \quad (4.19)$$

$$\hat{r}_\mathrm{I}^j = \frac{r_\mathrm{I}^j - \overline{r}_\mathrm{I}}{\sigma_r};\ j = 1, 2, \ldots, N_\mathrm{pat} \quad (4.20)$$

where, $\overline{k}_s$ and $\overline{r}_\mathrm{I}$ refer to the mean values of k_S and r_I, respectively, and their corresponding standard deviations are denoted by σ_k and σ_r. The magnitudes of these are as follows: $\overline{k}_s$ = 0.0125 (min^{-1}); $\overline{r}_\mathrm{I}$= 0.0261(min^{-1}); σ_k = 0.00762 (min^{-1}); σ_r= 0.0149 (min^{-1}).

4.3.4 CI-Based Models for the Prediction of CO_2 Gasification Rate Constant and Reactivity Index

The PCA-transformed four variables (v_1–v_4) defined by Equations (4.12) – (4.15) were used as inputs in developing the GP-, MLP-, and SVR-based k_S and r_I predicting models. For constructing and assessing the generalization ability of these models, the experimental data set (See Appendix 4.A) consisting of 108 input–output patterns was randomly partitioned in 3:1 ratio into training (81 patterns) and test (27 patterns) sets. While the former set was used in training the CI-based models, the latter was used in testing their generalization capability. The output prediction accuracy and generalization performance of each CI-based model were evaluated in terms of the *coefficient of correlation* (*CC*), *root mean squared error* (*RMSE*), and *mean absolute percent error* (*MAPE*) values pertaining to the experimental and model-predicted quantities of the char gasification rate constant and reactivity index.

GP-Based Modeling of Gasification Performance Variables

The two GP-based models predicting the gasifier performance variables, namely k_S and r_I, were developed using *Eureqa Formulize* software package (Schmidt and

Lipson, 2009). In the GP (Koza, 1992; Kinnear, 1994) implementation, the *RMSE*-dependent fitness function was used to assign the fitness values of the candidate expressions. The effects of the GP procedural parameters, such as the size of the training and test sets as also various input normalization schemes, were studied rigorously. The two overall best models, respectively predicting char *gasification rate constant* (k_S) (GP model-I) and *reactivity index* (r_I) (GP model-II), are as follows:

$$\hat{k}_s = 0.1408\, \upsilon_3^2 + 0.08651\upsilon_2\upsilon_3 - 0.1038\upsilon_3\upsilon_4 + 0.2912\upsilon_4 + 0.1316\upsilon_1 - 0.1657\upsilon_2 - 0.8297\upsilon_3 - 0.1657 \quad (4.21)$$

$$\hat{r}_I = 0.04857\, \upsilon_1^2 - 0.1374\upsilon_3\upsilon_4 + 0.3116\upsilon_4 + 0.2684\upsilon_1 - 0.07561\upsilon_2 - 0.7114\upsilon_3 - 0.1374 \quad (4.22)$$

where υ_i denotes i^{th} PCA-transformed variable, and $\hat{k}_s$ and $\hat{r}_I$, respectively, refer to the normalized values of k_S and r_I (see eqs. 4.19 and 4.20). As can be seen, both GP models have nonlinear forms. It is also observed that these models contain all the four PCA-transformed variables (υ_1– υ_4). This is noteworthy since the GP formalism is known to use only those inputs from the supplied data that significantly influence the dependent variable (Cheng and Worzel, 2015). From Equations (4.12)–(4.15), it is noticed that the four PCA transformed variables have been derived using as many subsets of the seven gasification variables and parameters defining the coal and char properties. The presence of all four PCA-transformed variables in the GP-based models in turn underlines the importance of the original seven variables and parameters in determining the values of the char gasification rate constant and reactivity index.

MLP- and SVR-Based Modeling of Gasification Performance Variable

The details of the heuristic procedure involved in obtaining an optimal MLP network (Freeman and Skapura, 1991; Bishop, 1994) model possessing good prediction and generalization performance has been explained in Chapter 2, section 2.2.1 ; a detailed description of the SVR (Vapnik, 1995; Burges, 1998) and its implementation has been provided in Chapter 2, section 2.3. In the present study, SVR-based models were developed using the ε-SVR module of the data-mining package known as *Rapid Miner* (2014) and MLP-based models were built using

IBM-SPSS (2011) package. The parameter values and other attributes of the optimal MLP-based models I and II, respectively predicting the k_s and r_I magnitudes, are listed in Table 4.2; the details of the corresponding SVR-based optimal models I and II, are given in Table 4.3.

Table 4.2: Details of the architecture of the optimal MLP-based models*and the corresponding EBP algorithm parameter values

Model no.	Output variable	Input nodes	Number of hidden layers	Number of hidden nodes	Transfer function for hidden nodes	Transfer function for output node	Momentum coefficient (μ_{ebp})	Learning rate (η)
I	k_S	4	1	3	tanh	identity	0.07	0.4
II	r_I	4	1	3	tanh	identity	0.004	0.2

*<u>*Other details of the EBP-based models*</u>: (a) *rescaling method used for the scale-dependent variables: standardized*; (b) *learning mode*: *batch*; (c) *the random number generator seed value with respect to the optimal MLP model*: *200*; (d) *maximum training epochs*:*100*.

Table 4.3: Details of the ε-insensitive loss function-based optimal SVR models and the corresponding parameter values

Model no.	Output variable	Kernel type	Kernel gamma	Kernel degree	Kernel cache	*C* (cost parameter)	ε
I	k_S	ANOVA	0.25	2.0	200	2.7	0.05
II	r_I	ANOVA	25	2.0	200	256	0.05

Comparison of the CI-Based Models Predicting the Gasification Rate Constant

The magnitudes of *CC*, *RMSE*, and *MAPE* pertaining to the k_s predictions made by the GP model-I, MLP model-I, and SVR model-I are specified in Table 4.4. It is seen in this table that the *CC* (*RMSE*/*MAPE*) magnitudes with respect to the experimental k_s values and those predicted by the CI-based models for both training and test set data are high (low) and comparable. This result indicates that the stated models possess an excellent k_s prediction accuracy and generalization capability. Figure 4.1 consists of the three parity plots displaying the experimental k_s values and

those predicted by the GP-, MLP-, and SVR-based models, respectively. As can be noticed in all the three (*a*-*c*) panels of Figure 4.1, there exists a good agreement between the experimental and model-predicted k_S values.

Table 4.4: Statistical analysis of the prediction and generalization performance of the gasification rate constant (k_S) predicting GP-, MLP-, and SVR-based models

Model	Training set			Test set		
	CC_{trn}	$RMSE_{trn}$	$MAPE_{trn}$	CC_{tst}	$RMSE_{tst}$	$MAPE_{tst}$
GP model-I	0.974	1.79×10^{-3}	11.197	0.987	9.75×10^{-4}	8.467
MLP model-I	0.984	1.37×10^{-3}	9.194	0.993	6.92×10^{-4}	8.687
SVR model-I	0.991	1.04×10^{-3}	3.342	0.989	8.77×10^{-4}	9.900

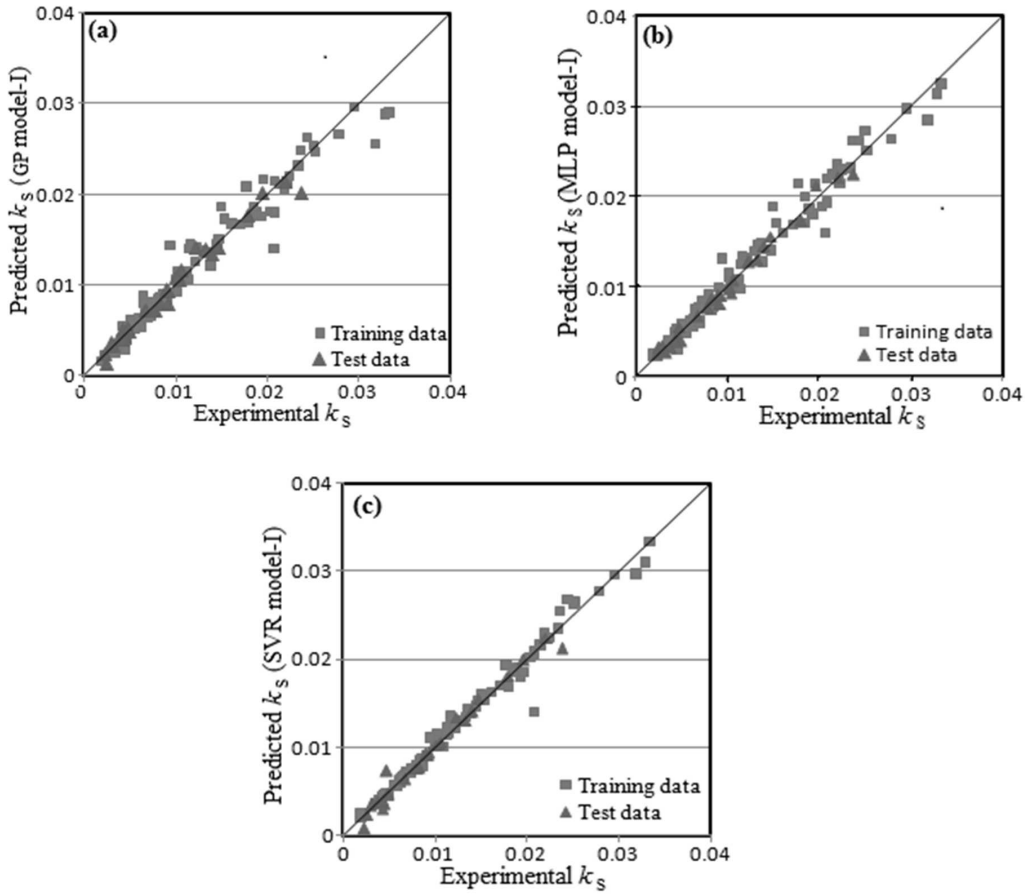


Figure 4.1: Parity plots of experimental versus model-predicted values of char gasification rate constant (k_S, min^{-1}); Panels (*a*), (*b*), and (*c*), respectively, depict plots pertaining to the k_S predictions made by GP-, MLP-, and SVR-based models.

Table 4.5: Statistical analysis of the prediction and generalization performance of the reactivity index (r_I) predicting GP-, MLP-, and SVR-based models

Model	Training set			Test set		
	CC_{trn}	$RMSE_{trn}$	$MAPE_{trn}$	CC_{tst}	$RMSE_{tst}$	$MAPE_{tst}$
GP model-II	0.961	4.38×10^{-3}	15.069	0.971	3.42×10^{-3}	13.307
MLP model-II	0.982	3.03×10^{-3}	9.478	0.971	2.89×10^{-3}	11.351
SVR model-II	0.991	2.02×10^{-3}	3.442	0.974	2.70×10^{-3}	14.134

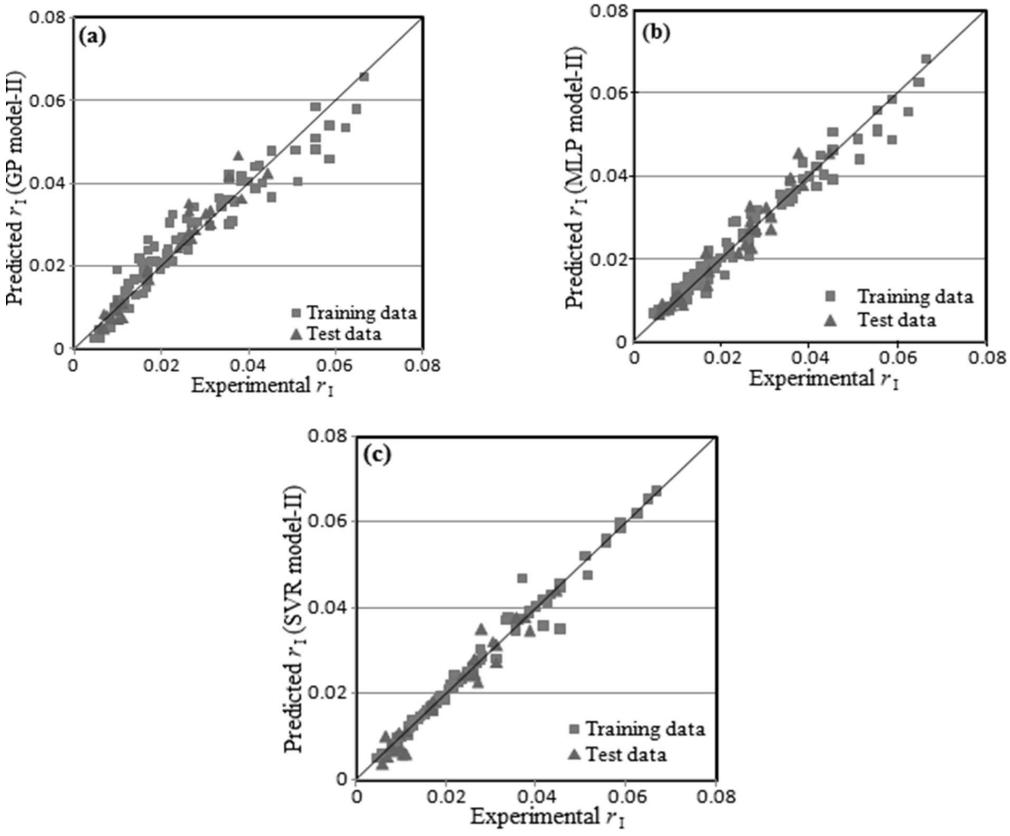


Figure 4.2: Parity plots of experimental versus model-predicted values of reactivity index (r_I, min^{-1}); panels (*a*), (*b*) and (*c*), respectively, depict plots pertaining to the r_I predictions made by GP-, MLP-, and SVR-based models.

Comparison of CI-Based Models for Reactivity Index Prediction

The magnitudes of the *CC*, *RMSE*, and *MAPE* pertaining to the r_I predictions made by the three CI-based models are listed in Table 4.5. From the table, it is clear that similar to the k_S-predicting models, each one of the three CI-based models for the reactivity index possesses an excellent prediction accuracy and generalization capability. Figure 4.2 contains three (*a*-*c*) panels, which respectively show how well the predictions of the GP-, MLP-, and SVR-based models match the corresponding experimental reactivity index values. In all these panels, it is noticed that there exists a good match between the experimental and model-predicted r_I values.

***Steiger's Test*:** A statistical test known as Steiger's z-test (Steiger, 1980) was performed for comparing the prediction performance of the GP-, MLP-, and SVR-based models. It tests the null hypothesis (H_0) that statistically two correlation coefficient magnitudes are not different, i.e., $CC_{AB} = CC_{AC}$, where CC_{AB} (CC_{AC}) refers to the correlation coefficient pertaining to the model B (model C) predicted outputs and their corresponding experimental counterparts. The results of the Steiger's z-test for the CI-based models predicting k_S and r_I are listed in Tables 4.6 and 4.7, respectively. It is seen in these tables that for all the six model pairs (three each for k_S and r_I) the *p*-values are less than 0.05. This indicates a uniform rejection of the null hypothesis (at 95% confidence level) about the statistical equivalence of the *CC* magnitudes pertaining to k_S and r_I predictions made by the model pairs GP–MLP, MLP–SVR, and GP–SVR. It can thus be concluded that the differences in the *CC* magnitudes of the stated model pairs are statistically significant. From the *CC* magnitudes listed in Tables 4.4 and 4.5, it is observed that among the three CI-based models, the MLP- based k_S prediction model, and the SVR-based r_I prediction model possess high prediction accuracies, and best generalization capabilities. Therefore, these models are more suited for the prediction of the gasification rate constant and reactivity index values. It may, however, be noted that there exist only minor differences between the prediction accuracies/generalization capabilities of the three CI-based models. Accordingly, the GP-based models, due to their simplicity and lower complexity, should be preferred if the convenience of usage is the main criterion for the utilization of a model.

Table 4.6: Results of Steiger's z-test testing the null hypothesis (H_0) pertaining to the equivalence of correlation coefficient (*CC*) magnitudes with respect to the model pairs predicting the gasification rate constant (k_S) values

Model pair (B-C)	*df*	CC_{AB}	CC_{AC}	CC_{BC}	*z*	*p-value*	H_0
GP-MLP	108	0.977	0.987	0.989	3.648	2.63×10^{-4}	Reject
MLP-SVR	108	0.987	0.991	0.991	2.032	4.21×10^{-2}	Reject
SVR-GP	108	0.991	0.977	0.986	5.301	1.14×10^{-7}	Reject

H_0: $CC_{AB} = CC_{AC}$, where A denotes experimental values of k_S; df refers to the degrees of freedom; reject H_0 if p-value < 0.05.

Table 4.7: Results of the Steiger's z-test testing the null hypothesis (H_0) pertaining to the equivalence of correlation coefficient (*CC*) magnitudes with respect to the model pairs predicting reactivity index (r_I) values

Model pair (B-C)	*df*	CC_{AB}	CC_{AC}	CC_{BC}	*z*	*p-value*	H_0
GP-MLP	108	0.960	0.979	0.980	-4.026	5.67×10^{-5}	Reject
MLP-SVR	108	0.979	0.989	0.976	-2.949	3.18×10^{-3}	Reject
SVR-GP	108	0.989	0.960	0.961	5.904	3.53×10^{-9}	Reject

H_0: $CC_{AB} = CC_{AC}$, where A denotes experimental values of r_I ; df refers to the degrees of freedom; reject H_0 if p-value < 0.05.

4.4 CONCLUDING REMARKS

The present chapter reports results of the CI-based data-driven modeling for the prediction of *char gasification rate constant* (k_S), *reactivity index* (r_I) magnitudes corresponding to the gasification of high ash Indian coals. The data for this modeling were collected by conducting gasification experiments in a TGA in the CO_2 atmosphere. These data were first fitted to the SUC model to obtain values of k_S and r_I, which were then correlated with the seven parameters (model inputs) consisting of

the coal and char properties, and the gasification conditions. The data-driven models possessing an excellent prediction accuracy and generalization capability were developed using three CI formalisms, namely GP, MLPNN, and SVR. Among these, the GP-based ones are less complex, easier to grasp, and more convenient to deploy in a practical setting. A notable feature of this study is that phenomenological (i.e., SUC) and data-driven approaches (GP, MLP, and SVR) have been integrated into developing comprehensive models for predicting two important kinetic parameters associated with the gasification of high ash coals. The models developed in this study can be gainfully employed in the design and operation of the gasifiers using high ash coals, which are available in abundance globally. Additionally, the models for determining the rate constant can be used for predicting the activation energies of the coal gasification reactions involving CO_2 in the temperature range of 900–1050^{o}C.

NOMENCLATURE

$\overline{k}_s$	mean values of k_S
$\overline{x}_q$	The mean value of x_q
x_q^j	j^{th} value of q^{th} un-normalized input variable, x_q
$\hat{x}_q^j$	Normal scores (standardized variables) pertaining to the values of the seven inputs
$\overline{r}_I$	mean values of r_I

Greek symbols

σ_k	standard deviations of k_S
σ_q	Standard deviation of x_q
σ_r	standard deviations of r_I

Appendix 4.A: Experimental data consisting of coal and char properties and gasification conditions, and the corresponding values of gasification rate constant and reactivity index utilized in building CI-based models

Expt. no.	Gasification temperature (T_G) (^oC)	CO_2 fraction (C_{CO_2})	Char preparation temperature (T_P) (^oC)	Surface area (BET) method (S_{N_2}) (m^2/g)	Ash (C_A) (%)	Surface area (CO_2) method (S_{CO_2}) (m^2/g)	Porosity (ϕ_P) (%)	Rate constant (k_S) (min^{-1})	Reactivity index (r_I) (min^{-1})	Activation energy kJ/mole
1	900	1	800	44.85	59.86	266.67	23.35	0.0091	0.0172	113.29
2	950	1	800	44.85	59.86	266.67	23.35	0.0140	0.0263	
3	1000	1	800	44.85	59.86	266.67	23.35	0.0238	0.0376	
4	1050	1	800	44.85	59.86	266.67	23.35	0.0329	0.0556	
5	900	1	900	24.67	59.86	246.38	21.46	0.0063	0.0143	119.06
6	950	1	900	24.67	59.86	246.38	21.46	0.0104	0.0238	
7	1000	1	900	24.67	59.86	246.38	21.46	0.0162	0.0357	
8	1050	1	900	24.67	59.86	246.38	21.46	0.0253	0.0454	
9	900	1	1000	18.19	59.86	200.01	20.10	0.0045	0.0100	133.23
10	950	1	1000	18.19	59.86	200.01	20.10	0.0082	0.0172	
11	1000	1	1000	18.19	59.86	200.01	20.10	0.0123	0.0278	
12	1050	1	1000	18.19	59.86	200.01	20.10	0.0220	0.0357	
13	900	0.7	800	44.85	59.86	266.67	23.35	0.0070	0.0151	115.33
14	950	0.7	800	44.85	59.86	266.67	23.35	0.0103	0.0260	
15	1000	0.7	800	44.85	59.86	266.67	23.35	0.0194	0.0357	

Appendix 4.A continued…

Expt. no.	Gasification temperature (T_G) (^oC)	CO_2 fraction (C_{CO_2})	Char preparation temperature (T_P) (^oC)	Surface area (BET) method (S_{N_2}) (m^2/g)	Ash (C_A) (%)	Surface area (CO_2) method (S_{CO_2}) (m^2/g)	Porosity (ϕ_P) (%)	Rate constant (k_S) (min^{-1})	Reactivity index (r_I) (min^{-1})	Activation energy kJ/mole
16	1050	0.7	800	44.85	59.86	266.67	23.35	0.0251	0.0556	
17	900	0.7	900	24.67	59.86	246.38	21.46	0.0050	0.0111	125.32
18	950	0.7	900	24.67	59.86	246.38	21.46	0.0084	0.0208	
19	1000	0.7	900	24.67	59.86	246.38	21.46	0.0123	0.0313	
20	1050	0.7	900	24.67	59.86	246.38	21.46	0.0223	0.0434	
21	900	0.7	1000	18.19	59.86	200.01	20.10	0.0039	0.0084	142.63
22	950	0.7	1000	18.19	59.86	200.01	20.10	0.0070	0.0167	
23	1000	0.7	1000	18.19	59.86	200.01	20.10	0.0114	0.0263	
24	1050	0.7	1000	18.19	59.86	200.01	20.10	0.0204	0.0339	
25	900	0.3	800	44.85	59.86	266.67	23.35	0.0043	0.0100	121.67
26	950	0.3	800	44.85	59.86	266.67	23.35	0.0087	0.0172	
27	1000	0.3	800	44.85	59.86	266.67	23.35	0.0133	0.0263	
28	1050	0.3	800	44.85	59.86	266.67	23.35	0.0178	0.0385	
29	900	0.3	900	24.67	59.86	246.38	21.46	0.0033	0.0077	140.00
30	950	0.3	900	24.67	59.86	246.38	21.46	0.0064	0.0128	
31	1000	0.3	900	24.67	59.86	246.38	21.46	0.0102	0.0217	

Appendix 4.A continued…

Expt. no.	Gasification temperature (T_G) (^oC)	CO_2 fraction (C_{CO_2})	Char preparation temperature (T_P) (^oC)	Surface area (BET) method (S_{N_2}) (m^2/g)	Ash (C_A) (%)	Surface area (CO_2) method (S_{CO_2}) (m^2/g)	Porosity (ϕ_P) (%)	Rate constant (k_S) (min^{-1})	Reactivity index (r_I) (min^{-1})	Activation energy kJ/mole
32	1050	0.3	900	24.67	59.86	246.38	21.46	0.0172	0.0313	
33	900	0.3	1000	18.19	59.86	200.01	20.10	0.0020	0.0048	163.72
34	950	0.3	1000	18.19	59.86	200.01	20.10	0.0048	0.0102	
35	1000	0.3	1000	18.19	59.86	200.01	20.10	0.0083	0.0172	
36	1050	0.3	1000	18.19	59.86	200.01	20.10	0.0137	0.0263	
37	900	1	800	50.35	67.5	296.25	24.54	0.0103	0.0364	101.25
38	950	1	800	50.35	67.5	296.25	24.54	0.0147	0.0444	
39	1000	1	800	50.35	67.5	296.25	24.54	0.0220	0.0588	
40	1050	1	800	50.35	67.5	296.25	24.54	0.0334	0.0667	
41	900	1	900	29.39	67.5	290.17	22.91	0.0067	0.0167	107.62
42	950	1	900	29.39	67.5	290.17	22.91	0.0105	0.0270	
43	1000	1	900	29.39	67.5	290.17	22.91	0.0154	0.0385	
44	1050	1	900	29.39	67.5	290.17	22.91	0.0237	0.0625	
45	900	1	1000	19.6	67.5	240.17	20.43	0.0043	0.0102	130.94
46	950	1	1000	19.6	67.5	240.17	20.43	0.0066	0.0156	
47	1000	1	1000	19.6	67.5	240.17	20.43	0.0095	0.0227	

Appendix 4.A continued…

Expt. no.	Gasification temperature (T_G) (^oC)	CO_2 fraction (C_{CO_2})	Char preparation temperature (T_P) (^oC)	Surface area (BET) method (S_{N_2}) (m^2/g)	Ash (C_A) (%)	Surface area (CO_2) method (S_{CO_2}) (m^2/g)	Porosity (ϕ_P) (%)	Rate constant (k_S) (min^{-1})	Reactivity index (r_I) (min^{-1})	Activation energy kJ/mole
48	1050	1	1000	19.6	67.5	240.17	20.43	0.0209	0.0417	
49	900	0.7	800	50.35	67.5	296.25	24.54	0.0093	0.0278	105.80
50	950	0.7	800	50.35	67.5	296.25	24.54	0.0139	0.0417	
51	1000	0.7	800	50.35	67.5	296.25	24.54	0.0209	0.0556	
52	1050	0.7	800	50.35	67.5	296.25	24.54	0.0319	0.0649	
53	900	0.7	900	29.39	67.5	290.17	22.91	0.0057	0.0139	110.09
54	950	0.7	900	29.39	67.5	290.17	22.91	0.0092	0.0254	
55	1000	0.7	900	29.39	67.5	290.17	22.91	0.0117	0.0357	
56	1050	0.7	900	29.39	67.5	290.17	22.91	0.0222	0.0588	
57	900	0.7	1000	19.6	67.5	240.17	20.43	0.0038	0.0086	135.92
58	950	0.7	1000	19.6	67.5	240.17	20.43	0.0068	0.0161	
59	1000	0.7	1000	19.6	67.5	240.17	20.43	0.0106	0.0270	
60	1050	0.7	1000	19.6	67.5	240.17	20.43	0.0190	0.0333	
61	900	0.3	800	50.35	67.5	296.25	24.54	0.0060	0.0263	111.22
62	950	0.3	800	50.35	67.5	296.25	24.54	0.0090	0.0313	
63	1000	0.3	800	50.35	67.5	296.25	24.54	0.0146	0.0400	

Appendix 4.A continued…

Expt. no.	Gasification temperature (T_G) (^oC)	CO_2 fraction (C_{CO_2})	Char preparation temperature (T_P) (^oC)	Surface area (BET) method (S_{N_2}) (m^2/g)	Ash (C_A) (%)	Surface area (CO_2) method (S_{CO_2}) (m^2/g)	Porosity (ϕ_P) (%)	Rate constant (k_S) (min^{-1})	Reactivity index (r_I) (min^{-1})	Activation energy kJ/mole
64	1050	0.3	800	50.35	67.5	296.25	24.54	0.0215	0.0454	
65	900	0.3	900	29.39	67.5	290.17	22.91	0.0046	0.0119	115.00
66	950	0.3	900	29.39	67.5	290.17	22.91	0.0071	0.0179	
67	1000	0.3	900	29.39	67.5	290.17	22.91	0.0102	0.0263	
68	1050	0.3	900	29.39	67.5	290.17	22.91	0.0181	0.0370	
69	900	0.3	1000	19.6	67.5	240.17	20.43	0.0024	0.0059	145.91
70	950	0.3	1000	19.6	67.5	240.17	20.43	0.0045	0.0100	
71	1000	0.3	1000	19.6	67.5	240.17	20.43	0.0074	0.0161	
72	1050	0.3	1000	19.6	67.5	240.17	20.43	0.0134	0.0250	
73	900	0.7	800	37.23	41.8	228.34	15.20	0.0048	0.0097	139.70
74	950	0.7	800	37.23	41.8	228.34	15.20	0.0115	0.0188	
75	1000	0.7	800	37.23	41.8	228.34	15.20	0.0186	0.0282	
76	1050	0.7	800	37.23	41.8	228.34	15.20	0.0245	0.0515	
77	900	0.7	900	27.65	41.8	221.99	14.61	0.0043	0.0112	145.83
78	950	0.7	900	27.65	41.8	221.99	14.61	0.0086	0.0172	
79	1000	0.7	900	27.65	41.8	221.99	14.61	0.0148	0.0250	

Appendix 4.A continued…

Expt. no.	Gasification temperature (T_G) (^oC)	CO_2 fraction (C_{CO_2})	Char preparation temperature (T_P) (^oC)	Surface area (BET) method (S_{N_2}) (m^2/g)	Ash (C_A) (%)	Surface area (CO_2) method (S_{CO_2}) (m^2/g)	Porosity (ϕ_P) (%)	Rate constant (k_S) (min^{-1})	Reactivity index (r_I) (min^{-1})	Activation energy kJ/mole
80	1050	0.7	900	27.65	41.8	221.99	14.61	0.0235	0.0454	
81	900	0.7	1000	7.33	41.8	163.20	12.02	0.0026	0.0060	165.10
82	950	0.7	1000	7.33	41.8	163.20	12.02	0.0064	0.0125	
83	1000	0.7	1000	7.33	41.8	163.20	12.02	0.0103	0.0185	
84	1050	0.7	1000	7.33	41.8	163.20	12.02	0.0186	0.0278	
85	900	1	800	37.23	41.8	228.34	15.2	0.0075	0.0160	118.55
86	950	1	800	37.23	41.8	228.34	15.2	0.0122	0.0250	
87	1000	1	800	37.23	41.8	228.34	15.2	0.0196	0.0385	
88	1050	1	800	37.23	41.8	228.34	15.2	0.0296	0.0510	
89	900	1	900	27.65	41.8	221.99	14.61	0.0056	0.0126	138.71
90	950	1	900	27.65	41.8	221.99	14.61	0.0104	0.0227	
91	1000	1	900	27.65	41.8	221.99	14.61	0.0180	0.0303	
92	1050	1	900	27.65	41.8	221.99	14.61	0.0279	0.0427	
93	900	1	1000	7.33	41.8	163.20	12.02	0.0036	0.0083	155.07
94	950	1	1000	7.33	41.8	163.20	12.02	0.0079	0.0168	
95	1000	1	1000	7.33	41.8	163.20	12.02	0.0130	0.0220	

Appendix 4.A continued…

Expt. no.	Gasification temperature (T_G) (^oC)	CO_2 fraction (C_{CO_2})	Char preparation temperature (T_P) (^oC)	Surface area (BET) method (S_{N_2}) (m^2/g)	Ash (C_A) (%)	Surface area (CO_2) method (S_{CO_2}) (m^2/g)	Porosity (ϕ_P) (%)	Rate constant (k_S) (min^{-1})	Reactivity index (r_I) (min^{-1})	Activation energy kJ/mole
96	1050	1	1000	7.33	41.8	163.20	12.02	0.0225	0.0357	
97	900	0.3	800	37.23	41.8	228.34	15.2	0.0030	0.0067	160.39
98	950	0.3	800	37.23	41.8	228.34	15.2	0.0066	0.0126	
99	1000	0.3	800	37.23	41.8	228.34	15.2	0.0115	0.0213	
100	1050	0.3	800	37.23	41.8	228.34	15.2	0.0197	0.0357	
101	900	0.3	1000	7.33	41.8	163.20	12.02	0.0020	0.0060	195.15
102	950	0.3	1000	7.33	41.8	163.20	12.02	0.0046	0.0094	
103	1000	0.3	1000	7.33	41.8	163.20	12.02	0.0080	0.0156	
104	1050	0.3	1000	7.33	41.8	163.20	12.02	0.0208	0.0217	
105	900	0.3	900	27.65	41.8	221.99	14.61	0.0023	0.0070	166.08
106	950	0.3	900	27.65	41.8	221.99	14.61	0.0051	0.0118	
107	1000	0.3	900	27.65	41.8	221.99	14.61	0.0109	0.0200	
108	1050	0.3	900	27.65	41.8	221.99	14.61	0.0151	0.0235	

REFERENCES

Adschiri, T., Shiraha, T., Kojima, T., and Furusawa, T. (1986). Prediction of CO_2 gasification rate of char in fluidized bed gasifier. *Fuel*, *65*(12), 1688-1693.

Ahn, D. H., Gibbs, B. M., Ko, K. H., and Kim, J. J. (2001). Gasification kinetics of an Indonesian sub-bituminous coal-char with CO_2 at elevated pressure. *Fuel*, *80*(11), 1651-1658.

Ballal, G., and Zygourakis, K. (1986). Gasification of coal chars with carbon dioxide and oxygen. *Chemical Engineering Communications*, *49*(1-3), 181-195.

Bhatia, S. K., and Gupta, J. S. (1992). Mathematical modelling of gas-solid reactions: Effect of pore structure. *Reviews in Chemical Engineering*, *8*(3-4), 177-258.

Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments*, *65*(6), 1803-1832.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

Chejne, F., Lopera, E., and Londoño, C. A. (2011). Modelling and simulation of a coal gasification process in pressurized fluidized bed. *Fuel*, *90*(1), 399-411.

Cheng, C., and Worzel, W. P. (2015). Application of machine-learning methods to understand gene expression regulation. In *Genetic Programming Theory and Practice* XII. Riolo, W. P., Worzel, M., and Kotanchek (Eds.), Chapter 1, Springer International Publishing, Switzerland, pp 1-15.

Chi, W. K., and Perlmutter, D. D. (1989). The effect of pore structure on the char-steam reaction. *AIChE Journal*, *35*(11), 1791-1802.

Feng, B., and Bhatia, S. K. (2003). Variation of the pore structure of coal chars during gasification. *Carbon*, *41*(3), 507-523.

Fermoso, J., Gil, M. V., Borrego, A. G., Pevida, C., Pis, J. J., and Rubiera, F. (2010). Effect of the pressure and temperature of devolatilization on the morphology and steam gasification reactivity of coal chars. *Energy & Fuels*, *24*(10), 5586-5595.

Freeman, J. A., and Skapura, D. M. (1991). *Neural Networks Algorithms, Applications, and Programming Techniques*. Addison-Wesley, Reading, M.A, USA.

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, *185*, 1-17.

Gururajan, V. S., Agarwal, P. K., and Agnew, J. B. (1992). Mathematical modelling of fluidized bed coal gasifiers: Chemical reaction engineering. *Chemical engineering research & design*, *70*(A3), 211-238.

IBM SPSS Neural Networks 20 manual, IBM: Chicago, 2011.

Irfan, M. F., Usman, M. R., and Kusakabe, K. (2011). Coal gasification in CO_2atmosphere and its kinetics since 1948: a brief review. *Energy*, *36*(1), 12-40.

Jayaraman, K., Gokalp, I., Bonifaci, E., and Merlo, N. (2015). Kinetics of steam and CO_2 gasification of high ash coal–char produced under various heating rates. *Fuel*, *154*, 370-379.

Kim, Y. T., Seo, D. K., and Hwang, J. (2011). Study of the effect of coal type and particle size on char–CO_2 gasification via gas analysis. *Energy & Fuels*, *25*(11), 5044-5054.

Kinnear, K. E. (1994). *Advances in Genetic Programming* (Vol. 1). MIT press, Cambridge, MA.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (Vol. 1). MIT press, Cambridge, MA.

Kristiansen, A. (1996). *Understanding Coal Gasification* (Vol. 86). IEA Coal Research report IEACR-86, International Energy Agency, London, U.K.

Liu, H., Zhu, H., Kaneko, M., Kato, S., and Kojima, T. (2009). High-temperature gasification reactivity with steam of coal chars derived under various pyrolysis conditions in a fluidized bed. *Energy & Fuels*, *24*(1), 68-75.

Miller, B. G. (2011). Clean coal technologies for advanced power generation, In *Clean Coal Engineering Technology*. Chapter 7, Elsevier, Burlington, pp 251–300.

Miura, K., Hashimoto, K., and Silveston, P. L. (1989). Factors affecting the reactivity of coal chars during gasification, and indices representing reactivity. *Fuel*, *68*(11), 1461-1475.

Molina, A., and Mondragon, F. (1998). Reactivity of coal gasification with steam and CO_2. *Fuel*, *77*(15), 1831-1839.

Moreea-Taha, R. (2000). *Modelling and Simulation for Coal Gasifier*. Report CCC=42, IEA Coal Research, Putney Hill, London. ISBN 92-9029-354-3.

Naredi, P., and Pisupati, S. V. (2007). Interpretation of char reactivity profiles obtained using a thermogravimetric analyzer. *Energy & Fuels*, *22*(1), 317-320.

Ng, S. H., Fung, D. P., and Kim, S. D. (1984). Some physical properties of Canadian coals and their effects on coal reactivity. *Fuel*, *63*(11), 1564-1569.

Ochoa, J., Cassanello, M. C., Bonelli, P. R., and Cukierman, A. L. (2001). CO_2 gasification of Argentinean coal chars: A kinetic characterization. *Fuel Processing Technology*, *74*(3), 161-176.

Parkash, S., and Chakrabartty, S. K. (1986). Microporosity in Alberta plains coals. *International journal of coal geology*, *6*(1), 55-70.

Patel, S. U., Kumar, B. J., Badhe, Y. P., Sharma, B. K., Saha, S., Biswas, S., Chaudhury, A., Tambe, S. S., and Kulkarni, B. D. (2007). Estimation of gross calorific value of coals using artificial neural networks. *Fuel*, *86*(3), 334-344.

Patil-Shinde, V., Kulkarni, T., Kulkarni, R., Chavan, P. D., Sharma, T., Sharma, B. K., Sharma, T., Tambe, S. S., and Kulkarni, B. D. (2014). Artificial intelligence-based modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier. *Industrial & Engineering Chemistry Research*, *53*(49), 18678-18689.

RapidMiner. (2014). http://rapid-i.com/content/view/181/190/lang.en/

Saha, S. (2013). *Studies on Physical Properties of Indian Coals and its Effect on Coal Gasification Kinetics*. Ph.D. Thesis, Indian School of Mines (ISM), Dhanbad, India.

Saha, S., Sahu, G., Datta, S., Chavan, P., Sinha, A. K., Sharma, B. K., and Sharma, T. (2013). Studies on CO_2 gasification reactivity of high ash Indian coal. *Int. J. Emerging Technol. Adv. Eng*, *3*, 29-33.

Saha, S., Sharma, B. K., Chavan, P., Datta, S., Sahu, G., Mall, B. K., and Sharma, T. (2011). Effect of inorganic matter on the reactivity of Indian coals. *Asian Journal of Chemistry*, *23*(10), 4335-4340.

Saha, S., Sharma, B. K., Kumar, S., Sahu, G., Badhe, Y. P., Tambe, S. S., and Kulkarni, B. D. (2007). Density measurements of coal samples by different probe gases and their interrelation. *Fuel*, *86*(10), 1594-1600.

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, *324*(5923), 81-85.

Silbermann, R., Gomez, A., Gates, I., and Mahinpey, N. (2013). Kinetic studies of a novel CO_2gasification method using coal from deep unmineable seams. *Industrial & Engineering Chemistry Research*, *52*(42), 14787-14797.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, *87*(2), 245-251.

Takarada, T., Tamai, Y., and Tomita, A. (1985). Reactivities of 34 coals under steam gasification. *Fuel*, *64*(10), 1438-1442.

Takarada, T., Tamai, Y., and Tomita, A. (1986). Effectiveness of K_2 CO_3 and Ni as catalysts in steam gasification. *Fuel*, *65*(5), 679-683.

Takematsu, T., and Maude, C. (1991). *Coal Gasification for IGCC Power Generation*, IEA Coal Research, International Energy Agency, London.

Van Heek, K. H., and Mühlen, H. J. (1987). Effect of coal and char properties on gasification. *Fuel processing technology*, *15*, 113-133.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Wu, Y., Wu, S., Gu, J., and Gao, J. (2009). Differences in physical properties and CO_2gasification reactivity between coal char and petroleum coke. *Process Safety and Environmental Protection*, *87*(5), 323-330.

Ye, D. P., Agnew, J. B., and Zhang, D. K. (1998). Gasification of a South Australian low-rank coal with carbon dioxide and steam: kinetics and reactivity studies. *Fuel*, *77*(11), 1209-1219.

Zhang, L., Huang, J., Fang, Y., and Wang, Y. (2006). Gasification reactivity and kinetics of typical Chinese anthracite chars with steam and CO_2. *Energy & fuels*, *20*(3), 1201-1210.

Zhao, M. M., Saulov, D. N., Cleary, M. J., and Klimenko, A. Y. (2012). Numerical simulation of coal gasification with CO_2 capture based on two-dimensional fluidized bed model. *International Journal of Chemical Engineering and Applications*, *3*(6), 466-470.

Chapter 5

Use Genetic Programming for Selecting Predictor Variables and Modeling in Process Identification

ABSTRACT

Availability of an accurate and robust dynamic model is essential for implementing the model dependent process control. When first principles based modeling becomes difficult, tedious and/or costly, a dynamic model in the black-box form is obtained (*process identification*) *by using the measured input-output process data. Such a dynamic model frequently contains a number of time delayed inputs and outputs as predictor variables. The determination of the specific predictor variables is usually done via a trial and error approach that requires an extensive computational effort. The computational intelligence* (*CI*) *based data-driven modeling technique, namely, genetic programming* (*GP*), *can search and optimize both the structure and parameters of a linear/nonlinear dynamic process model. It is also capable of choosing those predictor variables that significantly influence the model output. Thus, usage of GP for process identification helps in avoiding the extensive time and efforts involved in the selection of the time delayed input-output variables. This advantageous GP feature has been illustrated in this study by conducting process identification of two chemical engineering systems. The results of the GP-based identification when compared with those obtained using the transfer function based identification clearly indicates the outperformance by the former method.*

5.1 INTRODUCTION

Availability of an accurate, parsimonious, and robust dynamical process model is essential in various tasks such as model based process control, process monitoring, and optimization. The task of constructing mathematical models of dynamical processes from their measured input-output data is known as "process/system identification." It can be viewed as the interface between the real world of applications and the mathematical world of control theory and model abstractions (Ljung, 2010). There are two principal ways for conducting process identification namely, *phenomenological* (first-principles) and *empirical* / *black-box*. In the first approach, the physico-chemical phenomena underlying a chemical process is rigorously described in terms of the mass, energy and momentum balance equations. This type of modeling requires complete details of the governing phenomena, such as, kinetic rate constants, heat and mass transfer coefficients, and other thermodynamic information, which in most cases of practical interest are unavailable. Also, chemical processes very often exhibit complex nonlinear behavior, which makes the development of phenomenological models a tedious, costly and possibly even an impossible task to be completed in a reasonable time span. In such cases, the other approach i.e., empirical/black-box modeling is resorted to for process identification.

A black-box model representing the dynamics of a single input-single output (SISO) nonlinear process can be described using discrete time-variant inputs and outputs as given below:

$$y_{t+1} = f\ (y_t, y_{t-1}, y_{t-2},\ldots, y_{t-m+1}; u_t, u_{t-1}, u_{t-2},\ldots,\ u_{t-n+1}) \qquad (5.1)$$

where y_{t+1} refers to the *one-time-step-ahead* value of the output *y*, subscript *t* refers to the sampling instant, *u* is the manipulated variable (input), *f* denotes the functional relationship between y_{t+1} and the current and past (time delayed/lagged) values of the inputs and outputs, and *m* and *n*, respectively refer to the number of lags in the process output and input. In the above equation, the current and time-delayed inputs and outputs signify the predictor variables for the one-step-ahead-prediction of the output, i.e. y_{t+1}.

The principal advantage of the black-box modeling is that a model can be constructed solely from the measured process data without needing the details of the governing physico-chemical phenomena. In the conventional black-box modeling, the model structure is specified a-priori and the parameters associated with this model are

estimated using an appropriate linear/nonlinear strategy. Since several efficient linear/nonlinear parameter estimation methods are available, the real difficult part in the black-box modeling is the specification of the model structure. For linear systems model specification is easy; however, for nonlinear systems selection of model structure poses significant difficulties since it involves choosing an appropriate nonlinear model structure from numerous competing ones.

The complexities involved in the conventional black box approaches to system identification necessitated exploration of alternative modeling strategies that do not require a-priori specification of the model structure. This requirement is fulfilled, for example, by a number of computational intelligence (CI) based exclusively data-driven nonlinear modeling formalisms such as *artificial neural networks* (ANNs), and *support vector regression* (SVR). An excellent overview of various linear and nonlinear methods for process/system identification is given by Ljung (2010) (also see Garcia and Morari, 1982; Isidori, 1989; Narendra and Parthasarathy, 1990; Tambe et al., 1996).

Apart from ANNs and SVR, the discipline of CI comprises a novel exclusively data-driven modeling formalism, namely *genetic programming* (GP). The uniqueness of the GP methodology is that given an example input-output data set, it is capable of searching and optimizing both, the specific structure (form) and the associated parameters, of an appropriate linear/nonlinear data-fitting function; significantly, unlike ANNs and SVR methods, GP does this without making any assumption regarding the structure and parameters of the data-fitting function (Patil-Shinde et al., 2014). Despite its novelty, GP has not been used widely in process identification to the same extent as ANNs and SVR. The full details of GP (Koza,1990; Poli et al.,2008; Shrinivas et al.,2015) are provided in Chapter 2, Section 2.2.2 of this thesis.

Implementation of GP is a stochastic procedure and, therefore, it contains a strong random element. A typical characteristic of the best solution (data-fitting model) searched and optimized by the GP is that it contains only those predictor variables that yield an optimal data-fitting performance. In the context of process identification, this means that GP selects only those time delayed inputs and outputs as predictors in Eq. (5.1), which significantly influence the one-step ahead output (y_{t+1}). This automatic selection of the important predictor variables by the GP formalism is immensely beneficial in practice since it substantially reduces the computational time and effort required in identifying the specific time-delayed inputs

and outputs in Eq. (5.1). Accordingly, in this study two process identification case studies have been performed to demonstrate the stated ability of GP of simultaneously identifying the important time delayed inputs and outputs and performing the system identification. The two chemical engineering-specific systems chosen in the case studies are: (i) nonlinear height control system for a conical tank, and (ii) concentration control system for a nonlinear adiabatic CSTR. For convenience, the process input-output data for these systems have been obtained using their phenomenological models. In real practice, the process input-output data should be collected by performing open-loop experiments.

There have been studies wherein GP has been employed in system/process identification (see e.g. Kristinsson and Dumont (1992), Iba and Sato (1995), Yadavalli et al. (1999), Nandi et al. (2000), and Sankpal et al. (2001)). It may however be noted that these studies did not explore GP's feature of identifying influential predictor variables. In the present study, the performance of the process model identified by the GP has also been compared with that identified using a transfer function model. A novel feature of the GP formalism is that while searching for an optimal data fitting model, it can identify key predictors and their combinations in the example data. This GP property has been exploited in the present study for automatically choosing those lagged inputs and outputs, which significantly influence the one-time-step ahead output in the dynamic process model.

5.2 RESULTS AND DISCUSSION

5.2.1 Case study I: Nonlinear Height Control System for a Conical Tank

In this case study, a conical tank has been considered (see Figure 5.1) wherein F_{in} and F_{out} are the inlet and outlet flow rates, respectively. The control objective is to maintain the height of the tank, *h*, at a given set point by manipulating the inlet flow rate, F_{in}. The conical tank dynamics are described by following equations (Aravind et al., 2013):

Area of the tank is given by:

$$D = \pi r^2 \tag{5.2}$$

$$tan\,\alpha = \frac{r}{h} = \frac{R}{H} \tag{5.3}$$

According to law of conservation of mass:

inlet flow rate – outlet flow rate = accumulation

$$D\frac{dh}{dt} = F_{\text{in}} - F_{\text{out}} \tag{5.4}$$

$$F_{\text{out}} = L\sqrt{h} \tag{5.5}$$

$$D\frac{dh}{dt} = F_{\text{in}} - L\sqrt{h} \tag{5.6}$$

$$\frac{dh}{dt} = \frac{F_{\text{in}} - L\sqrt{h}}{D} \tag{5.7}$$

Where $\frac{dh}{dt}$ is the *rate of change of height.*

Therefore,

$$D = \frac{\pi R^2 h^2}{H^2} \tag{5.8}$$

$$\frac{dh}{dt} = \frac{(F_{\text{in}} - L\sqrt{h})H^2}{\pi R^2 h^2} \tag{5.9}$$

Here, H is the maximum height of the tank with R as the radius at that height, L is the discharge coefficient, and h is the height of the tank at any instant, t. In this tank, the inlet flow rate, F_{in}, is the manipulated variable and the height of the tank, h, is the controlled variable. Eq. (5.9) was integrated to generate an input-output dataset for identifying the process. A total of 1000 data points were generated while varying F_{in} randomly in the 50 to 200 $\text{cm}^3\text{sec}^{-1}$ range at every time step of one second. The parameter values used in simulating Eq. (5.9) are: probability of F_{in} variation at any instant = 0.22, H = 73 cm, L = 20 $\text{cm}^{2.5}\text{sec}^{-1}$, R = 19.25 cm and initial height of tank = 24 cm. The generated time profiles of F_{in} and h are shown in Figures 5.2 and 5.3, respectively.

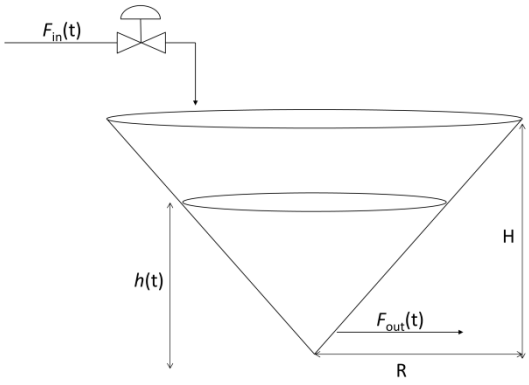


Figure 5.1: Schematic of a height control system for a conical tank.

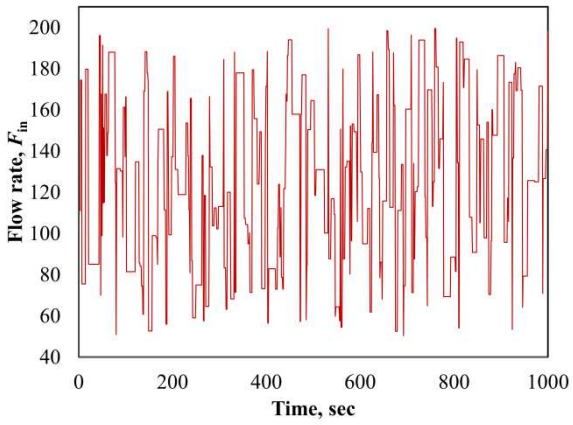


Figure 5.2: Random variations in manipulated variable, F_{in}.

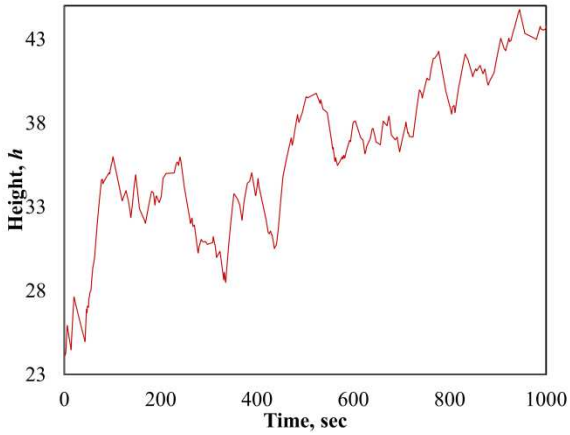


Figure 5.3: Controlled variable (h) response to the random variations in F_{in}.

GP-based identification

The generated data were arranged for the GP-based process identification by considering one current and two time delayed values of both h and F_{in}. Thus, a set of six predictor variables (h_{t}, $h_{\text{t}-1}$, $h_{\text{t}-2}$; $F_{\text{in}_{\text{t}}}$, $F_{\text{in}_{\text{t}-1}}$, $F_{\text{in}_{\text{t}-2}}$) was considered for the prediction of the one-time-step-ahead model output; the data set of 1000 points was split randomly in *training, test* and *validation* sets in 70:20:10 ratio. While the first set was used in generating the GP model, the second and third sets were respectively employed in testing the generalization ability of the model, and its validation. For generating the GP-based model, a software package named *Eureqa Formulize* (Schmidt and Lipson, 2009) was employed. The GP searched following fittest expression:

$$h_{\text{t}+1} = 2\left(h_{\text{t}} + \frac{6.6437}{34.7472 \times F_{\text{in}_{\text{t}}} - 93.5708 \times h_{\text{t}-1} - 134.1376 \times h_{\text{t}-2}} - 0.5 \times h_{\text{t}-1}\right) \qquad (5.10)$$

The values of the *correlation coefficient* (*CC*), *root mean squared error* (*RMSE*) and *mean absolute percentage error* (*MAPE*) for the training, test, and validation set data are listed in Table 5.1. The high (low) and comparable values of *CC* (*RMSE, MAPE*) for the training, test and validation sets indicate that the GP-based model (5.10) possesses excellent one-time-step-ahead prediction accuracy and generalization ability.

Table 5.1: Prediction accuracy and generalization performance of GP-based model (5.10) for conical tank height control system

	Training set	Test set	Validation set
CC	0.9993	0.9984	0.9937
RMSE	0.1269	0.0799	0.0638
MAPE	0.0015	0.0007	0.0005

Figure 5.4 shows the parity plot of the desired (target) versus GP model-predicted values of $h_{\text{t}+1}$. From the GP-based optimal data-fitting model (5.10), it is noticed that the model consists of four predictor variables (h_{t}, $h_{\text{t}-1}$, $h_{\text{t}-2}$and $F_{\text{in}_{\text{t}}}$). It can thus be seen that although the data supplied to the GP algorithm contained six

predictor variables (h_t, h_{t-1}, h_{t-2}; F_{in_t}, $F_{in_{t-1}}$, $F_{in_{t-2}}$) the method searched and optimized a model with only four predictor variables (h_t, h_{t-1}, h_{t-2}; F_{in_t}).

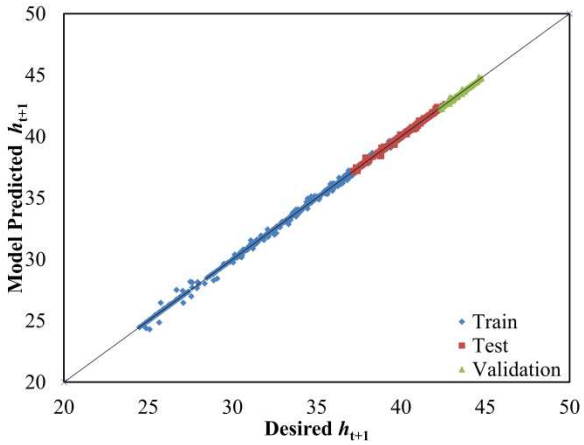


Figure 5.4: Desired versus GP-model predicted h_{t+1} values pertaining to the training, test, validation set data.

Transfer function based identification

Here the control system consists of an input (F_{in}), and output (h). The output is related to the input via a transfer function. The simulated data generated for the conical tank height control system described earlier appropriately arranged to conduct transfer Function based process identification by using values of both h and F_{in}. The same proportion of 70:20:10 was used to partition the dataset of 1000 points into training, test and validation sets. For developing the transfer function based model *Matlab Sysid* (*System identification*) toolbox was used. The continuous-time identified transfer function model is given as:

$$T_1(S) = \frac{0.005835\,S^3+0.002684\,S^2+0.0003145\,S+6.571\times 10^{-8}}{S^3+0.085\,S^2+0.001164\,S+4.852\times 10^{-8}} \tag{5.11}$$

In Laplace transform, the input is represented by F_{in}(S) and output is represented by h(S).

$$T_1(S) = \frac{h(S)}{F_{in}(S)} \tag{5.12}$$

The transfer function of the system multiplied by the input function produces the output function of the system. The values of *CC*, *RMSE* and *MAPE* pertaining to the transfer function-based model (5.11) are given in Table 5.2. Figure 5.5 shows the parity plot of the desired (target) versus transfer function model-predicted values of h_{t+1}. A comparison of these values with those corresponding to the predictions of the GP-based model (5.10) reveals that the CI-based model possesses superior prediction accuracy and generalization capability.

Table 5.2: Prediction accuracy and generalization performance of transfer function model (5.11) for conical tank height control system

	Training set	Test set	Validation set
CC	0.9828	0.9624	0.8760
RMSE	0.6549	0.4116	0.5471
MAPE	0.0161	0.0088	0.0098

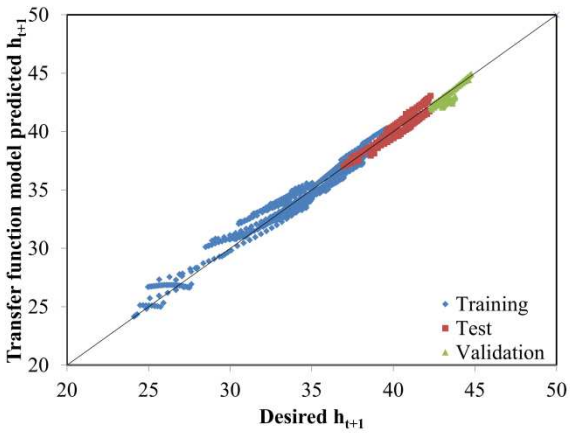


Figure 5.5: Desired versus transfer-function model predicted h_{t+1} values pertaining to the training, test, validation set data.

5.2.2 Case study II: Adiabatic Nonlinear CSTR Concentration Control System

This case study considers a continuously stirred tank reactor (CSTR) (Figure 5.6) in which a second order irreversible exothermic chemical reaction ($A \rightarrow B$) takes place. The control objective is to maintain the concentration of A, i.e. C_A, at a given set-point by manipulating the inlet flow rate, F. The adiabatic CSTR dynamics are described by the following ordinary differential equations (Luyben, 1996):

Reactor component A continuity:

$$\frac{dC_A}{dt} = \frac{F}{V}\left(C_{A_{in}} - C_A\right) - k\,C_A{}^2 \tag{5.13}$$

$$k = k_0\, e^{\frac{-Ea}{RT}} \tag{5.14}$$

$$\frac{dC_A}{dt} = \frac{F}{V}\left(C_{A_{in}} - C_A\right) - k_0\, e^{\frac{-Ea}{RT}} C_A{}^2 \tag{5.15}$$

where $\frac{dC_A}{dt}$ is the *rate of change of concentration of A*

Reactor energy balance equation is given as:

$$\frac{dT}{dt} = \frac{F}{V}(T_{in} - T) - \frac{H_R\, k\, C_A{}^2}{\rho C_p} \tag{5.16}$$

$$C_p = 4.184 - 0.002\,(T - 273) \tag{5.17}$$

$$\frac{dT}{dt} = \frac{F}{V}(T_{in} - T) - \frac{H_R\, k_0\, e^{\frac{-Ea}{RT}}\, C_A{}^2}{\rho\,(4.184 - 0.002\,(T - 273))} \tag{5.18}$$

where $\frac{dT}{dt}$ is rate of change of the outlet temperature, T, $C_{A_{in}}$ is the inlet concentration of species A, V is the volume of the CSTR, k_0 denotes the reaction rate constant, R is the gas constant, H_R represents the heat of reaction, T_{in} is fluid inlet temperature, ρ is the density of liquid, C_p is the fluid specific heat capacity, T is fluid outlet temperature, and E_a denotes the activation energy of the reaction.

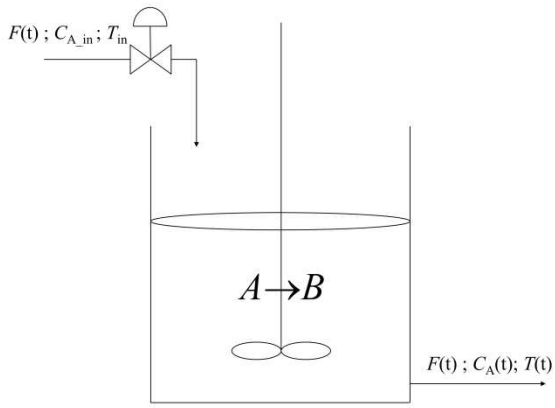


Figure 5.6: Schematic of an adiabatic CSTR control system.

In this CSTR process, the outlet concentration of *A*, i.e. C_A, and temperature, *T*, both vary with time, and *F* and C_A are the manipulated and controlled variables, respectively. For the purpose of illustration, the GP and transfer function based dynamic models have been built only for C_A. As in Case study I (section 5.2.1), process data were generated by integrating the phenomenological model described by Equations (5.15) to (5.18). Specifically, *F* was perturbed randomly at every time step of 0.1 min in the range, 10 to 150 lit min^{-1}, to generate an input-output dataset consisting of 1000 points. The parameter values used in simulating (5.15) to (5.18) were: probability of *F* variation at any instant = 0.2, $C_{A_{in}}$ = 6 mol lit^{-1}, *V*=100 lit, k_0 = 0.15 lit $mol^{-1}min^{-1}$, *R* = 8.314 $Jmol^{-1}K^{-1}$, H_R = –590 J mol^{-1}, T_{in}= 288 K, ρ = 1.050 kg lit^{-1}, E_a = 5000 J mol^{-1}, initial concentration of *A* = 2 mol lit^{-1}, and initial outlet temperature = 295 K. Figures 5.7, 5.8 and 5.9, respectively show the generated time profiles of *F*, C_A , and *T*.

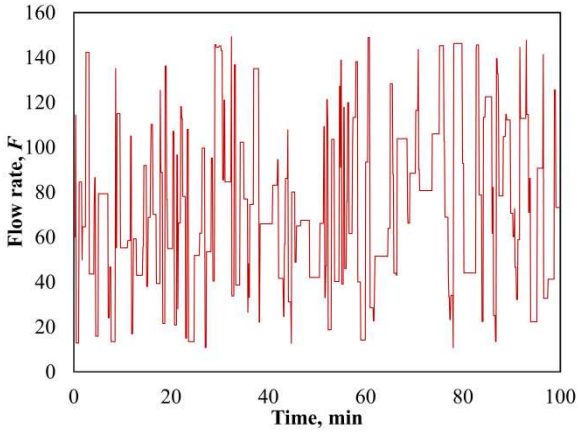


Figure 5.7: Random variations in manipulated variable, *F*.

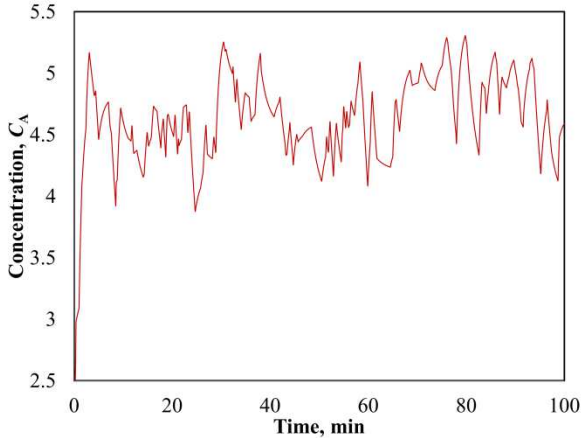


Figure 5.8: Controlled variable (C_A) response for random variations in *F*.

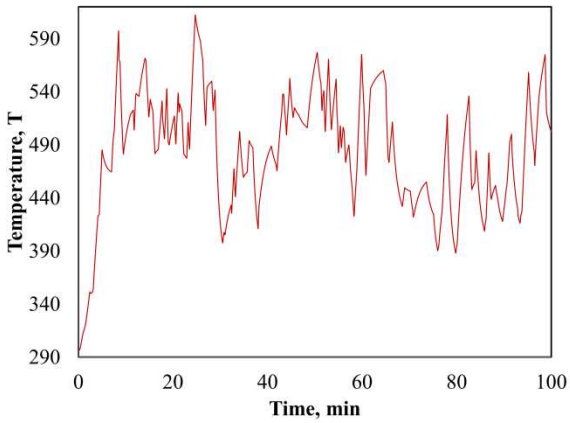


Figure 5.9: Outlet temperature (*T*) response for random variations in *F*.

GP- based identification

For developing the GP-based discrete time model, the current and two time-delayed values of both, *C*A and *F*, were utilized. Thus, the six predictor variable (C_{A_t} , $C_{A_{t-1}}$, $C_{A_{t-2}}$, F_t, F_{t-1}, F_{t-2}) set was employed in developing the GP-based model predicting the one-step-ahead process output, $C_{A_{t+1}}$. For developing the discrete time GP-based model the data set consisting of 1000 points was split randomly in *training*, *test* and *validation* sets in 70:20:10 ratio, respectively. The GP searched fittest expression obtained using the *Eureqa Formulize* software package is given as:

$$C_{A_{t+1}}=1.9125\left(C_{A_t}+\frac{\left(0.2213-0.0644\times C_{A_{t-2}}\right)}{1.9125\times F_t}-0.4771\times C_{A_{t-1}}\right) \quad (5.19)$$

Likewise Case Study I, in this case also it is seen that among the six predictor variables that were provided to it, the GP algorithm has used only four of them (C_{A_t},$C_{A_{t-1}}$,$C_{A_{t-2}}$, F_t) in obtaining model (5.19). The values of the *CC*, *RMSE* and *MAPE* corresponding to the $C_{A_{t+1}}$ predictions made by (5.19) for the training, test, validation set data are listed in Table 5.3. Figure 5.10 shows the parity plot of the desired versus GP-model predicted values of $C_{A_{t+1}}$. From this figure and the high (low) and comparable values of *CC* (*RMSE* and *MAPE*) for the training, test and validation set data it is clear that the GP based identification model (5.19) possesses an excellent one-time- step-ahead prediction accuracy and generalization ability.

Table 5.3: Prediction accuracy and generalization performance of GP-based model (5.19) for CSTR control system

	Training set	Test set	Validation set
CC	0.9937	0.9928	0.9939
RMSE	0.0362	0.0259	0.0308
MAPE	0.0030	0.0016	0.0024

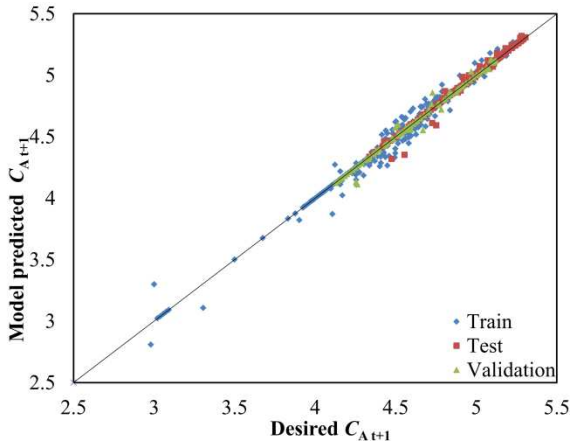


Figure 5.10: Desired versus GP-model predicted $C_{A_{t+1}}$ values pertaining to the training, test, validation set data.

Transfer function based identification

Here, the simulated CSTR generated data were arranged appropriately for the development of the transfer function based process identification, wherein both input (F_{in}), and output (C_A) signals were utilized. This data set of 1000 points was split randomly in training, test and validation sets in 70:20:10 ratio. The continuous-time transfer function based model obtained using *Matlab sysid* software is given below:

$$T_2(S) = \frac{0.00152S^3 + 0.125\,S^2 + 0.007943\,S + 0.0001468}{S^3 - 1.226\,S^2 + 0.49\,S + 0.00231} \tag{5.20}$$

In this Laplace transform based model, the input is represented by F_{in} (S) and output is represented by

$$T_2(S) = \frac{C_A(S)}{F(S)} \tag{5.21}$$

The transfer function defined in this equation when multiplied by the input function produces the output function of the CSTR system. The *CC*, *RMSE* and *MAPE* magnitudes pertaining to the $C_{A_{t+1}}$ predictions made using (5.20) are listed in Table 5.4. Figure 5.11 shows the parity plot of the desired (target) versus transfer function model-predicted values of $C_{A_{t+1}}$. A comparison of the prediction and generalization performance of the GP and transfer function based CSTR models indicates that the former model has outperformed the latter model.

Table 5.4: Prediction accuracies and generalization performance of transfer function model (5.20) for CSTR control system

	Training set	Test set	Validation set
CC	0.9888	0.9815	0.9920
RMSE	0.0526	0.0585	0.0449
MAPE	0.0090	0.0089	0.0074

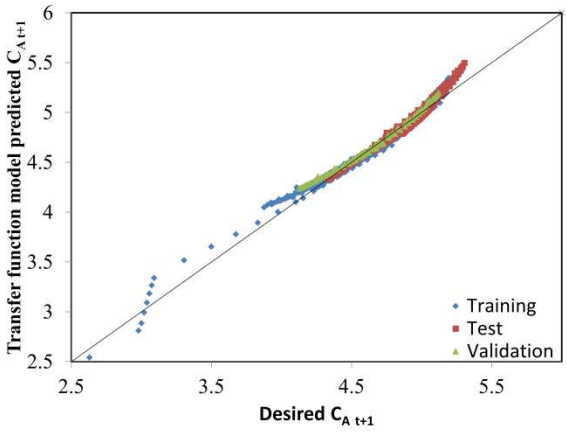


Figure 5.11: Desired versus transfer-function model predicted $C_{A_{t+1}}$ values pertaining to the training, test, validation set data.

5.2.3 Sensitivity Analysis of Predictor Variables

In this study, sensitivity analysis (also termed "importance" analysis) was also performed for the data used in the development of two GP-based models. It was conducted using the *IBM-SPSS package* (2011) to ascertain the extent of influence exerted by each predictor variable on the one-time-step-ahead value of the output (controlled) variable (see Chapter 2, section 2.6 for a detailed discussion of sensitivity analysis).

The importance analysis for the conical tank system in case study-I was conducted using the entire simulated data set of 1000 points. Figure 5.12 exhibits the importance and normalized importance chart indicating the extent of influence exerted by each predictor variable on the one-time-step-ahead-value of the controlled variable (h_{t+1}). In this figure, it is seen that the four predictor variables, namely, h_t, h_{t-1}, h_{t-2}, and F_{in_t} exert a significant influence on the h_{t+1} magnitude. Notably, the same four variables appear in the optimal model (5.10) thus demonstrating the ability of the GP formalism to select the most influential predictor variables during searching and optimizing a data-fitting model.

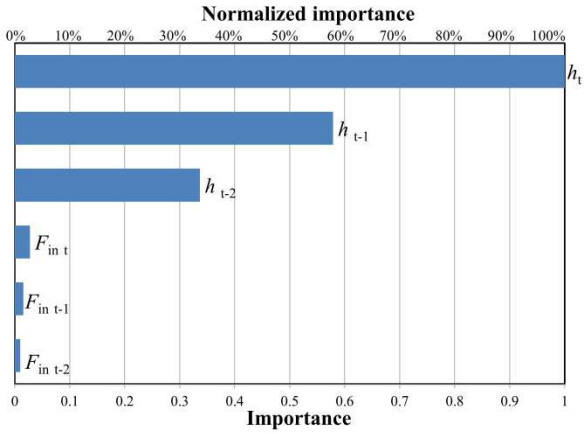


Figure 5.12: Normalized importance of six predictor variables on process output, h_{t+1}.

Figure 5.13 exhibits the importance and normalized importance charts pertaining to the adiabatic nonlinear CSTR concentration control system. It indicates that among the six predictor variables (C_{A_t}, $C_{A_{t-1}}$, $C_{A_{t-2}}$, F_t, F_{t-1}, F_{t-2}) those four that influence the one-time-step-ahead value of the control variable (C_A) most strongly are $C_{A_{t-1}}$, F_t C_{A_t}, and $C_{A_{t-2}}$. It is noteworthy that the same four predictor variables have been utilized by the GP formalism in obtaining the optimal model (5.19).

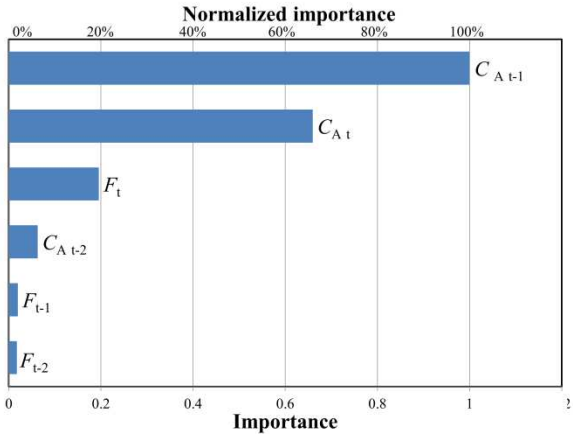


Figure 5.13: Normalized importance of six predictor variables on process output, $C_{A_{t+1}}$.

5.3 CONCLUSION

In the conventional process/system identification, it becomes tedious and computationally intensive to select the specific predictor variables that strongly correlate with the single- or multi-time-step-ahead values of the output. In this paper, a GP-based strategy has been suggested for simultaneously identifying the important predictor variables as also searching and optimizing an optimal data fitting function

and its parameters. The said strategy has been illustrated by conducting two process identification case studies wherein the GP formalism has been shown to (a) identify the influential time-delayed inputs and outputs, and (b) simultaneously perform system identification using these influential predictors. The two chemical engineering systems chosen in the case studies are: (i) nonlinear height control system for a conical tank, and (ii) adiabatic nonlinear CSTR concentration control system. From the GP-based models obtained in these case studies, it is noticed that although the data supplied to the GP algorithm contained six predictor variables, it searched and optimized models with only four predictor variables. Noticeably, these predictors were identified by the sensitivity analysis to be having most influence on the model output. Both the GP based process identification models (5.10) and (5.19) predict the one-time-step-ahead values of the output variables (h_{t+1} and $C_{A_{t+1}}$) with an excellent prediction and generalization performance as indicated by the high (low) magnitudes of the correlation coefficient (root mean squared error and mean absolute percentage error) pertaining to the training, test, and validation set data. It is also observed that the GP–based models possess better prediction accuracy and generalization capability than the continuous-time transfer function models. Moreover, the GP-based models are less complex than the transfer function models. This feature is important since usually less complex (i.e., parsimonious) models possess better at generalization than their more complex counterparts. To summarize, the GP-based system identification strategy—being computationally economical and much less tedious—has the potential to become an effective alternative to the conventionally used linear/nonlinear identification strategies. Having identified a process using the GP strategy the corresponding model can be gainfully utilized to implement the model predictive control (MPC) strategy.

NOMENCLATURE

f	functional relationship between y_{t+1} and the current and past (time delayed/ lagged) values of the inputs and outputs
m , n	number of lags in the process output and input, respectively
t	sampling instant
u	manipulated variable (input)
y_{t+1}	one-time-step-ahead value of the output y

REFERENCES

Aravind, P., Valluvan, M., and Ranganathan, S. (2013). Modelling and Simulation of Non Linear Tank. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, *2*(2), 842-849.

Garcia, C. E., and Morari, M. (1982). Internal model control. A unifying review and some new results. *Industrial & Engineering Chemistry Process Design and Development*, *21*(2), 308-323.

Iba, H., and Sato, T. (1995). A numerical approach to genetic programming for system identification. *Evolutionary computation*, *3*(4), 417-452.

IBM SPSS Neural Networks 20 Manual, Chicago: IBM, 2011.

Isidori, A. (1989). *Nonlinear Control Systems*. 2nd edition, Springer Science & Business Media, Berlin.

Koza, J. R. (1990). Genetically breeding populations of computer programs to solve problems in artificial intelligence. In *Tools for Artificial Intelligence, Proceedings of the 2nd International IEEE Conference*; pp. 819-827, IEEE.

Kristinsson, K., and Dumont, G. A. (1992). System identification and control using genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(5), 1033-1046.

Ljung, L. (2010). Perspectives on system identification. *Annual Reviews in Control*, *34*(1), 1-12.

Luyben, W. L. (1996). *Process Modeling, Simulation and Control for Chemical Engineers*. 2nd ed., McGraw-Hill Higher Education, New York, pp. 46-49.

Nandi, S., Rahman, I., Tambe, S. S., Sonolikar, R. L., and Kulkarni, B. D. (2000). Process identification using genetic programming: A case study involving fluidized catalytic cracking (FCC) unit. In *Petroleum Refining and Petrochemical-based Industries in Eastern India*; Saha, R. K., Ray, S., Maity, B. R., Bhattacharya, D., Ganguly, S., and Chakraborty, S. L., (Eds.), Allied Publishers Ltd.: , Mumbai, pp 195-201.

Narendra, K. S., and Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on neural networks*, *1*(1), 4-27.

Patil-Shinde, V., Kulkarni, T., Kulkarni, R., Chavan, P. D., Sharma, T., Sharma, B. K., Tambe, S. S., and Kulkarni, B. D. (2014). Artificial intelligence-based modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier. *Industrial & Engineering Chemistry Research*, *53*(49), 18678-18689.

Poli, R., Langdon, W. B., McPhee, N. F., and Koza, J. R. (2008). *A field guide to genetic programming*. Available at: http://lulu.com, http://www.gp-fieldguide.org.uk. Accessed: 23 May 2015.

Sankpal, N. V., Cheema, J. J. S., Tambe, S. S., and Kulkarni, B. D. (2001). An artificial intelligence tool for bioprocess monitoring: Application to continuous production of gluconic acid by immobilized Aspergillus niger. *Biotechnology letters*, *23*(11), 911-916.

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, *324*(5923), 81-85.

Shrinivas, K., Kulkarni, R. P., Shaikh, S., Ghorpade, R. V., Vyas, R., Tambe, S. S., Ponrathnam, S., and Kulkarni, B. D. (2015). Prediction of reactivity ratios in free radical copolymerization from monomer resonance–polarity (Q–e) Parameters: Genetic programming-based models. *International Journal of Chemical Reactor Engineering*, *14*(1), 361-372.

Tambe, S. S., Kulkarni, B. D., and Deshpande, P. B. (1996). *Elements of Artificial Neural Networks with Selected Applications in Chemical Engineering, and Chemical & Biological Sciences*. Simulation & Advanced Controls Inc., Louisville, K.Y.

Yadavalli, V. K., Tambe, S. S., Dahule, R. K., and Kulkarni, B. D. (1999). Consider genetic programming for process identification. *Hydrocarbon processing*, *78*(7), 89-97.

Chapter 6

Prediction of °API Values of Crude Oils by Use of Saturates/Aromatics/Resins/Asphaltenes Analysis: Computational- Intelligence-Based Models

ABSTRACT

The °API value is an important physicochemical characteristic of crude oils often used in determining their properties and quality. There exist models—predominantly linear ones—for predicting the °API magnitude from the molecular composition of a crude oil. This approach is tedious and time-consuming since it requires quantitative determination of numerous crude-oil components. Usually, the hydrocarbons present in a crude oil are grouped according to their molecular average structures into saturates, aromatics, resins, and asphaltenes (*SARA*). *An °API-value prediction model dependent on these four fractions is relatively easier to develop, although this approach has been rarely used. A rigorous scrutiny of the relevant data suggests that some of the dependencies between the individual SARA fractions and the corresponding °API-value could be nonlinear. Accordingly, in this study, SARA-fraction based nonlinear models have been developed for the prediction of °API values using three Computational Intelligence* (*CI*) *formalisms: genetic programming* (*GP*), *artificial neural networks* (*ANNs*), *and support vector regression* (*SVR*). *The SARA analyses and °API values of 403 crude-oil samples covering wide ranges have been used in developing these models. A comparison of the CI-based models with an existing linear model indicates that all the former class of models possess a significantly better °API-value prediction and generalization performance than those exhibited by the linear model. Also, the SVR-based model has been found to be the most accurate °API-value predictor. Because of their better prediction accuracy, CI-based models can be gainfully used to predict °API values of crude oils.*

6.1 INTRODUCTION

Crude oil is a complex mixture of hydrocarbons that also contains some "hetero" atoms such as oxygen, nitrogen, and sulfur. The knowledge of a crude oil's type and quality is essential because these characteristics determine its market value and ease of refining. The stated crude-oil attributes can be ascertained by the use of various properties, such as, standard specific gravity, pour point, and sulfur and metal contents. Often, the density of petroleum oils is expressed in terms of the ^{o}API value; this metric was devised by American Petroleum Institute (API) and National Bureau of Standards. It is known to be a crucial property because it directly affects the production and price of a crude oil. The ^{o}API value is used extensively in the classification and determining properties, such as the viscosity and compressibility factor, of petroleum oils, and also in setting the operating parameters of distillation columns in a refinery. It is a measure of the crude oil's "lightness" or "heaviness" or the standard specific gravity that compares the specific gravity of the oil to that of water; ^{o}API value is computed as

$$°\text{API} = \frac{141.5}{\text{specific gravity } (^{60°\text{F}}/_{60°\text{F}})} - 131.5 \tag{6.1}$$

In a commonly used classification scheme, crude oils are categorized on the basis of their °API magnitudes as follows (Strubinger et al. 2012): *extra heavy* (°API<10); *heavy* (10<°API<22.3); *medium* (22.3<°API<31.1), and *light* (°API>31.1). The °API value is measured using a standard hydrometer according to the American Society for Testing and Materials (ASTM) methods D287 (*ASTM D287-12* 2012) and D1298 (*ASTM D1298-12b* 2012). It varies strongly with temperature because of the significant volume expansion of the oil upon heating. Generally, the less processing a crude oil must undergo, it is regarded as more valuable. Considering the chemistry of oil refining, the denser the crude oil, the higher is its carbon/hydrogen ratio, and more intense and costly refinery processing is required for producing specific volumes of gasoline and distillate fuels. Thus, °API value of a crude oil significantly influences the quantum of investment and energy consumption in a refinery, which form the two largest components of the total refining cost (ICCT 2014). Commonly, the higher the °API magnitude, the lighter is the crude oil and higher is its demand; therefore, an accurate evaluation of °API becomes very important (Lammoglia and Filho, 2011).

In a widely used classification scheme, crude oils are categorized on the basis of their composition and types of hydrocarbons present in them (Albahri et al., 2003). Specifically, the hydrocarbons are classified into four groups on the basis of their molecular average structures, polarizability, and polarity: *Saturates* (alkanes and cycloparaffins); A*romatics* (hydrocarbons; mono, di, and polyaromatic); *Resins* (polar molecules with the heteroatoms nitrogen, oxygen, sulfur); and A*sphaltenes* (similar to resins but possessing higher molecular weight and a polyaromatic core). This method of classification is known as SARA analysis. The stated crude oil components are separated by use of the SARA-fractionation method (Speight and Ozum, 2002). Various techniques, such as the clay/gel-adsorption chromatography (basis of *ASTM D2007-93*), thin-layer chromatography (TLC) (Vela et al., 1995), and high-performance liquid chromatography (HPLC) (Suatoni and Swab, 1975; Chaffin et al., 1996) are used to perform the SARA analysis. Among these, HPLC has been demonstrated to be a very efficient alternative to the *ASTM 2007* (1993) procedure for SARA fractionation because it is achieved rapidly.

The ASTM tests for measuring the °API value need expensive equipment and are time-consuming to perform; thus, these are difficult to use in the on-line monitoring of the crude-oil quality (Muhammad and de Vasconcellos Azeredo, 2014). To overcome this difficulty, mathematical models that predict the °API value from the measured values of other oil-specific attributes, have been proposed. Accordingly, it has been shown that the quality of the crude petroleum and its derivatives, as assessed in terms of the °API values, could be predicted directly from the molecular composition of crude oils. A number of studies have followed this strategy to propose models that use data from various spectroscopic methods, such as Fourier transform infrared-attenuated total reflectance, absorption and synchronous ultraviolet fluorescence (Abbas et al., 2012), nuclear magnetic resonance (Muhammad and de Vasconcellos Azeredo, 2014), infrared (Pasquini and Bueno, 2007), and attenuated total reflection Fourier transform infrared spectroscopy (Filgueiras et al. 2014), for the prediction of °API magnitudes. Another approach to developing a model predicting the °API value is dependent on the use of SARA fractions as predictors. The principal advantage of this methodology is that because of the limited number of inputs (i.e., four), a SARA-fraction-based model can be developed relatively easily, and speedily compared to a model that takes into account a large number of hydrocarbons present

in a crude oil. Although attractive, this approach has been rarely used in developing the °API-value prediction models. An extensive literature search has revealed that a SARA-fraction-based model has been developed by Fan and Buckley (2002) to predict the °API magnitudes. The principal objective of their study, however, was to examine three methods (namely, gravity-driven chromatographic separation, TLC, and HPLC) used for separating crude oils and other hydrocarbon materials into SARA fractions. Fan and Buckley (2002) proposed the °API prediction model for differentiating the SARA data obtained by the use of ASTM and HPLC methods from those provided by the TLC-flame-ionization-detector (TLC-FID) method. Their model, possessing a linear form and valid over the °API-value range of 15–40, is given as

$$°\text{API} = 74.5 - 0.306\,\text{S} - 0.385\,\text{A} - 1.08\,\text{R} - 0.763\,\text{A}_\text{p} \qquad (6.2)$$

where *S, A, R,* and A_p, respectively, represent the weight percentages (wt%) of saturates, aromatics, resins, and asphaltenes. The magnitude of the *correlation coefficient* (*CC*) between the experimental, and Eq. (6.2) - predicted °API values for the HPLC-analyzed 87 crude-oil samples were found to be 0.825 (Fan and Buckley 2002).

To verify the true nature of the dependencies (whether linear or nonlinear) between the SARA constituents and the corresponding °API values, a large data set consisting of SARA analyses of 565 crude-oil samples was compiled from a number of publications, including a database. Different analytical methods, such as TLC-FID, ASTM, HPLC, gas chromatography-mass spectrometry (GC-MS), and open-column chromatography, have been used in conducting the SARA analyses. The compiled data contain a number of samples for which the wt% values of the individual SARA constituents do not add up exactly to 100. Thus, the data set was screened to select 403 samples for which the wt% magnitudes of SARA constituents add up to $100 \pm 2\%$. Here, the value of 2% was chosen to allow for small experimental errors in the SARA analyses. The screened data set and the respective data sources are tabulated in Appendix 6.A. This set contains data pertaining to the light (115 samples), medium (127 samples), heavy (127 samples), and very-heavy (34 samples)

crude oils. These data were used to generate the cross-plots (Figure 6.1), wherein °API values were plotted against the individual constituents of the SARA analyses. It is seen in Figure 6.1(*a*) that there exists an approximately linear relation between the °API values and the wt% of saturates. However, a similar inference of linear dependence cannot be drawn unambiguously because of the large scatter seen in Figures 6.1(*b*), 6.1(*c*), and 6.1(*d*), pertaining to the relationships existing between °API values and the wt% of aromatics, resins, and asphaltenes, respectively. Thus, it is quite plausible that these relationships, although not obvious to the naked eye, in reality are nonlinear. Hence, it is worthwhile to explore whether a nonlinear model would better capture the relationships between the °API values and the SARA components of crude oils and thereby make more accurate predictions than the linear model defined in Eq. (6.2). With this objective, three CI-based modeling formalisms—GP, ANNs, and SVR—have been used in this study for developing the SARA-fraction-based models for prediction of °API values of crude oils. In addition, the prediction and generalization performance of the CI-based models have been compared with those of the linear model (Eq. 6.2). The same large-sized data set used in generating the cross plots in Figure 6.1 was used in the development of the CI-based models for simulating the stated linear model. Moreover, the coefficients of the linear model of Fan and Buckley (2002) were freshly determined by use of the large-sized data set to further test whether a linear model is indeed the most-suitable predictor of °API. The results of all these modeling studies clearly indicate that the CI-based models possess significantly higher °API-value-prediction accuracy and -generalization capability than the original and the freshly fitted linear models.

This chapter is structured as follows. Section 6.2, titled “Data” provides details of the data used in the CI-based modeling of °API values. The next section 6.3 termed “Results and Discussion,” describes development of the three CI-based models for the prediction of °API values. This section also provides results of a comparison of the °API-value-prediction performance of the three CI-based models with that of the linear model by Fan and Buckley (2002) and its refitted version. Finally, “Conclusion” (section 6.4) summarizes the major findings of this study.

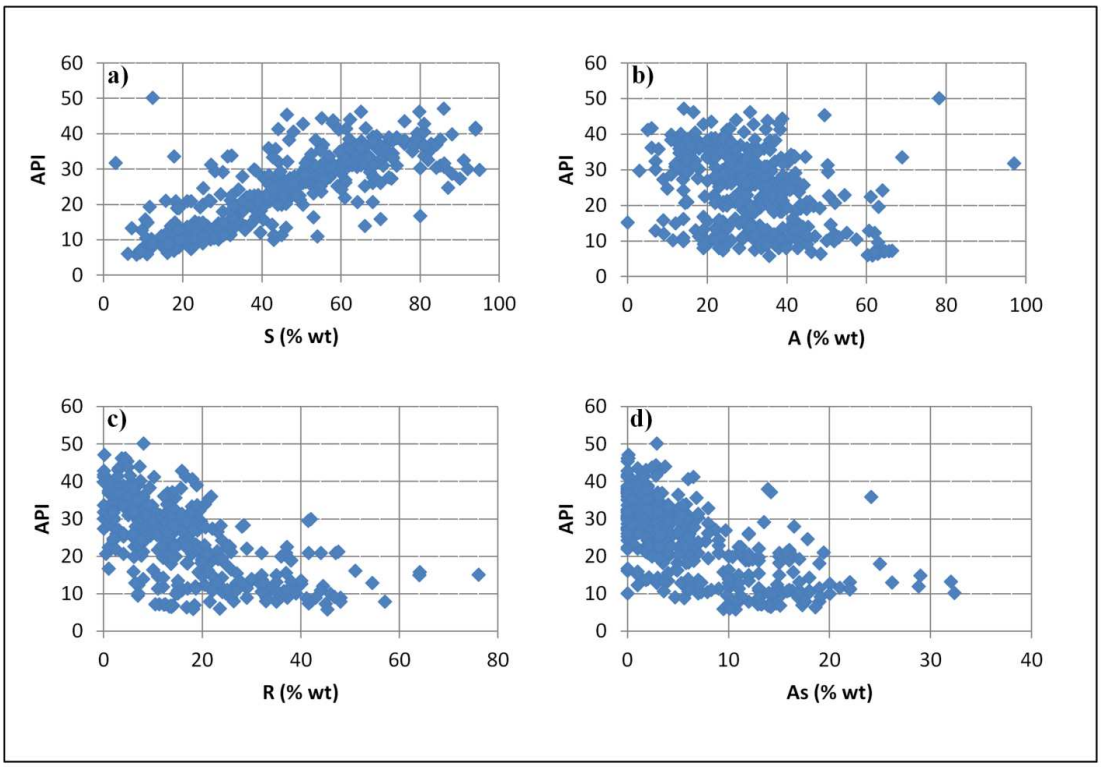


Figure 6.1: Cross-plots of °API values vs. percentages of individual SARA constituents.

6.2 DATA

The GP-, MLP-, and SVR-based models predicting the °API value of a crude oil were developed by use of a data set consisting of 403 input/output patterns (Appendix 6.A). Each pattern contains four predictor variables (model inputs) with wt% values of saturates (*S*), aromatics (*A*), resins (*R*), and asphaltenes (A_p), and the corresponding magnitude of °API (desired model output). For constructing the models, the inputs and °API values were normalized by use of the z-score method. The prediction accuracy and generalization capability of a model were assessed on the basis of the *CC*, *RMSE*, and mean-absolute-percent error (*MAPE*) values, which were computed by use of the experimental and the corresponding model-predicted °API values. The *MAPE* was evaluated according to the following expression:

$$MAPE_j(\%) = \frac{1}{N_p}\sum_{i=1}^{N_p}\left|\frac{y_i - \hat{y}_{i,j}}{y_i}\right| \times 100\ ; \quad j = 1, 2, \ldots, N_{pp} \tag{6.3}$$

where j denotes the index of candidate solution; $MAPE_j$ refers to the *MAPE* of the j^{th}-candidate solution in the population; y_i is the desired (target) output value corresponding to the i^{th}-input data pattern in the training/test data set; $\hat{y}_{i,j}$ is the model-predicted °API value when the i^{th}-input pattern is used to compute the output of the j^{th}-candidate solution, and N_p is the number of input/output patterns in the training/test set. All models were trained and their generalization performance was tested by use of the fivefold cross-validation scheme. In this method, the available example set was portioned into five subsets. Multiple training runs were conducted, and each time a different subset was used as the test set; the remaining four subsets were used as the training set. Finally, the statistical quantities—namely, *CC*, *RMSE*, and *MAPE*—corresponding to training and test sets obtained in multiple runs are averaged. The optimal model is selected on the basis of high (low) and comparable averaged values of *CC* (*RMSE*, *MAPE*) for both training and test sets. In the following subsections, details of the construction of the GP-, MLP-, and SVR-based models and a comparison of their °API-value-prediction and -generalization performance are presented.

6.3 RESULTS AND DISCUSSION

6.3.1 GP-Based Modeling

The GP-based °API-value prediction model was developed by use of the *Eureqa Formulize* software (Edwards, 2009; Schmidt and Lipson, 2009). This software has been optimized to search parsimonious models (i.e., with low complexity), which are expected to possess the much-desired generalization ability. The Eureqa Formulize software uses the plain “single train/test split” procedure, in which training and test sets of fixed sizes are, respectively, used in the construction and assessment of the generalization capability of a candidate expression.

The detailed procedure for GP (Koza, 1992; Poli et al., 2008) implementation has been explained in Chapter 2 (section 2.2.2). For obtaining a parsimonious °API-value prediction model possessing good prediction accuracy and generalization capability, several GP runs were conducted, each time using a different operator subset from the large set provided by the *Eureqa Formulize* package. The best solution in each run was documented. From several such solutions, those satisfying

the following criteria were screened to select an overall optimal model (Goel et al., 2015):

- Small and comparable magnitudes of *RMSE* and high and comparable magnitudes of *CC*s for both training and test set.
- Must contain all four input variables: *S*, *A*, *R*, and A_p.
- Should possess low complexity (i.e., should contain a small number of terms, which ensures better generalization by the model).

In GP simulations, the *RMSE* was evaluated as follows:

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{N_p}(y_i - \hat{y}_{i,j})^2}{N_p}} \quad ; j = 1, 2, \ldots, N_{\text{pp}} \tag{6.4}$$

where, $RMSE_j$ refers to the *RMSE* pertaining to the j^{th}-candidate solution. The optimal GP-based °API-value prediction model selected on the basis of previously described criteria is as follows:

$$°\text{API} = 10.204 \times \left(\hat{S} + \frac{(\hat{R}\times\hat{A}_p) + (\hat{A}\times\hat{S}^2) - \hat{R} + (\hat{A}\times\hat{A}_p\times\hat{S}^2)}{9.286 + (5.482\times\hat{A}\times\hat{A}_p) + (\hat{S}\times\hat{R}\times\hat{A}_p) + (5.482\times\hat{A}\times\hat{R}\times\hat{A}_p)}\right) + 24.61 \tag{6.5}$$

where, $\hat{S} = \frac{S-46.456}{21.194}$, $\hat{A} = \frac{A-30.813}{12.986}$, $\hat{R} = \frac{R-16.316}{12.714}$, and $\hat{A}_p = \frac{A_p-6.384}{6.365}$

The *CC*, *RMSE*, and *MAPE* magnitudes with respect to °API-value predictions made by Eq. (6.5) for both training and test sets are listed in Table 6.1.

6.3.2 MLP-Neural-Network-Based Modeling

A detailed description of the MLP training procedure and related issues is provided by, for example, Zurada (1992), Bishop (1995), and Tambe et al. (1996) (also see Chapter 2 (section 2.2.1)). The MLP-based optimal °API-value prediction model was trained by use of the *error back propagation* (EBP) algorithm from the *RapidMiner* data-mining suite (Mierswa et al. 2006; RapidMiner 2007). The model consists of four input nodes (N = 4), and a single output-layer node in its architecture; the four input-layer nodes represent weight percentages of the *S*, *A*, *R*, and A_p components, respectively, in the crude oils, and a single output node represents the

corresponding °API value. To obtain an optimal MLP model, its structural and training algorithm-specific parameters—such as the number of hidden layers, number of hidden nodes in each layer, learning rate (η), and momentum coefficient (μ_{ebp})—were systematically varied. The criterion for choosing an optimal model was minimum *RMSE* magnitude for the test set. The magnitudes of the MLP architectural and EBP specific parameters (η, μ_{ebp}) that led to an optimal MLP model were number of hidden layers = two; number of nodes in Hidden Layers 1 and 2: five each; η = 0.5; and μ_{ebp} = 0.01. The prediction accuracy and the generalization performance of the optimal MLP model have been evaluated in terms of *CC*, *RMSE*, and *MAPE* magnitudes with respect to the target and MLP-model-predicted °API values for the training- and test-set data; these are listed in Table 6.1.

6.3.3 SVR-Based Modeling

A rigorous description of the SVR (Vapnik, 1995, 1996, 1997) based development of a multiple input – single output model is provided in Chapter 2, section 2.3. In the present study, the SVR-based °API-value prediction model was also developed by use of the *RapidMiner* software (RapidMiner 2007). Specifically, the model was constructed by use of the ε-SVR algorithm; the kernel function used was the radial-basis function. The algorithm uses three parameters: regularization constant (C), kernel gamma (γ), and radius of the tube (ε). These were varied systematically to obtain an optimal SVR model possessing high °API-value prediction accuracy and generalization capability. The magnitudes of the stated ε-SVR parameters that led to an optimal SVR model are C = 1.0, γ = 1.0, and ε = 0.001. This optimal model is derived from 260 support vectors. Table 6.1 lists the *CC*, *RMSE*, and *MAPE* magnitudes with respect to the °API-value predictions made by the optimal SVR model for both training and test sets.

6.3.4 Comparison of °API-Value Models

The large-sized data set consisting of 403 data patterns covers a wide range of light, medium, heavy, and very-heavy crude oils. Thus, before comparing the performance of various °API prediction models, an exercise was conducted to improve the prediction and generalization performance of the linear model (Eq. 6.2) proposed by Fan and Buckley (2002). Specifically, the five parameters of Eq. 6.2

were refitted by use of the Marquardt (1963) algorithm. The same training and test data sets that were used in developing the CI-based models were used in this model refitting. The refitted linear model—or the modified Fan and Buckley (2002) (modified-FB) model—is given as

$$°API = -205.889 + 2.483\,S + 2.228\,A + 2.104\,R + 1.905\,A_p \qquad (6.6)$$

The prediction and generalization performance of the three CI-based, FB, and modified-FB models are provided in Table 6.1. This performance assessment is made in terms of the *CC*, *RMSE*, and *MAPE* values computed by use of the experimental and model predicted °API values. For the FB and modified-FB models, these statistical quantities were evaluated by use of the same training and test data sets as used in the development of the CI-based models. From the *CC*, *RMSE*, and *MAPE* values pertaining to the predictions of the modified-FB model (Eq. 6.6), it is observed that refitting the five parameters of the original FB model (Eq. 6.2) has indeed resulted in a significant improvement in the °API-value-prediction accuracy and -generalization capability of the original FB model. Specifically, the *CC* with respect to the training and test data have improved by 11.2 and 11.25%, respectively, whereas the corresponding *RMSE* and *MAPE* magnitudes have decreased by 28.9 and 29%, and 36.08 and 35.48%, respectively.

The *CC*, *RMSE*, and *MAPE* magnitudes listed in Table 6.1 also indicate that there exists a minor variation in the °API-value prediction accuracies and generalization capabilities of the three CI-based models. Here, it is noticed that among the three CI-based models, the prediction and generalization performance of the SVR model is marginally better than that of the GP- and MLP-based models. The high *CC* magnitude of 0.871 with respect to the °API-value predictions made by the SVR model by use of both training- and test-set data clearly indicates that the model possesses good prediction accuracy and generalization capability. This observation is also supported by the lower *RMSE* and *MAPE* magnitudes pertaining to the °API-value predictions by the SVR-based model compared with the predictions made by the GP- and MLP-based models.

A comparison of the °API-value-prediction and –generalization performance exhibited by the CI-based, FB, and modified-FB models reveals the following.

- The *CC* magnitudes corresponding to the predictions made by all three CI-based models by use of the training- and test set data are significantly higher than those for the FB and modified-FB models.
- The training- and test-set *RMSE* and *MAPE* values pertaining to the predictions by all three CI-based models are significantly lower than the corresponding values for the FB and modified-FB models.

Table 6.1: Prediction accuracy of °API values and generalization performance of GP, MLP, SVR, FB and modified FB models

°API-value Model	Training Set			Test Set		
	CC_{trn}	$RMSE_{\text{trn}}$	$MAPE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$	$MAPE_{\text{tst}}$
GP	0.840	5.436	18.01	0.841	5.544	18.19
MLP	0.859	5.220	19.37	0.859	5.192	19.59
SVR	0.871	4.811	13.45	0.871	4.995	13.51
FB	0.730	7.911	36.00	0.727	8.166	36.18
Modified-FB	0.820	5.625	23.01	0.818	5.811	23.34

The overall inferences from the results presented in Table 6.1 are that all three CI-based models outperform the FB and modified-FB models by a wide margin, and that the SVR-based model performs marginally better than the MLP- and GP-based models.

The Steiger (1980) z-test (for more details see Chapter 2, section 2.7) was performed to rigorously compare the prediction and generalization performance of the three CI-based and the modified-FB models. This test is used to examine whether the two *CC*s corresponding to the predictions of two competing models are significantly different. It tests the null hypothesis (H_0) that two *CC* magnitudes are not statistically different; that is, $CC_{\text{AB}} = CC_{\text{AC}}$, where subscripts A, B, and C, respectively—for the present study—denote the experimental °API values and those predicted by models B and C. The choice of the modified-FB model for comparison stems from the fact that its °API prediction and generalization performance is better than that of the original

FB model. The Steiger (1980) z-test has examined the validity of the null hypothesis ($CC_{\mathrm{AB}} = CC_{\mathrm{AC}}$) with respect to the following three pairs of °API values:

- Experimental/GP-model predicted and experimental/modified-FB model predicted
- Experimental/MLP-model predicted and experimental/modified-FB model predicted
- Experimental/SVR-model predicted and experimental/modified-FB model predicted

The results of the Steiger (1980) z-test are listed in Table 6.2. It is observed in this table that in all the three cases, the null hypothesis regarding the equivalence of the two *CC*s—one of which pertains to the predictions of the modified-FB model—has been uniformly rejected. Hence, it is possible to infer that there is a statistically nonsignificant difference in the *CC* magnitudes of the various model pairs. This inference, along with the *CC*, *RMSE*, and *MAPE* values listed in Table 6.1, is clearly indicative of the superior prediction and generalization performance of the three CI-based models compared to the modified-FB model.

Table 6.2: Results of the Steiger (1980) z-test comparing correlation coefficient (*CC*) values of GP, MLP and SVR models with the modified-FB model

Model pair (*B*-*C*)	CC_{AB}	CC_{AC}	Full example set			
			df	*Z*	*P*	H_0
GP/Modified-FB	0.840	0.816	401	2.305	2.11×10^{-2}	Reject
MLP/Modified-FB	0.857	0.816	401	3.881	1.03×10^{-4}	Reject
SVR/Modified-FB	0.870	0.816	401	5.638	1.71×10^{-8}	Reject
FB/Modified-FB	0.727	0.816	401	-6.393	1.62×10^{-10}	Reject

H_0 is rejected when $p < p_0$(where $p_0 = 0.05$). H_0 is the null hypothesis $CC_{AB} = CC_{AC}$, where A denotes experimental ºAPI values of crude oils; df = degrees of freedom.

The parity plots of the experimental °API values and those predicted by the SVR, MLP, GP and the modified-FB models, respectively, are shown in Figures

6.2(*a*) through 6.2(*d*). It can be clearly seen that the data points in Figures 6.2(*a*), 6.2(*b*), and 6.2(*c*), showing the predictions of SVR, MLP, and GP models, respectively, exhibit a lower scatter compared with the predictions of the modified-FB model. This observation also supports the earlier result that the SVR-, MLP-, and GP-based models are capable of predicting the °API values of crude oils with a better accuracy and generalization ability than both the FB and modified-FB models. Among the better-performing models, the GP model—because of its compact size and ease of computation—is more convenient to use and deploy in practical applications. However, when the highest °API-value prediction accuracy is the principal criterion of selection, then the SVR model should receive a preferential treatment.

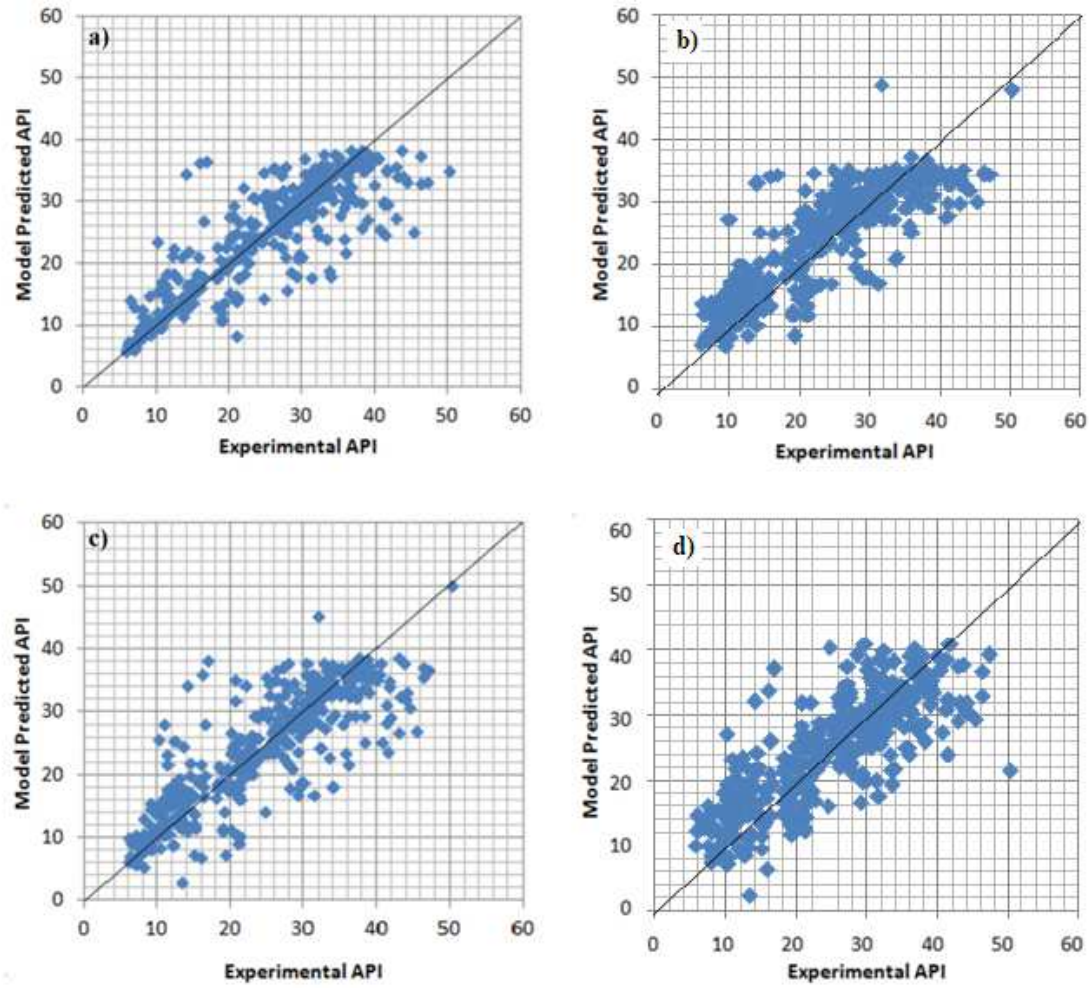


Figure 6.2: Parity plots of the experimental API gravity values and those predicted by the following models: (*a*) SVR, (*b*) MLP, (*c*) GP and, (*d*) Modified-FB.

A notable feature of the GP-based model (Eq. 6.5) is that it possesses a nonlinear structure. It may be noted that by its very character the GP formalism, can search and optimize a linear or a nonlinear function and all the associated parameters that would fit the example data optimally. The fact that the GP model has searched and optimized a nonlinear function for fitting the °API-value data of a large number of crude oils is indicative of °API values indeed being dependent nonlinearly on the weight percentages of the SARA fractions of crude oils. From the superior prediction and generalization performance of the three CI-based nonlinear models, it can be inferred that nonlinear equations are better suited than the linear ones for relating the °API values to the SARA composition of crude oils.

6.4 CONCLUSION

The °API value is an important physicochemical characteristic of crude oils and used routinely in the determination of their other properties and quality. Various models (predominantly linear) have been developed for predicting °API values from the molecular composition of crude oils. Because it requires determining the extent of a large number of crude oil components, the stated approach becomes tedious, costly, and time-consuming. In a practically convenient though rarely used approach, the wt% values of the molecular average structures in the crude oil—saturates (*S*), aromatics (*A*), resins (*R*), and asphaltenes (A_p)—have been used as model inputs to predict the °API values. A linear model derived from this approach was proposed earlier by Fan and Buckley (2002). Scrutiny of an extensive crude-oil database suggests that the relationships between the °API values and wt% values of some of the SARA constituents could be nonlinear. For capturing these nonlinearities and thereby developing models possessing better °API value-prediction accuracies, this study uses three CI-based and exclusively data-driven formalism: GP, MLP, and SVR. Similar to the linear model, these formalisms use the SARA composition of a crude oil for the prediction of its °API value. Among three CI-based methods, GP possesses several novel and attractive characteristics, but it remains a much-less-used data-driven modeling technique when compared with ANNs and SVR. The best-fitting GP based °API-value prediction model developed in this study possesses a nonlinear form. A comparison of the prediction and generalization performance of the

three CI-based models indicates that the SVR strategy has yielded an overall best-performing model. It has been also found that each of the three CI-based nonlinear models possesses a better °API-value prediction accuracy and generalization capability than not only the original linear FB model but also its improved linear version (the modified-FB model). This result clearly indicates that the nonlinear models using weight percentages of the SARA constituents are better-suited than the corresponding linear ones for predicting the °API values of crude oils. Other noteworthy characteristics of the CI-based models developed in this study are as follows.

- A large number of SARA-constituent data and the corresponding °API values pertaining to the light, medium, heavy, and very heavy crude oils have been used in the model development.
- The previously stated SARA analyses were performed by use of various methods, such as TLC-FID, ASTM, HPLC, GC-MS, and open-column chromatography.

These characteristics have imparted a wider applicability to the CI-based models. Because of their significantly higher prediction accuracies, these models possess a potential to be the preferred ones for predicting the °API value of crude oils.

NOMENCLATURE

A	wt% of aromatics
$\hat{A}$	scaled A
A_{p}	wt% of asphaltenes
$\hat{A}_p$	scaled A_{p}
N_{p}	number of patterns in the example data set; number input/output patterns in the training/test set for GP
N_{pp}	number of candidate solutions in a population for GP
R	wt% of resins
$\hat{R}$	scaled R

S wt% of saturates

$\hat{S}$ scaled S

y_i desired (target) output value corresponding to the i^{th}-input data pattern in the training/test data set

$\hat{y}_{\text{i,j}}$ magnitude of the model-predicted °API value when i^{th} -input pattern is used to compute the output of the j^{th}-candidate solution

Appendix 6.A: ºAPI-Value Models Data

Sr. No.	Saturates (wt %)	Aromatics (wt %)	Resins (wt%)	Asphaltenes (wt %)	ºAPI values	Reference
1	20.01	11.36	36.26	32.37	10.3	Sanchez - Minero et al. (2013)
2	25.38	17.04	31.41	26.17	13.1	Sanchez - Minero et al. (2013)
3	32.1	26.34	25.82	15.74	21.2	Sanchez - Minero et al. (2013)
4	43.15	29.95	18.2	8.7	27.1	Sanchez - Minero et al. (2013)
5	51.62	31.35	14.25	2.78	32.8	Sanchez - Minero et al. (2013)
6	27.9	31.7	32.1	7.3	10.5	Molina et al. (2010)
7	20.5	39.2	36.5	3.6	10.7	Molina et al. (2010)
8	35.8	30.3	32.1	1.8	13.8	Molina et al. (2010)
9	23	22	35	18	10.4	Hinkle et al. (2010)
10	38.4	29.8	25.8	4.8	16.5	Wang and Buckley (2003)
11	64.1	14.5	17.9	2.7	20.7	Wang and Buckley (2003)
12	49.5	21.5	25.6	2.8	22.6	Wang and Buckley (2003)
13	67.3	14.9	15.1	2.3	29.2	Wang and Buckley (2003)
14	70.6	15	12.9	1.3	31	Wang and Buckley (2003)
15	70.6	16.3	11.4	1.9	31.1	Wang and Buckley (2003)
16	62.8	15.8	18.7	2.6	31.2	Wang and Buckley (2003)
17	63.4	16.5	17.4	2.7	31.6	Wang and Buckley (2003)
18	65.2	18.3	13.9	1.3	37.2	Wang and Buckley (2003)
19	59.4	24.9	10.2	6.5	41.3	Wang and Buckley (2003)
20	46	25	15	12	13.5	Hernandez et al. (1983)
21	37	31	12	18	14.3	Hernandez et al. (1983)
22	11	12	64	15	15	Cendejas et al. (2013)
23	27	15	47	11	21	Cendejas et al. (2013)
24	38	15	42	5	30	Cendejas et al. (2013)
25	42.72	38.47	17.86	0.31	33.39	Alcazar-Vara and Buenrostro – Gonzalez (2011)
26	46.81	37.13	15.63	0.01	38.27	Alcazar-Vara and Buenrostro – Gonzalez (2011)
27	44.03	38.32	16.65	0.01	41.35	Alcazar-Vara and Buenrostro – Gonzalez (2011)

28	30.1	31.3	13.6	25	18	Kumar et al. (2005)
29	48.1	28.8	6.5	16.5	28	Kumar et al. (2005)
30	32.3	19.4	37.1	11.2	22.6	Nasr-El-Din and Taylor (1992)
31	36.77	46.72	6.12	10.39	19.85	Kord and Ayatollahi (2012)
32	43.67	52.09	0.49	3.75	20.75	Kord and Ayatollahi (2012)
33	46.12	37.24	7.57	9.07	20.93	Kord and Ayatollahi (2012)
34	38.99	50.59	6.17	4.25	22.71	Kord and Ayatollahi (2012)
35	42.68	40.69	7.63	9	24.46	Kord and Ayatollahi (2012)
36	62.4	17.2	6.2	14.2	37.2	Tang et al. (2005)
37	39.6	9.1	44.5	6.8	12.1	Subramaniam and Hanson (1998)
38	35.7	7	54.5	2.9	12.9	Subramaniam and Hanson (1998)
39	26.2	41.5	21.9	10.2	12	Hannisdal et al. (2006)
40	19	45	20	16	19	Freitas et al. (2009)
41	19	32	38	12	19	Islas- Flores et al. (2006)
42	55	30	13	2	36	Islas- Flores et al. (2006)
43	26.96	42.65	15.03	15.36	10.2	Rose et al. (2001)
44	52.49	41.04	5.48	0.99	34.24	Nokandeh et al. (2012)
45	29.53	54.52	12.04	3.91	23	Kazempour et al. (2013)
46	41.81	44.15	10.8	3.24	25.7	Kazempour et al. (2013)
47	18.17	28.97	41.52	11.31	21	Chávez-Miyauchi et al. (2013)
48	41.8	28.7	28.4	1.5	28.4	Alcazar-Vara et al. (2012)
49	41.7	34.2	21.8	2.3	36	Alcazar-Vara et al. (2012)
50	42	43	8	7	23.8	Amin et al. (2011)
51	53.48	34.45	8.5	5.3	31	Amin et al. (2011)
52	16.8	44.9	24.8	13.5	10.2	Clarke and Pruden (1997)
53	43	50	7	0	10.1	World Data Base (2005)
54	25	35	22	18	10.9	World Data Base (2005)
55	54	14	15	17	11	World Data Base (2005)
56	26	29	22	22	11.2	World Data Base (2005)
57	25	47	17	11	11.4	World Data Base (2005)
58	43	24	11	22	11.4	World Data Base (2005)

59	32	32	17	19	11.6	World Data Base (2005)
60	21	35	24	21	11.7	World Data Base (2005)
61	26	52	12	10	11.9	World Data Base (2005)
62	24	55	15	6	12.3	World Data Base (2005)
63	30	62	7	1	12.3	World Data Base (2005)
64	19	35	23	22	13.2	World Data Base (2005)
65	28	39	30	3	13.6	World Data Base (2005)
66	21	39	31	7	13.7	World Data Base (2005)
67	66	24	8	2	14	World Data Base (2005)
68	32	41	24	3	14.3	World Data Base (2005)
69	29	51	11	10	14.7	World Data Base (2005)
70	24	43	20	12	14.8	World Data Base (2005)
71	23	0	76	1	15.2	World Data Base (2005)
72	70	23	6	1	16	World Data Base (2005)
73	53	27	10	10	16.4	World Data Base (2005)
74	80	19	1	0	16.8	World Data Base (2005)
75	32	32	17	19	18.2	World Data Base (2005)
76	34	31	20	15	18.3	World Data Base (2005)
77	38	29	20	13	18.8	World Data Base (2005)
78	38	40	14	8	19.5	World Data Base (2005)
79	33	31	24	12	19.6	World Data Base (2005)
80	19	63	12	6	19.7	World Data Base (2005)
81	34	32	21	13	19.8	World Data Base (2005)
82	46	30	13	10	20.3	World Data Base (2005)
83	39	35	21	5	20.4	World Data Base (2005)
84	39	28	21	12	20.6	World Data Base (2005)
85	68	22	4	6	20.7	World Data Base (2005)
86	38	39	8	16	21.3	World Data Base (2005)
87	36	25	23	16	21.4	World Data Base (2005)
88	39	34	11	16	21.8	World Data Base (2005)
89	36	22	29	13	22.1	World Data Base (2005)
90	38	61	1	0	22.4	World Data Base (2005)

91	38	38	14	11	22.6	World Data Base (2005)
92	44	30	17	9	22.9	World Data Base (2005)
93	53	38	7	2	23.2	World Data Base (2005)
94	43	37	13	7	23.4	World Data Base (2005)
95	48	36	14	3	23.7	World Data Base (2005)
96	34	64	2	0	24.3	World Data Base (2005)
97	87	10	2	2	24.8	World Data Base (2005)
98	53	37	6	4	25	World Data Base (2005)
99	59	34	6	0	25.3	World Data Base (2005)
100	66	23	4	6	25.3	World Data Base (2005)
101	48	30	17	6	25.6	World Data Base (2005)
102	48	32	9	12	25.9	World Data Base (2005)
103	68	23	7	2	26.1	World Data Base (2005)
104	70	25	4	0	26.2	World Data Base (2005)
105	48	31	13	8	26.2	World Data Base (2005)
106	61	26	6	8	26.3	World Data Base (2005)
107	60	24	8	8	28.8	World Data Base (2005)
108	69	28	3	0	27.1	World Data Base (2005)
109	56	31	11	3	27.3	World Data Base (2005)
110	90	9	0	0	27.4	World Data Base (2005)
111	47	35	12	6	27.5	World Data Base (2005)
112	45	40	11	3	27.6	World Data Base (2005)
113	72	25	2	0	27.8	World Data Base (2005)
114	51	39	9	1	28.4	World Data Base (2005)
115	53	34	10	4	28.5	World Data Base (2005)
116	55	35	9	1	28.6	World Data Base (2005)
117	88	11	1	0	28.7	World Data Base (2005)
118	55	31	10	4	29.4	World Data Base (2005)
119	54	32	7	6	29.5	World Data Base (2005)
120	53	36	10	1	29.5	World Data Base (2005)
121	54	32	8	6	29.8	World Data Base (2005)
122	95	3	2	0	29.8	World Data Base (2005)

123	52	35	9	5	29.9	World Data Base (2005)
124	53	30	11	6	30	World Data Base (2005)
125	92	7	0	0	30.1	World Data Base (2005)
126	51	36	9	5	30.2	World Data Base (2005)
127	80	18	2	0	30.3	World Data Base (2005)
128	57	27	9	7	30.3	World Data Base (2005)
129	84	13	2	0	30.4	World Data Base (2005)
130	51	34	9	5	30.6	World Data Base (2005)
131	86	12	1	0	30.7	World Data Base (2005)
132	64	22	9	5	30.7	World Data Base (2005)
133	60	35	5	1	31	World Data Base (2005)
134	62	25	9	4	31	World Data Base (2005)
135	74	12	9	6	31	World Data Base (2005)
136	85	13	1	0	31.1	World Data Base (2005)
137	65	28	6	1	31.2	World Data Base (2005)
138	66	26	6	2	31.6	World Data Base (2005)
139	86	12	2	0	31.7	World Data Base (2005)
140	51	39	6	3	31.8	World Data Base (2005)
141	3	97	0	0	31.8	World Data Base (2005)
142	60	28	6	5	32	World Data Base (2005)
143	64	27	7	2	32	World Data Base (2005)
144	61	37	2	0	32.3	World Data Base (2005)
145	61	32	6	1	32.3	World Data Base (2005)
146	56	32	8	5	32.4	World Data Base (2005)
147	91	7	2	0	32.5	World Data Base (2005)
148	82	17	1	0	32.6	World Data Base (2005)
149	65	29	5	1	32.8	World Data Base (2005)
150	62	26	7	5	32.8	World Data Base (2005)
151	70	15	6	8	32.9	World Data Base (2005)
152	73	21	5	1	33	World Data Base (2005)
153	67	22	8	4	33.4	World Data Base (2005)
154	73	20	4	3	33.4	World Data Base (2005)

155	57	42	0	0	33.7	World Data Base (2005)
156	71	21	5	4	33.7	World Data Base (2005)
157	67	25	7	1	33.8	World Data Base (2005)
158	81	14	2	4	33.8	World Data Base (2005)
159	65	25	8	2	34.4	World Data Base (2005)
160	73	21	4	1	34.5	World Data Base (2005)
161	82	13	2	2	34.8	World Data Base (2005)
162	78	16	5	0	35.1	World Data Base (2005)
163	71	25	4	0	35.2	World Data Base (2005)
164	62	31	6	2	35.7	World Data Base (2005)
165	71	20	8	1	35.8	World Data Base (2005)
166	83	13	2	3	36	World Data Base (2005)
167	65	27	5	3	36.1	World Data Base (2005)
168	61	30	8	2	36.1	World Data Base (2005)
169	66	26	6	1	36.4	World Data Base (2005)
170	65	25	6	5	36.5	World Data Base (2005)
171	64	32	4	0	36.7	World Data Base (2005)
172	70	22	6	2	36.7	World Data Base (2005)
173	78	18	3	1	36.8	World Data Base (2005)
174	81	16	3	0	36.8	World Data Base (2005)
175	84	14	2	1	36.9	World Data Base (2005)
176	84	13	1	2	37	World Data Base (2005)
177	79	15	4	3	37.1	World Data Base (2005)
178	76	22	2	0	37.2	World Data Base (2005)
179	76	23	1	0	37.6	World Data Base (2005)
180	72	23	4	1	37.8	World Data Base (2005)
181	68	26	6	2	37.8	World Data Base (2005)
182	80	18	3	0	38	World Data Base (2005)
183	73	22	4	1	38.1	World Data Base (2005)
184	76	20	3	1	38.3	World Data Base (2005)
185	85	11	2	1	38.4	World Data Base (2005)
186	62	32	5	2	38.6	World Data Base (2005)

187	79	15	6	0	38.7	World Data Base (2005)
188	74	24	1	0	38.8	World Data Base (2005)
189	72	22	4	2	38.9	World Data Base (2005)
190	74	21	3	1	39	World Data Base (2005)
191	72	22	5	1	39	World Data Base (2005)
192	68	25	7	1	39.2	World Data Base (2005)
193	69	24	6	1	39.4	World Data Base (2005)
194	88	11	0	1	39.9	World Data Base (2005)
195	81	17	1	1	40.5	World Data Base (2005)
196	94	5	0	0	41.3	World Data Base (2005)
197	94	6	0	0	41.8	World Data Base (2005)
198	81	19	0	0	42.9	World Data Base (2005)
199	76	21	3	1	43.6	World Data Base (2005)
200	19	14	46	20	10	Hinkle et al. (2008)
201	35.7	24.6	32.4	7.3	13	Lammoglia and Filho (2011)
202	42.5	33.1	22.3	2.12	14.4	Lammoglia and Filho (2011)
203	40.2	33.3	23.4	3.1	19.4	Lammoglia and Filho (2011)
204	44.9	32.1	20.6	2.4	20	Lammoglia and Filho (2011)
205	49.6	28.6	20	1.76	21.3	Lammoglia and Filho (2011)
206	72.7	13.9	13.4	0.5	27.4	Lammoglia and Filho (2011)
207	68.1	17.6	14.3	0.5	27.7	Lammoglia and Filho (2011)
208	55.7	24.3	19.1	0.9	28.1	Lammoglia and Filho (2011)
209	50.4	28.1	19.7	1.8	29.4	Lammoglia and Filho (2011)
210	81.2	6	12.8	0.5	36.2	Lammoglia and Filho (2011)
211	79.2	13.4	7.4	0.5	40.2	Lammoglia and Filho (2011)
212	85.9	14.1	0.1	0.1	47.2	Lammoglia and Filho (2011)
213	20.74	39.2	24.81	15.25	10.71	Hsu and Robinson (2007)
214	15.83	36.74	18.61	28.82	12	Ancheyta (2013)
215	47.9	36.5	15.2	0.4	23.314	Ancheyta (2013)
216	48	37.5	14.2	0.3	22.98	Ancheyta (2013)
217	41.2	36.4	20.4	2.1	22.98	Ancheyta (2013)
218	82.7	13.4	3.9	0	37.2	Ancheyta (2013)

219	62.7	23.6	12.2	1.5	36.2	Ancheyta (2013)
220	35.3	36.8	24.5	3.5	18.23	Ancheyta (2013)
221	41.8	38.8	18.7	0.6	23.3	Ancheyta (2013)
222	50.9	34.6	14	0.5	28.4	Ancheyta (2013)
223	40.6	32.1	20.6	6.6	27.8	Ancheyta (2013)
224	79.8	16.5	3.6	0.1	46.3	Ancheyta (2013)
225	57.3	27.9	13.5	1.3	30.6	Ancheyta (2013)
226	60.6	30	9.2	0.2	33.62	Ancheyta (2013)
227	42.4	36.1	20.5	1	22.1	Ancheyta (2013)
228	65	30.7	4.3	0	46.3	Ancheyta (2013)
229	50.3	31.4	17.5	0.7	26.1	Ancheyta (2013)
230	55.4	28.3	12.9	3.4	37	Ancheyta (2013)
231	54.5	28.8	14.9	1.8	30.6	Ancheyta (2013)
232	24.4	43.4	19.9	12.4	19.2	Ancheyta (2013)
233	45	29	14	12	26.12	Kök et al. (1998)
234	18	31	22	29	14.95	Kök et al. (1998)
235	69	18	13	0	36	Al-Saffar et al. (2001)
236	43	40	13	6	25.1	Pantoja et al. (2011)
237	50	30	13	5	26.2	Pantoja et al. (2011)
238	51	30	13	5	26.4	Pantoja et al. (2011)
239	62	31	7	2	32.2	Pantoja et al. (2011)
240	61	29	8	2	32.2	Pantoja et al. (2011)
241	63	27	7	3	34.5	Pantoja et al. (2011)
242	18.5	31.9	37.9	11.7	19	Islas-Flores et al. (2005)
243	38.44	14.59	41.44	5.53	29.59	Castro and Vazquez (2009)
244	26.53	14.74	47.6	11.13	21.27	Castro and Vazquez (2009)
245	10.49	9	64.12	16.39	15.82	Castro and Vazquez (2009)
246	15	19.11	46.78	19.11	9.17	Castro and Vazquez (2009)
247	56.2	25.7	17.1	1	28.4	Khalil de Oliveira et al. (2012)
248	51.1	30.9	16.6	1.4	29.8	Khalil de Oliveira et al. (2012)
249	29	42.2	15.8	13	11	Tharanivasan et al. (2009)
250	17.8	46.2	18.4	17.3	7	Tharanivasan et al. (2009)

251	18.2	42.7	21.5	17.6	8	Tharanivasan et al. (2009)
252	50.3	30.5	14.6	4	20	Tharanivasan et al. (2009)
253	17.3	39.7	25.8	16.9	9	Tharanivasan et al. (2009)
254	61.1	29.6	5.3	4	32	Tharanivasan et al. (2009)
255	60.9	36.6	2.4	0	22	Tharanivasan et al. (2009)
256	16.43	34.91	41.12	5.61	8.89	Zhang et al. (2013)
257	55.1	38.8	3.6	2.8	44.45	Flego and Zannoni (2012)
258	57.8	38	2.9	1.8	43.11	Flego and Zannoni (2012)
259	12.3	78.2	8.1	2.9	50.21	Flego and Zannoni (2012)
260	61.2	34.7	2.8	1.7	41.53	Flego and Zannoni (2012)
261	32.3	41.8	20.5	5.8	33.94	Flego and Zannoni (2012)
262	46.3	49.4	4.8	0	45.38	Flego and Zannoni (2012)
263	53.5	35.8	9.3	1.4	38.3	Flego and Zannoni (2012)
264	63.8	7.8	4.4	24.1	35.88	Flego and Zannoni (2012)
265	11.6	48.3	23.2	17	19.34	Flego and Zannoni (2012)
266	17.8	68.9	9.8	3.6	33.6	Flego and Zannoni (2012)
267	25.1	41	16.3	17.8	24.7	Flego and Zannoni (2012)
268	27.1	50.2	15.8	7	31.35	Flego and Zannoni (2012)
269	9.6	30.5	40.1	20	12.7	Flego and Zannoni (2012)
270	7.1	22.3	38	32	13.3	Flego and Zannoni (2012)
271	15.6	45.7	19.3	19.4	21.1	Flego and Zannoni (2012)
272	47.9	29.2	18	6	40.7	Flego and Zannoni (2012)
273	43.4	35.8	14.5	6.8	35.7	Flego and Zannoni (2012)
274	50.4	31.6	15.8	3.2	42.9	Flego and Zannoni (2012)
275	31.3	44.6	18.9	5.5	33.7	Flego and Zannoni (2012)
276	62.2	27.3	7.3	3.7	44.1	Flego and Zannoni (2012)
277	57.9	35.1	4.9	2.4	43.8	Flego and Zannoni (2012)
278	46.4	30.5	19.4	4.4	32.2	Flego and Zannoni (2012)
279	28.1	50.3	16.8	5	29.4	Flego and Zannoni (2012)
280	49.1	31.7	17.6	2.3	26.5	Flego and Zannoni (2012)
281	15.2	33.4	35.1	16.3	10.9	Flego and Zannoni (2012)
282	65.13	16.86	4.13	13.88	38	Gui et al. (2010)

283	34.22	38.82	19.96	6.58	18.89	Hoshyargar and Ashrafizadeh (2013)
284	18.8	51.9	14.6	14.3	10	Poindexter and Marsh (2009)
285	10.7	57.4	24.1	7.9	10.5	Poindexter and Marsh (2009)
286	14.6	53.1	25	8.3	11.1	Poindexter and Marsh (2009)
287	68.3	17.1	9.4	3.2	31.2	Abudu and Goual (2008)
288	43.21	35.3	16.68	4.99	22.4	Rogel et al. (2003)
289	45.31	33.29	17.55	3.85	22.2	Rogel et al. (2003)
290	49.43	37.62	8.61	4.43	25.7	Rogel et al. (2003)
291	49.68	37.78	9.28	3.34	26.8	Rogel et al. (2003)
292	48.61	34.35	11.65	5.38	22.8	Rogel et al. (2003)
293	48.83	37.4	9.51	4.26	25.6	Rogel et al. (2003)
294	54.54	35.35	8.23	1.87	25	Rogel et al. (2003)
295	55.41	36.77	6.89	0.94	26.3	Rogel et al. (2003)
296	49.41	38.23	10.44	1.92	27.2	Rogel et al. (2003)
297	39.62	38.71	16.83	4.93	18.3	Rogel et al. (2003)
298	42.64	36.35	12.96	7.74	24.1	Rogel et al. (2003)
299	49.53	41.33	2.54	6.15	26.1	Rogel et al. (2003)
300	51.79	29.93	15.83	2.46	26.7	Rogel et al. (2003)
301	54.73	30.41	12.78	2.08	25.1	Rogel et al. (2003)
302	32.45	41.5	21.12	4.93	16.3	Rogel et al. (2003)
303	26.13	45.3	22.57	6.01	14	Rogel et al. (2003)
304	47.4	21.7	25.5	5.4	22.8	Cunha et al. (2008)
305	51.2	24.2	23.1	1.5	26.6	Khalil de Oliveira et al. (2012)
306	54.5	23	22	0.5	27.4	Khalil de Oliveira et al. (2012)
307	57.1	24.5	18	0.4	27.8	Khalil de Oliveira et al. (2012)
308	53.8	22	23.7	0.5	28.3	Khalil de Oliveira et al. (2012)
309	57.7	24.2	17.4	0.7	28.8	Khalil de Oliveira et al. (2012)
310	52.7	33.6	12.6	1.1	29	Khalil de Oliveira et al. (2012)
311	56.6	24.4	19	0.5	29.5	Khalil de Oliveira et al. (2012)
312	61.3	24.7	13.9	0.5	30.6	Khalil de Oliveira et al. (2012)
313	57.2	26.2	14.7	2	31.1	Khalil de Oliveira et al. (2012)

314	66.7	20.1	12.8	0.4	33.9	Khalil de Oliveira et al. (2012)
315	11.7	32	41.6	14.7	7.5	Angle and Hua (2011)
316	14.1	37.3	37.2	11.4	9	Poteau et al. (2005)
317	15	38	35	12	8	Acevedo et al. (2004)
318	17	44	30	10	14	Acevedo et al. (2004)
319	42	34	18	7	21	Acevedo et al. (2004)
320	41	45	12	2	21	Acevedo et al. (2004)
321	40.3	29	19.2	9.5	15.8	Maninpey et al. (2010)
322	12	37	33	17	9	Dalmazzone et al. (2012)
323	52.9	29.7	13.2	4	29.29	Panuganti et al. (2011)
324	66.26	25.59	5.35	2.8	41.6	Panuganti et al. (2011)
325	19	25	43	13	9	Marcano et al. (2011)
326	12	36	38	14	10.3	Marcano et al. (2011)
327	55	28	13	4	23.7	Marcano et al. (2011)
328	19	28	42	11	8	Marcano et al. (2011)
329	60	14	24	2	24.3	Marcano et al. (2011)
330	52	26	16	6	30.3	Marcano et al. (2011)
331	32.3	38.25	21.6	6.04	18.36	Khansari et al. (2012)
332	20	43	27	10	12.9	Linan et al. (2010)
333	17	41	29	13	11.6	Linan et al. (2010)
334	19	42	28	11	12.9	Linan et al. (2010)
335	61	20	19	0.59	27	Ferno et al. (2010)
336	25	33	29	13	9	Ocanto et al. (2009)
337	30	26	32	12	15	Ocanto et al. (2009)
338	10	23	48	19	8	Ocanto et al. (2009)
339	25	28	35	11	15	Ocanto et al. (2009)
340	11	19	57	13	8	Ocanto et al. (2009)
341	25	24	36	15	20	Ocanto et al. (2009)
342	21	27	37	15	21	Ocanto et al. (2009)
343	35	33	28	4	28	Ocanto et al. (2009)
344	44	25	21	10	20	Ocanto et al. (2009)
345	22	30	44	4	21	Ocanto et al. (2009)

346	35	24	32	9	21	Ocanto et al. (2009)
347	40.41	42.06	12.21	5.32	27.5	Juyal et al. (2012)
348	59.56	32.76	6.95	0.73	35	Juyal et al. (2012)
349	57.4	30.8	10.4	1.4	32	Gonzalez et al. (2005)
350	55.8	23.9	17.5	2.7	30	Kraiwattanawong et al. (2009)
351	59.6	26.5	10.1	3.8	30	Kraiwattanawong et al. (2009)
352	49.5	28.4	12.4	9.7	27	Kraiwattanawong et al. (2009)
353	45.3	26.8	24.9	3.1	22.4	Rocha et al. (2013)
354	52	30	12	6	26	Juyal et al. (2009)
355	30.1	42.1	13.36	13.5	29.17	Kord et al. (2012)
356	32.61	43.48	7.61	16.3	20.29	Jafari Behbahani et al. (2012)
357	22.6	33.6	32.9	10.8	8.1	Cinar et al. (2011)
358	44.65	34.55	17.9	2.86	29.3	Mendoza de la Cruz et al. (2009)
359	44.14	40.13	12.79	2.94	32.03	Gonzalez et al. (2007)
360	33	14	51	0	16.2	Abivin and Taylor (2012)
361	29	26	39	7	10.2	Abivin and Taylor (2012)
362	28	29	37	6	10.7	Abivin and Taylor (2012)
363	29	19	35	17	13.2	Abivin and Taylor (2012)
364	22	24	45	10	7.4	Abivin and Taylor (2012)
365	33	22	40	4	13.5	Abivin and Taylor (2012)
366	38	20	37	4	15.6	Abivin and Taylor (2012)
367	45	18	26	11	11.4	Abivin and Taylor (2012)
368	29	23	32	15	10.6	Abivin and Taylor (2012)
369	34	24	27	13	15.5	Abivin and Taylor (2012)
370	22	20	48	11	9	Abivin and Taylor (2012)
371	32	24	34	9	11.4	Abivin and Taylor (2012)
372	54.8	23.57	21.21	0.41	35.3	Mena-Cervantes et al. (2011)
373	10.9	61.5	18.1	9.5	6	Chang et al. (2003)
374	8.3	35.6	45.4	10.7	5.9	Chang et al. (2003)
375	12.4	45.1	35.9	4.7	9.11	Marques et al. (2011)
376	23	21.1	38.8	17.1	10.5	Angle et al. (2005)
377	21	19	44	16	11	Angle et al. (2005)

378	16.1	48.5	16.8	18.6	6.4	Fadaei et al. (2011)
379	58.4	26.2	14.61	0.79	29.9	Pacheco et al. (2011)
380	30.27	45.05	18.7	5.99	11.42	Long et al. (2011)
381	6.1	60.3	23.5	10.1	6.1	Fathi and Pereira-Almao (2011)
382	9.4	63.1	13.3	14.1	6.7	Fathi and Pereira-Almao (2011)
383	9.2	62.8	13.7	14.2	6.5	Fathi and Pereira-Almao (2011)
384	9.4	61.9	13.6	15.1	6.9	Fathi and Pereira-Almao (2011)
385	9.3	62.6	14.1	13.9	6.9	Fathi and Pereira-Almao (2011)
386	9.8	64.6	12.5	13.2	7.2	Fathi and Pereira-Almao (2011)
387	9.2	64.3	12.2	14.3	7.1	Fathi and Pereira-Almao (2011)
388	10.3	66.4	10.4	12.9	7.3	Fathi and Pereira-Almao (2011)
389	9.8	65.6	11.2	13.3	7.2	Fathi and Pereira-Almao (2011)
390	30.2	24.8	40.1	3.6	13.5	Bukka et al. (1992)
391	27.91	60.64	6.35	5.3	12.89	Bahzad et al. (2010)
392	20.35	62.92	6.88	9.85	9.58	Bahzad et al. (2010)
393	16.3	39.8	26.4	17.5	8.05	Peramanu et al. (2001)
394	19.4	38.1	26.7	15.8	10.7	Peramanu et al. (2001)
395	23.1	41.7	20.4	14.8	12.5	Peramanu et al. (2001)
396	20.8	41.1	22.1	16	11.1	Peramanu et al. (2001)
397	36.9	37.9	19.4	5.8	22.4	Leon et al. (2002)
398	32.3	42.2	19.8	5.8	18.3	Leon et al. (2002)
399	43.6	35.5	14.3	6.6	22.8	Leon et al. (2002)
400	44.3	38.9	11.6	5.2	25.6	Leon et al. (2002)
401	45.6	34.2	17	3.2	25	Leon et al. (2002)
402	51.9	38.9	8.1	1.1	26.3	Leon et al. (2002)
403	68.3	11.6	18.8	1.3	39	Mohammadi et al. (2012)

REFERENCES

Abbas, O., Rebufa, C., Dupuy, N., Permanyer, A., and Kister, J. (2012). PLS regression on spectroscopic data for the prediction of crude oil quality: API gravity and aliphatic/aromatic ratio. *Fuel*, *98*, 5-14. http://dx.doi.org/10.1016/j.fuel.2012.03.045.

Abivin, P., Taylor, S. D., and Freed, D. (2012). Thermal behavior and viscoelasticity of heavy oils. *Energy & Fuels*, *26*(6), 3448-3461. http://dx.doi.org/10.1021/ef300065h.

Abudu, A., and Goual, L. (2008). Adsorption of crude oil on surfaces using quartz crystal microbalance with dissipation (QCM-D) under flow conditions. *Energy & Fuels*, *23*(3), 1237-1248. http://dx.doi.org/10.1021/ef800616x.

Acevedo, S., Rodríguez, P., and Labrador, H. (2004). An electron microscopy study of crude oils and maltenes. *Energy & Fuels*, *18*(6), 1757-1763. http://dx.doi.org/10.1021/ef040044j.

Albahri, T. A., Riazi, M. R., and Alqattan, A. A. (2003). Analysis of quality of the petroleum fuels. *Energy & Fuels*, *17*(3), 689-693. http://dx.doi.org/10.1021/ef020250w.

Alcazar-Vara, L. A., and Buenrostro-Gonzalez, E. (2011). Characterization of the wax precipitation in Mexican crude oils. *Fuel Processing Technology*, *92*(12), 2366-2374. http://dx.doi.org/10.1016/j.fuproc.2011.08.012.

Alcazar-Vara, L. A., Garcia-Martinez, J. A., and Buenrostro-Gonzalez, E. (2012). Effect of asphaltenes on equilibrium and rheological properties of waxy model systems. *Fuel*, *93*, 200-212. http://dx.doi.org/10.1016/j.fuel.2011.10.038.

Al-Saffar, H. B., Hasanin, H., Price, D., & Hughes, R. (2001). Oxidation reactions of a light crude oil and its SARA fractions in consolidated cores. *Energy & Fuels*, *15*(1), 182-188. http://dx.doi.org/10.1021/ef000135q.

Amin, J. S., Nikooee, E., Ghatee, M. H., Ayatollahi, S., Alamdari, A., and Sedghamiz, T. (2011). Investigating the effect of different asphaltene structures on surface

topography and wettability alteration. *Applied Surface Science*, *257*(20), 8341-8349. http://dx.doi.org/10.1016/j.apsusc.2011.03.123.

Ancheyta, J. (2013). *Modeling of processes and reactors for upgrading of heavy petroleum*. CRC Press, Boca Raton, Florida.

Angle, C. W., and Hua, Y. (2011). Phase Separation and Interfacial Viscoelasticity of Charge-Neutralized Heavy Oil Nanoemulsions in Water. *Journal of Chemical & Engineering Data*, *56*(4), 1388-1396. http://dx.doi.org/10.1021/je101162n.

Angle, C. W., Lue, L., Dabros, T., and Hamza, H. A. (2005). Viscosities of heavy oils in toluene and partially deasphalted heavy oils in heptol in a study of asphaltenes self-interactions. *Energy & fuels*, *19*(5), 2014-2020. http://dx.doi.org/10.1021/ef0500235.

ASTM D1298-12b, Standard Test Method for Density, Relative Density, or API Gravity of Crude PEtroleum and Liquid Petroleum Products by Hydrometer Method. 2012. West Conshohocken, Pennsylvania: ASTM International. http://dx.doi.org/10.1520/D1298-12B.

ASTM D2007-93, Standard Test Method for Characteristic Groups in Rubber Extender and Processing Oils and Other Petroleum-Derived Oils by the Clay-Gel Absorption Chromatographic Method. 1993. West Conshohocken, Pennsylvania: ASTM International.

ASTM D287-12, Standard Test Method for API Gravity of Crude Petroleum and Petroleum Products (*Hydrometer Method*). 2012. West Conshohocken, Pennsylvania: ASTM International. http://dx.doi.org/10.1520/D0287-12.

Bahzad, D., Al-Fadhli, J., Al-Dhafeeri, A., and Abdal, A. (2010). Assessment of selected apparent kinetic parameters of the HDM and HDS reactions of two Kuwaiti residual oils, using two types of commercial ARDS catalysts. *Energy & Fuels*, *24*(3), 1495-1501. http://dx.doi.org/10.1021/ef9012104.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford university press, New York.

Bukka, K., Miller, J. D., Hanson, F. V., and Oblad, A. G. (1992). Fractionation and characterization of Whiterocks tar-sand bitumen. *Energy & Fuels*, *6*(2), 160-165. http://dx.doi.org/10.1021/ef00032a007.

Castro, L. V., and Vazquez, F. (2009). Fractionation and characterization of Mexican crude oils. *Energy & Fuels*, *23*(3), 1603-1609. http://dx.doi.org/10.1021/ef8008508.

Cendejas, G., Arreguín, F., Castro, L. V., Flores, E. A., and Vazquez, F. (2013). Demulsifying super-heavy crude oil with bifunctionalized block copolymers. *Fuel*, *103*, 356-363. http://dx.doi.org/10.1016/j.fuel.2012.08.029.

Chaffin, J. M., Lin, M. S., Liu, M., Davison, R. R., Glover, C. J., and Bullin, J. A. (1996). The use of HPLC to determine the saturate content of heavy petroleum products. *Journal of liquid chromatography & Related Technologies*, *19*(10), 1669-1682. http://dx.doi.org/10.1080/10826079608005500.

Chang, J., Fujimoto, K., and Tsubaki, N. (2003). Effect of initiative additives on hydro-thermal cracking of heavy oils and model compound. *Energy & Fuels*, *17*(2), 457-461. http://dx.doi.org/10.1021/ef020190u.

Chávez-Miyauchi, T. E., Zamudio-Rivera, L. S., Barba-López, V., Buenrostro-Gonzalez, E., and Martínez-Magadán, J. M. (2013). N-aryl amino-alcohols as stabilizers of asphaltenes. *Fuel*, *110*, 302-309. http://dx.doi.org/10.1016/j.fuel.2012.10.044.

Cinar, M., Castanier, L. M., and Kovscek, A. R. (2011). Combustion kinetics of heavy oils in porous media. *Energy & Fuels*, *25*(10), 4438-4451. http://dx.doi.org/10.1021/ef200680t.

Clarke, P. F., and Pruden, B. B. (1997). Asphaltene precipitation: Detection using heat transfer analysis, and inhibition using chemical additives. *Fuel*, *76*(7), 607-614. http://dx.doi.org/10.1016/S0016-2361(97)00052-5.

Cunha, R. E., Fortuny, M., Dariva, C., and Santos, A. F. (2008). Mathematical modeling of the destabilization of crude oil emulsions using population balance equation. *Industrial & Engineering Chemistry Research*, *47*(18), 7094-7103. http://dx.doi.org/10.1021/ie800391v.

Dalmazzone, C., Noïk, C., and Argillier, J. F. (2012). Impact of chemical enhanced oil recovery on the separation of diluted heavy oil emulsions. *Energy & Fuels*, *26*(6), 3462-3469. http://dx.doi.org/10.1021/ef300083z.

de Oliveira, M. C. K., Teixeira, A., Vieira, L. C., de Carvalho, R. M., de Carvalho, A. B. M., and do Couto, B. C. (2011). Flow assurance study for waxy crude oils. *Energy & Fuels*, *26*(5), 2688-2695. http://dx.doi.org/10.1021/ef201407j.

Edwards, L. 2009. Eureqa, the Robot Scientist, http://www.physorg.com/news179394947.html (accessed 20 November 2015).

Fadaei, H., Scarff, B., and Sinton, D. (2011). Rapid microfluidics-based measurement of CO_2 diffusivity in bitumen. *Energy & Fuels*, *25*(10), 4829-4835. http://dx.doi.org/10.1021/ef2009265.

Fan, T., and Buckley, J. S. (2002). Rapid and accurate SARA analysis of medium gravity crude oils. *Energy & Fuels*, *16*(6), 1571-1575. http://dx.doi.org/10.1021/ef0201228.

Fathi, M. M., and Pereira-Almao, P. (2011). Catalytic aquaprocessing of Arab light vacuum residue via short space times. *Energy & Fuels*, *25*(11), 4867-4877. http://dx.doi.org/10.1021/ef200936k.

Fernø, M. A., Torsvik, M., Haugland, S., and Graue, A. (2010). Dynamic laboratory wettability alteration. *Energy & Fuels*, *24*(7), 3950-3958. http://dx.doi.org/10.1021/ef1001716.

Filgueiras, P. R., Sad, C. M., Loureiro, A. R., Santos, M. F., Castro, E. V., Dias, J. C., and Poppi, R. J. (2014). Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel*, *116*, 123-130. http://dx.doi.org/10.1016/j.fuel.2013.07.122.

Flego, C., and Zannoni, C. (2012). Direct Insertion Probe–Mass Spectrometry (DIP–MS) Maps and Multivariate Analysis in the Characterization of Crude Oils. *Energy & Fuels*, *27*(1), 46-55. http://dx.doi.org/10.1021/ef301124s.

Freitas, L. S., Von Mühlen, C., Bortoluzzi, J. H., Zini, C. A., Fortuny, M., Dariva, C., and Caramão, E. B. (2009). Analysis of organic compounds of water-in-crude oil

emulsions separated by microwave heating using comprehensive two-dimensional gas chromatography and time-of-flight mass spectrometry. *Journal of Chromatography A*, *1216*(14), 2860-2865. http://dx.doi.org/10.1016/j.chroma.2008.09.076.

Goel, P., Bapat, S., Vyas, R., Tambe, A., and Tambe, S. S. (2015). Genetic programming based quantitative structure–retention relationships for the prediction of Kovats retention indices. *Journal of Chromatography A*, *1420*, 98-109. http://dx.doi.org/10.1016/j.chroma.2015.09.086.

Gonzalez, D. L., Hirasaki, G. J., Creek, J., and Chapman, W. G. (2007). Modeling of asphaltene precipitation due to changes in composition using the perturbed chain statistical associating fluid theory equation of state. *Energy & fuels*, *21*(3), 1231-1242. http://dx.doi.org/10.1021/ef060453a.

Gonzalez, D. L., Ting, P. D., Hirasaki, G. J., and Chapman, W. G. (2005). Prediction of asphaltene instability under gas injection with the PC-SAFT equation of state. *Energy & fuels*, *19*(4), 1230-1234. http://dx.doi.org/10.1021/ef049782y.

Gui, B., Yang, Q. Y., Wu, H. J., Zhang, X., and Lu, Y. (2010). Study of the effects of low-temperature oxidation on the chemical composition of a light crude oil. *Energy & fuels*, *24*(JANFEV), 1139-1145. http://dx.doi.org/10.1021/ef901056s.

Hannisdal, A., Ese, M. H., Hemmingsen, P. V., and Sjöblom, J. (2006). Particle-stabilized emulsions: effect of heavy crude oil components pre-adsorbed onto stabilizing solids. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, *276*(1), 45-58. http://dx.doi.org/10.1016/j.colsurfa.2005.10.011.

Hernandez, M. E., Vives, M. T., and Pasquali, J. (1983). Relationships among viscosity, composition, and temperature for two groups of heavy crudes from the eastern Venezuelan basin. *Organic Geochemistry*, *4*(3), 173-178. http://dx.doi.org/10.1016/0146-6380(83)90038-4.

Hinkle, A., Shin, E. J., Liberatore, M. W., Herring, A. M., and Batzle, M. (2008). Correlating the chemical and physical properties of a set of heavy oils from around the world. *Fuel*, *87*(13), 3065-3070. http://dx.doi.org/10.1016/j.fuel.2008.04.018.

Hoshyargar, V., and Ashrafizadeh, S. N. (2013). Optimization of flow parameters of heavy crude oil-in-water emulsions through pipelines. *Industrial & Engineering Chemistry Research*, *52*(4), 1600-1611. http://dx.doi.org/10.1021/ie302993m.

Hsu, C. S., and Robinson, P. (Eds.). (2007). *Practical Advances in Petroleum Processing* (Vol. 1). Springer Science & Business Media, New York.

International Council on Clean Transportation (ICCT). 2011. An Introduction to Petroleum Refining and the Production of Ultra Low Sulfur Gasoline and Diesel Fuel. Report, October 2011,http://www.theicct.org/sites/default/files/publications/ICCT05Refining_Tutorial_FINAL_R1.pdf (accessed 20 November 2015).

Islas-Flores, C. A., Buenrostro-Gonzalez, E., and Lira-Galeana, C. (2005). Comparisons between open column chromatography and HPLC SARA fractionations in petroleum. *Energy & Fuels*, *19*(5), 2080-2088. http://dx.doi.org/10.1021/ef050148.

Islas-Flores, C. A., Buenrostro-Gonzalez, E., and Lira-Galeana, C. (2006). Fractionation of petroleum resins by normal and reverse phase liquid chromatography. *Fuel*, *85*(12), 1842-1850. http://dx.doi.org/10.1016/j.fuel.2006.02.007.

Jafari Behbahani, T., Ghotbi, C., Taghikhani, V., and Shahrabadi, A. (2012). Investigation on asphaltene deposition mechanisms during CO_2 flooding processes in porous media: A novel experimental study and a modified model based on multilayer theory for asphaltene adsorption. *Energy & Fuels*, *26*(8), 5080-5091. http://dx.doi.org/10.1021/ef300647f.

Juyal, P., Ho, V., Yen, A., and Allenson, S. J. (2012). Reversibility of asphaltene flocculation with chemicals. *Energy & Fuels*, *26*(5), 2631-2640. http://dx.doi.org/10.1021/ef201389e.

Juyal, P., Yen, A. T., Rodgers, R. P., Allenson, S., Wang, J., and Creek, J. (2009). Compositional variations between precipitated and organic solid deposition control (OSDC) asphaltenes and the effect of inhibitors on deposition by electrospray ionization fourier transform ion cyclotron resonance (FT-ICR) mass

spectrometry. *Energy & Fuels*, *24*(4), 2320-2326. http://dx.doi.org/10.1021/ef900959r.

Kazempour, M., Manrique, E. J., Alvarado, V., Zhang, J., and Lantz, M. (2013). Role of active clays on alkaline–surfactant–polymer formulation performance in sandstone formations. *Fuel*, *104*, 593-606. http://dx.doi.org/10.1016/j.fuel.2012.04.034.

Khalil De Oliveira, M. C., and Gonçalves, M. A. (2012). An effort to establish correlations between Brazilian crude oils properties and flow assurance related issues. *Energy & Fuels*, *26*(9), 5689-5701. http://dx.doi.org/10.1021/ef300650k.

Khansari, Z., Gates, I. D., and Mahinpey, N. (2012). Detailed study of low-temperature oxidation of an Alaska heavy oil. *Energy & Fuels*, *26*(3), 1592-1597. http://dx.doi.org/10.1021/ef201828p.

Kök, M. V., Karacan, Ö., and Pamir, R. (1998). Kinetic analysis of oxidation behavior of crude oil SARA constituents. *Energy & fuels*, *12*(3), 580-588. http://dx.doi.org/10.1021/ef970173i.

Kord, S., and Ayatollahi, S. (2012). Asphaltene precipitation in live crude oil during natural depletion: Experimental investigation and modeling. *Fluid Phase Equilibria*, *336*, 63-70. . http://dx.doi.org/10.1016/j.fluid.2012.05.028.

Kord, S., Miri, R., Ayatollahi, S., and Escrochi, M. (2012). Asphaltene deposition in carbonate rocks: experimental investigation and numerical simulation. *Energy & Fuels*, *26*(10), 6186-6199.http://dx.doi.org/10.1021/ef300692e.

Koza, J. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, Massachusetts: MIT Press.

Kraiwattanawong, K., Fogler, H. S., Gharfeh, S. G., Singh, P., Thomason, W. H., and Chavadej, S. (2009). Effect of asphaltene dispersants on aggregate size distribution and growth. *Energy & Fuels*, *23*(3), 1575-1582. http://dx.doi.org/10.1021/ef800706c.

Kumar, K., Dao, E., and Mohanty, K. K. (2005). AFM study of mineral wettability with reservoir oils. *Journal of colloid and interface science*, *289*(1), 206-217. http://dx.doi.org/10.1016/j.jcis.2005.03.030.

Lammoglia, T., and de Souza Filho, C. R. (2011). Spectroscopic characterization of oils yielded from Brazilian offshore basins: Potential applications of remote sensing. *Remote Sensing of Environment*, *115*(10), 2525-2535. http://dx.doi.org/10.1016/j.rse.2011.04.038.

León, O., Contreras, E., Rogel, E., Dambakli, G., Acevedo, S., Carbognani, L., and Espidel, J. (2002). Adsorption of native resins on asphaltene particles: a correlation between adsorption and activity. *Langmuir*, *18*(13), 5106-5112. http://dx.doi.org/10.1021/la011394q.

Linan, L. Z., Lopes, M. S., Wolf Maciel, M. R., Nascimento Lima, N. M., Filho, R. M., Embiruçu, M., and Medina, L. C. (2010). Molecular distillation of petroleum residues and physical-chemical characterization of distillate cuts obtained in the process. *Journal of Chemical & Engineering Data*, *55*(9), 3068-3076. http://dx.doi.org/10.1021/je9010807.

Long, J., Shen, B., Ling, H., Zhao, J., and Lu, J. (2011). Novel solvent deasphalting process by vacuum residue blending with coal tar. *Industrial & Engineering Chemistry Research*, *50*(19), 11259-11269. http://dx.doi.org/10.1021/ie2004169.

Mahinpey, N., Murugan, P., and Mani, T. (2010). Comparative kinetics and thermal behavior: the study of crude oils derived from Fosterton and Neilburg fields of Saskatchewan. *Energy & Fuels*, *24*(3), 1640-1645. . http://dx.doi.org/10.1021/ef901470j.

Marcano, F., Flores, R., Chirinos, J., and Ranaudo, M. A. (2011). Distribution of Ni and V in A1 and A2 asphaltene fractions in stable and unstable Venezuelan crude oils. *Energy & Fuels*, *25*(5), 2137-2141. http://dx.doi.org/10.1021/ef200189m.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, *11*(2), 431-441. http://dx.doi.org/10.1137/0111030.

Marques, J., Maget, S., and Verstraete, J. J. (2011). Improvement of ebullated-bed effluent stability at high conversion operation. *Energy & Fuels*, *25*(9), 3867-3874. http://dx.doi.org/10.1021/ef2006047.

Mena-Cervantes, V. Y., Hernández-Altamirano, R., Buenrostro-González, E., Beltrán, H. I., and Zamudio-Rivera, L. S. (2011). Tin and silicon phthalocyanines molecularly engineered as traceable stabilizers of asphaltenes. *Energy and Fuels*, *25*(1), 224-231. http://dx.doi.org/10.1021/ef101023r.

Mendoza de la Cruz, J. L., Argüelles-Vivas, F. J., Matías-Pérez, V., Durán-Valencia, C. D. L. A., and López-Ramírez, S. (2009). Asphaltene-induced precipitation and deposition during pressure depletion on a porous medium: an experimental investigation and modeling approach. *Energy & Fuels*, *23*(11), 5611-5625. http://dx.doi.org/10.1021/ef9006142.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006, August). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935-940). ACM. http://dx.doi.org/10.1145/1150402.1150531.

Mohammadi, A. H., Eslamimanesh, A., and Richon, D. (2012). Monodisperse thermodynamic model based on Chemical+ Flory–hüggins polymer solution theories for predicting asphaltene precipitation. *Industrial & Engineering Chemistry Research*, *51*(10), 4041-4055. http://dx.doi.org/10.1021/ie202737p.

Molina, D., Uribe, U. N., and Murgich, J. (2010). Correlations between SARA fractions and physicochemical properties with 1 H NMR spectra of vacuum residues from Colombian crude oils. *Fuel*, *89*(1), 185-192. http://dx.doi.org/10.1016/j.fuel.2009.07.021.

Muhammad, A., and de Vasconcellos Azeredo, R. B. (2014). 1 H NMR spectroscopy and low-field relaxometry for predicting viscosity and API gravity of Brazilian crude oils–A comparative study. *Fuel*, *130*, 126-134. http://dx.doi.org/10.1016/j.fuel.2014.04.026.

Nasr-El-Din, H. A., and Taylor, K. C. (1992). Dynamic interfacial tension of crude oil/alkali/surfactant systems. *Colloids and surfaces*, *66*(1), 23-37. http://dx.doi.org/10.1016/0166-6622(92)80117-K.

Nokandeh, N. R., Khisvand, M., and Naseri, A. (2012). An artificial neural network approach to predict asphaltene deposition test result. *Fluid Phase Equilibria*, *329*, 32-41. http://dx.doi.org/10.1016/j.fluid.2012.06.001.

Ocanto, O., Marcano, F., Castillo, J., Fernández, A., Caetano, M., Chirinos, J., and Ranaudo, M. A. (2009). Influence of experimental parameters on the determination of asphaltenes flocculation onset by the titration method. *Energy & Fuels*, *23*(6), 3039-3044. http://dx.doi.org/10.1021/ef900106f.

Pacheco, V. F., Spinelli, L., Lucas, E. F., and Mansur, C. R. (2011). Destabilization of petroleum emulsions: evaluation of the influence of the solvent on additives. *Energy & Fuels*, *25*(4), 1659-1666. http://dx.doi.org/10.1021/ef101769e.

Pantoja, P. A., Lopez-Gejo, J., Le Roux, G. A., Quina, F. H., and Nascimento, C. A. (2011). Prediction of crude oil properties and chemical composition by means of steady-state and time-resolved fluorescence. *Energy & Fuels*, *25*(8), 3598-3604. http://dx.doi.org/10.1021/ef200567x.

Panuganti, S. R., Vargas, F. M., and Chapman, W. G. (2011). Modeling reservoir connectivity and tar mat using gravity-induced asphaltene compositional grading. *Energy & Fuels*, *26*(5), 2548-2557. http://dx.doi.org/10.1021/ef201280d.

Pasquini, C., and Bueno, A. F. (2007). Characterization of petroleum using near-infrared spectroscopy: Quantitative modeling for the true boiling point curve and specific gravity. *Fuel*, *86*(12), 1927-1934. http://dx.doi.org/10.1016/j.fuel.2006.12.026.

Peramanu, S., Singh, C., Agrawala, M., and Yarranton, H. W. (2001). Investigation on the reversibility of asphaltene precipitation. *Energy & Fuels*, *15*(4), 910-917. http://dx.doi.org/10.1021/ef010002k.

Poindexter, M. K., and Marsh, S. C. (2009). Inorganic Solid Content Governs Water-in-Crude Oil Emulsion Stability Predictions†. *Energy & Fuels*, *23*(3), 1258-1268. http://dx.doi.org/10.1021/ef800652n.

Poli, R., Langdon, W., and Mcphee, N. 2008. A Field Guide to Genetic Programming. http://www0.cs.ucl.ac.uk/staff/wlangdon/ftp/papers/poli08_fieldguide.pdf (accessed 20 November 2015).

Poteau, S., Argillier, J. F., Langevin, D., Pincet, F., and Perez, E. (2005). Influence of pH on stability and dynamic properties of asphaltenes and other amphiphilic molecules at the oil-water interface. *Energy & Fuels*, *19*(4), 1337-1341. http://dx.doi.org/10.1021/ef0497560.

RapidMiner, 2007. RapidMiner Studio: Modern Software for Lightning Fast Predictive Analytics, https://rapidminer.com/products/studio (accessed 20 November 2015).

Rocha, E. R., Lopes, M. S., Wolf Maciel, M. R., Maciel Filho, R., and Medina, L. C. (2013). Fractionation and Characterization of a Petroleum Residue by Molecular Distillation Process. *Industrial & Engineering Chemistry Research*, *52*(44), 15488-15493. http://dx.doi.org/10.1021/ie400669k.

Rogel, E., Leon, O., Contreras, E., Carbognani, L., Torres, G., Espidel, J., and Zambrano, A. (2003). Assessment of asphaltene stability in crude oils using conventional techniques. *Energy & Fuels*, *17*(6), 1583-1590. http://dx.doi.org/10.1021/ef0301046.

Rose, J. L., Monnery, W. D., Chong, K., and Svrcek, W. Y. (2001). Experimental data for the extraction of Peace River bitumen using supercritical ethane. *Fuel*, *80*(8), 1101-1110. http://dx.doi.org/10.1016/S0016-2361(00)00175-7.

Sanchez-Minero, F., Ancheyta, J., Silva-Oliver, G., and Flores-Valle, S. (2013). Predicting SARA composition of crude oil by means of NMR. *Fuel*, *110*, 318-321. http://dx.doi.org/10.1016/j.fuel.2012.10.027.

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, *324*(5923), 81-85. http://dx.doi.org/10.1126/science.1165893.

Speight, J. G., and O¨ zu¨m, B. (2002). *Petroleum Refining Processes*. Marcel Dekker, New York.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, *87*(2), 245.

Strubinger, A., Ehrmann, U., and León, V. (2012). Using the gas pycnometer to determine API gravity in crude oils and blends. *Energy & Fuels*, *26*(11), 6863-6868. http://dx.doi.org/10.1021/ef301193x.

Suatoni, J. C., and Swab, R. E. (1975). Rapid hydrocarbon group-type analysis by high performance liquid chromatography. *Journal of Chromatographic Science*, *13*(8), 361-366. http://dx.doi.org/10.1093/chromsci/13.8.361.

Subramanian, M., and Hanson, F. V. (1998). Supercritical fluid extraction of bitumens from Utah oil sands. *Fuel Processing Technology*, *55*(1), 35-53. http://dx.doi.org/10.1016/S0378-3820(97)00076-3.

Tambe, S. S., Kulkarni, B. D., and Deshpande, P. B. 1996. *Elements of Artificial Neural Networks with Selected Applications in Chemical Engineering, and Chemical & Biological Sciences*. Louisville, Kentucky: Simulation & Advanced Controls.

Tang, Y., Huang, Y., Ellis, G. S., Wang, Y., Kralert, P. G., Gillaizeau, B., and Hwang, R. (2005). A kinetic model for thermally induced hydrogen and carbon isotope fractionation of individual n-alkanes in crude oil. *Geochimica et Cosmochimica Acta*, *69*(18), 4505-4520. http://dx.doi.org/10.1016/j.gca.2004.12.026.

Tharanivasan, A. K., Svrcek, W. Y., Yarranton, H. W., Taylor, S. D., Merino-Garcia, D., and Rahimi, P. M. (2009). Measurement and modeling of asphaltene precipitation from crude oil blends. *Energy & Fuels*, *23*(8), 3971-3980. http://dx.doi.org/10.1021/ef900150p.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik, V. (1996). Structure of Statistical Learning Theory. In *Computational Learning and Probabilistic Reasoning*. Gammerman, A. (ed.), John Wiley and Sons, New York.

Vapnik, V., Golowich, S. E., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. In *Advances in neural information processing systems*; Mozer, M., Jordan, M., and Petsche, T. (Eds.), Cambridge, Massachusetts: MIT Press, pp 281-287.

Vela, J., Cebolla, V. L., Membrado, L., and Andrés, J. M. (1995). Quantitative hydrocarbon group type analysis of petroleum hydro conversion products using an improved TLC-FID system. *Journal of Chromatographic Science*, *33*(8), 417-425. http://dx.doi.org/10.1093/chromsci/33.8.417.

Wang, J., and Buckley, J. S. (2003). Asphaltene stability in crude oil and aromatic solvents the influence of oil composition. *Energy & Fuels*, *17*(6), 1445-1451. http://dx.doi.org/10.1021/ef030030y.

World Data Base. 2005. Environmental Technology Centre, Canada, www.oilproduction.net/files/Oil_Data_Base.zip (accessed 20 November 2015).

Zhang, L., Li, S., Han, L., Sun, X., Xu, Z., Shi, Q., and Zhao, S. (2013). Coking Reactivity of Laboratory-Scale Unit for Two Heavy Petroleum and Their Supercritical Fluid Extraction Subfractions. *Industrial & Engineering Chemistry Research*, *52*(16), 5593-5600. http://dx.doi.org/10.1021/ie302891b.

Zurada, J. (1992). *Introduction to Artificial Neural Systems*. St. Paul, Minnesota: West Publishing.

Chapter 7

The Removal of Arsenite [As(III)] and Arsenate [As(V)] Ions from Wastewater Using TFA and TAFA Resins: Computational Intelligence Based Reaction Modeling and Optimization

ABSTRACT

Being significantly toxic, removal of arsenic forms an important ingredient of the drinking- and waste-water treatment. Tannin is a polyphenol-rich substrate that adsorptively binds to the multivalent metal ions. In this study, tannin-formaldehyde (*TFA*) *and tannin-aniline-formaldehyde* (*TAFA*) *resins were synthesized and successfully used for an adsorptive removal of arsenite* [*As*(*III*)] *and arsenate* [*As*(*V*)] *ions from the contaminated water. Next, a computational intelligence* (*CI*) *based hybrid strategy was used to model and optimize the resin-based adsorption of As*(*III*) *and As*(*V*) *ions for securing optimal reaction conditions. This strategy first uses an exclusively reaction data driven modeling method, namely, genetic programming* (*GP*) *to predict the extent* (*%*) *of As*(*III*)*/As*(*V*) *adsorbed on TFA and TAFA resins. Next, the input space of the GP-based models representing reaction condition variables/parameters was optimized using genetic algorithm* (*GA*) *method; the objective of this optimization was to maximize the adsorption of As*(*III*) *and As*(*V*) *ions on the two resins. Finally, the sets of optimal reaction conditions given by GP-GA hybrid modeling-optimization method were verified experimentally, which indicate that the optimized conditions have lead to 0.3% and 1.3% increase in the adsorption of As*(*III*) *and As*(*V*) *ions on TFA resin. More significantly, the optimized conditions have increased the adsorption of As*(*III*) *and As*(*V*) *on TAFA resin by 3.02% and 12.77%, respectively. The GP-GA based strategy introduced here can be gainfully utilized for modeling and optimization of similar type of contaminant-removal processes.*

7.1 INTRODUCTION

Presence of toxic materials is a serious problem requiring an effective solution during the management and treatment of water and wastewater. Globally, one of the most toxic metals, namely, arsenic forms a common contaminant of the ground water, which is an important source of the drinking water (Ng et al., 2003). Depending upon arsenic's oxidation state, its toxicity varies significantly. In the non-processed natural drinking water, it is mostly found as arsenite [As(III)] and arsenate [As(V)] (Kandu and Gupta, 2006). Among these, As(III) is sixty times more toxic than As(V) (Kandu et al., 2004). According to World Health Organization (WHO), and United States' Environment Protection Agency (USEPA), the maximum allowed concentration of arsenic in the drinking water is 0.01 mg/L (WHO, 2004; USEPA, 1999). However, in most regions of the world arsenic concentration exceeds that limit by many folds. Mohan and Pittman (2007) have critically reviewed removal of arsenic from water and wastewater using various adsorbents.

Tannins, which are available widely, can be effective agents for the water treatment in developing countries. These occur in nature as a biomass containing multiple hydroxyl groups and exhibiting a specific affinity towards metal ions. Thus, tannins are potentially effective and efficient adsorbents for the recovery of metal ions. Their disadvantage, however, is that being water-soluble; they can get easily leached by water when used directly as an adsorbent for the recovery of metals from the aqueous systems. A number of attempts involving immobilization of tannins has been made to overcome the drawback alluded to above (Liao et al., 2004). Makeswari et al. (2014) synthesized a novel tannin gel adsorbent from the leaves of *Ricinus Communis* for removing chromium(VI) ions. A bio-adsorbent from the tannin immobilized collagen/cellulose has been synthesized by Zhang et al. (2015) for the adsorption of lead(II). Shirato and Kamei (1994) have patented synthesis of insoluble tannin; it is prepared by dissolving a hydrolysable tannin powder in aqueous ammonia and the resulting mixture is treated with a formaldehyde solution to form a precipitate. This is then subjected to the treatment of a mineral acid. The resulting polymer is used for the processing of waste liquids and recovery of heavy metals. A method for the preparation of an insoluble tannin adsorbent and the adsorption of nuclear fuel material, and iron ions thereof, has also been patented (Shirato and Kamei, 1994).

In an earlier study, Mulani et al. (2014) synthesized and characterized tannin-formaldehyde and tannin-aniline-formaldehyde resins; they also studied the adsorption kinetics of arsenic using the said two resins. Specifically, Mulani et al. (2014) investigated the effect of influential parameters such as pH and contact time on the kinetics of arsenic adsorption and desorption. The objective of the present study is twofold: (a) development of models predicting the extent of As(III)/As(V) adsorption on the tannin-formaldehyde (TFA) and tannin-aniline-formaldehyde (TAFA) resins, and (b) obtaining the optimal resin composition and reaction pH magnitudes leading to maximum adsorption of the stated metalloid ions.

In order to optimize the resin-based As(III)/As(V) removal reactions and thereby obtain the conditions resulting in the maximum adsorption of As(III) and As(V) ions, it is necessary to develop the respective reaction models. There exist two principal methods, namely, *phenomenological* and *empirical*, for modeling the stated adsorption reactions. Both these approaches possess significant difficulties, which are detailed in Chapter 1 (section 1.3). The difficulties encountered in the phenomenological and empirical (essentially regression-based) reaction modeling requires exploration of alternative nonlinear reaction modeling strategies.

Artificial neural networks (ANNs) (see, for example, Bishop, 1994; Zurada, 1992; Tambe et al., 1996) and support vector regression (SVR) (Vapnik, 1995; Zaid, 2012) are computational intelligence (CI) based exclusively data-driven nonlinear modeling formalisms; these have been used widely as alternatives to the regression based modeling. In addition to ANNs and SVR, the field of computational intelligence comprises a novel data-driven modeling strategy, namely genetic programming (GP). There exist a number of studies in chemistry and chemical engineering wherein the GP-based symbolic regression has been employed for developing data-driven predictive models (see, for example, Patil- Shinde et al.,2014; Goel et al.,2015; Pandey et al.,2015; Koç and Koç, 2015; Bahrami et al.,2016). It possesses several attractive characteristics, which are explained in Chapter 2 (section 2.2.2) as also by Vyas et al., 2015 and Verma et al., 2016. Due to its several attractive characteristics, in this study, GP has been employed first to develop models predicting the extent of As(III) and As(V) adsorption on the tannin-formaldehyde and tannin-aniline-formaldehyde resins. Next, the input space of the GP-based models

consisting of molar composition of the resin and reaction pH was optimized using the GA formalism with a view to maximize the extent of As(III)/As(V) adsorption by the resins. A detailed description of GA formalism (Holland, 1975; Goldberg, 1989; Deb, 1995) has been provided in Chapter 2, section 2.4.1. In the past, the GP-GA hybrid modeling-optimization strategy alluded to above has been employed in the optimization of glucose to gluconic acid fermentation (Cheema et al., 2002). There also exist studies in chemical engineering/technology wherein two other CI-based strategies namely ANNs and SVR are individually combined with the GA optimization method to formulate ANN-GA (Nandi et al., 2001; Huang et al.,2003; Rao et al.,2009) and SVR-GA (Nandi et al.,2004; Wu et al.,2009) hybrid modeling-optimization strategies, respectively.

In the present investigation, following four case studies have been performed using the hybrid GP-GA strategy.

- Case study I: Modeling and optimization of adsorption of As(III) on TFA resin
- Case study II: Modeling and optimization of adsorption of As(V) on TFA resin
- Case study III: Modeling and optimization of adsorption of As(III) on TAFA resin
- Case study IV: Modeling and optimization of adsorption of As(V) on TAFA resin

The inputs and outputs pertaining to the four GP-based models developed in this study are given in Table 7.1. In case studies I and II, experiments were conducted at a fixed tannin concentration and, therefore, it is not considered as a model input. In case study III, tannin and aniline concentrations are not included as model inputs since experiments were conducted at fixed tannin and aniline concentrations (Mulani et al., 2014). In all the four case studies, several sets of GA-optimized reaction conditions that were expected to result in the improved adsorption of As(III)/As(V) ions, were obtained. The overall best sets of conditions obtained thereby were subjected to experimental verification. Results of this experimentation indicate that the optimized reaction conditions have indeed succeeded in improving the extent of As(III)/As(V) adsorption on the TFA and TAFA resins.

Table 7.1: Inputs and the output of four GP-based models

Model	Model inputs (reaction operating variables/parameters)	Model output (extent of adsorption)
I	Moles of formaldehyde (x_1) and ammonia (x_2), and reaction *pH* (x_3)	Adsorption (%) (y) of As(III) on TFA resin
II	Moles of formaldehyde (x_1) and ammonia (x_2), and reaction *pH* (x_3)	Adsorption (%) (y) of As(V) on TFA resin
III	Moles of formaldehyde (x_1) and ammonia (x_2), and reaction *pH* (x_3)	Adsorption (%) (y) of As(III) on TAFA resin
IV	Moles of tannin (x_1), aniline (x_2) and formaldehyde (x_3), and reaction *pH* (x_4)	Adsorption (%) (y) of As(V) on TAFA resin

This chapter is organized as follows. The details of resin preparation and characterization as also the resin-based As(III) and As(V) adsorption experimentation, are provided in the " Materials and methods " (section 7.2). Next, the "Results and Discussion" (section 7.3) first describes the outcome of the adsorption experiments, which is followed by the presentation of the results and discussion pertaining to the four modeling-optimization case studies (section 7.3.2). Section 7.3.3 provides results of the experimental validation of the overall optimum reaction conditions yielded by the GP-GA hybrid strategy. Finally, "Concluding Remarks" (section 7.4) summarize the principal findings of the study.

7.2 MATERIALS AND METHODS

Tannic acid (LOBA CHEMIE, Mumbai, India), aniline, ammonia (25 wt% solution, MERCK, India), formaldehyde (37 wt% solution, QUALIGENS, India), sodium As(III), and sodium arsenate (LOBA CHEMIE, Mumbai, India) were used without further purification and distillation.

7.2.1 Preparation of Tannin-Formaldehyde (TFA) Resin

The composition of various synthesized tannin-formaldehyde monomer resins is presented in Table 7.2. Here, 10–50 mL of 37% formaldehyde solution was added to 4g of commercial tannin powder and the resultant mixture was stirred for five minutes to ensure a uniform mixing. Depending on the desired composition, 20–40 mL ammonia solution (25 wt%) was added to the above-stated mixture with continuous stirring, and the brown precipitate formed thereby was kept at an ambient

temperature (25^{o}C) for fifteen days. This precipitate was neutralized with 10.8 N hydrochloric acid solutions and the resultant precipitate was filtered through Whatman no.2 filter paper and treated with 1.2 M hydrochloric acid for making it insoluble in the acidic and basic media. It was further washed with de-ionized water and dried at 80^{o}C to obtain an insoluble solid tannin resin.

Table 7.2: Monomer composition of tannin-formaldehyde (TFA) resins

Resin.	Tannin (g)	Formaldehyde (mL)	Ammonia (mL)
TFA 01	4.0	50	40
TFA 02	4.0	40	40
TFA 03	4.0	30	40
TFA 04	4.0	20	40
TFA 05	4.0	10	40
TFA 06	4.0	50	20
TFA 07	4.0	40	20
TFA 08	4.0	30	20
TFA 09	4.0	20	20
TFA 10	4.0	10	20

7.2.2 Preparation of Tannin-Aniline-Formaldehyde (TAFA) Resin

The procedure for the TFA synthesis was repeated to prepare the tannin-aniline-formaldehyde (TAFA) resins except that tannin was partially substituted with aniline. Three sets of TAFA resins were prepared using the stated procedure by varying the tannin: aniline ratio as listed in Tables 7.3 – 7.5.

Table 7.3: Monomer composition of tannin-aniline-formaldehyde resins [tannin: aniline ratio 3:1 (w/w)]

Resin	Tannin (g)	Aniline (g)	Formaldehyde (mL)	Ammonia (mL)
TAFA 01	3.0	1.0	50	40
TAFA 02	3.0	1.0	40	40
TAFA 03	3.0	1.0	30	40
TAFA 04	3.0	1.0	20	40
TAFA 05	3.0	1.0	10	40
TAFA 06	3.0	1.0	50	20
TAFA 07	3.0	1.0	40	20
TAFA 08	3.0	1.0	30	20
TAFA 09	3.0	1.0	20	20
TAFA 10	3.0	1.0	10	20

Table 7.4: Monomer composition of tannin-aniline-formaldehyde resins [tannin: aniline ratio 2:2 (w/w)]

Resin	Tannin (g)	Aniline (g)	Formaldehyde (mL)	Ammonia (mL)
TAFA 01	2.0	2.0	50	40
TAFA 02	2.0	2.0	40	40
TAFA 03	2.0	2.0	30	40
TAFA 04	2.0	2.0	20	40
TAFA 05	2.0	2.0	10	40

Table 7.5: Monomer composition of tannin-aniline-formaldehyde resins [tannin: aniline ratio 1:3 (w/w)]

Resin No.	Tannin (g)	Aniline (g)	Formaldehyde (mL)	Ammonia (mL)
TAFA 06	1.0	3.0	50	40
TAFA 07	1.0	3.0	40	40
TAFA 08	1.0	3.0	30	40
TAFA 09	1.0	3.0	20	40
TAFA 10	1.0	3.0	10	40

7.2.3 As(III)/As(V) adsorption on TFA and TAFA resins

Arsenic standards for [As(III)] and [As(V)] were prepared from $NaAsO_2$ and $Na_2HAsO_4.7H_2O$ [Loba Chemie Pvt, Ltd, Mumbai, India], respectively. The As(III) and As(V) stock solutions were prepared by dissolving 173.30 mg of sodium arsenite and 450 mg of sodium arsenate, respectively in 100 mL distilled water. The intermediate and secondary standards of arsenic solutions were prepared freshly for each experiment. The working solutions containing arsenic were prepared by dissolving an appropriate amount of arsenic from the stock solutions in de-ionized water. The efficiency of TFA and TAFA resins in removing As(III) and As(V) was studied at different pH magnitudes ranging between 2 and 10.

7.2.4 Adsorption Measurements

The experiments pertaining to the adsorption of As(III)/As(V) ions on TFA and TAFA resins were performed in a batch mode. The extent of adsorption of

As(III)/As(V) was measured as a function of time and pH under vigorous agitation. Max-uptake capacity of As(III) and As (V) is mg/gm of resin.

***Analytical method*:** The residual arsenic in a water sample was determined using the molybdenum blue method (Johnson and Pilson, 1972). It was utilized to estimate the individual concentrations of As(III) and As(V) in the treated water samples to assess the efficiency of the oxidation step and subsequent adsorption-based removal of arsenic. Spectrophotometric measurements were performed at 865 nm wavelength using an absorbance cell of 1 cm path length for the determination of arsenic. The calibration curves for the total arsenic were prepared using solutions containing As(III) and As(V).

***Molybdenum blue method*:** This method allows for the routine analysis of As(III) and As(V) by the spectroscopic measurement of arsenic-molybdenum complexes. Since water used in the experiment contained no or negligible amount of phosphate, the method was modified for the determination of As(III) and As(V) only (Johnson and Pilson, 1972). This method requires a mixed reagent, which was prepared as given below.

***Preparation of mixed reagent*:** Mix thoroughly a solution of 25 mL of 5N sulfuric acid and 7.5 mL of 0.032 M ammonium molybdate; add to it 15 mL of 0.1 M ascorbic acid solution (freshly prepared) followed by the addition of 2.5 mL of 0.0082 M potassium antimony tartrate solution with a thorough mixing post each addition. This reagent was prepared freshly each time.

The mixed reagent when added to an untreated aliquot of a sample containing As(V) ions, produces blue color due to the formation of arsenomolybdate complex. It may, however, be noted that As(III) does not form the said complex. Accordingly, the intensity of the color formed and, hence, the absorbance of the untreated aliquot are proportional to the concentration of As(V) present. For converting As(III) to As(V), potassium iodate was used as an oxidizing agent. Thus the absorbance of an oxidized aliquot of the sample is proportional to the total concentration of arsenic (i.e., As(III) + As(V)). The concentration of As(III) is then calculated as the difference between the concentration of total arsenic (As(III) + As (V)) and that of the As(V).

Two sets of solutions respectively containing As(III) and As(V) ions in the concentration range of 1–15 mg/L were prepared from their standard solutions. One

set was used for the oxidation treatment of the As(III) aliquots while the other set formed solutions of the untreated As(V). One mL of 1 N hydrochloric acid and 1 mL of 0.017 M potassium iodate were added successively to oxidize each of the aliquots containing As(III) with a thorough mixing after each addition. Ten minutes were allowed for the oxidation of As(III) to As(V) and, thereafter, 4 mL of the mixed reagent was added to each of the treated (oxidized) and untreated aliquots with a thorough mixing. After two hours, a blue colored arsenic-molybdenum complex was formed. The amount of complex formed is directly proportional to the arsenic concentration, which was determined as a function of the absorbance measured at 865 nm wave-length with a UV spectrophotometer. Blank samples were run twice using the above-described procedure along with the samples.

7.3 RESULTS AND DISCUSSION

7.3.1 Experimental

***Effect of pH*:** The waste water containing metal ions is acidic in nature. Accordingly, the effect of pH on the adsorption of As(III) and As(V) on TFA and TAFA resins was studied in the pH range of 2–10. The values of the reaction operating variables and the corresponding magnitudes of the adsorption (%) of As(III) and As(V) ions on the TFA and TAFA resins are listed in Appendix 7. A. It is observed that at lower pH values, the phenolic group (–OH) of the TFA resin gets protonated to higher extent, which results in a strong repulsion to the positively charged arsenic ion in the solution; such a situation is not favorable for As(III) removal (Mohan and Pittman, 2007; Dutré and Vandecasteele, 1998; Dambies et al., 2002; Arai et al., 2005; Pena et al., 2006). However, As(V) ions were best adsorbed on TFA resin in the pH range of 3–5. It can thus be inferred that in this pH range the metal anions follow the anion exchange mechanism and, accordingly, get adsorbed by releasing protons from the phenolic –OH groups of tannin (Onyango et al., 2003; Zhang et al., 2007). The experimental results also show that an increase in the pH magnitude does not result in a significant change in the adsorption (%) of As(III)/As(V) ions. This may be due to the hydroxyl group not being abundantly present on the surface of TFA and TAFA resins.

***Comparison with other Adsorbents*:** TFA and TAFA resins exhibit higher adsorption capacity for As (III) and As(V) ions compared to the adsorbents prepared

from the waste rice husk (Amin et al., 2006). It has been also found that TFA and TAFA resins possess lower adsorption capacity for As(V) ions when compared with that of, for example, ferrihydride, mesoporous alumina, and amorphous aluminum hydroxide (Anderson et al., 1976; Thirunavukkarasu et al., 2001; Kim et al., 2004).

7.3.2 GP-Based Adsorption Reaction Modeling and GA-Based Optimization of the Reaction Conditions

Pre-Processing of Adsorption Data

The experimental data reported in Appendix 7. A, pertaining to the adsorption of As(III) and As(V) ions on TFA and TAFA resins were used to develop four GP-based models described below in case studies I, II, III, and IV, respectively. From the data, it is seen that there exists an order of magnitude difference between some of the reaction condition variables and, therefore, pre-processing of the input and output data was carried out through a normalization scheme. The normalized input variables were obtained as follows:

$$\hat{x}_i^j = \frac{x_i^j - \overline{x}_i}{\sigma_{x_i}};\ j = 1, 2, \ldots,\ N_\mathrm{p};\quad i = 1, 2, \ldots, I. \tag{7.1}$$

where, N_p represents the number of patterns in the example data set (N_p = 40 in all four case studies); I refers to the dimensionality of the input space (I = 3 for case studies I to III and = 4 for case study IV); x_i^j represents the i^{th} un-normalized input variable of j^{th} pattern; $\hat{x}_i^j$ denotes the normal score (standardized variable) pertaining to the i^{th} input variable of the j^{th} pattern/vector, and $\overline{x}_i$ and σ_{x_i}, respectively refer to the mean and standard deviation values of the i^{th} input variable. Similar to the model inputs, the outputs in all the four case studies were normalized as follows:

$$\hat{y}^j = \frac{y^j - \bar{y}}{\sigma_y};\ \ j = 1, 2, \ldots, N_\mathrm{p} \tag{7.2}$$

where, $\hat{y}^j$, denotes the normal score (standardized variable) pertaining to the j^{th} output pattern; y^j refers to the j^{th} output value, and $\bar{y}$ and σ_y, respectively refer to the mean and the standard deviation of the N_p number of outputs. The mean and standard deviation values used in the above-described normalization procedure are listed in Table 7.6.

For developing GP-based models possessing good prediction accuracy and generalization ability, each of the four experimental data sets, listed in Appendix 7.A

(Tables 7.A.1- 7.A.4) was split randomly in 75:25 ratio into training (30 patterns) and test (10 patterns) sets. The training set was used for developing a GP-based model while the test set data was used in testing the generalization ability of the trained model. In Appendix 7.A (Tables 7.A.1- 7.A.4), the test set data are marked using "*" symbol.

Table 7.6: Mean and standard deviation magnitudes in respect of inputs $\{x_i\}$ and the output {y} of four GP-based models

Case Study	Model inputs		Model output	
	Mean	Standard deviation	Mean (%)	Standard deviation (%)
I	$\bar{x}_1$ = 0.323 moles, $\bar{x}_2$ = 0.477 moles, $\bar{x}_3$ = 5.6	σ_{x_1}= 0.164 moles, σ_{x_2} = 0.144 moles, σ_{x_3} =3.045	$\bar{y}$ =74.925	σ_y= 16.057
II	$\bar{x}_1$= 0.323 moles, $\bar{x}_2$= 0.477 moles, $\bar{x}_3$= 6.6	σ_{x_1}= 0.164 moles, σ_{x_2} = 0.144 moles, σ_{x_3}= 3.112	$\bar{y}$ = 86.572	σ_y=15.728
III	$\bar{x}_1$ = 0.351 moles, $\bar{x}_2$ = 0.433 moles, $\bar{x}_3$ = 5.7	σ_{x_1}= 0.182 moles, σ_{x_2} = 0.148 moles, σ_{x_3}= 2.96	$\bar{y}$ = 74.585	σ_y=11.141
IV	$\bar{x}_1$= 0.000881 moles, $\bar{x}_2$ = 0.027 moles, $\bar{x}_3$ = 0.366 moles, $\bar{x}_4$ = 7	σ_{x_1}= 0.000298 moles, σ_{x_2}= 0.00542 moles, σ_{x_3}= 0.179 moles, σ_{x_4}= 3.097	$\bar{y}$= 71.684	σ_y= 7.348

GP-based Modeling

A detailed description of GP (Koza, 1992) formalism is given in Chapter 2 (section 2.2.2). The GP-based models were developed using *Eureqa Formulize* software package (Schmidt and Lipson, 2009) that has been optimized to construct parsimonious (i.e. with lower complexity) expressions possessing good generalization ability. This software offers to its users a number of options for preprocessing of the example input-output data and generation of candidate solutions. In all the case studies, these options were rigorously and systematically explored with the objective of securing models possessing high As(III)/As(V) adsorption prediction accuracy and generalization capability. A set containing five basic arithmetic operators, namely, *addition*, *subtraction*, *multiplication*, *division*, and *exponentiation*, was used in the

generation of candidate expressions. To secure an overall optimal data-fitting model, the GP procedure was repeated numerous times by employing different seed expressions and random number generator seed values. In each repeated run, the GP algorithm searched a different mathematical expression. The prediction accuracy and the generalization performance of a GP-based model was evaluated in terms of *coefficient of correlation* (*CC*) and *root mean squared error* (*RMSE*) between the experimental (target) and the corresponding model-predicted values of the adsorption (%) of As(III)/As(V) ions on a particular resin. These two statistical measures were evaluated separately for the training and test data sets. The overall best GP-model was selected on the basis of their high and comparable magnitudes of *CC* and low and comparable values of *RMSE* in respect of both the training and test set data.

Case Study-I: *GP-based Modeling and GA-based Optimization of Adsorption of As(III) on TFA Resin*

The input space of the GP-based model-I consists of three reaction operating variables, viz. molar concentrations of formaldehyde (x_1) and ammonia (x_2), and solution pH (x_3). The training and test data sets used in constructing this model are listed in Appendix 7.A (Table 7.A.1). The overall best GP-based model (*GP_Model-I*) relating the three normalized inputs ($\hat{x}_1, \hat{x}_2$ and $\hat{x}_3$) to the output ($\hat{y}$) predicting the normalized value of the adsorption (%) of As(III) on the TFA resin is given as:

$$\hat{y} = 0.6714\,\hat{x}_1 + \frac{0.2495}{1.05 + 1.05\hat{x}_3 + 0.03351\hat{x}_1 + \hat{x}_3^2 - \hat{x}_2} - 0.2767\hat{x}_1^2\hat{x}_2 - 0.6152 \qquad (7.3)$$

The predictions of *GP_Model-I* have yielded high and comparable magnitudes of the *coefficient of correlation* (CC_{trn} = 0.957; CC_{tst} = 0.949) and low and comparable values of the *root mean square error* ($RMSE_{\mathrm{trn}}$ = 4.820; $RMSE_{\mathrm{tst}}$ = 4.265) in respect of both the training and test set data. It thus clearly suggests that the model possesses reasonably good prediction accuracy and generalization capability. The parity plot of the experimental *y* values and those predicted by the GP-based model-I (obtained by de-normalizing $\hat{y}$ values) has been presented in Figure 7.1. In this plot, it is noticed that there exists a good match between the experimental and model-predicted values of the adsorption (%) of As(III) on the TFA resin.

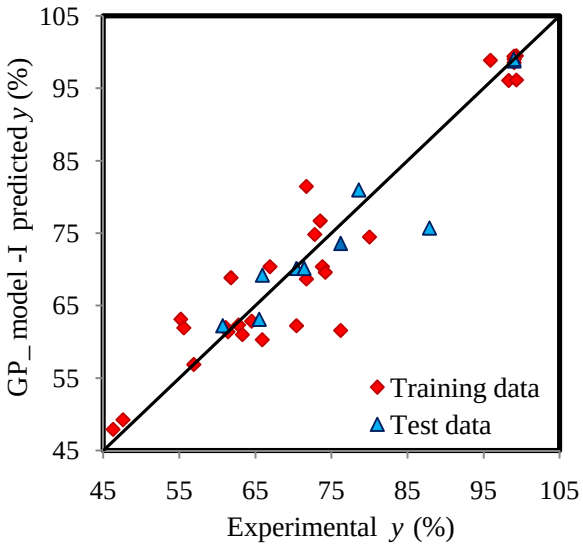


Figure 7.1: Parity plot of experimental versus GP-model predicted values of adsorption of As(III) on TFA resin (%) (y).

Having obtained the high performing *GP_ Model-I,* its input space consisting of the three reaction (decision) variables ($x_1 - x_3$) was optimized by employing GA formalism with the objective of maximizing the extent of adsorption (%) of As(III) on TFA resin. This optimization was performed using the *MendelSolve* (2016) software. While performing the said optimization, following values of the GA-specific parameters were used: population size (N_{A}) = 50, crossover rate (R_{cr}^{A}) = 1, mutation rate (R_{mut}^{A}) = 0.03125 and maximum number of generations (N_{gen}) = 100.

Using these parameter values several GA replicates were run by employing different random number generator seeds. The top three sets of optimized reaction operating variables obtained thereby are listed in Table 7.7.

Table 7.7: Optimized reaction variables given by GP-GA hybrid method for case study I

Optimal solution set	Optimized reaction variables			GA-maximized As(III) adsorption (%)	Experimentally validated As(III) adsorption (%)
	formaldehyde moles (x_1^{opt})	ammonia moles (x_2^{opt})	pH (x_3^{opt})		
1.	0.615	0.343	2.876	99.46	99.6
2.	0.544	0.577	4.997	91.52	91.2
3.	0.615	0.432	2.697	91.70	90.9

Case Study II: *GP-based Modeling and GA- based Optimization of Adsorption of As(V) on TFA Resin*

The input space of the GP-based model-II consists of the same three reaction operating variables/parameters considered in case study I. However, the model output describes adsorption (%) of As(V) on TFA resin. The training and test set data used in developing GP-based model-II are given in Appendix 7.A (Table 7.A.2). By implementing the GP procedure described in Chapter 2 (section 2.2.2), the following best data-fitting expression (*GP_Model-II*) was obtained:

$$\hat{y} = 1.686\,\hat{x}_1^4 - 0.8981\hat{x}_1^3 - 0.1817\,\hat{x}_1\hat{x}_2^2 - 4.037\hat{x}_1^2 - 0.1946\,\hat{x}_3 + 0.8573 \qquad (7.4)$$

The *CC* magnitudes in respect of the predictions by Eq. (7.4) for the training and test set data are CC_{trn} = 0.969 and CC_{tst}= 0.962, respectively; the corresponding *RMSE* magnitudes are $RMSE_{\text{trn}}$= 4.540 and $RMSE_{\text{tst}}$ = 6.565. The high (low) and comparable magnitudes of the training and test set *CCs* (*RMSEs*) clearly suggest an excellent prediction and generalization performance by *GP_Model-II*. Figure 7.2 shows the parity plot of the experimental versus *GP_Model-II* predicted magnitudes of the adsorption (%) (*y*) of As(V) on TFA resin. As can be noticed from this plot, the model predicted As(V) adsorption values exhibit a close match with their experimental counterparts; particularly in the region wherein *y* magnitudes are $\geq$ 70, the match is excellent.

The three inputs of *GP_Model-II* representing three reaction variables were optimized to secure their optimal values leading to maximization of the adsorption (%) of As(V) on TFA resin. The GA-specific parameters that yielded the top three sets of optimized reaction variables are: population size (N_{A}) = 55, crossover rate (R_{cr}^{A}) = 1, mutation rate (R_{mut}^{A}) = 0.03111 and maximum number of generations (N_{gen}) = 110. Using these parameter values several GA runs were conducted using each time a different value of the random number generator. Table 7.8 lists the top three optimized sets of the reaction conditions obtained using *MendelSolve* (2016) software. As can be seen, the best solution given by the hybrid GP-GA method is expected to result in the As(V) absorption of 99.8%, which is 2.8% higher than the maximum adsorption (%) of 97% obtained in experiments.

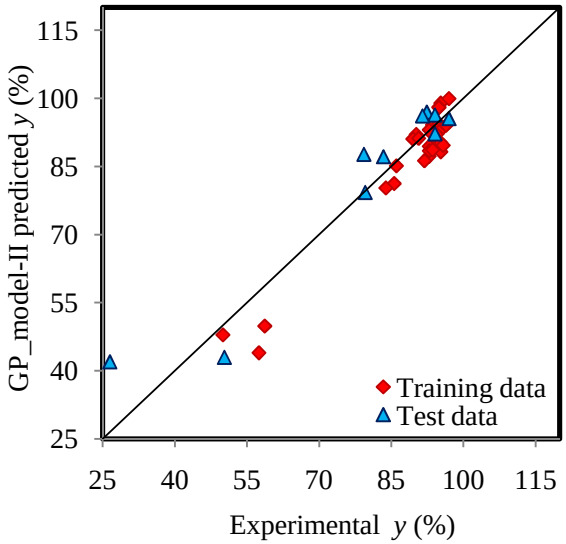


Figure 7.2: Parity plot of experimental versus GP-model predicted values of adsorption of As(V) on TFA resin (%) (*y*).

Table 7.8: Optimized reaction variables given by GP-GA hybrid method for case study II

Optimal solution set	Optimized reaction variables			GA-maximized As(V) adsorption (%)	Experimentally validated As(V) adsorption (%)
	formaldehyde moles (x_1^{opt})	ammonia moles (x_2^{opt})	pH (x_3^{opt})		
1.	0.123	0.294	2.008	99.809	98.3
2.	0.253	0.294	6.144	93.345	92.9
3.	0.123	0.426	5.778	91.539	90.7

Case Study III: GP-based Modeling and GA- based Optimization of Adsorption of As(III) on TAFA Resin

The input space of the GP-based Model-III predicting the adsorption of As(III) on TAFA resin contains three reaction operating variables, viz. molar concentrations of formaldehyde (x_1) and ammonia (x_2), and solution pH (x_3). The training and test set data used in developing this model are listed in Appendix 7.A (Table 7.A.3). The overall best model (*GP_model-III*) obtained by using Eureqa Formulize software is as follows:

$$\hat{y} = \hat{x}_1^4 - 2.7209\,\hat{x}_1^2 - 0.1831\,\hat{x}_3^2 + 0.3646\,\hat{x}_2\hat{x}_1^2 - 0.4914\,\hat{x}_3 - 0.2211\,\hat{x}_1\,\hat{x}_3 + 1.3232 \quad (7.5)$$

The *CC* and *RMSE* magnitudes in respect of the predictions made by Eq. (7.5) for the training and test set data are CC_{trn} = 0.859, CC_{tst} = 0.868, $RMSE_{\text{trn}}$ = 6.374 and $RMSE_{\text{tst}}$ = 5.593. These values indicate that *GP_model-III* is endowed with reasonably good prediction and generalization performance. This observation is also supported by the parity plot in Figure 7.3 depicting of the experimental versus *GP_ Model-III* predicted As(III) adsorption (*y*) values.

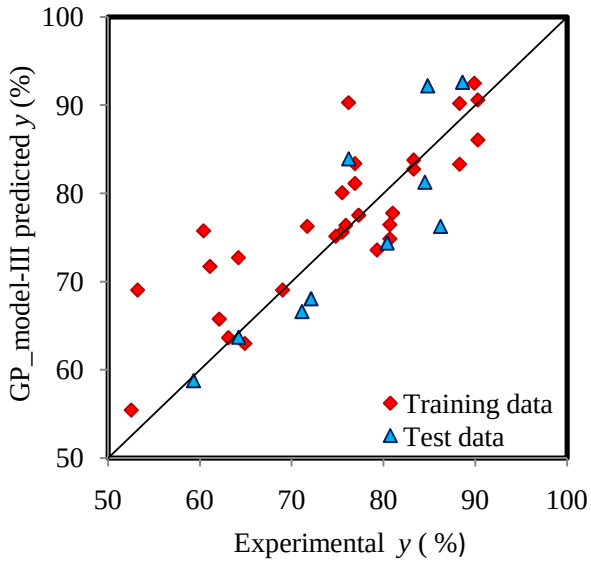


Figure 7.3: Parity plot of experimental versus GP-model predicted values of adsorption of As(III) on TAFA resin (%) (*y*).

Towards obtaining the optimal conditions leading to maximization of As(III) adsorption on TAFA resin, the input space of the *GP_Model-III* was subjected to GA-based optimization using *MendelSolve* (2016) software. The GA-specific parameters that yielded the three overall best sets of optimized reaction condition variables are: population size (N_{A}) = 50, crossover rate (R_{cr}^{A}) = 1, mutation rate (R_{mut}^{A}) = 0.03125 and maximum number of generations (N_{gen}) = 150. The top three GA-optimized sets of optimized reaction conditions, which are expected to maximize the adsorption of As(III) on TAFA resin are listed in Table 7.9. From the tabulated values, it is seen that the GA-searched best solution is capable of improving the extent of As(III) adsorption from 90.3% (best experimental As(III) adsorption value) to 94.07%.

Table 7.9: Optimized reaction variables given by GP-GA hybrid method for case study III

Optimal solution set	Optimized reaction variables			GA-maximized As(III) adsorption (%)	Experimentally validated As(III) adsorption (%)
	formaldehyde moles (x_1^{opt})	ammonia moles (x_2^{opt})	pH (x_3^{opt})		
1.	0.362	0.583	2.000	94.07	93.32
2.	0.615	0.586	3.001	90.59	89.73
3.	0.357	0.294	2.000	93.05	91.76

Case Study IV: GP-based Modeling and GA-based Optimization of Adsorption of As(V) on TAFA Resin

In this case study, the input space of the GP-based Model-IV contains four reaction operating variables, namely, molar concentrations of tannin (x_1), aniline (x_2), formaldehyde (x_3), and solution pH (x_4). The training and test set data used in developing this model are listed in Appendix 7.A (Table 7.A.4). The overall best model (*GP_model-IV*), predicting the extent (%) of [As(V)] adsorption on TAFA resin is as follows:

$$\hat{y} = 0.2459\ \hat{x}_3^2 - 0.2937\ \hat{x}_4^3 + 0.2459\ \hat{x}_1\hat{x}_3\hat{x}_4 - \hat{x}_1\hat{x}_2\hat{x}_4 + 0.4698\ \hat{x}_3 - 0.3624 \qquad (7.6)$$

The *CC* and *RMSE* magnitudes in respect of the predictions by this model for the training and test set data are CC_{trn} = 0.969, CC_{tst} = 0.976, $RMSE_{\text{trn}}$ = 2.406 and $RMSE_{\text{tst}}$ = 4.289. These values clearly indicate that *GP_model-IV* possesses very good prediction and generalization performance. Figure 7.4 shows the parity plot of the experimental versus *GP_ Model-IV* predicted magnitude of the As(V) adsorption (y) on TAFA resin. A close match between the model-predicted As(V) adsorption values and their experimental counterparts strongly supports the observation that *GP_Model-IV* is capable of reasonably accurate predictions, and also possesses good generalization capability.

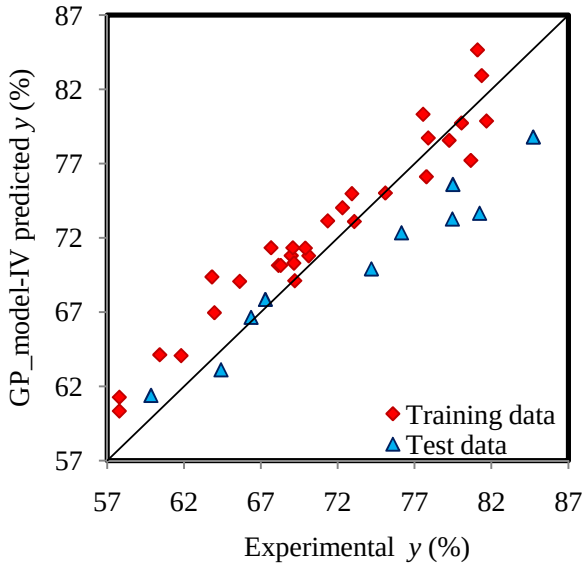


Figure 7.4: Parity plot of experimental versus GP-model predicted values of adsorption of As(V) on TAFA resin (%) (*y*).

In order to maximize the extent of As(V) adsorption on TAFA resin, the four inputs of *GP_Model-IV* were optimized using *MendelSolve* (2016) genetic algorithm software. The GA-specific parameters that yielded the three overall best sets of optimized reaction condition variables are: population size (N_{A}) = 60, crossover rate (R_{cr}^{A}) = 1, mutation rate (R_{mut}^{A}) = 0.03100 and maximum number of generations (N_{gen}) = 105. The top three GA searched sets of optimized reaction conditions resulting in the maximized As(V) adsorption on TAFA resin value are listed in Table 7.10. From the tabulated values, it is seen that the GA-searched best solution is capable of enhancing As(V) adsorption from 84.73% (best experimental As(V) adsorption value) to 99.83%.

Table 7.10: Optimized reaction variables given by GP-GA hybrid method for case study IV

Optimal solution set	Optimized reaction operating variables				GA-maximized As(V) adsorption (%)	Experimentally validated As(V) adsorption (%)
	tannin moles (x_1^{opt})	aniline moles (x_2^{opt})	formaldehyde moles (x_3^{opt})	pH (x_4^{opt})		
1.	0.000587	0.022	0.615	2.000	99.834	97.5
2.	0.001175	0.032	0.615	2.003	93.892	92.2
3.	0.000588	0.022	0.615	4.013	88.232	86.7

7.3.3 Experimental Validation of Optimized Reaction Operating Variables

In Tables 7.7–7.10, the overall best set of GA-optimized reaction conditions, which is expected to result in the maximum adsorption of As(III)/As(V) on TFA and TAFA resins is listed first. In order to validate the GA-optimized solutions, the overall best solution in each case study was subjected to experimental verification. The magnitudes of the adsorption (%) of As(III)/As(V) on TFA and TAFA resins measured in the respective validation experiments are listed in the last column of Tables 7.7–7.10. Here, it is seen that the experimentally validated adsorption magnitudes (99.6%, 98.3%, 93.32, 97.5%) are in reasonable to close agreement with the corresponding GA-maximized values of 99.46%, 99.81%, 94.07, and 99.83%, in case studies I, II, III, and IV, respectively. From the experimental data listed in Appendix 7.A, it is observed that the best adsorption values obtained using non-optimized reaction operating conditions in case studies I, II, III, and IV are 99.3% (experiment numbers 3 and 13), 97% (experiment numbers 31 and 36), 90.3% (experiment numbers 2 and 3), and 84.73% (experiment number 21), respectively. It can thus be seen that the optimized solutions provided by the GP-GA hybrid modeling-optimization strategy have enhanced the extent of As(III)/As(V) adsorption by 0.3%; 1.3%, 3.02, and 12.77% in case studies I to IV, respectively. In the absence of a reaction model, a non-assisted manual inspection of the reaction data provides no clues to the precise values of the optimized reaction conditions that are necessary for the maximization of adsorption of As(III)/As(V) on TFA and TAFA resins. This difficulty has been overcome by the usage of the GP-GA hybrid technique, which has provided the optimized conditions leading to reasonable to major improvements in the extent of As(III)/As(V) adsorption on TFA and TAFA resins.

7.4 CONCLUSION

Arsenic is one of the most toxic metalloid and very often forms a contaminant of the ground water globally. Since groundwater is an important source of the drinking water removal of arsenic therein has gained importance while managing and treating water and wastewater. Accordingly, this study reports usage of tannin-formaldehyde (TFA), and tannin-aniline-formaldehyde (TAFA) resins for the adsorptive removal of As(III) and As(V) ions. Moreover, the chapter presents results of a study wherein a hybrid method (termed ‘GP-GA’) integrating genetic

programming (GP) and genetic algorithms (GA) has been employed for modeling and optimization of the adsorptive removal of As(III)/As(V) ions using TFA and TAFA resins. The principal advantage of the GP-GA techniques is that modeling and optimization can be performed exclusively from the adsorption reaction data without invoking the detailed knowledge of the physicochemical phenomena underlying the reaction. Usage of the said hybrid formalism has provided a number of sets of optimized reaction conditions that are expected to maximize the adsorptive removal of As(III) and As(V) ions. The overall best of these optimized reaction conditions when verified experimentally, have resulted in 0.3% and 1.3% increases (over the corresponding best non-optimized experiments) in the TFA-based adsorption (%) of As(III) and As(V) metal ions, respectively. More significantly, improvements of 3.02% and 12.77% (over the respective best non-optimized experiments) have been witnessed in the adsorption of As(III) and As(V), respectively on the TAFA resin due to the application of GA-optimized reaction conditions. The GP-GA based hybrid modeling-optimization strategy presented here for the adsorptive removal of As(III)/As(V) ions can be gainfully utilized for the modeling and optimization of other contaminant removal processes.

NOMENCLATURE

I	dimensionality of the input space
N_{A}	population size in GA simulation
N_{gen}	maximum number of generations in GA evolution
N_{pat}	number of patterns in the example data set
R_{cr}^{A}	crossover rate in GA procedure
R_{mut}^{A}	mutation rate in GA procedure
$\hat{x}_i$	i^{th} normalized input variable
$\overline{x}_i$	mean values of the i^{th} input variable
x_i^{opt}	i^{th} optimized decision variable.

$\bar{y}$ mean value of output variable

$\hat{y}^j$ normal score (standardized variable) pertaining to the j^{th} output pattern

σ_{x_i} standard deviation values of the i^{th} input variable

σ_y standard deviation value of output variable

APPENDIX 7.A

Table 7.A.1: Experimental data for As(III) adsorption on TFA resin (case study I)

Expt. No.	Formaldehyde moles (x_1)	Ammonia moles (x_2)	pH (x_3)	As(III) adsorption (%) (y)
1	0.615	0.587	2	71.7
2	0.615	0.587	3	98.3
3	0.615	0.587	5	99.3
4*	0.615	0.587	8	87.9
5	0.615	0.587	10	80.0
6*	0.492	0.587	2	78.6
7	0.492	0.587	3	95.9
8	0.492	0.587	5	99.0
9	0.492	0.587	8	72.8
10*	0.492	0.587	10	76.2
11	0.369	0.587	2	73.5
12	0.369	0.587	3	99.0
13	0.369	0.587	5	99.3
14*	0.369	0.587	8	70.4
15	0.369	0.587	10	61.8
16	0.246	0.587	2	71.7
17	0.246	0.587	3	99.0
18*	0.246	0.587	5	99.0
19	0.246	0.587	8	76.2
20	0.246	0.587	10	65.9
21	0.123	0.587	2	56.9
22*	0.123	0.587	3	99.0
23	0.123	0.587	5	99.0
24	0.123	0.587	8	47.6
25	0.123	0.587	10	46.3
26*	0.369	0.294	2	71.4
27	0.369	0.294	3	66.9
28	0.369	0.294	5	73.8
29	0.369	0.294	8	74.2
30*	0.369	0.294	10	65.9
31	0.246	0.294	2	64.5
32*	0.246	0.294	3	65.5
33	0.246	0.294	5	55.2
34	0.246	0.294	8	62.8
35	0.246	0.294	10	55.6
36	0.123	0.294	2	61.1

Table 7.A.1 continued…

Expt. No.	Formaldehyde moles (x_1)	Ammonia moles (x_2)	pH (x_3)	As(III) adsorption (%) (y)
37	0.123	0.294	3	70.4
38*	0.123	0.294	5	60.7
39	0.123	0.294	8	61.4
40	0.123	0.294	10	63.3

*test data

Table 7.A.2: Experimental data for As(V) adsorption on TFA resin (case study II)

Expt. No.	Formaldehyde moles (x_1)	Ammonia moles (x_2)	pH (x_3)	As(V) adsorption (%) (y)
1*	0.615	0.587	2	83.4
2	0.615	0.587	4	86.1
3	0.615	0.587	8	85.6
4	0.615	0.587	9	83.9
5*	0.615	0.587	10	79.6
6	0.492	0.587	2	58.7
7	0.492	0.587	4	50.0
8	0.492	0.587	8	57.5
9*	0.492	0.587	9	50.3
10*	0.492	0.587	10	26.5
11	0.369	0.587	2	95.3
12*	0.369	0.587	4	92.4
13	0.369	0.587	8	93.2
14	0.369	0.587	9	90.2
15	0.369	0.587	10	89.5
16	0.246	0.587	2	93.6
17*	0.246	0.587	4	94.1
18	0.246	0.587	8	95.3
19	0.246	0.587	9	92.9
20	0.246	0.587	10	91.9
21*	0.123	0.587	2	94.1
22	0.123	0.587	4	94.9
23	0.123	0.587	8	94.9
24	0.123	0.587	9	92.9
25	0.123	0.587	10	92.9
26	0.369	0.294	2	94.9
27*	0.369	0.294	4	91.5
28	0.369	0.294	8	94.5
29	0.369	0.294	9	90.7
30	0.369	0.294	10	95.3
31*	0.246	0.294	2	97.0
32	0.246	0.294	4	95.8
33	0.246	0.294	8	95.8
34	0.246	0.294	9	93.6
35*	0.246	0.294	10	79.3
36	0.123	0.294	2	97.0
37	0.123	0.294	4	94.9

Table 7.A.2 continued…

Expt. No.	Formaldehyde moles (x_1)	Ammonia moles (x_2)	pH (x_3)	As(V) adsorption (%) (y)
38	0.123	0.294	8	96.4
39	0.123	0.294	9	92.9
40	0.123	0.294	10	93.6

*test data

Table 7.A.3: Experimental data for As(III) adsorption on TAFA resin (case study III)

Expt. No.	formaldehyde moles (x_1)	Ammonia moles (x_2)	pH (x_3)	As(III) adsorption (%) (y)
1*	0.615	0.587	2	84.8
2	0.615	0.587	3	90.3
3	0.615	0.587	5	90.3
4	0.615	0.587	8	60.4
5*	0.615	0.587	10	71.1
6*	0.369	0.587	3	88.6
7	0.369	0.587	5	76.2
8	0.369	0.587	8	76.9
9	0.369	0.587	10	80.7
10	0.246	0.587	2	83.3
11*	0.246	0.587	3	76.2
12	0.246	0.587	5	83.3
13	0.246	0.587	8	77.3
14	0.246	0.587	10	61.1
15	0.123	0.587	2	75.5
16	0.123	0.587	3	71.7
17*	0.123	0.587	5	86.2
18	0.123	0.587	8	64.2
19*	0.123	0.587	10	72.1
20	0.615	0.294	2	74.8
21	0.615	0.294	3	79.3
22	0.615	0.294	5	69
23*	0.615	0.294	8	59.3
24	0.492	0.294	3	81
25*	0.492	0.294	5	80.4
26	0.492	0.294	8	62.1
27	0.492	0.294	10	49.4
28	0.369	0.294	3	89.9
29	0.369	0.294	5	88.3
30	0.369	0.294	8	88.3
31	0.369	0.294	10	75.9
32	0.246	0.294	2	76.9
33*	0.246	0.294	3	84.5
34	0.246	0.294	5	75.5
35	0.246	0.294	8	80.7
36	0.246	0.294	10	53.2
37	0.123	0.294	2	64.9
38*	0.123	0.294	3	64.2

Table 7.A.3 continued…

Expt. No.	formaldehyde moles (x_1)	Ammonia moles (x_2)	pH (x_3)	As(III) adsorption (%) (y)
39	0.123	0.294	5	63.1
40	0.123	0.294	10	52.5

*test data

Table 7.A.4: Experimental data for As(V) adsorption on TAFA resin (case study IV)

Expt. No.	Tannin (mole) (x_1)	Aniline (mole) (x_2)	Formaldehyde (mole) (x_3)	pH (x_4)	As(V) adsorption(%) (y)
1	0.001175	0.0215	0.615	2	68.98
2*	0.001175	0.0215	0.615	4	74.20
3	0.001175	0.0215	0.615	8	77.56
4	0.001175	0.0215	0.615	9	81.38
5	0.001175	0.0215	0.615	10	81.10
6*	0.001175	0.0215	0.493	2	67.30
7	0.001175	0.0215	0.493	8	75.11
8	0.001175	0.0215	0.493	9	80.67
9	0.001175	0.0215	0.493	10	79.26
10	0.001175	0.0215	0.369	4	61.83
11	0.001175	0.0215	0.369	8	67.68
12	0.001175	0.0215	0.369	9	71.36
13	0.001175	0.0215	0.246	2	63.98
14	0.001175	0.0215	0.246	8	63.83
15	0.001175	0.0215	0.246	9	70.12
16	0.001175	0.0215	0.246	10	69.09
17	0.001175	0.0215	0.123	2	65.63
18	0.001175	0.0215	0.123	8	69.21
19	0.001175	0.0215	0.123	9	68.16
20	0.001175	0.0215	0.123	10	69.16
21*	0.000587	0.0322	0.615	2	84.73
22	0.000587	0.0322	0.615	8	77.90
23	0.000587	0.0322	0.615	9	80.07
24	0.000587	0.0322	0.615	10	81.69
25*	0.000587	0.0322	0.493	9	79.50
26	0.000587	0.0322	0.493	10	77.78
27*	0.000587	0.0322	0.369	2	66.37
28	0.000587	0.0322	0.369	4	60.43
29	0.000587	0.0322	0.369	8	69.90
30	0.000587	0.0322	0.369	9	73.09
31	0.000587	0.0322	0.369	10	72.31
32*	0.000587	0.0322	0.246	2	64.42
33*	0.000587	0.0322	0.246	4	59.86
34	0.000587	0.0322	0.246	8	68.30
35*	0.000587	0.0322	0.246	9	76.16
36*	0.000587	0.0322	0.246	10	81.24
37	0.000587	0.0322	0.123	2	57.80

Table 7.A.4 continued…

Expt. No.	Tannin (mole) (x_1)	Aniline (mole) (x_2)	Formaldehyde (mole) (x_3)	pH (x_4)	As(V) adsorption(%) (y)
38	0.000587	0.0322	0.123	4	57.80
39*	0.000587	0.0322	0.123	9	79.47
40	0.000587	0.0322	0.123	10	72.93

*test data

REFERENCES

Amin, M. N., Kaneco, S., Kitagawa, T., Begum, A., Katsumata, H., Suzuki, T., and Ohta, K. (2006). Removal of arsenic in aqueous solutions by adsorption onto waste rice husk. *Industrial & Engineering Chemistry Research*, *45*(24), 8105-8110.

Anderson, M. A., Ferguson, J. F., and Gavis, J. (1976). Arsenate adsorption on amorphous aluminum hydroxide. *Journal of Colloid and Interface Science*, *54*(3), 391-399.

Arai, Y., Sparks, D. L., and Davis, J. A. (2005). Arsenate adsorption mechanisms at the allophane-water interface. *Environmental Science & Technology*, *39*(8), 2537-2544.

Bahrami, P., Kazemi, P., Mahdavi, S., and Ghobadi, H. (2016). A novel approach for modeling and optimization of surfactant/polymer flooding based on Genetic Programming evolutionary algorithm. *Fuel*, *179*, 289-298.

Bishop, C. M. (1994). Neural networks and their applications. *Review of Scientific Instruments*, *65*(6), 1803-1832.

Cheema, J. J. S., Sankpal, N. V., Tambe, S. S., and Kulkarni, B. D. (2002). Genetic programming assisted stochastic optimization strategies for optimization of glucose to gluconic acid fermentation. *Biotechnology Progress*, *18*(6), 1356-1365.

Dambies, L., Vincent, T., and Guibal, E. (2002). Treatment of arsenic-containing solutions using chitosan derivatives: uptake mechanism and sorption performances. *Water Research*, *36*(15), 3699-3710.

Deb, K. (1995). *Optimization for Engineering Design: Algorithms and Examples*. Prentice-Hall, New Delhi.

Dutré, V., and Vandecasteele, C. (1998). Immobilization mechanism of arsenic in waste solidified using cement and lime. *Environmental Science & Technology*, *32*(18), 2782-2787.

Goel, P., Bapat, S., Vyas, R., Tambe, A., and Tambe, S. S. (2015). Genetic programming based quantitative structure–retention relationships for the

prediction of Kovats retention indices. *Journal of Chromatography A, 1420*, 98-109. doi:http://dx.doi.org/ 10.1016/j.chroma.2015.09.086.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning* . Addison-Wesley, Reading, MA.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. 1975. Univ. of Michigan Press, Ann Arbor.

Huang, K., Zhan, X. L., Chen, F. Q., and Lü, D. W. (2003). Catalyst design for methane oxidative coupling by using artificial neural network and hybrid genetic algorithm. *Chemical Engineering Science*, *58*(1), 81-87.

Johnson, D. L., and Pilson, M. E. (1972). Spectrophotometric determination of arsenite, arsenate, and phosphate in natural waters. *Analytica Chimica Acta*, *58*(2), 289-299.

Kim, Y., Kim, C., Choi, I., Rengaraj, S., and Yi, J. (2004). Arsenic removal using mesoporous alumina prepared via a templating method. *Environmental Science & Technology*, *38*(3), 924-931.

Koç, D. İ., and Koç, M. L. (2015). A genetic programming-based QSPR model for predicting solubility parameters of polymers. *Chemometrics and Intelligent Laboratory Systems*, *144*, 122-127. doi:http://dx.doi.org/10.1016/j.chemolab.2015.04.005.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (Vol. 1). MIT press,Cambridge, MA.

Kundu, S., and Gupta, A. K. (2006). Investigations on the adsorption efficiency of iron oxide coated cement (IOCC) towards As (V)—kinetics, equilibrium and thermodynamic studies. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, *273*(1), 121-128.

Kundu, S., Kavalakatt, S. S., Pal, A., Ghosh, S. K., Mandal, M., and Pal, T. (2004). Removal of arsenic using hardened paste of Portland cement: Batch adsorption and column study. *Water Research*, *38*(17), 3780-3790.

Liao, X., Lu, Z., Zhang, M., Liu, X., and Shi, B. (2004). Adsorption of Cu (II) from aqueous solutions by tannins immobilized on collagen. *Journal of Chemical Technology and Biotechnology*, *79*(4), 335-342.

Makeswari, M., Shanti, T., and Manonmani, S. (2014). Adsorption of Cr (VI) Ions from Aqueous Solutions on to Tannin Gel Prepared from Leaves of Ricinuscommunis. *International Journal of Research in Chemistry and Environment*, *4*(3), 90-100.

MendelSolve, A Genetic Algorithm Solver, (2016). (accessed 20.05.2016) http://www.bluestretch.com/mendelsolve/index.htm.

Mohan, D., and Pittman, C. U. (2007). Arsenic removal from water/wastewater using adsorbents—A critical review. *Journal of Hazardous Materials*, *142*(1), 1-53.

Mulani, K., Daniels, S., Rajdeo, K., Tambe, S., and Chavan, N. (2014). Tannin-aniline-formaldehyde resole resins for arsenic removal from contaminated water. *Canadian Chemical Transactions*, *2*(4), 450-466.

Nandi, S., Badhe, Y., Lonari, J., Sridevi, U., Rao, B. S., Tambe, S. S., and Kulkarni, B. D. (2004). Hybrid process modeling and optimization strategies integrating neural networks/support vector regression and genetic algorithms: study of benzene isopropylation on Hbeta catalyst. *Chemical Engineering Journal*, *97*(2), 115-129.

Nandi, S., Ghosh, S., Tambe, S. S., and Kulkarni, B. D. (2001). Artificial neural-network-assisted stochastic process optimization strategies. *AIChE Journal*, *47*(1), 126-141.

Ng, J. C., Wang, J., and Shraim, A. (2003). A global health problem caused by arsenic from natural sources. *Chemosphere*, *52*(9), 1353-1359.

Onyango, M. S., Matsuda, H., and Ogada, T. (2003). Sorption kinetics of arsenic onto iron-conditioned zeolite. *Journal of Chemical Engineering of Japan*, *36*(4), 477-485.

Pandey, D. S., Pan, I., Das, S., Leahy, J. J., and Kwapinski, W. (2015). Multi-gene genetic programming based predictive models for municipal solid waste

gasification in a fluidized bed gasifier. *Bioresource technology*, *179*, 524-533. doi:http:// dx.doi.org/10.1016/j.biortech.2014.12.048.

Patil-Shinde, V., Kulkarni, T., Kulkarni, R., Chavan, P. D., Sharma, T., Sharma, B. K., Tambe, S.S., and Kulkarni, B. D. (2014). Artificial intelligence-based modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier. *Industrial & Engineering Chemistry Research*, *53*(49), 18678-18689.

Pena, M., Meng, X., Korfiatis, G. P., and Jing, C. (2006). Adsorption mechanism of arsenic on nanocrystalline titanium dioxide. *Environmental Science & Technology*, *40*(4), 1257-1262.

Rao, K., Rangajanardhaa, G., Rao, H., and Rao, S. (2009). Development of hybrid model and optimization of surface roughness in electric discharge machining using artificial neural networks and genetic algorithm. *Journal of Materials Processing Technology*, *209*(3), 1512-1520.

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*(5923), 81-85.

Shirato, W., and Kamei, Y. (1994). Method of preparing insoluble hydrolysable tannin and method of treating waste liquid with the tannin. *U.S. Patent No. 5,296,629*. Washington, DC: U.S. Patent and Trademark Office.

Shirato, W., and Kamei, Y. (1994). Preparation of insoluble tannin and its applications for waste treatment and adsorption processes. *European Patent EP0438776 A1*.

Tambe, S. S., Deshpande, P. B., Kulkarni, B. D. (1996). *Elements of artificial neural networks with selected applications in chemical engineering, and chemical & biological sciences*. Simulation & Advanced Controls Inc., Louisville, K.Y.

Thirunavukkarasu, O. S., Viraraghavan, T., and Subramanian, K. S. (2001). Removal of arsenic in drinking water by iron oxide-coated sand and ferrihydrite-batch studies. *Water Quality Research Journal of Canada*, *36*(1), 55-70.

USEPA (1999). The United States Environmental Protection Agency, Analytical Methods Support Document for Arsenic in Drinking Water, USEPA, Washington DC.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Verma, D., Goel, P., Patil-Shinde, V., and Tambe, S. S. (2016, January). Use genetic programming for selecting predictor variables and modeling in process identification. In *IEEE explore, 2016 Indian Control Conference* (*ICC*) (pp. 230-237). IEEE.

Vyas, R., Goel, P., and Tambe, S. S. (2015). Genetic programming applications in chemical sciences and engineering. In *Handbook of Genetic Programming Applications*; Gandomi, A.H., Amir H., Alavi, Ryan, C. (Eds.), Springer International Publishing, Switzerland, pp 99–140.

WHO,World Health Organization. (2004). *Guidelines for drinking-water quality: Recommendations* (Vol. 1). 2nd ed., Geneva, 1993.

Wu, C. H., Tzeng, G. H., and Lin, R. H. (2009). A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems with Applications*, *36*(3), 4725-4735.

Zaid, S. (2012). Development of support vector regression (SVR)-based model for prediction of circulation rate in a vertical tube thermosiphon reboiler. *Chemical Engineering Science*, *69*(1), 514-521.

Zhang, J. S., Stanforth, R. S., and Pehkonen, S. O. (2007). Effect of replacing a hydroxyl group with a methyl group on arsenic (V) species adsorption on goethite (α-FeOOH). *Journal of Colloid and Interface Science*, *306*(1), 16-21.

Zhang, M., Ding, C., Chen, L., and Huang, L. (2015). Preparation of Tannin-immobilized Collagen/Cellulose Bead for Pb (II) Adsorption in Aqueous Solutions. *BioResources*, *10*(1), 1773-1789.

Zurada, J.M. (1992). *Introduction to Artificial Neural Network.* West Publ Co., St. Paul.

Chapter 8

Genetic Programming based Models for Prediction of Vapor-Liquid Equilibrium

ABSTRACT

In chemical industry, design, operation, and control of separation processes heavily rely on the knowledge of the vapor-liquid equilibrium (*VLE*). *It is not always feasible, convenient, and economical to carry out detailed experiments studying the effects of operating parameters on the separation behavior. Thus, commonly thermodynamic models such as the equation of state* (*EoS*), *and activity coefficient are used for the estimation of VLE. These models are mostly developed for binary, tertiary, and quaternary systems. Purely data-driven modeling approaches are also used to develop these models. This approach too has its own difficulties. This chapter presents a study wherein genetic programming* (*GP*) *has been introduced for the prediction of VLE. Specifically, three case studies have been performed wherein four GP-based VLE models have been developed using experimental data for predicting the vapor phase composition,* (y_i), *of a ternary, and two groups of non–ideal binary systems. The input space of these models consists of three attributes of pure components* (*acentric factor, critical temperature, and critical pressure*), *and three intensive thermodynamic parameters* (*liquid phase composition, pressure, and temperature*). *The prediction and generalization performance of the GP-based models was rigorously compared with that of the corresponding conventionally employed Van Laar, NRTL, and UNIQUAC models. The results obtained thereby indicate superior prediction accuracy and generalization performance of the GP-based models vis a vis that of the conventional thermodynamic models. The GP-based modeling method proposed in this study can be gainfully utilized in the prediction of VLE as also designing corresponding experiments in different pressure and temperature ranges.*

8.1 INTRODUCTION

An accurate prediction of the phase behavior of chemical species and their mixtures is essential for designing, optimizing and controlling separation processes, and other unit operations employed in the chemical industry. Predicting phase equilibrium properties, such as phase composition, and partition coefficients at temperatures and pressures of interest using reliable models offers an attractive alternative to costly and time consuming experimental measurements (Gebreyohannes et al., 2013). Phase equilibrium, and in particular vapor-liquid-equilibrium (VLE), is important in a number of process engineering applications. Designing an effective, efficient, and economical separation scheme is immensely essential since lack of an accurate knowledge of VLE poses significant difficulties in chemical process design and development work. It has been broadly recognized that a viable separation scheme is as important as good chemistry for the success of chemical processes on a commercial scale (Dohrn and Brunner ,1995).

Conducting VLE experiments and a precise measurement of data thereof is often tedious, time-consuming and expensive; for highly reactive systems, the task becomes even more difficult and complicated. For instance, it is not always feasible to carry out VLE experiments at all the ranges of operating temperatures and pressures of practical interest (Vaferi et al., 2013). To overcome this difficulty, mathematical models are developed for the prediction of VLE.

There exist two principal methods, namely, *phenomenological* and *empirical*, for modeling VLE. The phenomenological approach (also termed “mechanistic” or “first principles”) includes thermodynamic models such as, *equation of state*, and *activity coefficient* models (Lashkarbolooki et al., 2013). This approach needs complete knowledge of the underlying physico-chemical phenomena. The prediction of VLE data by conventional thermodynamic methods is tedious since it involves determination of various thermodynamic parameters, which is arbitrary in many ways and, in some cases also introduces significant inaccuracies (Sharma et al., 1999). For some of the components, determination of thermodynamic parameters such as *binary interaction parameter* (BIP) by itself can be an elaborate and time consuming exercise.

The second (i.e. *empirical*) approach to VLE modeling is exclusively data-driven and, therefore, can be employed in the absence of a detailed knowledge of the

physico-chemical phenomenon underlying the VLE. It utilizes linear/nonlinear regression methods in formulating the models. A significant requirement of this approach is that the exact structure (form) of the linear/nonlinear data-fitting model needs to be specified unambiguously prior to the estimation of the unknown model-parameters. In case of ideal systems exhibiting a linear VLE behavior, the specification of the corresponding linear data-fitting function is relatively easy. However, VLE behavior of a large number of systems displays a nonlinear dependence on the operating parameters. In such cases choosing an appropriate nonlinear data-fitting model, from numerous competitive models becomes a daunting task (Patil-Shinde et al., 2014). The above-stated difficulties, which are faced commonly during both the phenomenological and regression-based VLE modeling, require exploration of alternative nonlinear modeling strategies.

The two computational intelligence (CI) based exclusively data-driven nonlinear modeling formalisms, namely, *Artificial neural networks* (ANNs) (see e.g., Bishop, 1994; Zurada, 1992; Tambe et al., 1996), and *support vector regression* (SVR) (Vapnik, 1995; Zaid, 2012) are often used as alternatives to the regression based modeling. These have found numerous applications in the field of thermodynamics and prediction of transport properties. Table 8.1 reports a number of studies wherein ANNs and SVR have been employed in VLE predictions.

In addition to ANNs and SVR, the field of computational intelligence comprises a data-driven modeling strategy, namely *genetic programming* (GP). The GP formalism has been described in detail in Chapter 2 (section 2.2.2). In earlier applications the GP technique has been used, for instance, in estimation of solvent activity in polymer solutions (Tashvigh et al., 2015), process identification (Verma et al., 2016), gasification performance prediction (Patil-Shinde et al., 2016), prediction of Kovats retention indices (Goel et al., 2016). Since it possesses several attractive characteristics, in this study, the GP formalism has been utilized for developing data-driven models predicting the vapor phase composition of ternary, and a group of binary mixtures. An exhaustive literature search indicates that this is the first instance wherein GP has been used in VLE prediction. The systems studied here are industrially relevant. The three specific VLE modeling case studies that have been performed are listed in Table 8.2. In all, four GP-based models have been developed; the inputs and outputs pertaining to these models are given in Table 8.3.

Table 8.1: VLE studies by using Artificial Intelligence formalisms

System(s)	Components	Inputs	Outputs	AI formalism	Thermodynamic model	Reference
Binary	Nine binary systems containing ethanol	Temperature (T), critical temperature (T_C), critical pressure (P_C), acentric factor(ω), normal boiling point (T_b) and composition of the solutes in the liquid (x_i)	Bubble point pressure (P), and vapor phase composition (y_i)	Multi-layer perceptron (MLP) trained using back-propagation algorithm	Peng–Robinson equation of state	Vaferi et al. (2013)
Binary + alkanols	CO_2+ 1 – propanol, CO_2+ 2 – propanol, CO_2+ 1– butanol, CO_2+ 1– pentanol, CO_2+ 2– pentanol, CO_2+ 1– hexanol,and CO_2+ 1– heptanol	Equilibrium temperature (T), CO_2mole fraction in the liquid phase (x_1), critical temperature of alkanol (T_C), critical pressure of alkanol (P_C), and acetnric factor of an alkanol (ω).	Equilibrium pressure (P), and CO_2 mole fraction in the vapor phase (y_i)	Multilayer perceptron (MLP) trained using Levenberg-Marquardt back-propagation learning algorithm	Peng–Robinson EOS coupled with Van der Waals and Wong-Sandler mixing rules	Zarenezhad and Aminian (2011)
Ternary	CO_2, NH_3, and H_2O	Normalised concentrations of CO_2, NH_3, and H_2O in the liquid phase (m_i), and Temperature (T)	Partial pressure (P_i), and total pressure(P_{total})	Multi-layer perceptron (MLP) and Radial basis function(RBF)	Kurz et al. (1995); Muller et al. (1998) ; Goppert and Mauner (1988)	Ghaemi et al. (2008)

Table 8.1 continued…

System(s)	Components	Inputs	Outputs	AI formalism	Thermodynamic model	Reference
Binary	Chlorodifluoromethan + carbondioxide (R22–CO_2), trifluoromethan + carbondioxide (R23–CO_2), carbondisulfied + trifluoromethan(CS_2–R23), carbondisulfied + chlorodifluoromethan (CS_2–R22)	Temperature (*T*), pressure (*P*)	Mole fraction of CO_2, R22, and R23 in the liquid phase (x_i), and mole fraction of CO_2, R22, and R23 in the vapor phase (y_i)	Multilayer perceptron (MLP) trained by the Levenberg-Marquardt algorithm	Redlich-Kwang-Soave (RKS) equation of state	Karimi and Yousefi (2007)
CO_2(solvent)/ hydrocarbon binary mixtures	CO_2 + 1–Hexane , CO_2 + 2-Ethyl –1–butene, CO_2 + n –Hexane, CO_2 + Propyl acrylate , CO_2 + Propyl methacrylate, CO_2 + Decafluorobutane, CO_2 + Methyl methacrylate	Reduced temperature (T_r), hydrocarbon mole fraction in liquid phase (x_i), and hydrocarbon mole fraction in vapor phase (y_i), hydrocarbons acentric factor (ω), and critical pressure (P_C),.	Bubble point/dew point pressure	Least-Squares Support Vector Machine (LSSVM)	Peng–Robinson equation of state and SAFT	Mesbah et al. (2015)
Binary	CO_2 + ethyl caproate, CO_2 + ethyl caprylate, and CO_2 + ethyl caprate	Temperature (*T*), pressure (*P*),	Mole fraction of CO_2 in vapor phase (y_{CO_2}), and liquid phase (x_{CO_2}),	Multilayer perceptron neural network (MLP NN)	Soave-Redlich-Kwong (SRK) or Peng Robinsons equation of state	Mohanty (2005)

Table 8.1 continued…

System(s)	Components	Inputs	Outputs	AI formalism	Thermodynamic model	Reference
Binary systems of CO_2 + cyclic compounds	CO_2+ Bisphenol A (BPA), CO_2+ Diphenyl carbonate , CO_2+ Quinoline, CO_2+ Nicotine, CO_2+ Benzene, CO_2+ Tetrahydrofuran	Reduced temperature of the system (T_{r2}), critical pressure (P_{C2}) , acentric factor (ω_2) of cyclic compounds and composition of CO_2 in the vapor (y_i), and liquid (x_i) phases	Bubble- and dew-point pressures	Cascade-forward back-propagation artificial neural network	Not used thermodynamic model	Lashkarbol-ooki et al. (2013)
Binary systems of CO_2 + hydrocarbon	CO_2 + n–Pentadecane, CO_2 + Decafluorobutane, CO_2 + 1 Hexene, CO_2 + 2-Ethyl –1– butene, CO_2 + n-Hexane	Reduced temperature of non-CO_2 compound $T_r = T/T_C$; where T is the temperature of the binary system and T_C is the critical temperature of non-CO_2 compound, critical pressure of non-CO_2 compound (P_C), acentric factor of non-CO_2 compound (ω) and mole fraction of carbon dioxide in binary systems (x_{CO_2}) −for bubble point pressure and (y_{CO_2}) − for dew point pressure	Bubble and dew point pressures	cascade-forward back-propagation artificial neural network	Not used thermodynamic model	Lashkarbol-ooki et al. (2013)
Ternary	Water + ethanol +2–propanol saturated with $NaNO_3$, NaCl, KCl Ethanol +1– propanol + water saturated with NaCl, KCl, $CuSO_4$	Liquid mole fraction of solvents (x_1, x_2, x_3), critical temperature of solvent (T_C), critical pressure of solvent (P_C), acentric factor (ω), cation radius ($R+$) and anion radius ($R-$)	Vapor mole fraction of solvents (y_1, y_2, y_3), and bubble point temperature (T)	Multilayer perceptron neural network (MLP NN)	Tan–Wilson (modification of Wilson model)	Nguyen et al. (2007)

Table 8.1 continued…

System(s)	Components	Inputs	Outputs	AI formalism	Thermodynamic model	Reference
Binary systems of CO_2+ ester	CO_2 + ethyl caprate, CO_2 + ethyl caproate, CO_2 + ethyl caprylate, CO_2 + diethyl carbonate, CO_2 + ethyl butyrate and CO_2 + isopropyl acetate	Equilibrium temperature (T), CO_2 mole fraction in the liquid phase x_{CO_2}, critical temperature (T_C), critical pressure (P_C), and acentric factor $\omega_2$of esters	Equilibrium pressure (P), and CO_2 mole fraction in the vapor phase (y_{CO_2})	Feed forward, back propagation neural network	Peng-Robinson (PR) and Soave-Redlick-Kwong (SRK) EOS	Si-Moussa et al. (2008)
Seventeen binary systems	The binary systems composed of alkenes, aromatics, aldehydes, alcohols, amines, amides, carboxyl acids, nitriles,esters, ethers, ketones nitro compounds, water, and halogen compound	Critical volume(V_C), acentric factor (ω), dipole moment, entropy of vaporization and electronegativity of components 1 and 2.	Margules parameters	Multi-layer perceptron (MLP) trained using back-propagation algorithm	Margules activity coefficient equation	Yamamoto and Tochigi (2007)
Binary, and Ternary	Hexane + ethanol, Hexane + benzene, Carbon disulfide + acetone, Acetone + chloroform, Hexane-benzene + toluene, Acetone + methanol + chloroform	Mole fraction in liquid phase (x_i), temperature (T)	Mole fraction in vapor phase (y_i),and pressure (P)	Radial basis function (RBF) NN	Universal Quasi-chemical (UNIQUAC) model	Ganguly (2003)
Binary	Methane + ethane Ammonia +water	Temperature (T), pressure (P),	Mole fraction in vapor phase (y_i), mole fraction in liquid phase (x_i),	Multi-layer perceptron (MLP) trained using back-propagation algorithm	Peng-Robinson EOS	Sharma et al. (1999)

Table 8.2: Description of three case studies

Case study	System	Components	Objective of GP-based modeling
I	Ternary	(i) 1, 2-dichloroethane (1) (ii) trichloroethylene (2) (iii) 1-propanol (3)	Prediction of the mole fraction of 1-2 dichloroethane (y_1), and trichloroethylene (y_2) in vapor phase
II	Group of binary	(i) tetrachloromethane (1) – ethanol (2) (ii) tetrachloromethane (1) –1-propanol (2), (iii) tetrachloromethane (1) – 1-butanol (2)	Prediction of mole fraction of tetrachloromethane in vapor phase (y_1), using a single model for three binary systems.
III	Group of binary	(i) ethanol (1) – ethyl acetate (2) (ii) 1-propanol (1) – propyl acetate (2) (iii) 1-butanol (1) – butyl acetate (2), (iv) 1-pentanol (1) – pentyl acetate (2)	a. Prediction of mole fraction of ethanol, 1-propanol, and 1-butanol in vapor phase {(y_1)}, using a single model developed using data of first three binary systems for interpolation. b. To test the extrapolation ability of the developed model on fourth binary system, namely, 1-pentanol (1) – pentyl acetate (2) to predict vapor phase composition of 1-pentanol (y_1).

Table 8.3: The inputs and the outputs pertaining to the four GP-based models developed in this study.

Model	No. of inputs	Model inputs (intensive thermodynamic variables and pure component properties)	Model output
I (case study-I)	6	Temperature (T); mole fractions of 1, 2-dichloroethane (x_1), and trichloroethylene (x_2) in liquid phase; acentric factors of 1, 2-dichloroethane (ω_1), trichloroethylene (ω_2) and 1-propanol (ω_3)	Mole fraction of 1, 2-dichloroethane in vapor phase (y_1)
II (case study-I)	6	Temperature (T); mole fractions of 1, 2-dichloroethane (x_1), and trichloroethylene (x_2) in liquid phase; acentric factors of 1, 2-dichloroethane (ω_1), trichloroethylene (ω_2) and 1-propanol (ω_3)	Mole fraction of trichloroethylene in vapor phase, (y_2)
III (case study-II)	5	(i) Mole fraction of tetrachloromethane in liquid phase (x_1), (ii) pressure, (P), (iii) temperature (T), (iv) critical temperature of the second component, namely, ethanol/1-propanol/ 1-butanol (T_{c_2}) of the binary system, and (v) critical pressure the second component, namely, ethanol/1-propanol/ 1-butanol (P_{c_2}) of the binary system	Mole fraction of tetrachloromethane in vapor phase (y_1)
IV(case study-III)	5	(i) Acentric factor of the first component, namely, ethanol/1-propanol/ 1-butanol (ω_1) of the binary system, (ii) acentric factor of the second component, namely, ethyl acetate/propyl acetate/butyl acetate (ω_2) of the binary system, (iii) liquid phase mole fractions of first components namely, ethanol/1-propanol/1-butanol (x_1) of the binary system, (iv) pressure (P) (kP_a), and (v) temperature (T) (K),	Mole fraction of ethanol, 1-propanol, and 1-butanol, and 1-pentanol in vapor phase $\{y_1\}$

Once a properly validated optimal GP-based model is constructed, its parameters can be further refined using a standard nonlinear regression technique, such as, Marquardt's algorithm (Marquardt, 1963). In this study, all the four GP-based models were developed and their generalization capability was assessed using experimental data. Additionally, as reported in DECHEMA (Gmehling and Onken, 1986; Gmehling et al., 1986) data series, the same sets of VLE data were used to develop corresponding activity coefficient models; in the case studies a subset of following activity coefficient models was used: Wilson (Wilson, 1964; Smith et al., 2005), Van Laar (Wong et al., 1992), Non random two-liquid (NRTL) (Renon and Prausnitz, 1969), and Universal Quasi-chemical (UNIQUAC) (Anderson and Prausnitz, 1978). The prediction and generalization performance of the GP-based models was rigorously compared with that of the corresponding thermodynamic models. The results of this comparison indicate that that the GP-based models possess comparable or better VLE prediction ability than the conventional thermodynamic models.

The remainder of this chapter is structured as follows. Various thermodynamic models for the VLE predictions are described briefly in section 8.2 titled "Phase equilibria modeling." section 8.3, titled "Data" provides details of the data used in the GP-based modeling of vapor phase composition. The next, section 8.4 titled ''Results and Discussion,'' describes the three case studies wherein GP-based models have been developed for (i) ternary system (section 8.4.2), (ii) a group of three non-ideal binary systems (section 8.4.3), and (iii) a group of four non-ideal binary systems (section 8.4.4). Additionally, this section also provides results of the comparison of the prediction and generalization performance of the developed four GP–based models with their thermodynamic counterparts as also the fine-tuned genetic programming-Marquardt (GP- Marquardt) models. Finally, ''Concluding Remarks'' (section 8.5) summarize the principal findings of the study.

8.2 PHASE EQUILIBRIA MODELING

8.2.1 Activity Coefficient Models

A number of methods such as, the *regular solution theory*, *universal functional activity coefficient* (UNIFAC) (Fredenslund et al., 1977), or *analytical solution of groups* (ASOG) (Derr and Deal,1969) are available for the VLE prediction; however, none of these strategies can be regarded as a highly accurate predictor (Smith et al.,

2005). Thus, these methods are used only when no experimental VLE data are available for the system of interest. The notable features regarding the applicability of thermodynamic VLE models are given below (Prausnitz et al., 1998; Smith et al., 2005):

- For moderately non-ideal systems, all the major models (Van Laar, two constant Margules, Wilson, UNIQUAC, and NRTL) perform comparably well.
- For mixtures of very different species, such as polar or associating compounds (e.g. alcohols and other oxy hydrocarbons), the two-parameter VLE models, namely, Van Laar, two constant Margules, UNIQUAC, and Wilson equation, are preferred over the three parameter NRTL equation.
- For non-polar solvents (e.g. hydrocarbons), the Wilson, UNIQUAC, and NRTL models have been found to make superior predictions than the Van Laar and two-parameter Margules equations.
- The NRTL and UNIQUAC equations are useful whereas the Wilson equation is inapplicable for species which are dissimilar and are only partially soluble to form two liquid phases.

8.2.2 Equation of State Models

A commonly employed method for predicting/describing thermodynamic properties of fluids, mixtures of fluids, and solids is "*equations of state* (EoS)". It is an efficient tool for calculating also the phase equilibrium of systems in pure or mixture form. The EOSs are widely used in theoretical and practical studies involving chemical process design, petroleum industry, reservoir fluids, etc. The van der Waals equation of state (Van der Waals, 1910) was the first equation to predict vapor-liquid coexistence. Later, the Redlich-Kwong equation of state (Redlich and Kwong, 1949) improved the accuracy of the van der Waals equation by proposing temperature dependence for the attractive term. Soave (1972) and Peng and Robinson (1976) proposed additional modifications of the Redlich-Kwong equation to more accurately predict the vapor pressure, liquid density, and equilibrium ratios. Numerous equations of state have been proposed in the literature with either an empirical, semi empirical, or theoretical basis. There are some notable comprehensive reviews on equation of state and these can be found in the works of Martin (1979), Anderko (1990), and Sengers et al. (2000).

Since design, operation, and control of a large number of industrial chemical processes are based on the predictions of VLE, and thermodynamic property models, it is at most important that they are robust and capable of accurate predictions.

8.3 DATA

The experimental VLE data were sourced from the Chemistry Data Series, DECHEMA (Gmehling and Onken, 1986; Gmehling et al., 1986) (see Appendix 8.A (Tables 8.A.1, 8.A.2, and 8.A.3)) to develop four GP-based models in three case studies for the prediction of vapor phase composition. Each of the four example sets used in the four case studies was split randomly in (75:25) ratio in the training and test sets, respectively. Whereas 75% data were used in developing (training) the GP-based models, the test set data (25%) were used in testing the generalization ability of the developed models.

Table 8.4: Physical properties of the components used in this study

Component	Acentric factor (ω)	Critical Temperature (T_c) (K)	Critical Pressure (P_c) (kP_a)	Reference
Ethanol	0.6436	514.0	6137	Perry and Green (2007)
1-Propanol	0.6209	536.8	5169	Perry and Green (2007)
1-Butanol	0.5883	563.1	4414	Perry and Green (2007)
1,2-Dichloroethane	0.2866			Perry and Green (2007)
Trichloroethylene	0.2170			Yaws (1999)
1-Pentanol	0.5748			Perry and Green (2007)
Ethyl acetate	0.3664			Perry and Green (2007)
Propyl acetate	0.3889			Perry and Green (2007)
Butyl acetate	0.4394			Perry and Green (2007)
Pentyl acetate	0.4480			Yaws (1999)

8.4 RESULT AND DISCUSSION

8.4.1 GP-based Vapor-Liquid Equilibria Modeling

All GP-based models were developed using the *Eureqa Formulize* software package (Schmidt and Lipson, 2012). The package has a number of options for preprocessing of the example input-output data and generation of candidate solutions. While building each GP-based model, these options were rigorously and

systematically explored with the objective of obtaining models possessing high accuracy of predicting *vapor phase mole fraction* (y_i) and generalization capability. An operator set containing five arithmetic operators, namely, *addition*, *subtraction*, *multiplication*, *division* and *exponentiation*, was used in the generation of the candidate expressions. To obtain a best possible data-fitting model, the GP (Koza, 1992; Poli, 2008) procedure (see Chapter 2, section 2.2.2) was repeated a number of times by using every time different seed expressions and random number generator seed values. It is worth noting that in each repeated run, the GP algorithm converged to a different mathematical expression. The fitness of each candidate expression was evaluated using the *squared error* fitness function. The statistical measures, used in assessing the prediction accuracy and generalization performance of a GP-based model were *coefficient of correlation* (*CC*) and *root mean squared error* (*RMSE*); these were evaluated using the experimental (target) and the corresponding model-predicted values of *vapor phase mole fraction* (y_i). These two statistical quantities were calculated separately for the training and test data sets. The overall best GP-model was selected from those obtained in the multiple GP runs on the basis of its high and comparable magnitudes of *CC* and low and comparable values of *RMSE* in respect of both the training and test set data. Next, the parameters, β, of the overall best model were refined further by using a standard nonlinear regression technique, namely Marquardt's algorithm (Marquardt, 1963) with a view to improve its prediction and generalization performance

8.4.2. Case Study I: GP-Based VLE Modeling of Ternary System 1, 2 Dichloroethane (1), Trichloroethylene (2), 1-Propanol (3)

The objective of this case study is to develop two GP models (*GP_model-I* and -*II*) predicting mole fractions of *1, 2-dichloroethane* (y_1), and *trichloroethylene* (y_2), in the vapor phase. Towards this goal, a total of 58 isobaric VLE data points at high temperature (352.65 –358.55 K) were collected (Gmehling and Onken, 1986) for the ternary system given in Table 8.A.1. These data consisting of the physiochemical properties (acentric factor) and the experimental conditions (temperature, liquid and vapor phase compositions) are given in Tables 8.4 and 8.A.1, respectively. From these data the training set (43 data patterns) was selected in a way such that it covers all the ranges of the experimental data and operating conditions.

(A) GP-based model for predicting mole fraction of 1, 2-dichloroethane in vapor phase (y_1)

The input space of the *GP_model-I* predicting the mole fraction of 1, 2-dichloroethane in vapor phase (y_1), contains six variables, namely, temperature (*T*); mole fractions of 1, 2-dichloroethane (x_1), and trichloroethylene (x_2) in liquid phase, acentric factors of 1, 2-dichloroethane (ω_1), trichloroethylene (ω_2), and 1-propanol (ω_3). By implementing the GP procedure given in section 8.4.1, the overall best GP-based model (*GP_model-I*) predicting the value of mole fraction of 1, 2-dichloroethane in the vapor phase (y_1) that was secured is given as:

$$y_1 = 0.3253 - \frac{1.476 \times 10^{-2}}{x_1} + 1.32\, x_2^2 \omega_2 + 0.1997\ \mathrm{T}\, x_1 \omega_1 \omega_3 - 0.4189\, x_2 - 11.95\, x_1 \quad (8.1)$$

This model (8.1) when subjected to the nonlinear regression using Marquardt's method (Marquardt, 1963) yielded following expression (*GP–Marquardt_model–I*):

$$y_1 = 0.314 - \frac{1.4 \times 10^{-2}}{x_1} + 1.12\, x_2^2 \omega_2 + 0.219\ \mathrm{T}\, x_1 \omega_1 \omega_3 - 0.375\, x_2 - 13.183\, x_1 \quad (8.2)$$

As can be seen in Eq. (8.2), the Marquardt's method has fitted a different set of parameters to the GP-based model. The predictions by *GP_model-I* have yielded high and comparable magnitudes of the coefficient of correlation (CC_{trn} = 0.997; CC_{tst} = 0.998), and low and comparable values of the root mean square error ($RMSE_{trn}$= 1.13×10^{-2}; $RMSE_{tst}$ = 1.08×10^{-2}) in respect of both the training and test set data. A comparison of the prediction accuracies and generalization performance of *GP_model-I* with that of other four models, namely, GP-Marquardt, Wilson (Gmehling and Onken, 1986), NRTL (Gmehling and Onken, 1986), and UNIQUAC (Gmehling and Onken, 1986) is provided in Table 8.5. The *CC* and *RMSE* magnitudes listed in this table clearly reveal that the proposed *GP_Model-I* has better prediction accuracy and generalization capability than its all four competing models.

Table 8.5: Statistical analysis and comparison of prediction generalization performance of *GP_model-I* with other four models for estimation of mole fraction of 1, 2-dichloroethane in vapor phase (y_1)

Type of model	Training set		Test set	
	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$
GP_model–I	0.997	1.13×10^{-2}	0.998	1.08×10^{-2}
GP–Marquardt_model–I	0.997	1.46×10^{-2}	0.998	1.31×10^{-2}
Wilson_model–I	0.996	1.34×10^{-2}	0.998	1.02×10^{-2}
NRTL_model–I	0.995	1.43×10^{-2}	0.997	1.10×10^{-2}
UNIQUAC_model–I	0.995	1.42×10^{-2}	0.997	1.11×10^{-2}

Figure 8.1 shows a comparison of the predictions of the mole fraction of 1, 2-dichloroethane in vapor phase (y_1) by the *GP_model-I*, with their experimental counterpart. An excellent match between the experimental and model predicted values of the mole fraction of 1, 2-dichloroethane in vapor phase pertaining to both training and test set data clearly establishes an outstanding prediction and generalization performance by *GP_model-I*.

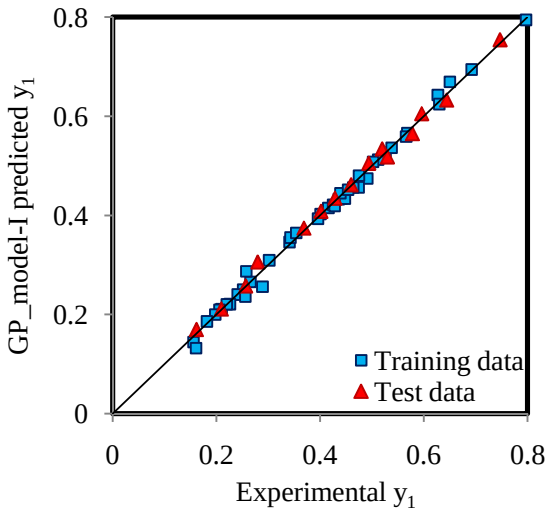


Figure 8.1: Parity plot of the experimental versus GP_model-I predicted mole fraction of 1, 2-dichloroethane in vapor phase (y_1) of case study I

(B) GP-based modeling of mole fraction of trichloroethylene in vapor phase (y_2)

The GP-based model (hereafter termed *GP_model-II*) for predicting the mole fraction of trichloroethylene in vapor phase (y_2) was developed using same inputs (see Tables 8.4 and 8.A.1) as employed in the development of the model for predicting the mole fraction of 1, 2-dichloroethane in the vapor phase (y_1). The overall best GP-based model (*GP_model-II*), which resulted in the high and comparable magnitudes of *CC* and low and comparable magnitudes of *RMSE* in respect of both training and test sets, is given as:

$$y_2 = 1.941 + 3.455 \times 10^{-2}\,\mathrm{T}x_2 + \frac{3.674\,\omega_3}{(8242x_1)^{\omega_1}} - 6.02 \times 10^{-3}\,\mathrm{T} - 11.13x_2 - 2.293x_2^2\omega_2 \quad (8.3)$$

The equation (8.3) when subjected to nonlinear regression using Marquardt's method (Marquardt, 1963) yielded following equation (*GP-Marquardt_model-II*) with a different set of parameter magnitudes.

$$y_2 = 1.24 + 0.025\,\mathrm{T}\,x_2 + \frac{2.342\,\omega_3}{(9667x_1)^{\omega_1}} - 4.0 \times 10^{-3}\mathrm{T} - 7.903\,x_2 - 2.106\,x_2^2\omega_2 \quad (8.4)$$

The *CC* magnitude in respect of the output (y_2) predicted by *GP_model-II* and the corresponding desired (experimental) values for the training and test sets are 0.998 and 0.994, respectively, and the corresponding *RMSE* magnitudes are 1.01×10^{-2} and 1.69×10^{-2}, respectively. From the high (low) and comparable values of *CC* (*RMSE*) for both the training and test set data, it can be concluded that the GP-based model has exhibited an excellent performance in predicting and generalizing the mole fraction magnitudes of trichloroethylene in the vapor phase. Next, performance of *GP_model-II* was compared with that of the corresponding *GP-Marquardt_model-II* and three activity coefficient models, namely, Wilson, NRTL, and UNIQUAC. This comparison made in terms of the *CC* and *RMSE* values is provided in Table 8.6. Here, the results reveal that the prediction and generalization performance of the *GP_model-II* is comparable with that of the competing four models.

Table 8.6: Statistical analysis and comparison of prediction generalization performance of *GP_model-II* with other four models for estimation of mole fraction of trichloroethylene in vapor phase (y_2)

Type of model	Training set		Test set	
	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$
GP_model–II	0.998	1.01×10^{-2}	0.994	1.69×10^{-2}
GP–Marquardt_model–II	0.995	1.28×10^{-1}	0.988	1.27×10^{-1}
Wilson_model–II	0.996	1.38×10^{-2}	0.998	6.96×10^{-3}
NRTL_model–II	0.996	1.45×10^{-2}	0.998	7.19×10^{-3}
UNIQUAC_model–II	0.996	1.48×10^{-2}	0.998	7.46×10^{-3}

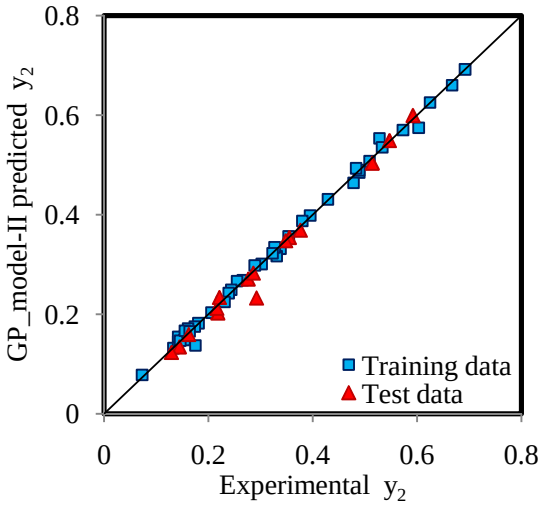


Figure 8.2: Parity plot of the experimental versus GP_model-II predicted mole fraction of trichloroethylene in vapor phase (y_2) of case study I

Figure 8.2 displays the parity plot of the *GP_model-II* predicted values of the mole fraction of trichloroethylene in the vapor phase (y_2) and their experimental counterparts. As can be seen, there is a very close agreement between the experimental and model predicted values pertaining to the training as also test set data thus supporting the earlier observation of an excellent prediction accuracy and generalization performance by *GP_model-II*. The results of this case study essentially indicate that the GP formalism can serve as an additional modeling method capable of yielding comparable or even better prediction accuracy and generalization capability when compared with existing methods for VLE prediction.

8.4.3 Case Study II: GP-Based VLE Modeling of Group of Three Binary Systems, namely, (i) Tetrachloromethane (1) – Ethanol (2), (ii) Tetrachloromethane (1) – 1 – Propanol (2), and (iii) Tetrachloromethane (1) -1-Butanol(2)

The objectives of this case study is as follows

- To develop a single optimal GP-based model for a group of three binary systems for predicting the vapor phase composition of the common first component, namely, tetrachloromethane (y_1) in each of the three systems.

In this case study, the second components of three binary systems belong to the homologous series of alcohol. The 96 experimental data points (Gmehling and Onken, 1986; Gmehling et al., 1986) pertaining to the three binary systems cover temperature and pressure ranges of 293.15–343.15 K, and 5.179–101.434 kP_a, respectively (also see Table 8.A.2). These data points were divided into the training (72 patterns) and test (24 patterns) data sets to, respectively, construct and test a single optimal *GP_model-III* predicting vapor phase composition of the common first component of the three binary systems.

The input space of proposed *GP_model-III* consisted of two critical properties, namely, critical temperature, (T_{c_2}) (K), and critical pressure, (P_{c_2}) (kP_a) of three components, namely, ethanol, 1-propanol, and 1-butanol, and three intensive thermodynamic variables, namely, the mole fraction of tetrachloromethane in liquid phase (x_1), pressure (P) (kP_a), and temperature (T) (K) (see Tables 8.4 and 8.A.2, respectively). By implementing the GP procedure described in sections 2.2.2 and 8.4.1, the following overall optimal GP-based model was obtained for the prediction of the mole fraction of tetrachloromethane in vapor phase (y_1):

$$y_1 = 5.016 + 2.028 \times 10^{-3} x_1^3 \mathrm{P} - 2.493\times10^{-4} P_{c_2} - 3.376\times10^{-3} T_{c_2} - 3.465 \times10^{-3} \mathrm{T} - 0.5851(1.06 \times 10^{-3})^{(2.028\times10^{-3} x_1 T_{c_2})} \quad (8.5)$$

This model when subjected to parameter fine-tuning by the nonlinear regression using Marquardt's method (Marquardt, 1963) yielded following expression (*GP-Marquardt_model-III*).

$$y_1 = 4.666 + 2\times10^{-3} x_1^3 \mathrm{P} - 1.85\times10^{-4} P_{c_2} - 3\times10^{-3} T_{c_2} - 4\times10^{-3} \mathrm{T} - 0.582(0.792)^{(0.064 x_1 T_{c_2})} \quad (8.6)$$

The performance of *GP_model-III*, in making accurate predictions pertaining to each of the three binary systems was compared with that of the two thermodynamic models, namely, VanLaar (Gmehling and Onken, 1986; Gmehling et al., 1986) and NRTL (Gmehling and Onken, 1986; Gmehling et al., 1986). The above stated comparison made in terms of *CC* and *RMSE* values pertaining to the predictions of mole fraction of tetrachloromethane in vapor phase (y_1), made by *GP_model-III*, *GP-Marquardt_model-III*, and models of VanLaar, and NRTL is provided in Table 8.7. The magnitudes of the stated statistical quantities clearly suggest that *GP_model-III* possesses an excellent prediction accuracy and generalization capability. It is also seen that (a) the prediction and generalization performance of GP_*Model-III* closely matches with that of the VanLaar, and NRTL models and (b) the performance of the *GP-Marquardt_model-III*, is only marginally inferior than the other three competing models.

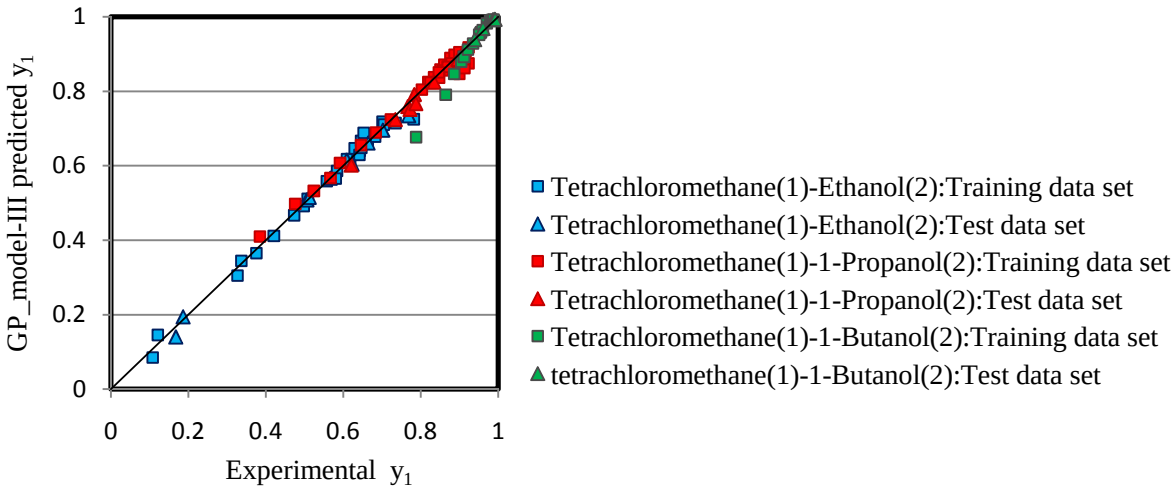


Figure 8.3: Parity plot of the experimental versus GP_model-III predicted mole fraction of tetrachloromethane (CCl_4) in vapor phase (y_1) of case study II

Figure 8.3 shows a comparison of the experimental values of the mole fraction of tetrachloromethane in vapor phase (y_1) for individual binary system with that the corresponding predictions made by *GP_model-III*. All points falling on or very close to the 45^o line indicates a very good match between the experimental and model predicted y_1 values. In this case study, it is noteworthy to note that a single GP-based model is capable of accurately predicting VLE for multiple binary systems thus saving the efforts involved in developing a separate model for each binary system in the group.

Table 8.7: Statistical analysis and comparison of prediction generalization performance of *GP_model-III* with other three models for estimation of mole fraction of tetrachloromethane (CCl_4) in vapor phase (y_1)

Type of model/ binary system	CCl_4(1)- ethanol(2)				CCl_4 (1) - 1-propanol(2)				CCl_4(1)- 1-butanol(2)			
	Training set		Test set		Training set		Test set		Training set		Test set	
	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$
GP_model–III	0.994	1.83×10^{-2}	0.997	1.54×10^{-2}	0.993	1.94×10^{-2}	0.995	1.41×10^{-2}	0.995	3.63×10^{-2}	0.996	4.35×10^{-3}
GP–Marquardt_model–III	0.994	7.02×10^{-2}	0.995	6.44×10^{-2}	0.987	3.08×10^{-2}	0.972	2.21×10^{-2}	0.994	4.04×10^{-2}	0.991	1.61×10^{-2}
VanLaar_model–III	0.997	1.21×10^{-2}	0.997	1.36×10^{-2}	0.996	1.69×10^{-2}	0.997	2.06×10^{-2}	0.995	5.35×10^{-3}	0.996	3.99×10^{-3}
NRTL_model–III	0.999	7.92×10^{-3}	0.999	6.75×10^{-3}	0.997	1.83×10^{-2}	0.997	2.05×10^{-2}	0.998	3.77×10^{-3}	0.999	1.42×10^{-3}

8.4.4 Case Study III: GP-Based VLE Modeling for Group of three Binary Systems, namely, (i) Ethanol (1) – Ethyl acetate (2), (ii) 1-Propanol (1) – Propyl acetate (2), and (iii) 1-Butanol (1) - Butyl acetate (2)

The objectives of this case study are as follows:

- To propose a single GP-based optimal model (*GP_model-IV*) to predict mole fractions in vapor phase (y_1) of the first components of a group of three binary systems; these components are: ethanol, 1-propanol, and 1-butanol. It may be noted, that (a) the first (second) components of all the three systems are homologs of the alcohol (acetate) series, and (b) conventionally, a separate model is developed for the VLE prediction pertaining to each binary system.
- To test the extrapolation ability of the developed GP-based model to predict mole fractions in vapor phase (y_1) of the fourth binary system, namely, 1-pentanol (1) – pentyl acetate (2). Note that the first and second components of this system are the higher homologs of corresponding components of the three alcohol-acetate binary systems whose data were considered in developing the GP-based model.

In this study, 130 experimental VLE data points pertaining to the group of above stated four binary systems belonging to alcohol-acetate homologous series were compiled from DECHEMA, VLE data series (Gmehling and Onken, 1986; Gmehling et al., 1986). Details of the experimental data used in the GP model building are given in Table 8.A.3. The GP-based model was developed using combined data of three binary systems, namely, (a) ethanol (1) – ethyl acetate (2), (b) 1 – propanol (1) – propyl acetate (2), and (c) 1-butanol (1) – butyl acetate (2). Apart from testing the model for its generalization ability using a test set consisting of data of the stated three binary systems, the model's extrapolation ability was tested using a validation set consisting of data of the fourth binary system, namely, 1-pentanol (1) –pentyl acetate (2).

For developing the GP-based model for the prediction of mole fraction of first components in vapor phase ($y_1's$), following variables and parameters were selected as inputs: (i) acentric factor of the first component, namely, ethanol/1-propanol/ 1-butanol (ω_1) of the binary system, (ii) acentric factor of the second component,

namely, ethyl acetate/propyl acetate/butyl acetate (ω_2) of the binary system, (iii) liquid phase mole fractions of first components namely, ethanol/1-propanol/1-butanol (x_1) of the binary system, (iv) pressure (P) (kP_a), and (v) temperature (T) (K), (see Tables 8.4 and 8.A.3, respectively). By implementing the GP procedure described in sections 2.2.2 and 8.4.1, the following best data-fitting expression (*GP_model-IV*) yielding high (low) magnitudes of *CC* (*RMSE*) in respect of its predictions pertaining to both training and test set data was obtained.

$$y_1 = 0.03333 + 0.5564\, x_1\omega_2 + 1.066\, x_1^3 + 1.736 \times 10^{-5} T^2 x_1\omega_1 - 0.0009673\, x_1 P - 0.004603\, T\, x_1^2 \quad (8.7)$$

For fine-tuning of its parameters, this model was subjected to the nonlinear regression using Marquardt's method (Marquardt, 1963), which resulted in following expression with small changes in the parameter values:

$$y_1 = 0.035 + 0.574\, x_1\omega_2 + 1.064 x_1^3 + 1.721 \times 10^{-5} \mathrm{T}^2 x_1\omega_1 - 0.001 x_1 \mathrm{P} - 0.005\ \mathrm{T}\ x_1^2 \quad (8.8)$$

The prediction and generalization performance of GP_ *model-IV* was compared with that of the two activity coefficient models, namely, Van Laar and NRTL, as also *GP-Marquardt_model-IV*. The results of this comparison in terms of *CC* and *RMSE* values are provided in Table 8.8. Specifically, the stated two statistical measures were evaluated considering separately the data of each of the three binary systems, namely, (a) ethanol and ethyl acetate, (b) 1 – propanol and – propyl acetate, and (c) 1-butanol – butyl acetate. It is observed in this table, that predictions made by *GP_model-IV* have yielded high *CC* (≈ 0.999) and low *RMSE* values (≈5.30×10^{-3}) in respect of both the training and test sets for each of the three binary systems. It is also noticed that GP_*model-IV* possesses better prediction accuracy and generalization capability than the activity coefficient models as also *GP-Marquardt_model-IV*.

Table 8.8: Statistical analysis and comparison of prediction generalization performance of *GP_model-IV* with other three models for estimation of mole fraction of ethanol, 1-propanol, and 1-butanol in vapor phase (y_1)

Type of model/ binary system	Ethanol (1) - Ethyl acetate (2)				1-Propanol(1) - Propyl acetate (2)				1-Butanol (1)-Butyl acetate (2)			
	Training set		Test set		Training set		Test set		Training set		Test set	
	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$	CC_{trn}	$RMSE_{\text{trn}}$	CC_{tst}	$RMSE_{\text{tst}}$
GP_model–IV	0.998	1.46×10^{-2}	0.998	1.78×10^{-2}	0.999	5.37×10^{-3}	0.999	5.30×10^{-3}	0.999	6.80×10^{-3}	0.998	9.88×10^{-3}
GP–Marquardt_model–IV	0.996	6.06×10^{-2}	0.996	4.94×10^{-2}	0.998	7.73×10^{-2}	0.999	7.30×10^{-2}	0.997	8.15×10^{-2}	0.998	7.62×10^{-2}
VanLaar_model–IV	0.999	9.25×10^{-3}	0.998	1.19×10^{-2}	0.999	8.79×10^{-3}	0.999	8.91×10^{-3}	0.998	1.04×10^{-2}	0.998	1.04×10^{-2}
NRTL_model–IV	0.999	9.19×10^{-3}	0.998	1.19×10^{-2}	0.999	8.63×10^{-3}	0.999	8.74×10^{-3}	0.998	9.81×10^{-3}	0.998	1.03×10^{-2}

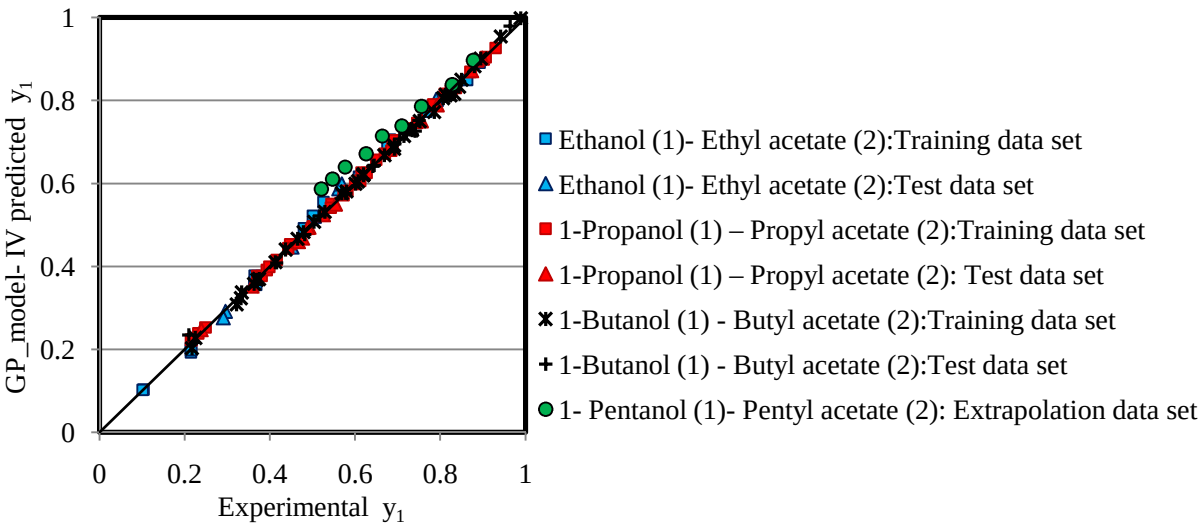


Figure 8.4: Parity plot of the experimental versus GP_model-IV predicted mole fraction of ethanol, 1-propanol, 1-butanol, and 1-pentanol in vapor phase (y_1) of case study III

The prediction and generalization ability of the proposed GP-model is shown graphically in terms of the parity plot in Figure 8.4. Here, experimental values of the mole fractions of component (1), namely, ethanol, 1-propanol, and 1-butanol in vapor phase, (y_1) are plotted against the corresponding *GP_model-IV* predicted values in respect of training and test set data. Figure 8.4 shows that all the data points of three binary systems (utilized in constructing *GP_model-IV*) are positioned on or close to the solid line indicating that model predictions closely match their targets. This figure also shows results pertaining to the extrapolation test of *GP_model-IV*. In the figure are plotted the experimental and model predicted values (circle symbol) of the mole fraction of 1-pentanol (1) (first component of the fourth binary system). As can be observed, there exists a reasonably close match between the experimental values and their model predictions, thus supporting the extrapolation ability of *GP_model-IV* to apply the learned trends in the data of three binary systems to make predictions for a new binary system.

Testing extrapolation ability of the GP_model-IV: Prediction of mole fraction of 1-pentanol in vapor phase (y_1)

In a special exercise, GP-*model-IV* was used for extrapolation. Specifically, it was utilized to make predictions of the mole fraction of the first component of the fourth binary system, namely, 1-pentanol and pentyl acetate. It may be noted that the training or test sets used in training and testing the *GP-model-IV* did not contain data of this binary system. The results of extrapolation exercise given in Table 8.9, indicate that *GP-model-IV* possesses a very good extrapolation capability with high magnitude for the *coefficient of correlation* (=0.998) and low magnitude of *root mean square error* (= 4.58×10^{-2}) (refer Table 8.9).

Table 8.9: Statistical analysis and comparison of prediction generalization performance of *GP_model-IV* with other three models to test its extrapolation capability on fourth binary system, namely, 1-pentanol (1) –pentyl acetate (2) to predict vapor phase composition of 1-pentanol (y_1).

Type of model/ binary system for extrapolation	1-Pentanol (1) –Pentyl acetate(2)	
	CC	*RMSE*
GP_model–IV	0.998	4.58×10^{-2}
GP–Marquardt_model–IV	0.998	5.97×10^{-2}
VanLaar_model–IV	0.998	6.36×10^{-3}
NRTL_model–IV	0.998	6.42×10^{-3}

8.5 CONCLUSION

In this work, an AI-based modeling strategy, namely, genetic programming has been utilized to develop models for the prediction of vapor liquid equilibria. Among various CI-based methods, GP possesses several novel and attractive characteristics and, yet, it remains an infrequently used data-driven modeling technique when compared with ANNs and SVR. In this investigation, three case studies have been conducted wherein four models have been developed for the prediction of vapor phase composition. The specific systems studied are as follows: (a) a ternary system (case study I), (b) a group of three binary systems wherein first component is common and

the second components are homologs of an alcohol series (case study II), and (c) a group of three binary systems consisting of first and second components belonging to alcohol and acetate homologous series (case study III). The predictors of these models include temperature, pressure, critical temperature, critical pressure, acentric factor, and liquid phase composition. The experimental data from DECHEMA chemistry data series were utilized to develop the stated four models. To examine whether model parameters could be fine-tuned further, the GP based models were subjected to nonlinear parameter estimation using Marquardt's method. The performance of the GP-based models was compared rigorously with that of a number of classical thermodynamic models, namely, Van Laar, Wilson, NRTL, and UNIQUAC as also GP-Marquardt models. Prediction accuracies and generalization performance of all developed models are verified to be better compared with the available prediction of thermodynamic models. Values of correlation coefficient (*CC*), root mean squared error (*RMSE*) show that in general, the developed GP-based models gives out better results than thermodynamic models, namely, Van Laar, Wilson, NRTL, and UNIQUAC as also GP- Marquardt's model for estimation of vapor phase composition of ternary and group of binary mixtures.

The novelty of this study is as follows.

- A rigorous search of the literature indicates that this is the first study, wherein GP strategy has been used innovatively for VLE predictions.
- A single optimal GP-based model (*GP_model-III*) has been developed for a group of three binary systems—with a common first component—to predict vapor phase composition of individual binary system.
- A single model (GP_*model-IV*) has been developed for a group of three binary systems—with their first and second components belonging to alcohol and acetate homologous series, respectively— to predict vapor phase composition of individual binary system. Also, the extrapolation capability of the model was tested on a totally different binary system containing homologs of alcohol and acetate. The results of this case study show that as regards with the three binary systems, *GP_model-IV* possesses an excellent prediction accuracy and generalization capability. Moreover, and notably, it also possesses extrapolation ability as confirmed by its closely matching predictions of the

mole fraction in the vapor phase of a totally different binary system containing higher homologs of the alcohol-acetate series.

- The prediction accuracies of the GP-based models reported here are as good as or better than the conventional thermodynamic models used in the VLE prediction. Also, GP-based models are less complex (parsimonious), easier to grasp, and more convenient to deploy in a practical setting.

The GP-based VLE modeling approach illustrated here can be gainfully extended to develop similar type of models for numerous other industrially important binary and ternary systems.

NOMENCLATURE

P_{ci} critical pressure of i^{th}component

T_{ci} critical temperature of i^{th}component

x_i liquid phase composition of i^{th}component

y_i vapor phase composition of i^{th}component

ω_i acentric factor of i^{th}component

Appendix 8.A

Table 8.A.1: Data source and ranges of experimental conditions regarding ternary system used in case study-I for generating GP- based model-I and II

System	Temperature (T) (K)	Pressure (P) (kP_a)	Mole fraction of 1,2-dichloroethane in liquid phase (x_1)	Mole fraction of trichloroethylene in liquid phase (x_2)	Mole fraction of 1,2-dichloroethane in vapor phase (y_1)	Mole fraction of trichloroethylene in vapor phase (y_2)	No of data patterns	Reference
1, 2-dichloroethane (1) trichloroethylene (2) 1-propanol (3)	352.65-358.55	101.325	0.103 – 0.819	0.051- 0.785	0.156 - 0.797	0.073 – 0.692	58	Gmehling and Onken (1986)

Table 8.A.2: Data source and ranges of experimental conditions regarding three different binary systems used in case study-II for generating GP based model-III

System	Temperature (T) (K)	Pressure (P) (kP_a)	Mole fraction of CCl_4 in liquid phase, (x_1)	Mole fraction of CCl_4 in vapour phase, (y_1)	Number of data patterns	Reference
CCl_4(1)- ethanol(2)	318.15	25.695 – 046.792	0.0212 – 0.9541	0.1210 – 0.7822	13	Gmehling and Onken (1986)
	323.15	40.543 – 057.315	0.1000 – 0.9000	0.3370 – 0.7050	09	Gmehling and Onken (1986)
	338.15	64.160 – 101.434	0.0237 – 0.9483	0.1075 – 0.7688	15	Gmehling and Onken (1986)
CCl_4 (1)- 1-propanol(2)	293.15	05.179 – 012.739	0.0906 – 0.9030	0.6210 – 0.9240	09	Gmehling and Onken (1986)
	303.15	08.653 – 020.132	0.0906 – 0.9030	0.5670 – 0.9100	09	Gmehling and Onken (1986)
	313.15	14.299 – 030.537	0.0906 – 0.9030	0.5240 – 0.9240	09	Gmehling and Onken (1986)
	343.15	49.183 – 091.966	0.0825 – 0.9630	0.3850 – 0.9130	11	Gmehling and Onken (1986)
CCl_4(1)- 1-butanol(2)	308.15	07.613 – 023.398	0.0989 – 0.9934	0.7885 – 0.9923	21	Gmehling et al. (1986)

Table 8.A.3: Data source and ranges of experimental conditions regarding four different binary systems used in case study-III for generating GP based model-IV

System	Pressure $(P)(kP_a)$	Temperature (T) (K)	Mole fraction of component (1) in liquid phase, (x_1)	Mole fraction of component (1) in vapor phase, (y_1)	No. of data patterns	Reference
Ethanol (1) – Ethyl acetate (2)	101.325	345.33 –349.85	0.101 – 0.8924	0.086 – 0.965	24	Gmehling and Onken (1986)
1-Propanol (1) – Propyl acetate (2)	101.325	367.85–371.15	0.136 – 0.9520	0.216 – 0.930	20	Gmehling and Onken (1986)
	079.993	361.21–363.88	0.162 – 0.9350	0.239 – 0.899	09	Gmehling and Onken (1986)
	053.329	350.22 – 353.03	0.162 – 0.9350	0.232 – 0.889	09	Gmehling and Onken (1986)
	026.66 5	333.13 – 337.17	0.162 – 0.9350	0.215 – 0.874	09	Gmehling and Onken (1986)
1-Butanol (1) – Butyl acetate (2)	101.325	389.35 – 394.90	0.109 – 0.9950	0.217– 0.989	33	Gmehling et al. (1986)
	022.065	349.55 –351.45	0.161 – 0.8730	0.210 – 0.807	07	Gmehling et al. (1986)
	006.666	323.85–327.85	0.180 – 0.9210	0.225 – 0.833	10	Gmehling et al. (1986)
1-Pentanol(1) – Pentyl acetate(2)	100.765	409.60 –413.70	0.456 – 0.8700	0.521 – 0.878	09	Gmehling et al. (1986)

REFERENCES

Anderko, A. (1990). Equation-of-State Methods for the Modelling of Phase Equilibria, *Fluid Phase Equilibria*, *61*, 145-225.

Anderson, T. F., and Prausnitz, J. M. (1978). Application of the UNIQUAC equation to calculation of multicomponent phase equilibria. 1. Vapor-liquid equilibria. *Industrial & Engineering Chemistry Process Design and Development*, *17*(4), 552-561.

Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments*, *65*(6), 1803-1832.

Derr, E. L., and Deal, C. H. (1969, September). Analytical solutions of groups: Correlation of activity coefficients through structural group parameters. In *Inst. Chem. Eng. Symp. Ser* (Vol. 32, No. 3, p. 40).

Dohrn, R., and Brunner, G. (1995). High-pressure fluid-phase equilibria: Experimental methods and systems investigated (1988–1993). *Fluid Phase Equilibria*, *106*(1), 213-282.

Fredenslund, A., Gmehling, J., and Rasmussen, P. (1977). *Vapor-Liquid Equilibria using UNIFAC: A Group-Contribution Method*. 1st ed., Elsevier Scientific Publishing Company, Amsterdam, Oxford, NewYork.

Ganguly, S. (2003). Prediction of VLE data using radial basis function network. *Computers & chemical engineering*, *27*(10), 1445-1454.

Gebreyohannes, S., Yerramsetty, K., Neely, B. J., and Gasem, K. A. (2013). Improved QSPR generalized interaction parameters for the nonrandom two-liquid activity coefficient model. *Fluid Phase Equilibria*, *339*, 20-30.

Ghaemi, A., Shahhoseini, S., Ghannadi Marageh, M., and Farrokhi, M. (2008). Prediction of vapor-liquid equilibrium for aqueous solutions of electrolytes using artificial neural networks. *Journal of Applied Sciences*, *8*, 615-621.

Gmehling, J., and Onken, U. (1986). *Vapor-Liquid Equilibrium Data Collection. Organic Hydroxyl Compounds: Alcohols*. Chemistry Data Series, Volume I, Part 2a, Frankfurt, Germany: DECHEMA.

Gmehling, J., Onken, U., and Arlt, W. (1986). *Vapor-Liquid Equilibrium Data Collection*. *Organic Hydroxyl Compounds: Alcohols and Phenols*. Chemistry Data Series, Volume I, Part 2b, Frankfurt, Germany: DECHEMA.

Goel, P., Bapat, S., Vyas, R., Tambe, A., and Tambe, S. S. (2015). Genetic programming based quantitative structure–retention relationships for the prediction of Kovats retention indices. *Journal of Chromatography A, 1420,* 98-109.

J. Koza, J. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA.

Karimi, H., and Yousefi, F. (2007). Correlation of vapour liquid equilibria of binary mixtures using artificial neural networks. *Chinese Journal of Chemical Engineering*, *15*(5), 765-771.

Lashkarbolooki, M., Shafipour, Z. S., Hezave, A. Z., and Farmani, H. (2013). Use of artificial neural networks for prediction of phase equilibria in the binary system containing carbon dioxide. *The Journal of Supercritical Fluids, 75*, 144-151.

Lashkarbolooki, M., Vaferi, B., Shariati, A., and Hezave, A. Z. (2013). Investigating vapor–liquid equilibria of binary mixtures containing supercritical or near-critical carbon dioxide and a cyclic compound using cascade neural network. *Fluid Phase Equilibria, 343*, 24-29.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, *11*(2), 431-441. http://dx.doi.org/10.1137/0111030.

Martin, J. J. (1979). Cubic Equations of State-Which? , *Ind. Eng. Chem. Fundam, 18,* 81-97.

Mesbah, M., Soroush, E., Azari, V., Lee, M., Bahadori, A., and Habibnia, S. (2015). Vapor liquid equilibrium prediction of carbon dioxide and hydrocarbon systems using LSSVM algorithm. *The Journal of Supercritical Fluids*, *97*, 256-267.

Mohanty, S. (2005). Estimation of vapor liquid equilibria of binary systems, carbon dioxide–ethyl caproate, ethyl caprylate and ethyl caprate using artificial neural networks. *Fluid Phase Equilibria*, *235*(1), 92-98.

Nguyen, V. D., Tan, R. R., Brondial, Y., and Fuchino, T. (2007). Prediction of vapor–liquid equilibrium data for ternary systems using artificial neural networks. *Fluid phase equilibria*, *254*(1), 188-197.

Patil-Shinde, V., Kulkarni, T., Kulkarni, R., Chavan, P. D., Sharma, T., Sharma, B. K., Tambe, S. S., and Kulkarni, B. D. (2014). Artificial intelligence-based modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier. *Industrial & Engineering Chemistry Research*, *53*(49), 18678-18689.

Patil-Shinde, V., Saha, S., Sharma, B. K., Tambe, S. S., and Kulkarni, B. D. (2016). High ash char gasification in thermo-gravimetric analyzer and prediction of gasification performance parameters using computational intelligence formalisms. *Chemical Engineering Communications*, *203*(8), 1029-1044.

Peng, D. Y., and Robinson, D. B. (1976). A new two-constant equation of state. *Industrial & Engineering Chemistry Fundamentals*, *15*(1), 59-64.

Perry, R. H., and Green, D. W. (2007). *Perry's Chemical Engineers' Handbook*. 8th ed., McGraw-Hill, NewYork.

Poli, R., Langdon, W. B., McPhee, N. F., and Koza, J. R. (2008). *A Field Guide to Genetic Programming*. Available via lulu: http://lulu.com, http://www.gp-field-guide.org.uk .

Prausnitz, J. M., Lichtenthaler, R. N., and de Azevedo, E. G. (1998). *Molecular Thermodynamics of Fluid-Phase Equilibria*. Pearson Education, Prentice-Hall Inc. Upper saddle River, New Jersey.

Redlich, O., and Kwong, J. N. (1949). On the thermodynamics of solutions. V. An equation of state. Fugacities of gaseous solutions. *Chemical reviews*, *44*(1), 233-244.

Renon, H., and Prausnitz, J. M. (1969). Estimation of parameters for the NRTL equation for excess Gibbs energies of strongly nonideal liquid mixtures. *Industrial & Engineering Chemistry Process Design and Development*, *8*(3), 413-419.

Schmidt, M., and Lipson, H. (2012). Eureqa Formulize (Version 0.97) [Computer software]. Nutonian Inc., Cambridge, MA.

Sengers, J. V., Kayser, R. F., Peters, C. J. and White Jr, H. J. (Eds.) (2000). *Equations of State for Fluids and Fluid Mixtures*, Elsevier, Amsterdam.

Sharma, R., Singhal, D., Ghosh, R., and Dwivedi, A. (1999). Potential applications of artificial neural networks to thermodynamics: vapor–liquid equilibrium predictions. *Computers & chemical engineering*, *23*(3), 385-390.

Si-Moussa, C., Hanini, S., Derriche, R., Bouhedda, M., and Bouzidi, A. (2008). Prediciton of high-pressure vapor liquid equilibrium of six binary systems, carbon dioxide with six esters, using an artificial neural network model. *Brazilian Journal of Chemical Engineering*, *25*(1), 183-199.

Smith, J. M., Van Ness, H. C., and Abbott, M. M. (2005). *Introduction to Chemical Engineering Thermodynamics*. McGraw-Hill, NY.

Soave, G. (1972). Equilibrium constants from a modified Redlich-Kwong equation of state. *Chemical Engineering Science*, *27*(6), 1197-1203.

Tambe, S. S., Deshpande, P. B., Kulkarni, B. D. (1996). *Elements of Artificial Neural Networks with Selected Applications in Chemical Engineering, and Chemical & Biological Sciences*. Simulation & Advanced Controls, Inc., Louisville. 1996.

Tashvigh, A. A., Ashtiani, F. Z., Karimi, M., and Okhovat, A. (2015). A novel approach for estimation of solvent activity in polymer solutions using genetic programming. *Calphad*, *51*, 35-41.

Vaferi, B., Rahnama, Y., Darvishi, P., Toorani, A., and Lashkarbolooki, M. (2013). Phase equilibria modeling of binary systems containing ethanol using optimal feed-forward neural network. *The Journal of Supercritical Fluids*, *84*, 80-88.

Van der Waals, J. D. (1910). The equation of state for gases and liquids. *Nobel lectures in Physics*, *1*, 254-265.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Verma, D., Goel, P., Patil-Shinde, V., and Tambe, S. S. (2016, January). Use genetic programming for selecting predictor variables and modeling in process identification. In *IEEE explore, 2016 Indian Control Conference* (*ICC*) (pp. 230-237). IEEE.

Wilson, G. M. (1964). Vapor-liquid equilibria, correlation by means of a modified Redlich-Kwong equation of state. In *Advances in Cryogenic Engineering* (pp. 168-176). Springer US.

Wong, D. S., Orbey, H., and Sandler, S. I. (1992). Equation of state mixing rule for non ideal mixtures using available activity coefficient model parameters and that allows extrapolation over large ranges of temperature and pressure. *Industrial & engineering chemistry research*, *31*(8), 2033-2039.

Yamamoto, H., and Tochigi, K. (2007). Prediction of vapor–liquid equilibria using reconstruction—learning neural network method. *Fluid phase equilibria*, *257*(2), 169-172.

Yaws, C. L. (1999). *Chemical Properties Handbook*. McGraw-Hill, NewYork.

Zaid, S. (2012). Development of support vector regression (SVR)-based model for prediction of circulation rate in a vertical tube thermosiphon reboiler. *Chemical engineering science*, *69*(1), 514-521.

Zarenezhad, B., and Aminian, A. (2011). Predicting the vapor-liquid equilibrium of carbon dioxide+ alkanol systems by using an artificial neural network. *Korean Journal of Chemical Engineering*, *28*(5), 1286-1292.

Zurada, J. M. (1992). *Introduction to Artificial Neural Systems*. (Vol. 8), West Publ. Co., St. Paul.

Chapter 9

Thesis Conclusion

9.1 INTRODUCTION

In this chapter, the principle findings and conclusions reached thereof of the studies presented in this thesis are reported. Additionally, suggestions for the further research are presented.

Mathematical models of chemical reactions and processes are required for a variety of tasks in chemical engineering and technology, such as, prediction of reaction's steady-state and dynamic behavior, equipment design, operation, scale-up, control, fault detection and diagnosis, optimization, etc. Conventionally, two approaches, namely, *phenomenological* (first principles/ mechanistic) and *empirical* are employed for chemical process modeling. Modern day chemical processes are complex and are difficult to model phenomenologically, since the complete knowledge regarding the physico-chemical phenomena underlying their behavior—which is absolutely necessary for this type of modeling—is usually not available or tedious, time-consuming and costly to acquire via experiments. The nonlinear nature of chemically reacting systems makes first-principles modeling even more daunting. Being not demanding in terms of the availability of mechanistic knowledge, empirical modeling is an attractive alternative to first principles modeling; however, it has its own drawbacks such as the requirement that the form of the model to be fitted must be specified a priori before estimating the function parameters. This is, in general, a difficult task since in many chemical processes multiple variables influence the nonlinear phenomenon and the precise interactions between them are not fully known.

Commonly, *deterministic* gradient-based methods are used in process optimization. Invariably, these approaches require that the objective function (to be maximized/minimized) must be continuous, differentiable and smooth—a criterion difficult to fulfill especially in the case of exclusively data-driven reaction/process models. Gradient-based optimization methods also have a tendency to get stuck in a local optimum leading to sub-optimal solutions.

To overcome the stated drawbacks of phenomenological/empirical modeling approaches and deterministic optimization techniques, in this thesis artificial intelligence (AI) based modeling methods, namely, *artificial neural networks* (ANN),

and *genetic programming* (GP), and a machine learning (ML) based method termed *support vector regression*, have been employed for modeling a number of complex chemical reactions and processes, and developing a property estimation relation. The processes for which models have been developed belong to diverse fields, namely, thermal energy production, polymers, petroleum, water treatment, and separation processes.

An important application of process models is in optimization. Accordingly, an AI-based hybrid modeling-optimization strategy integrating GP formalism and a stochastic optimization method, namely, *genetic algorithms* (GA) has been used in optimizing a resin based adsorptive waste-water treatment process.

In addition to the novel AI-based modeling and optimization methodologies, two conventional methods, namely, *principal component analysis* (PCA), and *sensitivity analysis* (SA) have been utilized for reducing the dimensionality of the models' input spaces, and identifying influential input variables, respectively.

In chapters 1 and 2, a broad objective of the thesis, the need for utilizing artificial intelligence based modeling and optimization methods, and their detailed description are provided. The following section provides the rationale, salient features, and highlights of the studies reported in chapters 3 to 8.

9.2 OVERALL CONCLUSION

Chapter 3 deals with experimentation and modeling of a coal gasifier using artificial intelligence based methods. Coal gasification is a cleaner and an efficient alternative to the coal combustion for producing the syngas. The high-ash coals are found in a number of countries, and they form an important source for the gasification. In India also a major portion of the electricity is generated in coal-based thermal power stations. Accordingly, in this study, extensive gasification experiments were conducted in a pilot-plant scale fluidized-bed coal gasifier (FBCG) using high-ash coals from India. Specifically, the effects of eight coal and gasifier process related parameters on the four gasification performance variables, namely $CO+H_2$ generation rate, syngas production rate, carbon conversion, and heating value of the syngas, were rigorously studied. The data collected from these experiments were used in the FBCG modeling by utilizing two artificial intelligence (AI) strategies namely genetic programming (GP) and artificial neural networks (ANNs). A comparison of the GP and ANN-based models reveals that their output prediction accuracies and the

generalization performance vary from good to excellent. The novelty of this study is that (a) modeling of a coal gasification process wherein currently mined coal in India has been utilized, was conducted using state-of-the-art artificial intelligence systems, (b) the models can be used in designing, operating, and optimizing environmentally friendly coal gasifiers that use high ash coals, and (c) a rigorous literature search shows that this is the first study wherein the GP strategy has been employed for the data-driven modeling in the coal sciences and engineering.

Although high ash coals are routinely used in the energy generation, systematic gasification kinetic studies using chars derived from these coals are scarce. Accordingly, chapter 4 reports the development of the data-driven models for the gasification of chars derived from the high ash coals. Specifically, the models predict two significant gasification performance parameters, viz. gasification rate constant, and reactivity index. These models have been constructed using three computational intelligence (CI) methods, namely genetic programming (GP), multilayer perceptron (MLP) neural network (NN), and support vector regression (SVR). The data used in the modeling were collected by performing extensive gasification experiments in the CO_2 atmosphere in a thermo-gravimetric analyzer (TGA), using char samples derived from Indian coals with high ash content. Values of the stated gasification performance parameters were obtained by fitting the experimental data to the shrinking un-reacted core (SUC) model. All the CI-based models developed in this study possess an excellent prediction accuracy and generalization capability. The notable features of this study are: (a) For the first time, models have been developed to predict the kinetic char gasification rate constant (k_S), and reactivity index (r_I) magnitudes corresponding to the gasification of high ash Indian coals being mined currently, and (b) phenomenological and AI-based modeling are integrated to predict the char gasification kinetic parameters. The models developed here can be gainfully employed in the design and operation of not only fluidized bed gasifiers but also of fixed bed ones using high ash Indian as also other coals. Additionally, the models for determining the rate constant can be used for predicting the activation energies of the coal gasification reactions involving CO_2 in the temperature range of 900–1050^oC.

Choosing inputs (independent/causal variables) of a mathematical model, which are influential and, thus, significantly affect its output (dependent/response variable) is a tedious, time consuming and trial and error procedure in conventional empirical

modeling. In Chapter 5, genetic programming-based strategy has been suggested for simultaneously identifying the important predictor variables as also searching and optimizing an optimal data fitting function and its parameters. The said strategy has been illustrated by conducting two process identification case studies wherein the GP formalism has been shown to (a) identify the influential time-delayed inputs and outputs, and (b) simultaneously perform system identification using these influential predictors. The two chemical engineering systems chosen in the case studies are: (i) nonlinear height control system for a conical tank, and (ii) adiabatic nonlinear CSTR concentration control system. Chapter 5, clearly establishes that GP method is capable of automatically identifying those inputs which significantly influence the dependent variable. Thus, efforts involved in identifying the influential inputs are greatly reduced. Additional benefit of GP-based models is that they are in most cases less complex owing to which they exhibit better generalization performance than their more complex counterparts, such as ANNs and SVR. The GP-based process identification demonstrated in Chapter 5 is significantly useful in implementing model based control strategies.

The API gravity (^{o}API) is an important physicochemical property of crude oils. It is used routinely in the determination of their quality and properties. In Chapter 6, GP, MLP, and SVR methods have been used for developing models for predicting ^{o}API values of crude oils. These models use rarely utilized SARA (*Saturates*, *Aromatics*, *Resins*, and *Asphaltenes*) composition as inputs for the prediction of ^{o}API gravity. It has been observed that all three CI-based nonlinear models possess a better °API-value prediction accuracy and generalization capability than the currently available only model (Fan and Buckley, 2002) and its improved linear version (the modified- Fan and Buckley model). This result clearly indicates that the CI-based models are currently the best models for the SARA fractions based prediction of API gravity of crude oils.

Groundwater is an important source of the drinking water globally and often contaminated with harmful arsenic metalloid ions. Thus, removal of arsenic has gained importance while managing and treating water and wastewater. Chapter 7, reports usage of tannin-formaldehyde (TFA), and tannin-aniline-formaldehyde (TAFA) resins for the adsorptive removal of As(III) and As(V) ions from the contaminated water. Moreover, a fully artificial intelligence based hybrid strategy (termed “GP-GA”) integrating genetic programming and genetic algorithms, has

been utilized for the modeling and optimization of resin-based adsorptive removal of As(III)/As(V) ions from water and waste-water. The said strategy has led to significant improvements in the resin based adsorptive removal of As(III) and As(V) ions over that observed in experiments before performing the reaction optimization. The hybrid methodology utilized in this study can also be extended for the modeling and optimization of other contaminant-removal processes.

Among various CI-based methods, genetic programming possesses several novel and attractive characteristics, and yet it remains an unused data-driven modeling technique for VLE predictions when compared with ANNs and SVR. Accordingly, Chapter 8 presents a study wherein the GP-based data-driven modeling approach has been successfully employed for the first time for predicting the vapor-liquid equilibria (VLE) of a ternary and groups of binary mixtures. The predictor variables of these models include temperature, pressure, critical temperature, critical pressure, acentric factor, and liquid phase composition, whereas the output (response) variable was vapor phase composition. The experimental data from various sources were used in VLE modeling. Prediction accuracies and generalization performance of all the GP-based models were verified and found to be better compared with the predictions of the existing thermodynamic models, namely, Van Laar, Wilson, NRTL, and UNIQUAC.

The novel features of this study are as follows.

- A new method of modeling has been successfully applied for VLE predictions.
- A single optimal GP-based model has been developed for a group of three binary systems to predict vapor phase composition; the developed model has been used for predicting mole fraction in the vapor phase of first components of individual binary systems.
- A single optimal GP-based model has been developed for a group of three binary systems, and the extrapolation capability of the developed model has been successfully tested on a different (i.e. fourth) binary system, where the four binary systems belong to homologous series of alcohols and acetates.

The advantage of GP-based VLE models is that as compared to the thermodynamic models these are less complex, easier to grasp, and more convenient to deploy in a practical setting. There exists an enormous scope for applying the GP-based VLE modeling approach to other binary and ternary systems.

9.3 SUGGESTIONS FOR FUTURE RESEARCH

This thesis presents applications of artificial intelligence (AI) and machine learning (ML) methodologies for process/reaction modeling, process identification, property prediction, and optimization of various chemical systems. Among the three AI-based methods used in the stated modeling, GP despite its attractive characteristics has been a least used method in chemistry and chemical engineering/technology as compared to ANNs and SVR. Its major advantage being depending upon the nature of relationship, it is capable of fitting a linear or nonlinear data fitting function and its parameters without making any assumptions. There is still a huge scope for using GP in chemical engineering for applications such as VLE prediction of multi-component systems, model predictive control (MPC), etc. On the fundamental side, a GP algorithm/software capable of solving multiple input – multiple output (MIMO) modeling problems does not seem to be available in the open domain or commercially. Availability of such a software/algorithm will make it possible simultaneous fitting of multiple data-fitting functions.

In the thesis, AI-based models have been developed for gasification of Indian coals containing a high percentage of ash. These coals are used extensively in combustion applications and, thus, AI-based modeling of coal combustion processes will be beneficial for industries where the said technology is used in equipment such as boilers and steel furnaces.

Deep learning algorithms are increasingly used in training of multi-layer ANNs with hundreds/thousands of neurons. These are capable of mining huge number of data and therefore find applications in computation intensive image recognition, speech recognition, robotics, etc. In chemical engineering/technology, deep learning based ANNs can be used in image recognition applications such as recognizing size and shape of bubbles in a fluidized bed and identifying flow behavior of liquids and fluids.

To summarize, the potential of AI and ML methods is limit-less. These are likely to find ever increasing applications in areas such as plant/equipment safety, emission control, design of new drug and other molecules with desired properties, and chemical process centric internet of things (IOT).

List of Publications

Publications Received from the Work Presented in the Thesis

1. Patil-Shinde, V., Kulkarni, T., Kulkarni, R., Chavan, P. D., Sharma, T., Sharma, B. K., Tambe, S. S., and Kulkarni, B. D. (2014). Artificial intelligence-based modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier. ***Industrial & Engineering Chemistry Research***, *53*(49), 18678-18689.

2. Patil-Shinde, V., Saha, S., Sharma, B. K., Tambe, S. S., and Kulkarni, B. D. (2016). High Ash Char Gasification in Thermo-Gravimetric Analyzer and Prediction of Gasification Performance Parameters Using Computational Intelligence Formalisms. ***Chemical Engineering Communications***, *203*(8), 1029-1044.

3. Verma, D., Goel, P., Patil-Shinde, V., and Tambe, S. S. (2016, January). Use genetic programming for selecting predictor variables and modeling in process identification. In ***IEEE explore, 2016 Indian Control Conference*** (*ICC*) (pp. 230-237). IEEE. (ISBN: 978-1-4673-7992-2), doi: 10.1109/INDIANCC.2016.7441133.

4. Goel, P., Saurabh, K., Patil-Shinde, V., and Tambe, S. S. (2016). Prediction of °API Values of Crude Oils by Use of Saturates/Aromatics/Resins/Asphaltenes Analysis: Computational-Intelligence-Based Models. ***SPE Journal,*** doi:10.2118/184391-PA

5. Patil-Shinde, V., Mulani, K. B., Donde, K., Chavan, N. N., Ponrathnam, S., and Tambe, S. S. (2016). The Removal of arsenite [As (III)] and arsenate [As (V)] ions from wastewater using TFA and TAFA resins: Computational intelligence based reaction modeling and optimization. ***Journal of Environmental Chemical Engineering***, *4*(4), 4275-4286.

6. Patil-Shinde, V., Tambe, S. S. (2016). Genetic programming formalism for prediction of vapor-liquid equilibrium. (to be communicated to ***Fluid phase Equilibria***).