INVESTIGATING RENEWABLE ENERGY SYSTEMS USING ARTIFICIAL INTELLIGENCE TECHNIQUES

By

Kamran Niroomand

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING

Department of Electrical and Computer Engineering

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

October 2023

St. John's, Newfoundland and Labrador, Canada

ABSTRACT

This research investigated applying Artificial Intelegence (AI) and Machine Learning (ML) to renewable energy through three studies. The first study characterized and mapped the recent research landscape in the field of AI applications for various renewable energy systems using Natural Language Prcoessing (NLP) and ML models. It considered published documetns at Scopus database in the period (2000-2021). The second study built hybrid Catboost-CNN-LSTM architecture pipeline to predict an industrial-scale biogas plant's daily biogas production and investigate the feedstock components importance on it. The third study investigated prediciting biogas yield of various subtrates and the significance of each organic component (carbohydrates, proteins, fats/lipids, and legnin) in biogas production using hybrid VAE-XGboost model.

The first study showed seven main metatopics and ascent of "deep learning (DL)" as a prominent methodology led to an increase in intricate subjects, including the optimization of power costs and the prediction of wind patterns. Also, a growing utilization of DL approaches for the analysis of renewable energy data, particularly in the context of wind and solar photovoltaic systems. The research themes and trends observed in the first study signify substantial recent investments in advanced AI learning techniques. The developed Catboost-CNN-LSTM pipeline achived a significant results and presented a superior approach when compared to previous relevant studies by eliminating the requirement for feature engineering, enabling direct prediction of biogas yield without the need for converting it into a classification task. The VAE-XGboost pipeline could ovcercome data limitation in the field and produced significant results. It has shown that the "fats" category is the most influential group on the methane production in biogas plants, however, "proteins" illustrated the lowest impact on biogas production.

Keywords: Machine Learning, Natural Language processing, Time series forecasting, Biogas, Renewable energy.

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Noori Saady and Dr. Carlos Bazan, for their guidance and feedback throughout my M.Eng. program. This work would not have been possible without their endless support and commitment during this journey.

The financial support for this work was provided by the Persistent, Emerging, and Organic Pollution in the Environment (PEOPLE), Memorial University of Newfoundland, the Department of Fisheries, Forestry, and Agrioculture, Government of Newfoundland and Labrador through the Canadian Agriculture Partnership Program.

Finally, my deep heartfelt gratitude goes to my parents, my sister, Parmida, who have always supported me and have encouraged me to move forward at most difficult times.

Table of Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS AND SYMBOLS

Acronym	Definition
AcoD	Anaerobic co-digestion
AD	Anaerobic digestion
ADM1	Anaerobic Digestion Model 1
ADP	Adenosine Diphosphate
ADP	Adenosine Diphosphate
AI	Artificial Intelligence
AMPTS	Automatic Methane Potential Test System
ANFIS	Adaptive Network-based Fuzzy Inference System
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ATP	Adenosine triphosphate
ATP	Adenosine Triphosphate
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short Term Memory
BiRNN	Bidirectional Recurrent Neural Network
BMP	Biochemical Methane Potential
BOD	Biological oxygen demand
c-TF-IDF	Class-based Term Frequency-Inverse Document Frequency

C/N	Carbon/ nitrogen ratio
CNN	Convolutional Neural Network
COD	Chemical oxygen demand
DIET	Direct Interspecies Electron Transfer
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
DTM	Dynamic Topic Modeling
EDA	Exploratory Data Analysis
FTIR	Fourier Transform mid-Infrared spectroscopy
GA	Genetic Algorithm
GAN	Generative Adversial Network
GHG	Greenhouse gas
GIS	Geographical Information System
GRU	Gated Recurrent Unit
HRT	Hydraulic retention time
IoT	Internet of Things
IR	Infrared
LSTM	Long Short Term Memory
MAE	Mean Absolute Error

MAPE	Mean Absolute Precision Error
ML	Machine Learning
MLP	Multi-layer Perception
MSE	Mean Square Error
NIR	Near Infrared
NMAE	Normlized Mean Absolute Error
NMAPE	Normlized Mean Absolute Precision Error
NMSE	Normlized Mean Square Error
NRMSE	Normlized Root Mean Square Error
OFMSW	Organic Fraction of Municipal Solid Waste
OLR	Organic Loading Rate
OP	Olive Pomace
PSO	Particle Swarm Optimization
PV	Photovoltaic
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RSM	Response Surface Methodology
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

TOC	Total Organic Carbon
TS	Total solid
TSF	Time Series Forecasting
TVS	Total volatile solids
UMAP	Uniform Manifold Approximation and Projection
UV	UltraViolet
VAE	Variational Auto Encoder
VFA	Volatile Fatty Acids
XGboost	eXtreme Gradient Boost

Symbols	**Definition**
b_j^n	The bias of n j^{th} of the feature map
o_i	Actual value
o_{max}	Maximum actual value
o_{min}	Minimum actual value
p_i	Predicted value
y_{ij}^n	Result of the n CL
$f_{t,c}$	Models the frequency of term *t* within a class *c*
$f_{t,d}$	Models the frequency of term *t* in document *d*
w	Kernels' weight
Z	Latent space

σ^2	Variance
μ	Mean vector
σ	Activation function
Σ	Diagonal covariance matrix

CHAPTER ONE

INTRODUCTION

1.1. BACKGROUND AND MOTIVATION

Renewable energy systems have become important with the ever-increasing demand for energy, limitations of fossil fuel resources, and concerns about sustainability. On the other hand, Artificial Intelligence (AI) is a booming sector. AI-based models are leveraged into renewable energy areas at an increasing rate due to their capability to handle complicated problems and high dimensional data (Jha et al., 2017).

AI is the study of how to build or program computers to enable them to do what human minds can do (Boden, 1996). Developments in AI methods, increasing the available dataset, and computer hardware have led to significant growth in leveraging AI modeling in renewable energy. Researchers and scientists in related areas have started to employ various AI techniques in different renewable energy systems for different purposes such as optimization, prediction, etc. One of the AI subfields that is mainly used in renewable energy systems is Machine Learning (ML). In general, classification modeling estimates a mapping function (f) from inputs to discrete output. However, regression modeling estimates a mapping function from input variables to a continuous output variable (Loh, 2011).

ML can be divided into three main categories: supervised, unsupervised, and semi-supervised learning (Figure 1.1). The term "supervised" is the learning system based on labels corresponding to training instances (Cunningham et al., 2008) and is suitable for classification and regression tasks. On the other hand, unsupervised learning does not need an annotated dataset;

thus, it is used of an unlabeled dataset and is suitable for clustering and dimension reduction (McAlpine and Michelow, 2022; Miorelli et al., 2021). Semi-supervised learning is a category between supervised and unsupervised where many unlabeled samples and a small number of labeled instances are considered together to build a better model (Ashfaq et al., 2017). Deep Learning (DL) is an advanced subfield of ML that is leveraged into complicated problems, like accurate Time Series Forecasting (TSF), and their results mostly are considerably outperforming their conventional ML counterparts (Sezer et al., 2020).

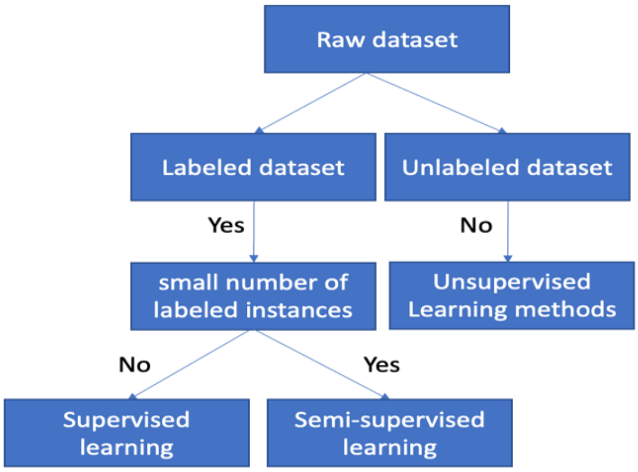

Figure 1.1. Main categories of the machine learning (ML) methods (Lim et al, 2021).

The purpose of using TSF is to predict the output variables at a future time point, considering the learning from historic time points. Various Deep Neural Network (DNN) architectures have been developed to fit a broad range of time series datasets within the different domains (Lim and Zohren, 2021), due to their capabilities in handling nonlinear features and data structures with high-level of invariance (Wang et al., 2019). One of the common types of deep learning architecture employed in renewable energy TSF tasks is Recurrent Neural Network (RNN) based models considering their ability to process sequential time series inputs (Hu and Chen, 2018). RNN is a class of Artificial Neural Networks (ANN) (Yu et al., 2018) in which the unit takes the current

and last step input data point at the same time, and the output depends on the previous data points (Sezer et al., 2020). Vanilla RNN is the simplest RNN algorithm that has limited application since the gradient vanishes during the training part of this method (Chung et al., 2014). Hence, there are types of RNN-based models such as Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), Gated Recurrent Unit (GRU) (Cho et al., 2014), bidirectional RNN (BiRNN), bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997), and bidirectional GRU (BiGRU) (Zhao et al., 2018), have been developed to mitigate the drawback. These developed models are leveraged into a wide variety of areas, like tool wear prediction, stress–strain behavior of soil, fog forecasting, stock price prediction, etc. (Wang et al., 2019; Zhang et al., 2020; Miao et al., 2020; Sethia and Raut, 2018).

Due to the increasing availability of data in renewable energy systems, improved quality of computational hardware, and development in AI techniques, academic and industrial researchers in various fields, including renewable energy systems, have begun to apply these techniques to enhance the results of their research. For instance, Bach-Andersen et al. (2017) leveraged Convolutional Neural Network (CNN) into large-scale wind turbine drivetrain monitoring. The results proved that deep learning algorithms outperformed human analysis and can provide robust fault detection on rotor bearings and planetary and helical stage gearbox bearings. Agga et al. (2022) leveraged a hybrid deep learning architecture, CNN-LSTM, to forecast short-term power production of a photovoltaic (PV) plant considering various look-back and look-forward time windows. The hybrid model gained a Mean Absolute Error (MAE) of 4.97 and outperformed conventional machine learning models and other solely deep learning models like multi-layer perception (MLP) with an MAE of 6.88.

Similarly, DL has various applications in power systems, such as online energy scheduling (Ji et al., 2021), power systems resilience improvement (Kamruzzaman et al., 2021), and adaptive power system emergency control (Huang et al., 2020). Cinar et al. (2022) employed several ML methods, like Support Vector Machine (SVM), and decision tree (DT), to optimize the biogas system considering the temperature feature. They used a lab-scale dataset and achieved an R^2 score of 0.93 by SVM. Hansen et al. (2020) used Gomptez, an ML model, and a blending of an ML-Gomptez model to predict the yield of biogas systems based on a laboratory dataset. They found that hybrid model performance is better than each single model with a Mean Absolute Precision Error (MAPE) of 4.52%, which is 53% more accurate than the Gomptez model. Similarly, Mahmoodi-Eshkaftaki and Ebrahimi (2021) used a lab dataset with DL algorithms with a feature selection method to enhance biogas system purification and find the optimal range for different reaction parameters such as biological oxygen demand/chemical oxygen demand ratio (BOD/COD), carbon/ nitrogen ratio (C/N), total solid (TS), and total volatile solids (TVS).

Despite the industrial application of ML and DL in a broad range of renewable energy systems, such as wind, power, solar energy, and other renewable energy fields, most research in biogas-related fields is focused on the experimental dataset. Considering the increasing need for renewable energy resources, more examples and case studies of how ML and DL can benefit and enhance industrial-scale biogas systems. Biogas production is a waste-based technology mainly for generating renewable energy and valorizing organic residues (Kougias and Angelidaki, 2018). The biogas systems process protects the air, water, and soil by recycling organic waste such as animal manure, food scrapes, wastewater biosolids, and organic by-products into renewable energy and soil products while decreasing greenhouse gas (GHG) emissions. According to the American Biogas Council, the U.S.A. has 2,300 sites producing biogas in 50 states, including more

than 300 farms, 1,200 water resource recovery facilities, 66 stand-alone systems that digest food scraps, and nearly 650 landfills. The U.S.A. has significant potential to build more than 15,000 new biogas industrial facilities, creating considerable economic, environmental, and energy benefits. American biogas council illustrated that building out the U.S. biogas infrastructure could produce approximately 100 trillion kilowatt-hours of electricity annually (sufficient to meet the electricity needs of 9.3 million homes), 33 trillion BTU of renewable heat per hour (4.3 million homes), or fuel for vehicles equivalent to 15.4 billion gallons per year (32 million vehicles). Mentioned biogas infrastructure would generate at least $45 billion in new capital deployment for the construction industry resulting in 375,000 short-term construction-related jobs and 25,000 permanent jobs.

Biogas is produced through anaerobic digestion (AD), where microorganisms carry out different sequences of biological reactions to degrade organic substrates and convert them to biogas. The AD process is a complex process occurring within four main steps (Adekunle and Okolie, 2015) (Figure 1.2). First, insoluble organic and higher molecular mass compounds, including but not limited to carbohydrates, fats, and proteins, are decomposed under the enzyme-mediated transformation to their smaller soluble parts. This step is called hydrolysis, and it is carried out by different types of microorganisms, such as *Bacteroides*, *Clostridia,* and facultative bacteria, such as *Streptococci,* etc. (Merlin et al., 2014). The next step is acidogenesis in which decomposed monomers are converted into organic acids such as butyric, propanoic, acetic, and alcohol (Gerardi, 2003). The next step is acetogenesis, where products that cannot be directly converted to methane by methanogenic microorganisms are converted into methanogenic substrates. In acetogenesis, volatile fatty acids (VFA) and alcohols are oxidized into intermediate compounds substrates like acetate, hydrogen, and carbon dioxide. VFAs with longer than one unit

of carbon chains are oxidized into acetate and hydrogen (Al Seadi et al., 2008). The final stage is methanogenesis, where produced intermediate compounds from the previous step are converted into biogas by methanogenic bacterial under strict anaerobic conditions (Aslanzadeh, 2014).

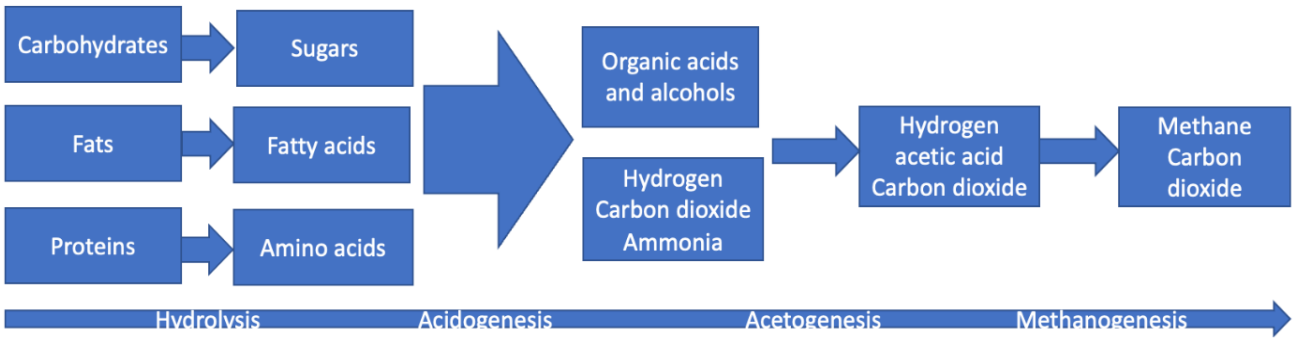

Figure 1.2. Steps of the anaerobic digestion (AD) process.

The anaerobic reactions generate biogas containing 55%-75% CH_4 (Kunatsa et al., 2022) .More than 17,400 biogas power units have been built in EU member states (Scarlat et al., 2018), reaching almost 20000 in 2022 (Bumharter et al., 2023). As a sustainable energy development in Indonesia, it has started to develop biogas units to generate electricity by converting animal manure to biogas (Khalil et al., 2019). China has recently adopted anaerobic digestion technology (Chen et al., 2017). China has 26576 biogas plants, 18000 small units, and 8576 large-scale units, generating approximately 9 billion m^3 of biomethane annually. However, the production of biogas facilities in China is impacted negatively by operational issues, such as maintaining optimum temperature and mixing ratio.

Moreover, Deng et al. (2017) found that most AD treatment plants had shut down because of the process complexity, understaffing, improper management, and reduced biogas production. Different parameters cause this reduction. For instance, maintaining suitable temperature and pH

for the microorganisms in each process step is necessary since they are sensitive to parameter changes. Obtaining optimum biogas production with reasonable cost cannot be achieved without a well-planned organic loading rate (OLR). Maintaining the OLR low can cause a reduction in the biogas production efficiency.

On the other hand, a high OLR can be a reason for process inhibition (Li et al., 2015; Sun et al., 2017). To obtain the optimum conditions for the specific biogas plant, OLR should be determined based on the feed substrate (Montingelli et al., 2015). Hydraulic retention time (HRT) determines the size of the biogas system, particularly the digester volume. Optimum biogas production can be obtained at different HRTs, depending on the used substrate (Ezekoye et al., 2011). Averagely, HRT should be between 10 and 25 days to prevent washouts of microorganisms from the digester (Schmidt et al., 2014). Also, using a combination of various substrates improves the content of nutrients, supplements, and phosphorus, and at the same time, it provides a balanced C/N ratio. The increase in the C/N ratio results in rapid nitrogen consumption before carbon digestion. Therefore, methane potential reduces (Al-Addous et al., 2017; Hills, 1979). However, the decrease in the C/N ratio results in ammonium accumulation, inhibiting the microorganisms (Al-Addous et al., 2017).

China is not the only region with issues related to biogas facility performance. Similar problems have also been found in countries such as South Korea (Kim et., 2012) and Brazil (De Clecq et al., 2017). Therefore, leveraging suitable ML and DL models is required to empower the path of reaching more sustainable biogas industrial facilities.

Due to the importance of replacing fossil fuels with renewable energy resources, they are increasingly gaining attention from governmental, industrial, and academic sectors worldwide. Hence, creating a clear technology roadmap is critical to integrate science and technology with a

business perspective in line with market movement and goals (Amer and Daim, 2010). Conducting intelligence investigations on forthcoming technologies and clear technology roadmaps will help governments and industries take smart investment steps and maintain their competitive advantage (Angelo et al., 2017). The sharp rise of employing AI in renewable energy systems has resulted in many scientific documents. Poege et al. (2019) stated that there is a high correlation and correspondence between scientific papers and patents. In other words, the quality of the patents can be determined by considering the referenced academic papers (Coupé, 2003). Therefore, analyzing the information in these textual datasets is beneficial for developing a technology roadmap. Natural language processing (NLP) is a subfield of AI and has a broad application in different domains, i.e., audio, text, video, picture, bioinformatic, etc. (Shukla and Kakkar, 2016; Gu et al., 2022; MacFarlane et al., 2022; Boorugu et al.,2020). With the powerful NLP methods in text analysis, an intellectual framework can be built to detect an informative pattern in the hidden layer of these texts over the timeline and provide an insightful perspective on renewable energy systems.

1.2. STATEMENT OF THE PROBLEM

This study aims to address three main problems in the field of renewable energy domain, namely, lack of comprehensive information management system, efficient time series-based pipeline, and data limitation, with a particular focus on biogas system leveraging AI-based methods.

1) Unstructured textual datasets are created in different forms of documents and are available in different clouds and online databases. The significant growth of the application of AI modeling in the renewable energy domain creates a massive amount of data and information within research papers, registered patents, reports, etc. A problem

in this domain is that investors and policymakers in governmental and industrial sectors can become uncertain about and among various generated information, specifically in AI with various techniques and approaches. Hence, analyzing these documents is necessary to develop clear and feasible technology road mapping. However, considering the tremendous amount of generated data daily at the intersection of computer science and renewable energy systems, manual analysis is time-consuming, labor-intensive, sometimes can be erroneous, and biased in final results. Humans may not consider some information deliberately or evaluate things differently, considering our knowledge and understanding. NLP techniques become valuable in providing insightful patterns and information from an extensive textual dataset with significant coherence for humans.

2) Biogas systems are complicated, and their power generation needs a significant reaction time (Chiu et al., 2022). However, researchers in this domain have begun to leverage AI models to analyze the input and process parameters and develop predictive models to forecast the performance of biogas systems. One of the missing points in the research conducted so far is the analysis of key input and intermediate factors, resulting in biogas or methane yield instability (Chiu et al., 2022). Considering the dimension of datasets in this field, especially industrial datasets, and the high degree of non-linearity and correlation between their features, finding the effect of key factors and forecasting the system's performance require complex calculations. Conventional mathematical and statistical modeling usually makes some errors and requires significant time. Hence, developing ML/DL algorithms that can analyze critical factors in industrial biogas reaction targeting optimizing output can result in developing more accurate TSF predictive models. Powerful AI techniques can be employed to enhance decision-making

in the operational sector of biogas plants and make them more economically sustainable. This will also positively impact the popularity of biogas plants in industry and make them one of the most privileged environmental-friendly approaches to generating electricity or heat worldwide.

3) As mentioned erlier there is a high level of uncertainty in biogas systems and ML models can play a significant role in predicting the system's yield. However, experimental and lab-scale data limitation in this domain considering measuring parameters can be challenging. Consequently, supervisded ML models usually cannot produce significant and generalized results for lab-scale biogas problems, since they require sufficient amount data points to be trained on to learn and capture the pattern of the data, enabaling them to make accurate predictions. An effective way to overcome this limitation is leveraging data augmatation techniques. More specifically, powerful AI-based techniques such as Variational Auto Encoder (VAE) can generate dummy datapoints that have the similar statistical charactrestics to the original data. Design engineers attempting to forecast the experimental methane yield during the initial phases of biogas project development can benefit from this solution.

1.3. RESEARCH OBJECTIVES

The research set the following objectives to help solve the research problems,

1) Developing a model to detect the trend of AI application in renewable energy systems and identify the most frequent computational algorithms in related domains from 2000 until 2021 from published English books and papers in Elsevier's Scopus database.

2) Leveraging Dynamic Topic Modeling (DTM) and NLP algorithms to characterize and map the recent AI applications landscape in renewable energy systems and investigate the evolution of this field over the considered 20-year timeframe. DTM and NLP will be used to analyze scientific publications (from Elsevier's Scopus) focusing mainly on AI applications in renewable energy systems.

3) Analyzing key input and intermediate factors of the industrial biogas plants by taking a case study of an industrial biogas plant in Shenzhen, China, over 440 days. A feature selection by ensemble learning methods, i.e., XGboost and random forest, will be conducted to determine the impact of key factors on the biogas system's production, aiming at optimizing output, resulting in the development of more accurate TSF models.

4) Developing hybrid DL-based TSF models such as LSTM, CNN-LSTM, and GRU for the industrial dataset to predict the yield of the biogas system. This method provides robust results without intensive feature engineering and can help other researchers in academia and industry develop more sustainable biogas systems.

5) Developing a hybrid approach to overcome data limitation in biogas domain via leveraging vertical autoencoder model and generating augmented data following the similar statectical charactrestics of real data.

6) Predicting bio methene potentional (BMP) using machine learning models.

1.4. STRUCTURE OF THE THESIS

This thesis consists of six chapters. Chapter 1 outlines the general research background, motivation, problem statement, research objectives, and thesis structure. Chapter 2 provides the literature reviews of the relevant topics, including (1) current widely used ML/DL algorithms in supervised learning and unsupervised learning and the challenges and limitations they face, (2) related research and challenges of biogas, (3) potential application of ML/DL on the challenges in biogas and bio hydrogen domains, (4) NLP methods with a focus on text mining domain, (5) image processing applications in biogas domain. Chapter 3 presents the development of DTM and NLP techniques, including a transfer learning model and feature extraction method, to intellectually investigate the AI application in renewable energy systems. Chapter 4 is a study to leverage AI in a biogas industrial plant to investigate the importance feedstock and did TSF of the systems' output via a hybrid DL model. Chapter 5 is a study which applied DL-based data augmatation method to address the data limitation facing ML modeling of biogas datasets. Specifically, it tried using ML to predict the biogas yield of various feedstocks using their organic components. Finally, Chapter 6 presents the conclusions of this research and recommendations for future work. The structure of the thesis is depicted in Figure 1.3.

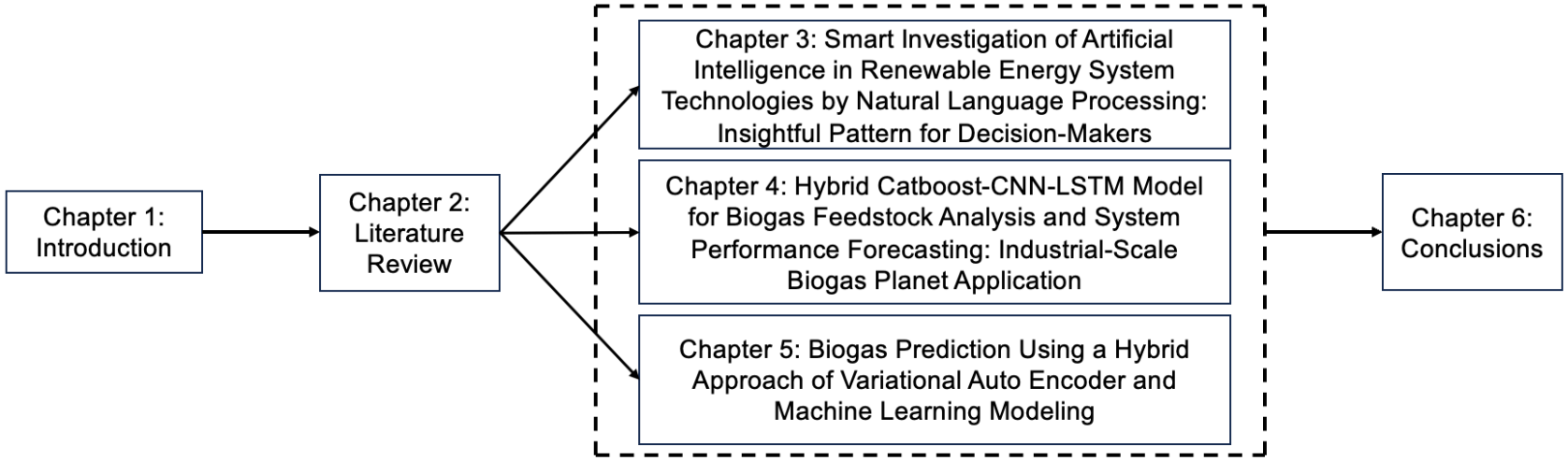

Figure 1.3. The research roadmap

1.5. CONTRIBUTION OF THIS THESIS

This research generated two journal papers that are under review:

Kamran Niroomand, Noori M. Cata Saady, Carlos Bazan,; Sohrab Zendehboudi, Amilcar Soares, Talib M. Albayati (2023) Smart Investigation of Artificial Intelligence in Renewable Energy System Technologies by Natural Language Processing: Insightful Pattern for Decision-Makers. *International Scientific Journal Engineering Applications of Artificial Intelligence.* Reference number: EAAI-22-3885R1. I developed the proposed approach, built the case study model, analyzed results, and wrote the draft of the paper.

Kamran Niroomand, Noori M. Cata Saady, Carlos Bazan, Sohrab Zendehboudi (2023) Hybrid Catboost-CNN-LSTM Model for Biogas Feedstock Analysis and System Performance Forecasting: Industrial-Scale Biogas Planet Application. *Journal of Energy for Sustainable Development.* Reference number: ESD-S-23-01172 (under review). I developed the proposed approach, built the case study model, analyzed results, and wrote the draft of the manuscript.

REFRENCES

Bumharter, C., Bolonio, D., Amez, I., Martínez, M. J. G., & Ortega, M. F. (2023). New opportunities for the European Biogas industry: A review on current installation development, production potentials and yield improvements for manure and agricultural waste mixtures. *Journal of Cleaner Production*, 135867.

Kunatsa, T., & Xia, X. (2022). A review on anaerobic digestion with focus on the role of biomass co-digestion, modelling and optimisation on biogas production and enhancement. *Bioresource technology*, *344*, 126311.

Jha, S. Kr., Bilalovic, J., Jha, A., Patel, N., & Zhang, H. (2017). Renewable energy: Present research and future scope of Artificial Intelligence. In Renewable and Sustainable Energy Reviews (Vol. 77, pp. 297–317). Elsevier BV. https://doi.org/10.1016/j.rser.2017.04.018

Boden, M. A. (Ed.). (1996). Artificial intelligence. Elsevier.

Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In Machine learning techniques for multimedia (pp. 21-49). Springer, Berlin, Heidelberg.

Miorelli, R., Kulakovskyi, A., Chapuis, B., D'almeida, O., & Mesnil, O. (2021). Supervised learning strategy for classification and regression tasks applied to aeronautical structural health monitoring problems. Ultrasonics, 113, 106372.

McAlpine, E. D., Michelow, P., & Celik, T. (2022). The utility of unsupervised machine learning in anatomic pathology. *American Journal of Clinical Pathology*, *157*(1), 5-14

Ashfaq, R. A. R., Wang, X.-Z., Huang, J. Z., Abbas, H., & He, Y.-L. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. In Information Sciences (Vol. 378, pp. 484–497). Elsevier BV. https://doi.org/10.1016/j.ins.2016.04.019

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. In Applied Soft Computing (Vol. 90, p. 106181). Elsevier BV. https://doi.org/10.1016/j.asoc.2020.106181

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. In Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering

Sciences (Vol. 379, Issue 2194, p. 20200209). The Royal Society. https://doi.org/10.1098/rsta.2020.0209

Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2019). A review of deep learning for renewable energy forecasting. In Energy Conversion and Management (Vol. 198, p. 111799). Elsevier BV. https://doi.org/10.1016/j.enconman.2019.111799

Hu, Y.-L., & Chen, L. (2018). A nonlinear hybrid wind speed forecasting model using LSTM network, hysteretic ELM and Differential Evolution algorithm. In Energy Conversion and Management (Vol. 173, pp. 123–142). Elsevier BV. https://doi.org/10.1016/j.enconman.2018.07.070

Yu, C., Li, Y., Bao, Y., Tang, H., & Zhai, G. (2018). A novel framework for wind speed prediction based on recurrent neural networks and support vector machine. In Energy Conversion and Management (Vol. 178, pp. 137–145). Elsevier BV. https://doi.org/10.1016/j.enconman.2018.10.008

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1412.3555

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. In Neural Computation (Vol. 9, Issue 8, pp. 1735–1780). MIT Press - Journals. https://doi.org/10.1162/neco.1997.9.8.1735

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Proceedings of SSST-8,

Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Association for Computational Linguistics. https://doi.org/10.3115/v1/w14-4012

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. In IEEE Transactions on Signal Processing (Vol. 45, Issue 11, pp. 2673–2681). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/78.650093

Zhao, W., Han, S., Hu, R. Q., Meng, W., & Jia, Z. (2018). Crowdsourcing and Multisource Fusion-Based Fingerprint Sensing in Smartphone Localization. In IEEE Sensors Journal (Vol. 18, Issue 8, pp. 3236–3247). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jsen.2018.2805335

Wang, J., Yan, J., Li, C., Gao, R. X., & Zhao, R. (2019). Deep heterogeneous GRU model for predictive analytics in smart manufacturing: Application to tool wear prediction. In Computers in Industry (Vol. 111, pp. 1–14). Elsevier BV. https://doi.org/10.1016/j.compind.2019.06.001

Zhang, N., Shen, S.-L., Zhou, A., & Jin, Y.-F. (2021). Application of LSTM approach for modelling stress–strain behaviour of soil. In Applied Soft Computing (Vol. 100, p. 106959). Elsevier BV. https://doi.org/10.1016/j.asoc.2020.106959

Miao, K., Han, T., Yao, Y., Lu, H., Chen, P., Wang, B., & Zhang, J. (2020). Application of LSTM for short term fog forecasting based on meteorological elements. In Neurocomputing (Vol. 408, pp. 285–291). Elsevier BV. https://doi.org/10.1016/j.neucom.2019.12.129

Sethia, A., & Raut, P. (2018). Application of LSTM, GRU and ICA for Stock Price Prediction. In Information and Communication Technology for Intelligent Systems (pp. 479–487). Springer Singapore. https://doi.org/10.1007/978-981-13-1747-7_46

Bach-Andersen, M., Rømer-Odgaard, B., & Winther, O. (2017). Deep learning for automated drivetrain fault detection. In Wind Energy (Vol. 21, Issue 1, pp. 29–41). Wiley. https://doi.org/10.1002/we.2142

Agga, A., Abbou, A., Labbadi, M., Houm, Y. E., & Ou Ali, I. H. (2022). CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production. In Electric Power Systems Research (Vol. 208, p. 107908). Elsevier BV. https://doi.org/10.1016/j.epsr.2022.107908

Cinar, S. Ö., Cinar, S., & Kuchta, K. (2022). Machine Learning Algorithms for Temperature Management in the Anaerobic Digestion Process. In Fermentation (Vol. 8, Issue 2, p. 65). MDPI AG. https://doi.org/10.3390/fermentation8020065

Hansen, B. D., Tamouk, J., Tidmarsh, C. A., Johansen, R., Moeslund, T. B., & Jensen, D. G. (2020). Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model. In Computational Science and Its Applications – ICCSA 2020 (pp. 734–745). Springer International Publishing. https://doi.org/10.1007/978-3-030-58799-4_53

Mahmoodi-Eshkaftaki, M., & Ebrahimi, R. (2021). Integrated deep learning neural network and desirability analysis in biogas plants: A powerful tool to optimize biogas purification. In Energy (Vol. 231, p. 121073). Elsevier BV. https://doi.org/10.1016/j.energy.2021.121073

Kougias, P. G., & Angelidaki, I. (2018). Biogas and its opportunities—A review. In Frontiers of Environmental Science & Engineering (Vol. 12, Issue 3). Springer Science and Business Media LLC. https://doi.org/10.1007/s11783-018-1037-8

Scarlat, N., Dallemand, J.-F., & Fahl, F. (2018). Biogas: Developments and perspectives in Europe. In Renewable Energy (Vol. 129, pp. 457–472). Elsevier BV. https://doi.org/10.1016/j.renene.2018.03.006

Khalil, M., Berawi, M. A., Heryanto, R., & Rizalie, A. (2019). Waste to energy technology: The potential of sustainable biogas production from animal waste in Indonesia. In Renewable and Sustainable Energy Reviews (Vol. 105, pp. 323–331). Elsevier BV. https://doi.org/10.1016/j.rser.2019.02.011

Chen, L., Cong, R.-G., Shu, B., & Mi, Z.-F. (2017). A sustainable biogas model in China: The case study of Beijing Deqingyuan biogas project. In Renewable and Sustainable Energy Reviews (Vol. 78, pp. 773–779). Elsevier BV. https://doi.org/10.1016/j.rser.2017.05.027

Scarlat, N., Dallemand, J.-F., & Fahl, F. (2018). Biogas: Developments and perspectives in Europe. In Renewable Energy (Vol. 129, pp. 457–472). Elsevier BV. https://doi.org/10.1016/j.renene.2018.03.006

Deng, L., Liu, Y., Zheng, D., Wang, L., Pu, X., Song, L., Wang, Z., Lei, Y., Chen, Z., & Long, Y. (2017). Application and development of biogas technology for the treatment of waste in China. In Renewable and Sustainable Energy Reviews (Vol. 70, pp. 845–851). Elsevier BV. https://doi.org/10.1016/j.rser.2016.11.265

Shukla, H., & Kakkar, M. (2016). Keyword extraction from Educational Video transcripts using NLP techniques. In 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). IEEE. https://doi.org/10.1109/confluence.2016.7508096

Gu, W., Yang, X., Yang, M., Han, K., Pan, W., & Zhu, Z. (2022). MarkerGenie: an NLP-enabled text-mining system for biomedical entity relation extraction. In C. Arighi (Ed.), Bioinformatics Advances (Vol. 2, Issue 1). Oxford University Press (OUP). https://doi.org/10.1093/bioadv/vbac035

MacFarlane, H., Salem, A. C., Chen, L., Asgari, M., & Fombonne, E. (2022). Combining voice and language features improves automated autism detection. In Autism Research (Vol. 15, Issue 7, pp. 1288–1300). Wiley. https://doi.org/10.1002/aur.2733

Boorugu, R., & Ramesh, G. (2020). A Survey on NLP based Text Summarization for Summarizing Product Reviews. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE. https://doi.org/10.1109/icirca48905.2020.9183355

Adekunle, K. F., & Okolie, J. A. (2015). A Review of Biochemical Process of Anaerobic Digestion. In Advances in Bioscience and Biotechnology (Vol. 06, Issue 03, pp. 205–212). Scientific Research Publishing, Inc. https://doi.org/10.4236/abb.2015.63020

Merlin Christy, P., Gopinath, L. R., & Divya, D. (2014). A review on anaerobic decomposition and enhancement of biogas production through enzymes and microorganisms. In Renewable and Sustainable Energy Reviews (Vol. 34, pp. 167–173). Elsevier BV. https://doi.org/10.1016/j.rser.2014.03.010

Gerardi, M. H. (2003). The microbiology of anaerobic digesters. John Wiley & Sons.

Al Seadi, T., Ruiz, D., Prassl, H., Kottner, M., Finsterwaldes, T., Volke, S. and Janssers, R. (2008) Handbook of Biogas. University of Southern Denmark, Esbjerg.

Aslanzadeh, S. (2014). Pretreatment of Cellulosic Waste and High Rate Biogas Production. Doctoral Thesis on Resource Recovery, University of Borås, Borås, 1-50.

Li, D., Liu, S., Mi, L., Li, Z., Yuan, Y., Yan, Z., & Liu, X. (2015). Effects of feedstock ratio and organic loading rate on the anaerobic mesophilic co-digestion of rice straw and pig manure. *Bioresource technology*, *187*, 120-127.

Sun, M. T., Fan, X. L., Zhao, X. X., Fu, S. F., He, S., Manasa, M. R. K., & Guo, R. B. (2017). Effects of organic loading rate on biogas production from macroalgae: Performance and microbial community structure. *Bioresource technology*, *235*, 292-300.

Montingelli, M. E., Tedesco, S., & Olabi, A. G. (2015). Biogas production from algal biomass: A review. *Renewable and Sustainable Energy Reviews*, *43*, 961-972.

Ezekoye, V. A., Ezekoye, B. A., & Offor, P. O. (2011). Effect of retention time on biogas production from poultry droppings and cassava peels. *Nigerian Journal of Biotechnology*, *22*, 53-59.

Schmidt, T., Ziganshin, A. M., Nikolausz, M., Scholwin, F., Nelles, M., Kleinsteuber, S., & Pröter, J. (2014). Effects of the reduction of the hydraulic retention time to 1.5 days at constant organic loading in CSTR, ASBR, and fixed-bed reactors–performance and methanogenic community composition. *biomass and bioenergy*, *69*, 241-248.

Al-Addous, M., Alnaief, M., Class, C., Nsair, A., Kuchta, K., & Alkasrawi, M. (2017). Technical possibilities of biogas production from olive and date waste in Jordan. *BioResources*, *12*(4), 9383-9395.

Hills, D. J. (1979). Effects of carbon: nitrogen ratio on anaerobic digestion of dairy manure. Agricultural wastes, 1(4), 267-278.

Chiu, M.-C., Wen, C.-Y., Hsu, H.-W., & Wang, W.-C. (2022). Key wastes selection and prediction improvement for biogas production through hybrid machine learning methods. In Sustainable Energy Technologies and Assessments (Vol. 52, p. 102223). Elsevier BV. https://doi.org/10.1016/j.seta.2022.102223

CHAPTER TWO

LITERATURE REVIEW

2.1. BASIC PRINCIPLES OF ANAEROBIC DIGESTION

Anaerobic digestion is a widely used biological process for treating industrial organic waste and wastewater (Jadhav et al., 2019). It involves the degradation of organic materials in the absence of oxygen to produce biogas, which consists of 60% methane, 40% carbon dioxide, and other gases. This process is also used for generating bioenergy from agricultural residue and animal manure (Jadhav et al., 2019). Removing carbon dioxide from biogas produces biomethane, which can be used as a fuel for heating or transportation. The process is carried out by microorganisms that convert organic matter into biogas through four steps: hydrolysis, acidogenesis, acetogenesis, and methanogenesis. The importance of these microorganisms in the degradation of organic materials and the production of methane/carbon dioxide has been well-established in the literature (Yuan and Zhu, 2016; Sonakya et al., 2001).

2.1.1. Hydrolysis

Anaerobic digestion process starts with hydrolysis, the critical step of converting complex polymer organic materials such as carbohydrates, lipids, and proteins into simpler monomers and substances (Menzel et al., 2020). Hydrolytic bacteria release extracellular enzymes that catalyze the parallel degradation processes of carbohydrates into simple sugars, proteins into amino acids, and lipids into long-chain fatty acids (Shrestha et al., 2017). Hydrolysis is considered the rate-limiting step when the substrate is complex, for instance, lignocellulosic biomass. Thus, the step significantly influences the rate of the biogas production process. Enzymes, including cellulase,

amylase, protease, and lipase, play the primary role in hydrolysis, breaking down the complex organic matter into smaller units.

2.1.2. Acidogenesis

During the hydrolysis step of the anaerobic digestion process, some of the large molecules are converted into hydrogen and acetate that methanogens can directly use to produce methane. However, most of the hydrolysis products remain relatively large and require further conversion into smaller molecules, such as acetic acid. Acidogenesis is the next step, where acidogenic bacteria transform the hydrolysis products into usable forms for methanogens. Simple sugars, amino acids, and fatty acids undergo degradation to produce acetate, carbon dioxide, and hydrogen. This step also generates short-chain volatile fatty acids (VFAs) and alcohols. The acidogenesis reactions are provided in Table 2.1.

2.1.3. Acetogenesis

After acidogenesis, the next step in the anaerobic digestion process is acetogenesis, where acetogenic bacteria convert the acidic products of the previous step into acetic acid, hydrogen, and carbon dioxide, which methanogens can utilize for methane production. The reactions that occur during acetogenesis are outlined in Table 2.1.

2.1.4. Methanogenesis

The last stage in the anaerobic digestion process is methanogenesis, where methane (CH_4) is produced from volatile fatty acids (VFAs) or hydrogen and carbon dioxide directly. Methane is a desirable end product in AD, as it is rich in electrons, making it an excellent energy source, and is not very soluble in water. Once it is produced, it leaves the liquid phase as a gas and does not have any further impact on the microbial community. The reactions that take place during methanogenesis are given in Table 2.1.

Table 2.I. The anaerobic digestion's biochemical reactions (Hajizadeh, 2021)

Step	Biochemical reaction
Acidogenesis	$C_6H_6O_6 + 2H_2 \rightarrow 2CH_3CH_2COOH + 2H_2O$ $C_6H_{12}O_6 \rightarrow 2CH_3CH_2OH + 2CO_2$
Acetogenesis	$CH_3CH_2COO^- + 3H_2O \rightarrow CH_3COO^- + H^+ + HCO_3^- + 3H_2$ $C_6H_{12}O_6 + 2H_2O \rightarrow 2CH_3COOH + 2CO_2 + 4H_2$ $CH_3CH_2OH + 2H_2O \rightarrow CH_3COO^- + 2H_2 + H^+$ $2HCO_3^- + 4H_2 + H^+ \rightarrow CH_3COO^- + 4H_2O$
Methanogenesis	$2CH_3CH_2OH + CO_2 \rightarrow 2CH_3COOH + CH_4$ $CH_3COOH \rightarrow CH_4 + CO_2$ $CH_3OH \rightarrow CH_4 + H_2O$ $CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$ $CH_3COO^- + SO_4^{2-} + H^+ \rightarrow 2HCO_3^- + H_2S$ $CH_3COO^- + NO^- + H_2O + H^+ \rightarrow 2HCO_3^- + NH_4^+$

2.2. BIOGAS PRODUCTION

The generation of biogas through anaerobic digestion (AD) offers a viable means of producing energy. AD is particularly popular among livestock farmers as it allows them to offset their electricity bills and reduce their carbon footprint. The end products of AD are biogas and digestate. Biogas is utilized to produce electricity and heat, while digestate is commonly used as a fertilizer, soil amendment, and/or bedding source. Various organic wastes can be utilized in AD for biogas production, including lignocellulosic wastes from agricultural and municipal activities, animal manure and slurry, sewage sludge and municipal solid waste, and food waste. Table 2.2 presents the typical energy production quantities from different waste sources.

Table 2.2 Biogas production from various substrates.

Substrate	Biogas production per ton (m^3 CH_4 kg^{-1} VS)	Ref.
Municipal and industrial		
Mechanically recovered organic fraction of municipal solid waste	0.344	Zhang and Banks (2010)
Food waste	0.461	Lu et al. 2017

Table 2.2 Continued.

Substrate	Biogas production per ton (m^3 CH_4 kg^{-1} VS)	Ref.
Banana peel	0.227	Zheng et al. (2013)
Cassava peels	0.205-0.25	Ghosh et al. (2020)
Sugarcane bagasse	0.075	Nwokolo et al. (2021)
Maize silage	0.2-0.25	Okonkwo et al. (2018)
Sewage sludge	0.2-0.4	Selormey et al. (2022)
Cheese whey	0.35-0.45	Bumbiere et al. (2020)
Agricultural waste		
Cattle manure	0.30-0.51	Lu et al. (2017)
Cow manure	0.15-0.30	Angelidaki and Ellegaard (2003)
Poultry manure	0.33	Johannesson et al. (2020)
Chicken manure	0.05-0.12	Nwokolo et al. (2021)
Pig manure	0.02-0.05	Bumbiere et al. (2020)
Grass silage	0.2-0.4	Ahmed & Kazda (2017)
Animal and slaughterhouse waste		
Animal waste	0.38	Selormey et al. (2022)
Rumen content	0.35	Selormey et al. (2022)
Stomach and gut contents	0.40-0.46	Limeneh et al. (2022)
Fish waste	0.45-0.60	Ijoma et al. (2021)

2.3. OVERVIEW OF DESIGN AND OPERATION PARAMETERS OF BIOGAS PLANTS

An anaerobic digester's successful design and operation depend on various parameters that should be considered. These parameters include the total solids content, volatile solids content (which refers to the organic matter amount), the Carbon-to-Nitrogen (C/N) ratio of the substrate, pH and alkalinity, temperature, hydraulic retention time, organic loading rate, and inoculum-to-

substrate ratio. Proper management of these parameters is crucial for optimizing biogas production and ensuring a stable operation of the digester.

2.3.1. pH and alkalinity

The pH level is an important factor affecting the performance of microorganisms in the AD process. Methanogens are particularly sensitive to low pH levels, while high pH levels can form toxic substances such as free ammonia. The pH inside the digester changes continuously throughout the AD process, and each group of microorganisms has an optimal pH range for maximum reaction rates (Liu et al., 2008). Hydrolytic bacteria and acidogens can operate within a wide pH range of 4-8.5, while methanogens require a narrow pH range of 6.5 to 7.2. Alkalinity is another crucial parameter in AD, as it determines the medium's resistance to pH changes (Zhai et al., 2015). Alkalinity is the equilibrium of CO_2 and bicarbonate ions, and it is more reliable than direct pH measurement in assessing digester imbalance. VFAs produced during acidogenesis can decrease the pH, but methanogens produce alkalinity as CO_2 and bicarbonate to neutralize this reduction. The concentration of CO_2 in the gas phase and bicarbonate in the liquid phase determines the pH value inside the digester. Low alkalinity can be addressed by reducing the OLR, adding salt to convert CO_2 to bicarbonate, or adding bicarbonate directly. Generally, the recommended alkalinity level for optimum methane production is between 1000 and 5000 mg $CaCO_3$/L (Issah et al., 2020).

2.3.2. Temperature

Temperature is a crucial factor that significantly impacts various parameters in the AD process. It influences the growth rate of microorganisms, diversity, thermodynamic equilibrium, stability, process kinetics, and methane yield. The minimum and maximum temperature range suitable for operating an anaerobic digester is between 20 ºC and 60 ºC. Based on the operation

temperature, the AD process is categorized as psychrophilic (20 ºC), mesophilic (35 ºC), and thermophilic (60 ºC) (Angelidaki et al., 2005). Temperature-phased AD is another configuration that utilizes the benefits of each temperature range.

2.3.3. Carbon to Nitrogen (C/N) ratio

Co-digestion is adding multiple substrates to the digester to adjust the C/N ratio and enhance the performance of the AD process. Co-digestion has several advantages such as balancing the nutrients, increasing the methane yield, reducing the hydraulic retention time (HRT), and enhancing the stability of the process. Co-digestion can also improve waste management by utilizing various organic waste streams unsuitable for single digestion (Rajendran et al., 2012). However, carefully selecting the co-substrates is necessary to avoid any potential inhibition or toxicity to the microorganisms in the digester (Wijesinghe et al., 2019).

2.3.4. Total solids content

The substrate's total solids (TS) content indicates the amount of moisture in the AD process. Moisture plays a crucial role in the AD process as it helps in the diffusion of soluble substrates and nutrients into microbial cells. There are two types of AD processes based on the total solids content: dry AD with 15-40% total solids and wet AD with 10-15% total solids (Orhorhoro et al., 2017). Dry AD has some advantages such as smaller reactor volume, less energy and water consumption, and fewer moving parts, and more than 60% of installed AD in Europe in 2005 were dry AD. However, in terms of specific methane production and process kinetics, wet AD is more efficient than dry AD.

2.3.5. Hydraulic retention time

The digester's hydraulic retention time (HRT) is the average duration taken by a water particle to travel between the digester's inlet and outlet, while the solid retention time (SRT) is the

average duration that microorganisms spend in the digester. Retention time is a vital parameter in the AD process because it directly influences the number of microorganisms (Qyyum et al., 2020). Methanogens, for example, double every 2-4 days. To have an effective SRT, process kinetics, substrate type, temperature, and OLR must be optimized (Westerholm et al., 2012). Low HRT raises the risk of biomass washout from the reactor, which may negatively impact the whole process. Low SRT, on the other hand, causes VFA accumulation and increases alkalinity, in addition to the issues associated with low HRT (Ioannou-Ttofa et al., 2021).

2.4. BASIC PRINCIPLES OF MACHINE LEARNING

ML uses data and analyzes them based on computational algorithms, producing understandable results for human beings. ML is one of AI's principal subfields and has grown since the early 1990s (Carleo et al., 2019). The short but accurate definition of ML is that it is one of the AI branches, a powerful tool to find insightful patterns in big data with high dimensionality (Biamonte et al., 2017). ML models learn like humans and aim to decrease human labor works. One way to categorize datasets for ML algorithms is based on their labels; labeled and unlabelled datasets. Labeled datasets are those with each instance accompanied by a target value (Wang et al., 2021), while unlabelled datasets only include features (Vinothkumar et al., 2022). In data science, two ML types are mainly applied, supervised and unsupervised learning (Chen, 2022). Supervised learning is one of the primary ML categories used in broad domains, from fake news detection (Alsubari et al., 2022) to laser machining (Behbahani et al., 2022). A part of the whole labeled dataset is split as a training part, and supervised learning algorithms are built upon that (Nasteski, 2017). In this regard, the model is trained firstly based on each target value and its corresponding features, and it gains informative knowledge about the data, such as the correlation between provided features. Eventually, the model can build a connection between different

variables within the labeled dataset (Zuranski et al., 2021). Other supervised learning models are widely leveraged in different such as linear regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and logistic regression. Linear regression tries to fit data by a linear equation with coefficients w = (w1,…, wp) to minimize the sum of squares of the distance between predicted and real values (Bourguignon and de Medeiros, 2022). Figure 2.1 illustrates how a simple regression model works and its decision boundary.

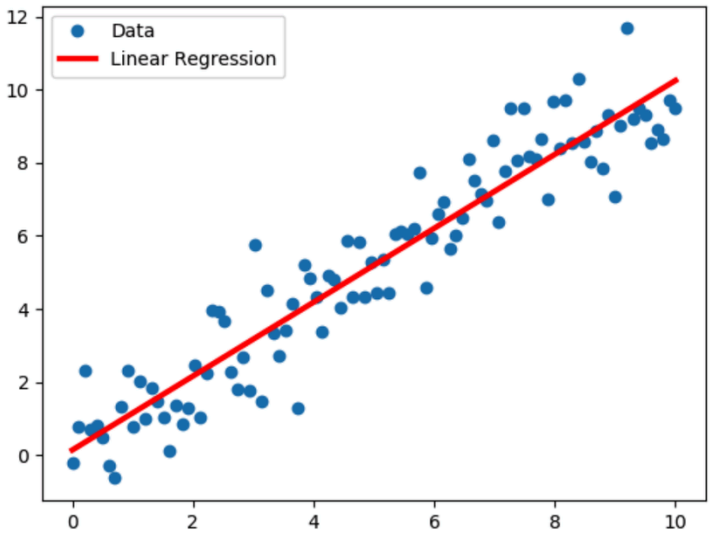

Figure 2.1. Linear regression decision boundary (Zuranski et al., 2021).

KNN is another supervised learning based on proximity (Yamac, 2021). KNN algorithm considers the distance (the method for distance measurement can be modified as a hyperparameter) of the quarry and all other instances based on determined local neighbors (a hyperparameter, K) (Mishra et al., 2021). Afterward, for the classification task, the model votes for the major class or average of the target value in terms of the regression task. KNN has a wide variety of applications in solving real-life data-driven projects, such as soil parameters estimation (Garg et al., 2020), adaptive thermal comforts (Xiong et al., 2021), dental fluorosis in groundwater prediction (Ataş et al., 2022), etc. Figure 2.2 shows how KNN can handle ML tasks.

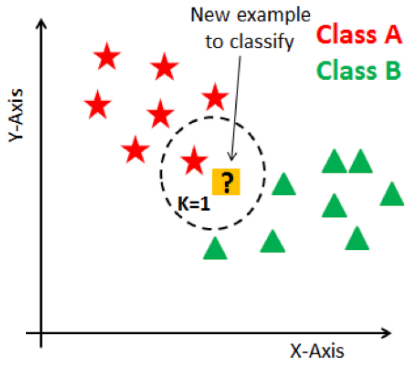

Figure 2.2. KNN algorithm functionality.

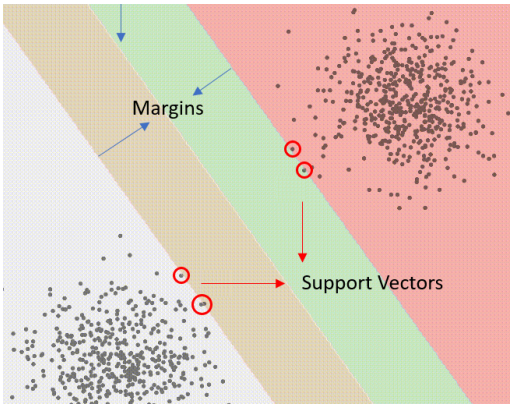

Figure 2.3. SVM algorithm functionality.

In the SVM algorithm, each instance is illustrated in an *n*-dimensional space, where *n* denotes the number of features. Following that, a decision boundary is built to distinguish classes (Noble, 2006). The constructed boundary, called the margin, attempts to increase the support vector's distance. If support vectors are in the margin zone, it can be considered noise and will subsequently reduce the model's generalization (Suthaharan, 2016). This model can be used for regression and classification tasks but is mainly considered a powerful classifier. It is especially useful in those datasets that have more features comparing instances. SVM significantly saves memory since the algorithm builds its decision boundary by considering a part of the training dataset. However, it negatively affects the computational efficiency of big datasets since the training time increases (Tanveer et al., 2022). Figure 2.3 illustrates how maximizing the margin between support vectors will build SVM's decision boundary.

Logistic regression is one of the main algorithms for binary classification tasks. This model is built upon a logistic function (equation 2.1), a sigmoid function, and log odds. As seen from Eq. 2.1, this model takes integer values and generates outputs ranging between 0 and 1 (Huang, 2022).

$$y = 1/1 + e^{-x} \tag{2.1}$$

where x and y denote the input and output values, respectively.

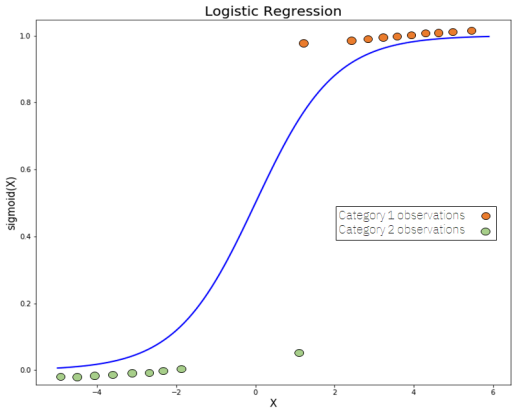

Figure 2.4. Logistic regression algorithm functionality (Verma et al., 2022).

Despite supervised methods, unsupervised learning algorithms do not need human forces to build the machine-readable dataset (labeling the dataset). Labels in the dataset help the machine to understand the connection and the mathematical relationships between each two data points (Alsubari et al., 2022). Unsupervised learning algorithms build structures based on hidden information from a dataset without human interference and will categorize similar data points based on built structures. Another characteristic of unsupervised learning models is that they can dynamically adapt to the data points by modifying built structures (Lee et al., 2022). This capability enables them to provide better deployment development comparing the supervised-based algorithms (Verma et al., 2022). Unsupervised learning has different tasks, such as clustering, association rules, and dimensionality reduction. Clustering is a data mining method that categorizes unlabeled data based on their similarities or differences. Clustering models are applied to process raw, unclassified data objects into groups based on structures or patterns in the information. Clustering models can be categorized into a few groups: exclusive, overlapping, hierarchical, and probabilistic. An association rule is a rule-based method that is used to find relationships between features in a given dataset. These methods are widely employed for market

basket analysis, enabling businesses to better perceive relationships between different products. Dimensionality reduction is a learning process used when the dataset has many features or dimensions. It decreases the number of data inputs to a reasonable size while preserving informative data. Often, this method is applied in the cleaning data section, such as when autoencoders remove noise from visual data to improve picture quality. Based on neural networks, Autoencoders compress data and then rebuild a new representation of the primary data points. As can be seen in Figure 2.5, the hidden layer is responsible for compressing the input layer before reconstructing it in the output layer. The step from the input layer to the hidden layer is called "encoding" while the step from the hidden layer is "decoding."

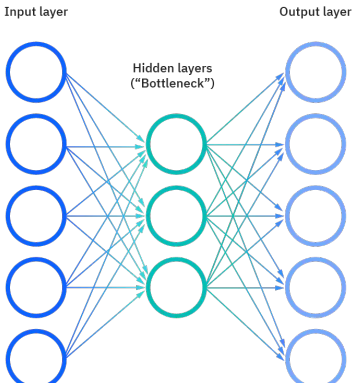

Figure 2.5. Autoencoders algorithm functionality (Alsubari et al., 2022).

Unsupervised learning models have a wide range of applications, such as anomaly detection, computer vision, customer personas, recommendation engines, medical imaging, etc. The main negative points of unsupervised learning models are that they can have wildly inaccurate results and require human intervention to validate the output variables (Su et al., 2022).

2.4.1. Data preprocessing

Data preprocessing is an important step in the data mining and analysis process that takes raw data and transforms it into a suitable format that machines can analyze. Raw, real-world data in various text types, images, videos, etc., are messy. Not only they may have several errors and inconsistencies, but they are often incomplete and unstructured. If the data do not have good quality, the model cannot provide significant results. Kubik et al. (2022) indicated the importance of data quality evaluation before machine learning deployment. There are four metrics to evaluate the data quality, completeness, validity, timeliness, and consistency (Jain et al., 2020). Figure 2.6 illustrates the four sections of data preprocessing: data reduction, data integration, data transformation, and data cleaning (Hameed and Naumann (2020)).

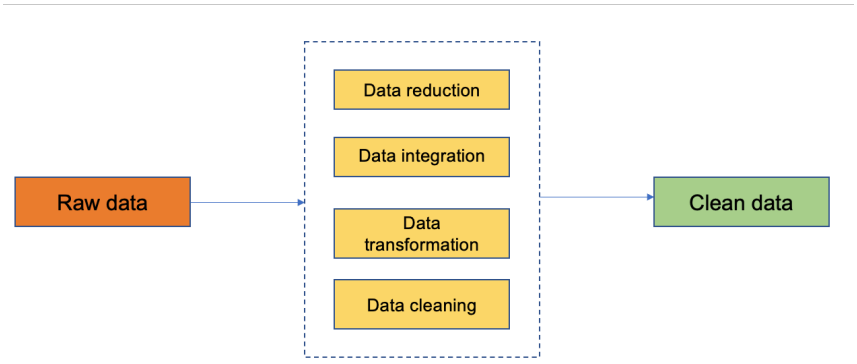

Figure 2.6. Significant steps of data preprocessing (Hameed and Naumann (2020)).

Data reduction cuts down the data size and the required storage space, and results in an simpler analysis while producing accurate results. Data reduction has different approaches, namely, numerosity reduction, dimensionality reduction, and data compression (Patel et al., 2015). Dimensionality reduction decreases the required storage space and computation time. Numerosity reduction puts smaller form of data representation instead of original data, without data loss,

aiming to reduce the data amount. Data compression compresses the data which by either encoding, reconstructing, or modifying data (Salomon and Motta, (2010)).

Data transformation process changes the format or the data structure and can be simple or complex considering the requirements and task. There are different methods in data transformation, namely, smoothing, aggregation, discretization, and normalization. Smoothing removes the noise from the dataset, and important features of the dataset can be identified (Gautam et al., 2015). Aggregation summarizes the data; it is an important section since the data quantity and quality play a key role in the results accuracy and significance (Korableva et al., 2018). Discretization process splits the continuous data into intervals. For instance, instead of indicating the class time, we can set an interval like (1-3 pm, 4-6 pm). Normalization is one the most famous method of scaling the data where data can be represented in a smaller range of -1.0 to 1.0. Data cleaning removes incomplete, inaccurate, and incorrect data, from the given dataset, and it also handles the missing values. There are different methods in terms of handling missing values. For example, while decision tree-based or regression models are leveraged, missing values can be replaced by the most probable value. Another solution for getting rid of missing values is to use attribute's mean value to replace the missing value when the dataset follows Gaussian distribution (Balakrishnan et al., 2009). There are three main solutions to handle noisy data: binning, regression, and clustering (Derczynski et al., 2013). Binning method sorts all data points, separates them, and stores them in the form of bins. Binning method accomplishes smotting through three approaches: smoothing by mean, median, and boundary. The first two methods replace the values in the bin with the bins mean and median, respectively (Evans and Grefenstette, 2018). Smoothing by bin boundary considers the minimum and maximum values of the bin values, and replaces them with the closest boundary value. Ridge Regression is employed to identify the suitable variables

for analysis. Clustering, an unsupervised learning technique, is leveraged to detect the outliers and group the data points (Mousavi et al., 2015).

Data integration is one of the main parts of data management where multiple data sources are merged into a single dataset. Data integration has three problems that should be taken into account during the process; schema integration, entity identification problem, and detecting and resolving data value concepts (Argelaguet et al., 2021). The problem of entity identification is to detect entities from a number of databases. Regarding detecting and resolving data value concepts, the difference between data taken from multiple databases should be considered during the merging process (Pang et al., 2022).

2.5. MACHINE LEARNING APPLICATIONS IN ENVIRONMENTAL FIELDS

ML has proven to be effective in handling complex data patterns or formats due to its strong predictive capabilities. As a result, ML, particularly deep learning, has seen widespread growth in various applications over the past decade, such as image classification and machine translation. Researchers in the field of environmental science and engineering have also embraced ML in various applications, including identifying environmental hazards (Tollefson et al., 2021), evaluating the health of water and wastewater infrastructure (Granata et al., 2017), enhancing treatment methods (Inoue et al., 2017), habitat stability modeling (DŲeroski (2009)), real-time decision support system for air quality (Masih, 2019) and species detection (Wäldchen and Mäder, 2018). ML is particularly well-suited for solving uncertain and dynamic environmental issues for several reasons. One of the key benefits of ML is its capability to consider a large number of factors that may have weak or nonlinear correlations with the desired outcome. Additionally, in cases where the important information is not contained in a single input variable and the essential

variables are not known beforehand, ML can be more effective than traditional statistical models at handling various data formats, such as text, images, and graphs, where some previously unknown combination of features is necessary to determine the outcome (Sarker, 2021). This makes ML particularly useful in situations where the data may be complex or multifaceted. Figure 2.7 illustrates different overall ML application in environmental domains.

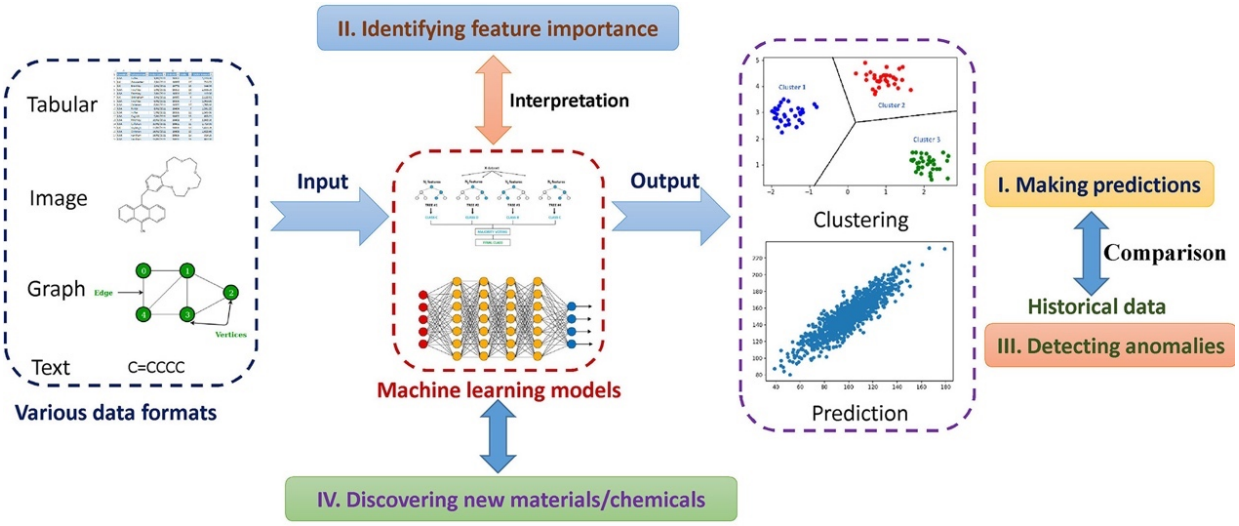

Figure 17. Schematic of the ML applications in environmental fields (Zhong et al., 2021).

2.5.1. Characterization of recent research landscape of AI applications in renewable energy systems

The number of publications and their diverse writing on AI applications has grown significantly recently (Kim et al., 2021). Stakeholders and researchers of each field can benefit from a comprehensive understanding of that domain's past and recent trends. It enables us to gain a more accurate insight into what topics researchers have been focusing on and how they evolved. A comprehensive insight into any research domain also empowers us to identify and work on essential problems (Johri et al., 2011). Conventional empirical efforts, such as surveys and

interviews to understand a research domain are challenging and produce unreliable results with a bias (Johri et al., 2011). Due to the inefficiency of conventional analysis methods for this large number of publications data, researchers have started applying data mining to unwrap the hidden information of scientific documents (Nie and Sun 2017). Text mining is a subfield of data mining focusing on analyzing and processing textual data (Choudhary et al., 2009). Text mining emerged in the late '80s and aims to discover hidden information and research trends by analyzing massive amounts of textual data (Hearst, 1999; Kostoff et al., 1999; Kostoff et al., 2000). Therefore, researchers can apply text mining techniques and bibliometric data to extract research trends in various domains (Viator and Pestorius 2001).

There have been many studies in which researchers have employed text-mining techniques to identify research trends in various fields. Viator and Pestorius (2001) used the Technology Opportunities Analysis of Scientific Information System (TECH OASIS) software to analyze text data from scientific papers on acoustic research from 1970 to 1999. TECH OASIS can be used for various text mining tasks, such as counting the term frequency, detecting the most frequent authors, and categorizing research topics. The purpose of their study was to identify trends in acoustic research. After using the software to analyze the scientific papers, they identified shifts in four areas of acoustic research between 1970 and 1999, including the proportion of US versus non-US affiliations, research areas by year, research areas by world region, and the Journal of the Acoustical Society of America's (JASA) coverage of three acoustic areas in 1999. Perez-Iratxeta et al. (2007) leveraged text mining techniques to investigate research trends in bioinformatics. They extracted papers on bioinformatics from the Medline database using a custom query that included various bioinformatics keywords from 1996 to 2005 and only considered the abstracts of the papers for text analysis. They analyzed and compared the frequency of bioinformatics terms in

the literature and found that microarray analysis was a popular topic among the bioinformatics community from 1996 to 2005.

Researchers have increasingly employed various topic modeling approaches for text mining, considering the development of AI and NLP techniques, including topic modeling. Topic modeling is a recent area of research that researchers have employed to find and investigate hidden topics in text documents (Vayansky and Kumar, 2020). Researchers have applied topic modeling to investigate research topics and trends in various fields, such as management research (Hannigan et al., 2019), transportation (Sun and Yin, 2017), marketing (Reisenbichler and Reutterer, 2019), communication research (Maier et al., 2018), hydropower (Jiang et al., 2016), renewable energies (Niroomand et al., 2022) and smart factories (Yang et al., 2018). For example, Johri et al. (2011) applied topic modeling methods to detect emerging and growing topics in engineering education from 2000 to 2008. They used the Latent Dirichlet Allocation (LDA) approach (Blei et al. 2003) to extract topics and the top 20 keywords associated with each topic in engineering education. They also extracted key phrases based on their frequency values to track their trends over time. Ayele and Juell-Skielse (2020) investigated the evolution of topics and trends in automated-driving vehicles. Their data includes 5425 publications on automated-driving vehicles research extracted from the Scopus database from 2000 to 2019. They leveraged the Dynamic Topic Model (DTM), since it considers the temporal aspect of topics, unlike LDA. Their results pointed out the evolution of twenty topics related to self-driving cars, including software system architecture and design, brake system and safety, and navigation in self-driving cars.

Researchers have carried out systematic studies investigating the direction of different types of renewable energy systems (Ding et al., 2021; Trappey et al., 2020; De Clercq et al., 2019; Ranjbari et al., 2021; Eldeeb and Mohamed, 2022). De Clercq et al. (2019) investigated the trend of biogas

invention and technology over time via leveraging text mining methods. They extracted 3186 biogas, waste management, and anaerobic digestion patents from the US patent database from 1990 to 2017. They applied LDA to calculate the term frequency-inverse document frequency (TF-IDF) score for unigrams, bigrams and trigrams. They showed the technologies' emergence annually by considering the top ten keywords based on the TF-IDF score. They identified 20 topics that domain experts could understand. However, identified topics were not labeled, which can be considered as one of the drawbacks of this study. Also, they have done some graph analysis, which showed the relation of different technologies. For example, they conducted a co-occurrence analysis of technology keywords with the most frequent groups of terms. Their results showed the characteristic technology concepts for the food waste corpus in 2016 and 2017 included "inhibitory secondary products", "hydrothermal low temperature", "Fenton reaction catalyst", "biomass pyrolyzing zone", and "concentrated organic waste", among others. Characteristic technology concepts for the biogas patent corpus in 2016 and 2017 included: "free nitrous acid", "hydrogen sulfide adsorbent", "biological treatment unit", "nanocarbon production method", "oxygen transport membrane", "pressure synthesis gas", and "micro turbine assembly". Characteristic technology concepts for the anaerobic digestion patent corpus in 2016 and 2017 included: "separation composite membrane", "waste processing tank", "polar biomass solution", "waste heat energy", "gas separation composite", and "gas separating layer".

Although many researchers investigated AI applications for different types of renewable energy systems (Dellosa and Palconit, 2021; Jha et al., 2017; Shin et al., 2021; AlShabi and Assad, 2021), they did not consider the temporal aspect in analyzing AI in renewable energy systems. To clarify, they did not show how these research topics have evolved. Investigating topic evolution is essential to accelerate research and development in renewable energy systems. It gives researchers a more

accurate insight into what methodologies and techniques were utilized within previous research at different periods. Moreover, researchers did not use an automatic or semi-automatic framework to characterize AI in renewable energy research, which might cause a less comprehensive understanding of this field, some biases, and human errors (Tao et al., 2020).

2.5.2. Machine learning applications in biogas systems

Renewable energy is becoming increasingly important in addressing the world's energy demands and mitigating the impact of climate change (Akadiri and Adebayo, 2022). Biogas systems, which produce energy from organic waste material, are crucial in the renewable energy mix (Mancini and Raggi, 2022). They have multiple benefits, such as reducing the amount of organic waste disposed in landfills and helping decrease methane emissions and other GHGs (Kougias and Angelidaki, 2018; Pöschl et al., 2010). Therefore, there is a significant potential to replace biogas as one of the main energy resources worldwide. However, biogas systems are complex due to various factors that influence the production and optimization of biogas, including feedstock composition, process parameters, microbial community, etc. The interplay between these factors, especially in large-scale systems, creates a challenge for maximizing biogas production (Hu et al., 2018) and ensuring the sustainability and efficiency of biogas system (Xu et al., 2018; Matuszewska et al., 2016; Westerholm et al., 2019; Mainardis et al., 2019). For instance, the availability of multiple feedstock for a biogas system often leads to anaerobic co-digestion (AcoD). While co-digestion has advantages, it can trigger a range of unwanted chemical reactions that can negatively impact the performance of the biogas system. (De Clercq et al., 2019). ML can help address these complexities in biogas systems since it offers various advantages in optimizing biogas production, leading to improved energy efficiency and increased biogas yield. ML algorithms can be used to analyze large amounts of data related to the biogas production process,

including process parameters, feedstock composition, and operational conditions. They can accurately predict future biogas production, allowing operators to make informed decisions about optimizing the production process to maximize the system's yield. Moreover, ML approaches can be utilized to predict and prevent potential system failures by providing real-time data analysis and process control, ensuring the stability and sustainability of biogas systems. Therefore, academic and applied researchers in the biogas domain have started applying AI to biogas systems to optimize the system and improve the yield. For example, le et al. (2022) used five different ML techniques: LR, RF, SVM, ANN, and XGBoost; the RF outperformed others with an R^2 score of 0.74. They measured the importance of three categories of routine monitoring indicators (feed amount, feedstock properties, and digester properties), individually or collectively. Feature importance analysis showed that the significance descended in the order of feed amount (45.9%), digester properties (38.6%), and feedstock properties (15.4%). In Xiao et al. (2021)'s study, a new two-stage model called NARX-BP hybrid neural networks was created to predict CH_4 production from in-situ biogas upgrading in biocathode microbial electrolysis cells through direct electron transfer. This model outperforms traditional one-stage models as it provides more accurate methane production predictions and insight into the mechanisms of biogas upgrading. The versatile model can be applied in various scenarios thanks to its ability to incorporate important intermediate variables. Furthermore, the model can support long-term predictions and optimal operation for anaerobic digestion or complex microbial electrolysis cells systems. In another study, Long et al. (2021) evaluated the effectiveness of 6 machine-learning algorithms to predict methane yield using genomic data and operational parameters from 8 research groups. For the classification models, RF showed the highest accuracy of 0.78 when using genomic data at the bacterial phylum level and 0.82 when operational parameters and genomic data were used. The regression models had a

low root mean square error of 0.04 using only genomic data at the bacterial phylum level. The feature importance analysis performed by RF indicated that *Chloroflexi, Actinobacteria, Proteobacteria, Fibrobacteres, and Spirochaeta* were the top 5 most significant phyla despite their relative abundances ranging from only 0.1% to 3.1%. The results of this study provide important information that can be used for early warning and proactive management of microbial communities. Ge et al. (2023) introduce the M-Anaerobic Digestion Model No.1 (ADM1) model for simulating anaerobic digestion, which employs a machine learning approach to predict the kinetic parameters of ADM1. Seventy-five biomass samples were used to develop the machine learning model, which considers the contents of C, H, O, N, S, and the digestion temperature. The sensitivity of 17 kinetic parameters was analyzed, and the seven parameters with the highest sensitivity were chosen as the model outputs. After optimization, the average R^2 for predicting the seven kinetic parameters was 0.92, and the root mean square error was 0.167. The overall accuracy of the M-ADM1, as expressed by Theil inequality coefficient, was 0.0163, 0.0327, and 0.0361 for municipal solid waste, kitchen waste, and sludge, respectively. These results support the hypothesis that incorporating machine learning models to predict crucial intermediate parameters can improve the performance of traditional ADM1. Baek et al. (2023) built AI pipelines using three algorithms - ANN, SVM, and RF - to predict the efficiency of AD in Direct Interspecies Electron Transfer (DIET)-stimulated environments. They focused on two key outputs: COD removal efficiency and methane production rate, which are important AD efficiency and stability indicators. The constructed ML models had high prediction efficiencies for both outputs (correlation coefficient > 0.934) as they utilized three operational time-based input parameters to capture the acclimation of microbial communities following changes in operating conditions. The results from the random forest model showed that the most important input variable was the time-

based parameter, which was recorded from the time of magnetite addition. Kowalczyk-Juśko et al. (2020) used a prediction model based on ANN to estimate methane production from various silage substrates using basic silage parameters. The model used input data such as silage type, pH, dry matter, dry organic matter, conductivity, and fermentation time, and the output data consisted of cumulative methane production. The resulting optimal prediction model was a Radial Basis Function (RBF) with 5 inputs, 2 neurons in a hidden layer, and 1 output. The model showed 73% quality of the network with a Root Mean Square Error (RMSE) of less than 3%, which is considered a satisfactory result. However, the model can be improved by adding a new analysis of silages. This prediction model can quickly estimate the energy value of different silages without the need for expensive, long-term analysis. Table 2.3. provides more research that applied ML techniques in the biogas domain.

Table 2.3. Previous scientific works at the intersection of AI and biogas.

Objective	Model(s)	Result	Reference
Peredicting daily biogas output from A set of waste inputs (municipal fecal residue, kitchen food waste)	Logistic regression, SVM, RF, XGBoost, and kNN	KNN showed the best results with accuracy of 0.87.	Clercq *et al.* (2019)
Predicting biogas production rate of food waste dry anaerobic digestion considering HRT, SRT, soluble chemical oxygen demand, total VFA, total solids and ammonia features.	RNN	Solid retention time and water content are important features in biogas reactions. Increasing intermediate materials, like VFAs, were easily converted into methane at higher water contents.	Seo et al., 2020
Modeling chemical processes within a biogas production system	SNN	Considering ten days of data points, the model can predict chemical processes up to the 100^{th} day with significant accuracy based on lab-scale data.	Capizzi et al., 2020
Predicting biogas production of fruits and vegetable waste considering different operational parameters	ANN	Predicted the performance with 85% accuracy	Gonçalves et al., 2021

Table 2.3. Previous scientific works at the intersection of AI and biogas.

Objective	Model(s)	Result	Reference
Identifying important operational parameters and predicting the biogas systems production rate	RF KNN SVM GLMNET	Total carbon was identified as the most important feature. KNN performed well in the regression task with a root mean square error of 26.6, and the logistic regression multiclass model gained an accuracy of 73%.	Wang et al., 2020
Predicting biogas systems performance of vegetables, fruits waste	ANFIS LSSVM	LSSVM performed better than ANFIS. LSSVM had a mean relative error (MRE %) and a mean squared error (MSE) of 2.951 and 0.0001, respectively, compared to 29.318 and 0.0039 for ANFIS.	Yang et al., 2021
Predicitng the Daily biomethane production considering Waste type and daily input volume, electricity and water consumption, and auxiliary chemical inputs of 4 years of operational data from an AcoD facility	Elastic net, RF, XGBoost	XGBoost outperformed with R^2 = 0. 88,	Clercq *et al.* (2020)
Predicting biogas production of spent mushroom compost in thermophilic and mesophilic laboratory conditions	ANN ANFIS	Root mean square error and r^2 in mesophilic condition: ANFIS are 0.1940 and 0.9998, ANN are 0.780 and 0.9981, and logistic model are 0.5111 and 0.9992, respectively. In the thermophilic condition, the Root mean square error and R^2 values were indicated as 0.3033 and 0.9997 for ANFIS, 0.3430 and 0.9992 for ANN, and 0.5506 and 0.9991 for the logistic model, respectively.	Najafi and Faizollahzadeh Ardabili, 2018
Predicting biogas production rate based on industrial data and finding important operation parameters	ACO GA ANN	R^2 = 0.9 and prediction error = 6.24%.	Beltramo et al., 2019

Note: DNN = Deep neural networks; RNN = Recurrent neural network; SNN = Spiking neural network; ANN = Artificial neural networks; RF = Random forest; KNN = K- nearest neighbor; SVM = Support vector machine; GLMNET = Generalized linear models fitting package via penalized maximum likelihood; LSSVM = Least square support vector machine; ML = Machine learning; HML = Hybrid of machine learning and Gompertz; ANFIS = Adaptive neuro-fuzzy inference system (ANFIS); ACO = Ant colony optimization (ACO); GA = Genetic algorithms (GA); AcoD: Anaerobic co-digestion.

TSF is a method used in statistical analysis based on analyzing a time-ordered sequence of data points to identify patterns and trends that can be used to make predictions about future values. It plays a critical role in the effective planning and operation of renewable energy and has become significantly important in industrial-scale projects in recent years, considering increasing the amount of time series data in renewable energy domains. AI-based architectures have demonstrated strong capability in TSF-related tasks and are widely employed in various renewable energy problems, such as energy production (Zheng et al., 2023), energy demand (Wang et al., 2023; Benali et al., 2019), resource availability (Victoria et al., 2021), etc. For example, Rahimilarki et al. (2023) developed a new deep learning-based method for fault detection and classification in wind turbine machines, utilizing time-series analysis and convolutional neural networks (CNNs). The method aims to address certain types of faults that are difficult to identify, such as those causing less than a 5% reduction in the performance of two actuators or four sensors of both inshore and offshore wind turbines in the presence of sensor noise.

Despite other renewable energy domains, the potential of AI has not been fully discovered in the biogas domain so far, especially within industrial-scale biogas systems. Considering the significant improvement in IoT equipment, the quality and quantity of industrial data have been increased considerably, and data-driven solutions can enhance the performance of industrial biogas units. Recently, researchers have started to leverage ML techniques toward industrial biogas systems. For example, De Clercq et al. (2019) aimed to improve biogas production in industrial settings by developing ML models, namely, LR, KNN, SVM, RF, and XGBoost that can predict biogas output based on specific waste inputs. The study involved using predictive algorithms on daily production data from two prominent biogas facilities in China to identify the key inputs that impact biogas production. Since biogas systems are so uncertain, authors had to conduct intensive

feature engineering, which is time-consuming and inflexible. More specifically, this inflexibility can be challenging in dynamic environments, like the biogas domain, where the nature of the data and the requirements of the machine learning model are constantly changing. Developing efficient end-to-end data-driven solutions for industrial biogas systems can benefit stakeholders, investors, and decision-makers before developing a biogas unit or enhancing the performance of existing biogas facilities. Moreover, they can help conserve resources and protect the environment by evaluating the impact of various factors on the system's performance and delivering reliable forecasting outcomes. In this regard, AI can significantly contribute to the popularity of biogas systems and make them more sustainable and reliable for societies.

2.5.3. Machine learning applications in biohydrogen systems

Biohydrogen systems are a promising renewable energy technology that can reduce GHG emissions by providing a cleaner alternative to fossil fuels and help mitigate the adverse effects of climate change. They are based on biological processes to produce hydrogen gas from organic materials such as biomass or wastewater (Li et al., 2022). Biohydrogen can be generated through different methods, namely, dark fermentation, photo fermentation, photo-dark fermentation, and microbiological electrolysis cells (Sharma et al., 2022). All biohydrogen systems use microorganisms such as bacteria and algae to break down organic materials and produce hydrogen gas as a by-product. For example, through dark fermentation process, anaerobic bacteria use glycolytic pathways to convert glucose into pyruvate, a process that involves the synthesis of adenosine triphosphate (ATP) from adenosine diphosphate (ADP) and the reduction of nicotinamide adenine dinucleotide (Chong et al., 2009). Pyruvate ferredoxin oxidoreductase and hydrogenase enzymes then convert pyruvate into acetyl coenzyme A, carbon dioxide, and H_2. Acetyl coenzyme A may also be converted to acetate, butyrate, and ethanol (Ntaikou et al., 2010).

The standard products during dark fermentation are acetate, butyrate, formate, and hydrogen, with a theoretical yield of 4 mol of hydrogen per mole of glucose (Sharma et al., 2022) (Table 2.3). However, the actual yield varies between 2 and 4 mol of H_2 per mole of glucose depending on the culture conditions (Sharma et al., 2022). Various types of anaerobic bacteria, including obligate anaerobes such as *Clostridium sp*. and facultative anaerobes such as *E. coli*, *Enterobacter*, and *Citrobacter*, have shown effectiveness in BioH_2 production. Spore-forming microbes like *C. butyricum*, *C. acetobutyricum*, *C. beijerinckii*, *C. thermolacticum*, *C. tyrobutyricum*, *C. thermocellum*, and *C. paraputrificum* have been studied in-depth due to their potential for scaling up the process (Chong et al., 2009; Sharma et al., 2022). Figure 2.8. illustrates the dark fermentation process for producing bioH_2.

Table 2.4. Associated reactions in the Ethanol, Butyrate, and Acetate pathways.

Step	Reaction
Ethanol pathway	$C_6\ O_{12}\ O_6 \rightarrow 2CH_3\ CH_2OH + 2CO_2 + 2H_2$
Butyrate pathway	$C_6\ O_{12}\ O_6 + 2H_2O \rightarrow 2CH_3\ CH_2CH_2COO^- + 2CO_2 + 5H_2$
Acetate pathway	$C_6\ O_{12}\ O_6 + 2H_2O \rightarrow 2CH_3COO^- + 4H_2 + 2CO_2$

However, similar to biogas systems, biohydrogen systems have high uncertainty (Katakojwala et al., 2022). For example, one of the key complexities associated with biohydrogen systems is the complexity of the biological processes involved. These processes are influenced by many factors, including temperature, pH, and nutrient availability (Ramírez-Díaz et al., 2022). Additionally, the microorganisms involved in biohydrogen production can be sensitive to environmental changes, making it challenging to maintain stable and consistent hydrogen

production over time (Cho et al., 2021). ML techniques can be used to analyze and interpret large datasets generated during the biohydrogen production process, enabling more accurate predictions and more efficient production process optimization. By leveraging advanced algorithms and computational power, ML can efficiently help identify patterns and correlations in data that might not be readily apparent through traditional statistical analysis. This can help researchers better understand the complex biological and environmental factors that impact biohydrogen production, leading to more effective strategies for improving yield and efficiency (Shen et al., 2022) (Figure 2.9).

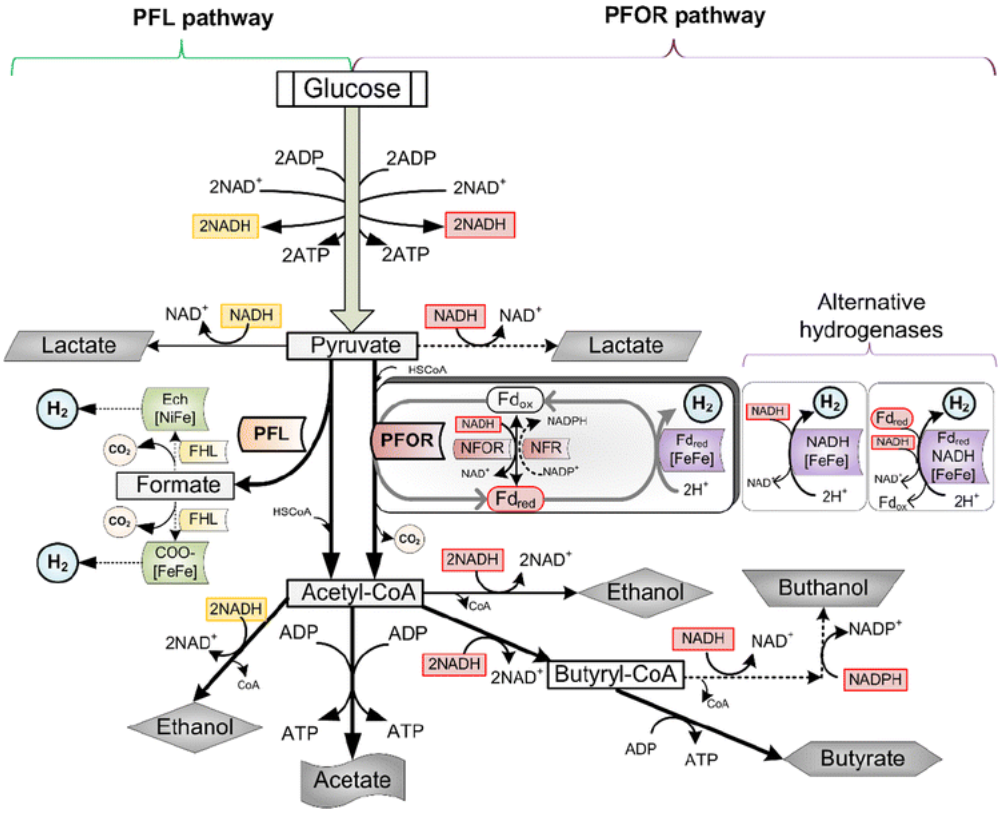

Figure 2.8. Biohydrogen production by dark fermentation pathway (Tapia-Venegas et al., 2015).

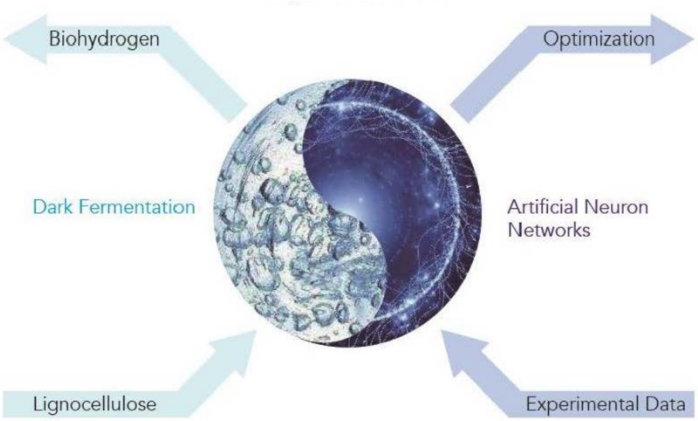

Figure 2.9. Artificial Intelligence application in biohydrogen domain (Liu et al. 2020).

Besides, ML can also be used to support real-time monitoring and control of biohydrogen production systems. By analyzing the data generated from Internet of Thing (IoT) equipment and other sources, ML models can detect and prevent potential faults, improving the overall reliability and stability of the production process (Hosseinzadeh et al., 2022). Therefore, researchers in biohydrogen domain have started leveraging AI-based methods to enhance the scope of their work. For instance, Sydney et al. (2020) examined the effectiveness of three different artificial neural network (ANN) models, which were based on the production and categorization of volatile fatty acids (VFA), in predicting the following three outcomes, accumulated hydrogen (H_2) production, hydrogen production rate, and H_2 yield. The study used data from a previous investigation that focused on the kinetics of biohydrogen and VFA production in a lab-scale setting, using this information to train and validate the models. The input variables included time and varying concentrations of acetate and butyrate (model 1), lactate, acetate, propionate, and butyrate (model 2), the sum of all VFA (model 3), and butyrate/acetate (model 4). All four models demonstrated high accuracy in predicting the aforementioned outcomes with R^2 score of greater than 0.987. They

suggested that using VFA as an input parameter is ideal for processes that uses pure cultures, whereas a model based on acetate and butyrate is recommended for more complex/mixed cultures. Hosseinzadeh et al. (2022) used ML approaches, SVM, GBM, AdaBoost, and RF, to measure the importance of parameters in the process and predict hydrogen production in the dark fermentation process from wastewater. They considered different key parameters including Fe, Ni, biomass proportion, acetate (A), butyrate (B), A/B, ethanol, pH, HRT and COD to predict hydrogen production from wastewater. The R^2 values obtained were 0.893, 0.902, 0.885, and 0.889 for GB, RF, SVM, and AdaBoost, respectively. Among these models, the RF approach was the most effective. By using permutation variable importance method, the relative importance of the effective factors in the process was determined. Figure 2.10. illustrates the relative importance of considered factors using mentioned ML models. Table 2.5. provides different ML models applied to the biohydrogen domain.

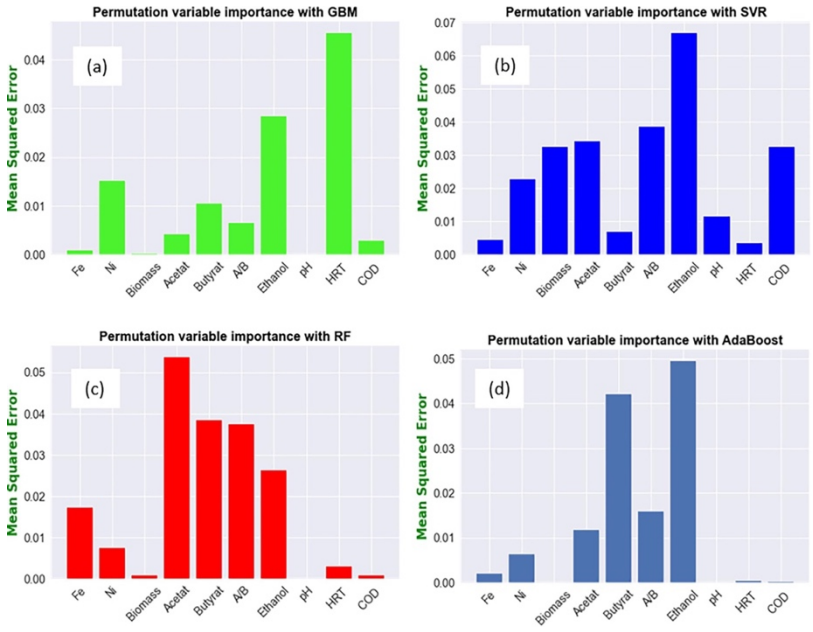

Figure 2.10. Permutation variable importance using ML algorithm (Hosseinzadeh et al. 2022).

Table 2.5. Application of ML in biohydrogen studies

Models	No of Data Points	Input(s)/ Features	Output(s)	No of hidden layers	Total Neurons	Results	References
ANN	Points = 205 Training =70.37% Testing = 19.44% Validation = 10.19%	Reactor/feed type; volatile solid; pH; OLR; HRT; temperature; reactor volume	Cumulative; biogas production	1	5 to 8	Among 24 networks tested; net10 shows best accuracy; R^2 = 0.9929	(Neto et al., 2021)
ANN	Points = 280 Training = 140 Testing = 140	Time; COD; effluent pH; VFA	HPR; maximum COD removal rate	2	12; 4	COD removal = 99% H_2 production = 6570 mL/d R^2 = 0.994	(Yogeswari et al., 2019)
ANN	Points =120 Training = 50% Testing = 25% Validation = 25%	Concentrations of COD, ALK, VFAs, and HRT, pH, and ORP	COD removal efficiency	1	4; 12; 20	Among the ten training algorithms, Levenberg-Marquardt algorithm was the best R^2 = 0.9704 MSE = 0.0150	(Yi-Fan et al., 2017)
ANN (Back Propagation)	Points = 313 Training = 60% Testing = 20% Validation = 20%	Initial substrate concentration, biomass concentration, temperature, initial pH, time	HPR	2	6; 4	R^2 = 0.988, 0.987, and 0.996 for each dataset	(Nasr et al., 2013)
ANN	Number of points not specified; Training = 70% Testing = 15% Validation = 15%	Not indicated	OLR; HPR; COD	1	5; 6; 9	Average R^2 = 0.92	(Ghasemian et al., 2019)
ANN	Points = 231	COD, pH, Dark fermentation time; VFAs	HPR, COD removal efficiency	1	20	R^2 = 0.607/0.907; 0.823/0.870	(Sridevi et al., 2014)

Table 2.5. Continued

Models	No of Data Points	Input(s)/ Features	Output(s)	No of hidden layers	Total Neurons	Results	References
ANN	Points = 50 Training = 41 Validation = 9	Concentration of substrate; Applied voltage; pH; temperature; reactor configuration	H_2 Yield	1	6; 8; 11; 12; 14	R^2 = 0.70–0.90	(Sewsynker et al., 2015)
MLP-ANN	Points =182	inoculum type; substrate type; substrate concentration; temperature; pH	H_2 yield (volume/g substrate) and (moles/mol substrate)	2	7; 7	volume/g MSE = 0.004; 0.42; R^2 = 0.90; mol/mol MSE = 0.006; 0.08; R^2 = 0.46	(Sewsynker and Kana, 2016)
ANN-ANOVA-RSM hybrid	Experimental samples from 29 batch experiments that are duplicated; Training = 80% Validation = 20%	Concentration of substrate containing sugar cane molasses; inoculum size; fermentation temperature; initial pH	Cumulative H_2 Production	1	6 to 10 neurons	12 generations; 29 population size; 60 % cross-over rate; 30 % parent size; 10 % mutation rate; R^2 = 0.91; Prediction Error = 15.12	(Whiteman and Gueguim Kana, 2014)
Generalization potential of ANN against RSM and other mechanistic models	Points = 30 Calibration = 80% Validation =20%	pH; concentrations of vanadium, iron, molybdenum, and light intensity	H_2 production	1	9	R^2 = 0.939	(Monroy et al., 2018)

Table 2.5. Continued

Models	No of Data Points	Input(s)/ Features	Output(s)	No of hidden layers	Total Neurons	Results	References
Hybrid ANN with fuzzy logic		Acidification pH; acidification time; COD of urban organic waste in the absence of inoculum	H_2 percentage in biogas; daily production	1	10	$R^2 = 0.8485$	(Moreno-Cárdenas et al., 2015)
ANN and ANFIS models	Points = 119 Training = 70% Testing = 15% Validation = 15%	OLR; effluent pH; mixed liquid SS; mixed liquid VSS	H_2 production	1	9	ANFIS $R^2 = 0.93$ MSE = 0.0073 ANN $R^2 = 0.88$ MSE = 0.00802	(Taheri et al., 2021)
Comparison between SVM; RF; GBM; AdaBoost	Points = 210 Training = 80% Testing = 20%	metal based catalysts (iron; nickel); biomass (inoculum) proportion; acetate to butyrate ratio; concentration of butyrate; acetate and ethanol; pH; COD; HRT	H_2 production	-	Doesn't specify:	MSE = 0.002–0.023; 0.023–0.032; $R^2 = 0.853–0.985$; 0.734–0.805	(Hosseinzadeh et al., 2022)

ANN = Artificial neural networks; RF = Random forest; SVM = Support vector machine; ANFIS = Adaptive neuro-fuzzy inference system; GA = Genetic algorithms; RSM = Response surface methodology ; ANOVA = Analysis of variance; MLP = Multilayer perceptron; GBM = Gradient Boosting Machines; ORP = Oxidation-reduction potential COD = Chemical oxygen demand; HRT: Hydraulic retention time; HPR = hydrogen production rate ; OLR = Organic loading rate; SS = suspended solid; VSS = Volatile suspended solid; MSE = Mean Square Error; VFA= Volatile fatty acids.

Genetic Algorithm (GA) is another useful model was used to optimize H_2 production process and can efficiently combine with other predictive ML methods (Kormi et al., 2018). A GA is a search technique that imitates natural selection and genetics to find optimal solutions. GA begins with a population of randomly generated potential solutions, called chromosomes, which are collections of symbols, often binary bits (Mirjalili and Mirjalili, 2019). Chromosomes are evaluated for fitness and then combined through crossover or mutated to produce the next generation of children (Mirjalili and Mirjalili, 2019). This process repeats, with physically fitter chromosomes having a higher probability of being selected. GA's significant nodes for optimization include the chromosome code, fitness function, selection, reproduction, crossover, and mutation mechanism (Kormi et al., 2018). Fig. 2.11. presents these critical features. GA ultimately aims to find the ideal chromosome that represents the best answer to the problem.

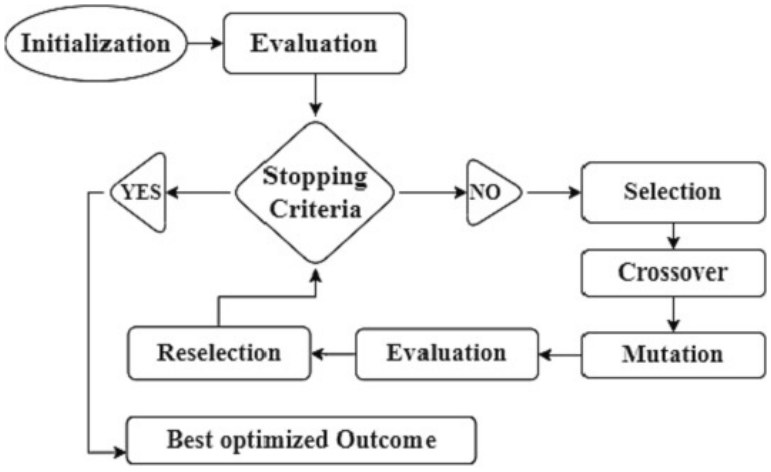

Figure 2.11. The optimization process in a Genetic Algorithm

For instance, Mahata et al. (2020) investigated the production of BioH_2 from organic waste, specifically starchy wastewater supplemented with groundnut de-oiled cake. They used mathematical tools such as response surface methodology (RSM) and AI, including ANNs and

SVM, to analyze the experimental results. The study found that SVM had better prediction abilities than ANN and RSM. The researchers then integrated these AI-based models with GA and particle swarm optimization (PSO) to determine the optimal process parameters. The ideal parameter value was found to be similar for both GA and PSO. However, PSO was discovered to be quicker than GA. Using an SVM-based model, the H_2 yield increased by 2.1 times compared to the unoptimized condition.

In 2014, Whiteman and Gueguim Kana conducted a study on hydrogen production and investigated four input variables: substrate concentration, temperature, inoculum size, and initial pH. They used the Box-Behnken design to collect data through 29 runs, which involved individual variations of the input variables. The study aimed to assess the importance of the mutual interactive contour plot, which should have high values of R^2, F-value, and signal-to-noise ratio, by evaluating the accuracy of the polynomial function of RSM using ANOVA. They also used five proposed topology models with different neuron densities in a single hidden layer to train the RMSE and reflect the predictability of ANN. The best ANN model was integrated into GA, and after 13 generations, a population size of 30, a cross-over rate of 65%, a parent size of 32%, and a mutation rate of 12%, the hybrid ANN-GA method significantly reduced the overestimation of RSM from a 119.08% error difference to 15.15%.

Prakasham et al. (2011) developed a hybrid model that combines an ANN architecture consisting of four input layers, ten hidden layers, and one output layer (GA), to predict the yield of the fermentation process and optimize it. They had sixteen experimental data points, 80% of which was used as the train part, and the rest was considered the test part. Their chosen ANN topology showed promising results with R^2 of 0.99 and very small error for both the training and testing parts. Their results are shown in Table 2.6.

Table 2.6. Prakasham et al. (2011)'s Artificial Neural Networks results

Part	MSE	MAE	MAPE
Train	9.1×10^{-8}	3.38×10^{-8}	2.81×10^{-10}
Test	3.33×10^{-8}	1.3×10^{-7}	5.7×10^{-8}

MSE: Mean Square Error; MAE: Mean Absolute Error; MAPE: Mean Absolute Precision Error

GA later optimized their ANN's results to get the most efficient amount of pH, inoculum aging, glucose to xylose ratio, and inoculum concentration (Table 2.7).

Table 2.7. Selected optimum fermentation conditions predicted by GA and experimental verification of biohydrogen yield.

	pH	Glucose: Xylose	Inoculum size (mg)	Age of inoculum (h)	Biohydrogen production (ml g^{-1} substrate)	
					GA-predicted	Experimental
1	6	2:3	80	12	350.12	357.43
2	5.8	2:3	84	13	380.35	378.29
3	5.5	3:2	90	11	360.35	334.18
4	6	2:3	83	15	370.58	329.37

2.5.4. Image Processing applications in biogas systems

Image processing is a useful method used in the biogas industry to identify the area's potential for producing biogas systems' substrate. It can also be used for real-time monitoring of the process to ensure optimal conditions for the microorganisms, which maximizes biogas production. Image processing approaches can also be used to monitor the quality of the biogas produced by analyzing images of the gas to detect impurities or changes in composition (Wiedemann et al., 2017). Dinova et al. (2018) used image processing to control of biogas production process. They analyzed the biogas production process at the 'Kubratovo, Bulgaria, wastewater treatment plant in two different seasons, and a correlation was established between the control parameters using various methods, including aerobic and anaerobic dehydrogenase

activities, chemical and technological indicators (temperature, pH, COD, PO_4^{3-}, NH_4^+, dry organic matter, the ratio of volatile fatty acids to the total alkalinity) and fluorescent image analysis. The fluorescent indicative system works by introducing a fluorescent dye into the anaerobic digestion process. As the microorganisms in the system consume the dye, it emits a fluorescent signal that can be measured and monitored over time. This signal provides information about the activity and health of the microbial community, which can be used to optimize the system's performance. The correlation analysis results indicated that the fluorescent image analysis parameters, such as clusters' number/mean size, fluorescence intensity, and area, were highly correlated with biogas production (Figure 2.12).

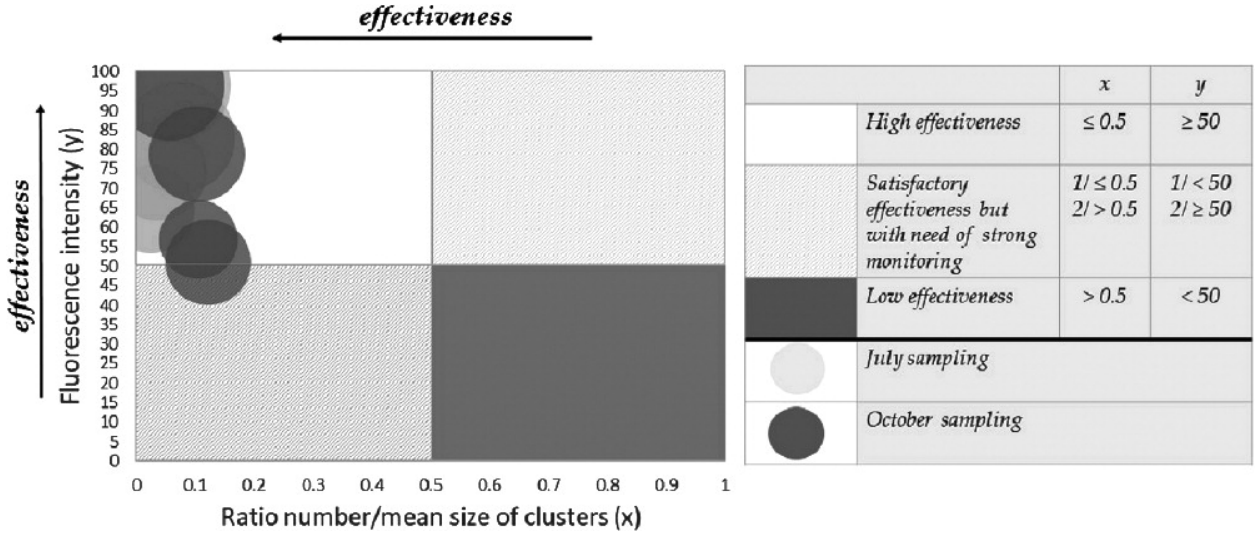

Figure 2.12. Fluorescent indicative system for assessment of the effectiveness of anaerobic digesters (Dinova et al., 2018).

This system measures the effectiveness of digester performance using two factors: the brightness of the biological system's fluorescence and the ratio of the number of clusters formed by the microbial consortium to their average size. The four digesters at the wastewater treatment

plant of Kubratovo were highly effective during two sampling periods, indicating that the entire facility generates 111% of the electricity it requires through AD. High fluorescence intensity suggests high biological activity and a low ratio of cluster number to average size indicates strong synergistic and syntrophic relationships within the microbial consortium. During the autumn sampling period, the performance of the third digester was marginally effective, falling between highly effective and satisfactorily effective. Satisfactorily effective zones are defined as areas where parameters are not within their optimal values but where the process can be monitored carefully. Ineffective zones are those with low fluorescence intensity and weak synergistic and syntrophic relationships in the biological system. Consequently, a fluorescent indicative system was suggested for controlling biogas production technologies, which would serve as a quick assessment tool for the effectiveness of anaerobic digestion.

Valenti et al. (2017) aimed to evaluate the availability of olive pomace (OP), the main waste produced by the olive oil industry, for use as a resource in biogas production. A geographic information system (GIS)-based model was used to compute indicators that describe the potential production of OP in specific geographic areas. Initially, the study focused on analyzing the spatial distribution of olive-producing areas in Sicily, a region that is highly representative of olive oil production in the Mediterranean Basin. The GIS-based model was applied to estimate the potential production of OP using data collected through surveys about olive oil industries. This included indicators such as the amount of olive oil produced and the amount of OP obtained.

The study's second phase focused on quantifying the amount of OP available for biogas production at a provincial level in areas with the highest potential for OP production. The results indicated a theoretical potential for producing 1.9 million Nm^3 of biogas from OP, demonstrating that it could be a valuable resource for renewable energy production. This finding could help

address the environmental burden of OP disposal and contribute to the sustainability of the biogas sector. The GIS-based model used in this study could also be used to build an information base for improving the sustainability of the biogas sector by identifying the best locations for new biogas plants in terms of optimizing the logistics of biomass supply. The study focused on utilizing information such as rural and livestock populations, land-use maps, and GIS to develop a model for evaluating biogas production from livestock manure and rural household waste in Iran. The model can identify suitable locations for constructing biogas production plants. This study's analysis process was more detailed than previous studies in Iran, allowing for a more accurate assessment of available biomass and suitable sites for biogas plant construction. The study showed that biogas production from livestock manure and rural waste could produce 2740 million m^3/year of methane. Lovrak et al. (2020) proposed a method for assessing the spatial distribution of biogas production potential, considering the seasonal variation in biomass production. The method combines statistical and spatial explicit methods, and uses a GIS approach. The case study is conducted in Croatia, and the results show that the proposed approach is more effective than current approaches. Their results demonstrate the importance of considering seasonality when assessing biogas potential for agricultural residues and show that the proposed approach can result in significant savings in storage facility capacity.

Refrences

Ntaikou, I., Antonopoulou, G., & Lyberatos, G. (2010). Biohydrogen production from biomass and wastes via dark fermentation: a review. *Waste and Biomass Valorization*, *1*, 21-39.

Chong, M. L., Sabaratnam, V., Shirai, Y., & Hassan, M. A. (2009). Biohydrogen production from biomass and industrial wastes by dark fermentation. *International journal of hydrogen energy*, *34*(8), 3277-3287.

Tapia-Venegas, E., Ramirez-Morales, J. E., Silva-Illanes, F., Toledo-Alarcón, J., Paillet, F., Escudie, R., ... & Ruiz-Filippi, G. (2015). Biohydrogen production by dark fermentation: scaling-up and technologies integration for a sustainable system. *Reviews in Environmental Science and Bio/Technology*, *14*, 761-785.

Sharma, A. K., Ghodke, P. K., Goyal, N., Nethaji, S., & Chen, W. H. (2022). Machine learning technology in biohydrogen production from agriculture waste: Recent advances and future perspectives. *Bioresource Technology*, 128076.

Mirjalili, S., & Mirjalili, S. (2019). Genetic algorithm. *Evolutionary Algorithms and Neural Networks: Theory and Applications*, 43-55.

Ahmed, S., & Kazda, M. (2017). Characteristics of on-demand biogas production by using sugar beet silage. *Anaerobe*, *46*, 114-121.

Bumbiere, K., Gancone, A., Pubule, J., Kirsanovs, V., Vasarevicius, S., & Blumberga, D. (2020). Ranking of bioresources for biogas production. *Rigas Tehniskas Universitates Zinatniskie Raksti*, *24*(1), 368-377.

Ijoma, G. N., Nkuna, R., Mutungwazi, A., Rashama, C., & Matambo, T. S. (2021). Applying PICRUSt and 16S rRNA functional characterisation to predicting co-digestion strategies of various animal manures for biogas production. *Scientific Reports*, *11*(1), 19913.

Okonkwo, U. C., Onokpite, E., & Onokwai, A. O. (2018). Comparative study of the optimal ratio of biogas production from various organic wastes and weeds for digester/restarted digester. *Journal of King Saud University-Engineering Sciences*, *30*(2), 123-129.

Ghatak, M. D., & Ghatak, A. (2018). Artificial neural network model to predict behavior of biogas production curve from mixed lignocellulosic co-substrates. *Fuel*, *232*, 178-189.

Nwokolo, N., Mukumba, P., Obileke, K., & Enebe, M. (2020). Waste to energy: A focus on the impact of substrate type in biogas production. *Processes*, *8*(10), 1224.

Prakasham, R. S., Sathish, T., & Brahmaiah, P. (2011). Imperative role of neural networks coupled genetic algorithm on optimization of biohydrogen yield. *international journal of hydrogen energy*, *36*(7), 4332-4339.

Mahata, C., Ray, S., & Das, D. (2020). Optimization of dark fermentative hydrogen production from organic wastes using acidogenic mixed consortia. *Energy Conversion and Management*, *219*, 113047.

Kormi, T., Mhadhebi, S., Ali, N. B. H., Abichou, T., & Green, R. (2018). Estimation of fugitive landfill methane emissions using surface emission monitoring and Genetic Algorithms optimization. *Waste management*, *72*, 313-328.

Sydney, E. B., Duarte, E. R., Burgos, W. J. M., de Carvalho, J. C., Larroche, C., & Soccol, C. R. (2020). Development of short chain fatty acid-based artificial neuron network tools applied to biohydrogen production. *International Journal of Hydrogen Energy*, *45*(8), 5175-5181.

Johannesson, G. H., Crolla, A., Lauzon, J. D., & Gilroyed, B. H. (2020). Estimation of biogas co-production potential from liquid dairy manure, dissolved air flotation waste (DAF) and dry poultry manure using biochemical methane potential (BMP) assay. *Biocatalysis and Agricultural Biotechnology*, *25*, 101605.

Limeneh, D. Y., Tesfaye, T., Ayele, M., Husien, N. M., Ferede, E., Haile, A., ... & Kong, F. (2022). A comprehensive review on utilization of slaughterhouse by-product: Current status and prospect. *Sustainability*, *14*(11), 6469.

Selormey, G. K., Barnes, B., Awafo, E. A., Kemausuor, F., & Darkwah, L. (2022). Development of mathematical model for predicting methane-to-carbon dioxide proportion in anaerobic

biodegradability of cattle blood and rumen content. *Energy Conversion and Management: X*, *16*, 100250.

Lovrak, A., Pukšec, T., & Duić, N. (2020). A Geographical Information System (GIS) based approach for assessing the spatial distribution and seasonal variation of biogas production potential from agricultural residues and municipal biowaste. *Applied energy*, *267*, 115010.

Zareei, S. (2018). Evaluation of biogas potential from livestock manures and rural wastes using GIS in Iran. *Renewable energy*, *118*, 351-356.

Kowalczyk-Juśko, A., Pochwatka, P., Zaborowicz, M., Czekała, W., Mazurkiewicz, J., Mazur, A., ... & Dach, J. (2020). Energy value estimation of silages for substrate in biogas plants using an artificial neural network. *Energy*, *202*, 117729.

Valenti, F., Arcidiacono, C., Chinnici, G., Cascone, G., & Porto, S. M. (2017). Quantification of olive pomace availability for biogas production by using a GIS-based model. *Biofuels, Bioproducts and Biorefining*, *11*(5), 784-797.

Dinova, N., Belouhova, M., Schneider, I., Rangelov, J., & Topalova, Y. (2018). Control of biogas production process by enzymatic and fluorescent image analysis. *Biotechnology & Biotechnological Equipment*, *32*(2), 366-375.

Wiedemann, L., Conti, F., Janus, T., Sonnleitner, M., Zörner, W., & Goldbrunner, M. (2017). Mixing in Biogas Digesters and Development of an Artificial Substrate for Laboratory-Scale Mixing Optimization. *Chemical Engineering & Technology*, *40*(2), 238-247.

Liu, L., Chen, H., & Han, Y. (2010). Determination and analysis of physical characteristics and fiber chemical composition of biogas residue. *Transactions of the Chinese Society of Agricultural Engineering*, *26*(7), 277-280.

Safarian, S., Ebrahimi Saryazdi, S. M., Unnthorsson, R., & Richter, C. (2021). Modeling of hydrogen production by applying biomass gasification: Artificial neural network modeling approach. *Fermentation*, *7*(2), 71.

Shen, M. Y., Torre, M., Chu, C. Y., Tratzi, P., Carnevale, M., Gallucci, F., ... & Petracchini, F. (2022). Green biohydrogen production in a Co-digestion process from mixture of high carbohydrate food waste and cattle/chicken manure digestate. *International Journal of Hydrogen Energy*, *47*(96), 40696-40703.

Hosseinzadeh, A., Zhou, J. L., Altaee, A., & Li, D. (2022). Machine learning modeling and analysis of biohydrogen production from wastewater by dark fermentation process. *Bioresource technology*, *343*, 126111.

Cho, B. A., Ross, B. S., du Toit, J. P., Pott, R. W. M., del Río Chanona, E. A., & Zhang, D. (2021). Dynamic modelling of Rhodopseudomonas palustris biohydrogen production: Perturbation analysis and photobioreactor upscaling. *International Journal of Hydrogen Energy*, *46*(74), 36696-36708.

Ramírez-Díaz, R. C., Prato-Garcia, D., & Vasquez-Medrano, R. (2022). How sustainable is the biohydrogen produced from sugarcane vinasse? An approach based on life cycle assessment. *Biomass Conversion and Biorefinery*, 1-21.

Katakojwala, R., & Mohan, S. V. (2022). Multi-product biorefinery with sugarcane bagasse: Process development for nanocellulose, lignin and biohydrogen production and lifecycle analysis. *Chemical Engineering Journal*, *446*, 137233.

Usman, T. M., Banu, J. R., Gunasekaran, M., & Kumar, G. (2019). Biohydrogen production from industrial wastewater: an overview. *Bioresource Technology Reports*, *7*, 100287.

Li, S., Li, F., Zhu, X., Liao, Q., Chang, J. S., & Ho, S. H. (2022). Biohydrogen production from microalgae for environmental sustainability. *Chemosphere*, *291*, 132717.

Ioannou-Ttofa, L., Foteinis, S., Moustafa, A. S., Abdelsalam, E., Samer, M., & Fatta-Kassinos, D. (2021). Life cycle assessment of household biogas production in Egypt: Influence of digester volume, biogas leakages, and digestate valorization as biofertilizer. *Journal of Cleaner Production*, *286*, 125468.

Westerholm, M., Hansson, M., & Schnürer, A. (2012). Improved biogas production from whole stillage by co-digestion with cattle manure. *Bioresource technology*, *114*, 314-319.

Qyyum, M. A., Haider, J., Qadeer, K., Valentina, V., Khan, A., Yasin, M., ... & Lee, M. (2020). Biogas to liquefied biomethane: Assessment of 3P's–Production, processing, and prospects. *Renewable and Sustainable Energy Reviews*, *119*, 109561.

Scarlat, N., Fahl, F., Dallemand, J. F., Monforti, F., & Motola, V. (2018). A spatial analysis of biogas potential from manure in Europe. *Renewable and Sustainable Energy Reviews*, *94*, 915-930.

Orhorhoro, E. K., Ebunilo, P. O., & Sadjere, G. E. (2017). Experimental determination of effect of total solid (TS) and volatile solid (VS) on biogas yield. *American Journal of Modern Energy*, *3*(6), 131-135.

Rajendran, K., Aslanzadeh, S., & Taherzadeh, M. J. (2012). Household biogas digesters—A review. *Energies*, *5*(8), 2911-2942.

Wijesinghe, D. T. N., Dassanayake, K. B., Sommer, S. G., Scales, P., & Chen, D. (2019). Biogas improvement by adding Australian zeolite during the anaerobic digestion of C: N ratio adjusted swine manure. *Waste and Biomass Valorization*, *10*, 1883-1887.

Angelidaki, I., Boe, K., & Ellegaard, L. (2005). Effect of operating conditions and reactor configuration on efficiency of full-scale biogas plants. *Water science and technology*, *52*(1-2), 189-194.

Issah, A. A., Kabera, T., & Kemausuor, F. (2020). Biogas optimisation processes and effluent quality: A review. *Biomass and Bioenergy*, *133*, 105449.

Zhai, N., Zhang, T., Yin, D., Yang, G., Wang, X., Ren, G., & Feng, Y. (2015). Effect of initial pH on anaerobic co-digestion of kitchen waste and cow manure. *Waste management*, *38*, 126-131.

Liu, C. F., Yuan, X. Z., Zeng, G. M., Li, W. W., & Li, J. (2008). Prediction of methane yield at optimum pH for anaerobic digestion of organic fraction of municipal solid waste. *Bioresource technology*, *99*(4), 882-888.

Hajizadeh, Abdollah (2021) *Biogas production by psychrophilic anaerobic digestion and biogas-to-hydrogen through methane reforming: experimental study and process simulation. Masters thesis, Memorial University of Newfoundland.*

Shrestha, S., Fonoll, X., Khanal, S. K., & Raskin, L. (2017). Biological strategies for enhanced hydrolysis of lignocellulosic biomass during anaerobic digestion: Current status and future perspectives. *Bioresource Technology*, *245*, 1245-1257.

Menzel, T., Neubauer, P., & Junne, S. (2020). Role of microbial hydrolysis in anaerobic digestion. *Energies*, *13*(21), 5555.

Sonakya, V., Raizada, N. and Kalia, V.C., 2001. Microbial and enzymatic improvement of anaerobic digestion of waste biomass. Biotechnology letters. 23, 1463-1466.

Yuan, H. and Zhu, N., 2016. Progress in inhibition mechanisms and process control of intermediates and by-products in sewage sludge anaerobic digestion. Renewable and Sustainable Energy Reviews. 58, 429-438.

Jadhav, P., Muhammad, N., Bhuyar, P., Krishnan, S., Abd Razak, A. S., Zularisam, A. W., & Nasrullah, M. (2021). A review on the impact of conductive nanoparticles (CNPs) in anaerobic digestion: Applications and limitations. Environmental Technology & Innovation, 23, 101526.

Zahedi, R., Aslani, A., Seraji, M. A. N., & Zolfaghari, Z. (2022). Advanced bibliometric analysis on the coupling of energetic dark greenhouse with natural gas combined cycle power plant for CO2 capture. *Korean Journal of Chemical Engineering*, *39*(11), 3021-3031.

Cao, J., Bucher, D. F., Hall, D. M., & Eggers, M. (2022). A graph-based approach for module library development in industrialized construction. *Computers in Industry*, *139*, 103659.

Levene, M., & Loizou, G. (1995). A graph-based data model and its ramifications. *IEEE Transactions on Knowledge and Data Engineering*, *7*(5), 809-823.

Riesen, K., & Bunke, H. (2008, December). IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning. In *SSPR/SPR* (Vol. 5342, pp. 287-297).

Beltramo, T., Klocke, M., & Hitzmann, B. (2019). Prediction of the biogas production using GA and ACO input features selection method for ANN model. In Information Processing in Agriculture (Vol. 6, Issue 3, pp. 349–356). Elsevier BV. https://doi.org/10.1016/j.inpa.2019.01.002

Najafi, B., & Faizollahzadeh Ardabili, S. (2018). Application of ANFIS, ANN, and logistic methods in estimating biogas production from spent mushroom compost (SMC). In

Resources, Conservation and Recycling (Vol. 133, pp. 169–178). Elsevier BV. https://doi.org/10.1016/j.resconrec.2018.02.025

Zareei, S., & Khodaei, J. (2017). Modeling and optimization of biogas production from cow manure and maize straw using an adaptive neuro-fuzzy inference system. In Renewable Energy (Vol. 114, pp. 423–427). Elsevier BV. https://doi.org/10.1016/j.renene.2017.07.050

Gonçalves Neto, J., Vidal Ozorio, L., Campos de Abreu, T. C., Ferreira dos Santos, B., & Pradelle, F. (2021). Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN). In Fuel (Vol. 285, p. 119081). Elsevier BV. https://doi.org/10.1016/j.fuel.2020.119081

Capizzi, G., Lo Sciuto, G., Napoli, C., Woźniak, M., & Susi, G. (2020). A spiking neural network-based long-term prediction system for biogas production. In Neural Networks (Vol. 129, pp. 271–279). Elsevier BV. https://doi.org/10.1016/j.neunet.2020.06.001

Victoria, M., Haegel, N., Peters, I. M., Sinton, R., Jäger-Waldau, A., del Cañizo, C., ... & Smets, A. (2021). Solar photovoltaics is ready to power a sustainable future. *Joule*, *5*(5), 1041-1056.

Benali, L., Notton, G., Fouilloy, A., Voyant, C., & Dizene, R. (2019). Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renewable energy*, *132*, 871-884.

Wang, D., Gan, J., Mao, J., Chen, F., & Yu, L. (2023). Forecasting power demand in China with a CNN-LSTM model including multimodal information. *Energy*, *263*, 126012.

Zheng, J., Du, J., Wang, B., Klemeš, J. J., Liao, Q., & Liang, Y. (2023). A hybrid framework for forecasting power generation of multiple renewable energy sources. *Renewable and Sustainable Energy Reviews*, *172*, 113046.

Rahimilarki, R., Gao, Z., Jin, N., & Zhang, A. (2022). Convolutional neural network fault classification based on time-series analysis for benchmark wind turbine machine. *Renewable Energy*, *185*, 916-931.

Baek, G., Lee, C., & Yoon, J. (2023). Machine learning approach for predicting anaerobic digestion performance and stability in direct interspecies electron transfer-stimulated environments. *Biochemical Engineering Journal*, 108840.

Ge, Y., Tao, J., Wang, Z., Chen, C., Mu, L., Ruan, H., ... & Chen, G. (2023). Modification of anaerobic digestion model No. 1 with Machine learning models towards applicable and accurate simulation of biomass anaerobic digestion. *Chemical Engineering Journal*, *454*, 140369.

Long, F., Wang, L., Cai, W., Lesnik, K., & Liu, H. (2021). Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data. *Water Research*, *199*, 117182.

Xiao, J., Liu, C., Ju, B., Xu, H., Sun, D., & Dang, Y. (2021). Estimation of in-situ biogas upgrading in microbial electrolysis cells via direct electron transfer: Two-stage machine learning modeling based on a NARX-BP hybrid neural network. *Bioresource technology*, *330*, 124965.

Li, C., He, P., Peng, W., Lü, F., Du, R., & Zhang, H. (2022). Exploring available input variables for machine learning models to predict biogas production in industrial-scale biogas plants treating food waste. *Journal of Cleaner Production*, *380*, 135074.

Mainardis, M., Buttazzoni, M., Gievers, F., Vance, C., Magnolo, F., Murphy, F., & Goi, D. (2021). Life cycle assessment of sewage sludge pretreatment for biogas production: From laboratory tests to full-scale applicability. *Journal of Cleaner Production*, *322*, 129056.

Westerholm, M., & Schnürer, A. (2019). Microbial responses to different operating practices for biogas production systems. *Anaerobic digestion*, 1-36.

Hu, Y., Scarborough, M., Aguirre-Villegas, H., Larson, R. A., Noguera, D. R., & Zavala, V. M. (2018). A supply chain framework for the analysis of the recovery of biogas and fatty acids from organic waste. *ACS Sustainable Chemistry & Engineering*, *6*(5), 6211-6222.

Pöschl, M., Ward, S., & Owende, P. (2010). Evaluation of energy efficiency of various biogas production and utilization pathways. *Applied energy*, *87*(11), 3305-3321.

Mancini, E., & Raggi, A. (2022). Out of sight, out of mind? The importance of local context and trust in understanding the social acceptance of biogas projects: A global scale review. *Energy Research & Social Science*, *91*, 102697.

Akadiri, S. S., & Adebayo, T. S. (2022). Asymmetric nexus among financial globalization, non-renewable energy, renewable energy use, economic growth, and carbon emissions: impact on environmental sustainability targets in India. *Environmental Science and Pollution Research*, *29*(11), 16311-16323.

Jha, S. Kr., Bilalovic, J., Jha, A., Patel, N., & Zhang, H. (2017). Renewable energy: Present research and future scope of Artificial Intelligence. In Renewable and Sustainable Energy Reviews (Vol. 77, pp. 297–317). Elsevier BV. https://doi.org/10.1016/j.rser.2017.04.018

Boden, M. A. (Ed.). (1996). Artificial intelligence. Elsevier.

Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In Machine learning techniques for multimedia (pp. 21-49). Springer, Berlin, Heidelberg.

Miorelli, R., Kulakovskyi, A., Chapuis, B., D'almeida, O., & Mesnil, O. (2021). Supervised learning strategy for classification and regression tasks applied to aeronautical structural health monitoring problems. Ultrasonics, 113, 106372.

McAlpine, E. D., Michelow, P., & Celik, T. (2022). The utility of unsupervised machine learning in anatomic pathology. American Journal of Clinical Pathology, 157(1), 5-14

Ashfaq, R. A. R., Wang, X.-Z., Huang, J. Z., Abbas, H., & He, Y.-L. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. In Information Sciences (Vol. 378, pp. 484–497). Elsevier BV. https://doi.org/10.1016/j.ins.2016.04.019

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. In Applied Soft Computing (Vol. 90, p. 106181). Elsevier BV. https://doi.org/10.1016/j.asoc.2020.106181

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. In Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (Vol. 379, Issue 2194, p. 20200209). The Royal Society. https://doi.org/10.1098/rsta.2020.0209

Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2019). A review of deep learning for renewable energy forecasting. In Energy Conversion and Management (Vol. 198, p. 111799). Elsevier BV. https://doi.org/10.1016/j.enconman.2019.111799

Hu, Y.-L., & Chen, L. (2018). A nonlinear hybrid wind speed forecasting model using LSTM network, hysteretic ELM and Differential Evolution algorithm. In Energy Conversion and

Management (Vol. 173, pp. 123–142). Elsevier BV. https://doi.org/10.1016/j.enconman.2018.07.070

Yu, C., Li, Y., Bao, Y., Tang, H., & Zhai, G. (2018). A novel framework for wind speed prediction based on recurrent neural networks and support vector machine. In Energy Conversion and Management (Vol. 178, pp. 137–145). Elsevier BV. https://doi.org/10.1016/j.enconman.2018.10.008

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1412.3555

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. In Neural Computation (Vol. 9, Issue 8, pp. 1735–1780). MIT Press - Journals. https://doi.org/10.1162/neco.1997.9.8.1735

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Association for Computational Linguistics. https://doi.org/10.3115/v1/w14-4012

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. In IEEE Transactions on Signal Processing (Vol. 45, Issue 11, pp. 2673–2681). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/78.650093

Zhao, W., Han, S., Hu, R. Q., Meng, W., & Jia, Z. (2018). Crowdsourcing and Multisource Fusion-Based Fingerprint Sensing in Smartphone Localization. In IEEE Sensors Journal (Vol. 18,

Issue 8, pp. 3236–3247). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jsen.2018.2805335

Wang, J., Yan, J., Li, C., Gao, R. X., & Zhao, R. (2019). Deep heterogeneous GRU model for predictive analytics in smart manufacturing: Application to tool wear prediction. In Computers in Industry (Vol. 111, pp. 1–14). Elsevier BV. https://doi.org/10.1016/j.compind.2019.06.001

Zhang, N., Shen, S.-L., Zhou, A., & Jin, Y.-F. (2021). Application of LSTM approach for modelling stress–strain behaviour of soil. In Applied Soft Computing (Vol. 100, p. 106959). Elsevier BV. https://doi.org/10.1016/j.asoc.2020.106959

Miao, K., Han, T., Yao, Y., Lu, H., Chen, P., Wang, B., & Zhang, J. (2020). Application of LSTM for short term fog forecasting based on meteorological elements. In Neurocomputing (Vol. 408, pp. 285–291). Elsevier BV. https://doi.org/10.1016/j.neucom.2019.12.129

Sethia, A., & Raut, P. (2018). Application of LSTM, GRU and ICA for Stock Price Prediction. In Information and Communication Technology for Intelligent Systems (pp. 479–487). Springer Singapore. https://doi.org/10.1007/978-981-13-1747-7_46

Bach-Andersen, M., Rømer-Odgaard, B., & Winther, O. (2017). Deep learning for automated drivetrain fault detection. In Wind Energy (Vol. 21, Issue 1, pp. 29–41). Wiley. https://doi.org/10.1002/we.2142

Agga, A., Abbou, A., Labbadi, M., Houm, Y. E., & Ou Ali, I. H. (2022). CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production. In Electric Power Systems Research (Vol. 208, p. 107908). Elsevier BV. https://doi.org/10.1016/j.epsr.2022.107908

Cinar, S. Ö., Cinar, S., & Kuchta, K. (2022). Machine Learning Algorithms for Temperature Management in the Anaerobic Digestion Process. In Fermentation (Vol. 8, Issue 2, p. 65). MDPI AG. https://doi.org/10.3390/fermentation8020065

Hansen, B. D., Tamouk, J., Tidmarsh, C. A., Johansen, R., Moeslund, T. B., & Jensen, D. G. (2020). Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model. In Computational Science and Its Applications – ICCSA 2020 (pp. 734–745). Springer International Publishing. https://doi.org/10.1007/978-3-030-58799-4_53

Mahmoodi-Eshkaftaki, M., & Ebrahimi, R. (2021). Integrated deep learning neural network and desirability analysis in biogas plants: A powerful tool to optimize biogas purification. In Energy (Vol. 231, p. 121073). Elsevier BV. https://doi.org/10.1016/j.energy.2021.121073

Kougias, P. G., & Angelidaki, I. (2018). Biogas and its opportunities—A review. In Frontiers of Environmental Science & Engineering (Vol. 12, Issue 3). Springer Science and Business Media LLC. https://doi.org/10.1007/s11783-018-1037-8

Scarlat, N., Dallemand, J.-F., & Fahl, F. (2018). Biogas: Developments and perspectives in Europe. In Renewable Energy (Vol. 129, pp. 457–472). Elsevier BV. https://doi.org/10.1016/j.renene.2018.03.006

Khalil, M., Berawi, M. A., Heryanto, R., & Rizalie, A. (2019). Waste to energy technology: The potential of sustainable biogas production from animal waste in Indonesia. In Renewable and Sustainable Energy Reviews (Vol. 105, pp. 323–331). Elsevier BV. https://doi.org/10.1016/j.rser.2019.02.011

Chen, L., Cong, R.-G., Shu, B., & Mi, Z.-F. (2017). A sustainable biogas model in China: The case study of Beijing Deqingyuan biogas project. In Renewable and Sustainable Energy Reviews (Vol. 78, pp. 773–779). Elsevier BV. https://doi.org/10.1016/j.rser.2017.05.027

Scarlat, N., Dallemand, J.-F., & Fahl, F. (2018). Biogas: Developments and perspectives in Europe. In Renewable Energy (Vol. 129, pp. 457–472). Elsevier BV. https://doi.org/10.1016/j.renene.2018.03.006

Deng, L., Liu, Y., Zheng, D., Wang, L., Pu, X., Song, L., Wang, Z., Lei, Y., Chen, Z., & Long, Y. (2017). Application and development of biogas technology for the treatment of waste in China. In Renewable and Sustainable Energy Reviews (Vol. 70, pp. 845–851). Elsevier BV. https://doi.org/10.1016/j.rser.2016.11.265

Shukla, H., & Kakkar, M. (2016). Keyword extraction from Educational Video transcripts using NLP techniques. In 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). IEEE. https://doi.org/10.1109/confluence.2016.7508096

Gu, W., Yang, X., Yang, M., Han, K., Pan, W., & Zhu, Z. (2022). MarkerGenie: an NLP-enabled text-mining system for biomedical entity relation extraction. In C. Arighi (Ed.), Bioinformatics Advances (Vol. 2, Issue 1). Oxford University Press (OUP). https://doi.org/10.1093/bioadv/vbac035

MacFarlane, H., Salem, A. C., Chen, L., Asgari, M., & Fombonne, E. (2022). Combining voice and language features improves automated autism detection. In Autism Research (Vol. 15, Issue 7, pp. 1288–1300). Wiley. https://doi.org/10.1002/aur.2733

Boorugu, R., & Ramesh, G. (2020). A Survey on NLP based Text Summarization for Summarizing Product Reviews. In 2020 Second International Conference on Inventive Research

in Computing Applications (ICIRCA). 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE. https://doi.org/10.1109/icirca48905.2020.9183355

Adekunle, K. F., & Okolie, J. A. (2015). A Review of Biochemical Process of Anaerobic Digestion. In Advances in Bioscience and Biotechnology (Vol. 06, Issue 03, pp. 205–212). Scientific Research Publishing, Inc. https://doi.org/10.4236/abb.2015.63020

Merlin Christy, P., Gopinath, L. R., & Divya, D. (2014). A review on anaerobic decomposition and enhancement of biogas production through enzymes and microorganisms. In Renewable and Sustainable Energy Reviews (Vol. 34, pp. 167–173). Elsevier BV. https://doi.org/10.1016/j.rser.2014.03.010

Gerardi, M. H. (2003). The microbiology of anaerobic digesters. John Wiley & Sons.

Al Seadi, T., Ruiz, D., Prassl, H., Kottner, M., Finsterwaldes, T., Volke, S. and Janssers, R. (2008) Handbook of Biogas. University of Southern Denmark, Esbjerg.

Aslanzadeh, S. (2014). Pretreatment of Cellulosic Waste and High Rate Biogas Production. Doctoral Thesis on Resource Recovery, University of Borås, Borås, 1-50.

Li, D., Liu, S., Mi, L., Li, Z., Yuan, Y., Yan, Z., & Liu, X. (2015). Effects of feedstock ratio and organic loading rate on the anaerobic mesophilic co-digestion of rice straw and pig manure. Bioresource technology, 187, 120-127.

Sun, M. T., Fan, X. L., Zhao, X. X., Fu, S. F., He, S., Manasa, M. R. K., & Guo, R. B. (2017). Effects of organic loading rate on biogas production from macroalgae: Performance and microbial community structure. Bioresource technology, 235, 292-300.

Montingelli, M. E., Tedesco, S., & Olabi, A. G. (2015). Biogas production from algal biomass: A review. Renewable and Sustainable Energy Reviews, 43, 961-972.

Ezekoye, V. A., Ezekoye, B. A., & Offor, P. O. (2011). Effect of retention time on biogas production from poultry droppings and cassava peels. Nigerian Journal of Biotechnology, 22, 53-59.

Schmidt, T., Ziganshin, A. M., Nikolausz, M., Scholwin, F., Nelles, M., Kleinsteuber, S., & Pröter, J. (2014). Effects of the reduction of the hydraulic retention time to 1.5 days at constant organic loading in CSTR, ASBR, and fixed-bed reactors–performance and methanogenic community composition. biomass and bioenergy, 69, 241-248.

Al-Addous, M., Alnaief, M., Class, C., Nsair, A., Kuchta, K., & Alkasrawi, M. (2017). Technical possibilities of biogas production from olive and date waste in Jordan. BioResources, 12(4), 9383-9395.

Hills, D. J. (1979). Effects of carbon: nitrogen ratio on anaerobic digestion of dairy manure. Agricultural wastes, 1(4), 267-278.

Chiu, M.-C., Wen, C.-Y., Hsu, H.-W., & Wang, W.-C. (2022). Key wastes selection and prediction improvement for biogas production through hybrid machine learning methods. In Sustainable Energy Technologies and Assessments (Vol. 52, p. 102223). Elsevier BV. https://doi.org/10.1016/j.seta.2022.102223

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). Machine learning and the physical sciences. In Reviews of Modern Physics (Vol. 91, Issue 4). American Physical Society (APS). https://doi.org/10.1103/revmodphys.91.045002

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, *549*(7671), 195-202.

Wang, R., Ye, S., Li, K., & Kwong, S. (2021). Bayesian network based label correlation analysis for multi-label classifier chain. *Information Sciences*, *554*, 256-275.

Vinothkumar, V., Muthukumaran, V., Rajalakshmi, V., Joseph, R. B., & Munirathnam, M. (2022). Efficient Data Clustering Techniques for Software-Defined Network Centres. In *Handbook of Research on Technologies and Systems for E-Collaboration During Global Crises* (pp. 201-217). IGI Global.

Chen, Yifu. (2022). Machine learning based approaches for classification of oil spills and microplastics in marine environments. Memorial University of Newfoundland. https://doi.org/10.48336/VTAJ-8C07

Alsubari, S. N., Deshmukh, S. N., Alqarni, A. A., Alsharif, N., Aldhyani, T. H., Alsaade, F. W., & Khalaf, O. I. (2022). Data analytics for the identification of fake reviews using supervised learning. *CMC-Computers, Materials & Continua*, *70*(2), 3189-3204.

Behbahani, R., Sarvestani, H. Y., Fatehi, E., Kiyani, E., Ashrafi, B., Karttunen, M., & Rahmat, M. (2022). Machine Learning-Driven Process of Alumina Ceramics Laser Machining. *arXiv preprint arXiv:2206.08747*.

Zuranski, A. M., Martinez Alvarado, J. I., Shields, B. J., & Doyle, A. G. (2021). Predicting reaction yields via supervised learning. *Accounts of chemical research*, *54*(8), 1856-1865.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51-62.

Bourguignon, M., & de Medeiros, R. M. (2022). A simple and useful regression model for fitting count data. *TEST*, 1-38.

Yamac, S. S. (2021). Reference evapotranspiration estimation With kNN and ANN Models using different climate input combinations in the semi-arid environment. *Journal of Agricultural Sciences*.

Mishra, S., Mallick, P. K., Tripathy, H. K., Jena, L., & Chae, G. S. (2021). Stacked KNN with hard voting predictive approach to assist hiring process in IT organizations. *The International Journal of Electrical Engineering & Education*, 0020720921989015.

Xiong, L., & Yao, Y. (2021). Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm. *Building and Environment*, *202*, 108026.

Ataş, M., Yeşilnacar, M. İ., & Demir Yetiş, A. (2022). Novel machine learning techniques based hybrid models (LR-KNN-ANN and SVM) in prediction of dental fluorosis in groundwater. *Environmental Geochemistry and Health*, *44*(11), 3891-3905.

Garg, A., Huang, H., Kushvaha, V., Madhushri, P., Kamchoom, V., Wani, I., ... & Zhu, H. H. (2020). Mechanism of biochar soil pore–gas–water interaction: gas properties of biochar-amended sandy soil at different degrees of compaction using KNN modeling. *Acta Geophysica*, *68*(1), 207-217.

Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, *24*(12), 1565-1567.

Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification (pp. 207-235). Springer, Boston, MA.

Tanveer, M., Rajani, T., Rastogi, R., Shao, Y. H., & Ganaie, M. A. (2022). Comprehensive review on twin support vector machines. *Annals of Operations Research*, 1-46.

Huang, F. L. (2022). Alternatives to logistic regression models in experimental studies. *The Journal of Experimental Education*, *90*(1), 213-228.

Alsubari, S. N., Deshmukh, S. N., Alqarni, A. A., Alsharif, N., Aldhyani, T. H., Alsaade, F. W., & Khalaf, O. I. (2022). Data analytics for the identification of fake reviews using supervised learning. *CMC-Computers, Materials & Continua*, *70*(2), 3189-3204.

Lee, J., Park, S., & Lee, J. (2022). Study on the Technology Trend Screening Framework Using Unsupervised Learning. *Applied Sciences*, *12*(17), 8920.

Verma, K. K., Singh, B. M., & Dixit, A. (2022). A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. *International Journal of Information Technology*, *14*(1), 397-410.

Su, H., Yang, X., Xiang, L., Hu, A., & Xu, Y. (2022). A novel method based on deep transfer unsupervised learning network for bearing fault diagnosis under variable working condition of unequal quantity. *Knowledge-Based Systems*, *242*, 108381.

Kubik, C., Knauer, S. M., & Groche, P. (2022). Smart sheet metal forming: importance of data acquisition, preprocessing and transformation on the performance of a multiclass support vector machine for predicting wear states during blanking. *Journal of Intelligent Manufacturing*, *33*(1), 259-282.

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., ... & Munigala, V. (2020, August). Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3561-3562).

Hameed, M., & Naumann, F. (2020). Data preparation: A survey of commercial tools. *ACM SIGMOD Record*, *49*(3), 18-29.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*(1), 259-268.

Salomon, D., & Motta, G. (2010). *Handbook of data compression*. London; New York: Springer,.

Gautam, R., Vanga, S., Ariese, F., & Umapathy, S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, *2*(1), 1-38.

Korableva, O. N., Kalimullina, O. V., & Mityakova, V. N. (2018, May). Innovation activity data processing and aggregation based on ontological modelling. In *2018 4th International Conference on Information Management (ICIM)* (pp. 1-4). IEEE.

Balakrishnan, N., Leiva, V., Sanhueza, A., & Cabrera, E. (2009). Mixture inverse Gaussian distributions and its transformations, moments and applications. *Statistics*, *43*(1), 91-104.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013, September). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the international conference recent advances in natural language processing ranlp 2013* (pp. 198-206).

Evans, R., & Grefenstette, E. (2018). Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, *61*, 1-64.

Mousavi, M., Bakar, A. A., & Vakilian, M. (2015). Data stream clustering algorithms: A review. *Int J Adv Soft Comput Appl*, *7*(3), 13.

Argelaguet, R., Cuomo, A. S., Stegle, O., & Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature biotechnology*, *39*(10), 1202-1215.

Pang, Z., Zhou, G., Ewald, J., Chang, L., Hacariz, O., Basu, N., & Xia, J. (2022). Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. *Nature Protocols*, *17*(8), 1735-1761.

Kim, B. J., Jeong, S., & Chung, J. B. (2021). Research trends in vulnerability studies from 2000 to 2019: Findings from a bibliometric analysis. *International Journal of Disaster Risk Reduction*, *56*, 102141.

Johri, A. (2011). The socio-materiality of learning practices and implications for the field of learning technology. *Research in Learning Technology*, *19*(3), 207-217.

Nie, B., & Sun, S. (2017). Using text mining techniques to identify research trends: A case study of design research. *Applied Sciences*, *7*(4), 401.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M. L., Rehman, M., Walther, T. C., ... & Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, *325*(5942), 834-840.

Hearst, M. A. (1999, June). Untangling text data mining. In Proceedings of the 37th Annual meeting of the Association for Computational Linguistics (pp. 3-10).

Kostoff, R. N. (1999). Science and technology innovation. *Technovation*, *19*(10), 593-604.

Losiewicz, P., Oard, D. W., & Kostoff, R. N. (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, *15*(2), 99-119.

Viator, J. A., & Pestorius, F. M. (2001). Investigating trends in acoustics research from 1970–1999. *The Journal of the Acoustical Society of America*, *109*(5), 1779-1783.

Ding, Z., Liu, R., & Yuan, H. (2021). A text mining-based thematic model for analyzing construction and demolition waste management studies. *Environmental Science and Pollution Research*, *28*(24), 30499-30527.

Trappey, A. J., Trappey, C. V., Wu, J. L., & Wang, J. W. (2020). Intelligent compilation of patent summaries using machine learning and natural language processing techniques. *Advanced Engineering Informatics*, *43*, 101027.

De Clercq, D., Wen, Z., & Song, Q. (2019). Innovation hotspots in food waste treatment, biogas, and anaerobic digestion technology: A natural language processing approach. *Science of the total environment*, *673*, 402-413.

Ranjbari, M., Saidani, M., Esfandabadi, Z. S., Peng, W., Lam, S. S., Aghbashlo, M., ... & Tabatabaei, M. (2021). Two decades of research on waste management in the circular economy: Insights from bibliometric, text mining, and content analyses. *Journal of Cleaner Production*, *314*, 128009.

Eldeeb, G., & Mohamed, M. (2022). Transit electrification state of the art: A machine-learning based text mining approach. *Transportation Research Part D: Transport and Environment*, *111*, 103446.

Dellosa, J. T., & Palconit, E. C. (2021, September). Artificial Intelligence (AI) in Renewable Energy Systems: A Condensed Review of its Applications and Techniques. In *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)* (pp. 1-6). IEEE.

Jha, S. K., Bilalovic, J., Jha, A., Patel, N., & Zhang, H. (2017). Renewable energy: Present research and future scope of Artificial Intelligence. *Renewable and Sustainable Energy Reviews*, *77*, 297-317.

Shin, W., Han, J., & Rhee, W. (2021). AI-assistance for predictive maintenance of renewable energy systems. *Energy*, *221*, 119775.

AlShabi, M., & Assad, M. E. H. (2021). Artificial Intelligence applications in renewable energy systems. In *Design and Performance Optimization of Renewable Energy Systems* (pp. 251-295). Academic Press.

Tao, J., Qiu, D., Yang, F., & Duan, Z. (2020). A bibliometric analysis of human reliability research. *Journal of Cleaner Production*, *260*, 121041.

Tollefson, J., Frickel, S., & Restrepo, M. I. (2021). Feature extraction and machine learning techniques for identifying historic urban environmental hazards: New methods to locate lost fossil fuel infrastructure in US cities. *Plos one*, *16*(8), e0255507.

Granata, F., Papirio, S., Esposito, G., Gargano, R., & De Marinis, G. (2017). Machine learning algorithms for the forecasting of wastewater quality indicators. *Water*, *9*(2), 105.

Inoue, A., Jiang, L., Lu, F., & Zhang, Y. (2017). Genomic imprinting of Xist by maternal H3K27me3. *Genes & development*, *31*(19), 1927-1932.

DŲeroski, S. (2009). Machine learning applications in habitat suitability modeling. In *Artificial intelligence methods in the environmental sciences* (pp. 397-411). Springer, Dordrecht.

Masih, A. (2019). Machine learning algorithms in air quality modeling. *Global Journal of Environmental Science and Management*, *5*(4), 515-534.

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, *9*(11), 2216-2225.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1-21.

CHAPTER THREE

SMART INVSTIGATION OF ARTIFICAL INTELIGENCE IN RENEWABLE ENERGY SYSTEM TECHNOLOGIES BY NATURAL LANGUAGE PROCESSING: INSIGHTFUL PATTERN FOR DECISION-MEKARES

3.1. ABSTRACT

This study aims to provide a framework which enables decision-makers and researchers to identify AI technology patterns in renewable energy systems from a massive data set of textual data. However, the study was challenged by the Scopus database limitation users to retrieve only 2000 documents per query. Therefore, we developed a search engine based on the Scopus application programming interface (API) that enables us to download an unlimited number of documents per query based on our desirable settings. The total number of 5661 renewable energy systems-related publications were extracted from Scopus database and Natural Language Processing (NLP) and unsupervised algorithms were leveraged to identify the most frequent computational science models and dense meta-topics and investigate their evolution throughout the period 2000-2021. The findings showed 7 meta-topics based on the class-based Term Frequency-Inverse Document Frequency (c-TD-IDF) score and term score decline graph. Emerging advanced algorithms, such as different deep learning architectures, directly impacted growing meta-topics involving problems with uncertainty and dynamic conditions.

Keywords: Natural language processing; Artificial intelligence; Text mining; Topic modeling; Pattern identification; Renewable energy.

3.2. INTRODUCTION

With the ever-increasing demand for energy, limitations of fossil fuel resources, and concerns about sustainability, renewable energy systems are increasingly gaining attention from governments, businesses, and research institutes worldwide. Hence, developing a clear technology roadmap is important to integrate science and technology with business planning meaningfully based on medium to long-term market direction and goals (Amer and Daim, 2010). Performing intelligence investigation on upcoming technologies and clear technology roadmaps will assist governments and industries in making smart investing decisions and maintaining their competitive edge (Angelo et al., 2017). This study focuses on the research direction of applying AI text modeling techniques for identifying the most common strategies employed by renewable energy systems literature, aiming at establishing a technology selection and research perspective in this domain. The significant growth of the application of AI modeling in the renewable energy domain creates a massive amount of data and information within research papers, registered patents, reports, etc.

In this regard, some researchers have focused on leveraging AI in biogas systems. Tufaner and Demirci (2020) employed a three-layer artificial neural network on lab-scale data to forecast biogas production rate by considering different features such as effluent alkalinity, organic loading rate, effluent chemical oxygen demand, etc. Chiu et al. (2022) applied a hybrid machine learning model, random forest, and long short-term memory by analyzing important feeds for biogas production for optimization based on a biogas plant dataset in China. They gained significant performance without conducting intensive feature engineering.

Table 3.1 presents recently published similar research leveraging various AI models for optimizing and forecasting biogas systems' performance.

Table 3.1. Previous scientific works at the intersection of AI and biogas.

Objective	Model(s)	Result	Reference
Optimizing biogas purification	DNN	The optimum ranges of: C/N (15.04–18.95), BOD/COD (0.763–0.818), TS (8.1–10.6%) and T.VS (38.19–49.46%). Large BOD/COD impacts biogas purification. pH > 7 can improve biogas purification.	Mahmoodi-Eshkaftaki and Ebrahimi, 2021
Predicting biogas production rate of food waste dry anaerobic digestion considering HRT, SRT, soluble chemical oxygen demand, total VFA, total and ammonia features.	RNN	Solid retention time and water content are important features in biogas reactions. Increasing intermediate materials, like VFAs, were easily converted into methane at higher water contents.	Seo et al., 2020
Modeling chemical processes within biogas production system	SNN	Considering ten days data points, model can predict chemical process up to the 100th day with significant accuracy based on lab-scale data.	Capizzi et al., 2020
Predicting biogas production of fruits and vegetables waste considering different operational parameters	ANN	Predicted the performance with 85% accuracy.	Gonçalves et al., 2021
Identifying important operational parameters and predicting the biogas systems production rate	RF KNN SVM GLMNET	Total carbon was identified as a most important feature. KNN performed well in the regression task with RMSE of 26.6 and logestic regression multiclass model gained accuracy of 73%.	Wang et al., 2020
Predicting biogas systems performance of vegetables, fruits waste	ANFIS LSSVM	LSSVM performed better than ANFIS.LSSVM had MRE % and MSE of 2.951 and 0.0001 respectively, compared to 29.318 and 0.0039 for ANFIS.	Yang et al., 2021
Predicting methane production in a biogas plant	Gompertz ML HML and Gompertz	HML was the best in predicting next-day biogas production and reduced the error by 53%. MAPE of HM (4.52%) < ML (4.84%) < the Gompertz model (9.61%).	Hansen et al., 2020
Building a predictive model of biogas yield and establishing optimal conditions for cow manure and maize straw biogas process	ANFIS	$R^2 = 0.99$ and the model suggested conditions increased the production by 8%.	Zareei and Khodaei, 2017

Table 3.1. Continued.

Objective	Model(s)	Result	Reference
Predicting biogas production of spent mushroom compost in thermophilic and mesophilic laboratory conditions	ANN ANFIS	RMSE and R^2 in mesophilic condition: ANFIS are 0.1940 and 0.9998, ANN are 0.780 and 0.9981, and logistic model are 0.5111 and 0.9992, respectively. In thermophilic condition, the values of RMSE and R^2 were indicated as 0.3033 and 0.9997 for ANFIS, 0.3430 and 0.9992 for ANN, and 0.5506 and 0.9991 for the logistic model, respectively.	Najafi and Faizollahzadeh Ardabili, 2018
Predicting biogas production rate based on industrial data and finding important operation parameters	ACO GA ANN	$R^2 = 0.9$ and prediction error = 6.24%.	Beltramo et al., 2019

Note: DNN = Deep neural networks; RNN = Recurrent neural network; SNN = Spiking neural network; ANN = Artificial neural networks; RF = Random forest; KNN = K- nearest neighbour; RMSE = Root mean square error; mean relative error = MRE ; Mean squared error = MSE; Mean absolute percentage error = MAPE; SVM = Support vector machine; GLMNET = Generalized linear models fitting package via penalized maximum likelihood ; LSSVM = Least square support vector machine; ML = Machine learning; HML = Hybrid of machine learning and Gompertz; ANFIS = Adaptive neuro-fuzzy inference system (ANFIS); ACO = Ant colony optimization (ACO); GA = Genetic algorithms (GA).

Like other domains, a problem that investors and decision-makers in governmental and industrial sectors face is that they can get confused among various generated information, specifically in the field of AI with a wide variety of techniques and approaches. Developing an almost automatic framework that can be employed to address the mentioned problem is crucial. More specifically, such a method can provide accurate insights quickly for various research domains, even complex areas, to empower decision-makers to better perceive the research dynamics and assist them in determining research and development (R&D) strategies. This study aimed to build a research landscape and give decision-makers and researchers in related domains an accurate insight into implementing AI algorithms in renewable energy systems from a scientific standpoint. There is a high level of correspondence between scientific papers and patents; in other words, patent quality can be measured by academic papers referenced (Poege et al., 2019; Coupé,

2003). Besides, the World Intellectual Property Organization specified that more than 90% of inventions are observed in patent papers (Souili et al., 2015). Hence, texts of scientific papers are beneficial resources for investigating technology development throughout a timeline and understanding it will benefit research and development. However, analyzing the vast textual literature and data is time-consuming and prone to mistakes. A powerful solution for this problem is using NLP techniques (Chowdhary, 2020). NLP techniques have various applications, such as information extraction, automated text summarization, question-answering systems, and speech recognition (Jusoh, 2018). Topic modeling is an unsupervised strategy that is a subfield of NLP, which is a valuable technique for achieving a high level of understanding of a large amount of unstructured text data (Hannigan et al., 2019). The basis of NLP is considering co-occurrences of a word in similar corpora (Daenekindt and Huisman, 2020). Co-occurrence rules empower machines to discover and group related concepts within the set of documents or records. This idea implies that when concepts are often found together in documents and records, that co-occurrence illustrates a hidden relationship that is probably of value in categorizing definitions. Various topic modeling methods have emerged in previous years, such as Latent Dirichlet Allocation (LDA), the most frequently leveraged algorithm in topic modeling (Jockers and Thalken, 2020). Also, there are other less frequent methods like Bidirectional Encoder Representations from Transformers Topic *(*BERTopic) (Grootendorst, 2021), Top2Vec (Angelov, 2020a), Structural Topic Modelling (STA) (Lindstedt, 2019), Correlation Explanation (CorEX) (Gallagher et al., 2017), Non-Negative Matrix Factorisation (NMF) (Wei et al., 2003), and Latent Semantic Analysis (LSA) (Landauer et al., 1998).

Previous research applied NLP models to scientific papers for topic modeling and trend identification. Mosallaie et al. (2021) employed NLP techniques to investigate the application of

AI in scientific papers on cancer-related domains. Tran et al. (2019) employed LDA in scientific papers between 1991 and 2018 to perfume topic modeling and provided static insight into the pattern of AI in the cancer domain. Jallan et al. (2019) applied the LDA model to detect current patterns in "construction-defect litigation cases." Lee et al. (2019) performed an accurate NLP-based model to extract contract risk, systematically detecting "poisonous" terms to help related companies manage their contracts. It performed well, achieving 81.8% area under the precise recall curve. Other studies utilized NLP approaches for technology forecasting. Kyebambe et al. (2017) clustered similar technologies considering patent characteristics and predicted new technologies one year forward. Lee et al. (2018) investigated the value of patents by leveraging feed-forward artificial neural networks and performed an evaluation analysis system for emerging technologies. Johri et al. (2011) applied topic modeling techniques to investigate emerging topics in engineering education between 2000 and 2008. More specifically, they leveraged LDA, an unsupervised learning method (Blei et al., 2003), as a topic modeling method to extract topics and their top 20 keywords correspondence in engineering education. They also extracted key phrases and their corresponding frequency values to quantitatively analyze their trends. Based on their results, some topics, like the "global and interaction aspect of engineering education," observed a considerable increase, while other topics remained almost constant over the period. Other researchers leveraged topic modeling in other domains like transportation (Sun and Yin, 2017), hydropower (Jiang et al., 2016), communication research (Maier et al., 2018), smart factory (Yang et al., 2018), marketing (Reisenbichler and Reutterer, 2019). To the best of the author's knowledge, this study is the first to comprehensively investigate the trend of AI topics in the renewable energy systems domain using NLP techniques. Although some studies appeared on trend analysis with limited scope to

only one area, such as solar energy or anaerobic digestion, they did not employ NLP techniques (Dong et al., 2012; Ren et al., 2018).

3.3. METHODOLOGY

Figure 3.1 presents the methodology employed in this study. First, a dataset has been built by collecting raw data from Scopus. This raw dataset contains all renewable energies scientific papers in which AI modeling has been used for 2000-2021. The collected dataset has been preprocessed within three steps specified in the "Preprocessing dataset" section. The next step is conducting different exploratory analyses on the preprocessed dataset. The preprocessed dataset was used as input for the BERTopic model to generate topics. Finally, the created topics have been merged into dense meta-topics by domain experts, and their evolution has been investigated over time by the dynamic topic modeling (DTM) method.

3.3.1. Collecting raw dataset

Scopus database has been chosen since its one of the most comprehensive databases for published papers. The Scopus database contains more than 22000 journals and books in renewable energy and computer science sectors, more specifically, 335 journals and 6699 books for renewable energies, and 1337 journals and 14111 books for computer science. The dataset has been extracted from the Scopus database by developing a search engine in Python and querying using keywords such as "Artificial Intelligence," "Machine Learning," "Deep Learning," and "Neural Network," in addition to "Renewable Energy," and "Green Energy" within the period 2000-2021. This study has focused on journals, conference papers, books, and book chapters in English. Scopus allows users to receive only 2000 per query. However, the developed search engine can retrieve the required data without any limitations. The developed search engine breaks

the dataset to a number of chunks; each can contain up to 25 documents. In this case, there are 226 chunks containing 25 documents and one chunk containing 11 documents. Another common method for data acquisition is employing SQL queries to search related keywords (Venugopalan and Rai, 2015; De Clercq et al., 2019). The extracted data and developed search engine are accessible on KamranNiroomand's GitHub account at: https://github.com/KamranNiroomand/Scopus-Search-Engine.git. The developed search engine can be used in future research, using the Scopus database, in various areas such as scientiometrics, intelligent decision support systems, etc.

3.3.2. Pre-processing dataset

The extracted raw dataset was then prepared for the all-MiniLM-L6-v2 algorithm, a sentence-based pre-trained model. Three features have been selected for this study: data, title, and abstract. Since each title has useful information about the context, the title, and abstract have been merged into a new feature for the analysis. Besides, the BERTopic model has been employed, and unlike other topic modeling algorithms, it does not require an intensive preparation BERTopic (Grootendorst, 2021). Given the nature of BERTopic, the sentence's primary structure is necessary (Egger and Yu, 2022). Finally, the required features have been converted from a "string" type to a "list" type for the next steps. However, conventional models like LDA need comprehensive data preparation steps like removing stop words, lemmetazion, tokenizing, etc. (Kadhim et al., 2014).

3.3.3. Data analysis

3.3.3.1. Exploratory data analysis (EDA)

Before building the BERTopic model, several data exploratory analyses have been conducted, such as the rate of renewable energy systems publications in which AI has been leveraged throughout the period to investigate the trend of AI in renewable energy systems (Figure

3.2). In Figure 3.2., the rate has been measured by dividing the number of publications that employed AI by the total number of publications in the renewable energies domain. Additionally, the most frequent modeling approaches used in renewable energy research have been detected and depicted over the specified period (Figure 3).

3.3.3.2. BERTopic model

This study employed BERTopic modeling algorithm (Grootendorst, 2020) to extract topics at the intersection of AI modeling and renewable energy systems and investigate them throughout 2000-2021. BERTopic is built based on Top2Vec (Angelov, 2020) and is an embedding-based model. The BERTopic model has been built in three steps. The first step vectorized the textual dataset to group close semantical terms (Egger, 2022) using a pre-trained sentence-based transformer algorithm (Reimers and Gurevych, 2019). Then, due to the high degree of sparsity within the generated vectors, a uniform various approximation and projection (UMAP) were utilized (McInnes et al., 2018) to reduce the dimensionality of vector space and keep global and local data structures. The vector space was reduced to 20 dimensions to create dense regions and employed Hierarchical Density-Based Spatial Clustering of Applications with Noise algorithm (hDBSCAN) (Campello et al., 2013; McInnes and Healy, 2017) as a clustering measure to identify these areas in the documents. Finally, this study considered the c-TD-IDF algorithm to create topics, where documents in a cluster are considered one document. Then TD-IDF score is calculated to show the importance of each word in a cluster. We considered the default parameters and algorithms of the BERTopic since they gave us the best results. For example, We tried different amounts of “n_neighbors” for UMAP algorithm, 15, which is the default value, 18, and 13. The result was almost the same. Also, we tried RoBERTa, a BERT-based model, instead of the all-

MiniLM-L6-v2 algorithm, which is the default algorithm of BERTopic, but the results were not satisfying. The TD-IDF (Joachims, 1996) can be calculated using Eq. 3.1:

$$W_{t,d} = tf_{t,d}.\log\left(N/df_t\right) \tag{3.1}$$

where the term frequency, $f_{t,d}$, models the frequency of term *t* in document *d*. The inverse document frequency indicates the amount of information that a term gives to a document and is measured by the logarithm function of the number of documents in a corpus that is denoted by *N* divided by the total number of documents that include *t*. The adjusted class-based TD-IDF (Grootendorst, 2022) can be expressed as in Eq. 3.2:

$$W_{t,c} = tf_{t,c}\,.\log\left(1 + \mathrm{A}/tf_t\right) \tag{3.2}$$

where the term frequency, $f_{t,c}$, models the frequency of term *t* within a class *c*. The higher the value of the c-TD-IDF, the more important the words in the clusters are. In addition, we merged topics to reduce their number from 98 to 7 dense and semantic meta-topics considering the result of the hierarchical clustering measured by the cosine distance matrix between clusters (Figure 3.4) and the domain expert's knowledge. Because of the limited capability of quantifying algorithms to provide sufficient contextual comprehension (Egger and Yu, 2022), topic modeling interpretation certainly needs human judgment (Hannigan et al., 2019) and domain expert knowledge (Egger and Yu, 2022). For the domain expert to choose interpretable labels for meta-topics, the top ten words of each meta-topic (Figure 3.5) and the "Term Score" graph (Figure 3.6), which shows the number of words representing each topic, have been taken into account. This study also includes DTM in which the identified topic fluctuation throughout the period has been investigated and evolved to ascertain emerged technologies by considering unigram (one world). Blei and Lafferty (2006) first employed the DTM method, built upon LDA, to solve the static concept of topic modeling. One of the BERTopic model functions is performing DTM based on

c-TD-IDF by creating global topics without considering their temporal nature (Grootendorst, 2022). To implement this, it was fitted to the whole textual dataset to create a global view of topics. Following that, local topics' representation can be generated by Eq. 3.3:

$$W_{t,c,i} = tf_{t,c,i} \,.\log\,(1 + \mathrm{A}/tf_t) \tag{3.3}$$

where the documents' term frequency is multiplied at timestep "*i*" considering the pre-calculated value of global IDF.

3.3.4. Computation

This study's programming parts was developed using the Python 3.10.2 language within the Google Colab notebook environment.

3.4. RESULTS AND DISCUSSION

3.4.1. Exploratory Data Analysis (EDA)

The proportion of leveraging AI toward renewable energy research is illustrated in Figure 3.2. Renewable energy research has experienced an increasing trend over the considered period. It is noticeable that from 2016 the slope of the graph has increased, which shows the power of AI modeling in this sector. However, the slope became almost flat in 2019-2020, likely due to the impact of the COVID-19 pandemic and its associated restrictions (Harper et al., 2020). During that year, most of the research was focused on treating methods for covid-19 or the effects of this virus on different aspects of our life (Verma et al., 2020; Herrera (2020)). Afterward, it started to grow with a sharper slope from 2020; the proportion of renewable energy research that employed AI peaked at 0.075 of total research within this domain, nearly double its 2015 value.

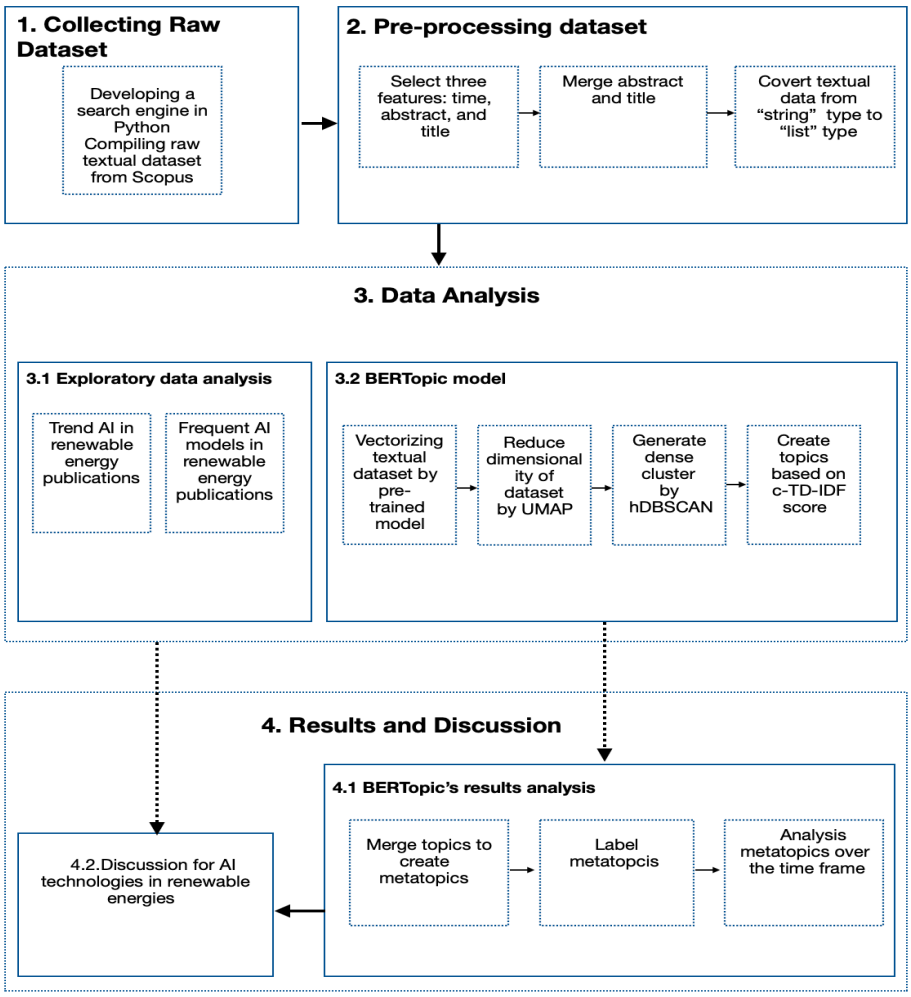

Figure 3.1. Schematic of the methodology

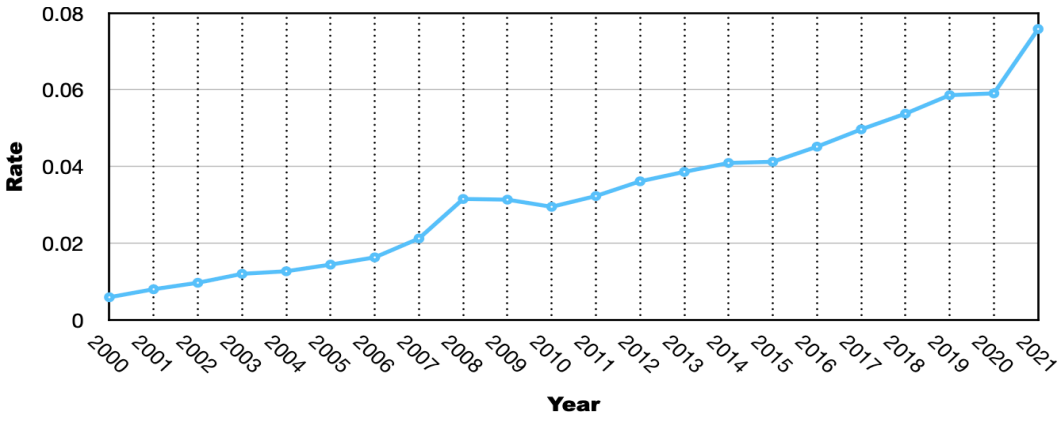

Figure 3.2. Trend of AI modeling in renewable energy systems publication.

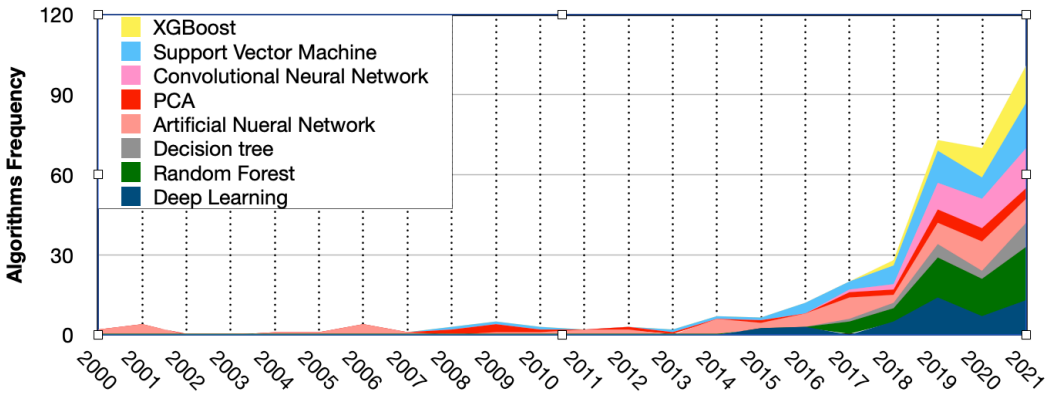

Figure 3.3. Most frequent computational algorithms in renewable energy systems

This study detected and investigated the most frequent computer science methods employed in renewable energy research over the specified period. The artificial neural network has been used from the beginning of the period (Figure 3.3). However, with time, researchers started to use other algorithms as well. For instance, from 2007, supervised and unsupervised methods using principal components analysis and support vector machine, respectively, emerged in renewable energy-related works. Gradually, various models appeared, such as deep learning methods in the research domain capable of handling dynamic situations like wind and wave energies with a high performance (Gu and Li, 2022). Also, deep learning algorithms, especially deep reinforcement learning) algorithms are utilized for complex renewable energy problems such as smart grid systems with uncertainty and nonlinearity and/or large structure datasets (Widodo et al., 2021). By 2016, the decision tree emerged alongside the random forest. The latter is an ensemble learning method with several advantages over the decision tree (Ahmad et al., 2018); thus, its usage rate exceeded that of the decision tree in 2016 and increased significantly throughout the rest of the period. Besides, XGBoost, a boosting model based on a decision tree, appeared in 2017 and grew

considerably from 2018 to 2021. Likewise, deep learning and convolutional neural network emerged in 2015 and 2016, respectively, and grew noticeably until the end of 2021.

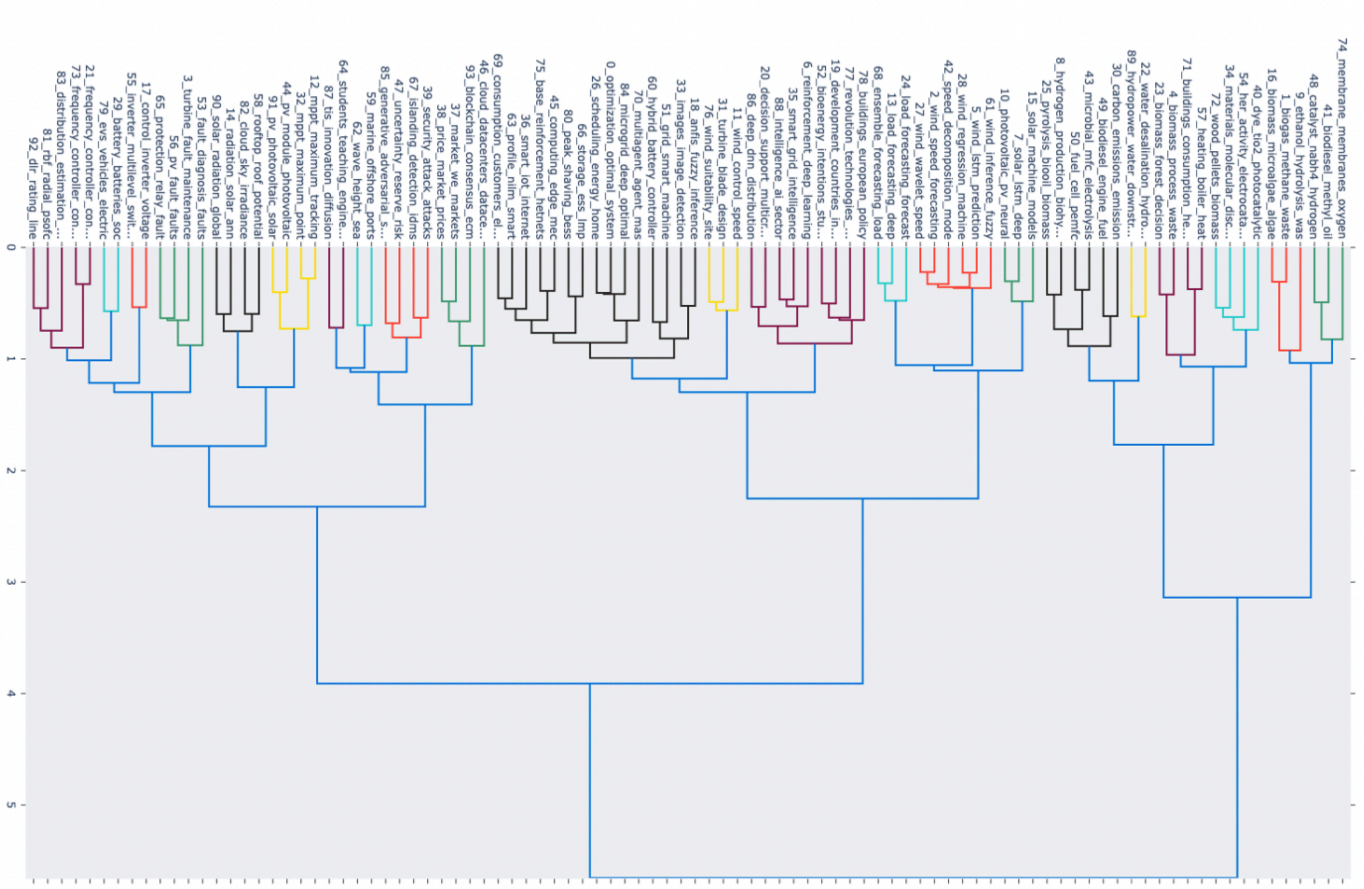

Figure 3.4. Hierarchical clustering to decrease the number of topics

3.4.2. Technology direction based on c-TD-IDF

Using the BERTopic model, similar topics have been merged due to clustering algorithms and a hierarchical graph. For instance, topics 12, 32, 44, 91, 58, 82, 14, 90, 10, 7, and 15 overlap, covering different angles of "solar energy as a renewable energy resource." Therefore these topics have been merged, and seven meta-topics (Prediction in Wind Systems, Power Systems Optimization, Biogas, Wind Turbine Fault Detection, Bio Hydrogen, Solar Energy and

Photovoltaic Cells, Biomass) have been identified. Figure 3.5 shows the top 10 words corresponding to each meta-topic based on their c-TD-IDF scores. Domain experts have chosen a human-interpretable label for each topic, considering the top-ten words of each meta-topic in terms of c-TD-IDF (Figure 3.5) and term rank graph (Figure 3.6). We had three domain experts, two professors from environmental engineering and one process engineering professor. Their opinions on how to merge topics to create meta-topics and how to choose a name for each meta-topic were almost in line with each other. To choose the interpretable name for each meta-topic, they rewrote the most useful keywords meaningfully. For example, Figure 3.5 shows all ten words of meta-topic 5 have a high level of relevancy to their cluster. Figure 3.5 also illustrates that meta-topic 4 includes words with high scores of c-TD-IDF, such as "turbine," "fault," "wind," and "detection," implying that this meta-topic is related to wind turbine fault detection. Besides, meta-topic 3 contains "biogas," "methane," "anaerobic," and "waste," illustrating that this meta-topic should be related to leveraging AI modeling in the biogas process. Words of Meta-topic 2 that have top ranks in terms of c-TD-IDF score contain "optimization," "microgrid," and "algorithm," implying power systems optimization by leveraging AI modeling techniques like deep reinforcement learning (Domínguez-Barbero et al., 2020; Ji et al., 2019). The seven identified meta-topics are 1) Prediction in Wind Systems, 2) Power Systems Optimization, 3) Biogas, 4) Wind Turbine Fault Detection, 5) Bio Hydrogen, 6) Solar Energy and Photovoltaic Cells, and 7) Biomass. One of the powerful features of BERTopic is that it does not need the preprocessing of raw data since it is an embedding-based algorithm, but this generates a large number of topics and merging them needs knowledge and significant attention.

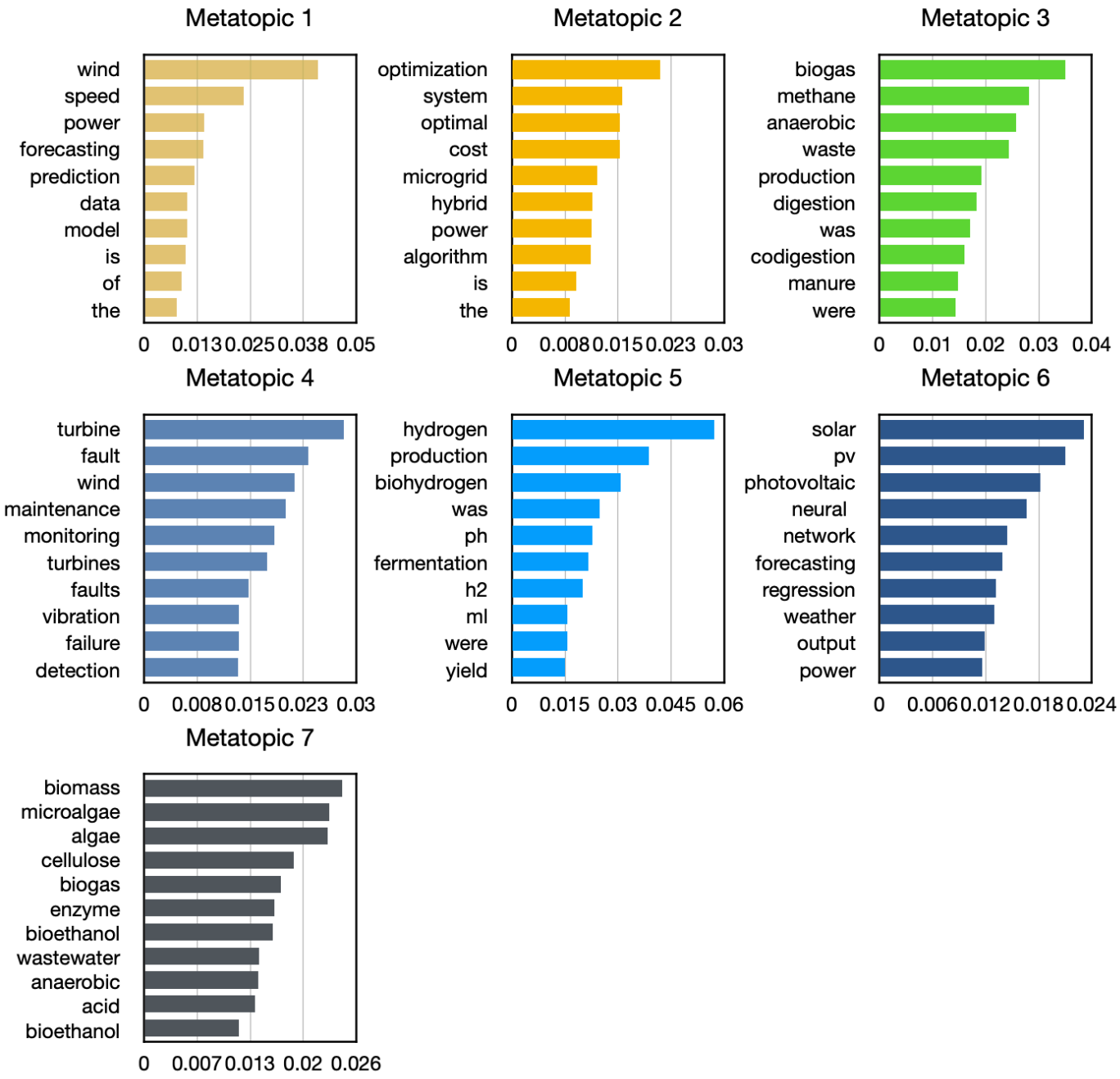

Figure 3.5. c-TD-IDF for each term in 7 identified meta-topics.

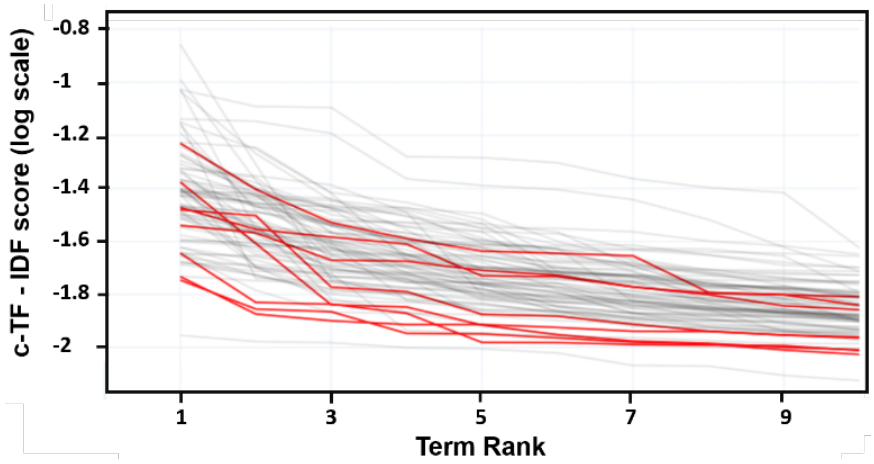

Figure 3.6. Term score decline per topic.

Figure 3.7. shows the evolution of identified meta-topics from 2000 to 2021. AI modeling emerged in renewable energy research in 2005. Afterward, leveraging computational algorithms increased at a similar rate in different types of renewable energy systems. The year 2014 was a turning point when researchers began to utilize more advanced techniques, such as different deep-learning algorithms. From 2014, the prediction in complicated systems like wind systems, solar energy and photovoltaic cells, and power systems optimization meta-topics experienced considerable growth. This aligns with our statement about implementing different deep learning algorithms, such as deep reinforcement learning, long short-term memory, and convolutional neural networks for dynamic and uncertain systems. Wind systems have been found to be a dominant and most attractive technology where deep learning models have been applied for wind speed prediction (Noorollahi et al., 2016; Moustris et al., 2016; Wang et al., 2015; Huang et al., 2021; Yeghikian et al., 2021; Shamshirband et al., 2019) and wind power prediction (Wu et al., 2016; Zameer et al., 2017; Dong et al., 2017). State-of-the-art AI techniques are vital in optimizing and controlling photovoltaic and solar energy technologies (Ghannam et al., 2019) and making them more economical (Youssef et al., 2017). For instance, photovoltaic systems damages can be

detected by deep convoluntinal neural network (Pierdicca et al., 2018), and their energy can be forecasted by a novel deep learning architecture (Abdel-Basset et al., 2021).

Similarly, deep learning has various applications in power systems, such as online energy scheduling (Ji et al., 2021), power systems resilience improvement (Kamruzzaman et al., 2021), and adaptive power system emergency control (Huang et al., 2020). Biogas is generated from wastethrough anaerobic digestion technology mainly for producing renewable energy and valorizing organic residues (Kougias and Angelidaki, 2018; Appels et al., 2011). Chen et al. (2017) mentioned that the optimal design of biogas plants depends on local conditions, namely, substrate supply and local infrastructure. Hence, developing machine learning/ deep learning algorithms that can determine optimal conditions for industrial reaction, the value of each feedstock (Chiu et al., 2022), and predicting the outcome of biogas reaction is necessary from an economic point of view. In other words, improving biogas plant economic sustainability will decrease operational costs. Considering the increasing importance of industrial biogas facilities, the significant growth of available data in this, and the capability of AI-based techniques in enhancing the economic sustainability of biogas facilities and, therefore, this area will probably emerge as one of the main prevalent domains of renewable energy systems. Biohydrogen is one the most environmentally friendly fuels since its combustion produces H2O as a carbon-free by-product and is also generated under anaerobic conditions and without consuming fossil fuels (Brentner et al., 2010). However, this process has several challenges, such as the cost of hydrogen as fuel, infrastructure facilities, distribution, and storage, etc. (Kamaraj et al., 2019). Despite optimizing AI modeling in biohydrogen systems (Wang et al., 2021; Khaleghi et al., 2021; Lian et al., 2021; Liu et al., 2021), these challenges showed their impact on the prevalence of biohydrogen topic. Continuous research and collaboration of data scientists and engineers would improve bio hydrogen technology and

utilize it in hydrogen-based vehicles (Manoharan et al., 2021). Regarding biomass, alongside forecasting biomass characteristics, process outcome, and performance of bio-energy end-use systems, one of the primary usages of AI is to generate synthetic datasets (Liao, and Yao, 2021). Data augmentation will increase the amount of labeled data and enhance performance of supervised machine learning algorithms (Bowles et al., 2018). The generative datasets in biomass are particularly useful regarding biomass properties, biofuel properties, kinetic parameters, engine performance, and Life Cycle Inventory (LCI) (Liao and Yao, 2021). Considering the fast pace of development of AI as well as growing sustainable systems data scientists, it can be predicted that analytic measures and AI-integrated models would significantly overcome the challenges like lack of sufficient high-quality data in the biomass domain, which leads to comprehensive assessment and optimization of biomass systems. The AI-based feature of this study that extracts topics automatically diminishes subjectivity from the exercise and enables consistent and comprehensive comparisons between topics and between time intervals. Seven extracted meta-topics provide comprehensive and logical coverage of the research field and have a high level of similarity to the topics being used to produce review paper studies (Lateef et al., 2022) or scientometirc research (Sohail et al., 2022) in renewable energy domains. DTM is a probabilistic time-series-based model capable of analyzing the evolution of topics over timeframes (Blei and Lafferty, 2006). DTM has various applications in different disciplines. For instance, Ayele and Juell-Skielse (2020) investigated the evolution of self-driving cars' topics and trends from 2000 to 2019. They chose DTM since, unlike LDA, DTM considers the temporal aspect of each topic (how topics have evolved over the period). Their results illustrated the evolution of twenty topics in the self-driving car domain, including software system architecture and design, brake system and safety and navigation in self-driving. Lee et al. (2016) applied dynamic topic modelling to toxicogenomics

data. It was used as an alternative technique to discover underlying patterns in time-series gene expression profiles which results in gaining a perception of the dynamic behavior of genes in the related systems. Besides, Morimoto and Kawasaki (2017) utilized dynamic topic modeling as a forecasting method for financial market volatility, which enhanced forecasting accuracy. Linton et al. (2017) used dynamic topic modeling into cryptocurrency community forums to investigate the evolution of different topics related to big events in the cryptocurrency society. Tabassum et al. (2021) used dynamic topic modeling to analyze social media by focusing on hashtags. Guldi (2019) applied a dynamic topic modeling algorithm to build a history record of British infrastructure.

Because the gap between scientific literature and commercialization is narrowing, text mining of academic papers can play a significant role for stakeholders intending to invest in renewable energy sectors and data-driven start-ups working on renewable energy systems. Notably, the most realistic technology read mapping with a broad scope can be achieved by validating the findings of text mining with the results of existing scientific publications and patent documents under the supervision of domain experts.

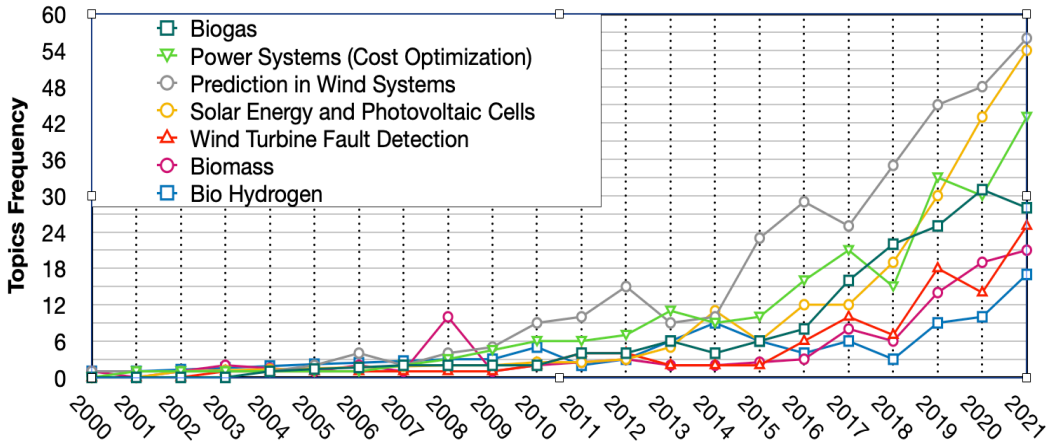

Figure 3.7. Topics evolution over the period

This current study is limited to scientific publications from the Scopus database from 2000 to 2021 to characterize the landscape of AI in renewable energy systems research. Future studies could consider other databases, such as the Web of Science, PubMed, IEEE Xplore, ScienceDirect or other data sources, such as patents, to perform complementary exploration. Besides, in the current study, we only analyzed uni-grams. Future studies can also consider the whole body of the paper for their analysis; particularly, the methodology section should contain informative text regarding methodological evolution. Additionally, we only considered English documents, and future studies can expand their scope by analyzing published documents in other languages and using other NLP techniques, such as GPT3 and WuDao 2.0, for their analyses and their results can be compared to current study's results.

3.5. SUMMARY

In this study, a capable search engine based on Scopus API has been developed, which can break the limit of the Scopus database and retrieve unlimited documentation per query. Besides, the study investigated AI in renewable energy systems publications throughout the 21's century from two points of view. First, by analyzing term frequency, the study investigated the trend of most frequent computational algorithms in renewable energy systems papers. The result showed that, within recent years, researchers started leveraging more diverse and complicated methods like deep learning techniques. However, still, conventional models like random forest are more popular. the study also uncovered the latent research topics of AI in renewable energy systems and considered their temporal aspect of them by employing the DTM method. Our DTM's results demonstrated the role of AI-based models in overcoming uncertainty and enhancing risk management in solar and wind systems (Tawn and Browell, (2022)). Besides, generated results reveal particular attention to modeling and simulation research projects. Matching and comparing

the results of both employed computational models and DTM analyses demonstrate that recent advancements in computational science have built new pathways for complicated renewable energy systems problems. Some examples can be energy storage optimization of wind, solar, and photovoltaic systems (Abualigah et al., 2022), monitoring and anomaly detection of wind turbines (Xiang et al., 2022), etc.

This feature that most steps of the proposed pipeline are automatic is significant. In other words, insights can be produced quickly, even for complex research fields involving a large number of papers annually, to assist policymakers and decision-makers in understanding the research dynamics better and help with research and development (R&D) strategies. This is of particular importance for R&D that needs high pace progress (Ebadi et al., 2022), for instance, disruptive technology development that can affect strategic stability (Sechser et al., 2019), national security (Ebadi et al., 2022), and economic development (Rifkin, 2011). Therefore, organizations that can better understand and monitor the research landscape will have a competitive edge.

3.6. CONCLUSIONS

The research characterized and mapped AI applications in renewable energy systems by leveraging a text mining technique to build semantic and dense structure clusters in a semantically continuous space. It used an unstructured dataset of 5661 scientific works between 2000 and 2021. This work comprehensively identified technologies at the intersection of AI and renewable energy systems and enhanced previous scientific works in pattern detection by leveraging novel algorithms for the NLP and the BERTopic method, resulting in a high level of coherency and efficiency. More specifically, BERTopic does not require time-consuming parts such as intensive data preprocessing and hyperparameter tuning to analyze textual datasets. The study investigated

technology trends by considering c-TD-IDF, annual analyses of topic evolution, most frequent AI algorithms, and the number of renewable energy systems papers that employed AI. BERTopic showed seven dense meta-topics covering all aspects of various renewable energy systems. The c-TD-IDF score and term score decline graphs provide insightful information to discriminate meta-topics and label them interpretably. For instance, meta-topics 7 is associated with terms such as biomass, microalgae, algae, and cellulose, which gained the highest c-TD-IDF. Additionally, the term score decline graph proves the previous statement as the first four terms of this meta-topics are reliable for labeling. The analysis showed that the development of AI modeling significantly impacted areas associated with high uncertainty, such as wind and microgrid systems. Future studies can expand their scope by considering other databases like US Patent or Science Direct and considering published documents in other languages. Also, they can analyze bi-gram and tri-gram instead of considering only uni-gram. Besides, instead of focusing only on the abstract and title, they can consider the whole body of the paper.

REFRENCES

Amer, M., & Daim, T. U. (2010). Application of technology roadmaps for renewable energy sector. In Technological Forecasting and Social Change (Vol. 77, Issue 8, pp. 1355–1370). Elsevier BV. https://doi.org/10.1016/j.techfore.2010.05.002

Angelo, A. C. M., Saraiva, A. B., Clímaco, J. C. N., Infante, C. E., & Valle, R. (2017). Life Cycle Assessment and Multi-criteria Decision Analysis: Selection of a strategy for domestic food waste management in Rio de Janeiro. In Journal of Cleaner Production (Vol. 143, pp. 744–756). Elsevier BV. https://doi.org/10.1016/j.jclepro.2016.12.049

Tufaner, F., & Demirci, Y. (2020). Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models. In Clean

Technologies and Environmental Policy (Vol. 22, Issue 3, pp. 713–724). Springer Science and Business Media LLC. https://doi.org/10.1007/s10098-020-01816-z

Chiu, M.-C., Wen, C.-Y., Hsu, H.-W., & Wang, W.-C. (2022). Key wastes selection and prediction improvement for biogas production through hybrid machine learning methods. In Sustainable Energy Technologies and Assessments (Vol. 52, p. 102223). Elsevier BV. https://doi.org/10.1016/j.seta.2022.102223

Mahmoodi-Eshkaftaki, M., & Ebrahimi, R. (2021). Integrated deep learning neural network and desirability analysis in biogas plants: A powerful tool to optimize biogas purification. In Energy (Vol. 231, p. 121073). Elsevier BV. https://doi.org/10.1016/j.energy.2021.121073

Seo, K. W., Seo, J., Kim, K., Ji Lim, S., & Chung, J. (2021). Prediction of biogas production rate from dry anaerobic digestion of food waste: Process-based approach vs. recurrent neural network black-box model. In Bioresource Technology (Vol. 341, p. 125829). Elsevier BV. https://doi.org/10.1016/j.biortech.2021.125829

Capizzi, G., Lo Sciuto, G., Napoli, C., Woźniak, M., & Susi, G. (2020). A spiking neural network-based long-term prediction system for biogas production. In Neural Networks (Vol. 129, pp. 271–279). Elsevier BV. https://doi.org/10.1016/j.neunet.2020.06.001

Gonçalves Neto, J., Vidal Ozorio, L., Campos de Abreu, T. C., Ferreira dos Santos, B., & Pradelle, F. (2021). Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN). In Fuel (Vol. 285, p. 119081). Elsevier BV. https://doi.org/10.1016/j.fuel.2020.119081

Wang, L., Long, F., Liao, W., & Liu, H. (2020). Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning

algorithms. In Bioresource Technology (Vol. 298, p. 122495). Elsevier BV. https://doi.org/10.1016/j.biortech.2019.122495

Yang, Y., Zheng, S., Ai, Z., & Jafari, M. M. M. (2021). On the Prediction of Biogas Production from Vegetables, Fruits, and Food Wastes by ANFIS- and LSSVM-Based Models. In A. Baghban (Ed.), BioMed Research International (Vol. 2021, pp. 1–8). Hindawi Limited. https://doi.org/10.1155/2021/9202127

Hansen, B. D., Tamouk, J., Tidmarsh, C. A., Johansen, R., Moeslund, T. B., & Jensen, D. G. (2020). Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model. In Computational Science and Its Applications – ICCSA 2020 (pp. 734–745). Springer International Publishing. https://doi.org/10.1007/978-3-030-58799-4_53

Zareei, S., & Khodaei, J. (2017). Modeling and optimization of biogas production from cow manure and maize straw using an adaptive neuro-fuzzy inference system. In Renewable Energy (Vol. 114, pp. 423–427). Elsevier BV. https://doi.org/10.1016/j.renene.2017.07.050

Najafi, B., & Faizollahzadeh Ardabili, S. (2018). Application of ANFIS, ANN, and logistic methods in estimating biogas production from spent mushroom compost (SMC). In Resources, Conservation and Recycling (Vol. 133, pp. 169–178). Elsevier BV. https://doi.org/10.1016/j.resconrec.2018.02.025

Beltramo, T., Klocke, M., & Hitzmann, B. (2019). Prediction of the biogas production using GA and ACO input features selection method for ANN model. In Information Processing in Agriculture (Vol. 6, Issue 3, pp. 349–356). Elsevier BV. https://doi.org/10.1016/j.inpa.2019.01.002

Reddy, K. S., Aravindhan, S., & Mallick, T. K. (2016). Investigation of performance and emission characteristics of a biogas fuelled electric generator integrated with solar concentrated

photovoltaic system. In Renewable Energy (Vol. 92, pp. 233–243). Elsevier BV. https://doi.org/10.1016/j.renene.2016.02.008

Ren, Y., Yu, M., Wu, C., Wang, Q., Gao, M., Huang, Q., & Liu, Y. (2018). A comprehensive review on food waste anaerobic digestion: Research updates and tendencies. In Bioresource Technology (Vol. 247, pp. 1069–1076). Elsevier BV. https://doi.org/10.1016/j.biortech.2017.09.109

Matuszewska, A., Owczuk, M., Zamojska-Jaroszewicz, A., Jakubiak-Lasocka, J., Lasocki, J., & Orliński, P. (2016). Evaluation of the biological methane potential of various feedstock for the production of biogas to supply agricultural tractors. In Energy Conversion and Management (Vol. 125, pp. 309–319). Elsevier BV. https://doi.org/10.1016/j.enconman.2016.02.072

Cherubini, E., Franco, D., Zanghelini, G. M., & Soares, S. R. (2018). Uncertainty in LCA case study due to allocation approaches and life cycle impact assessment methods. In The International Journal of Life Cycle Assessment (Vol. 23, Issue 10, pp. 2055–2070). Springer Science and Business Media LLC. https://doi.org/10.1007/s11367-017-1432-6

Poege, F., Harhoff, D., Gaessler, F., & Baruffaldi, S. (2019). Science quality and the value of inventions. In Science Advances (Vol. 5, Issue 12). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/sciadv.aay7323

Coupé, T. (2003). In The Journal of Technology Transfer (Vol. 28, Issue 1, pp. 31–46). Springer Science and Business Media LLC. https://doi.org/10.1023/a:1021626702728

Souili, A., Cavallucci, D., Rousselot, F., (2015). A lexico-syntactic pattern matching method to extract Idm- Triz knowledge from on-line patent databases. Procedia Eng 131, 418–425. https://doi.org/10.1016/j.proeng.2015.12.437.

Chowdhary, K. R. (2020). Natural Language Processing. In Fundamentals of Artificial Intelligence (pp. 603–649). Springer India. https://doi.org/10.1007/978-81-322-3972-7_19

Jusoh, S. (2018). A STUDY ON NLP APPLICATIONS AND AMBIGUITY PROBLEMS. Journal of Theoretical & Applied Information Technology, 96(6)

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic Modeling in Management Research: Rendering New Theory from Textual Data. In Academy of Management Annals (Vol. 13, Issue 2, pp. 586–632). Academy of Management. https://doi.org/10.5465/annals.2017.0099

Daenekindt, S., & Huisman, J. (2020). Mapping the scattered field of research on higher education. A correlated topic model of 17,000 articles, 1991–2018. In Higher Education (Vol. 80, Issue 3, pp. 571–587). Springer Science and Business Media LLC. https://doi.org/10.1007/s10734-020-00500-x

L. Jockers, M., & Thalken, R. (2020). Topic Modeling. In Text Analysis with R (pp. 211–235). Springer International Publishing. https://doi.org/10.1007/978-3-030-39643-5_17

Grootendorst, M. (2020, May 10). Topic modeling with BERT. | Towards data science. Retrieved from "https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6*"*

Angelov, D. (2020a). Top2Vec: Distributed Representations of Topics. Retrieved from "http://arxiv.org/pdf/2008.09470v1"

Lindstedt, N. C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017. In Social Currents (Vol. 6, Issue 4, pp. 307–318). SAGE Publications. https://doi.org/10.1177/2329496519846505

Gallagher, R. J., Reing, K., Kale, D., & Ver Steeg, G. (2017). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. In Transactions of the

Association for Computational Linguistics (Vol. 5, pp. 529–542). MIT Press - Journals. https://doi.org/10.1162/tacl_a_00078

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval - SIGIR '03. the 26th annual international ACM SIGIR conference. ACM Press. https://doi.org/10.1145/860435.860485

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. In Discourse Processes (Vol. 25, Issues 2–3, pp. 259–284). Informa UK Limited. https://doi.org/10.1080/01638539809545028

Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W. S., Cheung, N.-M., Nguyen, H. L. T., Ho, C. S. H., & Ho, R. C. M. (2019). Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis. In JMIR Medical Informatics (Vol. 7, Issue 4, p. e14401). JMIR Publications Inc. https://doi.org/10.2196/14401

Jallan, Y., Brogan, E., Ashuri, B., & Clevenger, C. M. (2019). Application of Natural Language Processing and Text Mining to Identify Patterns in Construction-Defect Litigation Cases. In Journal of Legal Affairs and Dispute Resolution in Engineering and Construction (Vol. 11, Issue 4, p. 04519024). American Society of Civil Engineers (ASCE). https://doi.org/10.1061/(asce)la.1943-4170.0000308

Lee, J., Yi, J.-S., & Son, J. (2019). Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP. In Journal of Computing in Civil Engineering (Vol. 33, Issue 3, p. 04019003). American Society of Civil Engineers (ASCE). https://doi.org/10.1061/(asce)cp.1943-5487.0000807

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. In Technological Forecasting and Social Change (Vol. 125, pp. 236–244). Elsevier BV. https://doi.org/10.1016/j.techfore.2017.08.002

Dong, B., Xu, G., Luo, X., Cai, Y., & Gao, W. (2012). A bibliometric analysis of solar power research from 1991 to 2010. In Scientometrics (Vol. 93, Issue 3, pp. 1101–1117). Springer Science and Business Media LLC. https://doi.org/10.1007/s11192-012-0730-9

Ren, Y., Yu, M., Wu, C., Wang, Q., Gao, M., Huang, Q., & Liu, Y. (2018). A comprehensive review on food waste anaerobic digestion: Research updates and tendencies. In Bioresource Technology (Vol. 247, pp. 1069–1076). Elsevier BV. https://doi.org/10.1016/j.biortech.2017.09.109

Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in patents. In Technological Forecasting and Social Change (Vol. 94, pp. 236–250). Elsevier BV. https://doi.org/10.1016/j.techfore.2014.10.006

De Clercq, D., Wen, Z., & Song, Q. (2019). Innovation hotspots in food waste treatment, biogas, and anaerobic digestion technology: A natural language processing approach. In Science of The Total Environment (Vol. 673, pp. 402–413). Elsevier BV. https://doi.org/10.1016/j.scitotenv.2019.04.051

Kadhim, A. I., Cheah, Y.-N., & Ahamed, N. H. (2014). Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering. In 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology. 2014 Artificial Intelligence with Applications in Engineering and Technology (ICAIET). IEEE. https://doi.org/10.1109/icaiet.2014.21

Angelov D. (2020). Top2Vec: Distributed Representations of Topics. Available online at: http://arxiv.org/pdf/2008.09470v1

Egger, R. (Ed.). (2022). Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications. Springer Nature. https://doi.org/16. 10.1007/978-3-030-88389-8_16

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. Available online at: https://arxiv.org/abs/1908.10084

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1802.03426

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In Advances in Knowledge Discovery and Data Mining (pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. In The Journal of Open Source Software (Vol. 2, Issue 11, p. 205). The Open Journal. https://doi.org/10.21105/joss.00205

Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Carnegie-mellon univ pittsburgh pa dept of computer science.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2203.05794

Egger, R., & Yu, J. (2021). Identifying hidden semantic structures in Instagram data: a topic modelling comparison. In Tourism Review. Emerald. https://doi.org/10.1108/tr-05-2021-0244

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning - ICML '06. the 23rd international conference. ACM Press. https://doi.org/10.1145/1143844.1143859

Harper, L., Kalfa, N., Beckers, G. M. A., Kaefer, M., Nieuwhof-Leppink, A. J., Fossum, M., Herbst, K. W., & Bagli, D. (2020). The impact of COVID-19 on research. In Journal of Pediatric Urology (Vol. 16, Issue 5, pp. 715–716). Elsevier BV. https://doi.org/10.1016/j.jpurol.2020.07.002

[53] Gu, C., & Li, H. (2022). Review on Deep Learning Research and Applications in Wind and Wave Energy. In Energies (Vol. 15, Issue 4, p. 1510). MDPI AG. https://doi.org/10.3390/en15041510

Widodo, D. A., Iksan, N., & Udayanti, E. D. (2021, March). Renewable energy power generation forecasting using deep learning method. In IOP Conference Series: Earth and Environmental Science (Vol. 700, No. 1, p. 012026). IOP Publishing.

Ji, Y., Wang, J., Xu, J., Fang, X., & Zhang, H. (2019). Real-Time Energy Management of a Microgrid Using Deep Reinforcement Learning. In Energies (Vol. 12, Issue 12, p. 2291). MDPI AG. https://doi.org/10.3390/en12122291

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2018). Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. In Energy (Vol. 164, pp. 465–474). Elsevier BV. https://doi.org/10.1016/j.energy.2018.08.207

Ji, Y., Wang, J., Xu, J., & Li, D. (2021). Data-Driven Online Energy Scheduling of a Microgrid Based on Deep Reinforcement Learning. In Energies (Vol. 14, Issue 8, p. 2120). MDPI AG. https://doi.org/10.3390/en14082120

Kamruzzaman, Md., Duan, J., Shi, D., & Benidris, M. (2021). A Deep Reinforcement Learning-Based Multi-Agent Framework to Enhance Power System Resilience Using Shunt Resources. In IEEE Transactions on Power Systems (Vol. 36, Issue 6, pp. 5525–5536). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tpwrs.2021.3078446

Huang, Q., Huang, R., Hao, W., Tan, J., Fan, R., & Huang, Z. (2020). Adaptive Power System Emergency Control Using Deep Reinforcement Learning. In IEEE Transactions on Smart Grid (Vol. 11, Issue 2, pp. 1171–1182). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tsg.2019.2933191

Domínguez-Barbero, D., García-González, J., Sanz-Bobi, M. A., & Sánchez-Úbeda, E. F. (2020). Optimising a Microgrid System by Deep Reinforcement Learning Techniques. In Energies (Vol. 13, Issue 11, p. 2830). MDPI AG. https://doi.org/10.3390/en13112830

Noorollahi, Y., Jokar, M. A., & Kalhor, A. (2016). Using artificial neural networks for temporal and spatial wind speed forecasting in Iran. In Energy Conversion and Management (Vol. 115, pp. 17–25). Elsevier BV. https://doi.org/10.1016/j.enconman.2016.02.041

Moustris, K. P., Zafirakis, D., Alamo, D. H., Nebot Medina, R. J., & Kaldellis, J. K. (2016). 24-h Ahead Wind Speed Prediction for the Optimum Operation of Hybrid Power Stations with the Use of Artificial Neural Networks. In Perspectives on Atmospheric Sciences (pp. 409–414). Springer International Publishing. https://doi.org/10.1007/978-3-319-35095-0_58

Wang, J., Qin, S., Zhou, Q., & Jiang, H. (2015). Medium-term wind speeds forecasting utilizing hybrid models for three different sites in Xinjiang, China. In Renewable Energy (Vol. 76, pp. 91–101). Elsevier BV. https://doi.org/10.1016/j.renene.2014.11.011

Wu, W., Chen, K., Qiao, Y., & Lu, Z. (2016). Probabilistic short-term wind power forecasting based on deep neural networks. In 2016 International Conference on Probabilistic

Methods Applied to Power Systems (PMAPS). 2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). IEEE. https://doi.org/10.1109/pmaps.2016.7764155

Zameer, A., Arshad, J., Khan, A., & Raja, M. A. Z. (2017). Intelligent and robust prediction of short term wind power using genetic programming based ensemble of neural networks. In Energy Conversion and Management (Vol. 134, pp. 361–372). Elsevier BV. https://doi.org/10.1016/j.enconman.2016.12.032

Dong, Q., Sun, Y., & Li, P. (2017). A novel forecasting model based on a hybrid processing strategy and an optimized local linear fuzzy neural network to make wind power forecasting: A case study of wind farms in China. In Renewable Energy (Vol. 102, pp. 241–257). Elsevier BV. https://doi.org/10.1016/j.renene.2016.10.030

Ghannam, R., Klaine, P. V., & Imran, M. (2019). Artificial Intelligence for Photovoltaic Systems. In Power Systems (pp. 121–142). Springer Singapore. https://doi.org/10.1007/978-981-13-6151-7_6

Youssef, A., El-Telbany, M., & Zekry, A. (2017). The role of artificial intelligence in photovoltaic systems design and control: A review. In Renewable and Sustainable Energy Reviews (Vol. 78, pp. 72–79). Elsevier BV. https://doi.org/10.1016/j.rser.2017.04.046

Brentner, L. B., Peccia, J., & Zimmerman, J. B. (2010). Challenges in Developing Biohydrogen as a Sustainable Energy Source: Implications for a Research Agenda. In Environmental Science & Technology (Vol. 44, Issue 7, pp. 2243–2254). American Chemical Society (ACS). https://doi.org/10.1021/es9030613

Kamaraj, M., Ramachandran, K. K., & Aravind, J. (2019). Biohydrogen production from waste materials: benefits and challenges. In International Journal of Environmental Science and

Technology (Vol. 17, Issue 1, pp. 559–576). Springer Science and Business Media LLC. https://doi.org/10.1007/s13762-019-02577-z

Wang, Y., Tang, M., Ling, J., Wang, Y., Liu, Y., Jin, H., He, J., & Sun, Y. (2021). Modeling biohydrogen production using different data driven approaches. In International Journal of Hydrogen Energy (Vol. 46, Issue 58, pp. 29822–29833). Elsevier BV. https://doi.org/10.1016/j.ijhydene.2021.06.122

Khaleghi, M. K., Savizi, I. S. P., Lewis, N. E., & Shojaosadati, S. A. (2021). Synergisms of machine learning and constraint-based modeling of metabolism for analysis and optimization of fermentation parameters. In Biotechnology Journal (Vol. 16, Issue 11, p. 2100212). Wiley. https://doi.org/10.1002/biot.202100212

Lian, Z., Wang, Y., Zhang, X., Yusuf, A., Famiyeh, Lord, Murindababisha, D., Jin, H., Liu, Y., He, J., Wang, Y., Yang, G., & Sun, Y. (2021). Hydrogen Production by Fluidized Bed Reactors: A Quantitative Perspective Using the Supervised Machine Learning Approach. In J (Vol. 4, Issue 3, pp. 266–287). MDPI AG. https://doi.org/10.3390/j4030022

Liu, Y., Liu, J., He, H., Yang, S., Wang, Y., Hu, J., Jin, H., Cui, T., Yang, G., & Sun, Y. (2021). A Review of Enhancement of Biohydrogen Productions by Chemical Addition Using a Supervised Machine Learning Method. In Energies (Vol. 14, Issue 18, p. 5916). MDPI AG. https://doi.org/10.3390/en14185916

Manoharan, Y., Hosseini, S. E., Butler, B., Alzhahrani, H., Senior, B. T. F., Ashuri, T., & Krohn, J. (2019). Hydrogen Fuel Cell Vehicles; Current Status and Future Prospect. In Applied Sciences (Vol. 9, Issue 11, p. 2296). MDPI AG. https://doi.org/10.3390/app9112296

Liao, M., & Yao, Y. (2021). Applications of artificial intelligence-based modeling for bioenergy systems: A review. In GCB Bioenergy (Vol. 13, Issue 5, pp. 774–802). Wiley. https://doi.org/10.1111/gcbb.12816

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D. A., Hernández, M. V., Wardlaw, J., & Rueckert, D. (2018). GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1810.10863

Lee, M., Liu, Z., Huang, R., & Tong, W. (2016). Application of dynamic topic models to toxicogenomics data. In BMC Bioinformatics (Vol. 17, Issue S13). Springer Science and Business Media LLC. https://doi.org/10.1186/s12859-016-1225-0

Morimoto, T., & Kawasaki, Y. (2017). Forecasting Financial Market Volatility Using a Dynamic Topic Model. In Asia-Pacific Financial Markets (Vol. 24, Issue 3, pp. 149–167). Springer Science and Business Media LLC. https://doi.org/10.1007/s10690-017-9228-z

Linton, M., Teo, E. G. S., Bommes, E., Chen, C. Y., & Härdle, W. K. (2017). Dynamic Topic Modelling for Cryptocurrency Community Forums. In Applied Quantitative Finance (pp. 355–372). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-54486-0_18

Tabassum, S., Gama, J., Azevedo, P., Teixeira, L., Martins, C., & Martins, A. (2021). Dynamic Topic Modeling Using Social Network Analytics. In Progress in Artificial Intelligence (pp. 498–509). Springer International Publishing. https://doi.org/10.1007/978-3-030-86230-5_39

Guldi, J. (2019). Parliament's Debates about Infrastructure: An Exercise in Using Dynamic Topic Models to Synthesize Historical Change. In Technology and Culture (Vol. 60, Issue 1, pp. 1–33). Project Muse. https://doi.org/10.1353/tech.2019.0000

Chen, L., Cong, R.-G., Shu, B., & Mi, Z.-F. (2017). A sustainable biogas model in China: The case study of Beijing Deqingyuan biogas project. In Renewable and Sustainable Energy Reviews (Vol. 78, pp. 773–779). Elsevier BV. https://doi.org/10.1016/j.rser.2017.05.027

Appels, L., Assche, A. V., Willems, K., Degrève, J., Impe, J. V., & Dewil, R. (2011). Peracetic acid oxidation as an alternative pre-treatment for the anaerobic digestion of waste activated sludge. In Bioresource Technology (Vol. 102, Issue 5, pp. 4124–4130). Elsevier BV. https://doi.org/10.1016/j.biortech.2010.12.070

Kougias, P. G., & Angelidaki, I. (2018). Biogas and its opportunities—A review. In Frontiers of Environmental Science & Engineering (Vol. 12, Issue 3). Springer Science and Business Media LLC. https://doi.org/10.1007/s11783-018-1037-8

Abdel-Basset, M., Hawash, H., Chakrabortty, R. K., & Ryan, M. (2021). PV-Net: An innovative deep learning approach for efficient forecasting of short-term photovoltaic energy production. In Journal of Cleaner Production (Vol. 303, p. 127037). Elsevier BV. https://doi.org/10.1016/j.jclepro.2021.127037

Pierdicca, R., Malinverni, E. S., Piccinini, F., Paolanti, M., Felicetti, A., & Zingaretti, P. (2018). DEEP CONVOLUTIONAL NEURAL NETWORK FOR AUTOMATIC DETECTION OF DAMAGED PHOTOVOLTAIC CELLS. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 42(2)

Mosallaie, S., Rad, M., Schiffauerova, A., & Ebadi, A. (2021). Discovering the evolution of artificial intelligence in cancer research using dynamic topic modeling. In COLLNET Journal

of Scientometrics and Information Management (Vol. 15, Issue 2, pp. 225–240). Informa UK Limited. **https://doi.org/10.1080/09737766.2021.1958659**

Johri, A., & Olds, B. M. (2011). Situated engineering learning: Bridging engineering education research and the learning sciences. Journal of Engineering Education, 100(1), 151-185.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Sun, L., & Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling. In Transportation Research Part C: Emerging Technologies (Vol. 77, pp. 49–66). Elsevier BV. https://doi.org/10.1016/j.trc.2017.01.013

Jiang, H., Qiang, M., & Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. In Renewable and Sustainable Energy Reviews (Vol. 57, pp. 226–237). Elsevier BV. https://doi.org/10.1016/j.rser.2015.12.194

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures, 12(2-3), 93-118.

Yang, H. L., Chang, T. W., & Choi, Y. (2018). Exploring the research trend of smart factory with topic modeling. Sustainability, 10(8), 2779.

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. Journal of Business Economics, 89(3), 327-356.

Ayele, W. Y., & Juell-Skielse, G. (2020, March). Eliciting evolving topics, trends and foresight about self-driving cars using dynamic topic modeling. In Future of Information and Communication Conference (pp. 488-509). Springer, Cham.

Lateef, A. A. A., Ali Al-Janabi, S. I., & Abdulteef, O. A. (2022). Artificial Intelligence Techniques Applied on Renewable Energy Systems: A Review. In Proceedings of International Conference on Computing and Communication Networks (pp. 297–308). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0604-6_25

Sohail, M., Afrouzi, H. N., Mehranzamir, K., Ahmed, J., Siddique, M. B. M., & Tabassum, M. (2022). A comprehensive scientometric analysis on hybrid renewable energy systems in developing regions of the world. Results in Engineering, 16, 100481.

Abualigah, L., Zitar, R. A., Almotairi, K. H., Hussein, A. M., Abd Elaziz, M., Nikoo, M. R., & Gandomi, A. H. (2022). Wind, solar, and photovoltaic renewable energy systems with and without energy storage optimization: a survey of advanced machine learning and deep learning techniques. Energies, 15(2), 578.

Xiang, L., Yang, X., Hu, A., Su, H., & Wang, P. (2022). Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks. Applied Energy, 305, 117925.

Tawn, R., & Browell, J. (2022). A review of very short-term wind and solar power forecasting. In Renewable and Sustainable Energy Reviews (Vol. 153, p. 111758). Elsevier BV. https://doi.org/10.1016/j.rser.2021.111758

Ebadi, A., Auger, A., & Gauthier, Y. (2022). On the evolution of research in hypersonics: application of natural language processing and machine learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2208.08507

Sechser, T. S., Narang, N., & Talmadge, C. (2019). Emerging technologies and strategic stability in peacetime, crisis, and war. Journal of Strategic Studies, 42(6), 727-735.

Rifkin, J. (2011). The third industrial revolution: how lateral power is transforming energy, the economy, and the world. Macmillan.

Verma, S., & Gustafsson, A. (2020). Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach. *Journal of Business Research*, *118*, 253-261.

Herrera Viedma, E. (2020). Global trends in coronavirus research at the time of Covid-19: A general bibliometric approach and content analysis using SciMAT.

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., & Zeng, Z. (2021). Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Frontiers in Earth Science*, *9*, 596860.

Yeghikian, M., Ahmadi, A., Dashti, R., Esmaeilion, F., Mahmoudan, A., Hoseinzadeh, S., & Garcia, D. A. (2021). Wind Farm Layout Optimization with Different Hub Heights in Manjil Wind Farm Using Particle Swarm Optimization. *Applied Sciences*, *11*(20), 9746.

Shamshirband, S., Rabczuk, T., & Chau, K. W. (2019). A survey of deep learning techniques: application in wind and solar energy resources. *IEEE Access*, *7*, 164650-164666.

CHAPTER FOUR

HYBRID CATBOOST-CNN-LSTM MODEL FOR BIOGAS FEEDSTOCK ANALYSIS AND SYSTEM PERFORMANCE FORECASTING: INDUSTRIAL-SCALE BIOGAS PLANET APPLICATION

4.1. ABSTRACT

Biogas plants are among the most environmentally friendly renewable energy-producing systems because they treat waste, generate energy and reduce greenhouse gas emissions. However, they suffer from unforeseen disruptions and lack of adequate production due to the variation in the feedstock characteristics and the complexity of Anaerobic Digestion (AD). Artificial Intelligence (AI)-based methods can provide insightful information about biogas systems and ways to optimize them. This study analyzes industrial-scale biogas plant's feedstock and predicts biogas production by developing end-to-end Machine Learning (ML) pipelines to benefit domain experts, stakeholders and decision-makers. I developed four ML models to analyze the effect of eleven substrates on the system's performance. The Catboost algorithm has been chosen among all models since it showed the least error and the most stable performance in 10-fold cross-validation. Afterward, we built three different deep learning algorithms, namely, Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN)-LSTM. The hybrid architecture fitted the best with the time series dataset with a normalized mean square error (NMSE) of 0.006. Feature importance results revealed that acid and oil waste contribute more in generating biomethane, while kitchen waste and fruit and vegetable waste significantly contribute to final production. The proposed approach benefits from large-scale quantitative analysis capabilities employable in other renewable energy time series problems.

Keywords: Artificial Intelligence; biogas systems; Machine learning; Time series Forecasting; long and short-term memory

4.2. INTRODUCTION

Relying on fossil fuels jeopardizes energy security and the well-being of communities because these fuels are nonrenewable and unsustainable, and their combustion produces greenhouse gases (GHG) and other emissions. These emissions cause health problems in humans, animals and plants, while GHG exacerbates global warming and negatively affects ecosystems worldwide. Therefore, many countries have started developing and utilizing various green, low-carbon and sustainable energies (Heydari et al., 2021). Biogas production is a waste-based recycling system mainly for producing renewable biofuel and bioenergy while valorizing organic residues (Kougias and Angelidaki, 2018; Pöschl et al., 2010). Biogas is produced through anaerobic digestion (AD), where microorganisms carry out a network of biological reactions to degrade and decompose various organic substrates such as animal manure, food scrapes, wastewater biosolids and organic by-products, and convert them to biogas (Naroznova et al., 2016). The AD process produces biogas usable to generate sustainable energy, such as electricity or heat, and can be used directly as vehicle fuel after enriching its methane content. At the same time, the GHG emission and storage need for organic waste decrease (Chiu et al., 2022; Islam et al., 2017). This process aligns with the definition of circular economy and sustainable development (Geissdoerfer et al., 2018; Daly, 2006).

Currently, biogas systems are gaining attention in different countries. For instance, according to the American Biogas Council, the US has around 2,300 biogas plants, and there is a potential to build more than 15,000 new biogas plants. The American Biogas Council indicated that focusing

on the US potential biogas infrastructure could produce up to approximately 100 trillion kilowatt-hours of electricity annually. Within EU countries, almost 17,400 biogas power plants have been utilized until 2018 (Scarlat et al., 2018), reaching more than 20,000 in 2022. Also, biogas energy production in the EU countries has grown to the equivalent of 6 million tonnes of oil, with a more than 20% annual growth rate (EurObserv, 2013). AD biogas production process contains four main steps: hydrolysis, acidogenesis, acetogenesis and methanogenesis (Appels et al., 2011; Adekunle and Okolie, 2015). The AD process is highly dynamic and complex because the various technical design and operation parameters and composition of the organic inputs affect the microorganisms' life cycle in each step (Li et al., 2021). The most influential technical parameters in biogas systems are indicated in Table 4.1. It has been found that the efficient control of the industrial-scale AD process is difficult, or sometimes the optimal conditions differ from lab-scale to industrial-scale process (Xu et al., 2018; Matuszewska et al., 2016; Westerholm et al., 2019; Mainardis et al., 2019). This difference is due to the complex interaction among the systems' inputs, microorganisms, and operation parameters, resulting in a nonlinear behavior and uncertainty of these systems' outputs. This behavior may negatively affect biogas plants' performance (Hu et al., 2018). For instance, the existence of different substrates within the feed of the biogas system usually leads to anaerobic co-digestion (AcoD), in which different unwanted reactions occur (De Clercq et al., 2019). Therefore, this system must monitor and control the feeds to achieve the desired efficiency and optimum production. However, optimizing such a system requires intensive mathematical calculation, and there is potential for significant errors resulting in poor results. Various conventional optimization methods can be applied to lab-scale biogas systems but optimizing industrial-scale biogas systems differ, and it is more complicated considering the technical and economic parameters involved (Westerholm et al., 2019; Mainardis et al., 2019). For

instance, Matuszewska et al. (2016) found that the optimum ratio of the biogas system's feed in a laboratory is different from the industrial scale due to longer retention time in the industrial unit.

Table 4.1. Important operational parameters in biogas systems.

Parameters	Description
Temperature	There are three different temperature zones: 1) Psychrophilic fermentation: < 20 °C (Dębowski et al., 2021). 2) Mesophilic fermentation: 20 to 40 °C (Huang et al., 2021). 3) Thermophilic fermentation: 40 to 70 °C (Shao et al., 2020). The highest biogas production rate is usually obtained at thermophilic fermentation (Yilmaz et al., 2018).
Carbon-to-nitrogen ratio (C/N ratio)	C/N ratio indicates the AD reactions stability, and it has a significant impact on the biogas yield (Xue et al., 2020).
pH	The optimal pH ranges from 6.6 to 7.6 (Budiyono et al., 2013).
HRT	HRT is the average time that the feed of the biogas system is kept in the digestors (Dong et al., 2022). Optimal HRT depends on the OLR and type of culture (pure or mixed culture) (Sravan et al., 2021).
OLR	OLR is equal to the mass of the feed substrate per unit of time and reactor volume (Dong et al., 2022).
Type of process	Dry or wet fermentation (Stolze et al., 2015).

In recent years, computer technologies such as the Internet of Things (IoT) and cloud applications have been developed, and researchers used them to enhance sustainable management systems of renewable energy (Alhasnawi et al., 2022; Bhoi et al., 2022; Tran et al., 2022; Bouali et al., 2021). AI is another powerful computational tool leveraging different fields of renewable energy systems such as wind systems (Sachit et al., 2022; Lee et al., 2021; Chatterjee and Dethlefs, 2021) and photovoltaic cells (Mellit and Kalogirou, 2021; Kurukuru et al., 2021; Serrano-Luján et al., 2022) to solve various complicated problems such as uncertain systems optimization's problems efficiently. AI has shown promising results in biogas systems, especially by developing Deep Learning (DL); thus, it has played a crucial role in biogas in recent years. For instance,

Tufaner and Demirci (2020) employed a three-layer Artificial Neural Network (ANN) architecture to build a predictive model for biogas systems performance considering different technical and operational parameters. They considered influent pH, effluent pH, influent alkalinity, effluent alkalinity, OLR, effluent COD, effluent total suspended solids (TSS), and effluent volatile suspended solids (VSS). Another study used ANN as a predictive model to predict the performance of a lab-scale biogas system. Their result shows that 333.4 NL/kgVS of biogas can be obtained by AcoD of agricultural waste and cow manure mixed at a ratio of 7 to 3 (Almomani and Bhosale, 2020). Cinar et al. (2022) employed ML algorithms such as Support Vector Machine (SVM) and Decision Tree (DT) for optimizing the AD process considering the temperature feature. They used a lab-scale dataset and obtained an R^2 score of 0.93 by SVM. Despite the large number of industrial-scale biogas plants, few studies use industrial-scale biogas data (Chiu et al., 2022). Two main characteristics of industrial-scale biogas datasets are being multivariate and time series. Sezer et al. (2020) indicated that Recurrent Neural Network (RNN)-based deep learning algorithms usually provide more significant and robust results than their conventional ML counterparts for time series forecasting problems. Since LSTM can learn patterns from sequential information, it suits any sequential dataset, including time series. Hybridizing LSTM with a one-dimensional CNN can enhance the performance of the proposed DL architecture, especially in processing long sequential data. It is because CNN provides more useful information for LSTM architecture by extracting informative features and learning knowledge from internal time-series instances. To our best knowledge, despite the high potential of hybrid CNN-LSTM algorithm in processing large-scale biogas systems' data, it has rarely been leveraged in biogas studies. However, researchers in other renewable energy domains, such as wind systems, solar energy and PV cells, have benefited from its capability. For instance, Shen et al. (2022) leveraged the CNN-LSTM model to predict

wind speed in wind power plants. Agga et al. (2022) employed the same model to forecast short-term photovoltaic power generation.

In this study, the study aim to provide an efficient end-to-end data-driven solution for industrial biogas systems that can benefit stakeholders, investors, and decision-makers before developing a biogas unit or enhancing the performance of existing biogas facilities. Furthermore, the studyaccurate and reliable method can prevent wasting resources and improve the environment by analyzing the effect of various factors on the system's performance and providing robust forecasting results.

4.3. MATERIALS AND METHODS

Conducting this study requires multiple steps, explained in detail within this section. The first part is collecting the raw data, followed by preprocessing them. We normalized the dataset to decrease the noise effect. We leveraged Catboost into the dataset to select features based on their importance on the system's performance to reduce the dimensionality of the dataset. Also, we employed deep learning models that required specific input shapes. Therefore, we made the models' input data points in a compatible format. Afterward, single deep learning models, namely LSTM and GRU, and hybrid deep learning architecture, CNN-LSTM, were built, and the selected features were fed to them. Finally, after evaluating all the models, the CNN-LSTM algorithm was chosen to forecast the performance of the Shenzhen industrial biogas plant. Figure 4.1 depicts the overall overview of this study.

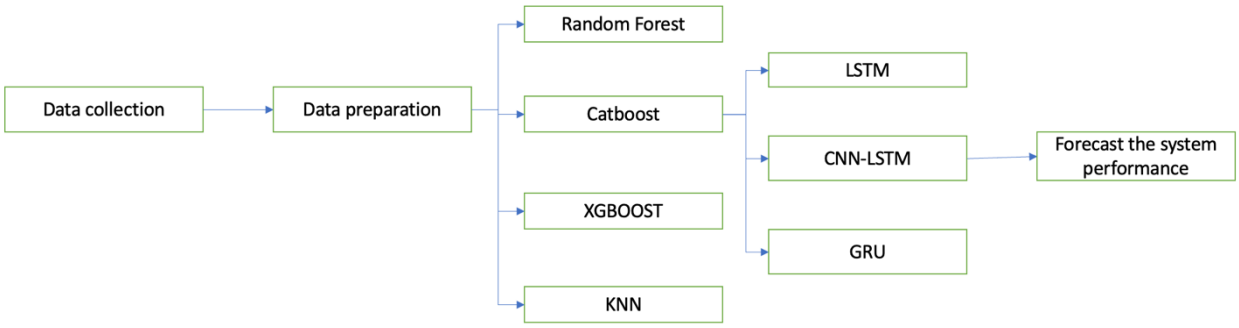

Figure 4.1. The overall methodology of the proposed data-driven pathway.

4.3.1. Raw data collection

The dataset of this study is from the Shenzhen biogas facility (Shenzhen, China), and it is publicly accessible through GitHub; Victoria 3467 BioDigest: Biogas Project, created in November 2018. 80% of the data has been used for training to measure the weights and biases of the hybrid deep learning model, while the remaining 20% has been used for testing (validating) the model performance.

4.3.2. Data preparation

The extracted dataset must be preprocessed and cleaned before being fed into AI models. The dataset included features for the inputs and outputs of two separate tanks. Since this study aims to analyze the Shenzhen biogas system, the study created five features: *acid feed*, *acid discharge*, *anaerobic feed*, *anaerobic cumulative*, and *daily system output*. Each feature is the sum of that feature for each tank. Moreover, although the dataset is a time series dataset considering its daily output target value, the date for each data point has not been indicated. Therefore, the study turned it into a time series dataset by utilizing built-in functions of the Pandas package in Python. A feature selection was conducted by leveraging the Catboost model, explained in section 2.2.1, to reduce the dimensionality of the dataset and remove unnecessary information. One of the main

differences between this study and that of De Clercq et al. (2019) is that the study has directly and efficiently calculated regression tasks' errors without performing intensive feature engineering and making regression to classification tasks by labeling target values. These errors included Normalized Root Mean Square Error (NRMSE), normalized mean square error (NMSE), and Normalized Mean Absolute Error (NMAE). In addition, the study considered a normalized error to facilitate the comparison among calculated errors. Finally, it created a three-dimensional input shape since our CNN-LSTM architecture requires a specific input shape.

4.3.2.1. Ensemble models

Ensemble learning is a fusion of a set of trained models aiming to enhance the predictive performance of a single model (Mendes-Moreira et al., 2012; Zhang and Ma, 2012; Rokach, 2016). Catboost, unbiased gradient boosting with categorical features (Dorogush et al., 2018; Prokhorenkova et al., 2018), is an improved gradient-boosted decision tree that prevents overfitting the model, a significant drawback of boosting-based models (González et al., 2020). This algorithm builds oblivious trees, called decision tables, which have the same splitting criterion for the whole level of the tree (Lou et al., 2017). These trees are symmetric, balanced, more resistant to overfitting, and learn faster in prediction (González et al., 2020) . A useful capability of the Catboost algorithm is measuring feature importance where the input variables' effect on a target value is quantitatively assessed (Long et al., 2021; Wang et al., 2020; Xu et., 2021).

4.3.3. Proposed method

Figure 4.2 illustrates the overall architecture of leveraged CNN-LSTM model to forecast the Shenzhen biogas facility biogas production. This study considered the time series data of the biogas facility inputs composed of different types of waste. It trained the hybrid DL architecture on daily input data by the sliding window algorithm (Kim and Cho, 2018). Since CNN networks

are specialized in extracting spatial features (Ketkar and Santana, 2017), our CNN architecture's convolutional layer extracted spatial characteristics of multivariate time series datasets and passed it to LSTM networks. On the other hand, LSTMs are specialized for temporal feature extraction, and LSTM networks model the temporal time series information employing the extracted spatial features. Hence, CNN-LSTM can be employed as a robust predictive model to forecast biogas production in a hierarchical, fully connected architecture. The performance of the model is evaluated by using different loss functions.

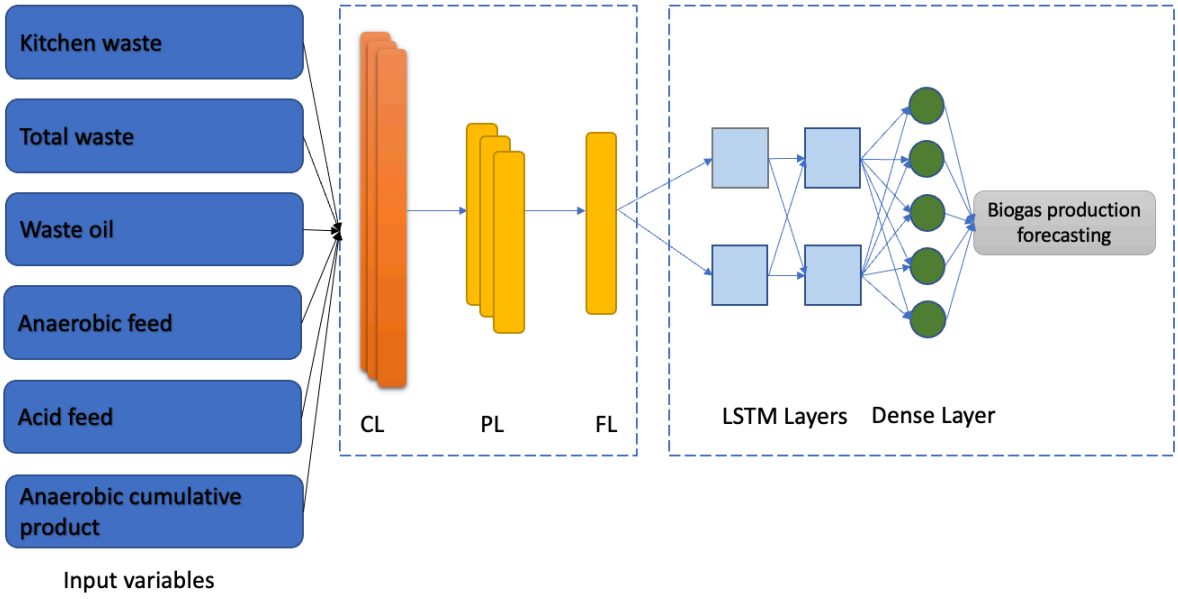

Figure 4.2. The overall proposed hybrid deep learning architecture.

4.3.3.1. CNN-LSTM

The proposed CNN-LSTM predictive model comprises a series of connections of deep learning architectures, with CNN placed at the upper layer. The CNN architecture receives different variables affecting biogas production, such as kitchen waste, acid feed and bread paste, alongside the date. The CNN architectures specialize in modeling obvious grid-like topology datasets (Brownlee, 2018). The CNN architectures are compatible with input formats like 1D, 2D,

and nD (Ketkar and Santana, 2017). The 1D CNN architectures can provide robust results in time series problems (Hussain et al., 2020). They consist of an input layer that receives the biogas measured variables in the Shenzhen biogas facility, an output layer extracting features to LSTM architecture, and hidden layers. The typical hidden layer structure consists of one Fully Connected Layer (CL), a specific type of linear operations, an activation function layer, usually Relu, and a Pooling Layer (PL). CL reads sequential time series multivariate inputs, generates feature maps, and passes them to the next layer. Within the process of linear operation, each neuron in the network processes biogas production data just for the receptive field. Also, this linear operation will result in parameter reduction and deepening of the CNN-LSTM hybrid architectures. Considering $x_i^0 = \{x_1, x_2, \dots, x_n\}$ is the biogas production input vector, and n is the number of scaled daily units per sliced window; Equation 4.1 shows the output of the first CL:

$$y_{ij}^1 = \sigma\left(b_j^1 + \sum_{m=1}^{M} w_{m=1,j}^1 \, x_{i+m-1,j}^0\right) \tag{4.1}$$

where y_{ij}^1 is the result of the first CL, calculated by the vector x_{ij}^1 from the previous layer, b_j^1 denotes the bias of j^{th} of the feature map, w represents kernels' weight by the index value of m, and σ shows the activation function. Equation 4.2 is the general formula for the result of n^{th} layer of the CL.

$$y_{ij}^n = \sigma\left(b_j^n + \sum_{m=1}^{M} w_{m=1,j}^1 \, x_{i+m-1,j}^0\right) \tag{4.2}$$

PL is employed to make a single neuron by combining the neuron's output of the previous layer. It enhances computational efficiency by reducing the representation's space size by using the maximum value of each neuron in the previous layer. Equation 4.3 shows the max PL process.

$$p_{ij}^n = max_{r \in R} \, y_{i \times T + r, j}^{n-1} \tag{4.3}$$

where T denotes the stride deciding the input data area, and *y* is the input size, which is greater than R, which is the pooling size. The last layer of the CNN network is a Fully Connected Layer (FCL), which flattens distilled feature maps into a single long vector (Géron, 2019; Pal and Prakash, 2017).

LSTM stores the main features of time series information of biogas production extracted by the CNN architecture. Within the LSTM network, the previous hidden state is updated by units, which leads to preserving long-term memory. Hence, LSTM is effective in understanding temporal features in the long-term sequence. LSTM network can provide robust performance in prediction biogas production since it overcomes vanished and explosive gradient problems controlling over new input and deciding how to update its memory. There are three gate units, input, output, and forget gate, enabling the network to control input and output information. The gating mechanism stores the memory by employing continuous values between 0-1. The mathematical process of LSTM input, forget, and output gates are explained in Equations 4.4, 4.5, and 4.6, respectively. Equations 4.7 and 4.8 explain the mathematics of the cell states and hidden states through the gating mechanism.

$$i_t = \sigma\left(W_{pi}p_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i\right) \tag{4.4}$$

$$f_t = \sigma\left(W_{pf}p_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f\right) \tag{4.5}$$

$$f_t = \sigma\left(W_{po}p_t + W_{ho}h_{t-1} + W_{co} \circ c_{t-1} + b_o\right) \tag{4.6}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma\left(W_{pc}p_t + W_{hc}h_{t-1} + b_c\right) \tag{4.7}$$

$$h_t = o_t \circ \sigma\left(c_t\right) \tag{4.8}$$

where σ is the employed activation function with non-linearity and accepts input ranges between -1 to 1, W shows the matrix weight, and b is the bias. Biogas production features, represented by p_t, that are extracted by PL of the CNN architecture at the time window of t and are employed as the input of the LSTM network. Leveraging the LSTM model results in significant performance in terms of time series modeling of signals and provides robust results in the Shenzhen biogas facility production. The last layer of CNN-LSTM architecture is a fully connected layer (dense layer) that predicts biogas production over a specific period. A dense layer connects neurons to every single neuron in the previous layers. It uses the output of LSTM as its input. The proposed CNN-LSTM model predicts biogas production daily. Equation 4.9 illustrates the mathematical mechanism of this part of the leveraged deep learning architecture.

$$d_i^n = \sum_j W_{ji}^{n-1} \left(\sigma \left(h_i^{n-1} \right) + b_i^{n-1} \right) \quad (4.9)$$

where σ denotes a nonlinear activation function such as tanh, n represents the number of LSTM units, W is i^{th} node weight of the $n-1$ layer, j^{th} denotes the layer n node, and finally b_i^{n-1} represents the bias.

4.3.3.2. Model architecture

Choosing the correct and efficient deep learning architecture is critical in deploying deep learning models (Karnuta et al., 2019). The typical structure of this hybrid architecture includes CL, PL, FCL, LSTM and a dense layer (Zhou et al., 2015). These layers have adjustable parameters, such as kernel size, filter size, and the number of units. Changing these parameters can affect the model's performance (He and Sun, 2015). The dataset's characteristics should be considered to develop the most efficient and robust model for biogas production. After the feature selection, the network's input is 440 × 8, 8 features consisting of a daily time series. The CL of the model has 64 feature maps with the kernel size of three-time steps to read sequences. The PL

simplifies the generated feature maps by maintaining the ¼ highest signal. After flattening feature maps by FC, two LSTM layers, with 50 units, processed the time-series information. Afterward, a dropout layer is responsible for better model generalization, with a rate of 20%. Finally, a fully connected layer connects its neurons to each AI unit of the preceding networks. I designed architecture and selected parameters for the CNN-LSTM hybrid topology are illustrated in Table 4.2.

Table 4.2. Proposed architecture, layers configuration.

CONV 1D	Filter	64
-	Kernel size	3
	Activation function	Relu
Max Pooling	-	2
Flatten	-	-
TimeDisterbuted	-	-
LSTM	Hidden nodes	50
	Activation function	tanh
LSTM	Hidden nodes	50
	Activation function	sigmoid
Dropout	Rate	0.25
Dense	-	1

4.3.3.3. Evaluation metrics

Three different metrics were considered to measure the performance of our leveraged models: NMSE, NRMSE, and NMAE. The study used the normalized errors to facilitate the comparison among them. Equations 4.10-4.15 describe the mathematics behind these errors.

$$MSE = \frac{1}{n} \times \sum_{i=1}^{n} (o_i - p_i)^2 \tag{4.10}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (o_i - p_i)^2}{n}} \tag{4.11}$$

$$MAE = \sum_{i=1}^{n} \frac{|o_i - p_i|}{n} \tag{4.12}$$

$$NMSE = \frac{MSE}{o_{max} - o_{min}} \quad (4.13)$$

$$NRMSE = \frac{RMSE}{o_{max} - o_{min}} \quad (4.14)$$

$$NMAE = \frac{MAE}{o_{max} - o_{min}} \quad (4.15)$$

where n denotes the number of instances; p_i ; is the predicted value, o_i is the actual value; o_{max} and o_{min} are the maximum and minimum actual values, respectively.

4.4. CASE STUDY

4.4.1. Shenzhen biogas facility dataset

Figure 4.3 depicts the overall structure of the Shenzhen biogas facility considering the raw dataset available on the public repository on GitHub (De Clercq et al., 2019). We did not employ imputation methods for this dataset since it has no missing values. The original dataset contains 18 variables where some inputs and outputs of tanks were measured separately. Since the study aimed to analyze the whole system, it considered some features and created new ones. For instance, the study summed both the "1_acidification_hydrolysis tank feed" and "2_acidification_hydrolysis tank feed" features to create a new feature, *acid feed* (Figure 4.4).

4.4.2. AI-based process optimization

Another advantage of the proposed method is that it can be used to enhance biogas system performance on an industrial scale. In this regard, the importance of different feed components was measured by Catboost. The study used Catboost since it showed more robust performance and stability than other employed machine learning models (the study elaborated more on this within section 4.3.4). Figure 4.4 depicts the importance of each component within the feed. As can be seen, cumulative anaerobic products (VFAs) and acid feed are the most important because

methanogens (methane-producing microorganisms) can use them as substrates directly. Besides, waste oil showed one of the highest ranks. Theoretically, the stochiometric methane yield per gram of volatile solids (VS) for fat, proteins, and carbohydrate is 1014, 496, and 415 NL CH_4 kg^{-1} VS, respectively. Obviously, fats and oils produce 2.44 methane compared to carbohydrates (Saady and Masse, 2015). Acid can be used as an effective chemical pretreatment in biogas production (Sarto et al., 2019; Syaichurrozi et al., 2019). The study's analysis illustrated that the critical factors of the complicated industrial-scale anaerobic digestion (biogas production) process could be identified. Industrial and academic researchers can benefit from the proposed method since it provides reliable insight into the process and allows them to avoid conducting comprehensive laboratory experiments or complex mathematical calculations. This model can optimize technical parameters in industrial anaerobic digestion, such as biological oxygen demand to chemical oxygen demand ratio, pH, temperature, etc.

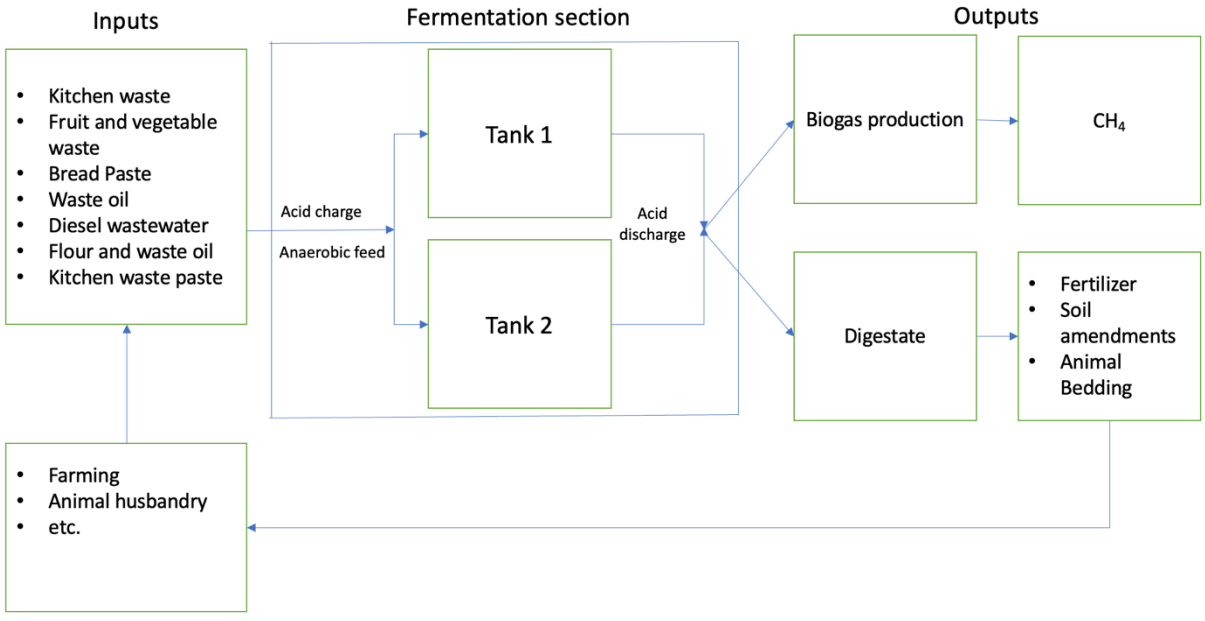

Figure 4.3. Shenzhen biogas plant overall process.

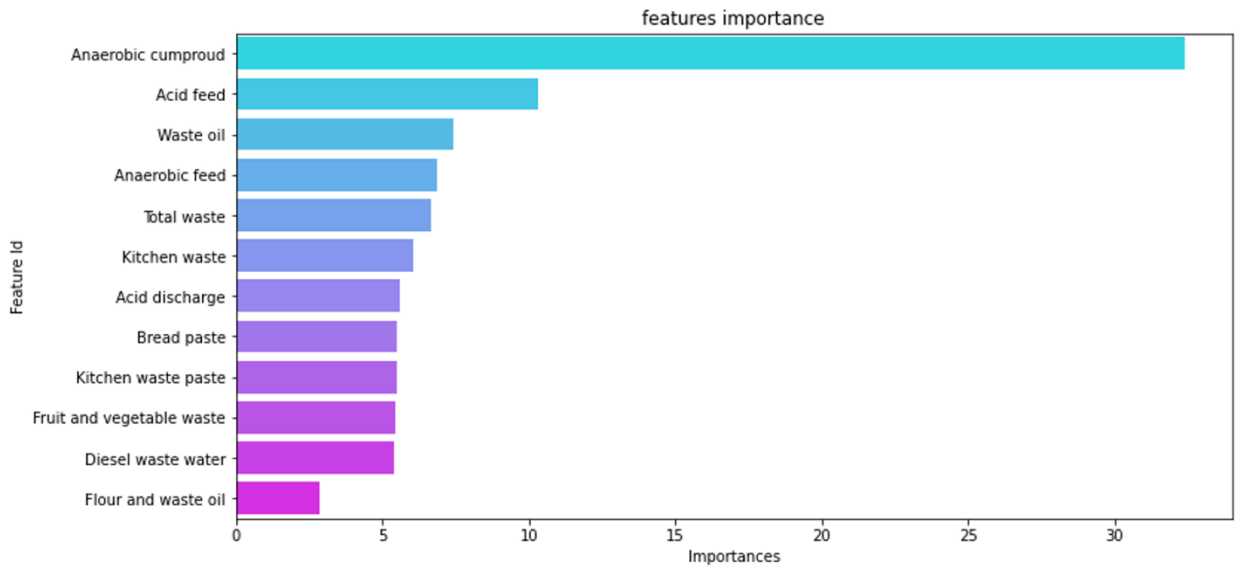

Figure 4.4. Identifying the feed categories' importance using Catboost feature importance.

4.4.3. Deep learning algorithms training and performance comparison

4.4.3.1. Learning curves

Figure 4.5 shows each algorithm's learning curves for training and test datasets. The learning behavior of each model is based on monitoring and recording the training and test loss, which is NMSE, in this case, per epoch. From a statistical point of view, there is a concept called overfitting, which means employing many parameters while adjusting a statistical model. The main consequence of overfitting is that it decreases the model's predictive capability. Therefore, the study used an early stop mechanism to optimize the number of epochs. This mechanism prevents overfitting and makes the model more efficient (Raskutti et al., 2014). This process prevents overfitting and contributes to the model's generalization (Chollet, 2021).

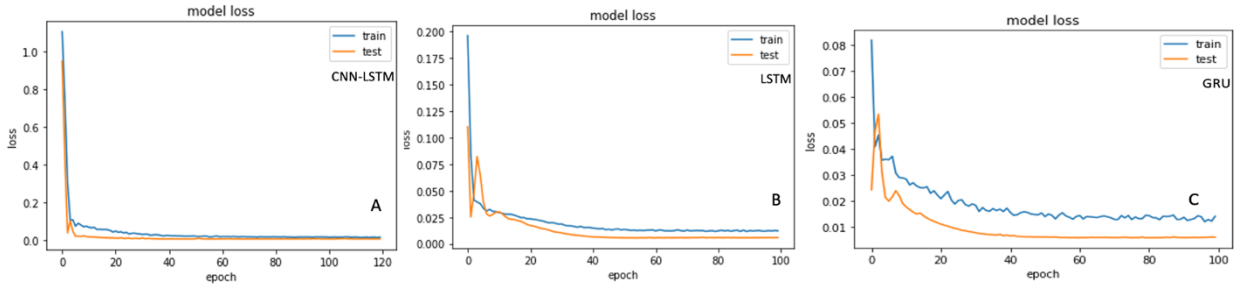

Figure 4.5. A,B, and C Learning curves for (a) CNN-LSTM, (b) LSTM, and (c) GRU architectures, respectively.

As shown in Figure 4.5, CNN-LSTM, LSTM, and GRU required 120, 100 and 113 epochs, respectively, determined by the early stop mechanism. The first two models were trained more stably compared to GRU. Moreover, our models did not experience overfitting since the learning graphs of training and test datasets are stable and close to each other (Figure 4.5).

4.4.3.2. Performance comparison between deep learning architectures

The studyhas applied different deep learning algorithms to the Shenzhen biogas dataset to ensure that the proposed architecture outperforms other models. Table 4.3 shows the performance of leveraged deep learning models for biogas production forecasting. LSTM, CNN-LSTM, and GRU were employed for time series prediction, and three metrics (NMSE, NRMSE and NMAE) were considered to evaluate the performance of algorithms. Computational results demonstrate that the proposed hybrid CNN-LSTM architecture outperformed conventional deep learning algorithms. Hence, the study selected the CNN-LSTM architecture for biogas production forecasting.

Table 4.3. Performance comparison of deep learning architectures.

Model	NMRSE	NMSE	NMAE
LSTM	0.089	0.008	0.065
CNN-LSTM	0.078	0.006	0.060
GRU	0.085	0.006	0.063

Figure 4.6 includes three time-series plots showing the predicted and actual Shenzhen biogas facility production performance. The blue line shows the actual amount of biogas produced, and the solid red line shows the predicted biogas production. The result of the leveraged CNN-LSTM model is more accurate than the other two deep learning-based models.

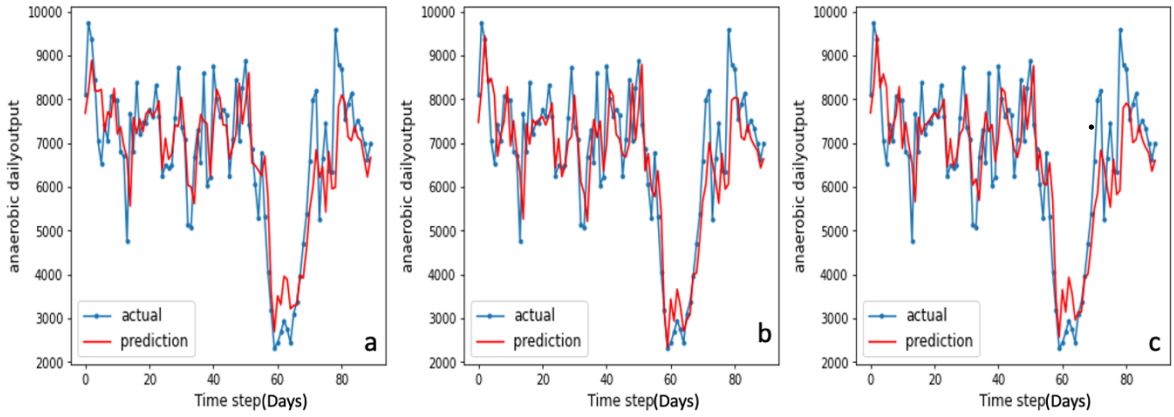

Figure 4.6. Comparison of actual and predicted values for (a) CNN-LSTM, (b) LSTM and, (c) GRU models respectively.

4.4.4. Performance comparison with convolutional machine learning models

The study applied different machine learning techniques to the dataset and compared their results to this study developed CNN-LSTM's result to prove the proposed model's efficiency and usefulness. The developed hybrid deep learning architecture showed the lowest error compared to all leveraged machine learning algorithms: RF, Extreme Gradient Boosting (XGBoost), Catboost, and KNN. This study used 10-fold cross-validation for performance comparison. Within a 10-fold cross-validation, the dataset is partitioned into 10 equal parts. Then, 10 iterations of training and testing are performed, where each fold will be the test dataset once (Refaeilzadeh et al., 2016). As a result, this study developed hybrid deep learning model gained the lowest NMSE followed by Catboost, XGBoost, RF, and KNN. The boxplot (Figure 4.7) illustrates the machine learning

methods' measured errors (NMSE) and stability. Moreover, the settings of leveraged machine learning models are indicated in Table 4.4 which were determined by gridsearch cross validation method.

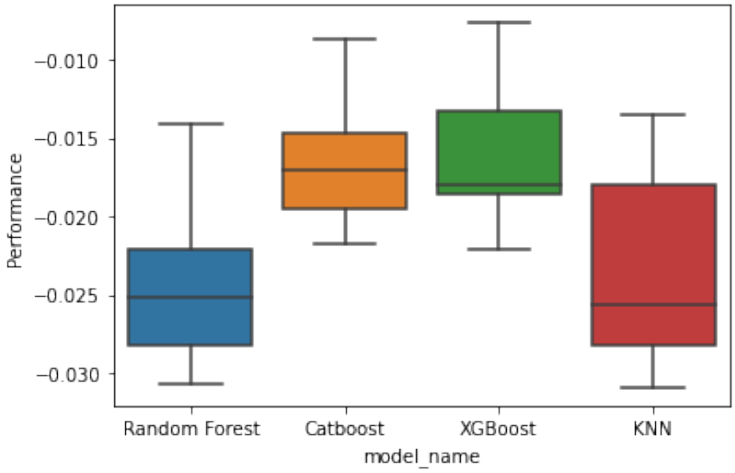

Figure 4.7. Comparison of machine learning models performance and stability

Table 4.4 Hyperparameters for machine learning models.

Model	Hyper parameters
Random Forest	No of estimators = 170, max depth = 4
Catboost	Grow policy = symmetric tree, max depth = 6
XGBoost	No of estimators = 1400, max depth = 5
KNN	No of neighbors = 6

4.5. DISCUSSION

The study built a hybrid ML-DL method for optimizing and forecasting the performance of an industrial-scale biogas system. Compared to ML or DL algorithms that are only focused on predicting the final production, The current study's proposed method provides generalized forecasting results with computational efficiency. It enables us to identify and investigate effective parameters in the system. Adjusting input parameters (feedstocks in our case) enhances the process and, subsequently, the final production (biogas yield). The dataset in this study was first used by De Clercq et al. (2019), who employed different ML algorithms alongside intensive feature

engineering to predict biogas system production. In this study, the Catboost part of the model identifies important parameters, and then the CNN-LSTM part predicts the system production. Figure 4.8 compares this study's proposed method and that of Clercq et al. (2019)'s AI-based method that focused on the same biogas systems.

Figure 4.8 shows that this study's developed pipeline is significantly smaller than Clercq et al. (2019)'s; thus, it is more flexible for different time series tasks. The drawback of this study is that operational and technical parameters have not been included in the Shenzhen biogas dataset which has been used in this study to build the model. However, it is vital to consider different operational parameters since biogas system production depends on the performance of the biogas digester, which is affected by different variable parameters (pH, temperature, OLR, HRT, mixing ratio, etc.). Future studies can benefit from combining IoT and AI to enhance biogas systems. Considering recent development in IoT, operational parameters, such as pH, OLR, HRT, temperature, etc., and microbiological features of the microorganisms of the biogas system can be measured with high accuracy on an hourly or daily basis by cutting-edge sensors. Expanding the biogas dataset by adding the mentioned variables with precisely recorded data will enhance data-driven analyses and make them more reliable in solving real-world problems. Besides, I can obtain more significant results and reduce the model bias by compiling more data points and increasing the dataset size. This study's proposed model can be utilized in various fields of time series forecasting, from different renewable energy systems to medical domains.

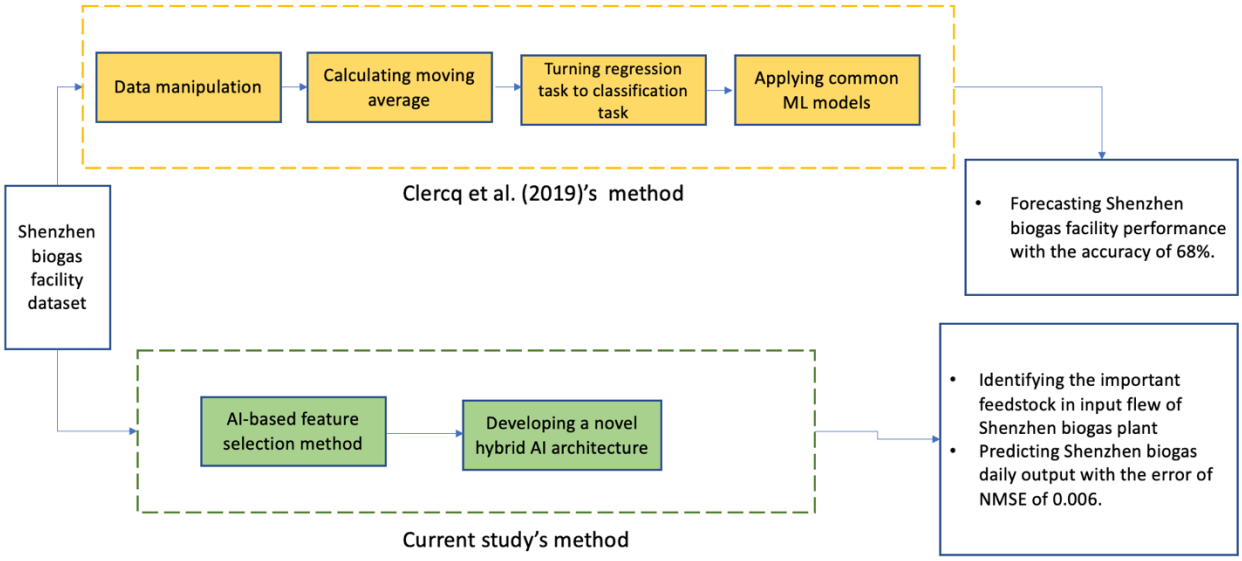

Figure 4.8. Comparison between Clercq et al. (2019) and my methods and results.

4.6. CONCLUSIONS

This research aimed to empower industrial-scale biogas systems' optimization and forecasting performance by developing a hybrid AI model based on time series data measured over more than one year. This study trained different ML algorithms, RF, Catboost, XGBoost, and KNN, to investigate the importance of each parameter on the system performance. This study chose Catboost since it showed more stability and the lowest error. Following the feature selection part, this study leveraged hybrid CNN-LSTM DL architecture to forecast system performance. It gained significant results with an NMSE of 0.078, compared to the other two single DL architectures, LSTM and GRU, with NMSEs of 0.089 and 0.144, respectively. Leveraging AI in the biogas plant domain can increase stakeholders' benefits by improving the system and reducing operational costs. Also, Future studies can consider a more comprehensive dataset and develop a user-friendly platform that can benefit decision-makers and investors. In this regard, stakeholders

can identify the pros and cons of their investments before constructing a biogas facility considering different environmental, economic and technical parameters.

REFERENCES

Adekunle, K. F., & Okolie, J. A. (2015). A Review of Biochemical Process of Anaerobic Digestion. In Advances in Bioscience and Biotechnology (Vol. 06, Issue 03, pp. 205–212). Scientific Research Publishing, Inc. https://doi.org/10.4236/abb.2015.63020

Agga, A., Abbou, A., Labbadi, M., El Houm, Y., & Ali, I. H. O. (2022). CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production. Electric Power Systems Research, 208, 107908.

Alhasnawi, B. N., Jasim, B. H., Mansoor, R., Alhasnawi, A. N., Rahman, Z. A. S. A., Haes Alhelou, H., ... & Siano, P. (2022). A new Internet of Things based optimization scheme of residential demand side management system. IET Renewable Power Generation.

Almomani, F. (2020). Prediction of biogas production from chemically treated co-digested agricultural waste using artificial neural network. Fuel, 280, 118573.

Almomani, F., & Bhosale, R. R. (2020). Enhancing the production of biogas through anaerobic co-digestion of agricultural waste and chemical pretreatments. Chemosphere, 255, 126805.

Appels, L., Assche, A. V., Willems, K., Degrève, J., Impe, J. V., & Dewil, R. (2011). Peracetic acid oxidation as an alternative pretreatment for the anaerobic digestion of waste activated sludge. In Bioresource Technology (Vol. 102, Issue 5, pp. 4124–4130). Elsevier BV. https://doi.org/10.1016/j.biortech.2010.12.070

Bhoi, A. K., Kabat, M. R., Nayak, S. C., & Palai, G. (2022). Renewable energy source based quality of service (QoS)-aware routing mechanism in cloud network. Wireless Networks, 28(4), 1703-1718.

Bouali, E. T., Abid, M. R., Boufounas, E. M., Hamed, T. A., & Benhaddou, D. (2021). Renewable Energy Integration Into Cloud & IoT-Based Smart Agriculture. IEEE Access, 10, 1175-1191.

Breiman, L. (2001). In Machine Learning (Vol. 45, Issue 1, pp. 5–32). Springer Science and Business Media LLC. https://doi.org/10.1023/a:1010933404324

Brownlee, J. (2018). Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.

Budiyono, B., Syaichurrozi, I., & Sumardiono, S. (2013). Biogas production from bioethanol waste: the effect of pH and urea addition to biogas production rate. Waste Technology, 1(1), 1-5.

Cano, J.-R., Gutiérrez, P. A., Krawczyk, B., Woźniak, M., & García, S. (2019). Monotonic classification: An overview on algorithms, performance measures and data sets. In Neurocomputing (Vol. 341, pp. 168–182). Elsevier BV. https://doi.org/10.1016/j.neucom.2019.02.024

Chatterjee, J., & Dethlefs, N. (2021). Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future. Renewable and Sustainable Energy Reviews, 144, 111051.

Chen, T., & Guestrin, C. (2016). XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD ’16: The 22nd ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. https://doi.org/10.1145/2939672.293978

Chiu, M.-C., Wen, C.-Y., Hsu, H.-W., & Wang, W.-C. (2022). Key wastes selection and prediction improvement for biogas production through hybrid machine learning methods. In Sustainable Energy Technologies and Assessments (Vol. 52, p. 102223). Elsevier BV. https://doi.org/10.1016/j.seta.2022.102223

Chollet, F. (2021). Deep learning with Python. Simon and Schuster.

Daly, H. E. (2006). Sustainable development—definitions, principles, policies. In The future of sustainability (pp. 39-53). Springer, Dordrecht.

De Clercq, D., Jalota, D., Shang, R., Ni, K., Zhang, Z., Khan, A., Wen, Z., Caicedo, L., & Yuan, K. (2019). Machine learning powered software for accurate prediction of biogas production: A case study on industrial-scale Chinese production data. In Journal of Cleaner Production (Vol. 218, pp. 390–399). Elsevier BV. https://doi.org/10.1016/j.jclepro.2019.01.031

Dębowski, M., Korzeniewska, E., Kazimierowicz, J., & Zieliński, M. (2021). Efficiency of sweet whey fermentation with psychrophilic methanogens. Environmental Science and Pollution Research, 28(35), 49314-49323.

Dong, R., Qiao, W., Guo, J., & Sun, H. (2022). Manure treatment and recycling technologies. In Circular Economy and Sustainability (pp. 161–180). Elsevier. https://doi.org/10.1016/b978-0-12-821664-4.00009-1

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). Catboost: gradient boosting with categorical features support (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1810.11363

EurObserv, E. R. The state of renewable energies in Europe. 13th EurObserv'ER report 2013; 4-9.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets(Vol. 10, pp. 978-3). Berlin: Springer

Geissdoerfer, M., Morioka, S. N., de Carvalho, M. M., & Evans, S. (2018). Business models and supply chains for the circular economy. In Journal of Cleaner Production (Vol. 190, pp. 712–721). Elsevier BV. https://doi.org/10.1016/j.jclepro.2018.04.159

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. "O'Reilly Media, Inc.".

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. In Information Fusion (Vol. 64, pp. 205–237). Elsevier BV. https://doi.org/10.1016/j.inffus.2020.07.007

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

He, K., & Sun, J. (2015). Convolutional neural networks at constrained time cost. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5353-5360).

Heydari, B., Abdollahzadeh Sharghi, E., Rafiee, S., & Mohtasebi, S. S. (2021). Use of artificial neural network and adaptive neuro-fuzzy inference system for prediction of biogas production from spearmint essential oil wastewater treatment in up-flow anaerobic sludge blanket reactor. In Fuel (Vol. 306, p. 121734). Elsevier BV. https://doi.org/10.1016/j.fuel.2021.121734

Hu, C., Yan, B., Wang, K., & Xiao, X. (2018). Modeling the performance of anaerobic digestion reactor by the anaerobic digestion system model (ADSM). In Journal of Environmental Chemical Engineering (Vol. 6, Issue 2, pp. 2095–2104). Elsevier BV. https://doi.org/10.1016/j.jece.2018.03.018

Huang, J., Guo, K., Shi, B., & Li, J. (2021). Mesophilic fermentation upgrades SCFA production from natural/raw henna plant biomass. Biomass Conversion and Biorefinery, 11(3), 795-801.

Hussain, D., Hussain, T., Khan, A. A., Naqvi, S. A. A., & Jamil, A. (2020). A deep learning approach for hydrological time-series prediction: A case study of Gilgit river basin. In Earth Science Informatics (Vol. 13, Issue 3, pp. 915–927). Springer Science and Business Media LLC. https://doi.org/10.1007/s12145-020-00477-2

Islam, M. T., Alam, M. Z., Al-Mamun, A., Elgharbawy, A. A., & Riyadi, F. A. (2017). Development of enzymatic pretreatment of palm oil mill effluent for monomers towards biogas production. Journal of Advanced Research in Materials Science, 38(1), 39-44.

Karnuta, J. M., Navarro, S. M., Haeberle, H. S., Helm, J. M., Kamath, A. F., Schaffer, J. L., Krebs, V. E., & Ramkumar, P. N. (2019). Predicting Inpatient Payments Prior to Lower Extremity Arthroplasty Using Deep Learning: Which Model Architecture Is Best? In The Journal of Arthroplasty (Vol. 34, Issue 10, pp. 2235-2241.e1). Elsevier BV. https://doi.org/10.1016/j.arth.2019.05.048

Ketkar, N., & Santana, E. (2017). Deep learning with Python(Vol. 1). Berkeley: Apress.

Kim, T.-Y., & Cho, S.-B. (2018). Web traffic anomaly detection using C-LSTM neural networks. In Expert Systems with Applications (Vol. 106, pp. 66–76). Elsevier BV. https://doi.org/10.1016/j.eswa.2018.04.004

Kurukuru, V. S. B., Haque, A., Khan, M. A., Sahoo, S., Malik, A., & Blaabjerg, F. (2021). A review on artificial intelligence applications for grid-connected solar photovoltaic systems. Energies, 14(15), 4690.

Lee, M., & He, G. (2021). An empirical analysis of applications of artificial intelligence algorithms in wind power technology innovation during 1980–2017. Journal of Cleaner Production, 297, 126536.

Li, Y., Wang, Z., Jiang, Z., Feng, L., Pan, J., Zhu, M., Ma, C., Jing, Z., Jiang, H., Zhou, H., Sun, H., & Liu, H. (2022). Bio-based carbon materials with multiple functional groups and graphene structure to boost methane production from ethanol anaerobic digestion. In Bioresource Technology (Vol. 344, p. 126353). Elsevier BV. https://doi.org/10.1016/j.biortech.2021.126353

Long, F., Wang, L., Cai, W., Lesnik, K., & Liu, H. (2021). Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data. In Water Research (Vol. 199, p. 117182). Elsevier BV. https://doi.org/10.1016/j.watres.2021.117182

Lou, Y., & Obukhov, M. (2017). BDT. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD ’17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. https://doi.org/10.1145/3097983.3098175

Louppe, G. (2014). Understanding Random Forests: From Theory to Practice (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1407.7502

Mainardis, M., Buttazzoni, M., Gievers, F., Vance, C., Magnolo, F., Murphy, F., & Goi, D. (2021). Life cycle assessment of sewage sludge pretreatment for biogas production: From laboratory tests to full-scale applicability. Journal of Cleaner Production, 322, 129056.

Matuszewska, A., Owczuk, M., Zamojska-Jaroszewicz, A., Jakubiak-Lasocka, J., Lasocki, J., & Orliński, P. (2016). Evaluation of the biological methane potential of various feedstock for the production of biogas to supply agricultural tractors. In Energy Conversion and Management (Vol. 125, pp. 309–319). Elsevier BV. https://doi.org/10.1016/j.enconman.2016.02.072

Mellit, A., & Kalogirou, S. (2021). Artificial intelligence and internet of things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: Challenges, recommendations and future directions. Renewable and Sustainable Energy Reviews, 143, 110889.

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression. In ACM Computing Surveys (Vol. 45, Issue 1, pp. 1–40). Association for Computing Machinery (ACM). https://doi.org/10.1145/2379776.2379786

Naroznova, I., Møller, J., & Scheutz, C. (2016). Global warming potential of material fractions occurring in source-separated organic household waste treated by anaerobic digestion or incineration under different framework conditions. In Waste Management (Vol. 58, pp. 397–407). Elsevier BV. https://doi.org/10.1016/j.wasman.2016.08.020

Pal, A., & Prakash, P. K. S. (2017). Practical time series analysis: master time series data processing, visualization, and modeling using Python. Packt Publishing Ltd.

Pöschl, M., Ward, S., & Owende, P. (2010). Evaluation of energy efficiency of various biogas production and utilization pathways. *Applied energy*, *87*(11), 3305-3321.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

Raskutti, G., Wainwright, M. J., & Yu, B. (2014). Early stopping and non-parametric regression: an optimal data-dependent stopping rule. The Journal of Machine Learning Research, 15(1), 335-366.

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-Validation. In Encyclopedia of Database Systems (pp. 1–7). Springer New York. https://doi.org/10.1007/978-1-4899-7993-3_565-2

Rokach, L. (2019). Ensemble learning: pattern classification using ensemble methods.

Roopnarain, A., & Adeleke, R. (2017). Current status, hurdles and future prospects of biogas digestion technology in Africa. In Renewable and Sustainable Energy Reviews (Vol. 67, pp. 1162–1179). Elsevier BV. https://doi.org/10.1016/j.rser.2016.09.08

Saady, N. M. C., & Massé, D. I. (2015). Impact of organic loading rate on psychrophilic anaerobic digestion of solid dairy manure. Energies, 8(3), 1990-2007.

Sachit, M. S., Shafri, H. Z. M., Abdullah, A. F., Rafie, A. S. M., & Gibril, M. B. A. (2022). Global Spatial Suitability Mapping of Wind and Solar Systems Using an Explainable AI-Based Approach. ISPRS International Journal of Geo-Information, 11(8), 422.

Sarto, S., Hildayati, R., & Syaichurrozi, I. (2019). Effect of chemical pretreatment using sulfuric acid on biogas production from water hyacinth and kinetics. In Renewable Energy (Vol. 132, pp. 335–350). Elsevier BV. https://doi.org/10.1016/j.renene.2018.07.121

Scarlat, N., Dallemand, J.-F., & Fahl, F. (2018). Biogas: Developments and perspectives in Europe. In Renewable Energy (Vol. 129, pp. 457–472). Elsevier BV. https://doi.org/10.1016/j.renene.2018.03.006

Serrano-Luján, L., Toledo, C., Colmenar, J. M., Abad, J., & Urbina, A. (2022). Accurate thermal prediction model for building-integrated photovoltaics systems using guided artificial intelligence algorithms. Applied Energy, 315, 119015.

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. In Applied Soft Computing (Vol. 90, p. 106181). Elsevier BV. https://doi.org/10.1016/j.asoc.2020.106181

Shao, W., Wang, Q., Rupani, P. F., Krishnan, S., Ahmad, F., Rezania, S., ... & Din, M. F. M. (2020). Biohydrogen production via thermophilic fermentation: A prospective application of Thermotoga species. Energy, 197, 117199.

Shen, Z., Fan, X., Zhang, L., & Yu, H. (2022). Wind speed prediction of unmanned sailboat based on CNN and LSTM hybrid neural network. Ocean Engineering, 254, 111352.

Škapa, S., & Vochozka, M. (2020). Towards Higher Moral and Economic Goals in Renewable Energy. Science and Engineering Ethics, 26(3), 1149-1158.

Sravan, J. S., Tharak, A., & Mohan, S. V. (2021). Status of biogas production and biogas upgrading: A global scenario. In Emerging Technologies and Biological Systems for Biogas Upgrading (pp. 3–26). Elsevier. https://doi.org/10.1016/b978-0-12-822808-1.00002-7

Stolze, Y., Zakrzewski, M., Maus, I., Eikmeyer, F., Jaenicke, S., Rottmann, N., Siebner, C., Pühler, A., & Schlüter, A. (2015). Comparative metagenomics of biogas-producing microbial communities from production-scale biogas plants operating under wet or dry fermentation

conditions. In Biotechnology for Biofuels (Vol. 8, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s13068-014-0193-8

Syaichurrozi, I., Villta, P. K., Nabilah, N., & Rusdi, R. (2019). Effect of sulfuric acid pretreatment on biogas production from Salvinia molesta. Journal of Environmental Chemical Engineering, 7(1), 102857.

Tran, M. K., Panchal, S., Khang, T. D., Panchal, K., Fraser, R., & Fowler, M. (2022). Concept review of a cloud-based smart battery management system for lithium-ion batteries: Feasibility, logistics, and functionality. Batteries, 8(2), 19.

Tufaner, F., & Demirci, Y. (2020). Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models. Clean Technologies and Environmental Policy, 22(3), 713-724.

Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. In Pattern Recognition Letters (Vol. 136, pp. 190–197). Elsevier BV. https://doi.org/10.1016/j.patrec.2020.05.035

Wang, L., Long, F., Liao, W., & Liu, H. (2020). Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. In Bioresource Technology (Vol. 298, p. 122495). Elsevier BV. https://doi.org/10.1016/j.biortech.2019.122495

Westerholm, M., Castillo, M. d. P., Chan Andersson, A., Jahre Nilsen, P., & Schnürer, A. (2019). Effects of thermal hydrolytic pretreatment on biogas process efficiency and microbial community structure in industrial- and laboratory-scale digesters. In Waste Management (Vol. 95, pp. 150–160). Elsevier BV. https://doi.org/10.1016/j.wasman.2019.06.004

Xu, F., Li, Y., Ge, X., Yang, L., & Li, Y. (2018). Anaerobic digestion of food waste – Challenges and opportunities. In Bioresource Technology (Vol. 247, pp. 1047–1058). Elsevier BV. https://doi.org/10.1016/j.biortech.2017.09.020

Xu, W., Long, F., Zhao, H., Zhang, Y., Liang, D., Wang, L., Lesnik, K. L., Cao, H., Zhang, Y., & Liu, H. (2021). Performance prediction of ZVI-based anaerobic digestion reactor using machine learning algorithms. In Waste Management (Vol. 121, pp. 59–66). Elsevier BV. https://doi.org/10.1016/j.wasman.2020.12.003

Xue, S., Wang, Y., Lyu, X., Zhao, N., Song, J., Wang, X., & Yang, G. (2020). Interactive effects of carbohydrate, lipid, protein composition and carbon/nitrogen ratio on biogas production of different food wastes. Bioresource Technology, 312, 123566.

Yilmaz, A., Ünvar, S., KOÇER, A., & Aygün, B. (2018). Factors affecting the production of biogas. Int. J. Sci. Eng. Res, 9(5), 59-62.

Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: methods and applications*. Springer Science & Business Media.

Zhou, C., Sun, C., Liu, Z., & Lau, F. C. M. (2015). A C-LSTM Neural Network for Text Classification (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1511.08630

Cinar, S. Ö., Cinar, S., & Kuchta, K. (2022). Machine Learning Algorithms for Temperature Management in the Anaerobic Digestion Process. *Fermentation*, *8*(2), 65.

CHAPTER FIVE

BIOGAS PREDICTION USING A HYBRID APPROACH OF VARIATIONAL AUTO ENCODER AND MACHINE LEARNING MODELING

5.1. ABSTRACT

This study aims to predict bio methene yeild of diverse subsrtrates considering their organic components (carbohydrates, protein, fat, and lignin) using machine learning models. In this regard, 75 data points from previous litretures were compiled. However, considering lack of sufficient data for training supervised machine leanring models, the study leveraged a deep-learning based data augmatation techniqe to build a dataset, with 500 data points, which has similar statistical charactrestics to the original dataset. To predict biogas yeild, the study used three ensamble models, XGboost, Catboost, and Random Forest that were hyper-tuned using nested cross-validation method. XGboost outperformed others with Root Mean Square Error (RMSE) of 0.165 and R^2 of 0.75. The study also investigated the effect of each organic component of substrates on the the biogas production. The results show that "fats" has the highest impact on biogas production while "protein" has the lowest effect. The developed pipeline can be used in other domains dealing with data limitation.

Keywords: Machine Learning, Data Augmatation, enasmble learning, biogas production

5.2. INTRODUCTION

Renewable energy is booming with growing importance because it can address several global challenges. Biogas is a renewable fuel that can supply energies (heat and electricity) that can replace those derived from fossil fuels (Heiker et al., 2021). Anaerobic digestion (AD) is a process in which bacteria break down organic substrates in the absence of oxygen to grow and produce biogas, which usually contains 50-70% methane (CH_4) and 30-50% carbon dioxide (CO_2) (Xu et al., 2018).

AD technology has been successfully applied in various sectors, including wastewater treatment plants, livestock farming, and treating the organic fraction of municipal solid waste (OFMSW). AD offers an effective solution for treating sludge in wastewater treatment plants, reducing its volume, eliminating pathogens, and producing biogas. Additionally, AD is a sustainable approach for managing and treating manure in livestock farming, minimizing emissions while simultaneously generating biogas for on-site energy needs. Moreover, applying AD in treating the OFMSW enables the diversion of organic waste from landfills, reducing methane emissions, and generating biogas that can be utilized for energy production, fostering a more sustainable waste management approach (Cruz et al., 2022).

According to the United States Environmental Protection Agency, 146 million tons of municipal and non-hazardous solid waste were landfilled in 2019, which was almost 50% of total generated solid. However, the recovery ratio for food waste was 11.8 % which accounted for almost 35 million tons of municipal and non-hazardous solid waste. Methane emissions from organic waste in landfills are a critical environmental issue (Nordin et al., 2022). Due to methane's 25-fold higher global warming potential compared to CO_2, landfills are significantly contributing to global warming (Sabour et al., 2020). Accordingly, interest is a growing in diverting organic

food waste from landfills to AD facilities, which are considered the most environmentally favorable option (Jaunich et al., 2020). The biogas generated through AD offers versatile applications, including electricity generation and household heating, serving as a sustainable alternative to traditional energy sources. Additionally, it can viably substitute fossil fuels in vehicles, further promoting the adoption of renewable energy in transportation. It is worth mentioning that AD exhibits comparatively lower capital, operational, and managerial costs (Mulu et al., 2021), while effectively addressing organic waste management, climate change mitigation, and bioenergy production (Bekchanov et al., 2019).

The AD process involves four primary stages (Figure 5.1): hydrolysis, acidogenesis, acetogenesis, and methanogenesis (Zabed et al., 2020). Hydrolytic bacteria release enzymes that break down particulate and colloidal biomass into soluble forms during hydrolysis. Carbohydrates, proteins, and lipids are enzymatically degraded into monosaccharides, amino acids, and long-chain fatty acids, respectively (Naik et al., 2021). In the acidogenesis stage, specific microorganisms metabolize hydrolysis products, producing hydrogen (H_2), CO_2, alcohols, and volatile fatty acids (VFAs). An excessive amount of VFA decreases pH and inhibits microorganisms' activities (Akbay et al., 2022). Acetogenesis is a transitional stage, where acetogenic bacteria oxidize VFAs and long-chain fatty acids to generate acetic acid, CO_2, H_2, and water (Deschamps et al., 2022). This stage facilitates the breakdown of long-chain VFAs into short-chain VFAs, such as acetic acid and butyric acid, which are readily utilized by methanogenic archaea (Deschamps et al., 2022). Methanogenesis, the final stage of AD, involves two types of microorganisms: acetoclastic methanogens, which convert acetate to methane, and hydrogenotrophic methanogens, which convert H_2 and CO_2 to methane (Kougias and Angelidaki, 2018).

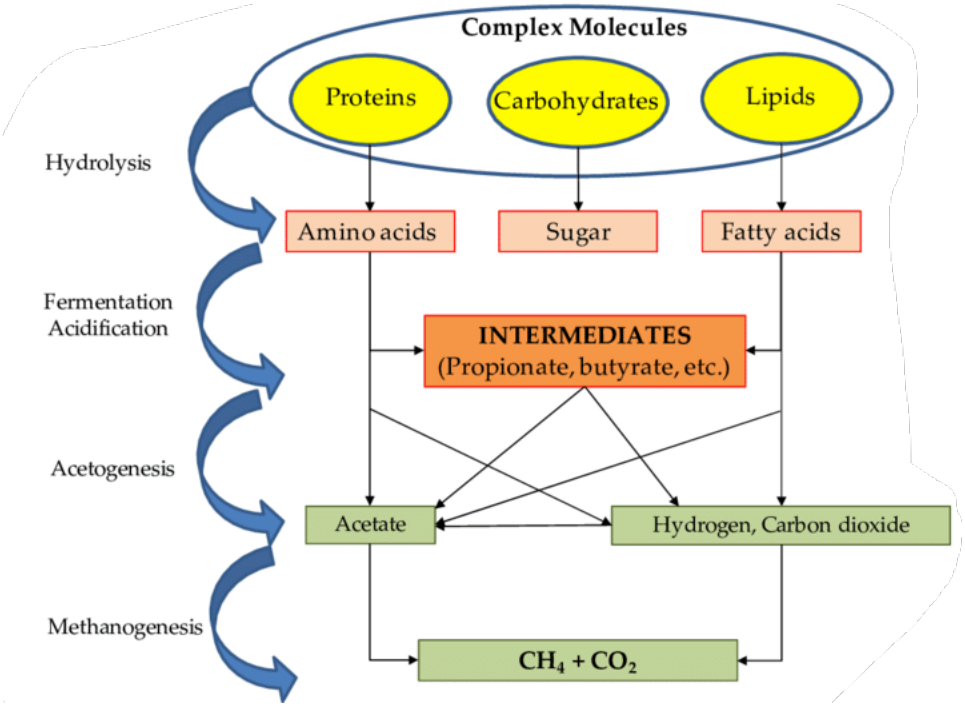

Figure 5.1. The Anaerobic digestion stages.

The biochemical methane potential (BMP) test shows the ultimate methane potential of substrates (Filer et al., 2019). The BMP test involves triplicate sets of serum bottles, including blank, control, and substrate-fed, to ensure the test accuracy and reliability. Substrate bottles contain inoculum, the substrate, and basal medium, while the blank bottle includes inoculum and a medium or water without a substrate to account for residual organic matter (Filer et al., 2019). Within control bottles, there are inoculum, a control substrate, usually a pure substrate such as glucose, and a nutrient medium. Biogas production is monitored over thirty or more days using the syringe method, liquid displacement, manometric measurement, mass loss, or biogas composition monitoring until biogas production ceases (Filer et al., 2019). The BMP results accurately indicate the quantity of methane generated from the substrate (Calabrò et al., 2018).

However, the BMP outcomes are influenced by different parameters such as microbial culture adaptation, organic loading, carbon-to-nitrogen ratio, pH, temperature, and other variables.

Furthermore, the BMP test is time-consuming and does not provide immediate results (Argiz et al., 2020). Thus, there is a need for a faster and smarter method to predict methane yield (Tsapekos et al., 2015). In order to overcome the conventional BMP test method limitations, new instruments have been developed, such as the automatic methane potential test system (AMPTS). The AMPTS removes CO_2 and other acid gas in the biogas before estimating the CH_4 yield; the instrument functionality is based on principles of the conventional BMP test (Shi, 2012). Therefore, in such a system, generated methane is directly measured on line using the liquid displacement and buoyancy method. Although the AMPTS can produce high-quality data and require minimum labour resources, it needs sound systems (Shi, 2012). There are also other methods to determine BMP, like, spectroscopy methods. These methods are employed to analyze how radiation interacts with matter in the ultraviolet (UV), visible, and infrared (IR) regions. These techniques assess properties such as absorbance, transmission, diffusion, or fluorescence. There are two main types of spectroscopy: atomic spectroscopy, which measures substances in a gaseous state after volatilization, and molecular spectroscopy, which directly analyzes substances in liquid form (Spanjers and van Lier, 2006; Esteves et al., 2012). Infrared (IR) spectroscopy techniques primarily focus on the interactions between chemical bonds. Two common examples of IR spectroscopic instruments are near-infrared spectroscopy (NIR) and Fourier transform mid-infrared spectroscopy (FTIR).

In determining BMP, the Envital® kit utilizes a fluorescence redox indicator. It is worth noting that this tool is still in its early stages of development and refinement (Bellaton et al., 2016). Researchers have utilized NIR spectroscopy in conjunction with chemometric modeling to predict

BMP values for a wide range of feedstocks, considering its ability to quantify compounds within the NIR radiation range (12,821 to 3,959 cm^{-1}) makes it well-suited for BMP estimation. This method is sensitive to interactions involving C-H, N-H, and O-H bonds, which are prevalent in many organic compounds. While NIR method has shown promising results for BMP prediction, there are challenges to address, particularly regarding the standard error of the laboratory reference method (Ward, 2016). The initial attempts to measure BMP using FTIR spectroscopy were conducted by Bekiaris et al. (2015). This technology has proven to be suitable for in-line determination of various parameters such as VFA, alkalinity, chemical oxygen demand (COD), and total organic carbon (TOC) (Spanjers and van Lier, 2006). FTIR spectroscopy requires only a small amount of the sample for analysis. However, it is costly and interpreting the obtained spectra can be more challenging than NIR spectroscopy due to overlapping overtones and combination bands (Bekiaris et al., 2015). Analyzing the chemical composition of the substrates provides an alternative approach for predicting its BMP, as it is influenced by the 'biomass's chemical characteristics (Godin et al., 2015). Buswell and Mueller (1952) developed elemental composition analytical equations (C, H, O, and N) of the substrate that have proven to predict stoichiometric methane production effectively. These stoichiometric equations exhibit high accuracy when applied to easily biodegradable substrates such as cellulose (Jingura et al., 2017). However, their reliability decreases when predicting the BMP of complex and slowly degradable compounds such as lignocellulosic biomasses (Thomsen et al., 2014). Consequently, these equations are mainly used to measure substrate biodegradability by comparing the methane yield from experimental BMP batch tests to the theoretical stoichiometric value.

Different studies used mathematical models to predict BMP; for instance, Hu et al. (2017) used two modeling methods as alternatives to the time-consuming BMP test for estimating the

ultimate specific methane yield. They considered elemental content and organic composition to calculate the theoretical methane yield of various leafy vegetables. Similarly, Zheng et al. (2013) investigated the biochemical composition of biodegradable solid wastes and predicted methane yield by BMP tests. Since BMP test showed slow results, they used a modified Gompertz equation to build a predictive model to estimate cumulative methane yield potential. These methods require intensive resources or cannot be used for various substrates.

Artificial intelligence techniques can be leveraged into this domain to build a generalized capability of predicting the methane yield of a wide range of substrates. However, there are not sufficient data points to train machine learning models in this domain. Therefore, data augmentation techniques can be a viable solution. In this research, we aim to overcome data limitations in the biogas domain and build a generalized model capable of predicting the biomethane yield of substrates based on their organic components by machine learning models.

5.3. METHODOLOGY

5.3.1. Data collection

The dataset used in this study is from Hegazy et al. (2023); a manuscript submitted to energies . This dataset contains 75 data points, including a wide range of substrates associated with their organic fractions of carbohydrate (cellulose, hemicellulose), protein, and fat, theoretical and experimental methane yields, and biodegradability collected from previously published papers.

5.3.2. Data preprocessing and augmentation

In this step, two features were dropped from the original dataset: “biodegradability” and “theoretical methane yield”. Afterward, the data input shape was made suitable for the variational autoencoder (VAE) model (Sønderby et al., 2016). Since we had a limited number of data points

for training the machine learning model, we utilized a high-level neural network API from Keras library to leverage the VAE model, a deep learning-based data augmentation technique. Autoencoders are a type of neural network that can learn a compressed representation or encoding of input data, which is then utilized to reconstruct the original input data (Bank et al., 2020). On the other hand, generative models can generate new samples that resemble the training data (Shaham et al., 2019). VAEs combine both architectures by learning to encode input data into a lower-dimensional latent space and then decoding it back into the original data space. During the training process, the model regularizes the encoding distribution to make that the latent space is regular enough and avoid overfitting (Sønderby et al., 2016). VAEs have been shown to be effective in generating complex generative models of data and have yielded state-of-the-art machine learning results in image generation and reinforcement learning. In a VAE, the encoding process involves mapping the input data to a distribution of latent variables, typically modeled as a multivariate Gaussian distribution (Sønderby et al., 2016). The encoding network learns the parameters of this distribution, including the mean (μ), and variance (σ^2), which represent the latent space (Akrami et al., 2022). Mathematically, the encoder takes an input x and produces a distribution q(z|x) over the latent variable z. This distribution is parameterized by the 'encoder's output, which is composed of the mean vector μ and the diagonal covariance matrix Σ (Akrami et al., 2022). The latent variable Z is sampled from this distribution using the reparameterization trick, where $Z = \mu + (\varepsilon \times \sigma)$ with ε sampled from a standard Gaussian distribution. During training, the VAE encourages the learned distribution to approximate a standard Gaussian distribution by minimizing the Kullback-Leibler (KL) divergence between q(z|x) and the standard Gaussian distribution p(z). This is represented by Eq. 5.1:

$$KL(q(z|x) \,||\, p(z)) = -0.5 \times \Sigma\,(1 + \log(\sigma^2) - \mu^2 - \sigma^2) \quad (5.1)$$

Within the decoding process, samples are taken from the latent space and are mapped back to the original data space to reconstruct the input data. The decoder network generates a conditional distribution p(x|z) over the input data, which models the reconstruction of the original data given the latent variable (Akrami et al., 2022). This distribution models the reconstruction of the original data given the latent variable, and the parameters of this distribution are learned during training. VAE aims to maximize the log-likelihood of the data under the decoder distribution while also minimizing the KL divergence between the encoder distribution and the prior distribution over the latent space. The objective function that combines these two components is called the evidence lower bound (ELBO) (Eq. 5.2).

$$\mathrm{ELBO}(x) = E[\log p(x|z)] - \mathrm{KL}(q(z|x) \,||\, p(z)) \tag{5.2}$$

where E is the average value of the expression inside the square brackets with respect to a certain distribution. The first part of Eq. 5.2 denotes the reconstruction term and measures how well the VAE can reconstruct the input data x given a latent variable Z. It computes the expected log-likelihood of the data under the decoder distribution p(x|z). The second term, KL(q(z|x) || p(z)), represents the regularization term and measures the divergence between the encoder distribution q(z|x) and the prior distribution p(z) over the latent space. It measures how much information is lost when approximating the true posterior distribution with the encoder distribution. Figure 5.2 shows the general architecture of VAE models.

The developed VAE's architecture consists encoded layer, defined as a dense layer with three units and a ReLU activation function, one latent space layer with two dense layers, and a decoded layer that is also a dense layer with seven units and a sigmoid activation function.

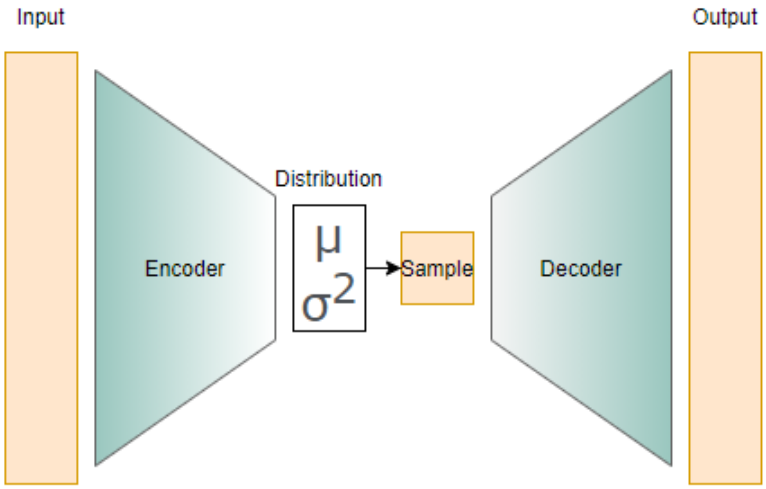

Figure 5.2. The architecture of variational autoencoder.

The VAE is trained using a combination of the mean squared error (MSE) loss for reconstruction and the divergence loss to regularize the latent space distribution (K). The study also leveraged early stopping techniques to enhance the model generalization. Afterward, the generated data were added to the original data, and the final dataset was standard scaling of generated data points.

5.3.3. Machine learning algorithms

This study used three ensemble models, Catboost, Random Forest (RF), and Extreme Gradient Boosting (XGBoost), to predict methane yield based on substates' chemical components and investigate the importance of components on each substate methane yield. According to Dorogush et al. (2018), Catboost is a boosted decision tree algorithm that effectively handles categorical features. It addresses a significant drawback of traditional boosting models, such as overfitting by utilizing decision tables, also known as oblivious trees (González et al., 2020) (Figure 5.3). These decision tables maintain the same splitting criterion for each level of the tree, resulting in symmetric and balanced trees. This approach enhances the model's resistance to overfitting and facilitates faster learning during prediction (Lou et al., 2017).

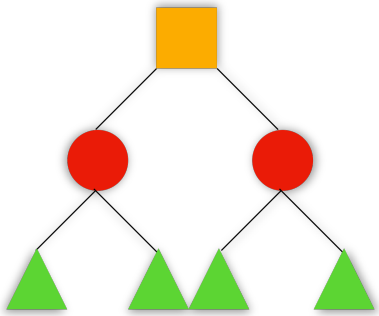

Figure 5.3. Catboost algorithm schematic.

XGBoost is a powerful algorithm for predicting the target value by comparing the predictions of a set of weaker models (Chen and Guestrin (2016)). To address the issue of overfitting, XGBoost incorporates LASSO and Ridge techniques. A notable characteristic of the XGBoost algorithm is its built-in cross-validation solution applied during each iteration. This feature helps assess the model's performance and prevents overfitting. XGBoost employs the gradient descent method and is categorized as an ensemble tree method.

Random Forest (RF) is an ensemble learning algorithm that combines the predictions of multiple decision trees to generate results (Badillo et al., 2020). It leverages the power of multiple algorithms by evaluating several decision trees simultaneously and then aggregating their results to obtain an optimal prediction. The RF algorithm begins by randomly selecting samples from the given dataset, where a decision tree is constructed for each sample, and the predictions of these individual trees are then averaged. This averaging process results in improving the robustness of the final prediction.

5.2.4 Optimizing hyperparameters

To hyper-tune and generalize the developed models, the study conducted nested cross-validation on each model (Figure 5.4). In nested cross-validation, there are outer and inner cross-validation loops (Cawley and Talbot, 2010). The outer loop divides the data into multiple folds where each fold acts as a holdout set for evaluating the model's performance. On the other hand, the inner loop, nested within each outer fold, performs cross-validation again to tune the model's hyperparameters. During the inner loop, the data within the outer fold is further divided into multiple folds, known as the inner folds. The inner loop is responsible for hyperparameter tuning, where different combinations of hyperparameters are tested using the inner folds. Once the optimal hyperparameters are determined, the model is trained on the entire outer fold to compute the model's performance. The mentioned process is repeated for each outer fold, and the evaluation metrics from all the folds are aggregated to provide an overall assessment of the model's performance (Cawley and Talbot, 2010).

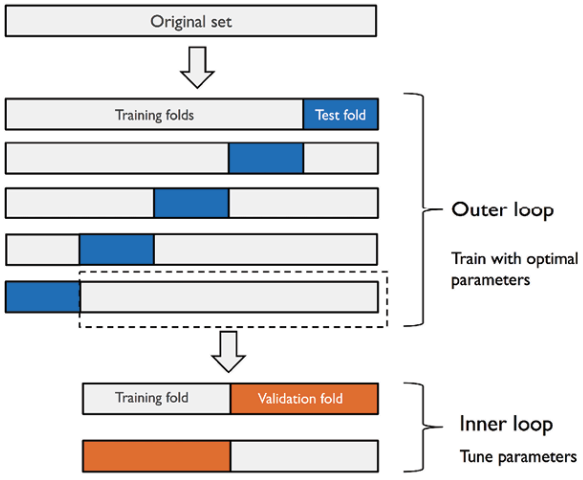

Figure 5.4. Nested cross-validation.

5.3.5. Model evaluation

The models were evaluated using three different metrics, R^2, normalized mean square error, and normalized root mean square error Calculated using Eqs. 5.3 to 5.5.

$$R^2 = \frac{Y_{pred}}{Y} \tag{5.3}$$

$$MSE = \frac{1}{n} \times \sum_{i=1}^{n} (o_i - p_i)^2 \tag{5.4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (o_i - p_i)^2}{n}} \tag{5.5}$$

where n denotes the number of instances; p_i ; is the predicted value, o_i is the actual value; o_{max} and o_{min} are the maximum and minimum actual values, respectively.

5.4. RESULTS AND DISCUSSION

5.4.1. Augmented data via VAE

The VAE generated 500 data points out of 56 given data points. In total, 556 data points, divided to train and test sections with a ratio of 0.8, were used for building machine learning pipelines. Table 5.1 provides the main statistical characteristics of the real and augmented dataset.

Table 5.1. Main statistical characteristics of the real and augmented data

	Real		**Augmented**	
Features	Mean	Standard deviation	Mean	Standard deviation
Carbohydrate	0.579661	0.236489	0.627699	0.199740
Protein	0.394000	0.167689	0.404619	0.148129
Fats	0.478964	0.162360	0.467246	0.147769
Cellulose	0.409964	0.197420	0.429572	0.168087
Hemicellulose	0.425929	0.202125	0.466284	0.177735
Lignin	0.486875	0.101977	0.443058	0.090414
Experimental methane yield	0.528732	0.118915	0.534889	0.103239

As it can be seen, real and augmented data had similar mean and standard deviation. Figure 5.5 shows the performance of the developed VAE model. Considering the improvement of the model and the closeness of both graphs to each other, the model has been generalized.

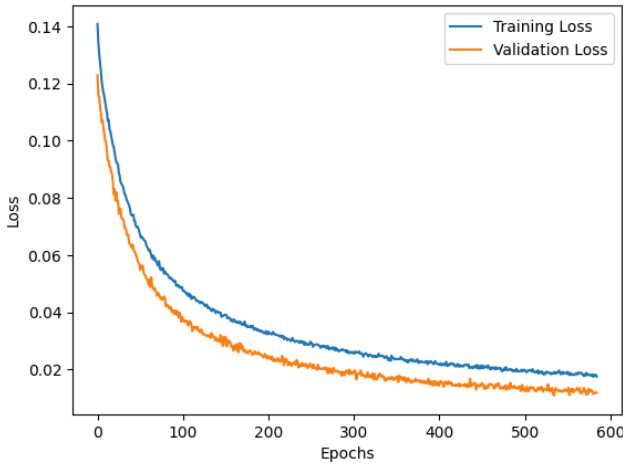

Figure 5.5. Training and loss validation of proposed variational autoencoder

5.4.2. Performance compression of machine learning models

In this study, three ensemble learning techniques were employed to predict the potential methane yield of each substrate. As mentioned earlier, the study conducted nested cross-validation to prevent the models from overfitting and improve their generalization. Table 5.2 provides the most important optimal tuned hyperparameters for each model.

Table 5.2. The most important tuned hyperparameters in developed machine-learning models

Model	Random Forest			Extreme Gradient Boosting		Catboost	
Hyper parameters	Number of estimators	Max depth	Min sample split	Max depth	Min child weight	Depth	Learning rate
	150	15	4	4	8	6	0.01

Table 5.3 illustrates the performance of each employed machine learning technique. XGBoost outperformed other developed models with MSE and RSME of 0.027, and 0.165, respectively. Moreover, Figure 5.6 shows the results of the developed predictive models in this study where XGBoost fits the data better than other models.

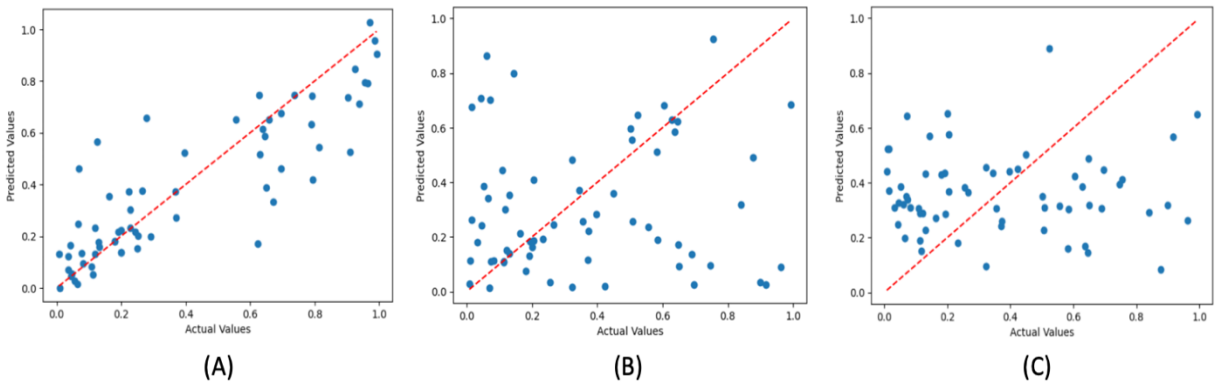

Figure 5.6. Predictive results versus actual values for (A) Extreme Gradient Boosting, (B) Random Forest, and (C) Catboost

Table 5.3. Models' performance comparison

Model	Error		
	MSE	RMSE	R^2
Extreme Gradient Boosting	0.027	0.165	0.75
Random Forest	0.13	0.36	0.69
Catboost	0.1	0.32	0.71

MSE = Mean Square Error; RMSE = Root Mean Square Error

5.4.3. AI-based investigation of chemical components importance

Considering that XGBoost performed better on the dataset in predicting the methane yield, the study considered its results to investigate the effect of each chemical component on the substrates' methene yield. Figure 5.7 illustrates the relative importance of each component of the substrate. As it can be seen in the Figure 5.7, the fat group is the most important in producing biomethane. This is because fats have more electrons, resulting in higher biomethane production (Saady and Masse, 2015). More specifically, methanogenic microorganisms such as lipolytic

bacteria and archaea can break down fats into simpler compounds, such as fatty acids, glycerol, and ultimately, methane (He et al., 2018). On the other hand, proteins undergo a more complex degradation process involving hydrolysis into amino acids and subsequent conversion into volatile fatty acids and ammonia. Moreover, the results show that fats contribute more than carbohydrates in producing biomethane, which is confirmed by previous studies' results. For instance, Saady and Masse (2015) indicated that the stochiometric methane yield per gram of volatile solids for fat and carbohydrate is 1014 and 415 NL CH_4 kg^{-1} VS, respectively, which means that fats can produce more than double the quantity of methane compared to carbohydrates.

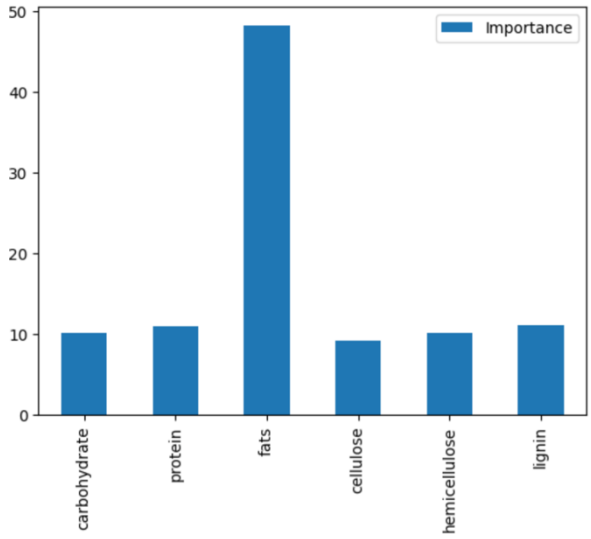

Figure 5.7. The relative effect of different chemical components in producing methane

The developed AI method in this study exhibits versatility beyond the biogas domain and holds potential for application in various environmental domains dealing with data limitations. The proposed method demonstrates adaptability across various industrial and laboratory scale tasks, from time series analysis to computer vision problems.

5.4.4. Application of AI-based models in the anaerobic digestion process

Previous studies employed different data augmentation techniques to overcome data limitations in the AD domain for laboratory- and industrial-scale problems. For example, Asadi and McPhedran (2021) aimed to determine GHG emission rate estimates from a cold-region biological nutrient removal municipal wastewater treatment plant (MWTP) using a hybrid approach combining a generative adversarial network (GAN) with regression modeling. The nonlinearity and complexity of the biological processes involved and limited data availability challenged them in modeling GHG emissions. To address these issues, they developed artificial data generation algorithms using GAN, which had not been previously applied to MWTP modeling. Laboratory-scale reactors and facility-monitored operating parameters were utilized to predict emission rates through the GAN and regression models. They achieved the best result by generating 100 data points where the RMSE was 3219. Their CH_4 and N_2O emission rate estimates modeling RMSE values were 1.35 and 0.588, respectively. Xu et al. (2020) applied hybrid random standard deviation sampling and a deep learning model to predict volatile fatty acids through the anaerobic fermentation. A random standard deviation sampling method was developed for virtual data augmentation using the mean values (x) and standard deviations (a) derived from multiple experimental determinations. Subsequently, deep learning models were established to acquire knowledge from the virtual data and make predictions regarding VFA production. The findings revealed that training the deep learning model with 16 hidden layers, 100 hidden neurons in each layer, and 20,000 virtual samples (comprising five input variables of the anaerobic fermentation process) resulted in improved outcomes. The model achieved an impressive correlation coefficient of 0.998 and a minimum mean absolute error of 3.28%. Table 5.4 provided several previous studies that used data-driven methods to overcome their data limitation in environmental fields.

Table 5.4. Application and performance efficiency of various AI-based models for the determination of biogas.

AD process	Input parameters	Output parameters	Compared ML algorithms	Most accurate model	Prediction performance		Reference
					R^2	RMSE	
AD of spent mushroom compost with wheat straw	C/N ratio, temperature, and retention time	Biogas production	Logistic model, ANN, and ANFIS	ANFIS	0.9996	0.1940	Najafi & Ardabili (2018)
Data from two AD facilities in south China (Hainan and Shenzhen)	A set of waste inputs (municipal fecal residue, kitchen food waste)	Biogas production	Logistic regression, SVM, RF, XGBoost, and kNN	kNN	0.87	–	Clercq et al. (2019)
4 years of operational data from an AcoD facility	Waste type and daily input volume, electricity and water consumption, and auxiliary chemical inputs	Daily biomethane production	Elastic net, RF, XGBoost	RF and XGBoost	0.88	–	Clercq et al. (2020)
17 samples with the same AD configuration from published works	Total carbon, total nitrogen, C/N ratio, cellulose, xylan, lignin and glucan content, and temperature	Methane yield	RF, SVM and kNN	GLMNET and KNN	0.73 (GLMNET)	26.6 (kNN)	Wang et al. (2020)
29 sets of experimental data from 9 published works on ZVI-based AD reactors	TSf, VSf, sCOD, ISR, Tsi, Vsi, pH of feedstock and inoculum, temperature, ZVI dosage, and ZVI particle size	Methane production	RF, XGBoost, and DL	XGBoost	–	21.09	Xu et al. (2021)
50 samples of data-sets from lab-scale	Genomic data, VFAs, temperature, OLRs, HRT, and waste types	Methane yield	RF, kNN, ANN, and XGBoost	RF	0.82	0.043	Long et al. (2021)
360 days data from a lab-scale reactor with food waste	pH, alkalinity, COD, VFA	Methane yield	RF, Xgboost, SVR, RNN	RNN	0.9731	0.023	Park et al. (2021)

VFA: Volatile fatty acids; ZVI: Zero-valent ion; ANFIS: Adaptive network-based fuzzy inference system; ANN: Artificial neural networks; C/N: Carbon-to-nitrogen ratio; COD: Chemical oxygen demand; GMLET: Logistic regression multiclass; HRT: Hydraulic retention time; kNN: k-nearest neighbors; OLR: Organic load ratio; RF: Random forest; RMSE: Root mean squared error; RNN: Recurrent neural network; SVM: Support vector machine; TAN: Total ammonia nitrogen; TVS: Total volatile solids; AcoD: Anaerobic co-digestion; AD: Anaerobic digestion; XGBoost: Extreme gradient boosting.

The dataset in this study has a small number of data points and has a sparse nature since data points have been compiled from different sources without considering the similarity between their chemical components or their reaction conditions. To clarify, each of the substrates can be protein-rich or fat-rich, or they can be converted to methane through different reaction pathways (thermophilic or mesophilic) or different types of reactors (anaerobic sequencing batch reactors, tubular reactors, baffled digesters, up-flow anaerobic sludge blanket), etc. Therefore, capturing the pattern of this dataset is difficult via machine learning models. However, by leveraging the VAE technique, this study achived a promising results with an RMSE of 0.165. While Xu et al. (2021) used data points that all are based on zero-valent ion AD reactions and chose the same machine learning model (XGBoost) as their best model, this study's result is way better than Xu et al. results with an RMSE of 21.09. Moreover, Wang et al. (2020) considered 17 samples with only the same AD configuration, and they achieved an RMSE of 26.6 with K nearest neighbor model. On the other hand, Park et al. (2021) and Clercq et al. (2020) modeled time seris dataset. Clercq et al. (2020) applied intensive feature engineering techniques and turned the task from regression to classification to achive a better results. Unlike Clercq et al. (2020) , Park et al. (2021) did not go through intensive data preprocessing and feature engineering leveraged Recurrent Neural Network (RNN) architecture and several machine learning models. They chose RNN since this model is more suitable for time seris tasks rather than other applied machine learning models and may not require that intensive feature engineering process.

Future studies could expand the scope and enhance the work by incorporating two key approaches. First, increasing the size of the initial dataset can provide broader coverage of substrates, thus facilitating a more robust analysis and ensuring the model's generalizability. Second, further investigation is warranted to consider the synergistic effects of mixed waste

substrates in real anaerobic digestion scenarios. Accounting for these interactions will enhance the accuracy and applicability of the pipeline for real-world applications.

5.5. CONCLUSION

This study employed XGBoost, Random Forest, and CatBoost models to predict the biogas yield of various substrates based on their chemical components. Additionally, it investigated the significance of each organic component in determining the methane yield of the biogas. Due to the limited availability of data in the biogas domain for AI modeling, a deep learning-based approach known as VAE was employed to generate synthetic data that mimicked the statistical characteristics of the original dataset. Following a meticulous hyperparameter tuning process using gridsearch cross validation, the XGBoost model outperformed all other models, achieving an MSE of 0.027 and an RMSE of 0.165. Notably, the feature importance analysis conducted by XGBoost revealed that the "fats" category emerged as the most influential group of chemical components in methane production in biogas while "proteins" group showed the lost effect on bio methene production.

REFERENCES

Heiker, M., Kraume, M., Mertins, A., Wawer, T., & Rosenberger, S. (2021). Biogas Plants in Renewable Energy Systems—A Systematic Review of Modeling Approaches of Biogas Production. In Applied Sciences (Vol. 11, Issue 8, p. 3361). MDPI AG. https://doi.org/10.3390/app11083361

Xu, F., Khalaf, A., Sheets, J., Ge, X., Keener, H., & Li, Y. (2018). Phosphorus Removal and Recovery From Anaerobic Digestion Residues. In Advances in Bioenergy (pp. 77–136). Elsevier. https ://doi.org/10.1016/bs.aibe.2018.02.003

Cruz, I. A., Chuenchart, W., Long, F., Surendra, K. C., Andrade, L. R. S., Bilal, M., … & Ferreira, L. F. R. (2022). Application of machine learning in anaerobic digestion: Perspectives and challenges. *Bioresource Technology*, *345*, 126433.

Levis, J. W.; Barlaz, M. A.; Themelis, N. J.; Ulloa, P., Assessment of the state of food waste treatment in the United States and Canada. *Waste Management* **2010,** 30, (8), 1486-1494.

Nordin, N. H., Kaida, N., Othman, N. A., Akhir, F. N. M., & Hara, H. (2020, June). Reducing Food Waste: Strategies for Household Waste Management to Minimize the Impact of Climate Change and Contribute to Malaysia's Sustainable Development. In *IOP Conference Series: Earth and Environmental Science* (Vol. 479, No. 1, p. 012035). IOP Publishing.

Sabour, M. R., Alam, E., & Hatami, A. M. (2020). Global trends and status in landfilling research: a systematic analysis. *Journal of Material Cycles and Waste Management*, *22*, 711-723.

Jaunich, M. K., Levis, J. W., DeCarolis, J. F., Barlaz, M. A., & Ranjithan, S. R. (2019). Solid waste management policy implications on waste process choices and systemwide cost and greenhouse gas performance. *Environmental science & technology*, *53*(4), 1766-1775.

Mulu, E.' M'Arimi, M. M., & Ramkat, R. C. (2021). A review of recent developments in application of low cost natural materials in purification and upgrade of biogas. *Renewable and Sustainable Energy Reviews*, *145*, 111081.

Bekchanov, M., Mondal, M. A. H., de Alwis, A., & Mirzabaev, A. (2019). Why adoption is slow despite promising potential of biogas technology for improving energy security and mitigating climate change in Sri Lanka?. *Renewable and Sustainable Energy Reviews*, *105*, 378-390.

Zabed, H. M., Akter, S., Yun, J., Zhang, G., Zhang, Y., & Qi, X. (2020). Biogas from microalgae: Technologies, challenges and opportunities. *Renewable and Sustainable Energy Reviews*, *117*, 109503.

Naik, G. P., Poonia, A. K., & Chaudhari, P. K. (2021). Pretreatment of lignocellulosic agricultural waste for delignification, rapid hydrolysis, and enhanced biogas production: A review. *Journal of the Indian Chemical Society*, *98*(10), 100147.

Akbay, H. E. G., Deniz, F., Mazmanci, M. A., Deepanraj, B., & Dizge, N. (2022). Investigation of anaerobic degradability and biogas production of the starch and industrial sewage mixtures. *Sustainable Energy Technologies and Assessments*, *52*, 102054.

Deschamps, L., Imatoukene, N., Lemaire, J., Mounkaila, M., Filali, R., Lopez, M., & Theoleyre, M. A. (2021). In-situ biogas upgrading by bio-methanation with an innovative membrane bioreactor combining sludge filtration and H2 injection. *Bioresource Technology*, *337*, 125444.

Kougias, P. G., & Angelidaki, I. (2018). Biogas and its opportunities—A review. *Frontiers of Environmental Science & Engineering*, *12*, 1-12.

Filer, J., Ding, H. H., & Chang, S. (2019). Biochemical methane potential (BMP) assay method for anaerobic digestion research. *Water*, *11*(5), 921.

Argiz, L., Reyes, C., Belmonte, M., Franchi, O., Campo, R., Fra-Vázquez, A., ... & Campos, J. L. (2020). Assessment of a fast method to predict the biochemical methane potential based on biodegradable COD obtained by fractionation respirometric tests. *Journal of Environmental Management*, *269*, 110695.

Tsapekos, P., Kougias, P. G., & Angelidaki, I. (2015). Biogas production from ensiled meadow grass; effect of mechanical pretreatments and rapid determination of substrate biodegradability via physicochemical methods. *Bioresource technology*, *182*, 329-335.

Bekiaris, G.; Triolo, J. M.; Peltre, C.; Pedersen, L.; Jensen, L. S.; Bruun, S., Rapid estimation of the biochemical methane potential of plant biomasses using Fourier transform mid-infrared photoacoustic spectroscopy. *Bioresource Technology* **2015,** 197, 475-481.

Godin, B., Mayer, F., Agneessens, R., Gerin, P., Dardenne, P., Delfosse, P., & Delcarte, J. (2015). Biochemical methane potential prediction of plant biomasses: comparing chemical composition versus near infrared methods and linear versus non-linear models. *Bioresource Technology*, *175*, 382-390.

Buswell, A. M., & Mueller, H. F. (1952). Mechanism of methane fermentation. *Industrial & Engineering Chemistry*, *44*(3), 550-552.

Jingura, R. M., & Kamusoko, R. (2017). Methods for determination of biomethane potential of feedstocks: a review. *Biofuel Research Journal*, *4*(2), 573-586.

Thomsen, S. T., Spliid, H., & Østergård, H. (2014). Statistical prediction of biomethane potentials based on the composition of lignocellulosic biomass. *Bioresource technology*, *154*, 80-86.

Doublet, J.; Boulanger, A.; Ponthieux, A.; Laroche, C.; Poitrenaud, M.; Cacho Rivero, J. A., Predicting the biochemical methane potential of wide range of organic substrates by near infrared spectroscopy. *Bioresource Technology* **2013,** 128, 252-258.

Calabrò, P. S., Catalán, E., Folino, A., Sánchez, A., & Komilis, D. (2018). Effect of three pretreatment techniques on the chemical composition and on the methane yields of Opuntia ficus-indica (prickly pear) biomass. *Waste Management & Research*, *36*(1), 17-29.

Yan, H., Zhao, C., Zhang, J., Zhang, R., Xue, C., Liu, G., & Chen, C. (2017). Study on biomethane production and biodegradability of different leafy vegetables in anaerobic digestion. *AMB Express*, *7*, 1-9.

Zheng, W., Phoungthong, K., Lü, F., Shao, L. M., & He, P. J. (2013). Evaluation of a classification method for biodegradable solid wastes using anaerobic degradation parameters. *Waste management*, *33*(12), 2632-2640.

Bank, D., Koenigstein, N., & Giryes, R. (2020). Autoencoders. *arXiv preprint arXiv:2003.05991*.

Shaham, T. R., Dekel, T., & Michaeli, T. (2019). Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4570-4580).

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). Ladder variational autoencoders. *Advances in neural information processing systems*, *29*.

Akrami, H., Joshi, A. A., Li, J., Aydöre, S., & Leahy, R. M. (2022). A robust variational autoencoder using beta divergence. *Knowledge-Based Systems*, *238*, 107886.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079-2107.

He, J., Wang, X., Yin, X. B., Li, Q., Li, X., Zhang, Y. F., & Deng, Y. (2018). Insights into biomethane production and microbial community succession during semi-continuous anaerobic digestion of waste cooking oil under different organic loading rates. *AMB Express*, *8*(1), 1-11.

Saady, N. M. C., & Massé, D. I. (2015). Impact of organic loading rate on psychrophilic anaerobic digestion of solid dairy manure. Energies, 8(3), 1990-2007.

Asadi, M., & McPhedran, K. N. (2021). Greenhouse gas emission estimation from municipal wastewater using a hybrid approach of generative adversarial network and data-driven modelling. *Science of The Total Environment*, *800*, 149508.

Xu, R. Z., Cao, J. S., Wu, Y., Wang, S. N., Luo, J. Y., Chen, X., & Fang, F. (2020). An integrated approach based on virtual data augmentation and deep neural networks modeling for VFA production prediction in anaerobic fermentation process. *Water Research*, *184*, 116103.

Long, F., Wang, L., Cai, W., Lesnik, K., & Liu, H. (2021). Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data. *Water Research*, *199*, 117182.

Xu, W., Long, F., Zhao, H., Zhang, Y., Liang, D., Wang, L., ... & Liu, H. (2021). Performance prediction of ZVI-based anaerobic digestion reactor using machine learning algorithms. *Waste Management*, *121*, 59-66.

Cata Saady, N. M., & Massé, D. I. (2015). Impact of Organic Loading Rate on Psychrophilic Anaerobic Digestion of Solid Dairy Manure. *Energies (19961073)*, *8*(3).

Shi, C. (2012). Potential Biogas Production from Fish Waste and Sludge.

Bellaton, S., Guérin, S., Pautremat, N., Bernier, J., Muller, M., Motellet, S., ... & Rocher, V. (2016). Early assessment of a rapid alternative method for the estimation of the biomethane potential of sewage sludge. *Bioresource Technology*, *206*, 279-284.

Spanjers, H., & van Lier, J. B. (2006). Instrumentation in anaerobic treatment–research and practice. *Water Science and Technology*, *53*(4-5), 63-76.

Esteves, S., Miltner, M., & Fletch, S. (2012). Monitoring review and guide for the optimisation of anaerobic digestion and biomethane plants. *Full Report*.

Najafi, B., & Faizollahzadeh Ardabili, S. (2018). Application of ANFIS, ANN, and logistic methods in estimating biogas production from spent mushroom compost (SMC). In Resources, Conservation and Recycling (Vol. 133, pp. 169–178). Elsevier BV. https://doi.org/10.1016/j.resconrec.2018.02.025

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1. SUMMARY

In this study, for the first time, AI applications in renewable energy systems were comprehensively investigated using machine learning and natural language processing methods. In this regard, An algorithm was developed to retrieve unlimited documents from the Scopus database and preprocess them for analysis. total number of 5561 documents for time interval of 2000-2021 were extracted and BERTopic was utilized to perform DTM and extract the main research themes. The temporal evolution of extracted 7 metatopics were assessed where wind and photovoltatic systems gaind the frist and second rank respectively while . biohydrogen gained the last rank. Also, the industrial scale biogas plant's daily output was predicted using a hybird deep learning architecture (CNN-LSTM) and the impact of each input on daily biogas production was determined using an ensemble learning model (Catboost). The original daily-based dataset consisted of 18 variables, including separate measurements of inputs and outputs from tanks measured over 441 days. To analyze the entire system, certain features were considered, and new ones were created. For example, "1_acidification_hydrolysis tank feed" and "2_acidification_hydrolysis tank feed" features were summed to create a new feature called "acid feed". In total, 12 featuers were extracted out of 18 features. Different machine learning algorithms, including RF, Catboost, and XGBoost, were trained to assess the importance of each parameter on the system's performance. Catboost was selected due to its stability and low error (NMSE= 0.015). The results revelead th"t "Acid Feed" had the greatest impact on the biogas

system and it followed by “waste oil” and “anaeroic feed” . Finally, six key features were selected out of twelve features. After the feature selection process, a hybrid CNN-LSTM deep learning architecture was utilized to forecast system performance. It achieved significant results with a normalized mean squared error (NMSE) of 0.078, outperforming the other single DL architectures, LSTM (NMSE of 0.089) and GRU (NMSE of 0.144). Moreover, a pipeline was developed to predict biomethen potetnional of various substrates. The original dataset was obtained from Chen and Saady (2023) study that contains 75 substrates with their chemical component and experimental methene yeild. Considering data limitation in this problem for training a supervised machine learning model, a deep learning approach called Variational Autoencoder (VAE) was employed to overcome data limitation. The generated dataset with 500 synthetic data points has the similar statistical characteristics to the real dataset. XGBoost, Random Forest, and CatBoost models were utilized in this study to predict the biogas yield of different substrates based on their organic components. The XGBoost model outperformed the other models, achieving a Mean Squared Error (MSE) of 0.027 and a Root Mean Squared Error (RMSE) of 0.165.

6.2. CONCLUSIONS

The major results obtained in this study are listed below:

- Most frequent emoloyed algorthims in renewable energy systems’ publications were investigated during the stuied preiod. The emergence ‘deep learning’ as a major technique was followed by a rise in complex topics with a high level of uncertanity such as power cost optimization and wind prediction.
- The results revealed the increasing use of machine/deep learning techniques in analyzing renewable energy data, specifically in wind and solar photovoltaic systems.

- The employed pipeline in this research, BERTopic, does not require an intensive data preproccessing which makes it more efficient than conventional models such as LDA.
- The research themes and trends reflected significant recent investment in advanced AI learning techniques, marking a shift from conventional methods.
- This systematic investigation can enhance the strategic decesion making in renewable eneryg systems.
- The utlized hybrid Catboost-CNN-LSTM pipeline achieved significant results in forecasting the system's performance. The developed method successfully predicted the industrial-scale biogas plant's yield and effectively identified the importance of each system input. This pipeline can handle time series problems in various environmental fields such as water, air, or soil.
- The developed pipeline introduced an improved method compared to previous relevant studies by eliminating the need for feature engineering and allowing direct prediction of biogas yield without conversion to a classification task.
- Developed prototype can be scaled to be utlized in production by consdiering more environmental and technical parameters in real-time using IoT devices.
- A deep learning approach called VAE was employed to overcome data limitation by generating synthetic data replicating the statistical characteristics of the original dataset of various substrates.
- The XGBoost model outperformed the other empployed enasmble models, The significance of each organic component in determining the methane yield of biogas was investigated. The analysis of feature importance revealed that the category of "fats"

emerged as the most influential group of chemical components in methane production in biogas.

- The developed pipeline can be used for different studies dealing with data limitation.
- The main limitation of this study was lack of d

6.3. RECOMMENDATION FOR FUTURE STUDIES

The current study investigated AI applications in renewable energy. It investigated the importance of daily inputs to an industrial biogas plant, predicted its daily biogas production, and predicted the potantional mehtene yeild of diverese substates using machine learning models based on their organic components. To increase the scope of each study and get more comprehensive and realistic results, future studies are recommended to:

- Explore other scientific databases, including Web of Science, PubMed, IEEE Xplore, ScienceDirect, or even consider additional data sources, such as patents, to complement the exploration of the research landscape.
- Consider the entire body of the accessible documents, focusing on the methodology section to gain insights into methodological evolution instead of only using abstract and title.
- Analyze published documents in other languages and incorporate other natural language processing (NLP) techniques such as GPT3 and WuDao 2.0 for their analyses.
- Consider the automatic topic labeling process by pre-trained models such as T5 or other encoder-decoder architectures.
- Consider operational parameters such as pH, Organic Loading Rate (OLR), Hydraulic Retention Time (HRT), temperature, and microbiological features of the microorganisms

into the dataset with precise and recorded data to enhance data-driven analyses and make it more reliable in addressing real-world problems.

- Use optimization algorithms such as GA (genetic algorithm) to find the most values for each feature to achieve the highest performance.
- Increase the size of the initial dataset to provide broader coverage of substrates, enabling a more comprehensive and robust analysis and enhancing the model's generalizability.
- Explore the synergistic effects of mixed waste substrates in real anaerobic digestion scenarios to make the pipeline applicable for real-world applications.