



- (51) **International Patent Classification:**
G06N 20/00 (2019.01) *G06N 7/01* (2023.01)
- (21) **International Application Number:**
PCT/US2024/037535
- (22) **International Filing Date:**
11 July 2024 (11.07.2024)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
18/352,873 14 July 2023 (14.07.2023) US
- (71) **Applicant (for US only):** **SCHLUMBERGER TECHNOLOGY CORPORATION** [US/US]; 300 Schlumberger Drive, Sugar Land, Texas 77478 (US).
- (71) **Applicant (for CA only):** **SCHLUMBERGER CANADA LIMITED** [CA/CA]; 125 – 9 Avenue SE, Calgary, Alberta T2G 0P6 (CA).
- (71) **Applicant (for FR only):** **SERVICES PETROLIERS SCHLUMBERGER** [FR/FR]; 42 rue Saint Dominique, 75007 Paris (FR).

- (71) **Applicant (for all designated States except CA, FR, US):** **SCHLUMBERGER TECHNOLOGY B.V.** [NL/NL]; Parkstraat 83, 2514 JG The Hague (NL).
- (72) **Inventors:** **MILES, Jeffrey**; One Hampshire Street, Cambridge, Massachusetts 02139 (US). **CRADDOCK, Paul Ryan**; One Hampshire Street, Cambridge, Massachusetts 02139 (US). **GO, Jinwoo**; 855 Piedmont Ave, Atlanta, Georgia 30308 (US).
- (74) **Agent:** **FLYNN, Michael L.** et al.; Schlumberger, 10001 Richmond Avenue, Room 4720, Houston, Texas 77042 (US).
- (81) **Designated States (unless otherwise indicated, for every kind of national protection available):** AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) **Title:** METHODS FOR CONFIDENCE ASSESSMENT WITH FEATURE IMPORTANCE IN DATA DRIVEN ALGORITHMS

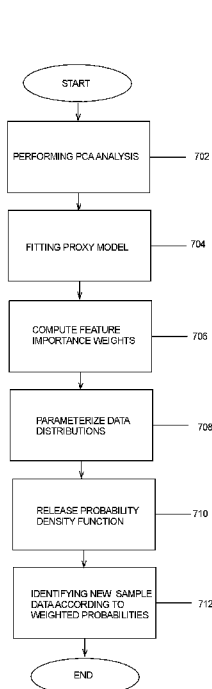


FIG. 7

(57) **Abstract:** Embodiments presented provide for a method for establishing a confidence assessment for data. Data may be segregated by features importance during the confidence assessment, allowing evaluators the ability to determine the quality of data being processed by the method. The method may comprise performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is an integer. The method may also comprise computing feature importance weights for each of the k principal component inputs. The method may also comprise identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

METHODS FOR CONFIDENCE ASSESSMENT WITH FEATURE IMPORTANCE IN DATA DRIVEN ALGORITHMS

CROSS-REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of U.S. Non-Provisional Application No. 18/352,873, entitled "METHODS FOR CONFIDENCE ASSESSMENT WITH FEATURE IMPORTANCE IN DATA DRIVEN ALGORITHMS" filed July 14, 2023, the disclosure of which is hereby incorporated herein by reference.

FIELD OF THE DISCLOSURE

[002] Aspects of the disclosure relate to verifying data quality and artificial intelligence modeling. More specifically, aspects of the disclosure relate to methods for confidence assessment for data using feature importance in data driven algorithms.

BACKGROUND

[003] Many methods of machine learning (ML) have been developed and can be broadly grouped under the umbrella of data-driven algorithms or statistical learning. A non-exhaustive list of ML methods includes Decision Tree, Random Forest, Support Vector Machine, K-Means Clustering, Logit Regression, Artificial Neural Networks, Convolutional Neural Networks, and Naive Bayes. The method can be based on supervised, unsupervised, semi-supervised, and reinforcement learning. These methods have been applied to a large number of classification (e.g., categorical dependent variables) and/or regression (e.g., continuous dependent variables) problems.

[004] However, many ML methods lack interpretability and can only make predictions without rigorous estimates of uncertainty and confidence in the predicted answers. Moreover, ML algorithms tend to perform poorly when extrapolating away from the domain or range of the data samples on which the algorithm was optimized (i.e., outside the range of the algorithm's training data).

[005] Several methods exist to assess feature importance—i.e., the sensitivity of a model output value to various input values. Feature importance is useful for identifying which model inputs (features) provide predictive strength in the model and those which carry no such predictive information. Feature importance is; therefore, useful for model interpretability. Methods for quantifying feature importance include Shapley additive explanations (SHAP), local interpretable model-agnostic explanations (LIME), accumulated local effects (ALE), mean decrease in impurity (MDI or Gini Importance), and mean decrease in accuracy (MDA or Permutation Importance). Generally speaking, these methods calculate a perturbation of a model output value from a perturbation, permutation, or elimination of each model input, and then rank or score the relative importance of each model input. A common approach of these methods is that features are interrogated independently, such that interpretation of feature importance scores or rankings becomes challenging for highly correlated features.

[006] There exist a variety of methods for outlier detection. Outlier detection is beneficial for estimating the confidence or reliability of a value output from a model prediction using feature data not included in the original model optimization. The standard score (also known as Z-score) and the box-and-whisker plot are two, widespread, statistics tools for outlier detection. The Z-score quantifies the number of standard deviations a sample datum is away from the mean of the dataset. Samples with Z-score values above a defined threshold (e.g., 2 standard deviations) are considered outliers. The score is calculated for univariate data (i.e., a single variable or feature) and works best when applied to data that is normally distributed (Gaussian) or nearly so. The box-and-whisker plot computes data percentiles from the dataset and determines outliers based on their individual percentile value compared to one or more thresholds, such as 25th percentile (1st quartile), 75th percentile (3rd quartile), inter-quartile range (IQR), and $IQR \times 1.5$. Percentile metrics are also calculated for single variables, but can be applied to non-

Gaussian distributions. Approaches for computing outliers in univariate data can be extended to multivariate data, with each variable having their own distribution. For example, in the case of Z-score, standard deviations can be represented as ellipses contoured around (uncorrelated) bivariate data. Other disclosed methods directed toward multivariate data distribution analysis include Minimum Covariance Determinant (MCD).

[007] There is a need to provide an apparatus and methods that are easier to operate than conventional apparatus and methods.

[008] There is a further need to provide apparatus and methods that do not have the drawbacks discussed above.

[009] There is a still further need to reduce economic costs associated with operations and apparatus described above with conventional tools.

SUMMARY

[010] So that the manner in which the above recited features of the present disclosure can be understood in detail, a more particular description of the disclosure, briefly summarized below, may be had by reference to embodiments, some of which are illustrated in the drawings. It is to be noted that the drawings illustrate only typical embodiments of this disclosure and are therefore not to be considered limiting of its scope, for the disclosure may admit to other equally effective embodiments without specific recitation. Accordingly, the following summary provides just a few aspects of the description and should not be used to limit the described embodiments to a single concept.

[011] In one example embodiment, a method is disclosed. The method may comprise performing a principal component analysis on k model original features to obtain k

principal components representing uncorrelated input data distributions, wherein k is an integer. The method may also comprise fitting a proxy model using k principal component inputs to an output property. The method may also comprise computing feature importance weights, for each of the k principal component inputs. The method may also comprise parameterizing k principal component input data distributions. The method may also comprise relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight. The method may also comprise identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs. The method may also comprise identifying sample outliers using visual cues. The method may also comprise displaying the sample outliers.

[012] In another example embodiment, a method is disclosed. The method may include steps of performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions. The method may also comprise fitting a proxy model using k principal component inputs to an output property. The method may also comprise computing feature importance weights for the each k principal component input. The method may also comprise parameterizing k principal component input data distributions as probability density functions. The method may also comprise relaxing the probability density functions of each k principal component input data distributions according to a normalized feature importance weight. The method may also comprise identifying any new sample data out-of-distribution according to weighted probabilities compared to assigned cut-offs. The method may also comprise identifying sample outliers using visual cues. The method may also comprise displaying the sample outliers.

[013] In another example embodiment, a method is disclosed. The method may comprise performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is an integer greater than 1. The method may also comprise fitting a proxy model using the k principal component inputs to an output property representing a geological feature. The method may also comprise computing feature importance weights for each of the k principal component inputs. The method may also comprise parameterizing k principal component input data distributions. The method may also comprise relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight. The method may also comprise identifying any new sample data in distribution to weighted probabilities compared to assigned cut-offs. The method may also comprise identifying sample outliers using visual cues. The method may also comprise saving the sample outliers in a non-volatile memory.

[014] In another example embodiment, an article of manufacture is disclosed, wherein the article of manufacture is configured to be performed on a computing device, wherein the performance is a method configured to include performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is a integer greater than 1. The method performed by the article of manufacture may include fitting a proxy model using the k principal component inputs to an output property representing a geological feature. The method performed by the article of manufacture may include computing feature importance weights for each of the k principal component inputs. The method performed by the article of manufacture may include parameterizing k principal component input data distributions. The method performed by the article of manufacture may include relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight. The method

performed by the article of manufacture may include identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs. The method performed by the article of manufacture may include identifying sample outliers using visual cues and saving the sample outliers in a non-volatile memory.

BRIEF DESCRIPTION OF THE DRAWINGS

[015] So that the manner in which the above recited features of the present disclosure can be understood in detail, a more particular description of the disclosure, briefly summarized above, may be had by reference to embodiments, some of which are illustrated in the drawings. It is to be noted; however, that the appended drawings illustrate only typical embodiments of this disclosure and are therefore not be considered limiting of its scope, for the disclosure may admit to other equally effective embodiments.

[016] FIG. 1 is a curve based representation of a data driven model featuring inputs and outputs.

[017] FIGS. 2A and 2B show an example dataset in which original data features are correlated and subsequently uncorrelated when rotated using principal components analysis onto their principal components.

[018] FIG. 3 is an example of feature sensitivity.

[019] FIG. 4 is an example of parameterizing a training data distribution with a Gaussian pdf to provide outlier detection that may be applied to new data samples.

[020] FIG. 5 is a visual example of a novel confidence assessment method for a hypothetical model using two input features, where outlier detection is performed using a

modified Gaussian parameterization that includes the sensitivity of the model to different input features.

[021] FIG. 6 is an illustrative use of a weighted out-of-distribution indicator.

[022] FIG. 7 is an example flow chart for one combination of outlier detection and feature importance for confidence assessment in a data-driven model prediction.

[023] To facilitate understanding, identical reference numerals have been used, where possible, to designate identical elements that are common to the figures (“FIGS”). It is contemplated that elements disclosed in one embodiment may be beneficially utilized on other embodiments without specific recitation.

DETAILED DESCRIPTION

[024] In the following, reference is made to embodiments of the disclosure. It should be understood; however, that the disclosure is not limited to specific described embodiments. Instead, any combination of the following features and elements, whether related to different embodiments or not, is contemplated to implement and practice the disclosure. Furthermore, although embodiments of the disclosure may achieve advantages over other possible solutions and/or over the prior art, whether or not a particular advantage is achieved by a given embodiment is not limiting of the disclosure. Thus, the following aspects, features, embodiments and advantages are merely illustrative and are not considered elements or limitations of the claims except where explicitly recited in a claim. Likewise, reference to “the disclosure” shall not be construed as a generalization of inventive subject matter disclosed herein and should not be considered to be an element or limitation of the claims except where explicitly recited in a claim.

[025] Although the terms first, second, third, etc., may be used herein to describe various elements, components, regions, layers and/or sections, these elements, components, regions, layers and/or sections should not be limited by these terms. These terms may be only used to distinguish one element, components, region, layer or section from another region, layer or section. Terms such as “first”, “second” and other numerical terms, when used herein, do not imply a sequence or order unless clearly indicated by the context. Thus, a first element, component, region, layer or section discussed herein could be termed a second element, component, region, layer or section without departing from the teachings of the example embodiments.

[026] When an element or layer is referred to as being “on,” “engaged to,” “connected to,” or “coupled to” another element or layer, it may be directly on, engaged, connected, coupled to the other element or layer, or interleaving elements or layers may be present. In contrast, when an element is referred to as being “directly on,” “directly engaged to,” “directly connected to,” or “directly coupled to” another element or layer, there may be no interleaving elements or layers present. Other words used to describe the relationship between elements should be interpreted in a like fashion. As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed terms.

[027] Some embodiments will now be described with reference to the figures. Like elements in the various figures will be referenced with like numbers for consistency. In the following description, numerous details are set forth to provide an understanding of various embodiments and/or features. It will be understood, however, by those skilled in the art, that some embodiments may be practiced without many of these details, and that numerous variations or modifications from the described embodiments are possible. As used herein, the terms “above” and “below”, “up” and “down”, “upper” and “lower”, “upwardly” and “downwardly”, and other like terms indicating relative positions above or

below a given point are used in this description to more clearly describe certain embodiments.

[028] The methods and systems described herein disclose a novel method of making a confidence estimate on the prediction from a data-driven model using inputs from any new sample data, based on the model's sensitivity to the different inputs (features) and the similarity in feature space between any new sample data and the original ensemble data used to train the model. Embodiments of the disclosure may be used in a variety of industries. One such non-limiting embodiment may be used in hydrocarbon recovery operations where data may be non-homogeneous but the need for accuracy of prediction is great. Other data intensive applications may use embodiments, such as computational fluid dynamics.

[029] FIG. 1 shows for a set of earth formation samples, the predicted values for a property of interest ("Target") obtained by fitting a set of measured input properties ("Features 1–3") to a data-driven machine learning model. The ground-truth values for the same property of interest are also shown. The model inputs, true target value, and predicted target value are represented as continuous curves, such as is commonly shown for wellbore logging plots. It is apparent that this model can make an accurate prediction of the true target value only for a subset of the samples. An investigation of the samples for which the model prediction is erroneous would reveal that those samples contain a set of input features whose distribution is different to that of the same features in the original ensemble of samples used to train the model. In other words, predictions have been made on outlier samples, and the predictions may fail for those samples. It is desired to make predictions of accuracy based upon any set of data given and to allow researchers the ability to easily determine if the data provided will be accurate.

[030] Several techniques of outlier detection are possible, such as by calculating data ranges as percentiles and standard deviations. In one embodiment, outliers are identified using a multi-variative Gaussian parameterization. In one example embodiment, this parameterization is done assuming uncorrelated model features (i.e., zero covariance between the model inputs). It is known; however, that many properties of earth formations are correlated to a greater or lesser extent and are not truly independent. This is illustrated in FIG. 2, where the same set of three features shown in FIG. 1 are now cross plotted on cartesian axes. The full covariance could be included in a more complicated Gaussian parameterization, but this is not preferred because of complexity, and it would not easily enable the use of feature sensitivities, described below. Alternatively, the original model features can be rotated onto uncorrelated bases using principal components analysis (PCA). As defined herein, PCA projects the original feature basis (which are correlated in the training data) onto the basis of principal components in which the data are statistically uncorrelated. Therefore, one embodiment of an outlier detection algorithm is the parameterization of a dataset using a multi-variate Gaussian distribution, where the component features of the data samples have been transformed by PCA such that they are statistically uncorrelated. Note that PCA is performed on the original ensemble of training samples (i.e., training dataset) and the extracted PCA parameters are then applied to any new samples in a testing dataset. FIG. 2 illustrates an example of PCA. The number of principal components may be the same as the number of original features, such that there is no dimensionality reduction.

[031] The probability of any new sample falling within the distribution space of the original ensemble data can then be represented by the joint probability density of the normal distribution in the space of k principal components:

$$P = \prod_{i=1}^k \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2}$$

[032] This probability density function (PDF) can be useful as an outlier indicator (e.g., standard score or Z-score), but the function does not fully capture the relative importance of the input features in the model. Depending on the relative sensitivity of the model to different features, the distance between the training data and a new test sample along one component may be much more important than the same distance along another component.

[033] In one embodiment, an appropriate vector of weights **w** can be used to modify the above outlier detection scheme. Conceptually, the weights describe how much the confidence in model predictions should be affected by deviations from the mean of the training data along each axis. The axes in this embodiment may represent principal component axes or the basis of the original input features. Several methods of sensitivity analysis are available. As an example, FIG. 3 shows relative feature sensitivities derived from a SHAP (Shapley additive explanations) analysis. The theory underlying SHAP is described elsewhere. Some comments regarding the example plot are worth noting. The model sensitivity is computed for each input feature separately, which are labeled “PC 1 – 3” in this example. Model sensitivity may be computed for any inputs to the model, including (as in this example) inputs that represent principal-component combinations of data attributes.

[034] Each data point on the plot represents a sample from an ensemble of data, in this case the same original ensemble of training samples (i.e., the training dataset). The impact on the model output is indicated by the position (i.e. the SHAP value) of the samples on the horizontal coordinate. The SHAP value for each sample represents the

difference in the model output induced by substituting the true value for a model input in place of the mean value for that input feature, where the mean is computed from the full ensemble. The further the deviation from zero, the greater the model sensitivity to the feature. Thus, in one or more embodiments, a feature importance can be derived from the spread of the SHAP values for each feature as calculated from the range, interquartile range, standard deviation, or any other appropriate metric or statistic. In one embodiment, the feature importance weights $\mathbf{w}$ are calculated as the standard deviation of the SHAP value distributions for each feature and then normalized with respect to the most important feature with unit weight ($\mathbf{w}_i = 1$).

[035] In one embodiment, the joint PDF above may then be modified for each model input (e.g., k principal component features from the training dataset) according to the relative feature importance weights $\mathbf{w}$ for that set of inputs. The modification may be justified by the fact that a certain amount of deviation along one axis of input features can be much more or less important for model confidence than the same deviation along another axis. Therefore, the feature importance weights $\mathbf{w}$ can be applied to the standard score of any data point, i.e.: $\frac{(z_i - \mu_i)}{\sigma_i}$, being its number of standard deviations from the mean along each principal component. In practice, this only requires a reweighting of the standard deviations that we fitted to the ensemble training data along each PC, e.g.: $\sigma_i \rightarrow \sigma_i / w_i$.

[036] The weighted PDF in this case becomes:

$$P = \prod_{i=1}^k \frac{w_i}{\sigma_i \sqrt{2\pi}} e^{-\frac{w_i^2}{2} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2}$$

[037] This is the coherent way of inserting the weights to penalize deviations along high-weight axes and to reduce the impact of deviations along less-important axes. The

weights serve to modify the effective distance in standard score space of a new sample with respect to the training dataset mean. FIG. 4 provides a conceptual illustration for a model using principal component features as inputs and for which weighting is applied to principal components 1 and 3, where principal component 1 has the most importance ($w_1 = 1$) and principal component 3 has lower importance ($w_3 \ll 1$). In the upper plot, the standard deviations of the sample distributions along each principal component are unweighted. In the lower plot, those standard deviations are relaxed for the principal component that has lower importance. In effect, some samples that would be classified as outliers (outside bounds of the ellipse) using unweighted criteria are now classified as in-distribution (inside bounds of the ellipse) because the distance metric is relaxed along the axis with low feature importance.

[038] FIG. 5 shows an analogous visual example of the novel confidence assessment method for a hypothetical model using two input features. The PDF of these hypothetical training data (circular points) are parameterized as Gaussian distributions (black contours, representing integer standard deviations of the distribution along each axis), the widths (i.e., standard deviations) of which are equal along the two feature dimensions. The Gaussian parameterization for outlier detection is modified (gray contours) using the above equation to account for the relative importance of the two input features for the algorithm predictions. This example illustrates a case where Feature 1 is less important than Feature 2 (i.e., Feature 1 has less impact on the model predictions), and so the confidence metric is penalized less (or “relaxed”) for distance along Feature 1 axis than Feature 2 axis. For the purposes of exemplifying unweighted and weighted outlier detection, a two-standard deviation (2σ) is used here as the cut-off between in-distribution and out-of-distribution. A new sample (diamond) would be classified as an outlier for an unweighted Gaussian distribution, because it falls outside of the 2σ value for Feature 1. However, noting here that Feature 1 has less importance in the model

prediction, the weighted Gaussian distribution is relaxed along Feature 1 axis, and the new sample (diamond) is classified as in-distribution. This novel weighting results in a more meaningful assessment of confidence in predictions for the new sample with respect to the training dataset.

[039] In defining cut-offs for whether any new sample is in-distribution or out-of-distribution, it is convenient to use the log-probability, taking the natural logarithm of the weighted PDF:

$$\log P_w = -\frac{k}{2} \log 2\pi + \sum_{i=1}^k \left[\log \left(\frac{w_i}{\sigma_i} \right) - \frac{w_i^2 (z_i - \mu_i)^2}{2\sigma_i^2} \right]$$

[040] If we define the cut-off (boundary) using a threshold of t weighted-standard deviations along each principal component axis, then the threshold on $\log P_w$ will be:

$$\log P_w \big|_{\text{cutoff}} = -\frac{k}{2} (t^2 + \log 2\pi) + \sum_{i=1}^k \log \left(\frac{w_i}{\sigma_i} \right)$$

[041] Those samples that are outliers have $\log P_w$ values less than the cutoff value.

[042] FIG. 6 illustrates the usefulness of the novel confidence indicator, here termed the weighted-out-of-distribution indicator (WOOD). FIG. 6 shows the same model prediction as in FIG. 1. To this prediction is now added a WOOD indicator, shown as a continuous, black curve for all samples. Overlaying the WOOD is a threshold or cut-off for discriminating in-distribution versus out-of-distribution samples. In this example, samples with WOOD values that fall to the right of the threshold are deemed to be outliers. An optional flag is shown in the right-most track which indicates whether confidence can be placed on the model prediction. The flag is colored for samples for which the prediction is not confident because those samples are identified as outliers. The flag for lacking confidence in the model prediction corresponds satisfactorily to those samples for which

the model prediction is shown to be in error. In the illustrated example, the flag is activated using only outlier detection as a criterion. However, extensions to this flag can be incorporated—for example, a flag that considers whether the estimated uncertainty about a point-estimate prediction exceeds a defined threshold.

[043] FIG. 7 discloses a visual workflow (flow chart) for one combination of outlier detection and feature importance for confidence assessment in a data-driven model. The method **700** provides several steps for completion. The method **700** illustrated provides only one example embodiment and should not be considered limiting. The method **700** starts at **702** with performing a principal component analysis (PCA) on k model original features (inputs) to obtain k principal components representing uncorrelated input data distributions. At **704**, the method continues fitting a proxy model using k principal component inputs to an output property (target) using linear regression. At **706**, the method continues at computing feature importance weights, w , for each k principal component inputs in the proxy linear regression model. At **708**, the method continues with parameterizing the k principal component input data distributions as Gaussian probability density functions (PDF). At **710** the method continues with relaxing the PDF of each k principal component input data distributions according to their normalized feature importance weight, w . At **712**, the method continues with identifying any new sample data in-distribution or out-of-distribution according to their weighted probabilities compared to assigned cut-offs. Sample outliers are identified using visual cues like a confidence red flag.

[044] In one example embodiment, a method is disclosed. The method may comprise performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is an integer. The method may also comprise fitting a proxy model using k principal component

inputs to an output property. The method may also comprise computing feature importance weights, for each of the k principal component inputs. The method may also comprise parameterizing k principal component input data distributions. The method may also comprise relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight. The method may also comprise identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs. The method may also comprise identifying sample outliers using visual cues. The method may also comprise displaying the sample outliers.

[045] In another example embodiment, the method may be performed wherein the displaying the sample outliers includes visually representing the sample outliers.

[046] In another example embodiment, the method may be performed wherein the displaying the sample outliers includes printing the sample outliers.

[047] In another example embodiment, the method may be performed wherein the fitting a proxy model using k principal component inputs to an output property is using linear regression.

[048] In another example embodiment, the method may be performed wherein the parameterized k principal component input data distributions is performed as Gaussian probability density functions.

[049] In another example embodiment, the method may be performed wherein the identifying sample outliers using visual cues includes a confidence red flag.

[050] In another example embodiment, a method is disclosed. The method may include steps of performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions. The method may also comprise fitting a proxy model using k principal component inputs to an output property. The method may also comprise computing feature importance weights for the each k principal component input. The method may also comprise parameterizing k principal component input data distributions as probability density functions. The method may also comprise relaxing the probability density functions of each k principal component input data distributions according to a normalized feature importance weight. The method may also comprise identifying any new sample data out-of-distribution according to weighted probabilities compared to assigned cut-offs. The method may also comprise identifying sample outliers using visual cues. The method may also comprise displaying the sample outliers.

[051] In another example embodiment, the method may be performed wherein the displaying the sample outliers includes visually representing the sample outliers.

[052] In another example embodiment, the method may be performed wherein the displaying the sample outliers includes printing the sample outliers.

[053] In another example embodiment, the method may be performed wherein the fitting a proxy model using k principal component inputs to an output property is using linear regression.

[054] In another example embodiment, the method may be performed wherein the parameterized k principal component input data distributions is performed as Gaussian probability density functions.

[055] In another example embodiment, the method may be performed wherein the identifying sample outliers using visual cues includes a confidence red flag.

[056] In another example, a method is disclosed. The method may comprise performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is an integer greater than 1. The method may also comprise fitting a proxy model using the k principal component inputs to an output property representing a geological feature. The method may also comprise computing feature importance weights, for each of the k principal component inputs. The method may also comprise parameterizing k principal component input data distributions. The method may also comprise relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight. The method may also comprise identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs. The method may also comprise identifying sample outliers using visual cues. The method may also comprise saving the sample outliers in a non-volatile memory.

[057] In another example embodiment, the method may be performed wherein the parameterized k principal component input data distributions is performed as Gaussian probability density functions.

[058] In another example embodiment, the method may be performed wherein the fitting the proxy model using k principal component inputs to an output property uses linear regression.

[059] In another example embodiment, the method may be performed wherein the method is configured to be performed on one a computer, a laptop and a server.

[060] In another example embodiment, an article of manufacture is disclosed, wherein the article of manufacture is configured to be performed on a computing device, wherein the performance is a method configured to include performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is a integer greater than 1. The method performed by the article of manufacture may include fitting a proxy model using the k principal component inputs to an output property representing a geological feature. The method performed by the article of manufacture may include computing feature importance weights, for each of the k principal component inputs. The method performed by the article of manufacture may include parameterizing k principal component input data distributions. The method performed by the article of manufacture may include relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight. The method performed by the article of manufacture may include identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs. The method performed by the article of manufacture may include identifying sample outliers using visual cues and saving the sample outliers in a non-volatile memory.

[061] In the following description, description is provided related to measurements obtained during wireline operations generally performed, as described above. As will be understood, various changes and alterations may be accomplished during the attainment of the desired measurements and as such, methods described should not be considered limiting.

[062] The foregoing description of the embodiments has been provided for purposes of illustration and description. It is not intended to be exhaustive or to limit the disclosure. Individual elements or features of a particular embodiment are generally not limited to that particular embodiment, but, where applicable, are interchangeable and can be used in a selected embodiment, even if not specifically shown or described. The same may be varied in many ways. Such variations are not to be regarded as a departure from the disclosure, and all such modifications are intended to be included within the scope of the disclosure.

[063] While embodiments have been described herein, those skilled in the art, having benefit of this disclosure, will appreciate that other embodiments are envisioned that do not depart from the inventive scope. Accordingly, the scope of the present claims or any subsequent claims shall not be unduly limited by the description of the embodiments described herein.

CLAIMS

What is claimed is:

1. A method, comprising:

performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is an integer;

fitting a proxy model using k principal component inputs to an output property;

computing feature importance weights, for each of the k principal component inputs;

parameterize k principal component input data distributions;

relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight;

identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs;

identifying sample outliers using visual cues; and

displaying the sample outliers.

2. The method according to claim 1, wherein the displaying the sample outliers includes visually representing the sample outliers.

3. The method according to claim 1, wherein the displaying the sample outliers includes printing the sample outliers.

4. The method according to claim 1, wherein the fitting a proxy model using k principal component inputs to an output property is using linear regression.

5. The method according to claim 1, wherein the parameterized k principal component input data distributions is performed as Gaussian probability density functions.

6. The method according to claim 1, wherein the identifying sample outliers using visual cues includes a confidence red flag.

7. A method, comprising:

- performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions;

- fitting a proxy model using k principal component inputs to an output property;

- computing feature importance weights for the each k principal component input;

- parameterize k principal component input data distributions as probability density functions;

- relaxing the probability density functions of each k principal component input data distributions according to a normalized feature importance weight;

- identifying any new sample data out-of-distribution according to weighted probabilities compared to assigned cut-offs;

- identifying sample outliers using visual cues; and

- displaying the sample outliers.

8. The method according to claim 7, wherein the displaying the sample outliers includes visually representing the sample outliers.

9. The method according to claim 7, wherein the displaying the sample outliers includes printing the sample outliers.

10. The method according to claim 7, wherein the fitting a proxy model using k principal

component inputs to an output property is using linear regression.

11. The method according to claim 7, wherein the parameterized k principal component input data distributions is performed as Gaussian probability density functions.

12. The method according to claim 7, wherein the identifying sample outliers using visual cues includes a confidence red flag.

13. A method, comprising:

- performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is an integer greater than 1;

- fitting a proxy model using the k principal component inputs to an output property representing a geological feature;

- computing feature importance weights, for each of the k principal component inputs;

- parameterizing k principal component input data distributions;

- relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight;

- identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs;

- identifying sample outliers using visual cues; and

- saving the sample outliers in a non-volatile memory.

14. The method according to claim 13, wherein the parameterized k principal component input data distributions is performed as Gaussian probability density functions.

15. The method according to claim 13, wherein the fitting the proxy model using k principal component inputs to an output property uses linear regression.

16. The method according to claim 13, wherein the method is configured to be performed on one of a computer, a laptop and a server.

17. An article of manufacture configured to be performed on a computing device, wherein the performance is a method configured to include:

- performing a principal component analysis on k model original features to obtain k principal components representing uncorrelated input data distributions, wherein k is an integer greater than 1;

- fitting a proxy model using the k principal component inputs to an output property representing a geological feature;

- computing feature importance weights, for each of the k principal component inputs;

- parameterizing k principal component input data distributions;

- relaxing the input data distributions of each of the k principal component input data distributions according to a normalized feature importance weight;

- identifying any new sample data in-distribution to weighted probabilities compared to assigned cut-offs;

- identifying sample outliers using visual cues; and

- saving the sample outliers in a non-volatile memory.

1/7

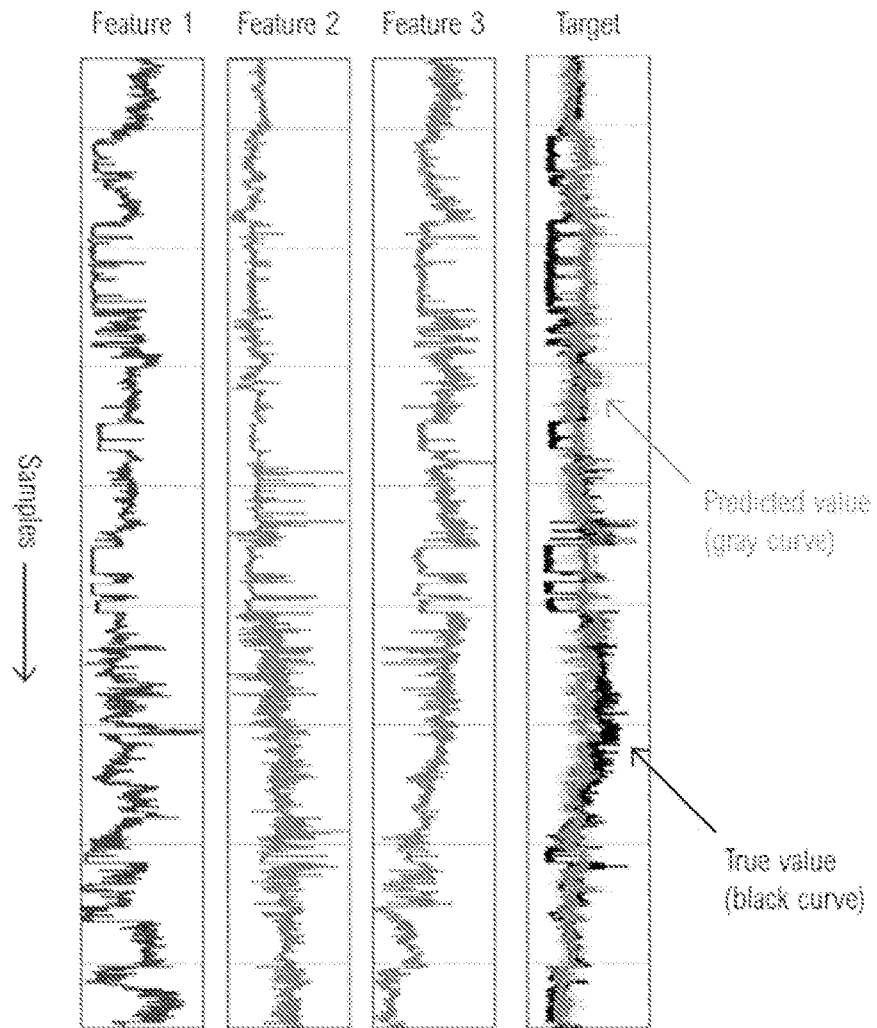


FIG. 1

2/7

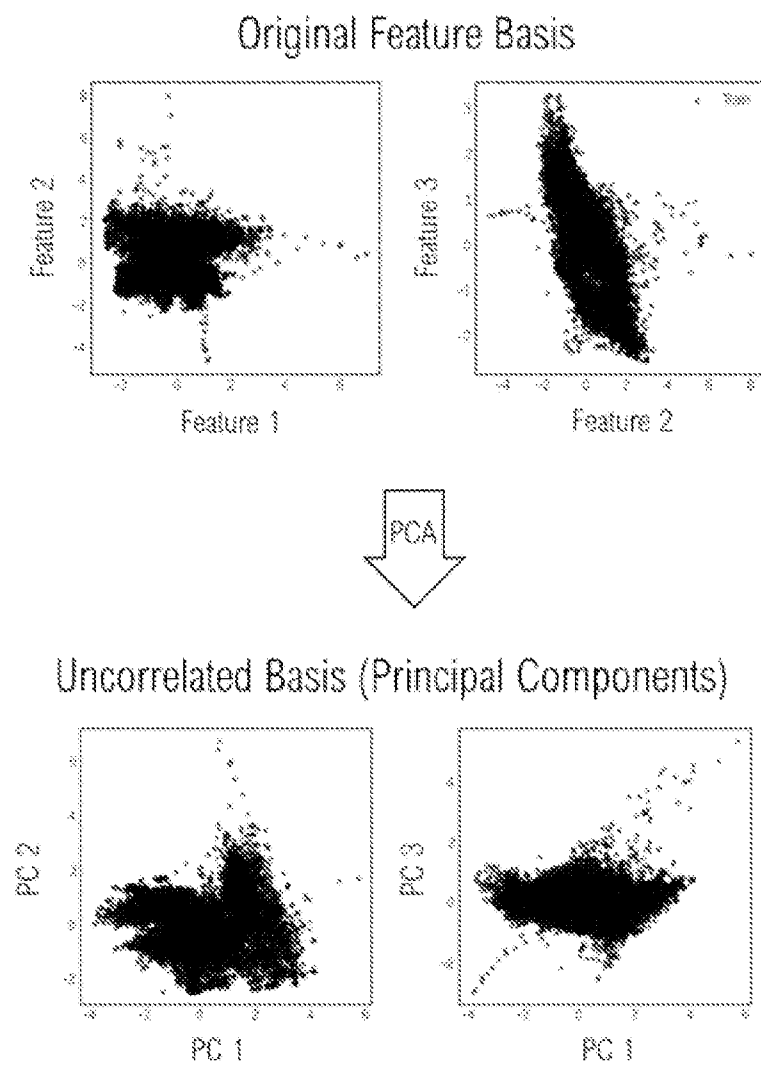


FIG. 2

3/7

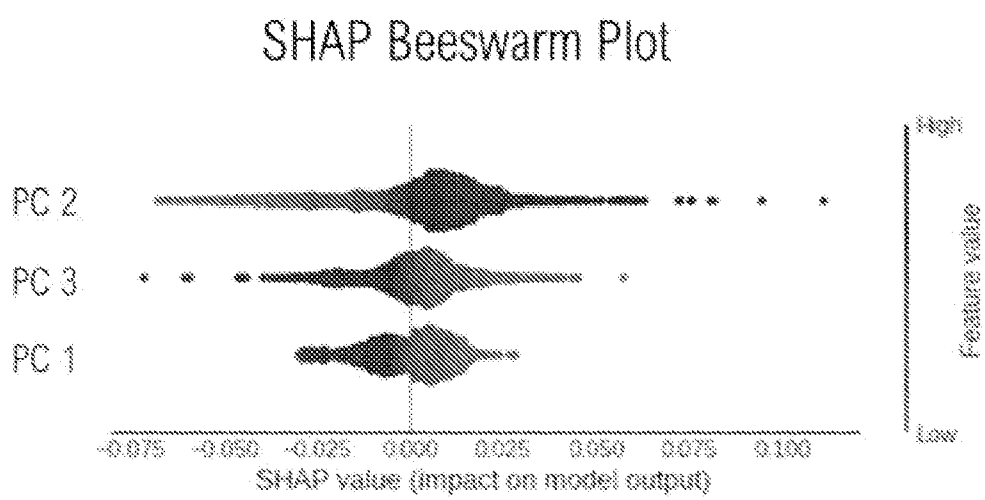


FIG. 3

4/7

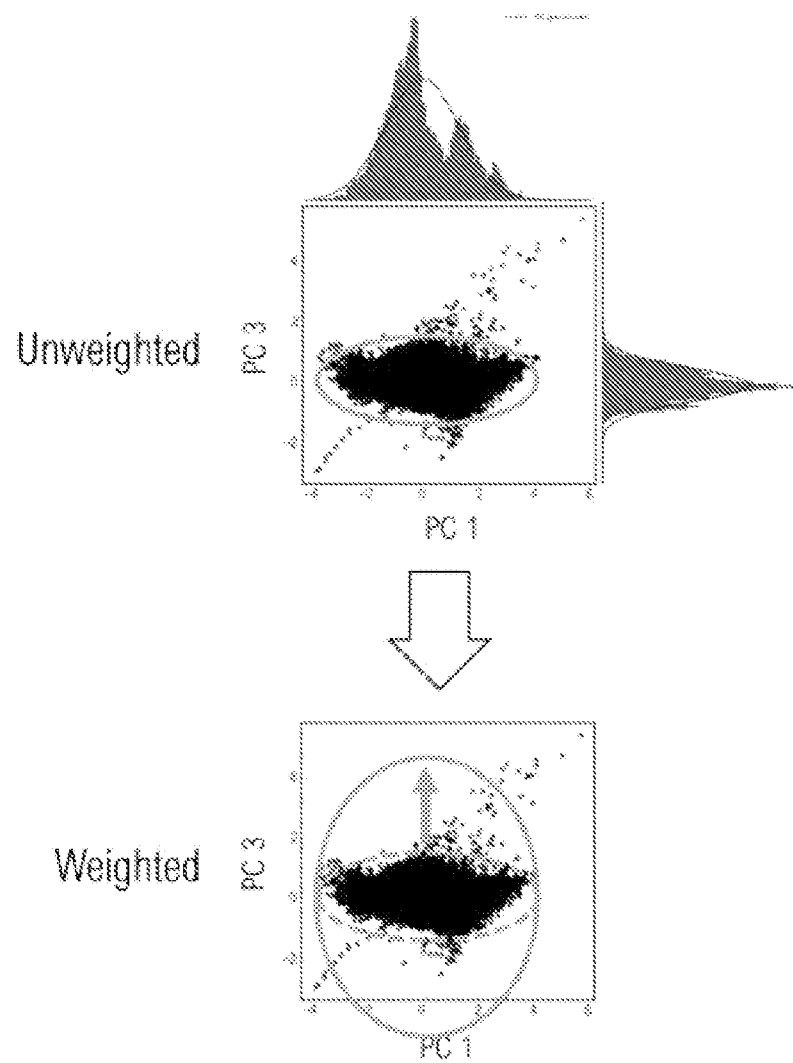


FIG. 4

5/7

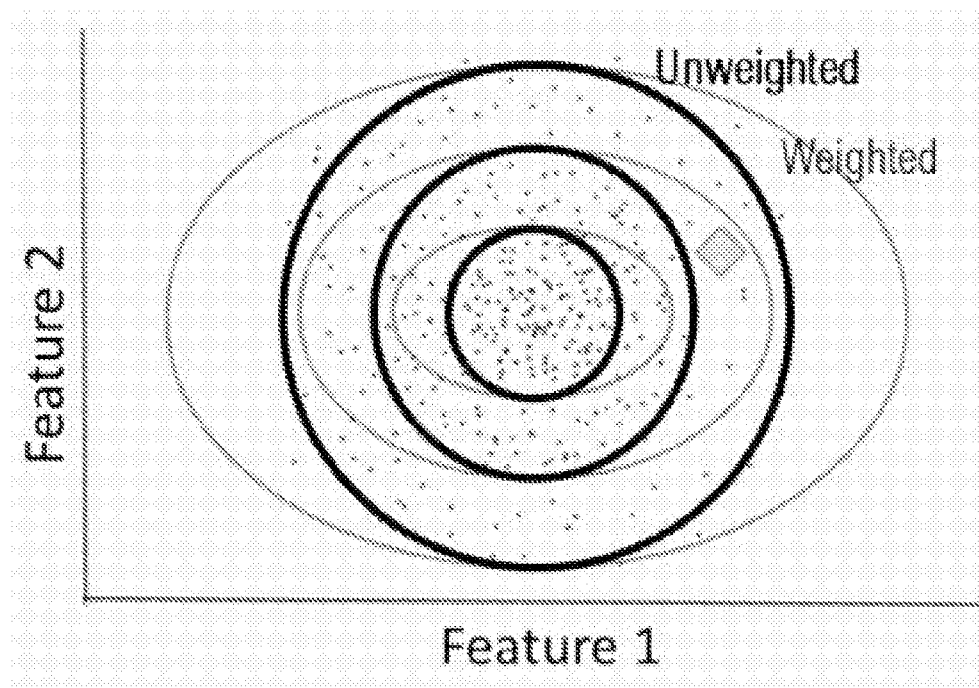


FIG. 5

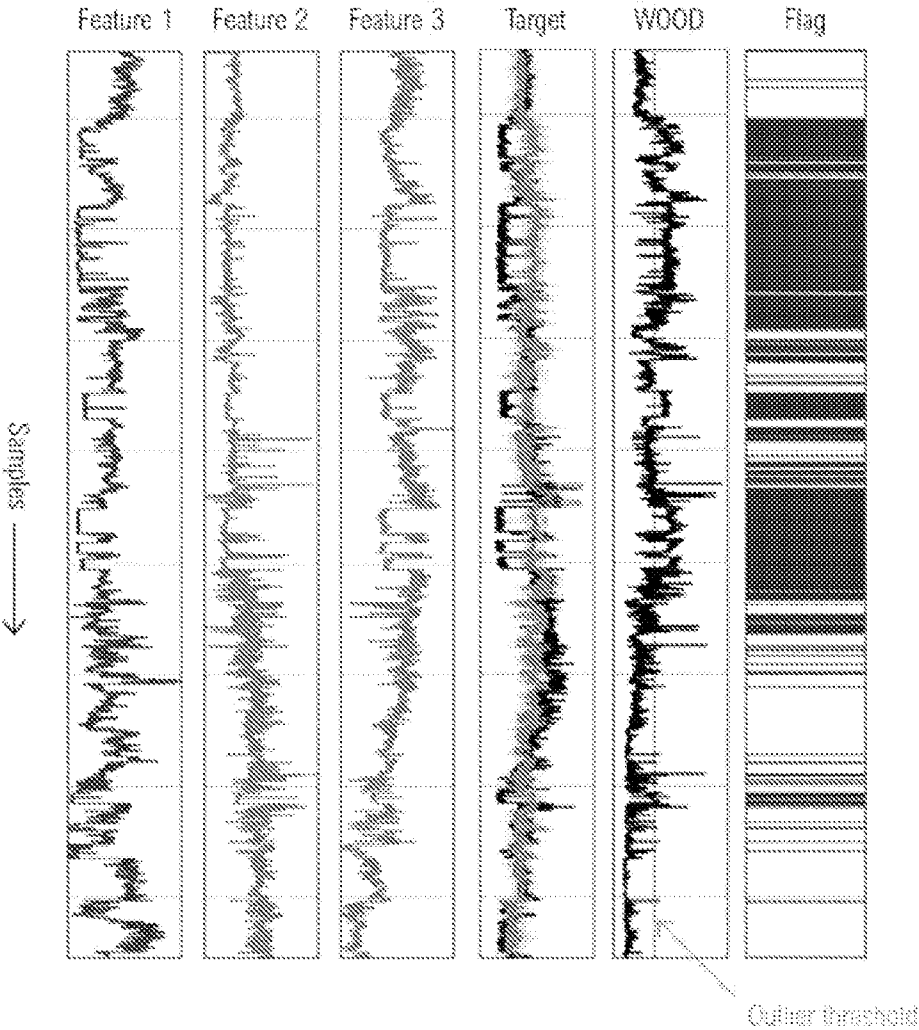


FIG. 6

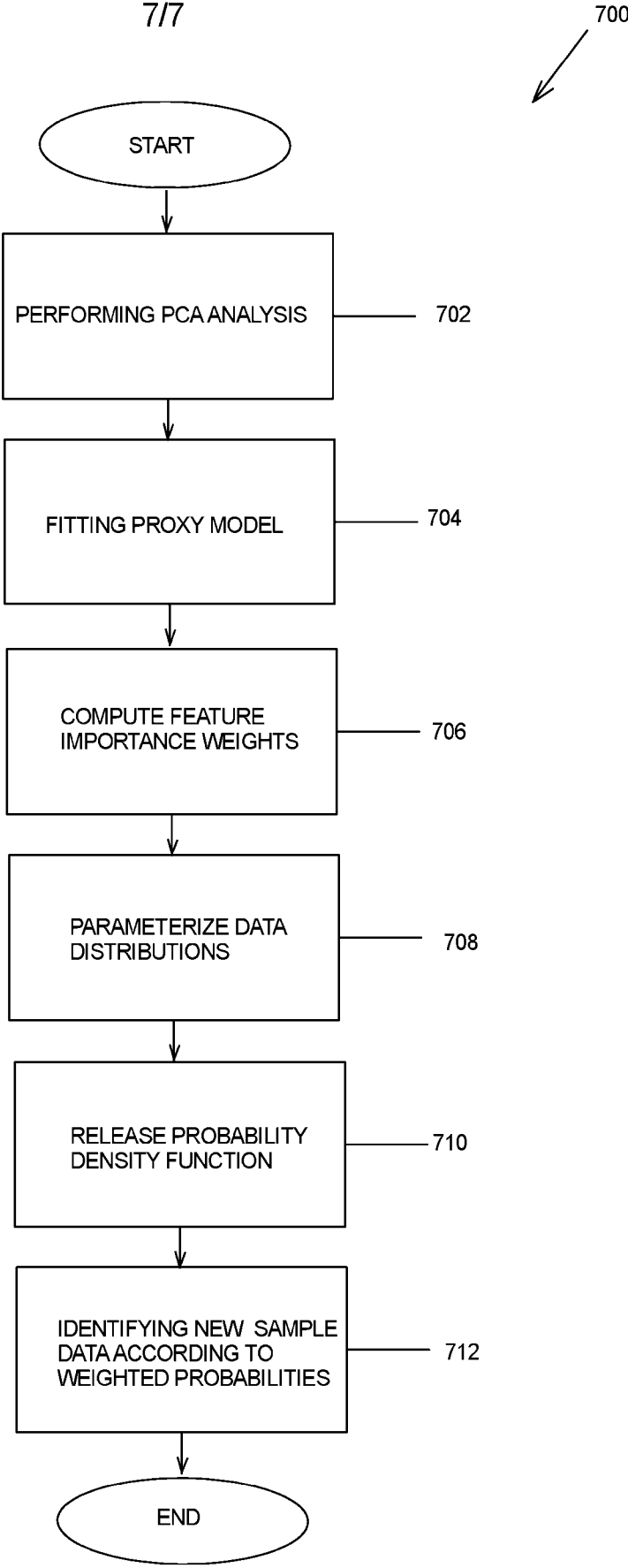


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/037535

A. CLASSIFICATION OF SUBJECT MATTER**G06N 20/00**(2019.01)i; **G06N 7/01**(2023.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N 20/00(2019.01); E21B 44/00(2006.01); G06F 17/18(2006.01); G06N 3/08(2006.01); G06N 5/02(2006.01);
G06N 5/04(2006.01); G06N 7/00(2006.01); G06Q 10/06(2012.01); G06Q 50/06(2012.01); H04L 12/24(2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: principal component analysis (PCA), feature importance, weight, outlier, probability
distribution, neural network, geological**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 110705859 A (CHINA THREE GORGES UNIVERSITY) 17 January 2020 (2020-01-17) paragraphs [0036]-[0037], [0039]; and claim 1	1-17
Y	US 2018-0189664 A1 (NATIONAL UNIVERSITY OF IRELAND, GALWAY) 05 July 2018 (2018-07-05) paragraphs [0022], [0048], [0066], [0176], [0186]-[0187]; and claim 1	1-17
Y	CN 114934766 A (CHINA UNIVERSITY OF GEOSCIENCES (WUHAN CITY)) 23 August 2022 (2022-08-23) paragraph [0017]	13-17
A	US 2021-0383248 A1 (SAP SE) 09 December 2021 (2021-12-09) claims 1-3	1-17
A	US 2023-0043993 A1 (ORACLE INTERNATIONAL CORPORATION) 09 February 2023 (2023-02-09) paragraphs [0021]-[0023], [0099]	1-17



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

21 October 2024

Date of mailing of the international search report

21 October 2024

Name and mailing address of the ISA/KR

**Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon
35208, Republic of Korea**

Facsimile No. +82-42-481-8578

Authorized officer

YANG, Jeong Rok

Telephone No. +82-42-481-5709

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2024/037535

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110705859	A	17 January 2020	None			
US	2018-0189664	A1	05 July 2018	EP	3109801	A1	28 December 2016
				WO	2016-207369	A1	29 December 2016
CN	114934766	A	23 August 2022	None			
US	2021-0383248	A1	09 December 2021	US	11568279	B2	31 January 2023
US	2023-0043993	A1	09 February 2023	None			