Republic of Iraq
Ministry of Higher Education and
Scientific Research
University of Babylon
College of Science for Women
Department of Computer Science

Improving XGBoost Algorithm using Optimization Technique

A Thesis
Submitted to the Council of the College of Science for Women at University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Science \ Computer Science

By

Hadeer Majed Abdul-Hussein

Supervised By

Prof. Dr. Samaher Al-Janabi **Assist.Prof.Dr. Saif Al-alak**

2022 A.D. **1444 A.H**

بسم الله الرحمن الرحيم

اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ ۞ خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ ۞ اقْرَأْ وَرَبُّكَ الْأَكْرَمُ ۞الَّذِي عَلَّمَ بِالْقَلَمِ ۞ عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ۞

صدق الله العلي العظيم

سورة العلق: الاية [1- 5]

Dedication

To "ALLAH" my Lord, my Creator, my dependents and my hope.

To the messenger who reached the valley of safety and advised the nation.

To the Prophet of mercy and the light of the worlds "Prophet Muhammad peace be upon him".

To my dear father, who taught me tender without waiting for someone who always prides me … asking God to extend in his life to witness my superiority and success after a long wait.

To my dear mother, whose invitations were the secret of my success and her tenderness.

To my brothers who support me and they were the secret of my success.

To my sisters who support me and they were the secret of my success.

To all members of my family and friends who have spared no effort in my encouragement and support.

To the College of Science for Girls, especially the Computer Department, this gave me the opportunity to complete the Masters and the teaching staff who were the reason for giving me knowledge.

Hadeer Majed Abdul-Hussein
2022

Acknowledgments

All Praises be to "ALLAH" Almighty who has enabled me to complete this task successfully and our utmost respect to His last Prophet "Mohammad",

I would like to express my gratitude and thanks to my supervisors "**Prof. Dr. Samaher Al-Janabi and Assist.Prof.Dr. Saif Al-alak**" who have never stopped encouraging me and always give my guidance, strength and help throughout this thesis.

I also would like to express my wholehearted thanks to **my family (father, mother, brothers, and sisters)** for their generous support they provided me throughout my entire life and particularly through the process of pursuing the master degree. Because of their unconditional love and prayers, I have the chance to complete this thesis.

Finally, I would also like to thank the College of Science for Girls, especially the Department of Computer, which gave me the opportunity to complete my masters and the teach staff who were the reason for giving me the knowledge.

Hadeer Majed Abdul-Hussein
2022

Supervisors Certification

We certify that this thesis entitled "**Improving XGBoost Algorithm using Optimization Technique**" was done by (**Hadeer Majed Abdul-Hussein**) under our supervision.

Committee Member (**Supervisor**)

Signature:

Name: **Samaher Al-Janabi**

Scientific order: Prof. Dr.

Address: University of Babylon/College of Science for Women

Date: / / 2022

Committee Member (Supervisor)

Signature:

Name: **Saif Al-alak**

Scientific order: Assist. Prof. Dr.

Address: University of Babylon/College of Science for Women

Date: / / 2022

The Head of the Department Certification

In view of the available recommendations, I forward the dissertation entitled "**Improving XGBoost Algorithm using Optimization Technique**" for debate by the examination committee.

Signature:

Name: **Assist.Prof.Dr.Saif M. Al-Alak**

Date: / / 2022

Address: University of Babylon/College of Science for Women

Abstract

Natural gas is one of the most important natural resources and a major source of energy, and given the importance of natural gas in many fields, including industry and commercial, and because of its direct impact on the life of living organisms, where gases vary in their degree of importance, some of them are considered essential for the continuation of life. Such as oxygen gas to humans and carbon dioxide to plants, and some are considered deadly because they contain a percentage of toxicity and as a result of the rapid spread of this, building systems with various purposes to deal with natural gases is one of the most important challenges facing the world today.

XGBoost is one of the best prediction algorithms related to data mining, as it gives high accuracy results and can deal with high sensitivity data such as health care data and rapidly spreading pollutants such as air, water and soil pollutants, but on the other hand, it contains many gaps, including that Its core / kernel is decision trees (DTs), which are characterized by several parameters, such as the root must be chosen, the depth of the tree, the number of terminal nodes. This thesis presents a solution to overcome these obstacles by replacing the XGBoost kernel with one of the agent optimization technique called GSK, after being developed with a new fitness function along with it.

This thesis presents a hybrid model to predict one of six types of natural gas called (HPM-STG). This model consists of four basic stages: ***The first stage*** is to obtain data from a source dedicated to scientific research related to natural gas. ***The second stage:*** conducting a preprocessing of the data. This stage was divided into several steps: (a) Checking the missing values and deleting any row that contains missing numbers. *(B) Calculating the correlation between the characteristics and the target.* As for ***the third stage***, the data was divided into training and testing and a prediction model named (***DGSK-XGB***) was built. ***The fourth stage***, evaluation the results based on five evaluation measures (i.e., accuracy, pression, recall, f-measurement, and Fb).

The designed hybrid model was characterized as a promising tool in classifying specific types of gas, include in six types, as it can analyze a large amount of data in a relatively short period of time. Also, replacing the XGBoost kernel with GSK gives results with relatively high accuracy compared to the traditional method of XGBoost.

To demonstrate the feasibility of the proposed hybrid model. The developed algorithm was compared with the traditional algorithm in terms of accuracy and execution time, and the developed method proved to be 90% accurate compared to the traditional method 50%, which confirms the preference of the developed method.

List of Tables

List of Figures		
3.1	Block diagram of DGSK-XGB Model	30
4.1	Data Merging Method	38
4.2	Compare traditional XGBoost with DXGBoost from aspect accuracy	51
4.3	Compare traditional XGBoost with DXGBoost from aspect time	52
4.4	Traditional XGBoost	52
4.5	DXGBoost-GSK	53

List of Algorithms		
2.1	GSK	15
2.2	XGBoost	22
3.1	Hybrid Prediction Model for Six Types of Gas (HPM-STG)	29
3.2	Preprocessing	32
3.3	DXGBoost-GSK	33

List of Abbreviation	
ALO	Ant Lion Optimization
ANN	Artificial Neural Networks
CPU	Central Processing Unit
CNN	Convolution Neural Network
DT	Decision Tree
ETC	Extra Trees Classifier
FPGA	field-programmable gate array
FAR	Functional autoregressive
GSK	Gaining-sharing knowledge
GPR	Gaussian Process Regression

GA	Genetic Algorithm
GBMs	Gradient Boosting Machines
GPU	Graphics processing unit
HPM-STG	Hybrid Prediction Model for Six Types of Gas
ISSA	Improved SSA
IWOA	improved whale optimization algorithm
IDA	Intelligent Data Analysis
LSTM	long short-term memory
MAD	mean absolute deviation
MAE	Mean Absolute Error
MAPE	mean absolute percentage error
MARNE	mean absolute range normalized error
MLP	multi-layer perceptron
PSO	Particle Swarm Optimization
RF	Random Forest
RNN	recurrent neural network
RMSE	Root Mean Square Error
R^2	the coefficient of determination
SSA	Singular spectrum analysis
SL	supervised learning
SVM	Support Vector Machine
SI	Swarm Intelligence
USCPs	Unique Standard Consumption Profiles
XGBoost	Extreme Gradient Boosting

Chapter One:

General Introduction

Chapter One: General Introduction

1.1 Introduction

The process of emission of gases in laboratories, or as a result of extracting some raw materials from the earth, or as a result of respiration of living organisms, is one of the most important processes for sustaining life. In general, these gases are divided into two types, some of them are poisonous and cause problems to the life of living organisms, and the other type is useful and necessary and used in many industries. Therefore, this thesis attempts to build a model that classifies six basic types of those gases, which are (Ethanol, Ethylene, Ammonia, Acetaldehyde, Acetone, and Toluene).

The basic components of natural gas are (Methane (c1), Non-hydrocarbons (H2O, CO2, H2S), NGL (Ethane (c2), pentane (c5), and heavier fractions), LPG (propane (c3), Butane(c4)). To leave solely liquid natural gas, shall eliminate both methane and non-hydrocarbons (water, carbon dioxide, hydrogen sulfide). Natural gas emits less CO2 than petroleum, which emits less CO2 than coal. The first choice is usually to save money and increase efficiency. One of the benefits of natural gas is that it burns completely when used, and unlike other traditional energy sources, the carbon dioxide produced when burning is absolutely non-toxic Natural gas is a pure gas by nature, and any contaminants that may be present in it may sometimes be simply and inexpensively eliminated. Natural gas stations are not generally distributed. Natural gas has a number of drawbacks, including the fact that extraction may be hazardous to the environment and necessitates the use of a pipeline, as well as the fact that methane leaks contribute to global warming. It asserts that increasing the pressure on gas at a constant temperature reduces the volume of the gas. In other words, Boyle's law asserts that volume is inversely proportional to pressure when the temperature and number of molecules stay constant [Rami J. Batrice and John C. Gordon,2021]

Intelligent Data Analysis (IDA) Among one of the pragmatic fields in computer science based on integration among the data Domain, Mathematical domain, and Algorithm domain; In general, to handle any problem through IDA must satisfy the following :(a) real problem: must find one of the real problems in one of the specific field of life, (b) design a new or a novel or hybrid model to solve it based on the integration among the above Three domains; (c) interpretation the result after analysis it to become understand and useful for any person not only for the person expert in the specific field of problem.

This thesis handle the main problem of Natural Gas that description in the above section by designing the hybrid model based on develop one of predict data mining technique through the optimization principle.

1.2 Problem Statement

The problem of this work is separated into sections: Part one is related to programming challenges while; the second part is related to application challenges; In general; the prediction techniques are split into two fields; prediction techniques related to data mining and predictions related to neurocomputing; this work deal with the first type of prediction technique.

- XGboost is one of the data mining prediction techniques that characterized by many features that make it the best. these features (include XGboost give high accuracy results and work with huge data/ stream data in real time but on other hand; the core of that algorithm is decision tree (DT) that have many limitations such as it requires choose the root of tree, determined the max number of levels of tree, also it have high computation and time of implementation. Therefore; the first challenge of this thesis is how can avoid these limitations (i.e., high computation and time of implementation) of this algorithm and befit from their features.

- The problem of prediction the types of natural gas need to high efficiency techniques; Therefore, the second challenge of this thesis is how can avoid these limitations thought build an efficient technique to predict multi types of gas coming from different sensors.

1.2.1 Research Questions

- How Gaining-Sharing Knowledge-based (GSK) can be useful in building a new predictor called DGSK-XGB?
- How can build a prediction model by replacing the kernel of XGboost with Gaining-Sharing Knowledge-based (GSK)?
- Is used enough measurement to assess the results of the proposed predictor?

1.3 Research Objectives

This section shows the main objectives attempt this work to reach of it:

- Solve the gap of XGboost through suggest a new kernel of it based on an optimization algorithm called GSK.
- Build predictor to find types of gas that are generated from different types of resources.
- Increase the accuracy through a combination of GSK and XGboost; at the same time reduce the time of implementation.

1.4 Literature Survey

The issue of perdition the types of Natural gas is one of the key issues related directly to people's lives and the continued of a healthful environment. Since the topic of this research is to find a recent predictive way to deal with types of data that is sensitive and performs within the range of data series, in this part of the thesis, we will try to review the works of past researchers in the same area of our issue and comparing works with seven basis points.

Ivan Smajla et.al.,2021 The optimal number of smart meters and the most appropriate distribution of consumption data to achieve acceptable accuracy using basic prediction algorithms. The findings show that the availability of high-resolution data on natural gas consumption generated from smart gas meters has a considerable influence on natural gas forecasting system accuracy. Demonstrate it as well. Especially with the partial installation of smart meters. It is necessary to properly locate smart meters while installing them in a region or city.The work is similar in terms of using the same structure but differs in minute details.

Weibiao Qiao et.al.,2020 A hybrid model based on the adaptive volterra filter and the modified whale optimization algorithm was proposed for estimating short-term natural gas demand. This work is similar to our work in design forecaster to predict the type of natural gas but use the short term based on Volterra.

Jiyuan Zhang et.al.,2020 Suggest extreme gradient boosting as a supervised learning approach (XGboost). The work is similar in terms of using XGboost technology, but the difference is that replaces the core of the algorithm with another kernel rather than DT.

Meng Meng et.al.2020 In this work Adsorption models are transformed into classic pressure/density dependent isothermal models, pressure and temperature unified models, and machine learning-based models. The work is similar in terms of using the XGBoost algorithm but the difference with the work is to replace its core with another algorithm.

Weibiao Qiao et.al.2020 The goal of this research is to integrate the multi-layer Perceptron (MLP) neural network method with five metaheuristic computational techniques. The work is similar in terms of creating a hybrid model by relying on Nature-inspired and the difference is that the work is the use of other techniques.

J. Ravnik and M. Hribersek, 2019 construct a strategy for implementing the defined profiles for predicting and preliminary gas allocation, as well as build unique standard gas consumption profiles for end gas users The work's outcome is based on an assessment of gas consumption, air temperature, and the creation of unique standard gas consumption profiles for different customer groups. also, this work used the sigmoid model function to create the consumption profiles. our work is specific to build predictor for multi type of natural gas therefore; it is similar that work in this point but, we used different techniques and evaluation measures.

Rok Hribar et.al.,2019 present a multi model to forecasting of home natural gas. The application was tested and compared in an urban environment. Models forecast gas consumption with hourly precision up to 60 hours in advance. The model's projections are based on historical and forecast temperatures, as well as temporal factors such as holiday markers and other episodic occurrences. The work is comparable to the prediction concept that exists, but the distinction is that the application is deployed in an urban area and compares.

Table 1: Compare the Previous Works

Name of authors	Dataset \ Database	Preprocessing	Methodology	Evaluation measures	Advantage	Disadvantages
(Ivan Smajla et.al.2021)	▪ Natural gas	▪ Data processing and filtering	▪ Developing an accurate model for natural gas consumption Forecasting.	▪ Accuracy= 0.9	▪ the appropriate number of smart meters and the best-fitting distribution of consumption data in order to produce good results in terms of accuracy by applying basic forecasting methods.	▪ The most significant disadvantage of installing smart meters is that it is an expensive and time-consuming process.

(Weibiao Qiao et.al.2020)	▪ natural gas	▪ Noize-filtering by using gaussian smoothing ▪ Chaotic character recognition by using a small data quantity method ▪ Phase space reconstruction by applying chaotic character (C-C) method ▪ Normalization	▪ Volterra adaptive filter and an improved whale optimization algorithm.	▪ R^2=0.99	▪ Short-term natural gas consumption is a critical fundamental indicator for the construction of natural gas pipeline networks, the creation of special natural gas planning, and the intelligent scheduling of natural gas.	▪ The developed prediction model offers good accuracy in projecting short-term natural gas usage but low efficiency.
(Jiyuan Zhang et.al.2020)	▪ gas	▪ The database is divided into subgroups, and each subgroup dataset is put into the model sequentially to generate trees. ▪ Trees are built with the goal of increasing the prediction accuracy of prior predictors by modifying weights to erroneously predicted points in the past.	▪ XGBoost trees to estimate the interfacial tension (IFT) between n-alkane and gases.	▪ R^2=0.999	▪ It gives high accuracy.	▪ This conclusion contradicts the generally accepted relevance correlation, which undervalues the influence of gas composition while exaggerating the effect of temperature.
(Meng Meng et.al.2020)	▪ Shale gas	▪ outliers' detection	▪ capability of machine learning for prediction of shale gas.	▪ R^2= 0.99	▪ The benefit of this unified model is that it can extend predictions beyond test ranges.	▪ It is still restricted to a certain type of shale.

(Weibiao Qiao et.al.2020)	▪ Gas	▪ Data collections ▪ Simulation ▪ Dataset arregement	▪ Combine the multi-layer Perceptron (MLP) neural network technique with five metaheuristic computational algorithms.	▪ R^2=0.9999	▪ The projected findings show that combining optimization methods can enhance prediction accuracy and help the MLP perform better.	▪ The planned network output values are close to the actual values. It remains difficult to provide the best-fit approximation approach based on the acquired training and prediction testing networks in order to assess all applied error performance systems.
(J. Ravnik and M. Hribersek, 2019)	▪ Gas	▪ Compute correlation coefficient	▪ The Unique Standard Consumption Profiles (USCPs).	▪ Accuarcy= 0.94	▪ The key benefit of this technique is its simplicity, which makes it simple to apply in the workflow of gas supply firms.	▪ Although knowledge of the temperature independent part of the consumption yields superior predictions of gas consumption, it should be avoided when data on temperature independent gas consumption is not available correctly.
(Rok Hribar et.al.2019)	▪ natural gas	▪ Compute correlation coefficient.	▪ different forecast models for natural gas.	▪ MAE ▪ MAPE	▪ An accurate forecast model may even be able to remove the requirement for costly energy reserves entirely.	▪ It should be highlighted that RNN has the drawback of requiring a lot of resources for training. Its implementation is additionally very expensive due to recurrent loops and instability caused by particularly lengthy sequences.

1.5 Thesis Layout

The remained chapters of this thesis organization as the following:

- ***Chapter two:*** gives an overview of the main theoretical background in this field.
- ***Chapter three:*** explains a build of the predictor based on the techniques of intelligent data analysis.
- ***Chapter four:*** illustrates the implementation of (DGSK-XGB) and the results of the case study.
- ***Chapter five:*** shows conclusions of this work together with some recommendations for future work in this field.

Chapter Two:

Theoretical Background

Chapter Two: Theoretical Background

2.1 Introduction

This chapter focuses on the main concepts, algorithms, and measures used to treat the problem presented in chapter one to determine the prediction of natural gas. In general, this chapter includes three main aspects:

The first aspect is an optimization which is used to obtain the best groups under certain conditions. The ultimate goal of all of these options is to either reduce the effort required or maximize the desired benefit. Many optimization algorithms have been clarified, namely "***(***Particle Swarm Optimization (PSO), Ant Lion Optimization (ALO), Genetic Algorithm (GA), and ***Gaining-Sharing Knowledge-based algorithm (GSK))***", then compare among these optimization algorithms to choose the most appropriate algorithm to solve the problem of the present thesis'.

The second aspect is prediction techniques, which also will explain the comparison among the main prediction techniques to determine the algorithm more suitable to solve the problem of that study (i.e., Decision Tree (DT), Random Forest (RF), Extra Trees Classifier (ETC) and ***Extreme Gradient Boosting (XGBoost)***.

The third aspect, discuss the most important criteria of efficiency, as choosing the appropriate standard is critical to ensure the performance of the model. Since no individual use of the efficiency criteria can provide us with a complete explanation of the performance of the model. So, this thesis used the following measures called "Accuracy, Precision, Recall, F-measure, and Fb".

2.2 Optimization Techniques

Optimization is one of the main model in computer science based on find the best values such as max, min or benefit values through optimization function; In general; the optimization model split into single object function model or multi objectives functions model also, some of these models based on constructions while the other not. There are many Techniques can used to find the optimal solution such as:

2.2.1 Particle Swarm Optimization algorithm (PSO)

Eberhart and Kennedy invented one of the swarm intelligence methods, particle swarm optimization (PSO), in 1995. It's a population-based, stochastic algorithm inspired by social behaviours seen in confined birds. It is one of the approaches to evolutionary optimization. Their biological inspiration comes from the concept of social interaction and communication in a flock of birds or a swarm of fish. In each of these groupings, a commander directs the movement of the entire squadron. Each individual's movement is guided by the leader and his own understanding. Individuals (i.e., particles) in The PSO algorithm prefer to follow the group leader, i.e., best performance because PSO is population reliant and evolutionary in nature. So, the particle swarming optimization (PSO) approach is one of the most common Swarm intelligent models. It has been successfully used and has already achieved good results in a range of scientific and technological sectors, with a wider class of complicated improvement issues, and has played a clear role in the improvement field in the previous time, For more detail see Table 2.1 [Kennedy, J., and Eberhart, R.,1995] [Samaher Al-Janabi et.al.,2020] [Samaher Al-Janabi et.al.,2021].

2.2.2 Genetic Algorithm (GA)

Genetic algorithms were developed in 1960 by John Holland at the University of Michigan but did not become popular until the 1990s. Their main goal is to address issues when deterministic techniques are too expensive, And the genetic algorithm is a type of evolutionary algorithm that is inspired by biological evolution. It is the selection of parents, reproduction, and mutation of offspring. The primary goal of evolution is to generate children that are biologically superior to their parents. A genetic algorithm is based on Darwin's Theory of Evolution by Natural Selection and attempts to imitate it. The underlying assumption is that the finest individuals from the population are chosen as parents, and they are asked to expand their generation by reproducing and having children. During the reproduction process, where genes from both parents' crossover, an error known as mutation occurs. These youngsters are urged to reproduce their offspring once again, and the cycle continues, resulting in healthier generations. This idea has influenced evolutionary computation to address issues such as optimization, feature selection, the classic knapsack problem, and many more [Holland, J. H.,1973] [Samaher Al-Janabi et.al.,2018].

The genetic algorithm involves several parameters, such as: [P. Pongcharoen et.al, 2002]

- The crossover: The crossover probability expresses how frequently a crossover be conducted. If no crossover occurs, the child is an identical replica of the parents. If there is a crossover, kids are created from chromosomal pieces from both parents.
- Mutation probability: The probability of a chromosome being mutated is expressed as a percentage. If there is no mutation, children are produced without change following crossover (or copy). When a mutation is accomplished, a portion of a chromosome is changed.

- Population size: The population size represents the number of chromosomes in a population (in one generation). GA has some crossover potential when there are few chromosomes, but only a small percentage of the search region is explored. GA, on the other hand, slows down when too many chromosomes are present.
- The number of generations: It shows the number of generations that are born.

2.2.3 Ant Lion Optimizer (ALO)

Mirjalili created ALO, a Metahorian swarm-based technique, in 2015 to imitate ant hunting behavior in nature. The lion-ant optimizer solves optimization issues by providing a heuristic after-factoring technique. It is an algorithm that is based on population. Antelopes and ants are the primary food sources for people. Ant larvae catch the ants by excavating holes in the sand, causing the ants to tumble toward the center while the antelope prepares to attack. Exploratory ants act as search agents with ants, so, Ant Lion Optimizer (ALO) is an algorithm that mimics the foraging behavior of lion ants. Recently, it has been utilized to handle a wide range of optimization problems. It has various advantages, including the fact that it is simple, scalable, and adaptable, as well as having a good balance between exploration and exploitation [Laith Abualigah et.al, 2021].

2.2.4 Gaining-Sharing Knowledge-based algorithm (GSK)

Nature-inspired algorithms have been widely employed in several disciplines for tackling real-world optimization instances because they have a high ability to tackle non-linear, complicated, and challenging optimization issues. Algorithm for knowledge acquisition and sharing; It is a great example of a modern algorithm influenced by nature that uses real-life behavior as a source of inspiration for problem solutions [Ali Wagdy Mohamed et. al.,2020], [Prachi Agrawal et. Al.,2021].

One of the optimization algorithms is the gaining sharing knowledge-based algorithm (GSK). GSK is based on the notion of acquiring and sharing knowledge over the course of a person's life [Prachi Agrawal et. al,2022].

The algorithm is divided into two stages: [Prachi Agrawal et. al,2022]

- Juniors learning and sharing the stage.
- The stage of senior learning and sharing.

Throughout the course of a person's life, they acquire and exchange knowledge or opinions with others. People in their early middle years obtain information through minor connections such as family members and relatives and desire to share their ideas or opinions with people who may or may not belong to their group. Similarly, adults in their forties and fifties gain information by engaging with coworkers, acquaintances, and so on. They have the knowledge and expertise to assess others and classify them as good or evil. They also discuss their ideas or thoughts with experienced or qualified individuals in order to broaden their knowledge. the process may be mathematically defined as follows: [Prachi Agrawal et. Al.,2021].

- The number of people is assumed in the first step (Number of population size N). Let $f(x_i)$, i= 1,2,… , N. represent an individual in a population $X_{ij} = (X_{i1}, X_{i2},\ldots, X_{iN})$, where N represents the branch of knowledge ascribed to that individual. and F_i i = 1, 2, N_{pop}) are the values of the appropriate objective function [Prachi Agrawal et. al,2022]
- It is necessary to compute the number of dimensions for the junior and senior stages. The number of dimensions that must be altered or updated throughout both phases must be specified, and this is computed using a non-linear decreasing and increasing equation:

$$\mathbf{D_{juniorphase}} = \mathbf{problemsize} * \left(\mathbf{1} - \frac{\mathbf{G}}{\mathbf{GEN}}\right)^{\mathbf{k}} \quad \textbf{(2.1)}$$

$$\mathbf{D_{seniorphase}} = \mathbf{problemsize} - \mathbf{D_{juniorphase}} \quad \textbf{(2.2)}$$

- Junior acquiring information sharing stage: People in their early to mid-twenties get knowledge through their tiny networks at this period. They share their thoughts or abilities with others who may or may not be members of their group because they are curious about others. As a result, people are kept up to date as follows: [Prachi Agrawal et. Al.,2021]

  - The people are ordered in ascending order according to objective function values as X_{best},..., X_{i1}, X_{i+1},...,X_{worst}.
  - For each X_i (i = 1, 2,N_{pop}), choose the best (X_{i1}) and worst (X_{i+1}) $X_i$to gather information, and then choose randomly (X_r) to share knowledge. As a result, the revised new individual is as follows:

$$\mathbf{X_{ij}^{new}} = \begin{cases} \mathbf{X_i + K_f[(X_{i-1} - X_{i+1}) + (X_r - X_i)]}, & \textbf{if } \mathbf{F(X_r) < F(X_i)} \\ \mathbf{X_i + K_f[(X_{i-1} - X_{i+1}) + (X_r - X_i)]}, & \textbf{otherwise} \end{cases} \quad \textbf{(2.3)}$$

where $k_f > 0$ represents the knowledge factor

- Senior acquiring knowledge sharing stage: This stage includes the impact and effect of other individuals (for better or worse) on a person. Individual updating can be calculated as follows: [Prachi Agrawal et. Al.,2021]

  - After sorting people in ascending order, the individuals are divided into three groups (best, middle, and worst) (based on the objective function values). Best individual = 100_P% (X_{P-best}), Middle person = N-2 100_P% (X_{middle}), Worst individual = 100_P% (X_{P-best}).

- Choose two random vectors from the top and bottom $100_P\%$ individuals for the gaining half for each individual X_i, and the third (middle person) is picked for the sharing part. As a result, the new person is [Prachi Agrawal et. al,2022].

$$\mathbf{X_{ij}^{new}} = \begin{cases} \mathbf{X_i + K_f[(X_{p-best} - X_{p-worst}) + (X_{middle} - X_i)]}, \textbf{ if } \mathbf{F(X_{middle}) < F(X_i)} \\ \mathbf{X_i + K_f[(X_{p-best} - X_{p-worst}) + (X_i - X_{middle})]}, \quad \boldsymbol{otherwise} \end{cases} \tag{2.4}$$

where p [0, 1] denotes the percentage of best and worst classes.

Show Algorithm 2.1 explain, the main step of GSK.

Algorithm #2.1: GSK [Ali Waqdy Mohamed, et. al.,2020]

Initialize Parameters: N,k_f, k_r k and p //**create a random initial population** $\mathbf{x_i, i = 1, 2, .. N}$

//Evaluate $\mathbf{f(x_i), \forall_i, i = 1, 2, ... N}$

1: ***For*** *G=1 to* $\mathrm{GEN}^{\max}$

Compute the number of (Gained and shared dims.of both phases)

2: $\mathrm{D_{juniorphase}} = \mathrm{problemsize} * \left(1 - \frac{\mathrm{G}}{\mathrm{GEN}}\right)^{\mathrm{k}}$ // **Junior GSK phase**

3: $\mathrm{D_{seniorphase}} = \mathrm{problemsize} - \mathrm{D_{juniorphase}}$ // **Senior GSK phase**

4: ***IF*** $\mathrm{F}(\mathrm{x_i^{new}}) \leq \mathrm{F}(\mathrm{F_i^{old}})$

5: $\mathrm{x_i^{old}} = \mathrm{x_i^{new}}$

6: $\mathrm{F}(\mathrm{x_i^{old}}) = \mathrm{F}(\mathrm{x_i^{new}})$

7: ***End IF*** // **update each vector**

8: ***IF*** $\mathrm{F}(\mathrm{x_i^{new}}) \leq \mathrm{F}(\mathrm{x_{best}^{G}})$

9: $x_{best}^{G} = x_i^{new}$

10: $\mathrm{F}(\mathrm{x_{best}^{G}}) = \mathrm{F}(\mathrm{x_i^{new}})$

11: ***End IF*** // **update global best**

12: ***End for***

End GSK

In this section, each variable is explained what it means:

- N is population size
- K_f is knowldge factor > 0
- K_r is knowldge ratio belong to (0,1)
- K is knowldge rate
- P is partition size
- G is generation number
- GEN is maximum number of generations.
- $D_{\textbf{junior phase}}$ and $D_{\text{senior phase}}$ is to compute the gaining and sharing dimenation of both phase
- $F\left(x_i^{old}\right)$ is the individuals at initial of gneration
- $F(x_i^{new})$ is the new individuals of generation afte update individual . and i is the number of iteration.
- $F\left(x_{best}^{G}\right)$ is the global best.

The algorithm steps are as follows:

- The data set, which is considered the initial generation.
- Compute Fitness Function for the initial generation.
- Determine the best solution for the initial generation, which is the first index.
- Specifies the worst solution for the initial generation which is the last index.
- The junior stage is calculated for the initial generation.
- Compute the senior stage.
- The step to update the best individuals comes by comparing the old and newer individuals. The best individuals are selected by determining the highest value, and the worst individuals are updated in the same way.

- This is the first generation and continues like this for the rest of the generations until reaches the last iteration.

Table 2.1: Analytic the Advantages and Disadvantages for OptimizationTechniques

O T	Advantage	Disadvantage
PSO [Zahraa Al-Barmani & Samaher Al-Janabi, 2020],[Samaher Al-Janabi & Ayad F. Alkaim,2020]	▪ Simple to put into action ▪ There are a limited number of settings that must be adjusted. ▪ It is possible to compute it in parallel. ▪ The end consequence of it validation ▪ Locate the worldwide best solutions ▪ Convergent quick method ▪ Do not mutate and overlap ▪ Demonstrating a short implantation time	▪ selecting the initial values for its parameters using the concept of trial and error / at random ▪ It only works with scattering issues. ▪ In a complicated issue, the solution will be locked in a local minimum.
GA [P. Pongcharoen et.al, 2002]	▪ It features a high number of parallel processors. ▪ It is capable of optimizing a wide range of problems including discrete functions. ▪ Continuous functions and multi-objective problems ▪ It delivers responses that improve with time. ▪ There is no requirement for derivative information in a genetic algorithm.	▪ GA necessitates less knowledge on the issue. ▪ However, defining an objective function and ensuring that the representation and operators are correct may be tricky. ▪ GA is computationally costly, which means it takes time.
ALO [Samaher Al-Janabi & Ali Hamza Salman, 2021], [Ali Hamza Salman & Samaher Al-Janabi,2020]	▪ The search region is examined using this technique by selecting at random and walking at random as well. ▪ The ALO algorithm has a high capacity to solve local optimization stagnation due to two factors: the first reason was the use of a roulette wheel, and the second component was the use of haphazard methods. ▪ relocates to a new location, and this site performs better throughout the optimization process, i.e. it retains search area areas ▪ It contains a few settings that you may change.	▪ The reduction in movement intensity is inversely related to the increase in repetitions. ▪ Because of the random mobility, the population has a high degree of variety, which causes issues in the trapping process. ▪ Because the method is not scaled, it is analogous to the black box problem.

GSK [Ali Wagdy Mohamed et. al,2021], [Prachi Agrawal et. al,2021]	▪ To resolve optimization issues ▪ GSK is a randomized, population-based algorithm that iterates the process of acquiring and sharing knowledge throughout a person's life. ▪ Use the GSK method to tackle a series of realistic optimization problems that have been suggested. ▪ In reality, it is simple to apply and a dependable approach for real-world parameter optimization.	▪ The algorithm is incapable of handling and solving multi-objective restricted optimization problems. ▪ The method cannot address issues with enormous dimensions or on a wide scale. ▪ Mixed-integer optimization issues cannot be solved.

2.3 Prediction Techniques

Prediction is find event / value will occur in the future based on the recent facts, the prediction based on law say the predictor give the real values if it is build based on facts otherwise will give virtual values. In general; The prediction techniques split into two types technique based on data mining while the other based on neurocomputing techniques. This Thesis works with the first type of that technique. as explain below:

2.3.1 Decision Tree (DT)

A decision tree is one of the simplest and most often used classification techniques. The Decision Tree method is part of the supervised learning algorithm family. The decision tree approach is also applicable to regression and classification issues. The goal of using a DT is to create a training model that can predict the class or value of a target variable by learning fundamental decision rules from past data (training data). To predict a class label for a record in DT, begin at the root of the tree. The root attribute values are compared to the record's attribute values. Based on the comparison, follow the branch that corresponds to that value and proceed to the next node. The kind of target variable determines the decision tree type [Jehad Ali et. al.,2012].

2.3.2 Extra Trees Classifier (ETC)

Extra Trees Classifier is a decision tree-based ensemble learning approach. Extra Trees Classifier, like Random Forest, randomizes some decisions and data subsets to reduce over-learning and overfitting. Extra Trees Classifier. Trailing trees have a classifier. This class employs a meta-estimator, which employs the mean to increase expected accuracy and control the fit of redundancy by a lot of decision trees have been fitted as a random method as additional trees on the samples, which are classes that are various subclasses of the data class, such that the total number of trees in a forest equals the number of trees in the forest [L Abhishek,2020].

2.3.3 Random Forest (RF)

Leo Breiman invented the random forest aggregation technique in 2001. According to Breman, "the generalization error of a forest of tree classifiers is dependent on the strength and interdependence of the individual trees in the forest." Random Forest, on the other hand, grows individual trees using an efficient method known as bootstrap clustering, as well as a random subspace method, to produce a very powerful clustered predictor capable of classification and regression, with better generalization error than a single decision tree. Bagging is employed alongside random feature selection. Each new training set is derived from the old training set, with replacements. Then, using random feature selection, a tree is formed on the new training set. The trees that have been planted have not been pruned. Bagging is employed for two purposes. The first is that using bagging appears to improve accuracy when using random characteristics. The second is that bagging may be used to provide ongoing estimates of the combined ensemble of trees' generalization error, as well as estimates of strength and correlation. So, Random Forest is an extremely adaptable and simple-to-use machine learning algorithm that produces great results most of the time without requiring unnecessary parameter adjustments. Despite its ease of use and

adaptability, it is one of the most extensively utilized algorithms (it can be used for both classification and regression tasks). It will construct a 'forest' from a set of decision trees that have been trained using the 'packing' method. The mobilization concept is founded on the assumption that merging several learning models in such a way that the total result improves. It is a classifier for groups. Contains numerous decision tree classifiers that are trained in parallel using booting and packing. Because it learns from only one pathway of decisions, a single decision tree frequently overfits the data from which it is learning. Predictions based on a single decision tree are seldom reliable when applied to new data. Random forest models lessen the danger of overfitting by incorporating randomization through the construction of several trees (n estimators). observing with a replacement (i.e., a bootstrapped sample) dividing nodes based on the optimal split among a random sample of the attributes chosen at each node [LEO BREIMAN, 2001].

2.3.4 Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting framework-based decision-tree-based ensemble Machine Learning approach. Artificial neural networks outperform all existing algorithms or frameworks in prediction problems involving unstructured data (images, text, etc.). Decision tree-based algorithms are the best [Tianqi Chen and Carlos Guestrin, 2016].

The algorithm distinguishes itself in the following ways:

- A diverse set of applications: It can solve regression, classification, and ranking issues.
- Portability: Compatible with Pc, Mac, and Linux.
- Languages: C++, Python, R, Java, and Julia are among the major programming languages supported.

2.3.4.1 What is the secret behind XGBoost's Success? [Tianqi Chen and Carlos Guestrin, 2016]

The XGBoost algorithm and the algorithm Gradient Boosting Machines (GBMs) algorithm are tree approaches that employ gradient ratio structures to assist weak learners. XGBoost, on the other hand, enhances the basic GBM algorithm framework with system and computer optimizations.

What are the advantages of the XGBoost algorithm over the regular GBM method? [Tianqi Chen and Carlos Guestrin, 2016]

- ***System Optimization:*** Parallelization: XGBoost employs a parallelized implementation to address the sequential tree approach method. This is possible because the loops used to create base learners are interchangeable; the outer loop enumerates a tree's leaf nodes, while the second inner loop computes the features. This layering of loops inhibits parallelization since the outer loop cannot be begun until the inner loop (the more computationally intensive of the two) is completed. To save time, the order of the loops is modified using initialization through a global scan of all instances and sorting using parallel threads. This decision increases algorithmic performance by offsetting any parallelization overheads in computation.
- ***Tree pruning***: Within the GBM framework, the stopping criteria for tree splitting is greedy and is based on the negative loss criterion at the split site. By utilizing the'max depth' option instead of the criterion, XGBoost begins pruning trees backward. This 'depth-first' approach greatly improves computation speed.
- ***Optimization of Hardware:*** This algorithm was developed to make the best use of hardware resources. This is accomplished by cache awareness, which entails assigning internal buffers in each thread to store gradient statistics. Out-of-core computing maximizes

available disk space when processing big data-frames that do not fit in memory.

- ***Improvements to Algorithms: [Tianqi Chen and Carlos Guestrin, 2016]***

A. ***Regulation:*** Uses both LASSO and Ridge regulation to penalize models that are increasingly complicated to minimize overfitting.

B. ***Scatter awareness:*** XGBoost takes sparse input features naturally by automatically learning the missing ideal value based on the training loss and manages various types of dispersion patterns in data more effectively.

C. ***Weighted Quantity Diagram***: To determine the best split points in weighted data sets, XGBoost employs a quantitative distributed graph.

D. ***Cross-validation:*** The method contains a cross-validation procedure for each iteration, which determines the exact number of rounds of reinforcement required in a single run.

Show algorithm 2.2 of XGBoost:

Algorithm#2.2: XGBoost [John Chambers Award 2016]

Input: training set $= \{(X_i, y_i)\}_i^N$

Output: $°F(x) = °F_M(x) \sum_{m=0}^{M} °F_m(x)$

Initialization: $°F_M(x) = arg_\theta \min \sum_{i=0}^{N} L(y_i, \theta)$ **// a model with a fixed value**

// Processing Stage

1: ***For*** m=1 to M

2: $$g_M(X_i) = \left[\frac{\partial\, L\,(y_i\,, F(X_i))}{\partial\, F(X_i)}\right]_{F(X)=\,°F_{(m-1)}\,(X)}$$ **// Calculate the 'gradients'**

3: $$h_M(X_i) = \left[\frac{\partial^2\, L\,(y_i\,, F(X_i))}{\partial\, F(X_i)^2}\right]_{F(X)=\,°F_{(m-1)}\,(X)}$$ **// Calculate the 'hessians'**

4: $$\theta_M = \arg_{\partial\, \in\, \theta} \sum_{i=1}^{N} \frac{1}{2}\, h_M(X_i) \left[-\frac{g_M(X_i)}{h_M(X_i)} - \theta\,(X_i)\right]^2 .°F_M = \partial\, \emptyset_m\,(X)$$

5:	$^{\circ}F_M\ (x) = {}^{\circ}F_{(m-1)}\ (x) + {}^{\circ}F_M\ (x)$	**// Model should update**
6:	***End for***	
End XGBoost		

The steps of algorithm are:

- loss function: $L\ (y, F(x))$
- Minimizing it in expectation: $^{\circ}F_M(x) = \arg_\theta \min \sum_{i=0}^{N} L(y_i, \theta)$
- Compute residuals: $g_M(X_i) = [\frac{\partial\ L\ (y_i, F(X_i))}{\partial\ F(X_i)}]_{F(X)= {}^{\circ}F_{(m-1)}\ (X)}$
- Compute residuals square: $h_M(X_i) = \left[\frac{\partial^2\ L\ (y_i, F(X_i))}{\partial\ F(X_i)^2}\right]_{F(X)= {}^{\circ}F_{(m-1)}\ (X)}$
- $\theta_M = arg_{\partial \in \theta}\ \sum_{i=1}^{N} \frac{1}{2}\ h_M(X_i) \left[-\ \frac{g_M(X_i)}{h_M(X_i)} -\ \theta\ (X_i)\right]^2 . {}^{\circ}F_M =\ \partial\ \emptyset_m\ (X)$
- training set: $\{(X_i, y_i)\}_i^N$
- The number of weak learners: M, learning rate: θ
- resolving the optimization problem:

  $\theta_M arg_{\partial \in \theta}\ \sum_{i=1}^{N} \frac{1}{2}\ h_M(X_i) \left[-\ \frac{g_M(X_i)}{h_M(X_i)} -\ \theta\ (X_i)\right]^2 . {}^{\circ}F_M =\ \partial\ \emptyset_m\ (X)$
- Update the model: $^{\circ}F_M\ (x) = {}^{\circ}F_{(m-1)}\ (x) + {}^{\circ}F_M\ (x)$

Table 2.2: Analytic the Advantages and Disadvantages for Prediction Techniques

PT	Advantage	Disadvantage
DT [[Samaher Al-Janabi,2021]], [Samaher Hussein Ali,2013]	▪ Decision trees take less work for data preparation during pre-processing as compared to other methods. ▪ Data normalization is not necessary for a decision tree. ▪ Data scaling is not required for a decision tree. ▪ Data missing values have no discernible impact on the decision tree generation process. ▪ The decision tree technique is highly natural and simple to interact with technical teams as well as stakeholders.	▪ A tiny change in the data causes a significant change in the structure of the decision tree, resulting in instability. ▪ When compare this approach to other algorithms, may see that the decision tree calculation become more complicated at times. ▪ A decision tree is rehearsal time is frequently lengthy. ▪ Because of the additional complexity and time required, decision tree training is more expensive. ▪ For forecasting continuous values and performing regression, the Decision Tree approach is unsuccessful.
ETC [L Abhishek,2020]	▪ A sort of collective learning in which the outcomes of numerous non-correlated decision trees gathered in the forest are combined. ▪ Increased predicting accuracy by using a meta-estimator. ▪ DT should be generated using the original training sample. ▪ Similar to the RF classifier, both ensemble learning models are used. ▪ The manner trees are built differs from that of RF. ▪ It chooses the optimum feature to partition the data based on the math Gini index criterion.	▪ bad performance when Overfitting is a difficult problem to tackle ▪ A huge number of uncorrelated DTs are generated by the random sample.

RF[LEO BREIMAN , 2001]	▪ Both regression and classification are possible using RF. ▪ The random forest generates accurate and understandable forecasts. ▪ It can also successfully handle massive data categories. ▪ In terms of accuracy in forecasting results, the random forest algorithm surpassed the decision tree method. ▪ Noise has a less influence on Random Forest. ▪ Missing values may be dealt with automatically using Random Forest. ▪ Outliers are frequently tolerated by Random Forest and handled automatically.	▪ Model interpretability: Random forest models are not easily understood because of the size of the trees, it can consume a large amount of memory. ▪ Complexity: Unlike decision trees, Random Forest generates a large number of trees and aggregates their results. ▪ Longer Training Period: Because Random Forest creates a large number of trees, it takes significantly longer to train than choice trees.
XGBoost [Tianqi Chen and Carlos Guestrin, 2016]	▪ The main benefit of XGB over gradient boosting machines is it has many hyperparameters that can be tweaked. ▪ XGBoost has a feature for dealing with missing values. ▪ It has several user-friendly features,including parallelization,distributed computing, cache optimization, and more. ▪ The XGBoost outperforms the baseline systems in terms of performance. ▪ It can benefit from out-of-core computation and scale seamlessly.	▪ XGBoost performs poorly on sparse and unstructured data. ▪ Gradient Boosting is extremely sensitive to outliers since each classifier is compelled to correct the faults of the previous learners. Overall, the approach is not scalable.

2.4 Evaluation Measures (EMs)

The evaluation assesses how successfully program activities meet expected objectives and how much variation in outcomes may be attributable to the program. Evaluation Measure crucial because it enables program implementers to make objectively based decisions about program operations and service delivery [Ahmed Patel et. al,2015], [David M. W. Powers , 2020].

Table 2.3: Measures of confusion matrix

	Prediction Value	
Actual Value	a (True Positive)	b (False Negative)
	c (False Positive)	d (True Negative)

2.4.1 Accuracy(A)

Accuracy: is the percentage of correct predictions in a classification method or model. It is the ratio of all "true" to all observations [Ahmed Patel et. al,2015].

$$AC = \frac{a+d}{a+b+c+d} \quad \textbf{(2.5)}$$

2.4.2 Precision (P)

It is the percentage of true positive outcomes to the total number of positive predictions made by the classifier [Ahmed Patel et. al,2015].

$$P = \frac{a}{a+c} \quad \textbf{(2.6)}$$

2.4.3 Recall (R)

Recall refers to how many real positives are accurately predicted; it is the ratio of the number of positives to the total number of components in the positive class [Ahmed Patel et. al,2015].

$$TP = \frac{a}{a+b} \quad \textbf{(2.7)}$$

2.4.4 F-Measurement (F)

This measure is based on both measures: precision and recall. That measure compute as eq(2.8) [Ahmed Patel et. al,2015].

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad \textbf{(2.8)}$$

2.4.5 F_β

is the ratio of beta-factor multiplied by Precision and Recall divided by beta-squared multiplied by Precision plus Recall [Ahmed Patel et. al,2015].

$$F_\beta = \frac{(1+\beta^2)*(\text{Precision} * \text{Recall})}{\beta^2 * \text{Precision} + \text{Recall}} \quad \textbf{(2.9)}$$

Chapter Three:

Hybrid Prediction Model for Six types of Natural Gas (HPM-STG)

Chapter Three: Hybrid Prediction Model for Six types of Natural Gas (HPM-STG)

3.1 Introduction

This chapter presents the main stages of building the new predictor and shows the specific details for each stage.

The hybrid Prediction Model for Six types of Natural Gas (HPM-STG) consist of four phases; ***The first stage*** collects real-time data from various natural gas sources. ***The second stage***, pre-processing is divided into multi steps including (a) Checking missing values. (b) Computing correlation among features and target. ***The third stage***; building a predictive algorithm (DGSK-XGB). ***The fourth stage*** uses five evaluation measures in order to evaluate the results of the algorithm DGSK-XGB. The ***HPM-STG*** block diagram is shown in Figure (3.1), and the steps of the model are shown in the algorithm (3.1).

Can summarize the main stages of this research below:

- Capture data from scientific location on internet where, these data collected from different sensors related to the natural gas.
- Through the pre-processing stage, check missing values and compute the correlation.
- Build a new predictor called (***HPM-STG***) by combining the benefits of GSK and XGBoost.
- Multi measures use to evaluate the predictor results include (accuracy, Precision, Recall, f-measurement, and Fb).

In addition, this chapter provides appropriate answers to the questions presented in section one of chapter one (1.2.1).

Algorithm#3.1: Hybrid Prediction Model for Six Types of Gas (HPM-STG)

Input: Stream of real-time data captured from 16 sensors, each sensor, gives 8 features; the total number of features are128 collected from 16 sensors

Output: *Predict the six types of Gas (Ethanol, Ethylene, Ammonia, Acetaldehyde, Acetone, and Toluene)*

*// **Pre-processing Stage***

1: ***For*** *every row in the dataset*

2: ***For*** *every column in the dataset*

3: ***Call*** *Examine for Missing Values*

4: ***Call*** *Correlation*

5: ***End for***

6: ***End for***

*// **Build DGSK –XGB Predictor***

7: ***For*** *i in range (1: total sample count in the dataset)*

8: *Split the dataset into training and testing datasets based on Five- Cross-Validation.*

9: ***End for***

10: ***For*** *each Training part*

11: ***Call*** *DGSK-XGB* ***//used Ackley Function as Function to test fitness function with GSK as kernel of XGBoost***

12: ***End for***

13: ***For*** *each Testing part*

14: *Test stopping conditions*

15: ***IF*** *maximum error generation < Emax*

16: ***Go to*** *step* ***21***

17: ***Else***

18: ***GO to*** *step* ***10***

19: ***End IF***

20: ***End for***

*// **Evaluation stage***

21: ***Call*** *Evaluation*

End HPM-STG

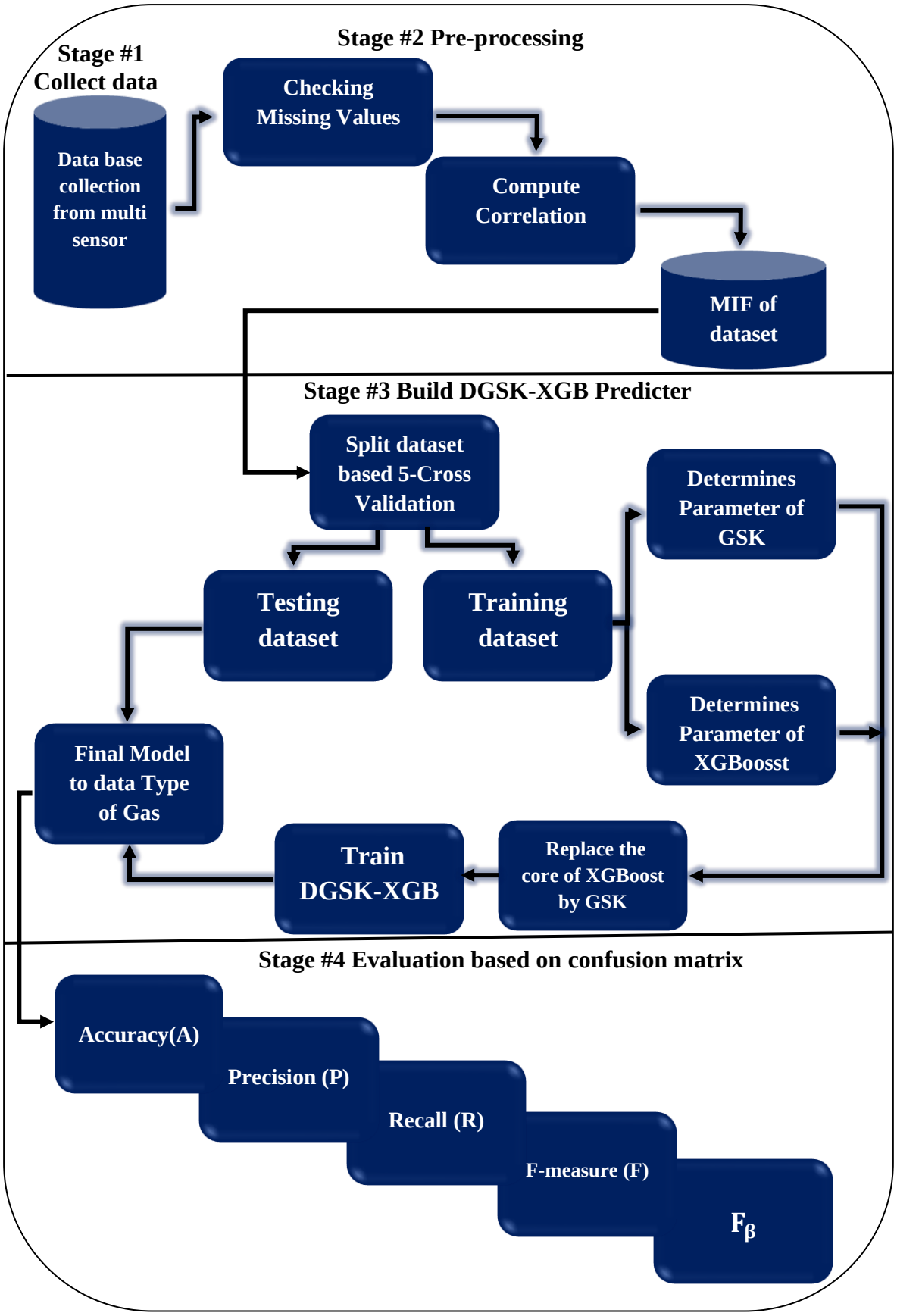

Fig (3.1): Block diagram of DGSK-XGB Model

3.2 The HPM-STG Stages

The initial stage of developing an efficient prediction model is the preparation of the dataset, which includes missing value drop, find correlation among Features and target. The second stage is developing the GSK algorithm through used Ackley function in addition to Junior and senior to grouping dataset into Multi intervals (groups) based on replace the kernel of XGBoost through into the final stage is test the performance of DGSK-XGB that find the type of gas for each group of data generated from the Second stage by addition one of six Labels of gas; The test achieves based Five measures related of the confusion matrix.

3.2.1 Preprocess stage

The data was gathered over a period of several months, that dataset handle as explained in Algorithm 3.2.

- The datasets are merged and stored into a single file.
- Check Missing values through dropping each column in the dataset. because, we attempting to build the Model predict the Label (i.e., Name) of each gas. In general; to satisfy the Law of prediction that say “the results of predictor are become accept and real if the predictor builds from facts otherwise the results become virtual and can’t dependent if it is build from virtual data. Therefore; dropping missing values rather than handle it.
- Find the correlation among all features coming from 16 different sensors with the target to determine which Features are most important and effect on target.

Algorithm#3.2: Preprocessing

Input: A stream of real-time data collected from multi sensors

Output: Most important Features affect in determine the type of gas.

// Checking Missing Value

1: ***For*** each r_i in the dataset ***// i=1… n, n Maximum number of Row***

2: ***For*** each c_j in the dataset ***// j=1… m, m Maximum number of Column***

3: ***IF*** U[i][j] =null Then ***//Check missing value***

4: Delete U[i,j]

5: ***Else***

6: V[i,j]=U[i,j]

7: ***End If***

8: ***End For***

9: ***End For***

// Compute Correlation

10: ***For*** each r_i in dataset

11: ***For*** each c_j in dataset

12: Compute Pearson Correlation $//C_{r_i,c_j} = \frac{\sum(r_i - \bar{r})(c_j - \bar{c})}{\sqrt{\sum(r_i - \bar{r})^2 \sum(c_j - \bar{c})^2}}$

13: ***End For***

14: ***End For***

End Preprocessing

3.2.2 Building Predictor Stage

This stage specific to determine the name of each gas through build the predictor combination between the advantages of GSK and XGBoost; that develop shown with all details in Algorithm 3.3 under title DXGBoost-GSK.

Algorithm #3.3: DXGBoost-GSK

Input: Preprocessed Dataset

Output: Gas Type prediction

Initialize Parameter: N,k_f , k_r, k and p ***//N= number of individuals in population, kf=Junior Phase kr= Knowledge ratio, and p= Senior Phase, K=number of iterations***

// Compute fitness function

1: Set Main Parameters: a=20; b=0.2; c=2π

2: ***For*** each row in dataset ***//i= number of rows***

3: ***For*** each column in dataset ***// j=d= Total number of features***

4: Fitness[i,j]=-a*$\left(-b\sqrt{\frac{1}{d}\sum_{l=1}^{d}v[i,j]^2}\right)-\exp\left(\frac{1}{d}\sum_{l=1}^{d}cos(c*v[i,j])\right)+a+\exp(1)$

5: ***End For***

6: ***End For***

7: Create a new sorted fitness function result

8: G=0

// Calculation of the Gained and Shared Dimension Phases

9: ***For*** G=1 to GEN^{max}

10: ***For*** every row in the dataset

11: ***For*** every column in the dataset

12: $D_{juniorphase} = problemsize * \left(1-\frac{G}{GEN}\right)^{k}$ ***//Junior Phase Equation***

13: $D_{seniorphase} = problemsize - D_{juniorphase}$ ***//Senior Phase Equation***

14: ***IF*** $F\left(v_{i,j}^{new}\right) \le F\left(v_{i,j}^{old}\right)$ ***// Every vector is updated***

15: $v_{i,j}^{old} = v_{i,j}^{new}$

16: $F\left(v_{i,j}^{old}\right) = F\left(v_{i,j}^{new}\right)$

17: ***End IF***

18: ***IF*** $F\left(v_{i,j}^{new}\right) \le F\left(v_{best}^{G}\right)$ ***// global best is updated***

19: $x_{best}^{G} = v_{i,j}^{new}$

20: $F(x_{best}^{G}) = F(v_{i,j}^{new})$

21: ***End IF***

22: ***End for***

23: ***End for***

24: ***For*** best individual in population $F(x_{best}^{G})$

25: $g_{best}(x_i)$= compute the derivative of question $F(x_{best}^{G})$

26: $h_{\text{best}}\ (x_i)$ = compute the second drivativt of question $F(x_{best}^{G})$

27: Prediction types of Gas:

28: $$Fit\ types\ of\ Gas = arg_{\partial \epsilon \theta} \sum_{i=1}^{N} \frac{1}{2}\ h_{best}\ (x_i) \left[- \frac{g_{best}\ (x_i)}{h_{best}\ (x_i)} - \theta\ (x_i) \right]^2 . ^{\circ}\mathrm{F}_M$$
$$= \partial\ \emptyset_{best}\ (x)$$

29: ***End For***

30: ***End For***

31: Return Gas Type

End DXGBoost-GSK

In this section, each variable is explained what it means:

- $v_{i,j}$: where i is the row and j is the column, i.e. the value of Index
- F ($v_{i,j}$): Fitness value for index
- G: is generation number
- Gen: is maximum number of generations.
- a, b, c and d are the main parameters of fitness function (ackley function).
- $F(x_{i,j}^{new})$: The fitness value of the new variable(individual).
- $F(x_{i,j}^{old})$: The fitness value of the old variable(individual).
- $F(x_{best}^{G})$: The fitness value of the global best.
- g_{best}: is the first derivation of the residuals for value of $F(x_{best}^{G})$
- h_{best}: is the second derivation of the residuals for value of $F(x_{best}^{G})$
- ∂: learning rate

3.2.3 Evaluation Stage

In this part, discuss how the predictor is evaluated using five measures; Accuracy, Precision, Recall, F-measure, and Fb how can compute of each that measure.

3.3 Summary

This section attempting to find answer into questions that described into section 1.2.1 in chapter one.

- ***How Gaining-Sharing Knowledge-based (GSK) can be useful in building a new predictor called DGSK-XGB?***

  GSK have a positive effect because it worked to determine the number of points belonging to each group based on an effective Fitness function that included both Junior and Senior, and one of its benefit reducing the execution time of XGBoost, In general; XGBoost required specify multi parameters such as: select root of DT, Max depth of DT, number of terminal nodes and this lead to high complexity.

- ***How can build a prediction model by replacing the kernel of XGBoost with Gaining-Sharing Knowledge-based (GSK)?***

  The core of XGBoost is DT that have many disadvantages as explain in point number one. therefore, in this work using a GSK as core of XGBoost rather than DT to reduce the Time of implementation and increase the accuracy but at the same time this increase the computations.

- ***Is used enough measurement to assess the results of the proposed predictor?***

  Yes, the confusion matrix has five different measures to compute the Performance of DXGBoost-GSK Model, and those measures are sufficient to determine the degree confidence of the predictor.

Chapter Four:

Implementation and Results of (HPM-STG)

Chapter Four: Implementation and Results of (HPM-STG)

4.1 Introduction

This chapter explain the results of each stage present in the design DXGBoost-GSK Model that is shown in chapter 3. Also, it gives Scientific Justification of the results for each stage.

4.2 Result of Implementation

This section of chapter explains the main results; In addition, described the details of database used to implementation the DXGBoost-GSk model.

4.2.1 Description of Dataset

The database* has 16 sensors; each sensor gives 8 features therefore, the total number of features equal to 128. The data is affiliated to 36 months divided into 10 divisions. Each division is called a batch and the data belongs to 6 types of gases, The first gas is called ammonia, the second gas is acetaldehyde, the third gas is acetone, the fourth gas is ethylene, the fifth gas is ethanol, and the sixth gas is toluene.

4.2.2 Result of Preprocessing

This stage begins form get the database from scientific internet site, where these database aggregation from multi sensors through different periods of time include 36 months. Split into ten groups. Then merging all datasets all datasets (groups) into single file to work on it as show in Figure 4.1.

*https://archive.ics.uci.edu/ml/datasets/gas+sensor+array+drift+dataset#

Table (4.1): Sample of dataset after Merging

	F01	F02	F03	F04	...	FF5	FF6	FF7	FF8	Target
0	15596	1.86824	2.37160	2.80367	...	1.4457	-0.54	-0.902	-2.654	1
1	26402	2.53240	5.41120	6.50990	...	1.9511	-0.889	-1.323	-1.749	1
2	42103	3.45418	8.19817	10.5084	...	3.0398	-1.334	-1.993	-2.348	1
3	42825	3.45119	12.1139	16.2668	...	4.0491	-1.432	-2.146	-2.488	1
4	58151	4.19483	11.4551	15.7153	...	4.4024	-1.930	-2.931	-4.088	1
5	79872	5.55359	18.6904	27.1917	...	4.4024	-1.930	-2.931	-4.088	1
6	94981	5.87293	22.3181	32.6731	...	6.5300	-2.688	-4.173	-4.808	1
7	10807	6.06471	25.6602	38.0592	...	7.9498	-3.407	-5.363	-5.870	1
8	12131	7.11836	28.6860	42.9793	...	9.4632	-4.029	-6.616	-7.203	1
9	13360	7.80608	30.7855	46.8975	...	10.349	-4.741	-7.752	-8.587	1
10	14338	8.10868	32.7197	52.8041	...	10.683	-5.452	-9.167	-9.748	1
11	15321	8.55041	35.0694	59.0104	...	11.803	-6.370	-10.77	-11.61	1
12	16300	8.87675	37.5601	64.5917	...	12.391	-7.119	-12.30	-13.49	1
13	17102	9.24244	38.8665	68.2393	...	13.085	-7.884	-13.90	-15.01	1
14	18492	2.03458	2.79174	3.27864	...	13.163	-8.682	-15.21	-16.19	1
15	32765	2.83066	6.41720	7.91038	...	2.5976	-1.144	-1.691	-2.102	1
16	43883	3.50994	9.54941	12.4542	...	3.9269	-1.576	-2.365	-2.790	1
17	53080	3.94959	12.2461	16.8724	...	5.1847	-2.003	-3.045	-3.539	1
18	61543	4.55606	14.7222	20.5638	...	6.1244	-2.353	-3.757	-4.082	1
19	80026	5.49087	19.3839	28.2716	...	8.3627	-3.193	-5.071	-5.566	1
20	96025	6.44952	23.0622	33.8838	...	9.7576	-4.047	-6.463	-7.110	1
21	10911	7.17743	26.5160	39.5908	...	11.489	-4.840	-7.789	-8.418	1
22	12123	7.78883	28.8286	42.0789	...	12.404	-5.522	-8.967	-9.596	1
23	13269	8.38129	31.5433	47.0881	...	13.359	-6.402	-10.74	-11.40	1
24	14247	8.59289	33.9558	51.8694	...	14.029	-7.146	-12.21	-12.93	1
...	...	...	...	...	...	...	...	...	...	...
13897	14618	12.4679	35.2246	48.3145	...	21.499	-9.868	-16.65	-36.36	5
13898	11108	1.6175	2.6187	3.8728	...	5.7291	-1.158	-1.841	-4.939	4
13899	97450	8.8981	21.2568	27.3432	...	9.41	-4.604	-7.461	-16.96	5
13900	22330	2.2167	5.1503	6.8132	...	7.1476	-2.089	-3.383	-7.367	4
13901	21720	18.104	55.606	78.6643	...	44.535	-22.02	-40.31	-82.35	5
13902	35192	2.8785	8.3185	10.8794	...	10.742	-3.696	-6.098	-12.57	4
13903	15017	12.6447	36.3565	49.6011	...	20.496	-9.848	-17.59	-32.26	5
13904	1024	1.056	1.1523	2.1886	...	5.8885	-0.284	-0.625	-3.923	4
13905	28929	3.3092	4.3933	5.057	..	5.3419	-0.842	-1.407	-4.447	5
13906	40392	3.28	10.2689	13.5859	...	11.591	-4.495	-7.655	-12.76	4
13907	18327	15.5945	45.4803	63.8775	...	33.366	-15.82	-28.51	-56.33	5
13908	29059	2.568	6.7142	8.7865	...	9.0262	-2.849	-4.734	-9.573	4
13909	23076	19.1666	60.9364	87.8557	...	60.609	-28.62	-54.25	-125.5	5

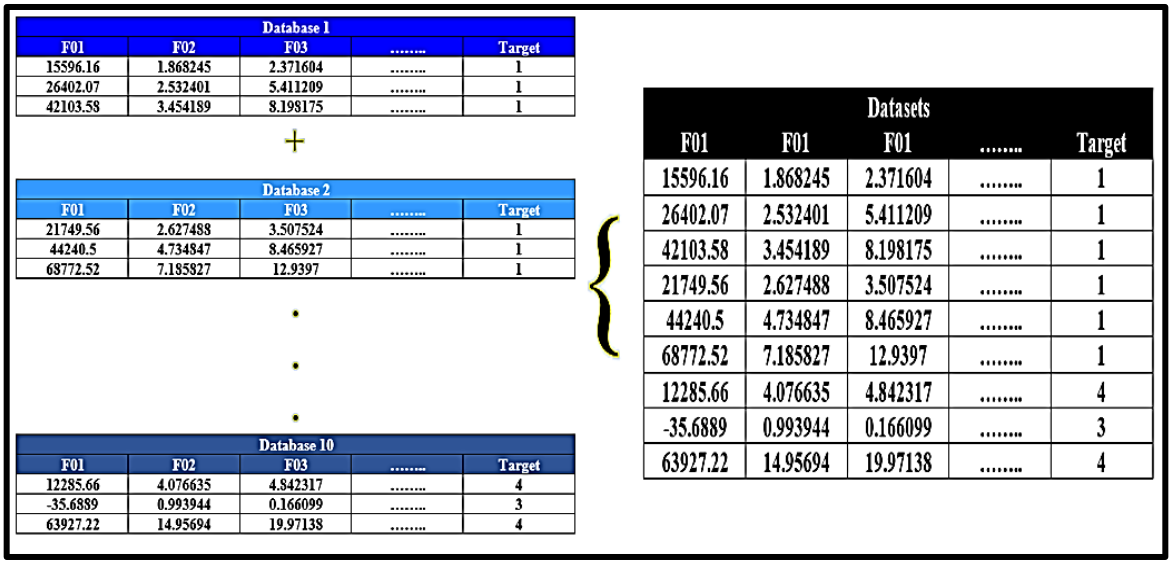

Fig (4.1): Data Merging Method

4.2.2.1 Checking Missing Value

After Merging all datasets in a single file; checking if that file has missing values or not; if found drop the record from that dataset to satisfy the Law of prediction otherwise continuous. In general, in this step not dropping any record.

4.2.2.2 Correlation

The correlation is computed among all the features with the target to determine the main features effect in specific type of gas. In general, there are three types of relationship among features and target; when the correlation forward in side (+1) this meaning the Positive relationship while If correlation value goes side (-1) this meaning the negative relationship between feature and target; otherwise, if correlation value is gone side (0) this meaning not found any relationship between feature and target.

The effects and relationships among features. When the value of the adopted threshold is greater than or equal to 0.80.and each type of gas is shown in tables (4.2), (4.3), (4.4), (4.5), (4.6) and (4.7).

Table (4.2): Correlation's Influence on the First Type of Gas

	F01	F02	F03	F04	...	FF6	FF7	FF8	Target
F01	1	0.16993	0.98272	0.941909	...	-0.62532	-0.48762	-0.0612	0.2618
F02	0.16993	1	0.18640	0.176823	...	-0.33036	-0.27067	-0.06084	-0.802
F03	0.98272	0.18640	1	0.978705	...	-0.66009	-0.52708	-0.09772	-0.658
F04	0.94190	0.17682	0.97870	1	...	-0.66517	-0.56594	-0.17047	-0.849
F05	0.58693	0.07599	0.63611	0.741052	...	-0.41503	-0.36067	-0.1207	-0.869
F06	-0.96451	-0.1771	-0.97081	-0.94842	...	0.693337	0.600748	0.187304	0.8458
F07	-0.7876	-0.1286	-0.81448	-0.83372	...	0.633635	0.714449	0.465428	-0.534
F08	-0.20933	0.01404	-0.2389	-0.30342	...	0.262469	0.609764	0.781228	-0.523
F11	0.87905	0.25510	0.84328	0.797083	...	-0.79819	-0.63104	-0.10267	-0.238
F12	0.41424	0.81531	0.44137	0.425014	...	-0.66086	-0.53091	-0.12287	0.640
F13	0.86349	0.28397	0.85672	0.819756	...	-0.8466	-0.68253	-0.14241	0.659
F14	0.84823	0.27807	0.85471	0.834129	...	-0.86612	-0.7415	-0.23365	0.720
F15	0.75125	0.20818	0.75637	0.74282	...	-0.78318	-0.68811	-0.24835	0.759
F16	-0.82618	-0.25593	-0.80844	-0.77785	...	0.88125	0.76461	0.24299	0.671
F17	-0.63253	-0.19809	-0.63048	-0.63177	...	0.8003	0.85554	0.47649	-0.730
F18	-0.10768	-0.01354	-0.12525	-0.18352	...	0.33832	0.68044	0.76985	-0.693
...	...	...	...	...	...	...	...	...	...
FE1	0.64936	0.33635	0.68801	0.69035	...	-0.98307	-0.83986	-0.30047	0.150
FE2	0.79249	0.32424	0.77158	0.7439	...	-0.68948	-0.57494	-0.15417	0.025
FE3	0.45536	0.34898	0.52390	0.55723	...	-0.93119	-0.84892	-0.40348	-0.151
FE4	0.27947	0.38191	0.36232	0.41408	...	-0.81154	-0.78311	-0.45211	-0.201
FE5	0.16257	0.42894	0.24835	0.33815	...	-0.6714	-0.65765	-0.38999	-0.356
FE6	-0.62797	-0.33029	-0.66197	-0.66806	...	0.99933	0.88534	0.34160	-0.282
FE7	-0.51947	-0.27925	-0.55872	-0.593	...	0.91286	0.99347	0.59539	-0.458
FE8	-0.0668	-0.06307	-0.10505	-0.18029	...	0.35254	0.67574	0.97721	-0.436
FF1	0.64777	0.33674	0.68793	0.68941	...	-0.98226	-0.8365	-0.29467	0.095
FF2	0.78725	0.32472	0.76469	0.73612	...	-0.68785	-0.57219	-0.14997	0.101
FF3	0.46017	0.37485	0.52932	0.56023	...	-0.93436	-0.84769	-0.39478	0.019
FF4	0.25372	0.69722	0.32951	0.37138	…	-0.73625	-0.70707	-0.39401	-0.063
FF5	0.1998	0.3430	0.2859	0.3621	...	-0.73876	-0.72162	-0.42286	-0.063
FF6	-0.6253	-0.3303	-0.6600	-0.6651	...	1	0.88429	0.33802	-0.203
FF7	-0.4876	-0.2706	-0.5270	-0.5659	...	0.88429	1	0.65771	-0.286
FF8	-0.0612	-0.0608	-0.0977	-0.1704	...	0.33802	0.657716	1	-0.462
Target	0.025521	-0.1191	-0.21221	-0.31558		-0.78115	-0.57556	-0.15445	1

The sensors more affect to determine the first gas is (FD1) in the first order and in the second order (F23, FC1) while the not important sensors are (F05, F24, F25, F32) therefore to reduce the computation can be neglected.

Table (4.3): Correlation's Influence on the Second Type of Gas

	F01	F02	F03	F04	...	FF6	FF7	FF8	Target
F01	1	0.344399	0.981737	0.956219	...	-0.68355	-0.4916	-0.1495	-0.056
F02	0.344399	1	0.40444	0.441592	...	-0.44338	-0.3338	-0.1007	0.050
F03	0.981737	0.40444	1	0.989887	...	-0.75504	-0.5657	-0.1949	0.024
F04	0.956219	0.441592	0.989887	1	...	-0.76019	-0.5999	-0.2509	0.002
F05	0.950936		0.967847	0.978915	...	-0.69469	-0.5447	-0.2361	0.585
F06	-0.9819	-0.38457	-0.98674	-0.97507	...	0.735879	0.57952	0.23440	0.982
F07	-0.88024	-0.33168	-0.89355	-0.90462	...	0.664752	0.64056	0.39386	0.939
F08	-0.71029	-0.16748	-0.71762	-0.74134	...	0.541595	0.68050	0.60939	0.960
F11	0.988728	0.331989	0.968172	0.942535	...	-0.67368	-0.4810	-0.1440	0.727
F12	0.303383	0.983259	0.367343	0.405587	...	-0.4333	-0.3245	-0.0941	0.450
F13	0.972307	0.403271	0.987593	0.976059	...	-0.74836	-0.5580	-0.1906	0.789
F14	0.946934	0.448491	0.976722	0. 981173	...	-0.75632	-0.5962	-0.2503	0.789
F15	0.944219	0.365693	0.954199	0.957843	...	-0.68418	-0.5343	-0.2324	0.726
F16	-0.98131	-0.36508	-0.98153	-0.96813	...	0.730996	0.56974	0.22284	-0.767
F17	-0.9323	-0.32896	-0.93738	-0.9399	...	0.705204	0.65879	0.35955	0.114
F18	-0.76983	-0.17164	-0.76922	-0.78453	...	0.567751	0.65941	0.53659	-0.103
...	...	...	...	...	...	...	...	...	…
FE1	0.730227	0.459082	0.788938	0.787489	⋯	-0.98525	-0.7746	-0.2777	0.71719
FE2	0.253303	0.532603	0.319691	0.334419	...	-0.63337	-0.4996	-0.1526	0.6144
FE3	0.331484	0.405302	0.433432	0.459471	...	-0.88351	-0.7846	-0.3778	0.9978
FE4	-0.09402	0.218697	-0.00371	0.031311	...	-0.56077	-0.5658	-0.3335	0.5382
FE5	-0.10313	0.184278	-0.01736	0.017159	...	-0.53466	-0.5434	-0.3241	0.5039
FE6	-0.68698	-0.43828	-0.75736	-0.7623	...	0.999266	0.84062	0.34456	-0.9852
FE7	-0.51654	-0.34825	-0.59025	-0.61994	...	0.869719	0.99434	0.65119	-0.7746
FE8	-0.15149	-0.10024	-0.19719	-0.25189	...	0.356078	0.72186	0.97304	-0.2777
FF1	0.730229	0.467987	0.789478	0.787447	...	-0.98636	-0.7730	-0.2726	0.5124
FF2	0.241141	0.534534	0.308845	0.324153	...	-0.62811	-0.4933	-0.1465	0.44395
FF3	0.356735	0.434986	0.461956	0.489598	...	-0.89404	-0.79	-0.3766	-0.8504
FF4	-0.06369	0.257711	0.033004	0.071941	...	-0.58498	-0.5884	-0.3480	-0.6294
FF5	-0.0809	0.220687	0.010947	0.049798	...	-0.54974	-0.5584	-0.3355	-0.1277
FF6	-0.68355	-0.44338	-0.75504	-0.76019	...	1	0.84042	0.34232	0.79764
FF7	-0.49164	-0.33389	-0.56579	-0.59997	...	0.840426	1	0.70148	0.75056
FF8	-0.1495	-0.10074	-0.19492	-0.25098	...	0.342326	0.70148	1	-0.8346
Target	0.73587	0.66475	0.54159	-0.67368	…	-0.01736	-0.7573	-0.5902	1

The sensors more affect to determine the second gas (F63, FF3) in the first order and in the second-order are (F73, FA3, FE3) while the not important sensor is (F58) therefore to reduce the computation can be neglected.

Table (4.4): Correlation's Influence on the Third Type of Gas

	F01	F02	F03	F04	...	FF6	FF7	FF8	Target
F01	1	0.67734	0.970291	0.902814	...	-0.39767	-0.3935	-0.4375	-0.1334
F02	0.67734	1	0.696187	0.662528	...	-0.47068	-0.4881	-0.4876	0.1050
F03	0.970291	0.696187	1	0.968945	...	-0.3775	-0.3773	-0.4160	-0.1149
F04	0.902814	0.662528	0.968945	1	...	-0.33087	-0.3334	-0.3683	-0.1207
F05	0.902912	0.587634	0.948309	0.977208	...	-0.33577	-0.3349	-0.3743	-0.1416
F06	-0.94828	-0.67095	-0.95548	-0. 90926	...	0.390051	0.39001	0.42227	0.8994
F07	-0.74604	-0.47364	-0.76662	-0.74634	...	0.307738	0.30443	0.33380	0.8709
F08	-0.88565	-0.46745	-0.86274	-0.8077	...	0.390246	0.38040	0.42556	0.8994
F11	0.773338	0.498846	0.713754	0.643182	...	-0.50337	-0.494	-0.5449	0.3832
F12	0.581212	0.90586	0.593075	0.561019	...	-0.57381	-0.5904	-0.5904	-0.491
F13	0.777131	0.550782	0.748151	0.690143	...	-0.51946	-0.5143	-0.5604	-0.571
F14	0.766599	0.558541	0.755509	0.708707	...	-0.48396	-0.4816	-0.5266	-0.505
F15	0.737638	0.443695	0.698821	0.642485	...	-0.46138	-0.4554	-0.5055	-0.469
F16	-0.76694	-0.54062	-0.7179	-0.65255	...	0.571925	0.56756	0.60858	-0.448
F17	-0.75637	-0.49892	-0.70064	-0.63256	...	0.55769	0.55118	0.59542	0.5624
F18	-0.68442	-0.31974	-0.60975	-0.53616	...	0.501959	0.4873	0.54117	0.5486
...	...	...	...	...	...	...	...	...	...
FE1	0.439714	0.452766	0.40608	0.349287	...	-0.98158	-0.9737	-0.9745	-0.7997
FE2	0.358079	0.500789	0.349394	0.303468	...	-0.85108	-0.8454	-0.8383	-0.7577
FE3	0.107565	0.281564	0.112196	0.086252	...	-0.87133	-0.8645	-0.8400	0.9983
FE4	-0.12111	0.123572	-0.10085	-0.10609	...	-0.6709	-0.6636	-0.6281	0.31206
FE5	-0.13341	0.105021	-0.11498	-0.12076	...	-0.65109	-0.6429	-0.6081	-0.6145
FE6	-0.39223	-0.44685	-0.36637	-0.3172	...	0.998574	0.9954	0.98653	-0.5874
FE7	-0.3858	-0.46945	-0.365	-0.31898	...	0.997871	0.9980	0.98730	-0.6272
FE8	-0.42065	-0.46531	-0.39625	-0.34739	...	0.987152	0.9863	0.98500	0.68881
FF1	0.459526	0.489692	0.432843	0.378153	...	-0.98127	-0. 974	-0.9753	0.52690
FF2	0.358837	0.527168	0.358293	0.316705	...	-0.84052	-0.8367	-0.8279	0.75735
FF3	0.133712	0.316884	0.140029	0.113857	...	-0.88849	-0.8830	-0.8575	0.99839
FF4	-0.09516	0.163825	-0.07419	-0.08011	...	-0.70252	-0.6969	-0.6597	0.75933
FF5	-0.1106	0.141532	-0.09127	-0.09717	...	-0.68122	-0.6747	-0.6381	0.74993
FF6	-0.39767	-0.47068	-0.3775	-0.33087	...	1	0.9981	0.98768	-0.8510
FF7	-0.39355	-0.48813	-0.37733	-0.33347	...	0.998179	1	0.98771	-0.8454
FF8	-0.4375	-0.4876	-0.41605	-0.36832	...	0.987682	0.9877	1	-0.8383
Target	-0.79975	-0.79975	-0.79975	-0.79975	…	0.16957	0.16512	0.23924	1

The sensors more affect to determine the third gas (FD3, FF3) in the first order and in the second-order is (FE3) while the not important sensors are (F06, F07, F08) therefore to reduce the computation can be neglected.

Table (4.5): Correlation's Influence on the Fourth Type of Gas

	F01	F02	F03	F04	...	FF6	FF7	FF8	Target
F01	1	0.536672	0.992808	0.97929	...	-0.68142	-0.68101	-0.66396	-0.785
F02	0.536672	1	0.497689	0.46759	...	-0.58358	-0.61584	-0.55421	-0.632
F03	0.992808	0.497689	1	0.993335	...	-0.64549	-0.64523	-0.63032	-0.619
F04	0.97929	0.46759	0.993335	1	...	-0.6122	-0.61166	-0.5988	-0.545
F05	0.950241	0.380151	0.967782	0.976469	...	-0.57699	-0.57348	-0.57235	0.6845
F06	-0.98981	-0.59105	-0.98432	-0.97202	...	0.675313	0.679695	0.655034	0.9955
F07	-0.98725	-0.56069	-0.9847	-0.97494	...	0.656568	0.660035	0.638363	0.9955
F08	-0.96893	-0.41594	-0.9732	-0.96766	...	0.605437	0.603002	0.598521	0.9699
F11	0.573491	0.301845	0.538615	0.511061	...	-0.87064	-0.85453	-0.88983	-0.562
F12	0.663011	0.885682	0.620979	0.583391	...	-0.76973	-0.78773	-0.74315	-0.562
F13	0.59799	0.33589	0.570564	0.545962	...	-0.88887	-0.87619	-0.90781	-0.495
F14	0.595743	0.339126	0.572464	0.55196	...	-0.87368	-0.86259	-0.89411	0.5856
F15	0.530194	0.251885	0.506764	0.486713	...	-0.8259	-0.81185	-0.8548	0.5628
F16	-0.61019	-0.40343	-0.57184	-0.54255	...	0.914912	0.90441	0.927824	0.4273
F17	-0.58955	-0.36856	-0.5518	-0.52399	...	0.900439	0.888751	0.916334	-0.493
F18	-0.46578	-0.16308	-0.43774	-0.41647	...	0.80205	0.78409	0.832032	-0.399
...	...	...	...	...	...	...	...	...	...
FE1	0.655177	0.475536	0.618869	0.585087	...	-0.98437	-0.97521	-0.98065	-0.984
FE2	0.642398	0.463001	0.608328	0.570971	...	-0.8893	-0.87984	-0.86578	-0.872
FE3	0.673593	0.655286	0.641977	0.609867	...	-0.95755	-0.95927	-0.93412	0.9986
FE4	0.623827	0.700009	0.597516	0.	...	-0.81798	-0.82456	-0.77941	0.7369
FE5	0.611998	0.693731	0.585827	0.558955	...	-0.81485	-0.82148	-0.77832	-0.883
FE6	-0.66926	-0.56049	-0.63272	-0.59928	...	0.997679	0.994739	0.99023	-0.723
FE7	-0.67189	-0.59035	-0.63574	-0.60202	...	0.997227	0.99717	0.98919	-0.800
FE8	-0.64846	-0.53275	-0.61384	-0.58192	...	0.989318	0.986433	0.98861	-0.627
FF1	0.670021	0.495534	0.634846	0.601339	...	-0.98829	-0.9801	-0.98495	-0.628
FF2	0.666366	0.519954	0.632442	0.595112	...	-0.89974	-0.89318	-0.87411	0.7838
FF3	0.681685	0.677211	0.649056	0.616805	...	-0.95893	-0.96195	-0.93484	0.9986
FF4	0.634726	0.738802	0.6057	0.578229	...	-0.82585	-0.83476	-0.78652	-0.873
FF5	0.626497	0.735939	0.597153	0.569106	...	-0.82439	-0.83328	-0.78621	-0.867
FF6	-0.68142	-0.58358	-0.64549	-0.6122	...	1	0.998303	0.99218	-0.761
FF7	-0.68101	-0.61584	-0.64523	-0.61166	...	0.998303	1	0.99023	0.7594
FF8	-0.66396	-0.55421	-0.63032	-0.5988	...	0.992183	0.990233	1	0.7877
Target	-0.63574	-0.60202	-0.56745	0.668872	...	-0.76763	-0.75217	-0.74159	1

The sensors more affect to determine the fourth gas (FF3) in the first order and in the second-order is (FE3) while the not important sensors are (F06, F07, F08) therefore to reduce the computation can be neglected.

Table (4.6): Correlation's Influence on the Fifth Type of Gas

	F01	F02	F03	F04	...	FF6	FF7	FF8	Target
F01	1	0.55071	0.99106	0.98226	...	-0.79572	-0.81748	-0.82079	0.394
F02	0.55071	1	0.54503	0.51402	...	-0.3934	-0.38795	-0.33961	0.251
F03	0.99106	0.54503	1	0.99652	...	-0.79358	-0.81755	-0.82388	0.419
F04	0.98226	0.51402	0.99652	1	...	-0.79907	-0.82557	-0.84115	0.427
F05	0.86679	0.34231	0.88205	0.90720	...	-0.77599	-0.81127	-0.87058	0.410
F06	-0.9938	-0.54831	-0.98867	-0.98114	...	0.79032	0.81153	0.80918	-0.397
F07	-0.9512	-0.42046	-0.95158	-0.96262	...	0.84606	0.87703	0.91886	-0.435
F08	-0.89136	-0.31953	-0.89172	-0.91025	...	0.83969	0.87459	0.93113	-0.428
F11	0.90937	0.48128	0.89315	0.88770	...	-0.84655	-0.85446	-0.84692	0.421
F12	0.04665	0.09468	0.04462	0.04053	...	-0.04174	-0.03965	-0.03312	0.019
F13	0.91045	0.50929	0.91351	0.91239	...	-0.86116	-0.86905	-0.85688	0.481
F14	0.88057	0.48073	0.89064	0.89472	...	-0.83424	-0.84433	-0.84014	0.474
F15	0.80685	0.31762	0.81679	0.84033	...	-0.8037	-0.83208	-0.8936	0.431
F16	-0.90608	-0.47958	-0.8924	-0.88854	...	0.85605	0.86113	0.84439	0.693
F17	-0.85224	-0.34804	-0.84852	-0.86249	...	0.84158	0.86516	0.90611	0.688
F18	-0.86162	-0.29455	-0.86057	-0.87975	...	0.84799	0.87974	0.93741	-0.878
...	...	...	...	...	...	...	...	...	...
FE1	0.72306	0.35782	0.71591	0.71968	...	-0.98458	-0.96922	-0.90496	-0.834
FE2	0.71029	0.46271	0.70275	0.69063	...	-0.89112	-0.88881	-0.82779	-0.679
FE3	0.60677	0.29972	0.62244	0.63153	...	-0.92248	-0.90242	-0.82538	0.998
FE4	0.39474	0.25193	0.41916	0.42712	...	-0.74441	-0.71248	-0.61584	0.695
FE5	0.43406	0.26538	0.45802	0.46717	...	-0.76032	-0.73333	-0.65001	0.699
FE6	-0.7878	-0.37899	-0.78596	-0.79239	...	0.99934	0.99581	0.95081	-0.886
FE7	-0.80566	-0.37611	-0.80591	-0.8147	...	0.995447	0.998828	0.97006	-0.884
FE8	-0.81078	-0.32375	-0.81528	-0.83452	...	0.938505	0.960283	0.99253	0.6756
FF1	0.72457	0.372426	0.716701	0.719442	...	-0.98359	-0.96695	-0.90009	0.6900
FF2	0.72006	0.48304	0.71191	0.69843	...	-0.8866	-0.88486	-0.82392	-0.885
FF3	0.62583	0.32638	0.64207	0.65013	...	-0.9285	-0.90825	-0.82903	0.9983
FF4	0.43840	0.29197	0.46506	0.47237	...	-0.76622	-0.73606	-0.63905	0.7013
FF5	0.45923	0.30333	0.48582	0.49374	...	-0.76786	-0.74149	-0.65388	0.7148
FF6	-0.79572	-0.3934	-0.79358	-0.79907	...	1	0.99594	0.94888	0.7597
FF7	-0.81748	-0.38795	-0.81755	-0.82557	...	0.99594	1	0.96901	0.9988
FF8	-0.82079	-0.33961	-0.82388	-0.84115	...	0.94888	0.96901	1	-0.741
Target	-0.70074	-0.27851	-0.70734	-0.71531	…	0.67472	0.99594	0.79546	1

The sensors more affect to determine the fifth gas are (F31, F63) in the first order and in the second-order (FE3, FF3, FF7) while the not important sensor is (F12) therefore to reduce the computation can be neglected.

Table (4.7): Correlation's Influence on the Sixth Type of Gas

	F01	F02	F03	F04	...	FF6	FF7	FF8	Target
F01	1	0.827722	0.996896	0.991918	...	-0.61345	-0.60564	-0.56495	-0.441
F02	0.827722	1	0.816031	0.799979	...	-0.58163	-0.57252	-0.53299	0.7816
F03	0.996896	0.816031	1	0.998122	...	-0.6219	-0.61513	-0.57624	0.6534
F04	0.991918	0.799979	0.998122	1	...	-0.61403	-0.60816	-0.57192	0.7074
F05	0.882512	0.619435	0.894015	0.911157	...	-0.46769	-0.46666	-0.45562	0.6721
F06	-0.58006	-0.56228	-0.57462	-0.55763	...	0.901622	0.895573	0.849866	0.7473
F07	-0.52663	-0.49916	-0.52648	-0.51498	...	0.896618	0.901221	0.906473	-0.480
F08	-0.49674	-0.45915	-0.50047	-0.49251	...	0.879275	0.888358	0.911965	-0.478
F11	0.967621	0.861111	0.96288	0.954858	...	-0.64712	-0.63864	-0.594	-0.437
F12	0.684363	0.960297	0.671271	0.654226	...	-0.53729	-0.52821	-0.48728	0.9602
F13	0.969599	0.853288	0.971326	0.966649	...	-0.66255	-0.6554	-0.61394	0.7204
F14	0.969883	0.839232	0.974511	0.972584	...	-0.65795	-0.65183	-0.61425	0.7532
F15	0.872413	0.670858	0.881237	0.891388	...	-0.55745	-0.55841	-0.56783	0.6953
F16	-0.62274	-0.61184	-0.61707	-0.60061	...	0.936488	0.92913	0.882264	0.7352
F17	-0.58498	-0.56172	-0.58629	-0.57689	...	0.934027	0.938994	0.955425	-0.617
F18	-0.56657	-0.53343	-0.57092	-0.5633	...	0.931048	0.938462	0.965093	-0.610
...	...	...	...	...	...	...	...	...	...
FE1	0.91316	0.767854	0.922928	0.923902	...	-0.73924	-0.73372	-0.68685	0.5591
FE2	0.844316	0.7675	0.853105	0.849607	...	-0.77327	-0.76849	-0.72074	0.5446
FE3	0.868439	0.719609	0.884873	0.889861	...	-0.74361	-0.73939	-0.69084	-0.705
FE4	0.816046	0.662245	0.83807	0.846828	...	-0.7312	-0.72841	-0.6807	0.9985
FE5	0.823799	0.655921	0.84509	0.855903	...	-0.70906	-0.70784	-0.67035	-0.59
FE6	-0.61266	-0.58029	-0.62157	-0.61408	...	0.999585	0.998917	0.96369	-0.573
FE7	-0.60543	-0.57329	-0.61538	-0.6088	...	0.997141	0.998449	0.97396	-0.517
FE8	-0.55209	-0.52257	-0.56389	-0.56027	...	0.949692	0.960481	0.99148	0.7221
FF1	0.912609	0.774064	0.921749	0.922107	...	-0.73731	-0.73132	-0.68238	0.7197
FF2	0.844972	0.775903	0.852937	0.848676	...	-0.77273	-0.76759	-0.71845	0.7947
FF3	0.869328	0.721019	0.885512	0.88976	...	-0.74312	-0.73837	-0.68709	0.7721
FF4	0.819507	0.663426	0.841597	0.849651	...	-0.72773	-0.72443	-0.67385	0.9985
FF5	0.82582	0.663984	0.848482	0.858406	...	-0.72101	-0.71893	-0.67774	-0.684
FF6	-0.61345	-0.58163	-0.6219	-0.61403	...	1	0.998802	0.959727	-0.733
FF7	-0.60564	-0.57252	-0.61513	-0.60816	...	0.998802	1	0.969092	-0.720
FF8	-0.56495	-0.53299	-0.57624	-0.57192	...	0.959727	0.969092	1	-0.734
Target	-0.55874	-0.44447	-0.58131	-0.58309	...	-0.59071	-0.5396	-0.51868	1

The sensors more affect to determine the sixth gas are (F21, F63, FE4) in the first order and in the second-order (F73, FB1, FF4) while the not important sensor is (F12) therefore to reduce the computation can be neglected.

4.2.3 Results of DXGBoost-GSK

This section of chapter applies the main steps of predictor after spilt the dataset into training and testing parts through 5-cross validation Then grouping dataset by GSK after that; specific Label for each group through DXGBoost; Final evaluation the results.

The data is divided into training data test data. Through five cross validation, build model based on certain percentage of the data, where this percentage of the data, where this percentage is for training and the rest for testing, and so on for the rest of the sections. Each time the error value is calculated, and which split gives the lowest error rate is depend on build the final model, The best division obtained is 80% for training and 20% for testing. In general; the total number of samples of these datasets are 13910.

XGBoost is a distributed gradient boosting library that is exceptionally efficient, versatile, and portable. It employs Gradient Boosting, and the fundamental parameters of the XGBoost method are listed in table 4.8.

For tackling continuous-space optimization problems The GSK algorithm simulates the process of acquiring and sharing knowledge. It is divided into two phases: (junior gaining and sharing) and (senior gaining and sharing). The current work quantitatively models these two phases in order to optimize the procedure. The main parameters of the DXGBoost-GSK algorithm are determined in table (4.8).

Table (4.8): Parameters of DXGBoost-GSK

Parameter	Value
Learning rate	0.1
n_estimators	100

Gamma	0
Base_score	0.5
Reg_lambda	1
population size(N)	Number of rows
knowledge factor (k_f)	0.5
knowledge ratio(k)	8
Knowledge rate (k_r)	0.9
Fitness function (Ackley)	$f = -a * e^{-\left(b * \sqrt{\frac{1}{d} * s^2}\right)} - \left(\left(\frac{1}{d}\right) * \cos(c * s)\right) + e^1 + 20$
Partition size (P)	0.1
a	20
b	0.2
c	2π
Features and fitness (d)	129
$D_{juniorphase}$	$D_{juniorphase} = problemsize * \left(1 - \frac{G}{GEN}\right)^k$
$D_{seniorphase}$	$D_{seniorphase} = problemsize - D_{juniorphase}$
g_{best}	compute the derivative of question $F(x_{best}^G)$
h_{best}	compute the second drivativt of question $F(x_{best}^G)$

The Ackley Function was not used as a fitness function for GSK, but it originally contains a fitness function, which is both Junior and Senior, but was used to test the fitness function and its working principle is:

- Take a uniform sample of a number of solutions X=[x,y].
- Determine the associated goal function values for each solution and order the solutions from highest to lowest.
- maintain solutions with lower objective values [lowest 20]. Get the x,y coordinates of the points you've chosen.
- Using the x,y coordinates of the chosen sites, estimate the multivariate normal distribution.

- Using the distribution, sample a fresh set of solutions. Return to step 2. Estimate the solutions' objective values. Iterate from step 2 to step 5 several t eventually locating the global highest point.

The table below shows all the results for junior, senior, and Ackley.

Table (4.9): the result of GSK

Iteration	Junior	Senior	Ackley
1	8.133	2.866	22.753
2	5.942	5.057	22.758
3	4.287	6.712	22.761
4	3.050	7.949	22.766
5	2.137	8.862	22.769
6	1.473	9.526	22.780
7	0.997	10.002	22.790
8	0.661	10.338	22.800
9	0.429	10.570	22.806
10	0.271	10.728	22.809
11	0.167	10.832	22.812
12	0.099	10.900	22.820
13	0.057	10.942	22.830
14	0.031	10.968	22.835
15	0.016	10.983	22.844
16	0.008	10.991	22.850
17	0.003	10.996	22.855
18	0.001	10.998	22.860
19	0.002	10.999	22.865
20	0.001	10.999	22.867

The GSK algorithm is applied to the data and depends on three main parameters (Junior, Senior, Ackley) where each parameter depends on a

certain law to be executed and indicates something where Junior means the amount of information to be obtained and Senior is the amount of information to be shared and they are the two principles The work of the GSK algorithm and the last parameter, Ackley, which is its work to test the fitness function, is based on the optimization principle, So it is suitable for the working principle of the GSK algorithm.

While; the results of DXGBoost after replacing their kernel with GSK are explained in table 4.10.

Table (4.10): The result of HPM-STG

Iteration	Initial Residuals	New Predictions	New Residuals
0	1.910	4.072	1.727
1	2.206	2.109	1.989
2	2.981	2.021	2.683
3	5.498	1.779	4.945
4	1.744	2.492	1.570
5	3.244	2.642	2.929
6	2.634	2.054	2.371
7	2.092	2.537	1.882
8	2.691	2.597	2.422
9	1.784	2.300	1.619
10	3.534	1.964	3.181
11	2.589	2.060	2.322
12	2.352	2.082	2.127
13	1.016	3.324	9.065
14	8.482	2.233	7.634
15	2.594	2.065	2.345
16	8.514	1.476	7.662
17	7.110	2.391	6.409

18	1.204	2.207	1.084
19	3.362	2.654	3.033
20	2.586	2.060	2.329

In table 4.10, the results of the developed method appeared, where it was found that the extent of convergence between Initial Residuals and New Residuals, as well as New Predictions, is the purpose of showing the value of the predictor to be closer to the real values, and whoever approaches the real values, the result is better, and each time the learning coefficient is added to expand the range It is useful to reach the real values by step by step, where if the jump is made quickly and the real values are reached, the results will be inaccurate, which is the reason for using the learning coefficient α and continuing until it approaches the real values.

Here show Evaluation measures for each gas with the program execution time.

Table (4.11): The result of Evaluation measures

Types of Gas	Accuracy	Precision	Recall	F-Measurement	F_β	Execution Time (Seconds)
Gas #1	0.9779	0.5032	0.7129	0.5900	0.5245	2.4878
Gas #2	0.5227	0.4982	0.5354	0.7523	0.5494	2.5358
Gas #3	0.5226	0.9455	0.5074	0.8961	0.6115	3.0889
Gas #4	0.6607	0.4798	0.4007	0.7148	0.5276	3.0782
Gas #5	0.4892	0.5023	0.4955	0.4989	0.5014	2.5627
Gas #6	0.4943	0.5004	0.9158	0.7524	0.5513	3.0828

In table 4.11, the results of the Evaluation measures are shown, as it examines the efficiency of the model for each of the six types of gas, where each scale has a certain number that shows the accuracy of the system, and the best measure was found for each type of gas.

Table (4.12): The compare between the traditional XGBoost and DXGBoost-GSK

# Iterations	XGBoost		DXGBoost	
	Time(seconds)	Accuracy	Time (seconds)	Accuracy
1	4.198	0.428	2.618	0.905
2	4.200	0.387	2.618	0.906
3	4.203	0.245	2.618	0.907
4	4.204	0.452	2.618	0.908
5	4.206	0.665	2.618	0.909
6	4.207	0.590	2.618	0.943
7	4.209	0.562	2.618	0.944
8	4.210	0.547	2.618	0.946
9	4.212	0.538	2.618	0.948
10	4.214	0.532	2.618	0.950
11	4.216	0.527	2.618	0.953
12	4.223	0.524	2.618	0.956
13	4.227	0.521	2.618	0.959
14	4.229	0.519	2.619	0.961
15	4.231	0.517	2.619	0.965
16	4.233	0.516	2.619	0.969
17	4.239	0.515	2.619	0.973
18	4.242	0.514	2.619	0.976
19	4.244	0.513	2.623	0.979
20	4.245	0.512	2.623	0.983
21	4.247	0.511	2.623	0.986
22	4.249	0.511	2.623	0.989
23	4.251	0.510	2.623	0.994
24	4.253	0.509	2.623	0.997
25	4.257	0.509	2.623	0.999

In table 4.12, the results were presented and it was a comparison between the developed method and the traditional method in terms of

accuracy and execution time, where in the developed method the accuracy appeared and the best accuracy was 0.999 And the worst accuracy is 0.905, and it is considered a good accuracy as it can be relied upon in testing the model to know the extent of the model's reliability, and the execution time took 2.623 It is an almost standard time in order to be useful in testing large models in a short time and useful in shortening the time when the data is large.

As for the traditional method, where the best accuracy was 0.665 The worst accuracy was 0.245, but its accuracy is less, it is basically unreliable, and the time it took to implement is 4.257 seconds, it took more time than the developed method, so it is not useful, on the one hand, to be slower, and on the other hand, if the volume of data is large, it takes longer.

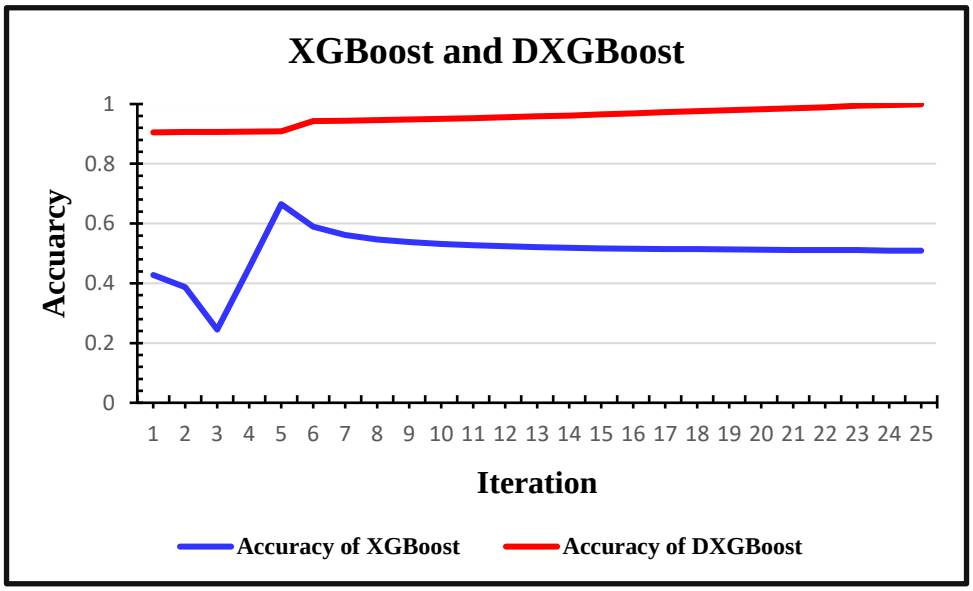

Fig (4.2): Compare traditional XGBoost with DXGBoost from aspect accuracy

Figure 4.2 shows the relationship between the developed method and the traditional method in terms of accuracy and was applied to the number of samples numbering 13910 and the number of columns 129 after applying the correlation to the data so that it becomes a matrix of 129 * 129. After applying the developed method to this matrix.

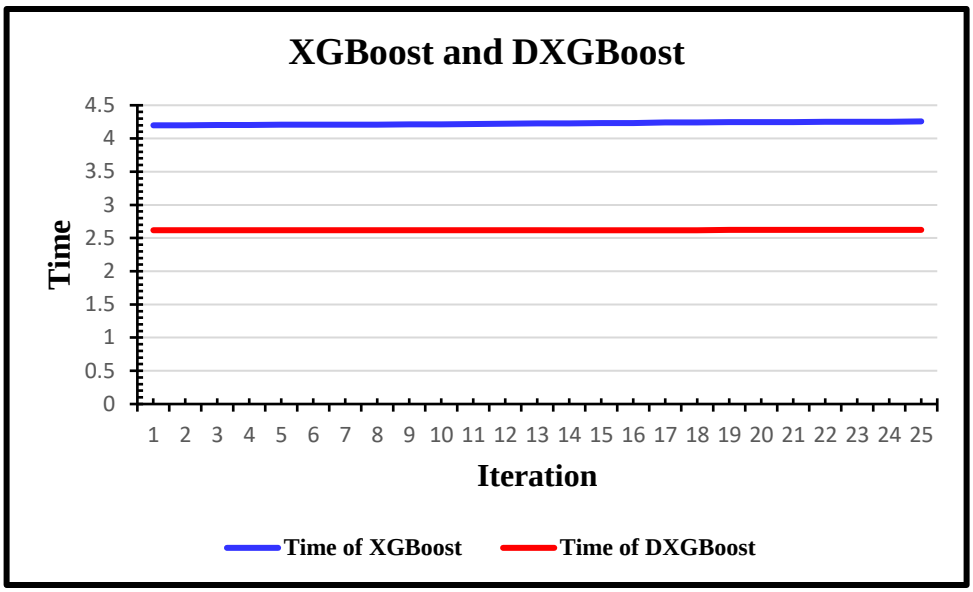

Fig (4.3): Compare traditional XGBoost with DXGBoost from aspect time

Figure 4.3 shows the execution time for both the developed method and the traditional method, and there is only a slight difference between the two methods, but certainly every number, even if it was a small number, affects the accuracy and quality of implementation.

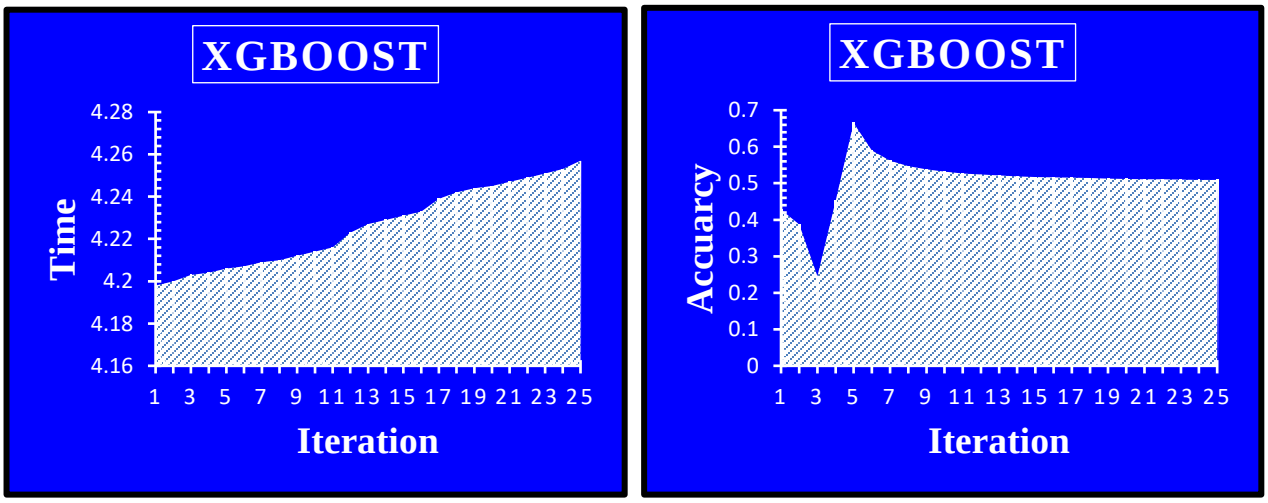

Fig (4.4): Traditional XGBoost

Figure 4.4 shows the traditional method of time and accuracy more clearly and accurately as the numbers in table 4.12 were applied, it turns out

that the time ranges between 4.198 and 4.257 and the accuracy ranges between 0.245 and 0.665.

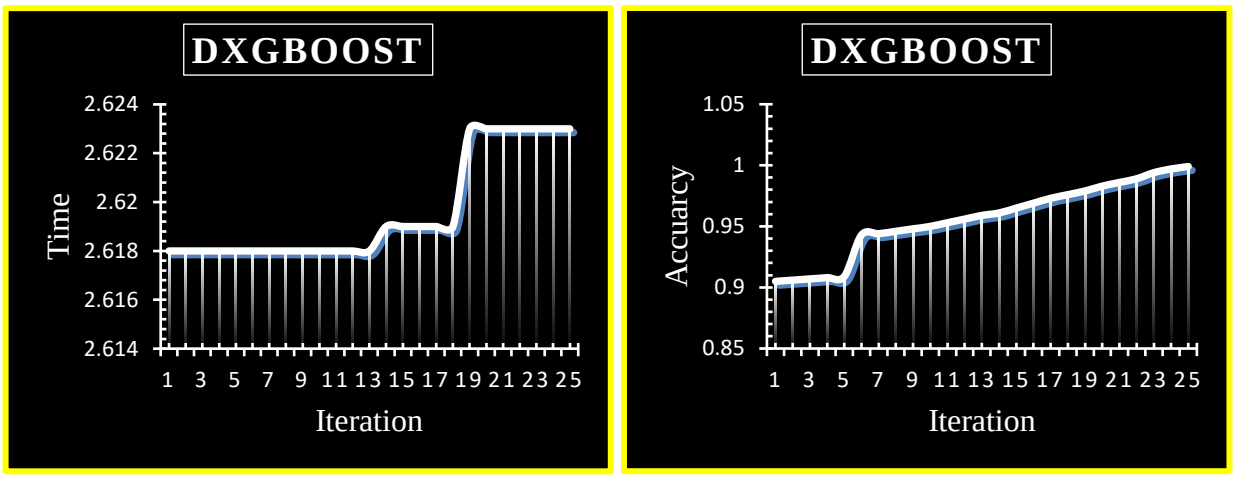

Fig (4.5): DXGBoost-GSK

The developed method is shown in figure 4.5, where both the accuracy and the time were clearly shown, as the numbers in table 4.12 were applied. The result is clear. The developed method was better in terms of its speed of performance and the accuracy of its results. Where the accuracy ranges from 0.905 and 0.999 The execution time ranges from 2.618 and 2.623 seconds.

Chapter Five:

Conclusions and Recommendations of Future

Chapter Five: Conclusions and Recommendations of Future

5.1 Introduction

This chapter reviews the most important conclusions reached through applying the HPM-STG into the dataset and focuses on how to address the challenges raised in the previous chapters, that include (programming challenges and application challenges). In addition, suggest a set of recommendations for researchers to work on it in the future.

5.2 Conclusions

A. The process of emission of gases as a result of chemical reactions is one of the most important problems that cause air pollution and affect living organisms, although the process of analyzing these gases is a very complex issue and requires a lot of time. But HPM-STG is able to process a large flow of data in a small time.

B. The data used in this research characteristic as very huge and split into multi groups related to 10 months, therefore at the first; aggregation of all data in a single dataset, and find the data have high duplication therefore handle this problem by take only the different interval to work on it, this step reduces the computation.

C. The correlation used in that model to determine which features from the 128 related to sensors are more affect in determining the type of gases. In general, found the following:

- The sensors more affect to determine the first gas is (FD1) in the first order and in the second order (F23, FC1) while the not important sensors are (F05, F24, F25, F32) therefore to reduce the computation can be neglected.

- The sensors more affect to determine the second gas (F63, FF3) in the first order and in the second-order are (F73, FA3, FE3) while the not important sensor is (F58) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the third gas (FD3, FF3) in the first order and in the second-order is (FE3) while the not important sensors are (F06, F07, F08) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the fourth gas (FF3) in the first order and in the second-order is (FE3) while the not important sensors are (F06, F07, F08) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the fifth gas are (F31, F63) in the first order and in the second-order (FE3, FF3, FF7) while the not important sensor is (F12) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the sixth gas are (F21, F63, FE4) in the first order and in the second-order (F73, FB1, FF4) while the not important sensor is (F12) therefore to reduce the computation can be neglected.

D. GSK is one of the pragmatic tools to work with real data, where, GSK characteristic thorny working in parallel and give high accuracy. In general; it is based on three parameters (Ackley function, Junior Phase, Senior Phase). Therefore, replacing the kernel of XGBoost with GSK are get high accuracy results but on the other side, the computation is increased. To reduce implementation time.

E. This work avoids the main drawbacks of XGBoost; where the kernel of XGBoost is the Decision tree, this makes it need to determine the

root; depth of the tree, In addition to high complexity. Through replace the kernel of it with GSK, enhance the performance of that algorithm from two points: reduce the implementation time and enhancement the performance.

5.3 Recommendations

A. It is possible to use another optimization algorithm that depends on the Agent principle as the kernel of the XGBoost algorithm, such as the Whale algorithm, the Lion algorithm, and the Practical swarm algorithm.

B. The HPM-STG implementation on CPU as hardware while; we can implement on other hardware such as GPU or FPGA.

C. It is also possible to use other types of sensors to study the effect of the emitted gas on the development of certain bacteria growth.

D. It is possible to use another technique for the classification process such as the Deep learning algorithm represented by Long Short-Term Memory (LSTM).

References

Reference

Agrawal, P., Alnowibet, K., & Mohamed, A. W. (2022). Gaining-sharing knowledge based algorithm for solving stochastic programming problems. Computers, Materials and Continua, 71(2), 2847. DOI:10.32604/cmc.2022.023126

Du, J., Zheng, J., Liang, Y., Lu, X., Klemeš, J. J., Varbanov, P. S., ... & Wang, B. (2022). A hybrid deep learning framework for predicting daily natural gas consumption. Energy, 257, 124689. https://doi.org/10.1016/j.energy.2022.124689

Kavzoglu, T., & Teke, A. (2022). Predictive Performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). Arabian Journal for Science and Engineering, 1-19. DOI: 10.1007/s13369-022-06560-8

Shi, J., Xie, W., Huang, X., Xiao, F., Usmani, A. S., Khan, F., ... & Chen, G. (2022). Real-time natural gas release forecasting by using physics-guided deep learning probability model. Journal of Cleaner Production, 368, 133201. https://doi.org/10.1016/j.jclepro.2022.133201

Wei, N., Yin, L., Li, C., Liu, J., Li, C., Huang, Y., & Zeng, F. (2022). Data complexity of daily natural gas consumption: Measurement and impact on forecasting performance. Energy, 238, 122090. https://doi.org/10.1016/j.energy.2021.122090

Al-Janabi, S. (2021, October). Overcoming the Main Challenges of Knowledge Discovery through Tendency to the Intelligent Data Analysis. In 2021 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 286-294). IEEE. DOI: 10.1109/ICDABI53623.2021.9655916

Al-Janabi, S., Alkaim, A., Al-Janabi, E., Aljeboree, A., & Mustafa, M. (2021). Intelligent forecaster of concentrations (PM2. 5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP). Neural Computing and Applications, 33(21), 14199-14229. https://doi.org/10.1007/s00521-021-06067-7

Al-Janabi, S., & Salman, A. H. (2021). Sensitive integration of multilevel optimization model in human activity recognition for smartphone and smartwatch applications. Big data mining and analytics, 4(2), 124-138. DOI: 10.26599/BDMA.2020.9020022

Batrice, R. J., & Gordon, J. C. (2021). Powering the next industrial revolution: transitioning from nonrenewable energy to solar fuels via CO 2 reduction. RSC advances, 11(1), 87-113. DOI: 10.1039/d0ra07790a

Reference

Abualigah, L., Shehab, M., Alshinwan, M., Mirjalili, S., & Elaziz, M. A. (2021). Ant lion optimizer: a comprehensive survey of its variants and applications. Archives of Computational Methods in Engineering, 28(3), 1397-1416. https://doi.org/10.1007/s11831-020-09420-6

Xiong, G., Li, L., Mohamed, A. W., Yuan, X., & Zhang, J. (2021). A new method for parameter extraction of solar photovoltaic models using gaining–sharing knowledge based algorithm. Energy Reports, 7, 3286-3301. https://doi.org/10.1016/j.egyr.2021.05.030

Rutherford, J. S., Sherwin, E. D., Ravikumar, A. P., Heath, G. A., Englander, J., Cooley, D., ... & Brandt, A. R. (2021). Closing the methane gap in US oil and natural gas production emissions inventories. Nature communications, 12(1), 1-12. https://doi.org/10.1038/s41467-021-25017-4

Wang, H., Qiao, L., Lu, S., Chen, F., Fang, Z., He, X., ... & He, T. (2021). A novel shale gas production prediction model based on machine learning and its application in optimization of multistage fractured horizontal wells. Frontiers in Earth Science, 675. doi: 10.3389/feart.2021.726537

Arul, R., Alroobaea, R., Mechti, S., Rubaiee, S., Andejany, M., Tariq, U., & Iftikhar, S. (2021). Intelligent data analytics in energy optimization for the internet of underwater things. Soft Computing, 25(18), 12507-12519. https://doi.org/10.1007/s00500-021-06002-x

Dashdondov, K., & Song, M. H. (2021). Factorial Analysis for Gas Leakage Risk Predictions from a Vehicle-Based Methane Survey. Applied Sciences, 12(1), 115. https://doi.org/10.3390/app12010115

Smajla, I., Sedlar, D. K., Vulin, D., & Jukić, L. (2021). Influence of smart meters on the accuracy of methods for forecasting natural gas consumption. Energy Reports, 7, 8287-8297. https://doi.org/10.1016/j.egyr.2021.06.014

Wang, H., Qiao, L., Lu, S., Chen, F., Fang, Z., He, X., ... & He, T. (2021). A novel shale gas production prediction model based on machine learning and its application in optimization of multistage fractured horizontal wells. Frontiers in Earth Science, 675. https://doi.org/10.3389/feart.2021.726537

Gao, F., & Shao, X. (2021). Forecasting annual natural gas consumption via the application of a novel hybrid model. Environmental Science and Pollution Research, 28(17),

Reference

21411-21424. https://doi.org/10.1007/s11356-020-12275-w

Salman, A. H., & Al-Janabi, S. (2020, December). Scheduling Activities of Smart Phone and Smart Watch Based on Optimal Pattern Model (SA-OPM). In International Conference on Innovations in Bio-Inspired Computing and Applications (pp. 164-178). Springer, Cham. https://doi.org/10.1007/978-3-030-73603-3_15

Al-Janabi, S., Mohammad, M., & Al-Sultan, A. (2020). A new method for prediction of air pollution based on intelligent computation. Soft Computing, 24(1), 661-680. doi:10.1007/s00500-019-04495-1

Al-Barmani, Z., & Al-Janabi, S. (2020, December). Intelligent Data Mining Techniques to Verification of Water Quality Index. In International Conference on Hybrid Intelligent Systems (pp. 590-605). Springer, Cham. https://doi.org/10.1007/978-3-030-73050-5_58

Al-Janabi, S., & Alkaim, A. F. (2020, December). A comparative analysis of DNA protein synthesis for solving optimization problems: a novel nature-inspired algorithm. In International Conference on Innovations in Bio-Inspired Computing and Applications (pp. 1-22). Springer, Cham. https://doi.org/10.1007/978-3-030-73603-3_1

Mohamed, A. W., Hadi, A. A., & Mohamed, A. K. (2020). Gaining-sharing knowledge based algorithm for solving optimization problems: a novel nature-inspired algorithm. International Journal of Machine Learning and Cybernetics, 11(7), 1501-1529. https://doi.org/10.1007/s13042-019-01053-x

Qiao, W., Moayedi, H., & Foong, L. K. (2020). Nature-inspired hybrid techniques of IWO, DA, ES, GA, and ICA, validated through a k-fold validation process predicting monthly natural gas consumption. Energy and Buildings, 217, 110023. https://doi.org/10.1016/j.enbuild.2020.110023

Qiao, W., Yang, Z., Kang, Z., & Pan, Z. (2020). Short-term natural gas consumption prediction based on Volterra adaptive filter and improved whale optimization algorithm. Engineering Applications of Artificial Intelligence, 87, 103323. https://doi.org/10.1016/j.engappai.2019.103323

Hassan, S. A., Ayman, Y. M., Alnowibet, K., Agrawal, P., & Mohamed, A. W. (2020). Stochastic travelling advisor problem simulation with a case study: A novel binary gaining-sharing knowledge-based optimization algorithm. Complexity, 2020.

Reference

https://doi.org/10.1155/2020/6692978

Meng, M., Zhong, R., & Wei, Z. (2020). Prediction of methane adsorption in shale: Classical models and machine learning based models. Fuel, 278, 118358. https://doi.org/10.1016/j.fuel.2020.118358

Chen, Y., Xu, X., & Koch, T. (2020). Day-ahead high-resolution forecasting of natural gas demand and supply in Germany with a hybrid model. Applied Energy, 262, 114486. https://doi.org/10.1016/j.apenergy.2019.114486

Miao, Y., Song, J., Wang, H., Hu, L., Hassan, M. M., & Chen, M. (2020). Smart micro-gas: A cognitive micro natural gas industrial ecosystem based on mixed blockchain and edge computing. IEEE Internet of Things Journal, 8(4), 2289-2299. DOI: 10.1109/JIOT.2020.3029138

Xiao, W., Liu, C., Wang, H., Zhou, M., Hossain, M. S., Alrashoud, M., & Muhammad, G. (2020). Blockchain for secure-GaS: Blockchain-powered secure natural gas IoT system with AI-enabled gas prediction and transaction in smart city. IEEE Internet of Things Journal, 8(8), 6305-6312. DOI: 10.1109/JIOT.2020.3028773

Assiri, A. S., Hussien, A. G., & Amin, M. (2020). Ant lion optimization: variants, hybrids, and applications. IEEE Access, 8, 77746-77764. DOI: 10.1109/ACCESS.2020.2990338

Guo, M. W., Wang, J. S., Zhu, L. F., Guo, S. S., & Xie, W. (2020). Improved ant lion optimizer based on spiral complex path searching patterns. IEEE Access, 8, 22094-22126. DOI: 10.1109/ACCESS.2020.2968943

Ghosh, A. M., & Grolinger, K. (2020). Edge-cloud computing for internet of things data analytics: embedding intelligence in the edge with deep learning. IEEE Transactions on Industrial Informatics, 17(3), 2191-2200. DOI: 10.1109/TII.2020.3008711

Abhishek, L. (2020, June). Optical character recognition using ensemble of SVM, MLP and extra trees classifier. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-4). IEEE. DOI: 10.1109/INCET49848.2020.9154050

Wei, N., Li, C., Peng, X., Li, Y., & Zeng, F. (2019). Daily natural gas consumption forecasting via the application of a novel hybrid model. Applied Energy, 250, 358-368. https://doi.org/10.1016/j.apenergy.2019.05.023

Potočnik, P., Šilc, J., & Papa, G. (2019). A comparison of models for forecasting the

Reference

residential natural gas demand of an urban area. Energy, 167, 511-522. https://doi.org/10.1016/j.energy.2018.10.175

Mirjalili, S. (2019). Genetic algorithm. In Evolutionary algorithms and neural networks (pp. 43-55). Springer, Cham. DOI: 10.1007/978-3-319-93025-1_4

Al_Janabi, S., Al_Shourbaji, I., & Salman, M. A. (2018). Assessing the suitability of soft computing approaches for forest fires prediction. Applied computing and informatics, 14(2), 214-224. https://doi.org/10.1016/j.aci.2017.09.006

Su, H., Zio, E., Zhang, J., Yang, Z., Li, X., & Zhang, Z. (2018). A systematic hybrid method for real-time prediction of system conditions in natural gas pipeline networks. Journal of Natural Gas Science and Engineering, 57, 31-44. https://doi.org/10.1016/j.jngse.2018.06.033

Akande, K. O., Owolabi, T. O., Olatunji, S. O., & AbdulRaheem, A. (2017). A hybrid particle swarm optimization and support vector regression model for modelling permeability prediction of hydrocarbon reservoir. Journal of Petroleum Science and Engineering, 150, 43-53. https://doi.org/10.1016/j.petrol.2016.11.033

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Ali, S. H. (2013, December). Novel approach for generating the key of stream cipher system using random forest data mining algorithm. In 2013 sixth international conference on developments in esystems engineering (pp. 259-269). IEEE. DOI: 10.1109/DeSE.2013.54

Debeljak, M., & Džeroski, S. (2011). Decision trees in ecological modelling. In Modelling complex ecological dynamics (pp. 197-209). Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-05029-9_14

Andres, C., & Lozano, S. (2006). A particle swarm optimization algorithm for part–machine grouping. Robotics and Computer-Integrated Manufacturing, 22(5-6), 468-474.https://doi.org/10.1016/j.rcim.2005.11.013

Pongcharoen, P., Hicks, C., Braiden, P. M., & Stewardson, D. J. (2002). Determining optimum genetic algorithm parameters for scheduling the manufacturing and assembly of complex products. International Journal of Production

Reference

Economics, 78(3), 311-322. https://doi.org/10.1016/S0925-5273(02)00104-4

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://link.springer.com/article/10.1023/a:1010933404324

Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks (Vol. 4, pp. 1942-1948). IEEE. DOI: 10.1109/ICNN.1995.488968

Holland, J. H. (1973). Genetic algorithms and the optimal allocation of trials. SIAM journal on computing, 2(2), 88-105. https://doi.org/10.1137/0202009

Appendix

Appendix A

Example: Gaining Sharing Knowledge

- Problem definition
  - The individuals at the beginning of the generation are the total number of individuals as a whole (i.e., the total number of rows) showing the first individual, then the total of the second row representing the second individual, then the total of the third row showing the third individual, and so on for the rest of the individuals and they represent members of the first generation.
  - For each individual found, the fitness of individuals for the first generation is calculated.
  - The initial global best is determined which is the first index.
  - Then determine the global initial worst which is the last index.
  - Junior phase steps are calculated:
    - Determine the random position of an individual (x_r).
    - Determine the random value.
    - Determine the neighbors of each individual (If the individual is in the last index, then the individual before the last is selected and the individual preceded by two places is selected. But if the individual is the first, then the one who comes next is chosen and the one who comes next is chosen two places. But if the individual is in the rest of the sites, then the one who precedes him and the one who follows him is chosen).
    - Apply the laws of junior, but according to the condition apply the law.

Appendix A

- After individuals appear after applying for the Junior, apply the steps of the Senior.
- Senior phase steps are calculated:
    - Find the best individual, worst individual, and better individual.
    - Apply the laws of senior, but according to the condition apply the law.
- Now a new generation is being generated, which appeared after the Senior step and calculated the value of Fitness for each individual.
- In the same way, the steps are repeated to generate a number of generations according to the number of generations specified.
- **The objective is the GSK method was used to tackle the set of real-world optimization problems provided for the competition of evolutionary algorithms.**

Table (A-1): Sample of population

F01	F02	F03	F04	F05	F06	F07	F08	Target
0.1	0.6	0.2	0.5	0.9	0.2	0.1	0.4	2
0.4	0.2	0.6	0.1	0.7	0.6	0.9	0.2	9
0.7	0.9	0.4	0.3	0.5	0.8	0.2	0.1	1
0.4	0.6	0.1	0.8	0.3	0.9	0.6	0.8	5
0.9	0.2	0.6	0.9	0.5	0.8	0.4	0.7	7

The individuals at beginning of generation:

First individual= (o.1 + 0.6 + 0.2 + 0.5 + 0.9 + 0.2 + 0.1 + 0.4 + 2) → 4.9

Second individual= (0.4 + 0.2 + 0.6 + 0.1 + 0.7 + 0.6 + 0.9 + 0.2 + 9) → 5.0

.
.
.

Table (A-2): First generation

First individual	Second individual	Third individual	Fourth individual	Fifth individual
4.9	5.0	9.5	12.0	12.7

Appendix A

Fitness of individuals at beginning of generation:

Fitness=-20*exp(-0.2*(sqrt((1/d) *(s**2)))) -((1/d) *cos(2*pi*s)) +(exp (1)) +20

Fitness=-20*exp (-0.2*(sqrt ((1/2) *(4.9**2)))) -((1/2) *cos(2*pi*4.9)) +(exp (1)) +20 → 12.311

Table (A-3): Fitness compute for each generation

Fitness of Generation1	Fitness of Generation 2	Fitness of Generation 3	Fitness of Generation 4	Fitness of Generation 5
12.311	12.356	17.999	18.553	19.553

- Junior Steps:

  If rand < kf:

    If fitness (individual [i]) < fitness (individual [xr])

      X= individual [i] + kf * ((b-a) + (individual [xr]- individual [i])

    Else

      X= individual [i] + kf * ((b-a) + (individual [i]- individual [xr])

  Else

    X= individual[i]

If 0.536 < 0.5

  If 12.311 < 0 → X= 4.9

  If 12.356 < 0 → X= 5.0

  If 17.999 < 0 → X= 9.5

  If 18.553 < 0 → X= 12.0

  If 19.553 < 0 → X= 12.7

- Senior Steps:
- Best =p →0.10 * problem size → 0.10* 5 → 0.5
- Worst = p → 0.10* 5 → 0.5
- Better = problem size -2* p → 5 -2*(0.10 * 5) → 4

  If rand<kr:

Appendix A

If fitness (individual [i]) <fitness(better):

X= individual [i]+kf*((best - worst) + (better- individual [i]))

else:

X= individual [i]+kf*((best - worst) + (individual [i]- better))

else:

individual [i]=X

If 0.536 < 0.9

If 12.311 < 10.858

X= individual [i]+kf*((best - worst) + (individual [i]- better))

X= 5.350

If 12.356 < 10.858 → X=5.5

If 17.999 < 10.858 → X= 12.25

If 18.553 < 10.858 → X= 16.0

If 19.553 < 10.858 → X= 17.049

- And so on to other individuals.

Table (A-4): Compute Junior phase and Senior phase for each generation

Number of generations	Junior Phase					Senior Phase				
1	4.9	5.0	9.5	12.0	12.7	5.35	5.5	12.2	16.0	17.0
2	5.35	5.5	12.25	16.0	17.04	6.02	6.25	16.3	22.0	23.5
3	0.96	1.18	13.6	27.5	41.8	2.48	2.59	18.5	39.3	60.7
4	2.48	2.59	18.5	60.7	39.3	3.24	3.29	25.7	89.1	57.0
5	3.24	3.29	25.7	57.0	89.1	3.62	3.64	36.6	83.5	131.6

Table (A-5): Compute Junior phase, Senior phase and Fitness Function (Ackley) for generation

Junior Phase	Senior Phase	Fitness Function (Ackley)
0.335	1.664	12.356
0.033	1.966	19.553
0.001	1.998	12.311
5.119	1.999	17.999
0.0	2.0	18.553

Appendix B

Example: Extreme Gradient Boosting (XGBoost)

- Problem definition
  - XGBoost creates new models on a regular basis and then combines them to form an ensemble model.
  - Create the first model and compute the error for each observation in the dataset.
  - Then you create a new model to forecast the residuals (errors).
  - The forecast from this model is then added to the ensemble of models.
  - XGBoost outperforms the gradient boosting technique because it strikes a reasonable compromise between bias and variance (Gradient boosting only optimized for the variance so tend to overfit training data while XGBoost offers regularization terms that can improve model generalization).
  - **The Objective Compare the predictions of the model to the actual weight. This signifies that the firm has completed the training data.**

Table (B-1): Initial Model (starting point)

Height	Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56
1.8	Red	Male	73
1.5	Green	Male	77
1.4	Blue	Female	57

$$\text{Average Weight} = \frac{88 + 76 + 56 + 73 + 77 + 57}{6}$$

Appendix B

$$\text{Error} = \text{True} - \text{Predicted}$$

$$88 - 71.2 = 16.8$$

$$76 - 71.2 = 4.8$$

$$56 - 71.2 = -15.2$$

$$73 - 71.2 = 1.8$$

$$77 - 71.2 = 5.8$$

$$57 - 71.2 = -14.2$$

Table (B-2): Calculate the errors based on the previous model (Residuals)

Height	Color	Gender	Weight (kg)	Residuals 1
1.6	Blue	Male	88	16.8
1.6	Green	Female	76	4.8
1.5	Blue	Female	56	-15.2
1.8	Red	Male	73	1.8
1.5	Green	Male	77	5.8
1.4	Blue	Female	57	-14.2

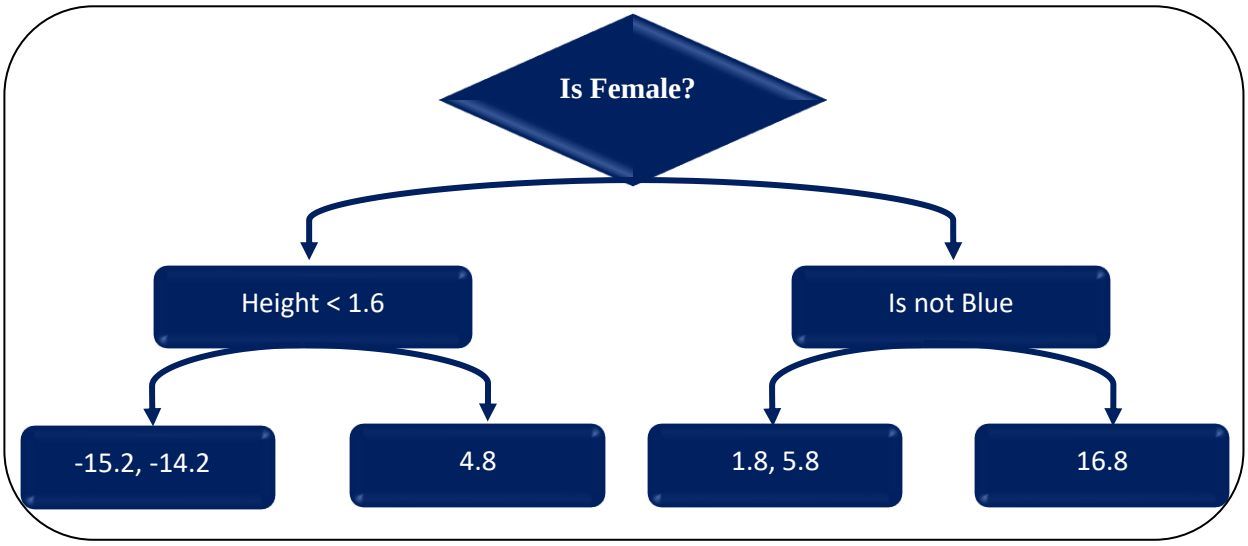

Fig (B-1): The results of Residuals

Appendix B

$$\text{Average } 1 = (-15.2 - 14.2)/2 \rightarrow -14.7$$

$$\text{Average } 1 = (1.8 + 5.8)/2 \rightarrow 3.8$$

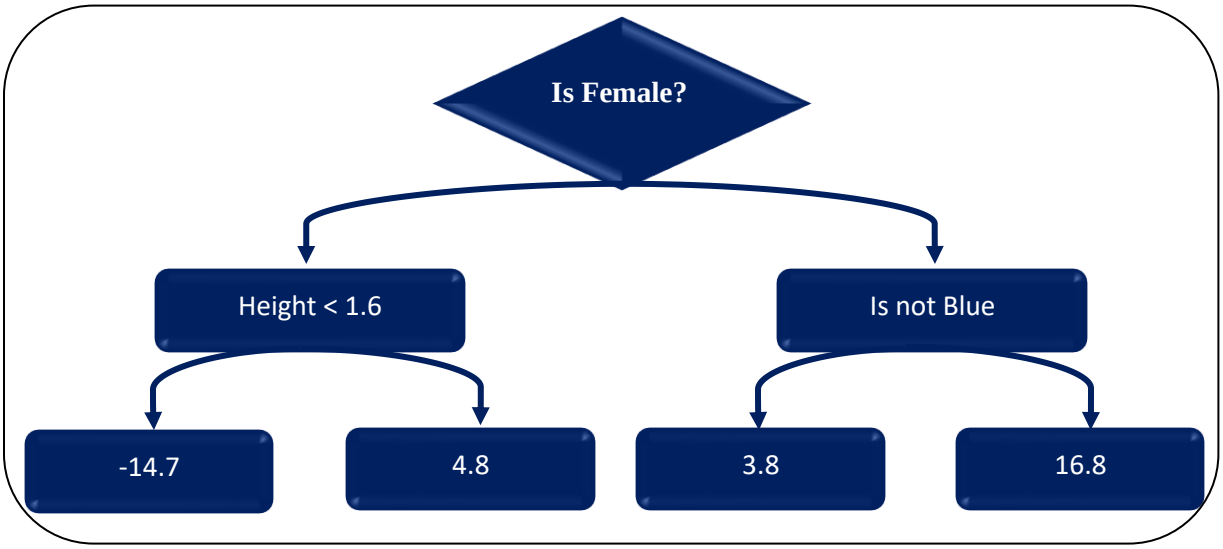

Fig (B-2): The results of final Residuals

- Let us now construct another tree using the new residual from the new forecast.

Table (B-3): Build a model to predict those errors

Initial Residuals	New Predictions	New Residuals
16.8	72.88	15.2
4.8	71.68	4.32
-15.2	69.68	-13.68
1.8	71.38	1.62
5.8	71.78	5.22
-14.2	69.78	-12.78

الخلاصة

يعتبر الغاز الطبيعي واحد من أهم الموارد الطبيعية ومصدر رئيسي للطاقة ، ونظراً لأهمية الغاز الطبيعي في العديد من المجالات منها الصناعة والتجارية ولما له من تاثير مباشر على حياة الكائنات الحية حيث تختلف الغازات في درجة اهميتها فمنها يعتبر اساسي لاستمرار الحياة مثل غاز الاوكسجين الى انسان وغاز ثاني اوكسيد الكاربون للنباتات والبعض الاخر يعتبر قاتل لاحتوائه على نسبة من السمية ونتيجة لسرعة انتشاره لذلك يعد بناء انظمة ذات اغراض متنوعة للتعامل مع الغازات الطبيعية هي واحدة من اهم التحديات التي يواجهها العالم اليوم.

تعتبر الـ XGBoost هي واحدة من افضل خوارزميات التنبىء المتعلقة بالتنقيب البيانات حيث تعطي نتائج ذات دقة عالية ويمكنها التعامل مع بيانات ذات حساسية عالية مثل بيانات العناية الصحية والملوثات سريعة الانتشار مثل ملوثات الهواء والماء والتربة ولكن من جانب اخر فانها تحتوي على العديد من الثغرات منها ان core/kernel لها هو اشجار القرار والتي تمتاز بعدة محددات منها يجب اختيار الجذر لها root عمق الشجر , عدد العقد الختامية . تقدم هذه الرسالة حل للتغلب على تلك المعوقات من خلال استبدال kernel للـ XGBoost بواحدة من تقنيات الامثلية المعتمدة مبدا agent هي GSK بعد تطويرها باستخدام دالة صلاحية جديدها معها.

تقدم هذه الرسالة نموذج هجين التنبؤ بنوع واحد من ستة انواع من الغاز الطبيعي سمي (HPM-STG) ويتالف من اربع مراحل اساسية : ***المرحلة الاولى استحصال*** البيانات من مصدرمخصصة للبحث العلمي تتعلق بالغاز الطبيعي. ***المرحلة الثانية:*** اجراء معالجة االولية للبينات وقسمت هذه المرحلة الى عدة خطوات: (ا) التحقق من القيم المفقودة وحذف اي صف يحتوي قم مفقودة .(ب) حساب الارتباط بين الخصائص والهدف. ***اما المرحلة الثالثة*** تم فيها تقسيم البيانات و بناء نموذج التنبىء المسمى (DGSK-XGB) . ***المرحلة الرابعة*** تم فيها استخدام خمس مقايس تقييم وهي (الدقة ، الدقة ، الاسترجاع ، القياس f ، و Fb).

امتاز النموذج الهجين المصمم بانه اداة واعدة في تصنيف انواع محددة من الغاز تلخصت بست انواع حيث بامكانه تحليل كمية كبيرة من البيانات في فترة زمنية قصيرة نسبيا. كما أن استبدال نواة XGBoost بـ GSK يعطي نتائج ذات دقة عالية نسبيا قياسا بالطريقة التقليدية لل XGBoost.

لا ثبات مدى جدوة النموذج الهجين المقترح تمت مقارنة الخوارزمية المطورة مع الخوارزمية التقليدية من حيث الدقة والوقت المستغرق للتنفيذ واثبتت طريقة المطورة دقة 90٪ مقارنة مع الطريقة التقليدية 50٪ مما يؤكد تفضيل الطريقة المطورة .

جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة بابل

كلية العلوم للبنات

قسم علوم الحاسوب

تحسين خوارزمية XGBoost بناءاً على تقنية التحسين

رسالة مقدمة

إلى مجلس كلية العلوم للبنات – جامعة بابل كجزء من متطلبات نيل درجة الماجستير في العلوم/ علوم الحاسوب

من قبل

هدير ماجد عبدالحسين الهنداوي

بإشراف

أ.د.سماهرحسين علي الجنابي أ.م.د. سيف محمود العلاك