New Tool (DGSK-XGB) for Forecasting Multi Types of Gas (T2E3A) Based on Intelligent Analytics

Hadeer Majed
Department of Computer Science,
Faculty of Science for Women (SCIW),
University of Babylon,
Babylon, Iraq

Samaher Al-Janabi *
Department of Computer Science,
Faculty of Science for Women (SCIW),
University of Babylon,
Babylon, Iraq
samaher@itnet.uobabylon.edu.iq
[0000-0003-2811-1493]

Saif Mahmood
Department of Computer Science,
Faculty of Science for Women (SCIW),
University of Babylon,
Babylon, Iraq

Abstract—***Natural gas is one of the alternative energy sources for oil. It is a highly efficient, low-cost, low emission fuel. Natural gas is an essential energy source for the chemical industry and is a critical component of the world's energy supply. It is also considered one of the cleanest, safest, and most beneficial sources of energy available. Therefore, this paper attempts to build a natural gas forecast optimization model as well as find out what type of gas is connected to the networks using intelligent data analysis. In this study; We will build the HPM-STG system, which consists of integrating two technologies together: XGBoost prediction technology and GSK optimization technology. This study will solve some problems, including the problem of data coming from natural gas that is collected manually, so the error rate will be because the prediction principle must be correct data, so we will solve it by collecting data through sensors, and the second problem that we will solve is the algorithm problem. Although the XGBoost algorithm is one of the best prediction algorithms, it faces many problems, so its core will be replaced by one of the optimization algorithms, and these two technologies together will give more accurate results.***

Keywords— E2T3A, DGSK-XGB, XGboost, GSK, Ethanol, Ethylene, Ammonia, Acetaldehyde, Acetone and Toluene

I. INTRODUCTION

Intelligent data analysis is one of the most important areas in real-world applications as well as computer science. We learn the artificial intelligence-based tools for finding information patterns by intelligently analyzing data to provide various techniques of exhibiting throughout the discovery or recovery pattern planning. The outcomes of the data evaluation and processing can be used in the application. As a result, a specific real-world problem must be addressed, realistic data must be assessed, and the best logic approach must be selected. To create a model that can assess data once it has been discovered. The reason for constructing rules, troubleshooting optimization, resulting in data, forecasting results, or providing a concise and relevant summary is the objective of the analysis. To succeed in today's demanding technological environment, gas and oil firms must build a diverse range of hybrid capabilities that allow production processes to interact with information technology. As a result of these requirements, IOFs emerge as a commercial capacity extension, with the leadership controlling the entire value chain rather than just the equipment. Large production environments with a wide assortment of assets open up new possibilities. (Wang, et al. 2022). It improves company operations, such as real-time visibility and process coordination, to bring assets closer to their optimal operating position.

The main problem of this paper is natural gas is the most important source in Iraq's economy and play a prominent role in controlling the country's development in various directions. Therefore, the question of forecasting its production rate for subsequent years is a very important point for drawing plans for a country according to rules and values that are closer to reality. Therefore, this paper attempts to build an optimization model to predict the gas associated with those networks using Intelligent data analysis. In this study, we will construct a five-step system: (a) Collect data from the Natural Gas network through IOT Platform in real-time, (b) Pre-processing that data based on split it into different intervals, and determining the main limitation and rules on it. (c) build predictive model called (HPM-STG). This predictor is based on Extreme Gradient Boosting (XGboost) typically using Decision Trees (DTs) to produce the predictor. However, it will replace the DTs with a Gaining-Sharing Knowledge-based Algorithm (GSK) since it has the ability to provide more accurate and optimal outcomes than DTs alone. (d) Finally, the HPM-STG outcomes would be assessed using five confusion matrix measures known as "AC, TP, P, F-

measure, and Fb". In addition, Cross-Validation will use to validate the accuracy of HPM-STG.

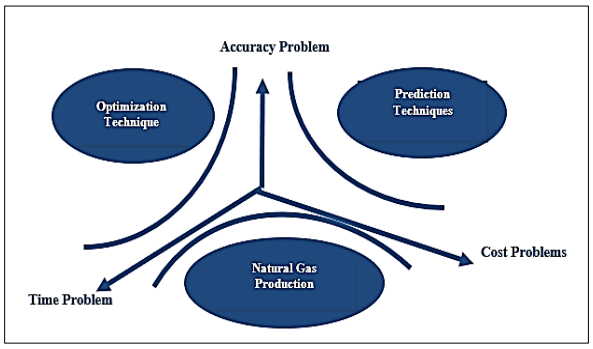

Fig.1. Relationships among the Main Three Challenges

II. RELATED WORK

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. (Taiyong Li et.al.2021) VMD-RSBL prediction technology will be used in conjunction with variable mode decomposition (prediction based on RSBL, SBL with delays that are random and random samples). It would be good to watch how VMD-RSBL corresponds to forecasting workloads in future time series, such as forecasting exchange rate, forecasting loads, and also forecasting wind speed. Simultaneously, we will evaluate the utility of the proposed approach for multivariate price forecasting and give policy recommendations based on the outcomes of the study. Predicting crude oil prices based on raw pricing data is a significant and time-consuming task in our sector. The disparity is due to the fact that we forecast using different methods.

(Huijun Wang et.al.2021) By merging machine learning and numerical tank, a program is used to build and analyze data-driven forecasting methods such as GPR, CNN, and SVM models. The GPR and tank models will come after that.The simulation was ran using both the evolutionary method and the standard optimization approach. On the basis of optimization, a data-driven model was constructed A procedure that is both quick and accurate. A replacement for the algorithm of aided numerical simulation optimization, Our approach is similar in that we applied an Optimization Algorithm, The new production of shale gas machine-based forecasting model learning differs from our business.

(Nalini Gupta and Shobhit Nigam .2020) The use of an artificial neural network to anticipate crude oil prices is a fresh and creative method (ANN) The key benefit of ANN's technique is that it continuously reflects the dynamic pattern of crude oil pricing that has been included during discovery. The optimal delay and delay effect number that governs crude oil prices, Our objective is equivalent to accurate prediction until there is a large and quick change in the real data, at which point it becomes impossible to successfully anticipate the new price. In contrast to our findings, the suggested model successfully accounts for these inclinations.

(David J.X.Gonzalez et.al.2022) The purpose of the study is to determine if oil and gas production pollutes the environment by examining the impact on primary oil and gas production (the number of drilling sites) and production activities (total volume of oil and gas). This flexibility to geographical, meteorological, environmental, and temporal aspects, as well as continual changes in wind direction as an external source of variation.

(Yue Su et.al.2021) DNN studies the influence of defect size on pipeline failure pressure, suggesting that the deep learning model outperforms empirical equations in terms of prediction accuracy. Simultaneously, a multi-layer ANN deep learning model outperforms a FEM simulation by at least two orders of magnitude. Our method is similar to deep learning, but we employ the FEM model in a different way.

III. HYBRID PREDICTION MODEL FOR NATURAL GAS (HPM-STG)

This paper presents the main stages of building the new predictor and shows the specific details for each stage. The HPM-STG is divided into five phases. The first collects data from the natural gas network in real-time utilizing devices linked to the network represented by the Internet of Things. The second stage, Pre-processing is the initial phase in the text mining process and plays an essential role in text mining techniques and applications. And that the pre-processing of the data (which was collected through devices, sensors, and tools) is based on dividing it into different periods and conducting some algorithms on it to filter it from impurities, and to determine the main restrictions and rules on it. We are going to build a predictive model called (HPM-STG). Figure (2) shows the main stages of HPM-STG and algorithm (1) shows the Building Prediction model.

We can summarize the main steps of this study as follows:

- Collecting data from the natural gas network through the IoT platform using devices, tools, and sensors.
- Through the pre-processing step, combine the datasets. and the basis of my work is the use of forecasting in data mining
- Create a new Hybrid predictor (HPM-STG) by combining the benefits of GSK and XGBoost

- Several metrics will be used to evaluate the prediction results as they are (accuracy, accuracy, reconnection, f-measurement, Fb).

Algorithm#1: Hybrid Prediction Model for Six Types of Gas (HPM-STG)

Input: Stream of real-time data capture from 16 sensors, each sensor, each give 8 features; the total number of features 128 collect from 16 sensors

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

// Pre-Processing Stage

1: ***For*** each row in gas dataset
2: ***For*** each column in gas dataset
3: **Call** Check Missing Values
4: **Call** Correlation
5: ***End for***
6: ***End for***

// Build DXGBoost-GSK Predictor

7: ***For*** i in range (1: total number of samples in Gas dataset)
8: Split dataset according to 5- Cross-Validation into Training and Testing dataset
9: ***End for***

10: **For** each Training part not used
11: **Call** DXGBoost-GSK //predictive the types of Gas
12: ***End for***

13: ***For*** each Testing part not used
14: Test stopping condition
15: ***IF*** max error generation < Emax
16: Go to step 21
17: ***Else***
18: GO to step 9
19: ***End IF***
20: End for

// Evaluation stage

21: ***Call Evaluation***

End HPM-STG

A. The HPM-STG Stages

The initial stage of developing an efficient prediction model in this part is dataset preparation, which comprises Drop Missing Value, Remove Duplication interval, and Five cross Validation. Using the GSK algorithm to forecast six different types of gases. The prediction model (HPM-STG) provides us with accurate findings, and the knowledge sharing acquisition procedure is employed to solve optimization issues in a continuous space. The final stage is to analyze the results using a variety of metrics.

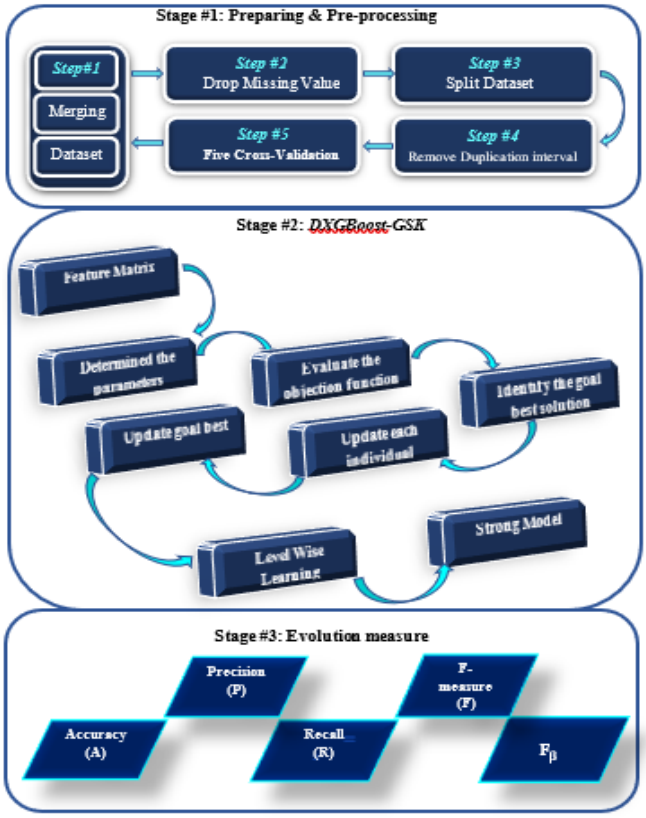

Fig .2. *Block diagram of Prediction Model*

1) Data Pre-process stage

The information was gathered over a period of several months as explained in algorithm 2.

- The data sets are merging and being stored into a single file.
- Drop Missing values. We utilize the ones in the database and delete the missing values to produce a proper forecast since error values would emerge if it is predicted using default values.
- Finally, for each column in the dataset, apply the correlation.
- Algorithm outlines the stage's major steps

Algorithm#2: Pre-processing

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

Output: The gas optimization approach is applied

// 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** j=null Then *//Check missing value*
4: Delete W[i,j]
5: **Else**
6: V[i,j]=W[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**

15: End Pre-processing

2) Building HPM-STG Predictor

The goal of developing a prediction model based on the combination of two technologies is to identify gases and then determine the kind of gases found. More details of that predictor with their parameters explain in algorithm 3.

Algorithm #3: DXGBoost-GSK

Input: Preprocessed Dataset

Output: Predictive Types of Gas

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=find iterations***

// Compute fitness function based on Ackley

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

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

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

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

5: ***End For***

6: ***End For***

7: Sort Population based on fitness function

8: G=0

// Calculation of the Gained and Shared Dimension Phases

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

10: ***For*** each row in gas dataset

11: ***For*** each column in gas 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*** $F(v_{i,j}^{new}) \leq F(v_{i,j}^{old})$ ***// Every vector is updated***

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

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

17: ***End IF***

18: ***IF*** $F(v_{i,j}^{new}) \leq\leq F(v_{best}^{G})$ ***// 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***

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

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

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

28: Prediction types of Gas:

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

30: Return Type of Gas

31: ***End For***

32: ***End For***

End DGSK-XGB

3) Evaluation Measures

The evaluation assesses how successfully program activities meet expected objectives and how much variation in outcomes may be attributable to the program. M&E is crucial because it enables program implementers to make objectively based decisions about program operations and service delivery.

TABLE I: MEASURES OF CONFUSION MATRIC

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

- ACCURACY

The accuracy of a classification technique or model is defined as the percentage of correct predictions. It is the proportion of "true" observations to all other observations.

$$\mathbf{AC} = \frac{a+d}{a+b+c+d} \quad (1)$$

- PRECISION (P)

It is the proportion of real positive outcomes to the total number of positive predictions made by the classifier.

$$\mathbf{P} = \frac{a}{a+c} \quad (2)$$

- RECALL (R)

Recall relates to how many true positives are correctly anticipated; it is the ratio of positives to the total number of positive class components.

$$\mathbf{TP} = \frac{a}{a+b} \quad (3)$$

- F-MEASUREMENT (F)

This measure is based on both measures: precision and recall. That measure compute as eq(4)

$$\mathbf{F} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

- FB

is the ratio of beta-factor multiplied by Precision and Recall divided by beta-squared multiplied by Precision plus Recall.

$$F_{\beta} = \frac{(1+\beta^2) * (Precision * Recall)}{\beta^2 * Precision + Recall} \quad (5)$$

IV. IMPLEMENTATION AND RESULT HPM-STG

Dataset contains data from 16 chemical sensors that were used in drift correction simulations in a discriminating test comprising 6 gases of varying concentrations. The data was collected at the gas distribution platform facility during a certain period of time (36 months). The measurement system platform allows for the versatility of obtaining desired concentrations of chemicals of interest with high accuracy and reproducibility, reducing common errors caused by human intervention and allowing the measurement system platform to focus solely on chemical sensors for truly meaningful compensation. the resulting data set comprises measurements of six distinct pure gaseous compounds: ammonia, acetaldehyde, acetone, ethylene, ethanol, and toluene.

The results for each step of the HPM-STG are shown in this section. All outcomes will also be justified. The table below shows the record number, the sample number, the properties number, and the target, as all records are equal to 6 numbers except for the third, fourth, and fifth records consisting of only 5 and missing the number which is 6.

Table 2 The Features of samples

Number of records	Number of samples	Number of features	# Gas
1	445	128	6
2	1244	128	6
3	1586	112	5
4	161	112	5
5	197	112	5
6	2300	128	6
7	3613	128	6
8	294	128	6
9	470	128	6
10	3600	128	6

A. Collection of Dataset

At this step, the data set is utilized to put the suggested model to the test*.

B. Pre-Processing

The outcomes of each step of the HPM-STG are shown in this section. Furthermore, all results must be supported by evidence. This section presents the main steps to preprocessing the dataset.

- ***Step #1: Merging***

In order to build a predictor of high accuracy with minimal computational complications, we will combine the following data from 36 different months together and deal with them as a single block to build the forecaster.

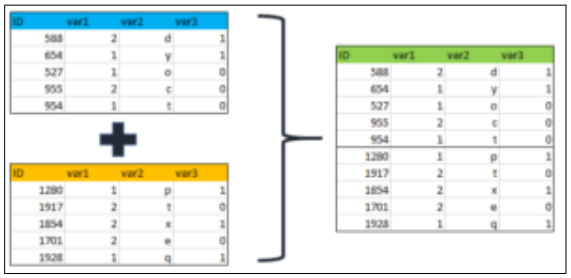

Fig 3 the way of merging datasets

After merging all the 10 data together, the record number as a whole will be equal to 13910.

- ***Step #2 Check Missing Value***

  We check the data if it contains missing values or not. All data has been checked after merging and no missing value appears, since we don't have a problem then there is no need to address it.

- ***Step #3: Correlation***

  The correlation relationship in the database to know the relationship that ties each value with its results, where the darker is closer to the outcomes and the lighter is after, and the checking process is easier for us.

In this part, we show tables of the relationship of each sensor with the target and the pictures that illustrate this, in addition to the tables that show each gas with the number of its repetitions.

Table 3 The first sensor correlation with the target

	F01	F02	F03	F04	F05	F06	F07	F08	F11	F12	F13	F14	F15	F16	F17	F18
F01	1	0.16993	0.98272	0.94190	0.58693	-0.96451	-0.7876	-0.20933	0.87905	0.414247	0.863494	0.84823	0.751252	0.75125	-0.82618	-0.10768
F02	0.169931	1	0.186405	0.176823	0.075992	-0.17717	-0.12863	0.014044	0.255107	0.815315	0.283978	0.278076	0.208185	-0.25593	-0.19809	-0.01354
F03	0.98272	0.186405	1	0.978705	0.636119	-0.97081	-0.81448	-0.2389	0.84328	0.441371	0.856729	0.854715	0.75637	-0.80844	-0.63048	-0.12525
F04	0.941909	0.176823	0.978705	1	0.741052	-0.94842	-0.83372	-0.30342	0.797083	0.425014	0.819756	0.834129	0.74282	-0.77785	-0.63177	-0.18352
F05	0.586937	0.075992	0.636119	0.741052	1	-0.58034	-0.52862	-0.24899	0.457099	0.188711	0.470362	0.482534	0.511706	-0.44695	-0.37487	-0.16637
F06	-0.96451	-0.17717	-0.97081	-0.94842	-0.58034	1	0.886767	0.342334	-0.84008	-0.43981	-0.84401	-0.84612	-0.72081	0.845855	0.708914	0.233659
F07	-0.7876	-0.12863	-0.81448	-0.83372	-0.52862	0.886767	1	0.678346	-0.65174	-0.3345	-0.6684	-0.7079	-0.63002	0.72306	0.790577	0.570898
F08	-0.20933	0.014044	-0.2389	-0.30342	-0.24899	0.342334	0.678346	1	-0.10021	0.009184	-0.12221	-0.20951	-0.29157	0.24493	0.564364	0.94018
F11	0.87905	0.255107	0.84328	0.797083	0.457099	-0.84008	-0.65174	-0.10021	1	0.567534	0.983361	0.958425	0.843955	-0.96602	-0.7484	-0.12794
F12	0.414247	0.815315	0.441371	0.425014	0.188711	-0.43981	-0.3345	0.009184	0.567534	1	0.618731	0.61855	0.451911	-0.57518	-0.44759	-0.0319
F13	0.863494	0.283978	0.856729	0.819756	0.470362	-0.84401	-0.6684	-0.12221	0.983361	0.618731	1	0.988515	0.867594	-0.96742	-0.76362	-0.14973
F14	0.84823	0.278076	0.854715	0.834129	0.482534	-0.84612	-0.7079	-0.20951	0.958425	0.61855	0.988515	1	0.88824	-0.95996	-0.79348	-0.23194
F15	0.751252	0.208185	0.75637	0.74282	0.511706	-0.72081	-0.63002	-0.29157	0.843955	0.451911	0.867594	0.88824	1	-0.8534	-0.73981	-0.3427

F16	-0.82618	-0.25593	-0.80844	-0.77785	-0.44695	0.84585	0.72306	0.24493	-0.96602	-0.57518	-0.96742	-0.95996	-0.8534	1	0.85964	0.29466
F17	-0.63253	-0.19809	-0.63048	-0.63177	-0.37487	0.708914	0.790577	0.564364	-0.7484	-0.44759	-0.76362	-0.79348	-0.73981	0.859647	1	0.650944
F18	-0.10768	-0.01354	-0.12525	-0.18352	-0.16637	0.233659	0.570898	0.94018	-0.12794	-0.0319	-0.14973	-0.23194	-0.3427	0.294666	0.650944	1
Targ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

- The sensors more affect to determine the first gas are (F13, F14) in the first order and in the second-order (F01, F03, F11) while the not important sensors are (F02, F05, F07, F08, F12, F17, F18) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the second gas (F23, F33) in the first order and in the second-order are (F24 , F25 , F34, F35) while the not important sensors are (F22, F28, F32, F38) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the third gas (F43 , F53) in the first order and in the second-order are (F44 , F51) while all other senses are not important therefore to reduce the computation can be neglected.
- The sensors more affect to determine the fourth gas (F63, F73) in the first order and in the second-order are all other
- The sensors more affect to determine the fifth gas are (F81, F83, F84, F85, F91, F93, F94, F95) in the first order and in the second-order (F86, F87, F88, F96, F97, F98) while the not important sensors are (F82, F92)) therefore to reduce the computation can be neglected.
- The sensors more affect to determine the six gas (FA1, FA2, FA3, FA4, FA5, FB1, FB2, FB3, FB4, FB5) in the first order while all other senses are not important therefore to reduce the computation can be neglected.

In general, we determined the sensor that have 0.80 or more than as correlated with target as important features.

C. Implementation and Result of DGSK-XGB Stage

Choose the settings that are acceptable for the learning algorithm under consideration. One of the greatest obstacles in science is that XGBoost takes too long to implement and provide results, thus this part discusses how DGSK-XGB handles this problem and overcomes this challenge.

In other words, determining the weights and model number (M) are critical elements that primarily influence DGSK –XGB performance. Table 4 displays the primary characteristics of DGSK –XGB.

Table 4 shows the parameters used in DGSK-XGB

Parameter	Value
Max depth	6
Learning rate	0.3
n_estimators	100
colsample_bytree	1
Subsample	1
population size	13911
knowledge factor	0.5
knowledge ratio	0.9
Knowledge rate	10
P	0.1
A	20
B	0.2
C	2π

1) results of GSK

This approach, which is based on the aggregation principle, is used to address optimization issues. Following the application of the algorithm, six groups were displayed, each displaying a different gas.

TABLE 5: RESULTS OF GSK OPTIMIZATION ALGORITHM REPRESENT THE NUMBER OF POINTS IN EACH GROUP

Number of Group	Number of points related to that group
Group #1	2565
Group #2	2926
Group #3	1641
Group #4	1936
Group #5	3009
Group #6	1833

2) Result XGBoost

The categorization is determined using this algorithm. After determining the assembly using the previous technique, we will determine its classification in order to determine the classification of each gas.

TABLE 6: RESULT XGBOOST

Target	Average	Initial Residuals	New predictions	New Residuals
1	0.184386457	0.815613543	0.265947811	0.734052189
2	0.210337143	1.789662857	0.389303429	1.610696571
3	0.117964201	2.882035799	0.406167781	2.593832219
4	0.139170441	3.860829559	0.525253397	3.474746603
5	0.216303645	4.783696355	0.69467328	4.30532672
6	0.131766228	5.868233772	0.718589605	5.281410395

D. Evaluation Measure

The extent to which changes in outcomes may be ascribed to the program is measured by evaluating how successfully the program activities meet expected objectives. M&E is crucial because it enables program implementers to make objectively based decisions about program operations and service delivery. When we create a predictive model, we must assess

it using the following metrics: Precision, Accuracy, Recall, F-Measurement, Fb.

TABLE: EVALUATION MEASURE OF CONFUSION MATRIX-BASED ON HPM-STG

Target	Accuracy	Precision	Recall	F-measure	Fβ
1	0.625548127	0.405871511	0.26594781	0.321338302	0.353567872
2	0.59662713	0.489399772	0.19465171	0.27852437	0.287894611
3	0.554416341	0.436073165	0.13538926	0.20662644	0.186880451
4	0.536731124	0.430561448	0.13131335	0.201249339	0.179999917
5	0.543280334	0.491538614	0.13893466	0.216636459	0.206070719
6	0.119764934	1	0.11976493	0.21391085	0.359294803

V. CONCLUSIONS

In this study; We will construct a five-step system:(a) Collect data from multi senseor of Gas network through IOT Platform in real-time, (b) Pre-processing that data based on split it into different intervals. (c) build predictive model called (HPM-STG). This predictor is based on Extreme Gradient Boosting (XGboost) typically using Decision Trees (DTs) to produce the predictor. However, it will replace DTs with a Gaining-Sharing Knowledge-based Algorithm (GSK) since it has the ability to provide more accurate and optimal outcomes than DTs alone. (d) Finally, the HPM-STG outcomes would be assessed using five confusion matrix metrics referred to as "AC, TP, P, F-measure, and Fb".

- ***How Gaining-Sharing Knowledge-based (GSK) can be useful in building a new predictor called (HPM-STG)?***
  It had a positive effect because it worked to determine the number of points belonging to each group based on an effective activation function that included both Junior and Senior, and one of its benefits was that it cut the execution time of XGboost, (which had the advantage of requiring many parameters to be specified, such as depth Tree, root selection, and be of great complexity).
- ***How can build an optimal prediction model by replacing the kernel of XGboost with Gaining-Sharing Knowledge-based (GSK).?***
  The basis of XGboost is DT, which has several problems as explained in point one. As a result, in this study, one of the options (HPM-STG) was employed, with GSK as the core at XGboost rather than DT, to minimize time complexity and enhance accuracy while increasing the number of calculations.
- ***Is the assessment measure utilized sufficient to evaluate the results of the suggested predictor?***
  Yes, the confession matrix has five different measures to compute the repost of new prediction XGboost and those measures are sufficient to determine the degree confidence of the predictor.
- ***What are the advantages of developing a predictor using a combination of GSK and XGboost?***
  Through the development of a new predictor known as HPM-STG, which combines GSK and XGboost, where GSK was used to discover the best group of each Gas and the quantity of points associated with it. while XGboost predicts the kind of gas.

REFERENCES

[1] Abad, A. R. B., Ghorbani, H., Mohamadian, N., Davoodi, S., Mehrad, M., Aghdam, S. K. Y., & Nasriani, H. R. (2022). Robust hybrid machine learning algorithms for gas flow rates prediction through wellhead chokes in gas condensate fields. Fuel, 308, 121872. https://doi.org/10.1016/j.fuel.2021.121872

[2] Akhmedova, S., & Stanovov, V. (2021, July). Success-History Based Position Adaptation in Gaining-Sharing Knowledge Based Algorithm. In International Conference on Swarm Intelligence (pp. 174-181). Springer, Cham. DOI: 10.1007/978-3-030-78743-1_16

[3] Al-Janabi, S., & Mahdi, M. A. (2019). Evaluation prediction techniques to achievement an optimal biomedical analysis. International Journal of Grid and Utility Computing, 10(5), 512-527. https://doi.org/10.1504/ijguc.2019.102021

[4] Al-Jlibawi, A., Othman, M. L. B., Al-Huseiny, M. S., Aris, I. B., & Bahari, S. (2020, May). The efficiency of soft sensors modelling in advanced control systems in oil refinery through the application of hybrid intelligent data mining techniques. In Journal of Physics: Conference Series (Vol. 1529, No. 5, p. 052049). IOP Publishing. doi:10.1088/1742-6596/1529/5/052049

[5] Alkaim, A. F., & Al_Janabi, S. (2019, April). Multi objectives optimization to gas flaring reduction from oil production. In International conference on big data and networks technologies (pp. 117-139). Springer, Cham. https://doi.org/10.1007/978-3-030-23672-4_10

[6] Al-Janabi, S., Alkaim, A., Al-Janabi, E. et al. (2021) Intelligent forecaster of concentrations (PM2.5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP). Neural Comput & Applic 33, 14199–14229. https://doi.org/10.1007/s00521-021-06067-7

[7] Al-Janabi, S. & Alkaim, A.F. (2020). A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation, Springer, Soft Comput, Volume 24, Issue 1, pp 555–569. DOI 10.1007/s00500-019-03972-x

[8] Al-Janabi, S., Alkaim, A.F. & Adel, Z. (2020). An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft Comput 24, 10943–10962. https://doi.org/10.1007/s00500-020-04905-9

[9] Samaher Al_Janabi, Ibrahim Al_Shourbaji, Mahdi A. Salman (2018), Assessing the suitability of soft computing approaches for forest fires prediction, Applied Computing and Informatics, Volume 14, Issue 2,PP 214-224 ISSN 2210-8327, https://doi.org/10.1016/j.aci.2017.09.006

[10] Caiza, G., Garcia, C. A., Naranjo, J. E., & Garcia, M. V. (2020). Flexible robotic teleoperation architecture for intelligent oil fields. Heliyon, 6(4), e03833. https://doi.org/10.1016/j.heliyon.2020.e03833

[11] Chung, D. D. (2020). Materials for electromagnetic interference shielding. Materials Chemistry and Physics, 123587. https://doi.org/10.1016/j.matchemphys.2020.123587

[12] Cotfas, L. A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D. S., & Tajariol, F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement. IEEE Access, 9, 33203-33223. DOI: 10.1109/ACCESS.2021.3059821

[13] da Veiga, A. P., Martins, I. O., Barcelos, J. G., Ferreira, M. V. D., Alves, E. B., da Silva, A. K., ... & Barbosa Jr, J. R. (2022). Predicting thermal expansion pressure buildup in a deepwater oil well with an annulus partially filled with nitrogen. Journal of Petroleum Science and Engineering, 208, 109275. https://doi.org/10.1016/j.petrol.2021.109275

[14] Fernandez-Vidal, J., Gonzalez, R., Gasco, J., & Llopis, J. (2022). Digitalization and corporate transformation: The case of European oil & gas firms. Technological Forecasting and Social Change, 174, 121293. https://doi.org/10.1016/j.techfore.2021.121293

[15] Foroudi, S., Gharavi, A., & Fatemi, M. (2022). Assessment of two-phase relative permeability hysteresis models for oil/water, gas/water and gas/oil systems in mixed-wet porous media. Fuel, 309, 122150. https://doi.org/10.1016/j.fuel.2021.122150

[16] Gao, Q., Xu, H., & Li, A. (2022). The analysis of commodity demand predication in supply chain network based on particle swarm optimization algorithm. Journal of Computational and Applied Mathematics, 400, 113760. https://doi.org/10.1016/j.cam.2021.113760

[17] Gonzalez, D. J., Francis, C. K., Shaw, G. M., Cullen, M. R., Baiocchi, M., & Burke, M. (2022). Upstream oil and gas production and ambient air pollution in California. Science of The Total Environment, 806, 150298. https://doi.org/10.1016/j.scitotenv.2021.150298

[18] Guo, W., Jiang, M., Li, X., & Ren, B. (2018). Using a genetic algorithm to improve oil spill prediction. Marine pollution bulletin, 135,386-396. https://doi.org/10.1016/j.marpolbul.2018.07.026

[19] Gupta, N., & Nigam, S. (2020). Crude oil price prediction using artificial neural network. Procedia Computer Science, 170, 642-647. https://doi.org/10.1016/j.procs.2020.03.136

[20] Hao, P., Di, L., & Guo, L. (2022). Estimation of crop evapotranspiration from MODIS data by combining random forest and trapezoidal models. Agricultural Water Management, 259, 107249. https://doi.org/10.1016/j.agwat.2021.107249

[21] Heidari, A. A., Faris, H., Mirjalili, S., Aljarah, I., & Mafarja, M. (2020). Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks. Nature-Inspired Optimizers, 23-46. DOI: 10.1007/978-3-030-12127-3_3

[22] Houssein, E. H., Gad, A. G., Hussain, K., & Suganthan, P. N. (2021). Major Advances in Particle Swarm Optimization: Theory, Analysis, and Application. Swarm and Evolutionary Computation, 63, 100868. https://doi.org/10.1016/j.swevo.2021.100868

[23] Johny, J., Amos, S., & Prabhu, R. (2021). Optical Fibre-Based Sensors for Oil and Gas Applications. Sensors, 21(18), 6047.https://doi.org/10.3390/s21186047

[24] Kodan, R., Rashed, M. S., Pandit, M. K., Parmar, P., & Pathania, S. (2022). Internet of things in food industry. In Innovation Strategies in the Food Industry (pp. 287-303). Academic Press. DOI: 10.1007/978-3-030-88095-8_1

[25] Kumar, A., Vohra, M., Pant, S., & Singh, S. K. (2021). Optimization techniques for petroleum engineering: A brief review. International Journal of Modelling and Simulation, 1-9.https://doi.org/10.1080/02286203.2021.1983074

[26] Larestani, A., Mousavi, S. P., Hadavimoghaddam, F., & Hemmati-Sarapardeh, A. (2022). Predicting formation damage of oil fields due to mineral scaling during water-flooding operations: Gradient boosting decision tree and cascade-forward back-propagation network. Journal of Petroleum Science and Engineering, 208, 109315. https://doi.org/10.1016/j.petrol.2021.109315

[27] Larestani, A., Mousavi, S. P., Hadavimoghaddam, F., & Hemmati-Sarapardeh, A. (2022). Predicting formation damage of oil fields due to mineral scaling during water-flooding operations: Gradient boosting decision tree and cascade-forward back-propagation network. Journal of Petroleum Science and Engineering, 208, 109315. https://doi.org/10.1016/j.petrol.2021.109315

[28] Li, H., Yu, H., Cao, N., Tian, H., & Cheng, S. (2020). Applications of artificial intelligence in oil and gas development. Archives of Computational Methods in Engineering, 1-13. https://doi.org/10.1007/s11831-020-09402-8

[29] Mahdi, M. A., & Al-Janabi, S. (2019, April). A novel software to improve healthcare base on predictive analytics and mobile services for cloud data centers. In International conference on big data and networks technologies (pp. 320-339). Springer, Cham. https://doi.org/10.1007/978-3-030-23672-4_23

[30] Mohamed, A. W., Abutarboush, H. F., Hadi, A. A., & Mohamed, A. K. (2021). Gaining-Sharing Knowledge Based Algorithm With Adaptive Parameters for Engineering Optimization. IEEE Access, 9, 65934-65946. DOI: 10.1109/ACCESS.2021.3076091

[31] 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

[32] Samaher Al-Janabi, Ayad Alkaim, A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis, Egyptian Informatics Journal, 2022, https://doi.org/10.1016/j.eij.2022.01.004.

[33] Ossai, C. I. (2020). Modified spatio-temporal neural networks for failure risk prognosis and status forecasting of permanent downhole pressure gauge. Journal of Petroleum Science and Engineering, 184, 106496.https://doi.org/10.1016/j.petrol.2019.106496

[34] Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. Journal of Petroleum Science and Engineering, 208, 109520. https://doi.org/10.1016/j.petrol.2021.109520

[35] Phan, H. C., Bui, N. D., Pham, T. D., & Duong, H. T. (2022). Predicting capacity of defected pipe under bending moment with data-driven model. In Modern Mechanics and Applications (pp. 830-840). Springer, Singapore. DOI: 10.1007/978-981-16-3239-6_64

[36] Priyanka, E. B., Maheswari, C., & Thangavel, S. (2021). A smart-integrated IoT module for intelligent transportation in oil industry. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, 34(3), e2731. https://doi.org/10.1002/jnm.2731

[37] 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

[38] S. H. Ali, (2012) "A novel tool (FP-KC) for handle the three main dimensions reduction and association rule mining," IEEE,6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Sousse, 2012, pp. 951-961. doi: 10.1109/SETIT.2012.6482042

[39] Samaher Al-Janabi, Sarvesh Rawat, Ahmed Patel, Ibrahim Al-Shourbaji (2015), Design and evaluation of a hybrid system for detection and prediction of faults in electrical transformers, International Journal of Electrical Power & Energy Systems, Volume 67, Pages 324-335, https://doi.org/10.1016/j.ijepes.2014.12.005.

[40] Su, Y., Li, J., Yu, B., Zhao, Y., & Yao, J. (2021). Fast and accurate prediction of failure pressure of oil and gas defective pipelines using the deep learning model. Reliability Engineering & System Safety, 216, 108016. https://doi.org/10.1016/j.ress.2021.108016

[41] Suma, V., & Hills, S. M. (2020). Data mining based prediction of demand in Indian market for refurbished electronics. Journal of Soft Computing Paradigm (JSCP), 2(02), 101-110. https://doi.org/10.36548/jscp.2020.2.007

[42] Taiyong Li, Zijie Qian, Wu Deng, Duzhong Zhang, Huihui Lu,(2021) Shuheng Wang,Forecasting crude oil prices based on variational mode decomposition and random sparse Bayesian learning,Applied Soft Computing ,108032,ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021.108032

[43] Thethy, B., Kaebe, B., Honnery, D., Edgington-Mitchell, D., & Kleine, H. (2020). Influence of pressure transducer protrusion depth on pressure measurements of shock waves in shock tubes. Review of Scientific Instruments, 91(10), 106101. https://doi.org/10.1063/5.0016593