University of
South-Eastern Norway

www.usn.no

FMH606 Master's Thesis 2023

Energy and Environment Technology

Application of Machine Learning
in
Biogas Process

Ranjan Gaida

Faculty of Technology, Natural sciences and Maritime Sciences
Campus Porsgrunn

Course: FMH606 Master's Thesis, 2023

Title: Application of Machine Learning in Biogas Process

Number of pages: 67

Keywords: anaerobic digestion, machine learning, biogas production, artificial neural network, random forest, k-nearest neighbor.

Student: Ranjan Gaida

Supervisor: Gamunu L. Samarakoon Arachchige, Zahir Barahmand, Carlos Dimnarca

Summary:

Anaerobic digestion has become increasingly popular due to its potential to recover value-added resources from organic waste. To overcome the challenges of complex and non-linear relationships between the micro-organisms, machine learning-based approaches can be used for prediction, fault detection, optimization, and management of the overall process. Literature review suggested scoping review as the most effective method for understanding the topic. Search strings were developed on Scopus, Web of Science and Google scholar website to find research papers relevant to the thesis work, and 30 articles were selected. Limiting of research papers was followed by screening and trimming duplicate lists. Artificial Neural Network (ANN), Random Forest (RF) and K-Nearest Neighbor (KNN) were found to be best suitable method. Python code was written for ANN, RF and KNN, and data was analysed using packages like numpy, pandas, matplotlib and scikit-learn. Input variables such as influent sludge flow rate, total solids content, volatile solids content, alkalinity and volatile fatty acids were considered, and biogas production was the output variable. The prediction of the model's accuracy was done using two parameters, determination coefficient (R^2) and mean squared error (MSE). R^2 values for ANN, RF, KNN were found to be 59.3%, 62% and 51.5%, respectively, while MSE values were 6032492.244, 6695312.177 and 6854866.264. The best condition for R^2 is given by RF method, while MSE values favours ANN method. Literature review suggested R^2 gives the most accurate prediction as compared to other methods and hence RF method was considered to be the most suitable method for prediction of biogas.

Preface

This master's thesis presents the application of machine learning in biogas process and how we can use the machine learning algorithm in terms of effective prediction of biogas. This thesis was done under the guidance of Supervisor Prof. Gamunu L. Samarakoon Arachchige and Co-supervisor Zahir Barahmand and Carlos Dinmarca. I would also like to thank my friend Manjil Bista for the advice relating to coding in python. In this project, we fulfilled our tasks by effectively defining all the task objectives and demonstrating a model by using Artificial Neural Network, Random Forest and K-Nearest Neighbor method for the prediction of biogas.

Porsgrunn, 2023-05-15

Ranjan Gaida

Contents

List of Figures

List of Tables

Nomenclature

AD	Anaerobic Digestion
ADM1	Anaerobic Digestion Model No.1
ML	Machine Learning
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANFIS	Adaptive Neuro-Fuzzy Inference System
KNN	K-Nearest Neighbors
RF	Random Forest
LR	Logistic Regression
SVM	Support Vector Machine
RSM	Response Surface Methodology
ELM	Extreme Learning Machine
XGBoost	Extreme Gradient Boosting
POME	Palm Oil Mill Effluent
TMF	Total Mixed Feed
TSC	Total Solids Content
TVS	Total Volatile Solids Content
HRT	Hydraulic Retention Time
$CaCO_3$	Calcium Carbonate
CH_3COOH	Acetic Acid
R^2	Determination Coefficient
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
C	Carbon
N	Nitrogen
ReLU	Rectified Linear Activation Unit
RR	Recirculation Ratio
BMP	Bio-Methane Potential

1 Introduction

1.1 Background

Due to the waste produced by both domestic and industrial activity, both developed and emerging nations are searching for alternate energy sources. Currently, the majority of the primary energy supply in the globe comes from fossil fuels. However, due to the adverse effects of fossil fuels on the environment and the misuse of natural resources, public focus has shifted toward renewable energy sources to assure a sustainable future for energy production. In recent years, there has been an increase in interest in the use of biogas as a viable energy source to reduce greenhouse gas emissions.

Biogas production from the anaerobic digestion processes depends on parameters like retention time, pH, composition of medium, temperature inside the digestor tank, working pressure of the digestor tank, volatile fatty acids, etc [1]. For studying of models to investigate the complex and nonlinear relationship, Machine learning is considered having the potential to be used for predicting and controlling the performance of anaerobic digesters and has emerged to be an innovative tool for development of models [2]. Machine learning allows computers to find hidden information's without being explicitly programmed on where to look by using algorithms that iteratively learn from data. In the literature, several researchers have suggested several creative methods as successful and promising strategies for modeling biogas process. For simulating the complex and nonlinear interactions of the AD process, a number of machine learning techniques have been used, like support vector machines, adaptive neuro-fuzzy inference systems, k-nearest neighbors , random forests , and artificial neural networks [3]. In the study conducted by Tufaner and Demirci et al., a three-layer artificial neural network and nonlinear regression models were used to predict the performance of biogas production in controlled laboratory-scale experiment [4]. In an industrial-scale co-digestion facility, random forest and extreme gradient boosting-XGBoost has been used effectively by De Clercq et al., [5] while Zareei and Khodaei et al., used adaptive neuro-fuzzy inference system to model and optimize the biogas production from cow manure and maize straw in a pilot-scale study [6].

Chen et al. conducted an experiment where ANN and RSM methodology was used for modelling of methane production and H_2S content by using two years of industrial-scale plant data from Pahang, Malaysia. Input parameters like pH, temperature, and recirculation ratio were used for total treated effluent and bottom sludge to raw POME. Determination coefficient (R^2) and root mean squared error (RMSE) was used to evaluate the fitness of all models and result demonstrated that ANN was superior to RSM model with R^2 of 0.9762 and 0.85 respectively. The key interaction factors of methane yield were found to be temperature and recirculation ratio (RR), which was verified by a Pareto chart. Maintaining RR ratio at optimum level is key to achieving high methane yield with good stability, considering the trade-off between operating cost and revenue [7]. Khashaba et al., conducted the study where the cumulative methane production (CMP) from anaerobic digestion of sewage sludge altered with biochar was modelled using an ANN based on data compiled from 51 biomethane potential tests (BMP). With an R^2 of 0.9924, various forms of sewage sludge and biochar have been successfully predicted under both mesophilic and thermophilic environments. According to the findings, operating conditions have a greater impact on CMP, and CMP is strongly connected with both the physical characteristics and chemical composition of biochar, with chemical composition having the main influence [8]. As per the review done by Guo et al., the studies examine anaerobic digestion, thermal treatment, composting, and landfilling after concentrating on municipal solid waste management between 2003 and 2020. The artificial neural network is the most popular model that has been successfully utilized to solve non-linear organic solid waste problems [9]. Similarly, review done by Joshi et al., for the solid waste management for the activities like composting, incineration, pyrolysis, gasification, landfill, and anaerobic digestion came up to the conclusion that ANN is implemented majorly in this field for the better results. Main challenges identified from the review was data scarcity, customized AI models and presence of black box models and concluded that integration of edge and fog computing can be done to overcome these challenges [10].

Five machine learning algorithms, XGBoost, SVM, ANN, RF and LR, were used in a study performed by Li et al. to create models to forecast biogas production in an industrial-scale biogas plant handling food waste. As separate or combined input variables, three kinds of standard monitoring indicators (feed amount, feedstock qualities, and digester parameters) were used. The outcomes showed that when all of the indicators were present in the dataset, the random forest model performed the best, with an average R^2 of 0.74. Except for the RF model, which demonstrated the potential to forecast biogas output for the following day (R^2 = 0.73), the performance of the predictive models declined with lag time [11]. De Clercq et al. used models like elastic net, random forest, and extreme gradient boosting for predicting biomethane production in industrial-scale anaerobic co-digestion in time horizons for 1-day, 3-day, 5-day, 10-day, 20-day, 30-day, and 40-day. The result illustrated that random forest and extreme gradient boosting completely dominated the performance of elastic net with the value of R^2 ranging between 0.80 to 0.88 depending on time horizons. He found that food waste co-digested with percolate had strong positive interaction effects. XGBoost and random forest algorithms applied to industrial-scale anaerobic co-digestion data provide dependable prediction results and may be used as a useful complement for experimental and mechanistic/theoretical models of anaerobic digestion. However, these models have limitations and suggestions for deriving additional value from these methods are proposed [5]. Long et al. investigated the viability of utilizing six machine learning algorithms to forecast methane yield using genomic data and the corresponding operational factors from eight research groups. Accuracy of random forest classification models with values of 0.77 for operational parameters and 0.78 for genomic data at the bacterial phylum level was found. He also concluded that increasing the data amount and specific input features has the ability to significantly increase prediction accuracy [12].

Chong et al., for modeling the biogas generation and methane yield from the anaerobic digestion of palm oil mill effluent (POME) in a local-scale anaerobic covered lagoon, used ML methods like as response surface methodology, adaptive neuro-fuzzy inference system, and artificial neural network. Results showed that these models had a high coefficient of determination of up to 0.98 and were well fitted to two years of operational data. With an R^2 of 0.9791, the MAE of 0.0730, and the lowest root mean squared error, ANFIS has the highest prediction accuracy. Sensitivity study reveals that pH has the most significant impact on methane output and also concluded that in order to record more observations and quantitative findings for later analysis, more data sets with longer operational periods (> 3 years) are needed [13]. In an experiment conducted by Olatunji et al. to model the biogas and methane yield from anaerobic digestion of *Arachis hypogea* Shells with combined pretreatment techniques, fuzzy c-means (FCM)- clustered adaptive neuro-fuzzy inference system and optimized artificial neural network were used using significant operating parameters like temperature, retention time, pretreatment methods as an input variable. The FCM-ANFIS model with ten clusters performed better than the ANN model with R^2 value of 0.985 and concluded that biogas yield of pretreated *Arachis hypogea* can be predicted satisfactorily and is recommended for other similar studies [14].

Zhang et al. proposed a hybrid extreme learning machine to improve prediction accuracy by solving imbalanced data. Obtained results suggests that ELM model has a good prediction accuracy for real plant data with R^2 = 0.993. Feed volume and total volatile fatty acids of anaerobic digestion were the two important parameters that positively affected biogas production. The findings demonstrate that the challenges of machine learning in forecasting plant data imbalances are resolved by combining data balancing methods and optimization algorithms, enabling the precise prediction of plant biogas generation under various loads [15]. Pei et al. explored the methane yield and consequent microbial community in mixed high-solid anaerobic digestion (HS-AD) with spray-enhanced circumstances by machine learning and 16S rRNA gene sequencing to further examine the impact of the interaction between species and their compositional niches. Extreme learning machine, artificial learning network and random forest were used for analysing and modelling anaerobic digestion of dry fermentation. The best prediction model was ELM which predicted the material biogas production with a mean absolute error of 0.678 and coefficient of determination of 0.9574. He also suggested that ML algorithms can handle large datasets of text, images, strings, and the creation of internet databases will open new opportunities for microbial analysis. Future studies should expand the scope of the dataset to facilitate training and evaluation of ML models and learn more about the metabolic pathways of microorganisms [16].

It can be considered that there is still a significant literature gap in terms of artificial intelligence-based modeling studies for the estimation of biogas production from a real full-scale sludge digestion process in a biological treatment plant. It has been observed that most of the researchers develop their models using reactors that are lab- or pilot-scale. The current study focuses on using an artificial neural network, random forest, and k-nearest neighbor algorithm on data of wastewater collected from eight rural districts of Istanbul Metropolitan area with 4.5m diameter tunnel with collector lines and treated by a fully operational, entirely mixed anaerobic sludge digester system. More specifically, this study is aimed to predicts the biogas production rate using well-trained artificial intelligence models and a multiple regression model, and then evaluates the performance of the models' predictions using a variety of statistical performance indicators.

Hence, this thesis presents an overview of machine learning process applied in the field of anaerobic digestion. Some of the most widely used machine learning algorithms are discussed thoroughly and a demonstration of model for ANN, RF and KNN method is presented and analyzed accordingly.

1.2 Objectives

The broad objective of the thesis is to explore the feasibility of application of machine learning in biogas process. The specific aim of the study was (Appendix F provides information about project thesis descriptions):

1. Review of the literature on application of machine learning in the biogas process.
2. To identify the relevant requirements, benefits, and challenges.
3. To identify the most appropriate machine learning methods and tools for industrial applications.
4. And discussion on the demonstration of model for the further improvement of machine learning algorithm in biogas production.

1.3 Approach and Methodology

This work focuses on using machine learning algorithm for the prediction of biogas. Machine learning may be an effective way to overcome the limitations of the current modelling approaches and provide better process monitoring tools and might improve the sustainability of the plant's operations. To address the above research objectives, detailed discussion of several ML algorithms is done to get detail knowledge and its application in AD process. This process includes the use of scoping method for the literature review techniques. Various search strings were developed for searching the research papers on application of machine learning in biogas process in Scopus, Web of Science, Google Scholar, and several papers were sorted out based on the title, abstract and keywords. Limiting of research papers in terms of language, article and review paper followed by screening and trimming the duplicate lists, 30 articles were selected as they were most relevant for the thesis work. Various machine learning algorithms with their applications in several fields are discussed thoroughly and frequently used machine learning based on the experiments done by several researchers is used for the demonstration of the model for our work. Finally, the process and the results will be reviewed and evaluated, and some suggestions will be given for the future relevant work.

1.4 Target Group

This thesis work is especially interesting for the researchers in the field of machine learning and its application in biogas process. Anaerobic digestion method is being used globally to produce biogas and has also been thought as a suitable substitute for fossil fuels and energy generation. Several countries have been working effectively to produce biogas with the motivation of reducing greenhouse gases, recycling of food waste, producing renewable energy and fertilizers. This work is also interesting for all the customers who are working continuously to maximize biogas production and if suitable machine learning algorithms can be used, the business will rise with less environmental pollution.

1.5 Outline

The following chapter 2 will give you the detailed insight on theoretical background concerning this work with focus on the knowledge and discovery process, machine learning approaches including its requirements, benefits, and challenges. It will also discuss the importance of modelling in biogas processes and how to optimize the mathematical model including the several machine learning types and techniques used till date. Chapter 3 will discuss the literature review methodology and demonstration of the most relevant model using machine learning algorithm. Chapter 4 will discuss the results and discussion for the classification criteria of best machine learning algorithm and why we selected the model for our thesis work. Chapter 5 will discuss the conclusion as well as examine the validity and reliability of presented results followed by summarizing the work and giving overall outlook for possible future research and expansion on the topic.

2 Theoretical Background

This section gives an overview of relevant theoretical foundations about the anaerobic digestion process and how machine learning methods can be used successfully for the biogas processes. Several types of machine learning methods are discussed with their applicability in biogas production. The objectives of this thesis will also be answered in this section.

2.1 Anaerobic Digestion and Biogas Process

Anaerobic digestion is the process where decomposition of organic matters occurs by a wide range of microorganisms in absence of oxygen. Many naturally occurring anoxic habitats like watercourses, moist soils, sediments, etc exhibits such mechanisms. Additionally, it can also be used with a variety of feedstocks, including as agricultural plant residue, municipal, and food sector wastes, as well as industrial municipal waste streams [17]. Anaerobic digestion is one of the well-known techniques for organic waste treatment and has several advantages such as generation of bioenergy (i.e., methane-rich biogas), no need for aeration, low sludge yield, effective pathogen removal and organic fertilizer. Numerous initiatives have been taken for increasing the generation of biogas and enhancing the energy balance of AD processes in response to the rising need for sustainable energy sources [18].

Biogas is a clean and renewable form of energy source having full potential to replace conventional energy sources like fossil fuels and oil, which are the potential cause for harming the environment and is also declining more quickly [19].The main composition of biogas can be divided into two constituents: combustible and non-combustible components. Methane (about 55–70% of the volume) and carbon dioxide (30–40%) are most likely the main constituents of bio gas while other components like carbon monoxide, hydrogen, carbon dioxide, nitrogen and hydrogen sulphide, depends on the source materials and processing techniques used [20].

To create biogas through anaerobic digestion, the following procedures are normally taken:

- First, the organic waste material is collected and pre-treated for removing of any large particles or inorganic matters.
- An anaerobic digestor, a sealed tank which is maintained at a particular temperature and pH level to encourage the growth of microorganisms, receives the pre-treated organic material.
- The organic matter is broken down by the microorganisms and results in production of biogas, which is composed primarily of methane and carbon dioxide.
- Biogas thus, collected can be used in a generator to create power or heat as a source of sustainable energy and the remaining digestate can be used as a fertilizer for agriculture.

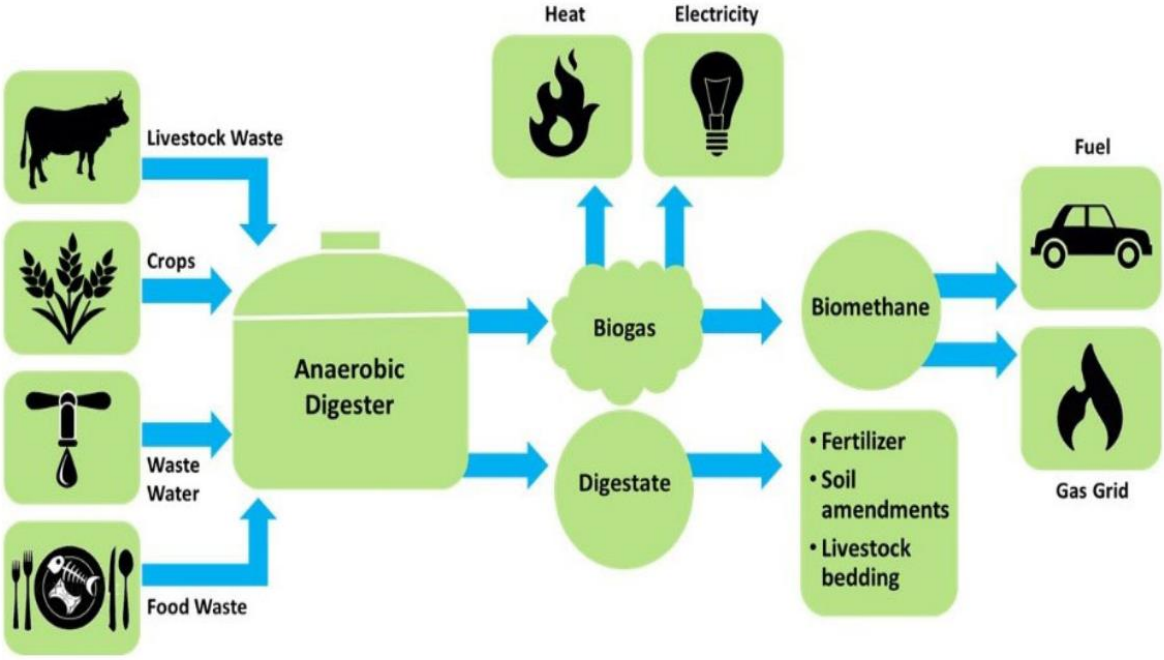


Figure 2.1: Biogas production from anaerobic digestion process [21].

Utilizing specific types of biomasses to partially meet energy needs is made highly appealing by biogas technology. A well working biogas system can offer consumers and the community a number of advantages, resulting in resource conservation and protecting the environment. Biogas, being the product of anaerobic degradation of organic substrates, it is also one of the oldest process used for the treatment of industrial waste and stabilisation of sludges.

2.1.1 Techniques for Enhancing Biogas Production

In order to enhance the biogas production, several methods have been used and can be classified into the following categories:

- By using additives.
- Recycling of slurry and slurry filtrate.
- Difference in operational parameters like temperature, hydraulic retention time (HRT) and particle size of the substrate.
- Used of fixed film/ biofilters [19].

2.1.2 Challenges in Biogas production from Anaerobic Digestion Process

Biogas production is relatively lengthy process since several microorganisms operate together and it depends on a number of variables including pH, temperature, C/N ratio, etc. Anaerobic digestion, being widely adopted for remediating diverse organic waste and producing digestate that is rich in nutrient and renewable energy, suffers from instability and adversely affects biogas production [21]. For the successful production of biogas, proper monitoring and control is required to improve efficiency and to keep the process stable. Additionally, biogas plants frequently operate in sub-optimal conditions to avoid the instability process and suffer from overload or inhibition resulting in changing of feedstock. As a result, it is crucial to use the right controller to keep biogas plants from failing [22]. Anaerobic process modelling might be an effective method for forecasting crucial process performance factors, such as methane generation.

2.1.3 Why Modelling is Important?

Biogas production from anaerobic digestion is probably the most versatile and efficient biofuel in terms of utilisation of feedstocks and energy application. Numerous mathematical models have been developed to monitor, optimize, and control the anaerobic digestion process. Despite the development of literature on the use of these models and reviews of them, no comprehensive classification criteria for these models have been proposed based on an assessment of their differences [23]. For the operation of efficient anaerobic digestion system, Important operational parameters like co-substrate ratio, their composition, volatile fatty acids/alkalinity ratio, organic loading rate, and solids/hydraulic retention time, etc are required. In addition to that, it is frequently challenging to achieve optimization, prediction and control, and early identification of system instability of anaerobic digestion process through laborious human monitoring techniques. So, the demand for the adoption of mathematical modelling is increasing to overcome the challenges like high complexity, involvement of nonlinear parameters, and high-dimensional conversion of the process [24].

2.1.4 Research on Mathematical Model

Till now, several models have been developed for reflecting various process occurring in the anaerobic digestion and these models are based on theoretical, analytical and statistical methods to explain the anaerobic process [2]. The mechanistic models is considered to predict the cumulative biogas yield and compositions considering the conservation of mass and energy [25]. Anaerobic Digestion Model No. 1 (ADM1) is one of the most widely used mathematical models for the AD process which consists of a variety of reaction kinetic equations. The ADM1 framework is particularly useful for process design and dynamic simulation. However, because of its fixed thermodynamics approach, some processes applicability would require significant structural modifications. And also due to incomplete understanding of AD microbiomes and physio-chemical process involved, the latent correlations between reactor performance and kinetic parameters have not been fully incorporated in the ADM1. Due to the tremendous difficulty of separating the highest specific absorption rates from the specific biomass concentration, it was not possible to identify all the factors and coefficients [26]. And due to the lack of a thorough knowledge of anaerobic digestion process, mechanistic AD models struggled to assess and predict digestive function which are frequently erroneous. Thus, there is a critical need for creative approaches to successfully forecast the results of digestion [2].

In the phase of development of various models for analysis, Machine learning has emerged to be an innovative tool for model creation and has the potential to be used to estimate and control the performance of anaerobic digesters [21]. Comparing ML to mechanistic models (such as ADM1), the drawbacks of ML-based AD process modelling are mitigated by:

(a) shorter execution time

(b) not requiring knowledge across various disciplines in bio-kinetics, microbiome, or heat/mass transfer

(c) avoiding the model re-calibration if trained utilizing large datasets [25].

2.1.5 Optimization of Mathematical Models

Mathematical model of the structures should be chosen according to the four basic principles:

- Simplicity, the model should be simple.
- Causality, the model should depict the most relevant cause-and-effect relationships.
- Identifiability, the value of unknown parameter from the data at hand should be determined.
- Predictive capability, the model should maintain its validity under potential future conditions.

For optimizing biogas process control and design strategies, improving laboratory and real world studies, the correct mathematical definition of the anaerobic digestion process must be developed [4]. Anaerobic digestion process is not completely understood regardless the fact that it has been existing for a longer period of time and is known widely and the complexity of microbiological, chemical and physical process involved is mostly to be blamed for this. Therefore, development of a reliable mathematical method to predict reactor performance based on historical information of some key factors might lead to improved control of an anaerobic digestion process [27].

2.2 Machine Learning

Machine learning is the process which enables computers to simulate human learning, identify and acquire knowledge from the real world and improve the performance of the task based on the knowledge available. Learning is the continuous process of gaining knowledge. Being capable of thinking, humans naturally learn from their experiences whereas computers use algorithms to learn rather than reasoning. There are many ML algorithms that have been developed till now. Popularity of machine learning is increasing nowadays because of its tremendous processor speed and memory size and the field has now a huge number of algorithms which uses mathematical or statistical analysis to learn, draw conclusions or infer data. The number of scientific publications that suggest modifications or fusions of ML algorithms shows that this number keeps rising [28]. Also, one advantage using this method is, it is entirely based on readily available online data or historical readings of the process. ML involves a cycle of:

- Training – giving the algorithm a training set of data to help it discover previously undiscovered patterns within the information.
- Validation - adjusting the classifier's hyper-parameters on a different data set improves the model's performance.
- Testing – To improve the ultimate accuracy of the model, a different sample of data is used [29].

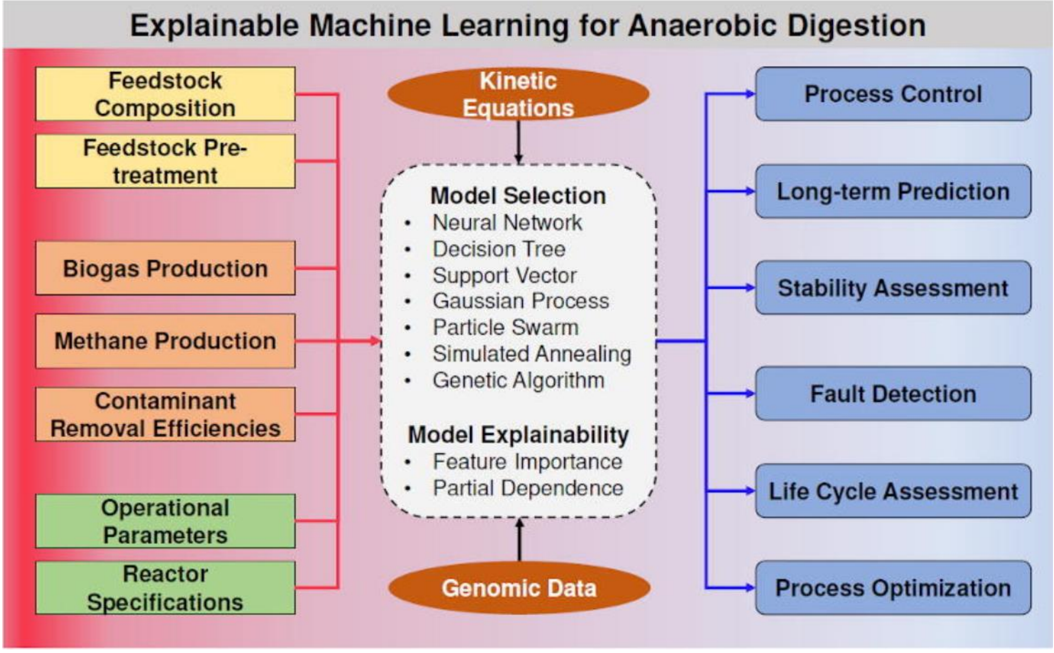


Figure 2.2: Machine Learning process for anaerobic digestion [26].

To explore the feasibility of application of machine learning in biogas process, the present study aims to address the following three tasks:

1. What is the application of machine learning in biogas process?
2. What are the relevant requirements, benefits, and challenges of using ML in the AD process?
3. How do we identify the most appropriate ML methods and tools for industrial application?

2.3 Types of Machine Learning Techniques

2.3.1 Artificial Neural Network

Artificial Neural Network abstracts the human brain's neuron network from the viewpoint of information processing, which is characterized by nonlinearity, non-limitation, great flexibility, and fault tolerance. Being the most popular ML model, ANN is capable of solving a wide range of challenging non-linear environmental issues, and they have a long history of success in the domains of municipal solid waste management, composting, anaerobic digestion, thermal treatment and disposal [9]. It consists of an input layer with input neurons (one for each input), hidden layers with hidden neurons (the number of which is user-specified), and an output layer with output neurons (one for each output) and the neurons between layers are connected via weights in such a way that only neurons in adjacent layers are connected [30]. Any ANN with more than one hidden layer is termed as deep neural network. Each hidden and output neuron has an associated activation function which reacts to linear combinations of present value from the layer of neurons before it [31]. The prediction accuracy and calculation speed of ANNs are often superior than those of other models when handling the huge amount of data and several ANN models have shown their excellency in various fields owing to their special characteristics [9]. The ability to model complex non-linear behaviour is the main advantage of ANN as compared to other machine learning approaches [32].

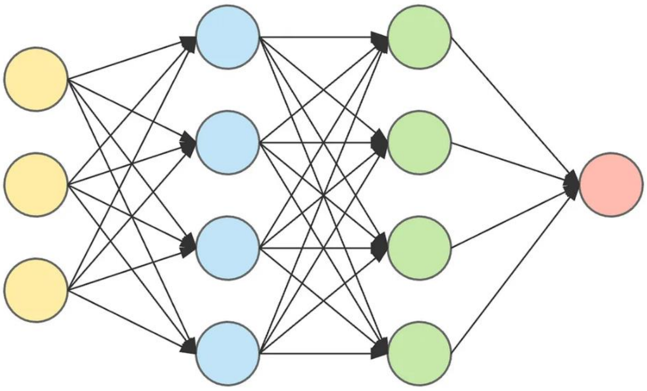


Figure 2.3: Diagram of Artificial Neural Network [34].

However, there are several drawbacks of ANN method that should not be ignored, with the "black box" aspect being the most important [33]. Incapability of explaining the reasoning process and providing reasoning basis is the main drawbacks of using ANN. Due to the fact that it can only simulate the process of change based on empirical data and does not advance our understanding of the fundamental causes of change, this limitation is truly disadvantageous to its application in natural science research, particularly in studies on the outlined mechanisms. When using ANN models in their research, researchers should pay close attention and consider these constraints [9].

2.3.2 Random Forest

It is a technique which combines the result predicted from several algorithms to obtain a better result. Random forest is an ensemble method which make use of bootstrap aggregation for generating decision trees and thus obtained final output is an aggregation of the prediction based on decision tree. By using this approach, it is possible to appropriately consider all available attributes and avoid having excessively coupled trees [34].

Random forests are a group of tree predictors where each tree depends on the values of a random vector sampled independently and with the same distribution for all the trees in the forest. The generalization error depends on the strength of the individual trees in the forest and their co-relation between them. A more modern approach that addresses these issues is the use of random forest models, which offer an appealing addition to nonlinear approximations of statistical correlations between variables [35]. The key benefit of the RF technique is its ability to operate well with larger and more dimensional data and its ability to rank models through an internal variable important measure [36].

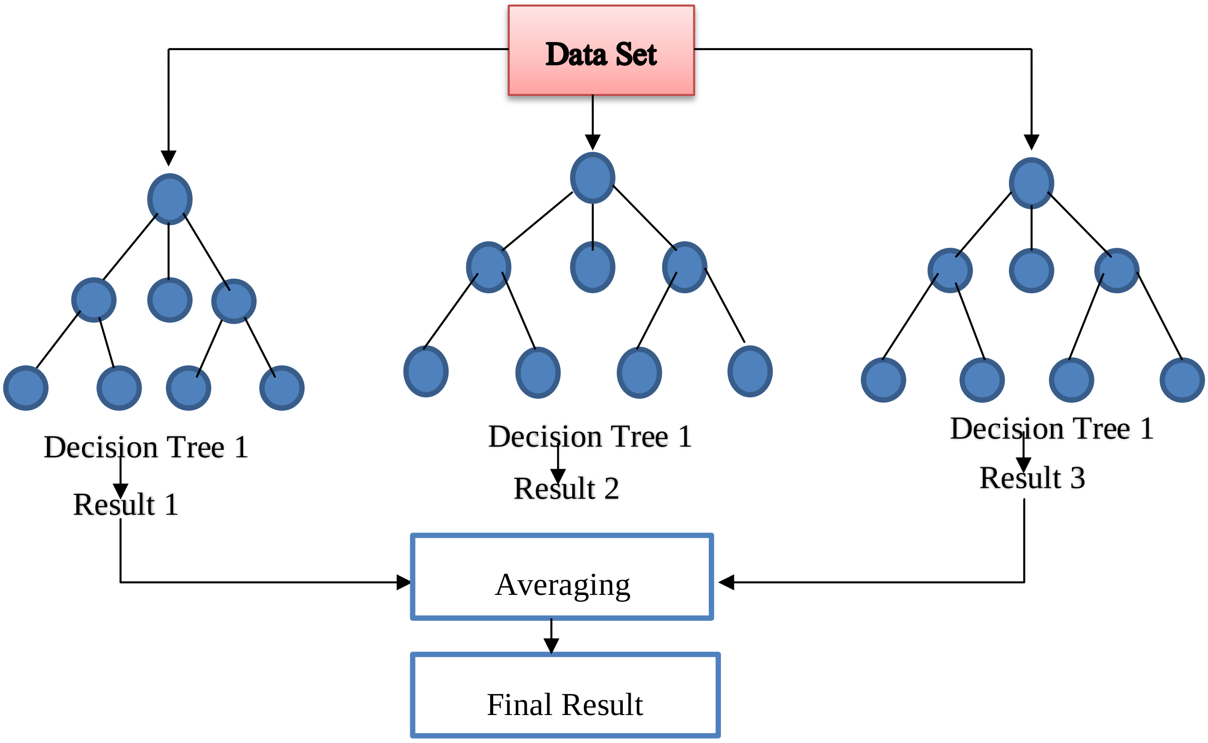


Figure 2.4: Diagram of Random Forest Method [37].

2.3.3 Support Vector Machine

It is an advanced supervised machine learning technique that Cortes and Vapnik introduced in 1995 and is now regarded as one of the most effective techniques for tackling regression and classification problems. The main idea behind the SVM algorithm is to map input data by using a non-linear mapping function (kernel) into a multi-dimensional feature space; as a consequence, the data are linearly separable. The linear regression is then applied in the multi-dimensional feature and the input data are mapped [38]. Frequently used kernel functions are the exponential radial basis, the polynomial kernel, and the multi-layer perceptron kernel functions [39].

2.3.4 Extreme Learning Machine

It can be viewed as a unique feedforward neural network model with one or more hidden layers. In contrast to ANNs, the model's initial parameters are randomly assigned and not iteratively updated. The output layer and hidden layer weights are learned all at once from the training data [40]. As a result, the ELM model's training takes substantially less time than a standard ANN since it only needs to maximize the quantity of hidden layer neurons. An ELM is just a generalized version of regularized linear regression (for the output layer coefficients) and have been show to outperform SVM and SVR in some cases [41].

2.3.5 K-Nearest Neighbor Regression

It is a simple method that conserves all of the existing cases and categorizes new data or cases using a similarity measure. This algorithm is frequently used to categorize a data point based on the categorization of its neighbors. KNN algorithm's "k" parameter is dependent on feature similarity. To improve accuracy, a technique known as parameter tuning involves determining the appropriate value of K [42]. For a given sample data point, it finds the k nearest data points based on a certain distance metric method. The forecast is done by averaging the k closest neighbours, or more generally by creating a weighted average with, for example, the Manhattan distance or the inverse square of the distance. The method is similar to kernel regression with a break for include points in the linear and smoother prediction when the weights are set by a kernel as well. The value of k is usually determined using a leave-one out cross-validation [31].

2.3.6 Adaptive Network-Based Fuzzy Inference System

ANFIS combines the features of neural networks using its self-learning, adaptiveness, parallel processing, and generalization capabilities with the advantages of fuzzy logic systems for extracting useful information from uncertainty. The ANFIS parameters have distinct physical meanings. Additionally, a hybrid approach that combines gradient descent and the least-squares estimate is applied during ANFIS training. The network's convergence rate is increased by the hybrid method, which also successfully prevents the network from getting trapped in local minima [43].

The network structure is made up of nodes and directional links, whose outputs depend on node-specific parameters, and the learning rules specify how to adjust these parameters to minimize a specified error measure [44]. It consists of two parts called as premise and consequent parts and training of the model determines the parameters which belongs to these parts utilizing an optimization algorithm [45]. These systems consist of five layers (fuzzification, rules, normalization, consequent, and addition), which performs a mathematical process [21].

3 Research Methodology

There are several ways of doing literature reviews. At first, searching for the keywords like methods and techniques of literature review, types of literature review and literature review methods was done on the google website. From thirty search results (10 webpages for each search), it was found that fourteen literature methods were commonly used which are listed in the table below:

3.1 Literature Review Procedure

Table 3.1: Literature Review Type and Methodology.

No.	Literature review methods	References
1	Narrative or traditional literature review	[46],[47],[48],[49],[50],[51]
2	Systematic literature review	[46],[48],[49],[50],[51],[52]
3	Critical review	[46],[47],[52]
4	Theoretical framework review	[46]
5	Descriptive or mapping reviews	[47]
6	Scoping reviews	[35],[36],[38],[39],[40]
8	Realist reviews	[35]
9	Argumentative literature review	[36]
10	Integrative literature review	[36]
11	Theoretical literature review	[36]
12	Meta-analysis literature review	[37],[40]
13	Meta-synthesis review	[37]
14	Rapid review	[38],[40]

After analysing various literature review methodologies like traditional review, systematic review, scoping review, critical review, etc., we came up with the conclusion of choosing scoping review method as a part of thesis literature review. This approach was used because it enables the careful analysis of the literature on a subject by distinguishing important components, theories, and evidence sources that influence field practice [53]. There are several primary steps in the scoping review protocol, including selecting studies, choosing research questions, charting data, summarizing findings, and reporting them [54]. The following measures were done in accordance with the scoping review protocol:

1. At first, three research questions were defined.
2. Numerous trial-and-error searches were performed using scientific database like Scopus, Web of science, to begin the search and its search strings are listed below:
   Scopus: TITLE-ABS-KEY (((anaerobic AND digestion) AND ((machine AND learning) OR (artificial AND intelligence)) AND biogas)).
   Web of Science: ALL= (((anaerobic digestion) and (machine learning) or (artificial

intelligence) and (biogas))).

Initially title, abstract and keywords were searched with no limit throughout all databases. As a result, 141,112 studies in all categories were listed in Scopus and web of science database. Then the strings were modified to get the accurate listings and 33 papers were selected for the review form Scopus and Web of Science.

3. Then the title strings were limited to time from 2019 to 2023 and the articles remained to 33 for both Scopus and web of Science.
4. By limiting the research paper in terms of article, review paper and language, 29 from Scopus and 33 from web of science was finalised.
5. The lists contained many duplicates and after trimming the lists and removing the duplicate articles using Microsoft Excel, only 52 remained.
6. By screening the titles and full text of studies, only 30 articles were found to relevant for the thesis topic work.
7. Finally, these 30 articles were accessed carefully for finding out the most suitable and appropriate method for machine learning application in biogas production.

The screening process was used to categorize the collected articles according to characteristics such as year, country, types of research paper, method, and tools. The figure below illustrates the results of the screening process.

Table 3.2: Flow diagram showing the screening process.

	Process	Description	Scopus	Web of Science
Identification	Searching	Initial search results	n = 141	n = 112
Screening	Screening 1	Limiting to publication year (2019-2023)	n = 33	n = 33
	Screening 2	Limiting to article and review paper	n = 29	n = 33
	Screening 3	Refining articles in English language	n = 29	n = 33
Eligibility	Eliminating duplicates	Eliminating common articles in databases	n = 52	
	Title screening	Screening based on the titles only	n = 30	
	Full-text assessing	Screening based on article's full/text	n = 30	

3.2 Model Building and Evaluation

The dataset obtained is from an experiment performed in the year 2008 which is shown in Appendix A. It was divided into categories namely, influent flow rate of the feed sludge, total solids content, total volatile solids content, alkalinity, volatile fatty acids, and total biogas production. With a ratio of 50%:50%, the dataset was split into training and test sets, with model validation performed on the training sets. The effectiveness of prediction models was then assessed using the coefficient of determination (R^2) and mean squared error (MSE). Influent flow rate is represented by TMF in the CSV data in appendix table A.

3.3 Demonstration of Model Used

A data sample used for the prediction of total produced biogas flow (Appendix A) was taken from GitHub website and was used for our analysis purpose for several machine learning algorithms. Originally, the experiment was done for SVM method. The python code that was found on the website presents support vector regression (SVR) model used on a dataset which was used for prediction of total biogas flow.

Initially, it was planned to use the python code for just ANN and RF method, but it was concluded to assign the code for various machine learning algorithms and analyze the result obtained as per the discussion with supervisors. Python was used for the coding of various ML algorithms. The code obtained from the website was modified as per the algorithms. Initially, it had only one hidden layer. So, the coding was done for two hidden layers and increased the number of iterations to obtain the desired convergence of the biogas production. It was found that determination coefficient (R^2) was increased with the increase in hidden layer number. The prediction of biogas was much better with the increase in hidden layer than that with single layer network.

3.4 Measuring Accuracy of the Estimation

After training of the model, it is important to measure the accuracy of prediction. Two metrices were used to evaluate the model accuracy namely determination coefficient (R^2) and mean squared error (MSE). R^2 is defined as the amount of variance in the predicted variable that can be related to the model input parameters. Higher value of R^2 indicates that the model includes significant input parameters and is trained well for predicting the experimental value in the dataset [8]. The coefficient of determination takes any values between 0 to 1 [55]. Mean squared error (MSE) assesses the average squared difference between the observed and predicted values and measures the amount of error in statistical models. Model having no error means MSE value is equal to zero and as the model error increases, so does its value [56]. The selection criteria for the network are maximum value of R^2 and minimum value of MSE for both testing and validation phase ensuring the model to fit and predict accurately.

4 Results and Discussion

This section discusses the descriptive information associated with the latest studies and trends about the application of machine learning in biogas prediction.

4.1 Classification Based on Publication Year

In this section, the reader is informed about the status of the study through a year-by-year analysis and emphasizes the researcher's interest in this area. Several research studies have been done in this field, but analysis of the results starting from 2019 has been done which is shown in the bar chart below. In 2019, there were two publications, and we can see the increase in trend at alarming rate. Because of the covid pandemic in 2020, the publications could not reach up to the trend line, but we can see rise in year 2022 with twelve number of publications. And there are already seven numbers of publication just in year 2023 till January and the trend will keep rising in upcoming days. Initially, Because of non-linearity behaviour or microbiomes and complex mathematical models and incomplete understanding, the interest seems to be less in this field but as the interest in Machine learning is increasing nowadays, the application of machine learning algorithm to produce biogas might increase in coming days.

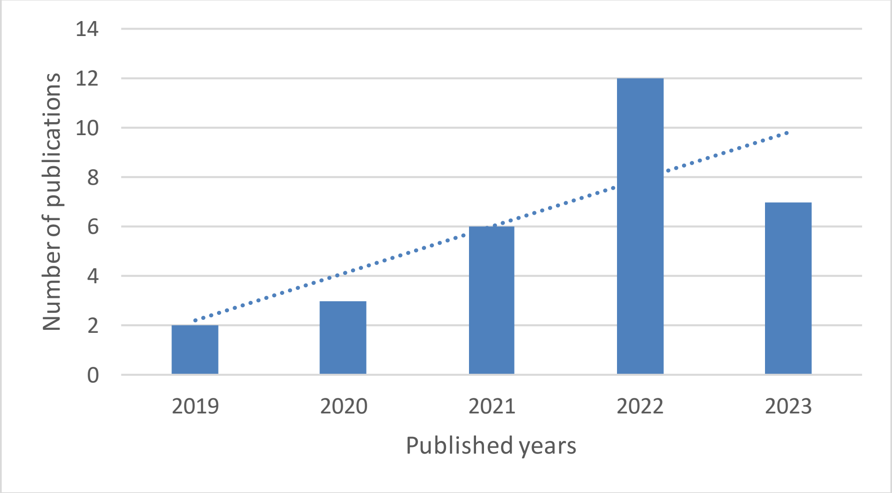


Figure 4.1: number of selected studies and overall trend from 2019 to January 2023.

4.2 Classification Based on Publication Country

From the analysis of selected publications for thesis work, thirteen countries contributed to this topic (2019-2023). From the bar graph below, we can see that the highest contribution is from China with fifteen publications followed by India and USA with three and two numbers. Similarly, as per the selected papers, other countries like Egypt, Belgium, Germany, etc. contributed with only one publication.

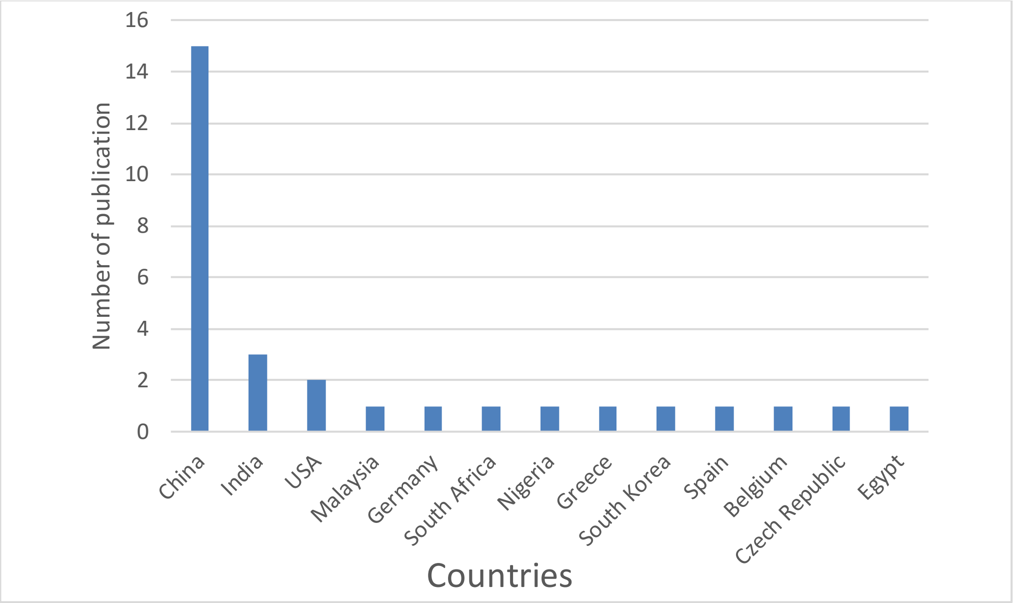


Figure 4.2: Contributions from different countries from 2019 to January 2023.

4.3 Classification Based on Document Type

The document based on only articles and review paper was considered for the discussion. We can see that articles have the highest rank with twenty-six numbers of publications and the number of studies for review paper was only four from the year 2019 to 2023. Other categories like conference papers, book chapters, etc. were not addressed in thirty selected papers.

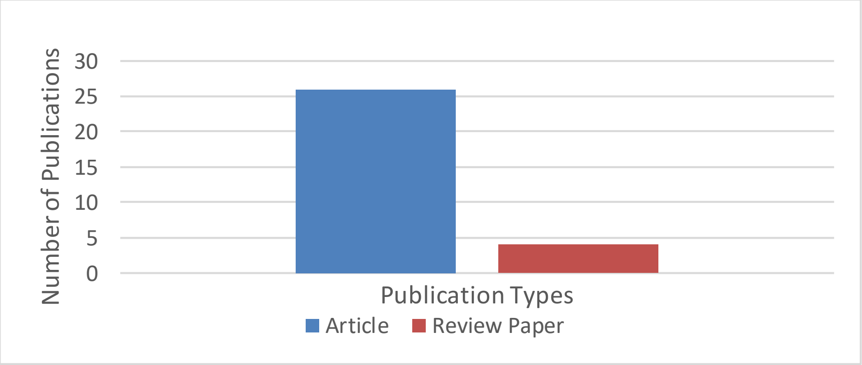


Figure 4.3: Categorization based on the document type.

4.4 Classification Based on Methodology Used for ML

From the bar chart, we can see that artificial neural network is the most recommended methodology and was selected by four publications. Random forest holds the second position as recommended by three studies. ANFIS, ELM, SVR, GBR and SVM method was followed by two publications and other methods like NSFM, DA-LSTM, XGBoost and KNN had one recommendation among the selected papers.

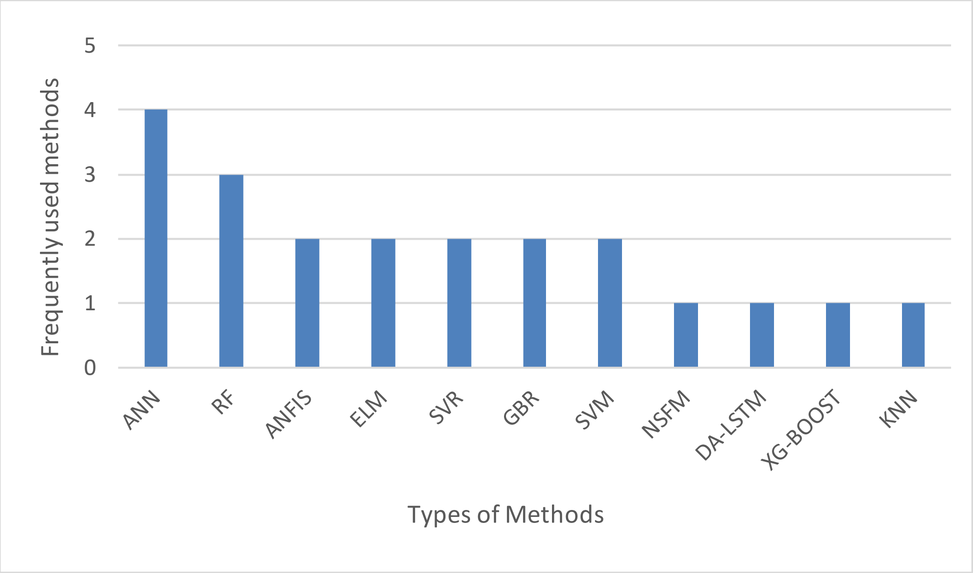


Figure 4.4: Types of methodology used for ML.

4.5 Selection of Machine Learning Algorithm based on the Methodology Used

After the several literature reviews done for the application of machine learning in biogas process, we found out the methodologies that have been used significantly for the effective prediction of biogas in various industrial scales. Machine learning have been proven as a powerful tool for dealing with non-linearity and highly complex relationship between variables and has been successful to solve various environmental problems with effective production of biogas [57]. The prediction could be entirely based on the readily available data since ML technique is more independent of solving interactions involved in the anaerobic digestion process. Many techniques, including ANN, RF, ANFIS, SVM, KNN etc., have been employed successfully for diverse purposes in AD applications, but no single algorithm has a clear advantage in terms of having the ability to generalize other scientific objectives. The varying performances may be due to the variable dataset format as well as the biological system diverse design [58]. And for every individual experiment whether large industrial or small co-pilot scale, the machine learning algorithm for different projects works differently and it is difficult to decide the supremacy of one algorithm compared to others.

From table 4.1, we can clearly see the use of various ML algorithms for various scientific tasks and accuracy of the system is calculated accordingly with the major parameters. Higher value of determination coefficient (R^2) indicates that the model has considerable input parameters and has been trained properly to forecast the experimental values in the dataset [8] and it would be difficult to determine the best suitability of a particular algorithm for the prediction of biogas. So, several machine learnings algorithms were used for the data sample obtained from the Github software and results were analysed for the biogas prediction.

Table 4.1: Accuracy obtained from various ML algorithms.

Biomass	**Scientific Tasks**	**Major parameters**	**Algorithms**	**Best accuracy**	**References**
Agriculture and forest waste, Algae	Prediction of gaseous products	Highest Temperature, Heating rate, Particle size, gas flow rate	RF, SVM	RF, R^2:*0.85-0.87*	[59]
Lignocellulosic biomasses	Prediction of biochar yield and carbon contents	Pyrolysis conditions, Particle size, Cellulose	RF	RF, R^2: 0.8548	[60]
Waste peanut shells	Optimization of methyl levulinate yield	Ratio of waste peanut shell methanol, Reaction time and temperature	ANN-GA	ANN, R^2: 0.89	[61]
Fermentation Biomasses	Prediction of remaining residuals	Heating rate, Temperature, weight loss at various heating rates	SVM	SVM, R^2 : 0.9999	[62]
wheat straw	Co-digestion design	C/N ratio, temperature, retention time	ANN, ANFIS, LR	ANFIS R^2: 0.9996	[63]
Mixed streams	Identification of key feedstock composition for biogas prediction	C/N ratio, cellulose, lignin, temperature	RF, GLMNET SVM, KNN	GLMNET R^2: 0.73	[2]
Mixed streams	Biogas Prediction for Industrial -scale Digestor	Loading rates, waste types	RF, XGBoost	XGBoost R^2: 0.88	[5]

4.6 Performance of ML Models

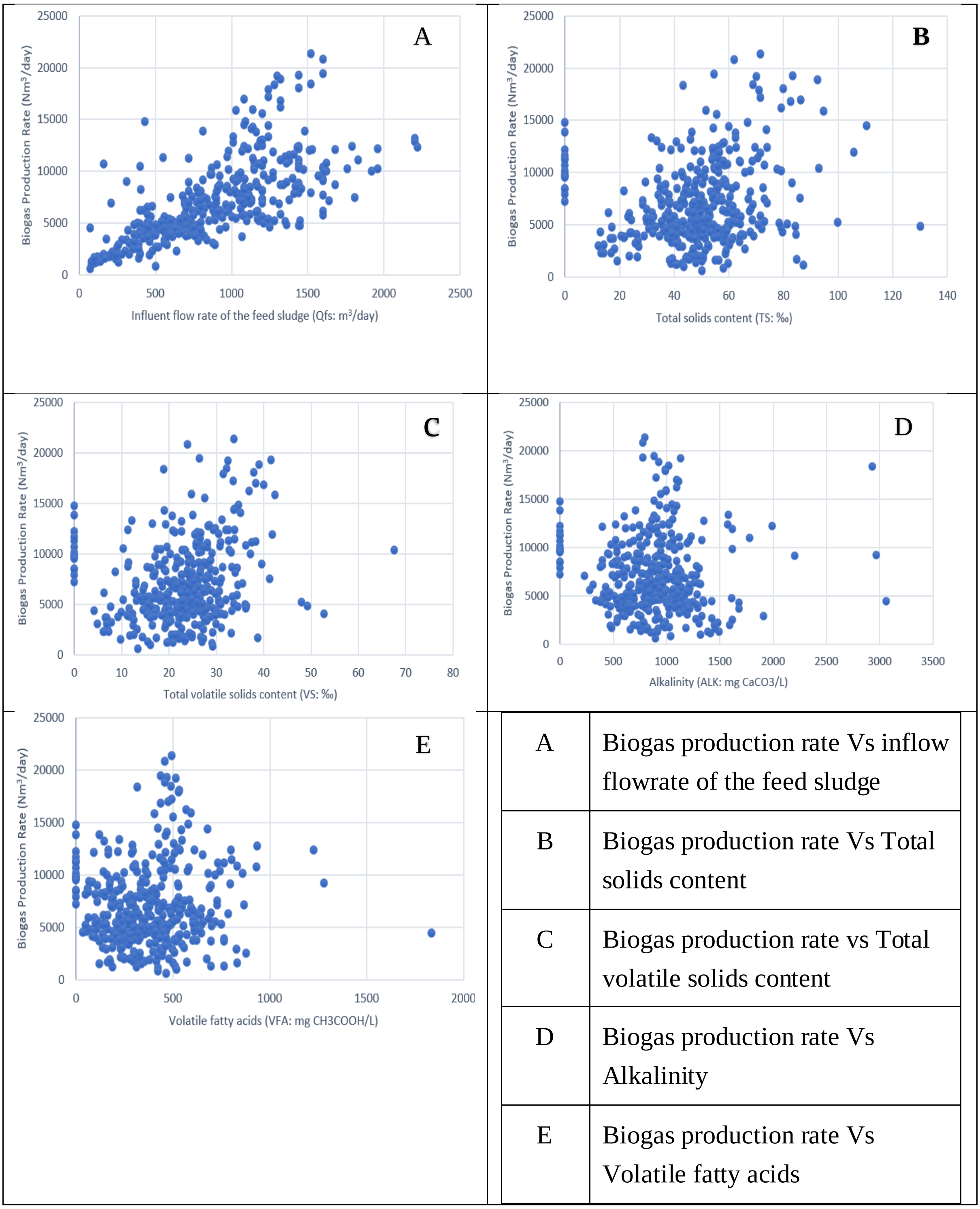
In this study, three different machine learning algorithms- artificial neural network, random forest and k-nearest neighbors are employed to generate predictive models. These are all frequently employed ML algorithms that have been found to accurately predict biogas.

4.6.1 Input Variables.

A total of 394 data were taken from the sample available on the website for the total prediction of biogas flow which is given in Appendix A. These values were divided into two parts i.e., samples and scores in the python code for analysis purpose. Samples contain the data of influent flow rate of the feed sludge, total solids content, total volatile solids content, alkalinity and volatile fatty acids. Scores contain the total biogas production rate, and these data were provided to the python code in CSV format. For the evaluation of models, libraries like pandas, numpy, matplotlib, sklearn are used.

Scatter diagram of the biogas production rate as a function of each estimator is shown in Figure 4.5. The naming of A, B, C, D and E in the figure shows the plot of various components with respect to biogas production and how well the individual features contribute to accurate prediction of biogas. Linear regression analysis shows that the plot of influent flow rate of the feed sludge vs biogas production rate provides more accurate prediction as compared to other parameters. Total solid content, total volatile solids content, alkalinity and volatile fatty acids show somewhat same graphical features and tends to be less accurate in terms of the prediction accuracy if compared individually in terms of biogas prediction.

4.6.2 Scatter Plots of Biogas Production Rate as a Function of Predictors



A	Biogas production rate Vs inflow flowrate of the feed sludge
B	Biogas production rate Vs Total solids content
C	Biogas production rate vs Total volatile solids content
D	Biogas production rate Vs Alkalinity
E	Biogas production rate Vs Volatile fatty acids

Figure 4.5: scatter plot of biogas production range as a function of predictors.

Table 4.2 presents the statistical data of the overall samples displaying the units of the individual components, maximum and minimum value of the overall data, its mean and standard deviations.

Table 4.2: statistics of the model components used in analysis (n= 394 for each variable).

Component	Unit	Max	Min	Mean	Standard deviation
Influent flow rate of the feed sludge	$(\mathrm{m^3/day})$	2220	70	871	434
Total solids content	(%)	130.08	12.27	49	19
Total volatile solids content	(%)	67.51	4.17	23	9
Alkalinity	$(\mathrm{mg/LCaCO_3})$	3061.11	226	896	385
Volatile fatty acids	$(\mathrm{mg/LCH_3COOH})$	1835.34	35.77	381	219
Biogas production rate	$(\mathrm{Nm^3/day})$	21393	608	6983	3954

4.6.3 Performance of ML Models to Predict Biogas Yield.

In this section we will discuss about the performance of ML model for the given data samples in Appendix A using several algorithms:

4.6.3.1 ANN Method

Appendix C shows the code for the artificial neural network model to predict biogas values based on some input features. Two hidden layers are considered for the simulation because it can easily affect network's ability to learn the complex data patterns. Khashaba et al., used two hidden layers for modelling of biochar enhanced sludge digestion and R^2 value was found to be 0.9922. Additionally, machine learning architectures with two hidden layers have demonstrated outstanding performance in resolving difficult issues like image and speech recognition, simple processing [8]. The appropriate number of hidden layers is always dependent on the particular problem and is frequently established through trial and testing. There is no set number of hidden layers that is generally advised for ANN. It consists of 60 nodes each in two hidden layers in the algorithm and these nodes are responsible for processing the input data and performing computations to produce output values. Rectified linear activation function (ReLU) is used for training the network. When the function receives any negative input then it returns to zero and for positive case, it returns that value with output ranging from 0 to infinity. It is also the most used activation function in neural networks and in most of the case utilized as default activation function [64]. The import statements are used at the beginning of the code to import the required libraries. The target variables are extracted from the data frame and saved in separate arrays referred to as samples and scores. The data is loaded from a CSV file using the pandas library's read_csv() method. Splitting of data between training and testing sets was done using the train_split() function of Scikit-learn, and the model was then trained using MLP Regressor. The maximum number of iterations was set to 1000. Following training of the MLP Regressor model, predictions are made on the testing set, the determination coefficient (R^2) and mean squared error values were obtained, and finally a plot of the actual vs. predicted biogas is generated.

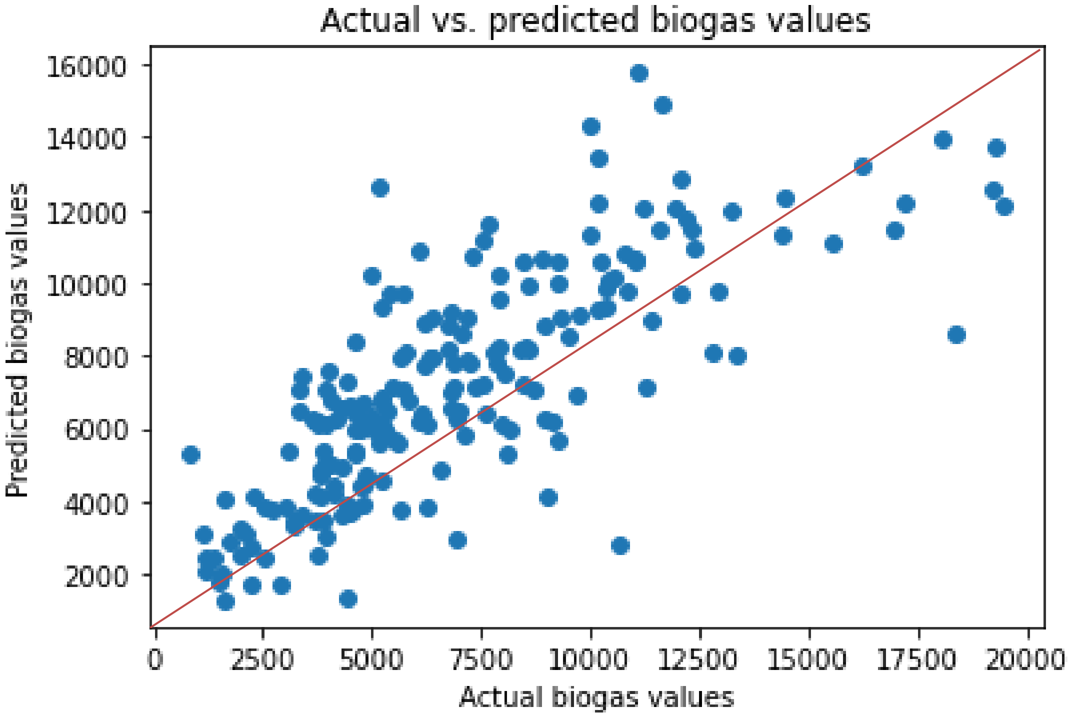


Figure 4.6: Actual vs predicted biogas values for ANN method.

Results showed that the model's mean square error was 6032492.224 and its determination coefficient value was 0.593. This graph displays a scatter plot of the test dataset's actual biogas values against the expected values. In the graph above, if we draw a 45-degree line (y=x), we can see that the values are scattered and most of the values lie above the diagonal line and few spots are below the line. For the accurate prediction, most of the data should align to the line which is not seen in our case.

4.6.3.2 Random Forest Method

Appendix D shows the code for the random forest model to predict biogas values. This program implements a random forest regressor model in Python. Necessary libraries like pandas, numpy, matplotlib, RandomForestRegressor are imported from sklearn.ensemble and various evaluation metrics and data splitting functions form sklearn. A while loop is used for recursive feature elimination and the loop continues until the convergence condition is met by the code. The model is configured with 1000 decision trees and a maximum depth of each tree is taken as 20 and random state is set to 50 for reproducibility. Predictions are made on testing set using trained model. After the model has been trained on a randomly divided training set, the mean squared error and determination coefficient were used to evaluate the model. Each feature in the model has its importance evaluated, and the component with the lowest importance is removed. Until there is just one feature remaining or the model has converged, this process was repeated. The mean squared error and determination coefficient score are included in lists for plotting after each iteration. The model's performance over iterations is then illustrated graphically using the mean squared error and R^2 score, along with a scatter plot of the actual versus projected biogas values.

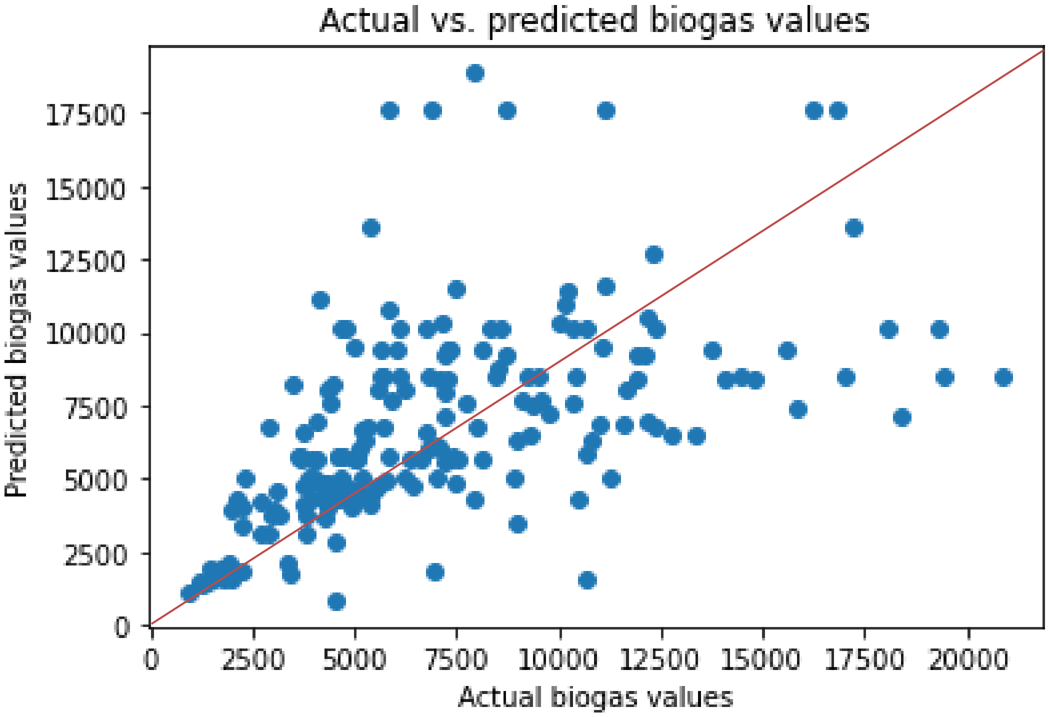


Figure 4.7: Actual vs predicted biogas values for Random Forest method.

Here, the diagonal line touches most of the values in the graph. We can see scatterings of the value that’s almost evenly distributed above the diagonal line. The initial data represents a strong positive relationship between actual and predicted values. The result also showed that the R^2 score for this method is 0.620 and mean squared error of 6695312.177 making it the most accurate model as compared to ANN and K-NN method.

4.6.3.3 KNN Method

Appendix E is the code of KNN method for predicting the biogas. The code implements a k-nearest neighbor regression model using the KNeighborsRegressor class from the scikit-learn library. The feature and target variables are extracted from the data frame and stored in the ‘samples’ and ‘scores’ variables and the number of features is determined on the basis of shape of ‘samples’ array. A while loop is used for recursive feature elimination, and it continues till the convergence is met. The data is split into training and testing set and test size is set to 50 percent of data and predictions are made in testing set using the trained model. The feature with the lowest absolute weight is found and removed from the dataset and the process is repeated until the model has converged, or all least important features have been removed. Feature removal is typically done in order to improve the machine learning model's performance by removing unnecessary characteristics that could cause over fitting and poor generalization in newly collected data.

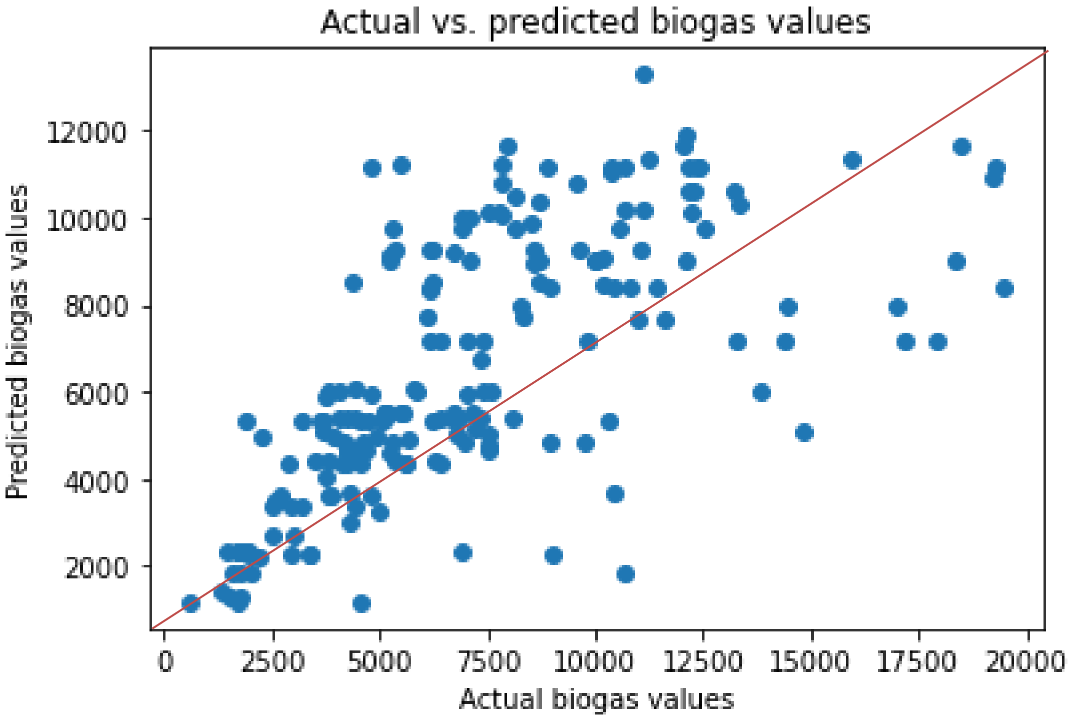


Figure 4.8: Actual vs predicted biogas values for KNN method.

Here we can see that most of the scattered points are above the diagonal lines, and few are below which might be due to noise and variability in the data or because of limitation in the model. After recursive elimination, this method recognized date as a most important feature for the prediction of biogas.

Table 4.3: summary of the results obtained.

ML Algorithms	R^2 values	MSE values
SVM method	A	8615491.909
ANN method	0.593	6032492.224
RF method	0.620	6695312.177
KNN method	0.515	6854866.264

A- Originally, the algorithm was made for SVM method and biogas prediction accuracy was checked only using MSE value (shown in Appendix A) and R^2 was not calculated while obtaining data and code from Github. Trying to run the python code which was obtained from the website resulted in problems and the code was not able to run. In that code, the loop was run four times and each time dropping the lowest feature. Initially the lowest feature weight was found to be for total volatile solid content percent, and it was dropped. Similarly, it was followed by alkalinity being the lowest feature for the second time, total solid content percent was the third one and fatty acid being the fourth dropping features by making the least impact on the model's performance. Finally, the influent flow rate of the feed sludge became the most important feature or predictor variable making the greatest influence on the output target variable.

The R^2 values in table 4.3 shows that the random forest approach outperforms the ANN and KNN methods. However, if we look at the MSE values, it becomes clear that the ANN technique outperforms other algorithms. The MSE value provides an average of squared difference between the predicted values and actual values in the dataset and lower MSE values results in model predictions that are closer to the actual value. The R^2 value shows how well the model fits the data and it ranges from 0 to 1, with higher values indicating a better fit. As per the research conducted by Chicco et al, the result given by coefficient of determination gives more accurate information than any other methods like Mean squared error, Root mean squared error, Mean absolute error [65].

To sum up, Taking R^2 as a major dominating factor for a linear regression, we can conclude that random forest method gives the better prediction as compared to ANN, SVM and KNN regression models.

5 Conclusion

Anaerobic digestion has played an important role in recycling organic waste and producing renewable source of energy and has been suggested as a sensible strategy for increasing biogas production, utilizing existing infrastructures and overcoming the challenges associated with it. Machine learning methods have been considered as a key to success for the anaerobic digestion process. The anaerobic digestion process and its performance as a result of different feed substrates or operating circumstances are better understood by researchers when they use machine learning models to mathematically predict and capture the behavior of complex systems.

The present study demonstrates a model for the application of machine learning in biogas processes. The challenges, benefits, and requirements of machine learning algorithms for biogas prediction are discussed in detail. From the literature review, scoping review was found out to be the most suitable method and review was done accordingly. Web of Science, Google Scholar and Scopus website were used for searching the relevant research work and thirty papers were finalized for the further review procedure.

Three different types of machine learning algorithms like artificial neural network, random forest and k-nearest neighbors were considered for the demonstration of the model, python code was developed accordingly, and model accuracy was checked using R^2 and MSE method. Random forest model gave a high prediction accuracy with R^2 value of 0.62 compared to other models and MSE value of 6032492.244 was given by ANN method. R^2 is the dominant factor for model accuracy as per literature review and was chosen to be the best parameter. R^2 of 62% for random forest is a bit low value and might be due to less number of data available for the simulation. Optimization with the help of backpropagation algorithm can be done to further improve the model accuracy as per the review done. Literature reviews suggested that the interest in this field is accelerating every year and researchers are committed to producing renewable sources of energy and biogas being a prominent solution.

6 Further Studies

Despite the advancement in ML modelling of AD process, the area is still in its early stage of development and the problems might be due to insufficient data, lack of consistent principles for model selections, etc. The use of online monitoring systems to continuously update predictions based on real-time data may enable more accurate and timely adjustment to the biogas production process. Also, since the nature of biogas production is dynamic, further studies can be improved for the accuracy and robustness of these models and should also investigate in using the machine learning algorithms to find the optimal setting that maximizes biogas yield and efficiency of the process by minimising resource consumption and environmental impacts. Machine learning techniques can be used effectively to improve process efficiency, sustainability, and overall performance as the field is constantly growing.

References

[1] ‘lucr 1_Dobre Paul_Main factors affecting biogas production_revistaRBL_2014 _1_.pdf’. Accessed: Apr. 23, 2023. [Online]. Available: https://rombio.eu/vol19nr3/lucr%201_Dobre%20Paul_Main%20factors%20affecting%20biogas%20production_revistaRBL_2014%20_1_.pdf

[2] L. Wang, F. Long, W. Liao, and H. Liu, ‘Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms’, *Bioresource Technology*, vol. 298, p. 122495, Feb. 2020, doi: 10.1016/j.biortech.2019.122495.

[3] L. Alejo, J. Atkinson, V. Guzmán-Fierro, and M. Roeckel, ‘Effluent composition prediction of a two-stage anaerobic digestion process: machine learning and stoichiometry techniques’, *Environ Sci Pollut Res*, vol. 25, no. 21, pp. 21149–21163, Jul. 2018, doi: 10.1007/s11356-018-2224-7.

[4] F. Tufaner and Y. Demirci, ‘Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models’, *Clean Techn Environ Policy*, vol. 22, no. 3, pp. 713–724, Apr. 2020, doi: 10.1007/s10098-020-01816-z.

[5] D. De Clercq, Z. Wen, F. Fei, L. Caicedo, K. Yuan, and R. Shang, ‘Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion’, *Science of The Total Environment*, vol. 712, p. 134574, Apr. 2020, doi: 10.1016/j.scitotenv.2019.134574.

[6] S. Zareei and J. Khodaei, ‘Modeling and optimization of biogas production from cow manure and maize straw using an adaptive neuro-fuzzy inference system’, *Renewable Energy*, vol. 114, pp. 423–427, Dec. 2017, doi: 10.1016/j.renene.2017.07.050.

[7] J. W. Chen, Y. J. Chan, S. K. Arumugasamy, and S. K. Yazdi, ‘Process modelling and optimisation of methane yield from palm oil mill effluent using response surface methodology and artificial neural network’, *Journal of Water Process Engineering*, vol. 52, p. 103493, Apr. 2023, doi: 10.1016/j.jwpe.2023.103493.

[8] N. H. Khashaba, R. S. Ettouney, M. M. Abdelaal, F. H. Ashour, and M. A. El-Rifai, ‘Artificial neural network modeling of biochar enhanced anaerobic sewage sludge digestion’, *Journal of Environmental Chemical Engineering*, vol. 10, no. 4, p. 107988, Aug. 2022, doi: 10.1016/j.jece.2022.107988.

[9] H. Guo, S. Wu, Y. Tian, J. Zhang, and H. Liu, ‘Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: A review’, *Bioresource Technology*, vol. 319, p. 124114, Jan. 2021, doi: 10.1016/j.biortech.2020.124114.

[10] L. M. Joshi, R. K. Bharti, and R. Singh, ‘Internet of things and machine learning-based approaches in the urban solid waste management: Trends, challenges, and future directions’, *Expert Systems*, vol. 39, no. 5, p. e12865, 2022, doi: 10.1111/exsy.12865.

[11] C. Li, P. He, W. Peng, F. Lü, R. Du, and H. Zhang, ‘Exploring available input variables for machine learning models to predict biogas production in industrial-scale biogas

plants treating food waste', *Journal of Cleaner Production*, vol. 380, p. 135074, Dec. 2022, doi: 10.1016/j.jclepro.2022.135074.

[12] F. Long, L. Wang, W. Cai, K. Lesnik, and H. Liu, 'Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data', *Water Research*, vol. 199, p. 117182, Jul. 2021, doi: 10.1016/j.watres.2021.117182.

[13] D. J. S. Chong, Y. J. Chan, S. K. Arumugasamy, S. K. Yazdi, and J. W. Lim, 'Optimisation and performance evaluation of response surface methodology (RSM), artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) in the prediction of biogas production from palm oil mill effluent (POME)', *Energy*, vol. 266, p. 126449, Mar. 2023, doi: 10.1016/j.energy.2022.126449.

[14] K. O. Olatunji, D. M. Madyira, N. A. Ahmed, O. Adeleke, and O. Ogunkunle, 'Modeling the Biogas and Methane Yield from Anaerobic Digestion of Arachis hypogea Shells with Combined Pretreatment Techniques Using Machine Learning Approaches', *Waste Biomass Valor*, Sep. 2022, doi: 10.1007/s12649-022-01935-2.

[15] Y. Zhang *et al.*, 'Plant-scale biogas production prediction based on multiple hybrid machine learning technique', *Bioresource Technology*, vol. 363, p. 127899, Nov. 2022, doi: 10.1016/j.biortech.2022.127899.

[16] Z. Pei *et al.*, 'Understanding of the interrelationship between methane production and microorganisms in high-solid anaerobic co-digestion using microbial analysis and machine learning', *Journal of Cleaner Production*, vol. 373, p. 133848, Nov. 2022, doi: 10.1016/j.jclepro.2022.133848.

[17] A. J. Ward, P. J. Hobbs, P. J. Holliman, and D. L. Jones, 'Optimisation of the anaerobic digestion of agricultural resources', *Bioresource Technology*, vol. 99, no. 17, pp. 7928–7940, Nov. 2008, doi: 10.1016/j.biortech.2008.02.044.

[18] G. Choi, H. Kim, and C. Lee, 'Long-term monitoring of a thermal hydrolysis-anaerobic co-digestion plant treating high-strength organic wastes: Process performance and microbial community dynamics', *Bioresource Technology*, vol. 319, p. 124138, Jan. 2021, doi: 10.1016/j.biortech.2020.124138.

[19] Yadvika, Santosh, T. R. Sreekrishnan, S. Kohli, and V. Rana, 'Enhancement of biogas production from solid substrates using different techniques—a review', *Bioresource Technology*, vol. 95, no. 1, pp. 1–10, Oct. 2004, doi: 10.1016/j.biortech.2004.02.010.

[20] Y. Qian, S. Sun, D. Ju, X. Shan, and X. Lu, 'Review of the state-of-the-art of biogas combustion mechanisms and applications in internal combustion engines', *Renewable and Sustainable Energy Reviews*, vol. 69, pp. 50–58, Mar. 2017, doi: 10.1016/j.rser.2016.11.059.

[21] I. Andrade Cruz *et al.*, 'Application of machine learning in anaerobic digestion: Perspectives and challenges', *Bioresource Technology*, vol. 345, p. 126433, Feb. 2022, doi: 10.1016/j.biortech.2021.126433.

[22] P. Ghofrani-Isfahani, B. Valverde-Pérez, M. Alvarado-Morales, M. Shahrokhi, M. Vossoughi, and I. Angelidaki, 'Supervisory control of an anaerobic digester subject to drastic substrate changes', *Chemical Engineering Journal*, vol. 391, p. 123502, Jul. 2020, doi: 10.1016/j.cej.2019.123502.

[23] S. Emebu, J. Pecha, and D. Janáčová, ‘Review on anaerobic digestion models: Model classification & elaboration of process phenomena’, *Renewable and Sustainable Energy Reviews*, vol. 160, p. 112288, May 2022, doi: 10.1016/j.rser.2022.112288.

[24] M. Khan, W. Chuenchart, K. C. Surendra, and S. Kumar Khanal, ‘Applications of artificial intelligence in anaerobic co-digestion: Recent advances and prospects’, *Bioresource Technology*, vol. 370, p. 128501, Feb. 2023, doi: 10.1016/j.biortech.2022.128501.

[25] R. Gupta *et al.*, ‘Review of explainable machine learning for anaerobic digestion’, *Bioresource Technology*, vol. 369, p. 128468, Feb. 2023, doi: 10.1016/j.biortech.2022.128468.

[26] A. Donoso-Bravo, J. Mailier, C. Martin, J. Rodríguez, C. A. Aceves-Lara, and A. V. Wouwer, ‘Model selection, identification and validation in anaerobic digestion: A review’, *Water Research*, vol. 45, no. 17, pp. 5347–5364, Nov. 2011, doi: 10.1016/j.watres.2011.08.059.

[27] F. Tufaner, Y. Avşar, and M. T. Gönüllü, ‘Modeling of biogas production from cattle manure with co-digestion of different organic wastes using an artificial neural network’, *Clean Techn Environ Policy*, vol. 19, no. 9, pp. 2255–2264, Nov. 2017, doi: 10.1007/s10098-017-1413-2.

[28] I. Portugal, P. Alencar, and D. Cowan, ‘The use of machine learning algorithms in recommender systems: A systematic review’, *Expert Systems with Applications*, vol. 97, pp. 205–227, May 2018, doi: 10.1016/j.eswa.2017.12.020.

[29] G. S. Fanourgakis, K. Gkagkas, E. Tylianakis, and G. Froudakis, ‘A Generic Machine Learning Algorithm for the Prediction of Gas Adsorption in Nanoporous Materials’, *J. Phys. Chem. C*, vol. 124, no. 13, pp. 7117–7126, Apr. 2020, doi: 10.1021/acs.jpcc.9b10766.

[30] ‘Accurate prediction of chemical exergy of technical lignins for exergy-based assessment on sustainable utilization processes | Elsevier Enhanced Reader’. https://reader.elsevier.com/reader/sd/pii/S0360544221032904?token=B2F5FF20AFDBFECF72548FA53DDEA1F0C692559EEAB441F8828BA941FF01C4C73AC400217DEED4D7BF62A6D5755D7A07&originRegion=eu-west-1&originCreation=20230209160436 (accessed Feb. 09, 2023).

[31] Z. Wang *et al.*, ‘Comparison of machine learning methods for predicting the methane production from anaerobic digestion of lignocellulosic biomass’, *Energy*, vol. 263, p. 125883, Jan. 2023, doi: 10.1016/j.energy.2022.125883.

[32] S. Dreiseitl and L. Ohno-Machado, ‘Logistic regression and artificial neural network classification models: a methodology review’, *Journal of Biomedical Informatics*, vol. 35, no. 5, pp. 352–359, Oct. 2002, doi: 10.1016/S1532-0464(03)00034-0.

[33] M. Kannangara, R. Dua, L. Ahmadi, and F. Bensebaa, ‘Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches’, *Waste Management*, vol. 74, pp. 3–15, Apr. 2018, doi: 10.1016/j.wasman.2017.11.057.

[34] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar, and N. Chilamkurti, ‘Random forest for big data classification in the internet of things using

optimal features', *Int. J. Mach. Learn. & Cyber.*, vol. 10, no. 10, pp. 2609–2618, Oct. 2019, doi: 10.1007/s13042-018-00916-z.

[35] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[36] S. S. Matin and S. C. Chelgani, 'Estimation of coal gross calorific value based on various analyses by random forest method', *Fuel*, vol. 177, pp. 274–278, Aug. 2016, doi: 10.1016/j.fuel.2016.03.031.

[37] 'What is a Random Forest?', *TIBCO Software*. https://www.tibco.com/reference-center/what-is-a-random-forest (accessed May 12, 2023).

[38] C. Cortes and V. Vapnik, 'Support-vector networks', *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[39] L. Liu and Y. Lei, 'An accurate ecological footprint analysis and prediction for Beijing based on SVM model', *Ecological Informatics*, vol. 44, pp. 33–42, Mar. 2018, doi: 10.1016/j.ecoinf.2018.01.003.

[40] S. Yin and H. Liu, 'Wind power prediction based on outlier correction, ensemble reinforcement learning, and residual correction', *Energy*, vol. 250, p. 123857, Jul. 2022, doi: 10.1016/j.energy.2022.123857.

[41] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, 'Extreme Learning Machine for Regression and Multiclass Classification', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, Apr. 2012, doi: 10.1109/TSMCB.2011.2168604.

[42] D. Subramanian, 'A Simple Introduction to K-Nearest Neighbors Algorithm', *Medium*, Jul. 12, 2021. https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e (accessed May 10, 2023).

[43] Y. Zhang, T. Chai, Y. Fu, and H. Niu, 'Nonlinear adaptive control method based on ANFIS and multiple models', in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, Dec. 2009, pp. 1387–1392. doi: 10.1109/CDC.2009.5399518.

[44] J.-S. R. Jang, 'ANFIS: adaptive-network-based fuzzy inference system', *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665–685, May 1993, doi: 10.1109/21.256541.

[45] D. Karaboga and E. Kaya, 'Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey', *Artif Intell Rev*, vol. 52, no. 4, pp. 2263–2293, Dec. 2019, doi: 10.1007/s10462-017-9610-2.

[46] K. A. Chetty Priya, 'Different types of literature review techniques followed in a research', *Knowledge Tank*, Dec. 15, 2021. https://www.projectguru.in/different-types-of-literature-review-techniques-followed-in-a-thesis-research/ (accessed Apr. 08, 2023).

[47] G. Paré and S. Kitsiou, *Chapter 9 Methods for Literature Reviews*. University of Victoria, 2017. Accessed: Feb. 22, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK481583/

[48] ‘Types of Literature Review’, *Research-Methodology*. https://research-methodology.net/research-methodology/types-literature-review/ (accessed Apr. 15, 2023).

[49] ‘Chapter 3-d0a27b4fc2eac406e4509dad1dedecfc.pdf’. Accessed: Apr. 15, 2023. [Online]. Available: https://www.goodfellowpublishers.com/free_files/Chapter%203-d0a27b4fc2eac406e4509dad1dedecfc.pdf

[50] T. Bowman, ‘Library Guides: Literature Review: Types of literature reviews’. https://libguides.csu.edu.au/review/Types (accessed Apr. 15, 2023).

[51] M. Tucker, ‘Library guides: Systematic style literature reviews for education and social sciences: Different types of literature review’. https://libraryguides.griffith.edu.au/c.php?g=451351&p=3333115 (accessed Apr. 15, 2023).

[52] U. Library, ‘Research Guides: Systematic Reviews: Types of Literature Reviews’. https://guides.library.ucla.edu/c.php?g=224129&p=1485355 (accessed Apr. 15, 2023).

[53] C. Cooke, ‘LibGuides: Understanding Review Types: Scoping Reviews’. https://libguides.lib.umanitoba.ca/reviewtypes/scoping (accessed Jan. 27, 2023).

[54] Z. Barahmand and G. Samarakoon, ‘Sensitivity Analysis and Anaerobic Digestion Modeling: A Scoping Review’, *Fermentation*, vol. 8, no. 11, Art. no. 11, Nov. 2022, doi: 10.3390/fermentation8110624.

[55] ‘Coefficient of Determination’, *Corporate Finance Institute*. https://corporatefinanceinstitute.com/resources/data-science/coefficient-of-determination/ (accessed Apr. 23, 2023).

[56] J. Frost, ‘Mean Squared Error (MSE)’, *Statistics By Jim*, Nov. 12, 2021. https://statisticsbyjim.com/regression/mean-squared-error-mse/ (accessed Apr. 23, 2023).

[57] ‘Machine learning in natural and engineered water systems | Elsevier Enhanced Reader’. https://reader.elsevier.com/reader/sd/pii/S0043135421008617?token=C5BA86A60E83DC11DF7FE82C5095FE2D768343BF7640CE31C975372565FF9BF4CB0D00BB181D4140E12ABB678B8257F2&originRegion=eu-west-1&originCreation=20230223130258 (accessed Feb. 23, 2023).

[58] ‘Machine learning and circular bioeconomy: Building new resource efficiency from diverse waste streams | Elsevier Enhanced Reader’. https://reader.elsevier.com/reader/sd/pii/S0960852422017783?token=3B19D4A35DB35D9971C1F6A17CF94CD35C455A866A108B975175D8A21B32997BEA8DF25BD3931090C32C37D005B43190&originRegion=eu-west-1&originCreation=20230223142755 (accessed Feb. 23, 2023).

[59] Q. Tang *et al.*, ‘Machine learning prediction of pyrolytic gas yield and compositions with feature reduction methods: Effects of pyrolysis conditions and biomass characteristics’, *Bioresource Technology*, vol. 339, p. 125581, Nov. 2021, doi: 10.1016/j.biortech.2021.125581.

[60] X. Zhu, Y. Li, and X. Wang, ‘Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions’,

Bioresource Technology, vol. 288, p. 121527, Sep. 2019, doi: 10.1016/j.biortech.2019.121527.

[61] ‘Fuel properties of hydrochar and pyrochar_ Prediction and exploration with machine learning | Elsevier Enhanced Reader’. https://reader.elsevier.com/reader/sd/pii/S0306261920306784?token=3CBC1E94B4FCC1660D503225A1452192A30A5848F6BB457192EB1B5B50AB0A975401853AD7BCBD18343D02DE6299DCFB&originRegion=eu-west-1&originCreation=20230223180129 (accessed Feb. 23, 2023).

[62] H. Shahbeig and M. Nosrati, ‘Pyrolysis of biological wastes for bioenergy production: Thermo-kinetic studies with machine-learning method and Py-GC/MS analysis’, *Fuel*, vol. 269, p. 117238, Jun. 2020, doi: 10.1016/j.fuel.2020.117238.

[63] B. Najafi and S. Faizollahzadeh Ardabili, ‘Application of ANFIS, ANN, and logistic methods in estimating biogas production from spent mushroom compost (SMC)’, *Resources, Conservation and Recycling*, vol. 133, pp. 169–178, Jun. 2018, doi: 10.1016/j.resconrec.2018.02.025.

[64] ‘What is Rectified Linear Unit (ReLU)’, *Deepchecks*. https://deepchecks.com/glossary/rectified-linear-unit-relu/ (accessed May 12, 2023).

[65] D. Chicco, M. J. Warrens, and G. Jurman, ‘The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation’, *PeerJ Comput. Sci.*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[66] ‘Prediction_Of_Total_Produced_Biogas_Flow/Recursive_SVM_Feature_selective_for_Biogas_Flow_Prediction.ipynb at main · kenanmorani/Prediction_Of_Total_Produced_Biogas_Flow’, *GitHub*. https://github.com/kenanmorani/Prediction_Of_Total_Produced_Biogas_Flow (accessed Apr. 12, 2023).

Appendices

Appendix A : csv file data for the total prediction of Biogas flow [66].

No.	Date	TMF	TSC Percent	TVS Percent	Alkalinity	Fatty Acid	biogas
1	1/1/2008	240	46.26	33.15	787	244.36	2113
2	1/2/2008	120	44.77	22.02	983	413.54	1754
3	1/3/2008	200	44.45	22.61	740	378.6	1884
4	1/4/2008	160	44.31	26.64	761	353.66	1768
5	1/5/2008	200	50.58	23.23	748	343.96	1715
6	1/6/2008	140	47.96	23.35	759	311.76	1387
7	1/7/2008	240	50.12	24.83	790	309.85	1757
8	1/8/2008	200	49.93	24.3	734	331.69	1863
9	1/9/2008	120	50.18	23.5	808	337.63	1516
10	1/10/2008	160	55.09	22.08	918	501.6	1562
11	1/11/2008	140	42.91	20.03	914	512.35	1309
12	1/12/2008	110	40.41	19.6	850	311.41	1183
13	1/18/2008	70	50	13.46	892	463.8	608
14	4/1/2008	80	43.58	16.01	1317	516.42	962
15	4/3/2008	100	87.22	29.07	1408	421.51	1140
16	4/4/2008	80	59.63	27.02	1386	764.35	1271
17	4/6/2008	100	38.79	15.43	1491	697.46	1248
18	4/8/2008	170	45.44	23.08	818	163.65	1696
19	4/11/2008	220	58.74	29.15	1481	325.07	2217
20	4/12/2008	250	65.85	31.08	1423	297.12	2649
21	4/13/2008	160	46.99	21.73	1593	478.41	1998
22	4/14/2008	170	56.29	27.31	1454	831.08	1599
23	4/15/2008	330	50.69	24.87	1450	676.35	2009
24	4/16/2008	340	43.1	16.41	1910	495.66	2902
25	4/17/2008	370	58.28	28.3	1195.56	695.23	3686
26	4/18/2008	310	57.94	27.64	1189	547.95	3368
27	4/19/2008	329	60.18	29.92	1193	558.21	3039
28	4/21/2008	400	48.06	24.3	1308	230.66	2090
29	4/22/2008	420	37.87	17.31	1343	449.15	4488
30	4/24/2008	520	57.44	23.8	3061.11	1835.34	4429
31	4/25/2008	550	57.14	26.9	1044	646.19	4402
32	4/26/2008	550	49.41	24.43	1041	609.6	4558
33	4/27/2008	570	46.63	22.7	1117	610.18	4828
34	4/28/2008	590	48.44	23.53	931	492.92	5367
35	4/29/2008	610	39.75	19.69	1104	645.12	5217
36	4/30/2008	610	39.01	19.49	1025	529.16	4242
37	5/1/2008	690	46.26	23.33	1100	718.63	5478
38	5/2/2008	690	43.62	21.86	815.56	415.1	5094
39	5/3/2008	740	38.25	19.49	810	427.11	4995

40	5/8/2008	640	54.66	25.62	847	356.25	4371
41	5/9/2008	640	52.53	24.93	883.33	469.3	4895
42	5/10/2008	650	46.71	22.72	822	465.49	4613
43	5/11/2008	660	53.35	26.67	819	373.73	4480
44	5/12/2008	660	26.21	12.25	765	245.72	4049
45	5/13/2008	660	47.51	23.35	1047.78	552.56	5143
46	5/14/2008	690	53.02	26.01	1143.33	514.57	5812
47	5/16/2008	690	59.98	30.73	1281.25	867.99	7112
48	5/17/2008	690	50.84	25.05	1237.5	729.5	7132
49	5/18/2008	690	35.42	14.84	1171.25	546.42	6839
50	5/19/2008	740	54.06	25.85	891	521.65	7546
51	5/22/2008	770	50	22.02	996	504.68	7006
52	5/23/2008	800	52.51	24.48	802	372.15	6392
53	5/24/2008	850	49.36	23.02	954	486.25	5243
54	5/25/2008	850	41.69	18.52	1013	479.99	6735
55	5/26/2008	900	67.7	30.68	1301	785.34	6300
56	5/27/2008	960	61.5	28.06	988	644.39	6752
57	5/28/2008	980	58.97	27.88	1113	628.81	6254
58	5/29/2008	1030	24.26	10.37	1051	619.49	5464
59	5/30/2008	1070	39.93	15.73	1218	603.95	5769
60	6/1/2008	1180	31.77	14.42	887	408.95	5247
61	6/2/2008	1170	38.43	18.1	880	446.15	5380
62	6/3/2008	720	43.33	20.81	857	268.95	5294
63	6/5/2008	560	32.55	15.98	790	162.75	4823
64	6/6/2008	440	46.32	22.09	1219	514.58	3698
65	6/7/2008	480	46.34	23.26	1293	601.96	3724
66	6/8/2008	440	21.57	9.13	1118	620.62	3757
67	6/9/2008	540	55.54	26.83	965	354.23	4660
68	6/10/2008	280	48.44	20.38	1160	463.05	3353
69	6/11/2008	700	36.29	16.96	873	233.81	4519
70	6/13/2008	720	59.7	26.39	1060	448.2	5217
71	6/14/2008	720	59.55	27.21	1047	430.61	4583
72	6/15/2008	720	48.27	21.5	1035	412.96	4207
73	6/16/2008	720	13.09	4.17	654	215.78	4334
74	6/17/2008	720	45.37	19.73	1133	469.97	4814
75	6/18/2008	720	65.42	30.61	1283	529.03	4019
76	6/19/2008	800	17.77	6.58	591	119.84	3686
77	6/21/2008	880	14.24	4.92	1025	187.67	3035
78	6/22/2008	640	16.84	6.14	985	444.94	2277
79	6/23/2008	360	46.95	13.29	1618	876.14	2510
80	6/24/2008	800	58.41	25.16	1191.11	464.93	3941
81	6/26/2008	880	59.84	27.2	1319	534.69	6224
82	6/27/2008	960	37.66	16.38	694.79	82.69	4642
83	6/28/2008	1040	38	16.61	1307	527.55	7838
84	6/30/2008	1120	38.44	18.03	905	223.96	5875

85	7/1/2008	1200	30.05	13.09	809	140.56	5324
86	7/2/2008	1280	30.62	15.22	801	138.43	5232
87	7/3/2008	1360	130.08	49.29	583	104.58	4823
88	7/4/2008	1440	17.18	7.59	576	78.74	4778
89	7/5/2008	1450	33.09	15.6	985	634.07	4773
90	7/6/2008	1600	32.52	14.91	813	339.95	5739
91	7/7/2008	1070	16.61	6.85	560	253.09	3694
92	7/8/2008	1600	15.82	6.26	663	141.68	6143
93	7/9/2008	1380	37.47	18.15	815	197.98	7205
94	7/10/2008	1600	48.11	22.54	2968.75	1277.91	9231
95	7/11/2008	1600	34.7	17.51	653.26	357.33	10411
96	7/12/2008	1600	55.94	26.67	586	314.53	10262
97	7/13/2008	1600	29.52	11.41	637	335.03	9122
98	7/14/2008	1440	36.9	17.73	822	242.21	8869
99	7/15/2008	1440	34.77	16.17	782.29	225.06	8579
100	7/16/2008	1120	41.05	18.96	733.33	207.1	8439
101	7/23/2008	970	72.5	34.46	1094	333.15	8576
102	7/24/2008	1200	55.82	26.15	992	418.24	7328
103	7/25/2008	1200	54.22	24.52	922.22	355.99	8099
104	7/26/2008	1200	53.23	24.82	890	334.63	8153
105	7/27/2008	1200	28.35	12.93	603	226.8	7277
106	7/28/2008	1200	28.35	12.93	603	226.8	6890
107	7/30/2008	1170	31.94	12.76	744.44	314.41	6158
108	7/31/2008	1200	36.62	17.19	814	237.45	5615
109	8/2/2008	1060	57.84	28.89	1263.33	580.15	6400
110	8/3/2008	1080	50.36	24.65	1130	541.08	6830
111	8/4/2008	1080	49.28	25.68	710.53	183.54	7641
112	8/5/2008	890	39.25	19.65	758	293.43	5731
113	8/9/2008	1080	48.15	24.7	861	258.19	5663
114	8/10/2008	1080	22.98	13.99	854	261.25	5732
115	8/11/2008	400	50.81	26.21	1109.78	323.5	4914
116	8/12/2008	400	53.23	26.54	1211.11	536.66	4294
117	8/13/2008	260	52.63	26.52	1329.17	826.18	2915
118	8/16/2008	540	51.32	24.8	1286	699.58	5144
119	8/17/2008	470	31.55	13.93	1061	353.57	5639
120	8/19/2008	500	57.88	29.16	1035	420.84	841
121	8/20/2008	370	56.02	27.29	1185	508.79	3821
122	8/21/2008	560	43.92	20.59	1160	294.94	4118
123	8/22/2008	560	60.56	30.73	1167	336.53	4977
124	8/24/2008	560	68.68	31.99	1208.7	488.72	5434
125	8/25/2008	560	52.46	26.57	1112	502.78	5601
126	8/28/2008	380	54.34	36.04	1101	372.12	4983
127	8/29/2008	600	41.05	24.34	1169	425.72	5194
128	8/30/2008	600	79.74	16.99	1157	405.45	4260

129	8/31/2008	600	49.31	23.13	1252.08	321.99	4360
130	9/4/2008	590	54.55	28.83	1077	321.06	3764
131	9/5/2008	450	30.73	11.52	819	189.86	4576
132	9/6/2008	450	42.42	23.32	994.9	238.04	4576
133	9/7/2008	600	56.12	29.42	930	217.17	4616
134	9/8/2008	70	49.57	27.55	913	312.78	4560
135	9/18/2008	240	19.11	9.7	655	118.53	1480
136	9/19/2008	260	40.94	20.75	882	188.6	1190
137	9/20/2008	280	47.2	20.57	970	283	1951
138	9/21/2008	240	39.22	14.36	1022	295.56	1720
139	9/22/2008	390	38.42	17.34	483	176.41	1629
140	9/23/2008	460	26.57	11.75	472	174.46	1876
141	9/29/2008	560	64	28	630	474.69	3352
142	9/30/2008	610	51.43	23.74	612	207.86	3643
143	10/2/2008	710	39.26	16.08	637	332.41	4800
144	10/3/2008	770	50.92	22.95	769.79	329.81	3995
145	10/4/2008	430	58.59	24.96	602	226.86	4175
146	10/5/2008	820	37.63	12.04	668	220.17	3781
147	10/6/2008	890	26.92	12.44	610	231.63	2902
148	10/7/2008	840	41.32	20.14	657.61	271.41	3362
149	10/8/2008	720	32.18	15.34	648	218.78	4194
150	10/9/2008	700	40.98	21.65	589	246.05	4726
151	10/10/2008	780	40.5	19.51	570	226.21	4807
152	10/11/2008	800	40.88	19.15	541	243.76	4067
153	10/12/2008	800	24.37	11.7	654	260.44	4228
154	10/13/2008	780	39.03	19.47	651	268.5	3323
155	10/14/2008	460	25.86	7.77	914	231.29	3209
156	10/15/2008	760	38.93	16.2	804	235.44	3769
157	10/16/2008	800	38.8	19.11	531	268.01	4413
158	10/17/2008	800	35.33	15.55	947	302.8	5089
159	10/18/2008	1000	48.04	23.25	920	309.72	6729
160	10/19/2008	700	29.83	14.45	745	260.08	6434
161	10/20/2008	800	48.92	23.3	957.78	511.41	6797
162	10/21/2008	800	40.86	20.72	986.73	479.78	6817
163	10/23/2008	700	23.69	9.91	1001	244.32	4213
164	10/24/2008	890	35.31	19.15	657.14	154.31	5320
165	10/25/2008	770	63	30.24	1211	641.16	6124
166	10/26/2008	890	54.19	26.26	1014	427.39	6779
167	10/27/2008	900	52.31	23.71	889.8	448.23	6164
168	10/28/2008	510	43	20.32	1346	204.68	4276
169	11/2/2008	520	18.25	6.97	542	396.05	2688
170	11/4/2008	650	40.42	18.83	706	301.89	3962
171	11/13/2008	360	49.14	25.47	582	432.99	2953
172	11/15/2008	500	45.52	16.9	658	467.71	3844

173	11/16/2008	540	36.04	14.54	1423	318.14	4410
174	11/20/2008	360	29	15.38	825	257.78	3940
175	11/23/2008	350	51.26	22.66	1273	506	2559
176	12/1/2008	460	46.37	23.32	535	299.29	4521
177	12/6/2008	920	36.75	18.11	1136	685.18	5845
178	12/7/2008	890	49.24	24.89	1132	399.11	7523
179	12/9/2008	990	45.04	23.73	1278	279.79	8043
180	12/10/2008	1030	34.01	16.9	1110	329.54	7851
181	12/11/2008	1050	38.28	20.15	1137	293.73	7226
182	12/12/2008	960	23.56	12.71	549	222.25	6111
183	12/13/2008	700	29.58	16.61	741	273.47	5598
184	12/14/2008	740	37.9	19.95	823	292.06	6244
185	12/15/2008	900	38.37	21.22	992	546.95	7021
186	12/16/2008	800	42.52	21.06	1087	275.79	7327
187	12/18/2008	900	35.11	19.55	840	291.01	7391
188	12/21/2008	750	64.46	33.48	936	521.51	5780
189	12/22/2008	920	52.91	25.62	555	311.19	6207
190	12/25/2008	720	34.71	18.49	469	235.06	3909
191	12/27/2008	720	21.86	11.48	648	221.4	3850
192	12/29/2008	480	20.49	11.22	682	179.91	3907
193	12/30/2008	390	12.27	6.78	452	136.29	3022
194	12/31/2008	480	14.31	7.29	535	299.29	2312
195	1/2/2009	300	13.24	6.23	780	418.16	2252
196	1/4/2009	400	48.91	21.29	692	276.48	2004
197	1/5/2009	240	23.48	12.61	971	248.79	1952
198	4/10/2009	1160	55.28	27.87	980	529.3	8919
199	4/11/2009	1200	46.02	10.28	811.67	511.65	10555
200	4/12/2009	1240	31.57	12.18	879.31	545.83	13310
201	4/13/2009	1160	62.37	20.66	1066.07	462.83	13775
202	4/14/2009	1320	79.08	36.85	1090.74	569.37	16213
203	4/15/2009	1320	82.62	40.03	1114.29	436.41	16844
204	4/17/2009	1440	83.38	41.46	775	467.77	19272
205	4/18/2009	1520	71.44	33.7	794	493.48	21393
206	4/19/2009	1600	54.51	26.43	882.14	436.64	19450
207	4/20/2009	1600	61.93	23.85	776.67	456.9	20842
208	4/21/2009	430	66.73	34.73	880.36	577.47	14832
209	4/23/2009	1200	56.17	29.79	1022	526.72	12512
210	4/25/2009	1440	79.92	37.95	990	533.94	18043
211	4/26/2009	1240	71.46	33.54	901.85	493.45	17211
212	4/27/2009	1520	68.92	32.2	1016	490.31	18454
213	4/28/2009	1320	92.42	38.94	925	458.53	18861
214	4/29/2009	1300	69.98	32.39	1128	514.93	19244
215	5/1/2009	1240	70.92	31.47	986	528.93	17909
216	5/2/2009	1200	55.57	27.53	946	499.88	15558

217	5/3/2009	1140	51.49	24.77	996	591.36	15930
218	5/4/2009	160	73	28.11	936	484.73	10691
219	5/7/2009	680	57.29	24.39	934	584.67	7591
220	5/8/2009	870	77.73	32.73	916	575.27	10333
221	5/9/2009	1170	69.55	29.38	960	609.34	12398
222	5/10/2009	1130	73.77	35.08	1000	469.07	14079
223	5/11/2009	1080	74.1	32	1024	537.1	12387
224	5/12/2009	1080	110.34	33.86	1048.33	422.65	14463
225	5/13/2009	1080	86.28	38.31	1096.43	475.04	16990
226	5/15/2009	980	105.66	41.77	876.79	394.75	11945
227	5/16/2009	1030	94.55	42.35	992.59	405.17	15853
228	5/17/2009	1140	54.28	18.96	1092	543.42	14286
229	5/19/2009	1190	40.89	19.3	1057.41	425.49	12936
230	5/20/2009	400	54.5	18.25	1196	458.21	10451
231	5/21/2009	210	46.77	22.07	1054	361.88	6929
232	5/24/2009	360	84.33	36.39	1128.85	430.82	4815
233	5/26/2009	480	51.89	21.44	1066.07	425.31	6595
234	5/27/2009	720	50.3	18.37	1194.44	441.38	8930
235	5/28/2009	720	47.89	17.08	1016.07	414.32	8111
236	5/29/2009	600	59	28.03	1096.3	447.42	4706
237	5/30/2009	620	48.77	22.28	1080.77	448.94	4555
238	5/31/2009	720	49.66	23.83	1150	445.28	6951
239	6/1/2009	720	43.64	21.35	841.07	379.18	7005
240	6/2/2009	180	26.78	12.5	830	379.18	3457
241	6/3/2009	540	47.33	19.38	1086.67	415.91	5114
242	6/4/2009	420	45.21	16.99	965	489.07	6258
243	6/5/2009	1000	51.88	24.95	993.33	417.28	5629
244	6/6/2009	910	55.59	26.55	963.33	384.1	8235
245	7/17/2009	1440	63.84	28.16	1120	357.82	11057
246	7/18/2009	1380	64.24	26.48	1779	304.09	10998
247	7/19/2009	1440	60.21	22.69	1992	214.79	12184
248	7/20/2009	1440	58.53	19.59	1214	580.56	10696
249	7/21/2009	1440	55.51	22.78	1029	313.51	10333
250	7/22/2009	1440	54.85	23.29	1056	435.31	10347
251	7/26/2009	600	72.88	27.49	1113	728.63	7497
252	7/27/2009	450	68.15	27.3	1088	533.16	6515
253	8/3/2009	600	72.95	28.64	980	602.85	5329
254	8/4/2009	450	61.96	25.86	737	349.89	5854
255	8/5/2009	450	64.1	27.1	1079	394.49	5054
256	8/6/2009	455	81.32	33.63	896	337.47	5074
257	8/7/2009	450	79.27	34.6	1613	240.76	4717
258	9/15/2009	390	58.87	27.28	895	329.65	2500
259	9/16/2009	500	47.01	19.09	933	445.51	2683
260	9/23/2009	560	52.42	21.93	732	237.24	3067

261	10/5/2009	320	59	27.28	1180	572.15	4310
262	10/6/2009	700	84.56	52.78	802	98.97	4078
263	10/7/2009	580	47.94	22.32	1098	296.13	3795
264	10/9/2009	575	41.64	19.31	758	154.21	2881
265	10/17/2009	680	64.58	31.05	1118	460.39	4047
266	10/18/2009	700	61.46	28.19	1021	511.02	3770
267	10/19/2009	750	72.07	36.26	903	407.96	4624
268	10/20/2009	400	69.16	31.74	1024	432.03	3942
269	10/21/2009	580	67.83	29.9	1677	278.85	4299
270	10/22/2009	400	64.22	29.82	1208	764.8	3914
271	11/5/2009	100	84.7	38.74	1180	572.15	1695
272	11/17/2009	360	40.68	20.01	1118	460.39	3191
273	11/19/2009	420	61.06	27.47	903	407.96	3517
274	11/21/2009	450	38.52	16.12	1677	278.85	3644
275	11/22/2009	450	47.31	22.44	1208	764.8	3683
276	1/9/2010	660	62.48	30.5	1103	750.3	5253
277	1/10/2010	690	50.13	24.73	1188	655.71	5538
278	1/11/2010	660	61.49	30.45	841	204.26	5401
279	3/4/2010	1210	70.77	30.87	510	295.58	4963
280	3/5/2010	1350	78.5	22.55	688	473.74	5180
281	8/19/2010	1350	55.99	26.53	917	493.78	11433
282	8/21/2010	1200	33.75	16.51	909	504.78	13015
283	8/22/2010	1080	61.19	28.77	805	387.73	8477
284	8/23/2010	1510	54.3	25.82	684	287.48	12051
285	8/24/2010	1380	54.19	26.52	780	431.3	11584
286	8/27/2010	1010	57.84	28.19	1350	933.8	12793
287	8/28/2010	1360	59.37	25.7	1331.67	930.56	10774
288	8/29/2010	1680	35.15	15.33	398	58.18	8705
289	8/30/2010	405	21.7	8.66	533	131.3	8224
290	8/31/2010	1250	37.15	19.94	389	50.86	4607
291	9/1/2010	820	59.9	27.25	1103	695.1	6354
292	9/2/2010	1640	43.51	21.93	492.11	279.95	7119
293	9/3/2010	1600	60.51	28.49	892.86	389.07	7670
294	9/4/2010	1400	58.32	28.85	721	232.67	9344
295	9/5/2010	1160	61.43	29.83	802	385.29	6070
296	9/6/2010	1240	58.84	27.37	914	525.49	9169
297	9/7/2010	1320	47.53	11.28	1012	687.57	8722
298	9/12/2010	1440	43.85	22.57	718	361.57	7900
299	9/13/2010	1520	43.4	22.66	705	307.2	7931
300	9/15/2010	920	59.15	33.29	376	138.34	4337
301	9/16/2010	1340	62.61	31.36	694	399.82	8686
302	9/17/2010	1620	56.32	29.02	773	368.69	10024
303	9/18/2010	1680	50.19	26.69	907	507.53	12093
304	9/19/2010	1620	57.49	30.8	521	294.9	10767

305	9/21/2010	1040	43.17	21.6	942	345.14	9590
306	9/22/2010	1460	55.35	29.82	852	355.94	8305
307	9/23/2010	1760	51.47	26.9	749	249.33	10215
308	9/26/2010	1960	38.92	21.28	399	92.27	12172
309	9/27/2010	2200	45.82	22.69	599	145.62	13213
310	9/28/2010	2200	58.13	28.45	860.29	288.99	12864
311	9/29/2010	2220	42.42	20.82	746.25	293.5	12297
312	10/3/2010	980	67.95	33.19	1025.71	681.09	10163
313	10/5/2010	400	49.68	25.28	815	322.1	6267
314	10/7/2010	1320	36.79	19.42	647.37	94.11	5873
315	10/10/2010	500	48.71	22.83	452.22	60.97	4822
316	10/12/2010	600	39.02	20.25	426.04	85.57	4120
317	10/17/2010	1450	62.84	25.86	623.26	211.51	5206
318	10/19/2010	1360	57.12	25.18	604.44	238.39	8967
319	10/21/2010	1310	49.4	21.72	513.04	186.09	6872
320	10/24/2010	1570	51	24.57	531.11	175.78	9524
321	10/26/2010	1420	55.95	26.71	514.44	173.88	8541
322	11/2/2010	1240	48.14	22.72	464.89	100.06	5377
323	11/4/2010	1810	67.19	31.05	852.33	560.87	7499
324	11/7/2010	1920	42.69	20.26	530.61	159.43	9985
325	11/9/2010	1960	48.52	23.47	480.43	142.42	10283
326	11/21/2010	800	42.01	22.96	855.56	605.27	6629
327	11/23/2010	1085	53.9	28.7	753.33	360.56	9263
328	11/25/2010	1114	49.93	24.27	857.61	259.09	7765
329	12/12/2010	600	50.1	22.53	631.52	153.09	3914
330	12/16/2010	560	50.72	27.77	333.33	35.77	4528
331	12/19/2010	1070	47.27	22.8	546.74	195.44	7186
332	12/23/2010	1410	54.41	28.16	629.07	189.77	7815
333	12/30/2010	1020	67.89	34.03	398.61	48.95	8143
334	1/2/2011	1270	49.94	25.21	675.58	287.7	7184
335	1/4/2011	1130	49.04	25.15	761.63	511.48	7310
336	1/6/2011	840	57.44	29.61	776.19	449.42	7311
337	1/9/2011	1170	44.19	21.43	732.61	412.4	8566
338	1/11/2011	1040	54.28	26.32	2201.22	797.23	9131
339	1/18/2011	1610	65.92	37.24	1070	716.83	9985
340	1/20/2011	1480	46.46	24.97	704.65	120.97	13858
341	1/23/2011	1790	62.72	33.78	1574.39	1226.66	12406
342	1/25/2011	970	70.56	33.77	1045	803.74	11424
343	1/27/2011	1150	58.26	36.15	1065.38	832.12	10833
344	1/30/2011	1440	35.35	11.31	1017.05	798.9	12400
345	2/2/2011	1210	49.69	25.77	743.48	457.42	11046
346	2/6/2011	1470	79.14	26.29	1126.92	858.64	10157
347	2/8/2011	1150	92.84	67.51	954.65	743.94	10377
348	2/13/2011	1530	47.49	25.96	839.58	459.59	12132

349	2/15/2011	1430	54.15	32.61	521.25	166.17	12363
350	2/17/2011	1240	60.05	33.65	934.09	678.64	14389
351	2/24/2011	1830	68.63	37.5	1227.38	731.21	11112
352	2/27/2011	1270	33.97	19.17	446	65.73	9362
353	3/6/2011	920	59.25	33.32	538.46	174.2	8697
354	3/8/2011	1320	58.06	30.91	965.91	763.24	11115
355	3/10/2011	900	45.59	24.01	378.89	150.13	8002
356	3/13/2011	960	44.13	23.96	540	64.32	8351
357	3/17/2011	1080	36.52	19.21	487	147.26	8257
358	3/22/2011	1120	68.92	34.53	1045.45	594.17	6897
359	3/24/2011	1010	44.86	24.51	461.22	85.97	9276
360	3/27/2011	1280	43.12	18.91	2931.91	316.96	18387
361	3/29/2011	1070	56.64	30.42	613.64	167.05	11951
362	3/31/2011	1270	49.59	24.05	1619.15	207.8	9840
363	4/3/2011	1270	71.51	25.49	1613.75	653.78	11900
364	4/7/2011	870	40.19	16.99	1076.25	450.48	9715
365	4/10/2011	800	46.17	26.58	571.74	99.67	8100
366	4/12/2011	820	43.22	27.71	593.75	109.17	7428
367	4/17/2011	1010	62.35	31.13	1578.05	223.76	13364
368	4/19/2011	310	82.97	39.52	961.45	697.49	8990
369	4/21/2011	905	55.91	27.26	691.3	306.45	10658
370	4/24/2011	810	47.89	24.06	633.71	175.03	9141
371	4/26/2011	720	53.93	32.63	829.55	449.26	11281
372	5/1/2011	1090	0	0	0	0	14776
373	5/3/2011	1110	0	0	0	0	12201
374	5/5/2011	550	0	0	0	0	11315
375	5/8/2011	1210	0	0	0	0	9600
376	5/10/2011	780	0	0	0	0	7912
377	5/15/2011	770	0	0	0	0	8423
378	5/24/2011	1060	0	0	0	0	9830
379	5/26/2011	795	0	0	0	0	7232
380	5/29/2011	1180	0	0	0	0	10675
381	5/31/2011	1150	0	0	0	0	10157
382	6/2/2011	1410	0	0	0	0	11655
383	6/9/2011	860	0	0	0	0	9743
384	6/14/2011	810	0	0	0	0	13845
385	6/16/2011	910	0	0	0	0	8493
386	6/23/2011	920	0	0	0	0	9534
387	6/30/2011	1140	0	38.33	0	0	11226
388	7/3/2011	1130	74	35.52	226	179.1	7078
389	7/5/2011	1030	45.01	21.61	892.5	110.47	8974
390	7/7/2011	910	46.88	22.5	308.33	162.4	6132
391	7/10/2011	940	38.19	18.33	277.17	109.25	5562
392	7/24/2011	810	71.55	34.34	430	63.43	5870
393	7/26/2011	1100	99.82	47.91	410	48.98	5207
394	7/28/2011	1110	86.02	41.29	1088.3	380.23	7551

Appendix B: Machine Learning algorithm by using Support Vector Machine

```
import pandas as pd
from sklearn import svm
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import warnings
from sklearn import preprocessing

# Load data from CSV file
df = pd.read_csv(r'C:\Users\Ranjan Gaida\Desktop\data.csv', encoding='utf-8')

df.head()
#Function to test model performance changes with feature elimination MSE
def Mean_Square_Error(model, x_test, y_test):
    prediction = model.predict(x_test)
    print ("Mean Square error of model:", mean_squared_error(y_test, prediction))

#Setting a parameter for SVM model
C = 1.0

# Identifying the target feature by splitting the dataset
samples = df.filter(['TMF', 'TKM_percent', 'TUKM_percent', 'Alcantine', 'Fatty_Accid'])
scores = df.filter(['biogas'])

# Deleting the 'Date' column from the dataset as supposed 'irrelevent' or 'unprocessible'
del df['Date']

# Defining the number of features to investigate
nFeatures = len(df.columns) - 1

rfeIndex = nFeatures

#Recursively eliminate features based on the lowest weight
while True:
    #Split into training and testing
    x_train, x_test, y_train, y_test = train_test_split(samples, scores, test_size = 0.50,
train_size=0.50)

    #Create SVM model using a linear kernel
    model = svm.SVR(kernel='linear', C=C).fit(x_train, y_train)
    coef = model.coef_

    #Print co-efficients of features
    for i in range(0, nFeatures):
        print(samples.columns[i-1],":", coef[0][i-1])
#Find the minimum weight among features and eliminate the feature with the smallest weight
    min = coef[0][0]
```

```
        index = 0
        for i in range(0, rfeIndex):
                if min > coef[0][i-1]:
                        index = index + 1
                        min = coef[0][i-1]
        if len(samples.columns) == 1:
                print("After recursive elimination we have the", samples.columns[index],
"feature with a score of:", min)
                Mean_Square_Error(model, x_test, y_test)
                break
        else:
                print ("Lowest feature weight is for", samples.columns[index], "with a value
of:", min)
                print ("Dropping feature", samples.columns[index])

                #Drop the feature in the 'samples' dataframe based on the lowest feature index
                samples.drop(samples.columns[index], axis = 1, inplace = True)
                Mean_Square_Error(model, x_test, y_test)
                print ("\n")
                rfeIndex = rfeIndex - 1
                nFeatures = nFeatures - 1
```

Results:

```
Fatty_Acid : 1.6436548300928848
TMF : 6.657213896331768
TKM_percent : 5.457184279436092
TUKM_percent : 29.474537096000077
Alcantine : 0.08988235420656565
Lowest feature weight is for TUKM_percent with a value of: 0.08988235420656565
Dropping feature TUKM_percent
Mean Square error of model: 8121951.551265039

Fatty_Accid : 2.1127632609003513
TMF : 5.596732539106597
TKM_percent : 24.183147355131958
Alcantine : -0.44507553501216535
Lowest feature weight is for Alcantine with a value of: -0.44507553501216535
Dropping feature Alcantine
Mean Square error of model: 9260311.29445238

Fatty_Accid : 0.365987946351197
TMF : 5.838201771366585
TKM_percent : 26.53703460708777
Lowest feature weight is for TKM_percent with a value of: 0.365987946351197
Dropping feature TKM_percent
Mean Square error of model: 8077565.72189119
```

```
Fatty_Accid : 1.3184196023751724
TMF : 5.890918980730476
Lowest feature weight is for Fatty_Accid with a value of: 1.3184196023751724
Dropping feature Fatty_Accid
Mean Square error of model: 10573677.283554532

TMF : 6.715555555555966
After recursive elimination we have the TMF feature with a score of: 6.715555555555966
Mean Square error of model: 8615491.909850966
```

Appendix C: Machine Learning algorithm by using Artificial Neural Network

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split

# Load data from CSV file
df = pd.read_csv(r'C:\Users\Ranjan Gaida\Desktop\data.csv', encoding='utf-8')

# Print first 5 rows of dataframe and its shape
print(df.head())
print(df.shape)

# Extract features and target variable from dataframe
samples = df[['TMF', 'TKM_percent', 'TUKM_percent', 'Alcantine', 'Fatty_Accid']].values
scores = df[['biogas']].values

# Print all of samples and scores arrays
print(samples[:394])
print(scores[:394])

nFeatures = samples.shape[1]

rfeIndex = nFeatures
converged = False

while not converged:
    # Split data into training and testing sets
    x_train, x_test, y_train, y_test = train_test_split(samples, scores, test_size=0.50,
train_size=0.50)

    # Train MLP regressor model
    model = MLPRegressor(hidden_layer_sizes=(60,60), activation='relu', solver='adam',
max_iter=1000, tol=1e-5).fit(x_train, y_train.ravel())

    # Make predictions on testing set
    prediction = model.predict(x_test)

    # Compute and print evaluation metrics
    mse = mean_squared_error(y_test, prediction)
    r2 = r2_score(y_test, prediction)

    print("MSE:", mse)
    print("R² score:", r2)

    # Get model coefficients
    coef = model.coefs_

    # Find index of feature with smallest absolute weight
    min_weight = np.abs(coef[0]).min()
    index = np.where(np.abs(coef[0]) == min_weight)[0][0]

    if samples.shape[1] == 1:
        print("After recursive elimination we have the", df.columns[index], "feature with a score
of:", min_weight)
        converged = True
    else:
        print("Lowest feature weight is for", df.columns[index], "with a value of:", min_weight)
        print("Dropping feature", df.columns[index])

        samples = np.delete(samples, index, axis=1)
        print("\n")
        rfeIndex = rfeIndex - 1
        nFeatures = nFeatures - 1

    # Check if the model has converged
    if model.n_iter_ >= model.max_iter:
        print("Model has not converged after", model.max_iter, "iterations.")
        break
    else:
        print("Model has converged after", model.n_iter_, "iterations.")
        converged = True

# Make predictions on testing set
prediction = model.predict(x_test)

# Create scatter plot of actual vs. predicted values
plt.scatter(y_test, prediction)

plt.xlabel('Actual biogas values')
plt.ylabel('Predicted biogas values')
plt.title('Actual vs. predicted biogas values')
plt.show()

mse_list = []
r2_list = []

while True:
    # your existing code here
    # ...

    mse_list.append(mse)
    r2_list.append(r2)
    if converged:
        break

# plot the MSE and R$^2$ score over iterations
iterations = range(len(mse_list))

fig, ax1 = plt.subplots()

ax1.plot(iterations, mse_list, 'b-', label='MSE')
ax1.set_xlabel('Iterations')
ax1.set_ylabel('MSE', color='*')
ax1.tick_params('y', colors='*')

ax2 = ax1.twinx()
ax2.plot(iterations, r2_list, 'r-', label='R$^2$ score')
ax2.set_ylabel('R$^2$ score', color='r')
ax2.tick_params('y', colors='r')

plt.title('Model performance over iterations')
fig.legend(loc='upper left')
plt.show()
```

Results:

```
MSE: 6032492.224757484
R$^2$ score: 0.5931522505321214
Lowest feature weight is for Date with a value of: 6.6571525218522074e-21
Dropping feature Date
```

Appendix D: Machine Learning Algorithm for Random Forest method

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split

# Load data from CSV file
df = pd.read_csv(r'C:\Users\Ranjan Gaida\Desktop\data.csv', encoding='utf-8')

# Print first 5 rows of dataframe and its shape
print(df.head())
print(df.shape)

# Extract features and target variable from dataframe
samples = df[['TMF', 'TKM_percent', 'TUKM_percent', 'Alcantine', 'Fatty_Accid']].values
scores = df[['biogas']].values

# Print all of samples and scores arrays
print(samples[:394])
print(scores[:394])

nFeatures = samples.shape[1]

rfeIndex = nFeatures
converged = False

while not converged:
    # Split data into training and testing sets
    x_train, x_test, y_train, y_test = train_test_split(samples, scores, test_size=0.50,
train_size=0.50)

    # Train Random Forest regressor model
    model = RandomForestRegressor(n_estimators=1000, max_depth=20,
random_state=50).fit(x_train, y_train.ravel())

    # Make predictions on testing set
    prediction = model.predict(x_test)

    # Compute and print evaluation metrics
    mse = mean_squared_error(y_test, prediction)
    r2 = r2_score(y_test, prediction)

    print("MSE:", mse)
    print("R² score:", r2)
   # Get feature importances
```

```
    importances = model.feature_importances_

    # Find index of feature with smallest feature importance
    index = np.argmin(importances)

    if samples.shape[1] == 1:
        print("After recursive elimination we have the", df.columns[index], "feature with an importance score of:", importances[index])
        converged = True
    else:
        print("Lowest feature importance is for", df.columns[index], "with a value of:", importances[index])
        print("Dropping feature", df.columns[index])

        samples = np.delete(samples, index, axis=1)
        print("\n")
        rfeIndex = rfeIndex - 1
        nFeatures = nFeatures - 1

    # Check if the model has converged
    if converged:
        break

# Make predictions on testing set
prediction = model.predict(x_test)

# Create scatter plot of actual vs. predicted values
plt.scatter(y_test, prediction)

plt.xlabel('Actual biogas values')
plt.ylabel('Predicted biogas values')
plt.title('Actual vs. predicted biogas values')
plt.show()

mse_list = []
r2_list = []

while True:
    # your existing code here
    # ...

    mse_list.append(mse)
    r2_list.append(r2)

    if converged:
        break

# plot the MSE and R2 score over iterations
```

```
iterations = range(len(mse_list))

fig, ax1 = plt.subplots()

ax1.plot(iterations, mse_list, 'b-', label='MSE')
ax1.set_xlabel('Iterations')
ax1.set_ylabel('MSE', color='*')
ax1.tick_params('y', colors='*')

ax2 = ax1.twinx()
ax2.plot(iterations, r2_list, 'r-', label='R2 score')
ax2.set_ylabel('R2 score', color='r')
ax2.tick_params('y', colors='r')

plt.title('Model performance over iterations')
fig
```

Results:
MSE: 6695312.177294152
R^2 score: 0.6204934804271711
Lowest feature importance is for Alcantine with a value of: 0.08181945079976062
Dropping feature Alcantine

MSE: 6740958.4483600715
R^2 score: 0.5154171094167677
Lowest feature importance is for TKM_percent with a value of: 0.10739103927710365
Dropping feature TKM_percent

MSE: 6311125.255567056
R^2 score: 0.5842450796421896
Lowest feature importance is for TKM_percent with a value of: 0.1747513373147001
Dropping feature TKM_percent

MSE: 7455388.267482883
R^2 score: 0.5006024555558546
Lowest feature importance is for TMF with a value of: 0.31251848816533406
Dropping feature TMF

MSE: 12790646.975849813
R^2 score: 0.21101056873340185
After recursive elimination we have the Date feature with an importance score of: 1.0

Appendix E: Machine learning algorithm for K-Nearest Neighbours

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split

# Load data from CSV file
df = pd.read_csv(r'C:\Users\Ranjan Gaida\Desktop\data.csv', encoding='utf-8')

# Print first 5 rows of dataframe and its shape
print(df.head())
print(df.shape)

# Extract features and target variable from dataframe
samples = df[['TMF', 'TKM_percent', 'TUKM_percent', 'Alcantine', 'Fatty_Accid']].values
scores = df[['biogas']].values

# Print all of samples and scores arrays
print(samples[:394])
print(scores[:394])

nFeatures = samples.shape

rfeIndex = nFeatures
converged = False

while not converged:
   # Split data into training and testing sets
   x_train, x_test, y_train, y_test = train_test_split(samples, scores, test_size=0.50,
train_size=0.50)

   # Train KNN regressor model
   model = KNeighborsRegressor(n_neighbors=5).fit(x_train, y_train.ravel())

   # Make predictions on testing set
   prediction = model.predict(x_test)

   # Compute and print evaluation metrics
   mse = mean_squared_error(y_test, prediction)
   r2 = r2_score(y_test, prediction)

   print("MSE:", mse)
   print("R² score:", r2)

   # Find index of feature with smallest variance
```

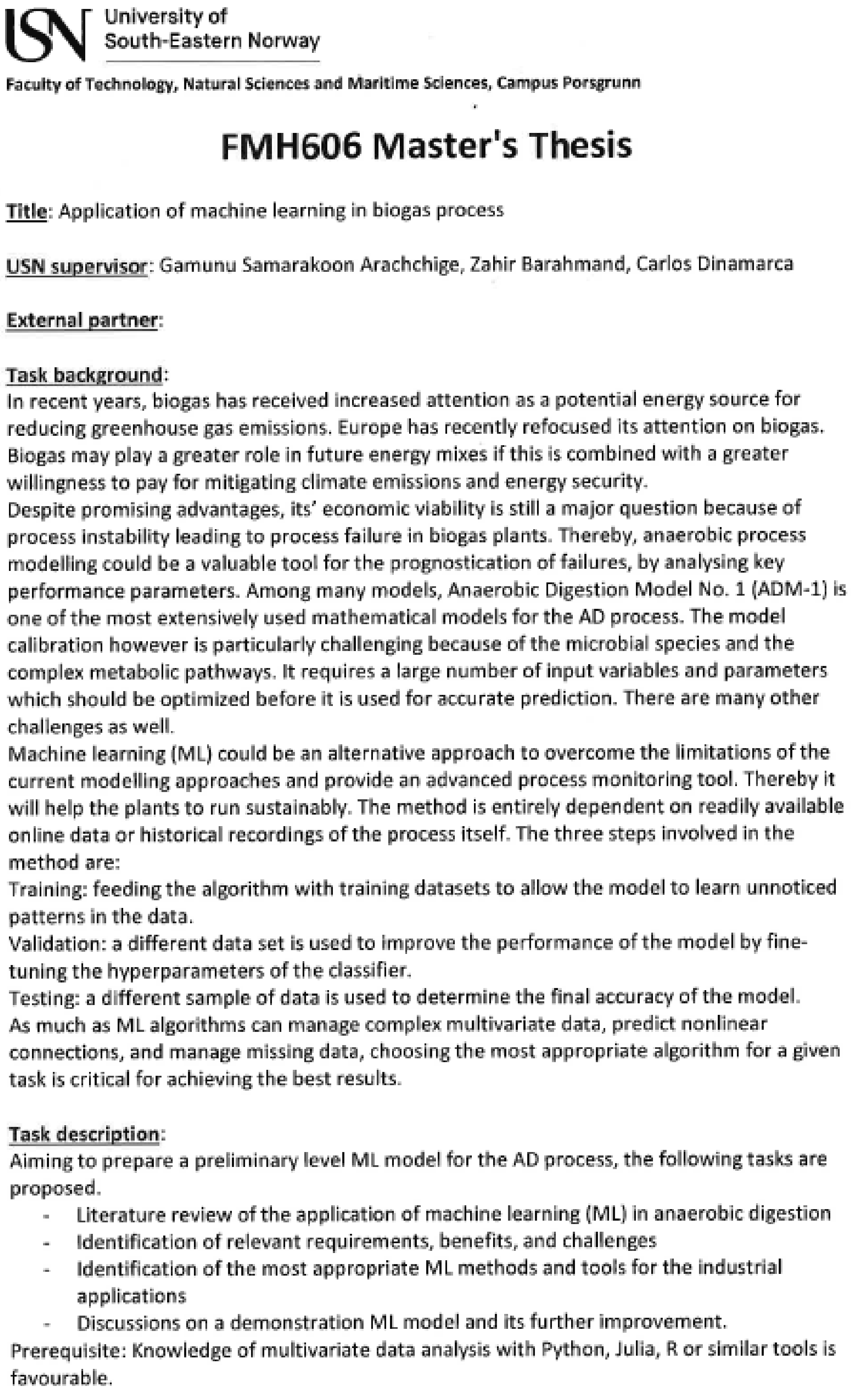
```
    variances = np.var(x_train, axis=0)
    index = np.argmin(variances)

    if samples.shape[1] == 1:
        print("After recursive elimination we have the", df.columns[index], "feature with a
variance of:", variances[index])
        converged = True
    else:
        print("Lowest variance is for", df.columns[index], "with a value of:", variances[index])
        print("Dropping feature", df.columns[index])

        samples = np.delete(samples, index, axis=1)
        print("\n")
        rfeIndex = rfeIndex - 1
        nFeatures = nFeatures - 1

    # Check if the model has converged
    if converged:
        break

# Make predictions on testing set
prediction = model.predict(x_test)

# Create scatter plot of actual vs. predicted values
plt.scatter(y_test, prediction)

plt.xlabel('Actual biogas values')
plt.ylabel('Predicted biogas values')
plt.title('Actual vs. predicted biogas values')
plt.show()

mse_list = []
r2_list = []

while True:
    # your existing code here
    # ...
  mse_list.append(mse)
    r2_list.append(r2)

 if converged:
        break

# plot the MSE and $R^2$ score over iterations
iterations = range(len(mse_list))
fig, ax1 = plt.subplots()

ax1.plot(iterations, mse_list, 'b-', label='MSE')
```

```
ax1.set_xlabel('Iterations')
ax1.set_ylabel('MSE', color='*')
ax1.tick_params('y', colors='*')

ax2 = ax1.twinx()
ax2.plot(iterations, r2_list, 'r-', label='R$^2$ score')
ax2.set_ylabel('R$^2$ score', color='r')
ax2.tick_params('y', colors='r')

plt.title('Model performance over iterations')
fig.legend(loc='upper left')
plt.show()
```

Results:

```
MSE: 6854866.264568527
R$^2$ score: 0.5158138094287092
Lowest variance is for TKM_percent with a value of: 85.28819688731996
Dropping feature TKM_percent

MSE: 7175282.987005074
R$^2$ score: 0.49049031272392896
Lowest variance is for TMF with a value of: 402.4094443247701
Dropping feature TMF

MSE: 7783316.454822334
R$^2$ score: 0.42312191132573573
Lowest variance is for TKM_percent with a value of: 35992.62251995157
Dropping feature TKM_percent

MSE: 8342607.588832487
R$^2$ score: 0.442177016325216
Lowest variance is for TMF with a value of: 117994.20537433065
Dropping feature TMF

MSE: 9304338.14680203
R$^2$ score: 0.4234968046411981
After recursive elimination we have the Date feature with a variance of:
172866.55765415236
```

Appendix F: Project Thesis Description.



University of South-Eastern Norway

Faculty of Technology, Natural Sciences and Maritime Sciences, Campus Porsgrunn

FMH606 Master's Thesis

Title: Application of machine learning in biogas process

USN supervisor: Gamunu Samarakoon Arachchige, Zahir Barahmand, Carlos Dinamarca

External partner:

Task background:

In recent years, biogas has received increased attention as a potential energy source for reducing greenhouse gas emissions. Europe has recently refocused its attention on biogas. Biogas may play a greater role in future energy mixes if this is combined with a greater willingness to pay for mitigating climate emissions and energy security.

Despite promising advantages, its' economic viability is still a major question because of process instability leading to process failure in biogas plants. Thereby, anaerobic process modelling could be a valuable tool for the prognostication of failures, by analysing key performance parameters. Among many models, Anaerobic Digestion Model No. 1 (ADM-1) is one of the most extensively used mathematical models for the AD process. The model calibration however is particularly challenging because of the microbial species and the complex metabolic pathways. It requires a large number of input variables and parameters which should be optimized before it is used for accurate prediction. There are many other challenges as well.

Machine learning (ML) could be an alternative approach to overcome the limitations of the current modelling approaches and provide an advanced process monitoring tool. Thereby it will help the plants to run sustainably. The method is entirely dependent on readily available online data or historical recordings of the process itself. The three steps involved in the method are:

Training: feeding the algorithm with training datasets to allow the model to learn unnoticed patterns in the data.

Validation: a different data set is used to improve the performance of the model by fine-tuning the hyperparameters of the classifier.

Testing: a different sample of data is used to determine the final accuracy of the model.

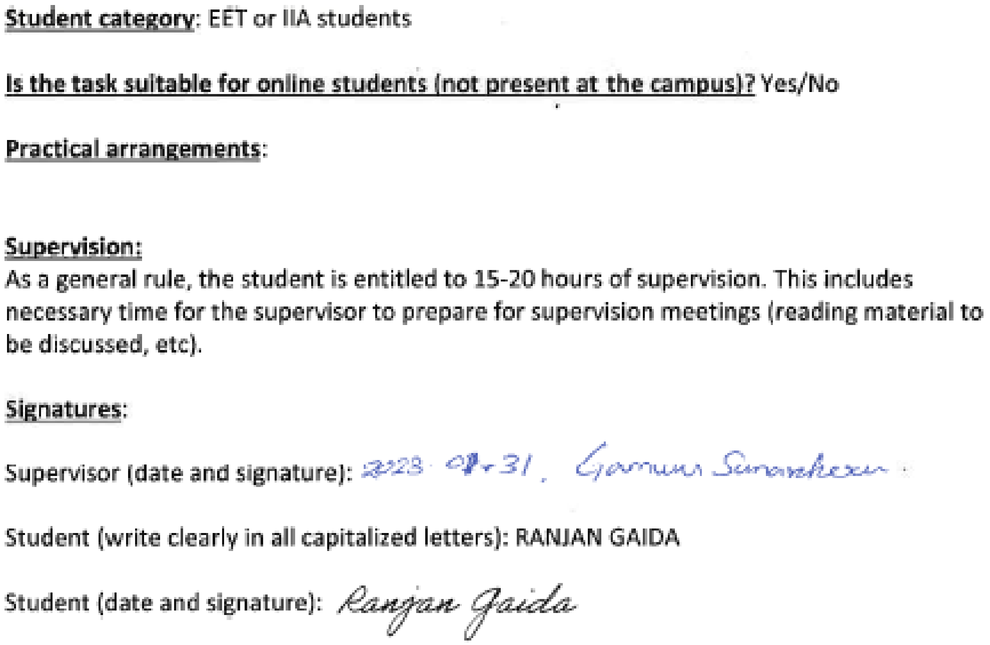
As much as ML algorithms can manage complex multivariate data, predict nonlinear connections, and manage missing data, choosing the most appropriate algorithm for a given task is critical for achieving the best results.

Task description:

Aiming to prepare a preliminary level ML model for the AD process, the following tasks are proposed.

- Literature review of the application of machine learning (ML) in anaerobic digestion
- Identification of relevant requirements, benefits, and challenges
- Identification of the most appropriate ML methods and tools for the industrial applications
- Discussions on a demonstration ML model and its further improvement.

Prerequisite: Knowledge of multivariate data analysis with Python, Julia, R or similar tools is favourable.



Student category: EET or IIA students

Is the task suitable for online students (not present at the campus)? Yes/No

Practical arrangements:

Supervision:
As a general rule, the student is entitled to 15-20 hours of supervision. This includes necessary time for the supervisor to prepare for supervision meetings (reading material to be discussed, etc).

Signatures:

Supervisor (date and signature): 2023.08.31, Gamini Samarakoon

Student (write clearly in all capitalized letters): RANJAN GAIDA

Student (date and signature): Ranjan Gaida