Hybridized Deep Learning Model with Optimization Algorithm: A Novel Methodology for Prediction of Natural Gas

Hadeer Majed, Samaher Al-Janabi(✉), and Saif Mahmood

Faculty of Science for Women (SCIW), Department of Computer Science, University of Babylon, Hillah, Iraq
samaher@itnet.uobabylon.edu.iq

Abstract. This paper handles the main problem of natural gas through design the hybrid model based on developing one of predict data mining techniques. The model consists of four stages; The first stage collects data from a different source related to natural gas in real-time. 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. As a results; we found DGSK-XGB give high accuracy reach to 93% compare with the tractional XGBoost; also, it reduces implementation time. And improving the performance.

Keywords: Natural Gas · XGboost · GSK · Optimization techniques

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 paper attempts to build a model that classifies six basic types of those gases, which are (Ethanol, Ethylene, Ammonia, Acetaldehyde, Acetone, and Toluene) [1, 2].

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, we shall eliminate both methane and non-hydrocarbons (water, carbon dioxide, hydrogen sulfide). That 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 advantages of natural gas is that it burns entirely when used, and unlike other traditional energy sources, the carbon dioxide produced when burning is absolutely non-toxic [3, 4]. Natural gas is a pure gas by nature, and any contaminants that may be present in it may sometimes be simply and inexpensively

A. Abraham et al. (Eds.): HIS 2022, LNNS 647, pp. 79–95, 2023.
https://doi.org/10.1007/978-3-031-27409-1_8

eliminated. Natural gas stations are not generally distributed and 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 constant temperature reduces the volume of the gas [5]. In other words, Boyle's law asserts that volume is inversely proportional to pressure when the temperature and number of molecules stay constant. Natural gas is composed of hydrocarbon components such as methane, but also ethane, propane, butane, and pentane, all of which are referred to as natural gas liquids (NGLs), as well as impurities such as carbon dioxide (CO_2), hydrogen sulfide (H_2S), water, and nitrogen [6].

Intelligent Data Analysis (IDA) [7, 15, 26] is 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 found 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 & useful for any person not only for the person expert in the specific field of problem.

This paper will 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.

The problem of this work is divided into parts: The first part 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 called XGboost; in general; 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 paper is how can avoid these limitations (i.e., high computation and time of implementation) of this algorithm and befit from their features. On other side; The problem of application can summarization by need of high efficiency prediction techniques; Therefore, the second challenge of this paper is how can avoid these limitations thought build an efficient technique to predict multi types of gas coming from different sensors.

2 Main Tools

Optimization [7, 15] is one of the main models 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 objective's function 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 [8].

2.1 Optimization Techniques [9–11]

2.2 Particle Swarm Optimization Algorithm (PSO)

Eberhart and Kennedy devised one of the swarm intelligence methods, particle swarm optimization (PSO), in 1995. It's a population-based, stochastic algorithm inspired by social behaviors seen in confined birds. It is one of the approaches to evolutionary optimization.

2.3 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.

2.4 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.

2.5 Gaining-Sharing Knowledge-Based Algorithm (GSK) [12, 16]

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 (see Table 1).

Table 1. Analytic the Advantages and Disadvantages for Optimization Techniques.

O T	Advantage	Disadvantage
PSO	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	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	Implementing GA is still a work in progress 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	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	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

2.6 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 paper works with the first type of that technique. as explain below.

2.7 The 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 [13].

2.8 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 [14].

2.9 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" [17].

2.10 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 [18] (see Table 2).

Table 2. Analytic the Advantages and Disadvantages for Prediction Techniques.

PT	Advantage	Disadvantage
DT [24]	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

(*continued*)

Table 2. (*continued*)

PT	Advantage	Disadvantage
ETC [25]	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 [26]	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	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

3 Proposed Method (HPM-STG)

This section 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 stages; The first stage collects data from a different source related to natural gas in real-time. 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 Fig. 1, and the steps of the model are shown in the algorithm (1). We can summarize the main stages of this research below:

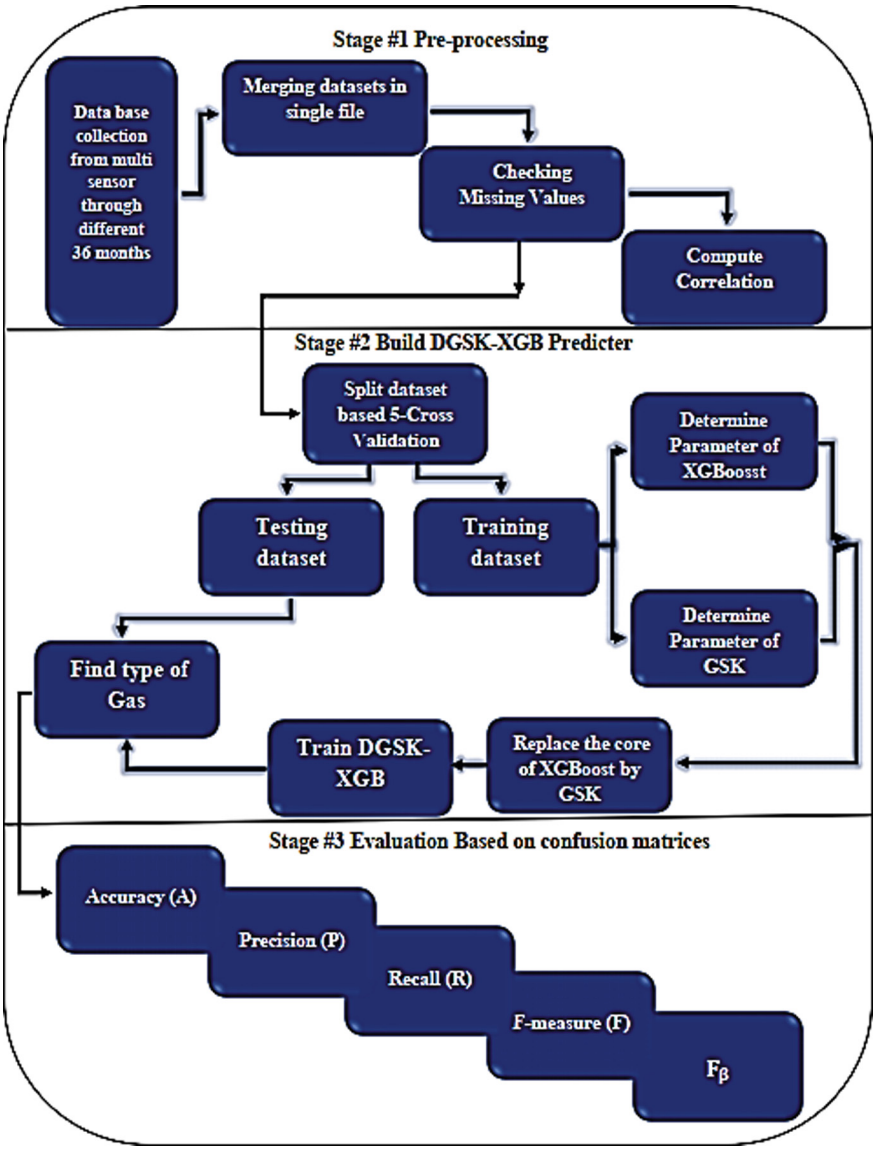

Fig. 1. Block diagram of DGSK-XGB Model

- Capture data from scientific location on internet where, these data collection 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).

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

Input: Stream of real-time data capture from 16 sensors, each sensor, 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 dataset*

2: ***For*** *each column in dataset*

3: ***Call*** *Check Missing Values*

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

5: ***End*** *for*

6: ***End*** *for*

*// **Build DGSK –XGB Predictor***

7: ***For*** *i in range (1: total number of samples in dataset)*

8: *Split dataset according to Five- Cross-Validation into Training and Testing dataset*

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*** *max 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

4 Results

This section of the paper plain the main results; In addition, described the details of a database used to implement the DXGboost-GSk model.

4.1 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 called Ammonia, Acetaldehyde, Acetone, Ethylene, Ethanol, and Toluene.

4.2 Result of Preprocessing

This stage begin form get the database from scientific internet sit, where these database aggregation from multi sensors through different periods of time include 36 months. Split into ten groups.

4.3 Checking Missing Value [21]

After Merging all datasets in single file; we 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.

4.4 Correlation [19, 20]

The correlation is computed among all the features with the target to determine the main features effect in specific type of gas. In general, we Found 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 go 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.

4.5 Results of DXGBoost-GSk

This section of chapter will apply 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 as shown in Table 3. Through five cross validations, where; we 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 who split gives the lowest error rate is depend on build the final model. In general; the total number of samples of these datasets are 13910.

Table 3. Number of samples of training and testing dataset based on five cross validations

Rate training dataset	# samples	Rate testing dataset	# samples
80%	11128	20%	2782
60%	8346	40%	5564
50%	6955	50%	6955
40%	5564	60%	8346
20%	2782	80%	11128

The Table 4 shows results of GSK based on three equations: junior, senior, and Ackley.

Table 4. The Result of GSK

It	Junior	Senior	Ackley
1	10.15019163870712	0.8498083612928795	22.753980010395882
2	9.35839324839964	1.64160675160036	22.725819840576897
3	8.621176953814654	2.378823046185346	22.627559663453333
4	7.935285368822167	3.064714631177833	22.739134598174868
5	7.297624744179685	3.702375255820315	22.63180468736198
6	6.705258323951894	4.294741676048106	22.736286138420425
7	6.155399906315438	4.844600093684562	22.73751724165138
8	5.64540760451318	5.35459239548682	22.678015204137193
9	5.1727778037666745	5.8272221962333255	22.7683895201492
10	4.735139310000001	6.264860689999999	22.732904122147605
11	4.33024768627229	6.66975231372771	22.730777667271113
12	3.9559797728608257	7.044020227139175	22.801723818612935
13	3.610328386980833	7.389671613019167	22.63505095191573
14	3.291397198172441	7.708602801827559	22.627375053302202
15	2.997395775429687	8.002604224570312	22.785544848141853
16	2.7266348021907447	8.273365197809255	22.77595747749035
17	2.477521455352944	8.522478544647056	22.7058029631555
18	2.248554944520475	8.751445055479525	22.687643377769465
19	2.038322207737026	8.961677792262973	22.701816441723256
20	1.845493760000001	9.15450624	22.763066773233398
21	1.6688196908972177	9.331180309102782	22.773781043057618

(*continued*)

Table 4. (*continued*)

It	Junior	Senior	Ackley
22	1.50712580775145	9.49287419224855	22.647336109599276
23	1.3593099207024493	9.640690079297551	22.682470735962827
24	1.2243382662004736	9.775661733799527	22.80833444732085
25	1.1012420654296875	9.898757934570312	22.65764842443089

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 [22, 23], 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 XGBoost after replacing their kernel with GSK are explained in Table 4.

In Table 5, 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.

Table 5. The result of HPM-STG

Iteration	Initial residuals	New predictions	New residuals
0	1.272424	8.187216	1.145182
1	−6.909718	6.223820	−6.218746
2	−6.913936	6.223398	−6.222542
3	−6.910175	6.223774	−6.219158
4	−6.772750	6.237517	−6.095475
5	−2.514800	6.663311	−2.263320
6	−6.914639	6.223328	−6.223175
7	−5.742731	6.340518	−5.168458
8	4.543536	7.369145	4.089182
9	−6.870089	6.227783	−6.183080

(*continued*)

Table 5. (*continued*)

Iteration	Initial residuals	New predictions	New residuals
10	−6.846299	6.230162	−6.161669
11	−3.359608	6.578831	−3.023647
12	2.267459	9.182251	2.040713
13	−6.912200	6.223572	−6.220980
14	−6.908514	6.223940	−6.217662
15	−6.914647	6.223327	−6.223182
16	−6.497434	6.265048	−5.847690
17	−6.683213	6.246470	−6.014892
18	−5.932537	6.321538	−5.339283
19	−6.914216	6.223370	−6.222794
20	−6.914572	6.223334	−6.223115
21	−6.893094	6.225482	−6.203784
22	−6.914734	6.223318	−6.223261
23	−5.683538	6.346438	−5.115184
24	−6.826928	6.232099	−6.144235
25	1.272424	8.187216	1.145182

In Table 6, 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, as the results are shown in the above table.

Table 6. The result of Evaluation measures

Types of gas	Accuracy	Precision	Recall	F-measurement	F_β	Execution time (second)
Gas #1	0.4779	0.5032	0.7129	0.5900	0.5245	2.4878
Gas #2	0.5227	0.4982	1.5354	0.7523	0.5494	2.5358
Gas #3	1.2226	0.5455	2.5074	0.8961	0.6115	3.0889
Gas #4	0.6607	0.4798	1.4007	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	1.5158	0.7524	0.5513	3.0828

In Table 7, 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 the accuracy appeared and the accuracy was 0.93, 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 4.70 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.

Table 7. The compare between the traditional XGBoost and DXGBoost-GSk

# Iteration	XGBoost		DXGBoost	
	Time	Accuracy	Time	Accuracy
1	2.9409520626068115	0.428063104	4.701775074005127ms	0.9368374562608915
2	2.956578493118286	0.387859209	4.702776193618774	0.9368374562608907
3	2.956578493118286	0.245783248	4.702776193618774	0.9368374562608898
4	2.956578493118286	1.452326905	4.702776193618774	0.9368374562608889
5	2.956578493118286	0.665733854	4.702776193618774	0.9368374562608881
6	2.956578493118286	0.59076485	4.702776193618774	0.9368374562608872
7	2.9658281803131104	0.562495346	4.702776193618774	0.9368374562608863
8	2.966827392578125	0.547653308	4.702776193618774	0.9368374562608854
9	2.9678261280059814	0.538508025	4.702776193618774	0.9368374562608847
10	2.9698259830474854	0.532307752	4.702776193618774	0.9368374562608838
11	2.970825433731079	0.527827222	4.702776193618774	0.9368374562608829
12	2.9728243350982666	0.52443808	4.702776193618774	0.936837456260882
13	2.973823070526123	0.521784852	4.702776193618774	0.9368374562608811
14	2.974822998046875	0.51965132	4.702776193618774	0.9368374562608803
15	2.975822925567627	0.517898412	4.702776193618774	0.9368374562608794
16	2.9768221378326416	0.516432615	4.702776193618774	0.9368374562608786
17	2.9778265953063965	0.515188728	4.702776193618774	0.9368374562608777
18	2.978820562362671	0.514119904	4.702776193618774	0.9368374562608769
19	2.9798214435577393	0.513191617	4.702776193618774	0.936837456260876
20	2.980821371078491	0.512377857	4.702776193618774	0.9368374562608751
21	2.981818675994873	0.511658661	4.702776193618774	0.9368374562608742
22	2.9828171730041504	0.511018452	4.702776193618774	0.9368374562608733
23	2.984816312789917	0.510444894	4.702776193618774	0.9368374562608726
24	2.9868156909942627	0.509928094	4.703782081604004	0.9368374562608717
25	2.9878153800964355	0.509460023	4.705773115158081	0.9368374562608708

As for the traditional method, where the best accuracy was 1.45 The worst accuracy was 0.24, which is ok, but its accuracy is less, it is basically unreliable, and the time it took to implement is 2.98. Although it took less implementation time than the developed method and also the accuracy was less than the proposed method, it is not useful, to be accurate.

Figure 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

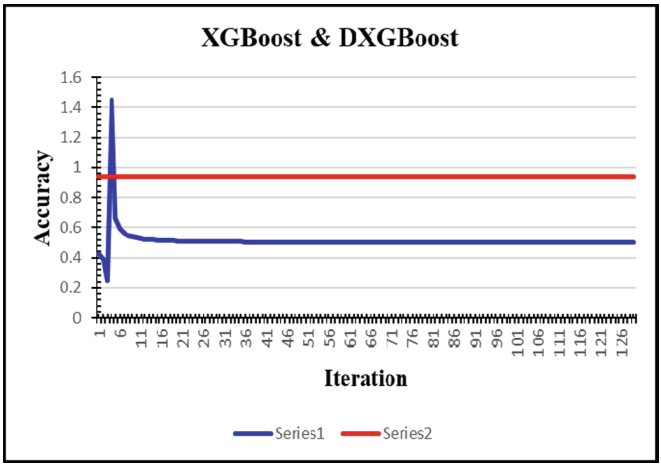

Fig. 2. Compare traditional XGBoost with DXGBoost from aspect accuracy

it becomes a matrix of 129 * 129. After applying the developed method to this matrix, the results shown in the above figure appear.

5 Conclusions

This section presents the most important conclusions reached through applying the HPM-STG into the dataset and focuses on how to avoid the both challenges (programming challenges and application challenges). In addition, we will suggest a set of recommendations for researchers to work on it in the future.

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.

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.

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, we 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 second 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 second 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 fifth 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.

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.

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. We can used the following idea for development this work in the futures

- 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.
- The HPM-STG implementation on CPU as hardware while; we can implement on other hardware such as GPU or FPGA.
- 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.
- It is possible to use another technology for the classification process such as the Deep learning algorithm represented by Long Short-Term Memory (LSTM).

References

1. Abad, A.R.B., et al.: Robust hybrid machine learning algorithms for gas flow rates prediction through wellhead chokes in gas condensate fields. Fuel **308**, 121872 (2022). https://doi.org/10.1016/j.fuel.2021.121872
2. Al-Janabi, S., Mahdi, M.A.: Evaluation prediction techniques to achievement an optimal biomedical analysis. Int. J. Grid Util. Comput. **10**(5), 512–527 (2019).https://doi.org/10.1504/ijguc.2019.102021
3. Alkaim, A.F., Al_Janabi, S.: Multi objectives optimization to gas flaring reduction from oil production. In: International Conference on Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems, pp. 117–139. Springer, Cham (April 2019). https://doi.org/10.1007/978-3-030-23672-4_10

4. 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. Appl. **33**, 14199–14229.https://doi.org/10.1007/s00521-021-06067-7
5. Al-Janabi, S., Alkaim, A.F.: A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft. Comput. **24**(1), 555–569 (2020)https://doi.org/10.1007/s00500-019-03972-x
6. Al-Janabi, S., Alkaim, A.F., Adel, Z.: An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft. Comput. **24**, 10943–10962 (2020)https://doi.org/10.1007/s00500-020-04905-9
7. Al_Janabi, S., Al_Shourbaji, I., Salman, M.A.: Assessing the suitability of soft computing approaches for forest fires prediction. Appl. Comput. Inf. **14**(2): 214–224 (2018). ISSN 2210-8327https://doi.org/10.1016/j.aci.2017.09.006
8. Chung, D.D.: Materials for electromagnetic interference shielding. Mater. Chem. Phys., 123587 (2020)https://doi.org/10.1016/j.matchemphys.2020.123587
9. Cotfas, L.A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D.S., Tajariol, F.: The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. IEEE Access **9**, 33203–33223 (2021).https://doi.org/10.1109/ACCESS.2021.3059821
10. 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., et al.: Predicting thermal expansion pressure buildup in a deepwater oil well with an annulus partially filled with nitrogen. J. Petrol. Sci. Eng. **208**, 109275 (2022)https://doi.org/10.1016/j.petrol.2021.109275
11. Fernandez-Vidal, J., Gonzalez, R., Gasco, J., Llopis, J. (2022). Digitalization and corporate transformation: the case of European oil & gas firms. Technol. Forecast. Soc. Chang. **174**, 121293.https://doi.org/10.1016/j.techfore.2021.121293
12. Foroudi, S., Gharavi, A., Fatemi, M.: 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 (2022). https://doi.org/10.1016/j.fuel.2021.122150
13. Gao, Q., Xu, H., Li, A.: The analysis of commodity demand predication in supply chain network based on particle swarm optimization algorithm. J. Comput. Appl. Math. **400**, 113760 (2022). https://doi.org/10.1016/j.cam.2021.113760
14. Gonzalez, D.J., Francis, C.K., Shaw, G.M., Cullen, M.R., Baiocchi, M., Burke, M.: Upstream oil and gas production and ambient air pollution in California. Sci. Total Environ. **806**, 150298 (2022). https://doi.org/10.1016/j.scitotenv.2021.150298
15. Al-Janabi, S., Alkaim, A.: A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis. Egypt. Inf. J. (2022).https://doi.org/10.1016/j.eij.2022.01.004
16. Al-Janabi, S.: 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 (2021)https://doi.org/10.1109/ICDABI53623.2021.9655916
17. Gupta, N., Nigam, S.: Crude oil price prediction using artificial neural network. Procedia Comput. Sci. **170**, 642–647 (2020). https://doi.org/10.1016/j.procs.2020.03.136
18. Hao, P., Di, L., Guo, L.: Estimation of crop evapotranspiration from MODIS data by combining random forest and trapezoidal models. Agric. Water Manag. **259**, 107249 (2022).https://doi.org/10.1016/j.agwat.2021.107249
19. Al-Janabi, S., Rawat, S., Patel, A., Al-Shourbaji, I.: Design and evaluation of a hybrid system for detection and prediction of faults in electrical transformers. Int. J. Electr. Power Energy Syst. **67**, 324–335 (2015)https://doi.org/10.1016/j.ijepes.2014.12.005
20. Houssein, E.H., Gad, A.G., Hussain, K., Suganthan, P.N.: Major advances in particle swarm optimization: theory, analysis, and application. Swarm Evol. Comput. **63**, 100868 (2021). https://doi.org/10.1016/j.swevo.2021.100868

21. Johny, J., Amos, S., Prabhu, R.: Optical fibre-based sensors for oil and gas applications. Sensors **21**(18), 6047 (2021). https://doi.org/10.3390/s21186047
22. Mahdi, M. A., & Al-Janabi, S.: 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. BDNT 2019. Lecture Notes in Networks and Systems, pp. 320–339. Springer, Cham (April 2019). https://doi.org/10.1007/978-3-030-23672-4_23
23. Kadhuim, Z.A., Al-Janabi, S.: Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis. Results Eng. **17** (2023). https://doi.org/10.1016/j.rineng.2022.100847
24. Mohammadpoor, M., Torabi, F.: Big Data analytics in oil and gas industry: an emerging trend. Petroleum **6**(4), 321–328 (2020). https://doi.org/10.1016/j.petlm.2018.11.001
25. Mohammed, G.S., Al-Janabi, S.: An innovative synthesis of optmization techniques (FDIRE-GSK) for generation electrical renewable energy from natural resources. Results Eng. **16** (2022). https://doi.org/10.1016/j.rineng.2022.100637
26. Ali, S.H.: A novel tool (FP-KC) for handle the three main dimensions reduction and association rule mining. In: IEEE,6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Sousse, pp. 951–961 (2012).https://doi.org/10.1007/978-90-313-8424-2_10