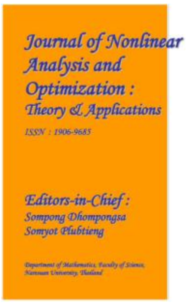

Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2, No.5 : 2024
ISSN : **1906-9685**

AN UNVEILING MODEL FOR DISTRIBUTED FEATURE MAPPING AND DATA DISTRIBUTION USING INTELLIGENCE APPROACH

P.HEMALATHA Research Scholar, PG & Research Department of Computer Science
Adaikalamatha college,Vallam , Affiliated to Bharathidasan University, Tiruchirappalli India
Dr. J. LAVANYA, Assistant professor and Head, Department of Artificial Intelligence and Machine Learning, Queens College of Arts and Science for Women, Punalkulam, Pudukottai, India. Affiliated to Bharathidasan University, Tiruchirappalli. lavanyamona01@gmail.com

Abstract- Artificial Intelligence (AI) is already an integral component of every individual's daily life, there is a problem of trust in these systems, which makes it more important than ever to describe black-box forecasts, particularly in the financial, healthcare, and military sectors. Although benchmark datasets are the focus of contemporary explainable Artificial Intelligence (XAI) methodologies, memory or computation limitations remain to determine the cognitive usability of such solutions in large-scale data scenarios. To handle high-volume datasets, we expand the model-agnostic XAI approach referred as Clustering Ensemble with Intersecting k-means (CEIK) in the present investigation. By integrating both local and global data, the proposed CEIK model seeks to explain the properties of predictive techniques. Specifically, the local reasoning gives instructions for changing the probability of the predicted class and a prediction-based reasoning for the identification. To manage the large volume, diversity, and velocity of massive datasets, our extension makes use of contemporary big data methodologies. The dataset were used to assess the framework's efficiency, description quality, and the models' significances. Our findings show that the suggested method effectively handles big datasets while maintaining the high-quality descriptions linked to the proposed model. Crucially, it shows a sub-linear reliance on dataset size instead of an exponential one, which makes it scalable for large datasets or any big data situation.

Keywords- big data, data representation, prediction, feature requirements, clustering, ensemble

1. Introduction

Our everyday lives have been drastically altered by the introduction of artificial intelligence (AI), which has a significant impact on business logic and gives those who were the first to pursue this new path a clear competitive edge. According to the International Data Corporation (IDC), 37.5 billion dollars were spent on artificial intelligence in 2019, which is over 44 percent greater [1]. The IDC also projects that 97.9 billion dollars will be invested on AI in 2023. All of these expenditures have resulted in what is known as AI-driven advancement [2], which aims to establish best practices and techniques for integrating AI into applications. This not only makes it easier for AI-powered solutions to be approved across a variety of enterprise fields, but it also makes AI more accessible to all. As a result, AI has significantly influenced society, particularly when it comes to accessing our personal information and making judgments for us [3]. Indeed, AI is already influencing every aspect of our everyday lives, and we have grown adapted to it before understanding it. AI makes decisions on a regular basis, whether it be through personalized advertisements on Google search engine pages, friend suggestions on Facebook, or purchase and movie suggestions on Amazon and Netflix [4] – [5]. However, businesses can depend on traditional ML methods, which gain knowledge from the extracted feature and produce ambiguous predictions, as long as we are talking about movie suggestions or customized ads. However, it is crucial to understand the rationale behind such a crucial decision under life-altering circumstances, like a military operation or an illness diagnosis [6]. Unfortunately, the inherent complexity of AI-based systems is a major barrier to their general adoption. Because of their black-box nature, these systems are competent of provisioning powerful predictions, but they are also difficult to directly explain. Thus, this dilemma has sparked a growing conversation on eXplainable AI (XAI), a new field of investigation with significant promise to enhance the reliability and accessibility of systems relies on AI [7]. Unquestionably, XAI is considered as the essential precondition for AI to continue on its unbroken path of advancement [8]. Finding a clear model that explains the reasoning underlying the prediction method without compromising accuracy is the major objective. Unfortunately, the most sophisticated models such as intricate neural networks with numerous hidden layers—also tend to be the most accurate, whereas decision trees and other more straightforward and intuitive models do not necessarily do as well.

As demonstrated in [9], XAI is a new topic who's significance is occasionally associated with the following various concepts: (i) Explainability, which refers to an AI architecture's capacity to communicate its judgments to humans in a way that they can understand, and (ii) Interpretability, which refers to the recognition of a feature set which has influenced a final decision [10]. The DARPA states that the two primary objectives of XAI are to make models more visible while maintaining a high degree of learning efficiency (predictive exactness, for example) and to allow consumers to recognize, and trust. In summary, explainability is a useful tool for helping AI systems justifies their choices. Its importance encompasses a number of important areas, such as prediction validation, model improvement, and gaining new information about the particular issue at hand. As a result, AI systems become more trustworthy and have access to a wider range of applications. This work rely on AI systems, XAI has the potential to provide significant benefits [11]. In this investigation, we handle the challenge of elucidating the reasoning behind a particular AI system choice. In comparison with universal explainers, which attempt to clarify the entire system's functioning, local explainers typically anticipate that the intricate decision function that governs the activities of the framework and can be approximated by an interpretable approach in the immediate vicinity of the target instance that the research team attempting to describe [12]. Although the aforementioned researches focus on the local explanation challenge, we firmly believe that clarification should not only provide information about how the model behaves close to a final instance but also enable users to understand how the model functions more broadly, including in situations with unknown inputs. In a different way, combining local and global explanations will help us better grasp the model's prediction in accordance with the information it has gained throughout training [13] – [15]. Furthermore, the majority of earlier methods are assessed using standard datasets; nevertheless, their scalability to huge data is frequently disregarded, which hinders the solutions' practicality.

The XAI methodology "Clustering Ensemble with Intersecting k-means (CEIK)," is extended in this research. To meet the challenges of managing large datasets, this development aims to combine local and global data. A clustering phase is used to interpret the global operations of the AI system. Its purpose is to find areas in the domain of instances where data points are consistently categorized by the decision framework and accurately reflect the actual data distribution. In order to make interpretation easier, we provide the consumer with hyper-rectangle that contains the examples in the cluster and is displayed as the logical norm. Additionally, the framework's local behavior is used to help the user understand what influences or weakens a prediction. In particular, we suggest an instance space transformation that allows us to deduce the effect of changing feature values on the forecasting possibility. In marked comparison to the initial version of proposed CEIK model, the enormous difficulties presented by large, heterogeneous, and dynamic datasets led to the creation of a unique framework, based on massive data methods, designed especially for proposed CEIK implementation. The success of the suggested framework is demonstrated by our tests on five large-scale datasets with respect to of model significance, cluster quality, and temporal efficiency. In conclusion, the following are our contributions:

✓ To handle large datasets, we expand the explainable AI (XAI) approach as the proposed Clustering Ensemble with Intersecting k-means (CEIK).

✓ Unlike existing concept, our method uses distributed computing and large-data data technologies to make the processing of the explanations more scalable and accurate using the clustering and ensemble model.

✓ To demonstrate the efficacy of the suggested extension, we do a thorough analysis on dataset with various big data attributes (volume, variety, and velocity).

The work is organized as: section 2 gives wider analysis on diverse approaches. The methodology CEIK is drafted in section 3. The numerical results are given in section 4 with conclusion in section 5.

2. Related works

The focus of current XAI developments has been on understanding the internal functioning of black-box designs, mostly using rule-based approaches and feature importance. The goal of feature importance techniques is to explain model behavior in particular cases. To give each feature significance, these approaches usually use models that are inherently interpretable, like decision tree structures or linear models. One of the most well-known of these is the Local Interpretable Model-agnostic Explanations (LIME) approach [16] which produces insights for certain identifications regardless of the fundamental machine learning technique that is employed. This is accomplished by altering the original data to provide fresh samples that aid in the formation of explanations. Another noteworthy approach in this category is SHapley Additive exPlanations (SHAP) [17] – [18]. It uses Shapley values, which are obtained from the theory of games, to clarify individual predictions. By taking into account the existence or lack of characteristics in a combination that supported the forecast and analyzing how the addition of a feature changes the prediction result, Shapley values evaluate the influence of every attribute on the prediction. Additionally, by looking at local performance and using regular pattern collections from training dataset to find locally discriminative characteristics close to a test instance, PALEX offers instance-level explanations [19]. Rule-based approaches, on the other side, provide clear guidelines that improve users' comprehension of choice boundaries. This method is demonstrated by Interpretable Decision Sets (IDS) which optimize the complexity, coverage, and accuracy of a set of standalone "if-then" principles. A model-agnostic approach that evolved and incorporates rule-based interpretability and local explanations to create anchors that secure predictions locally was presented by [20].

Furthermore, other methods presented in the investigations [21] use global information to produce local explanations, including counterfactual as well as supporting principles. However, the localized character of their explanations presents difficulties for these approaches. For instance, the reliability of linear model explanation decreases beyond a certain radius around the target instance. This restriction results from the lack of data regarding the size of the neighborhood where the explanation is still applicable [22]. The model incorporates "global data" from the design into the local explanation procedure to resolve these issues. To ensure uniform categorization, this is accomplished by a clustering phase that produces pure clusters with high inter-cluster variation and low intra-cluster variation. These improved rule-based descriptions use directional data to support the assumption as well as feature importance scores. This method offers a thorough grasp of model behavior in a variety of contexts by defining opposing and supporting approaches for the estimation in addition to providing rule-based explanations [23].

XAI is to generate intelligent systems that explains how they make decisions. Numerous thorough studies on XAI have emerged in the past decades [24], with an scientific emphasisas well as comparative study of diverse approaches. By using a local-to-global structure, these studies aid in the development of AI or ML applications that move from non-transparent to apparent. The necessity for XAI techniques has been highlighted by the increasing reliance on "black box" decision-support systems, especially in vital industries like medical care, security, and defense. By making such systems more transparent and predictable, these methods increase user trust [25]. In the medical field, the author looks into XAI techniques meant to increase reliability, accountability, and transparency in healthcare applications. A visual explanation technique for breast tumor diagnosis was proposed by the author in [26]. It is based on the quantitative and qualitative synchronization of consumer requests with extracted examples. The use of XAI techniques for Alzheimer's disease (AD) recognition is investigated by the researcher. Similar to this, the author discuss about how XAI approaches can be integrated with ICTs (information and communication technologies) in the financial sector to reduce risks and increase efficiency at the same time. Additionally, a thorough assessment of XAI applications in the field of social science is presented by [27], who highlight important discoveries and persisting difficulties.

The flexibility and applicability of prominent explanation strategies in large data systems remain a significant research gap in current XAI research [28]. Although established approaches are reliable in typical contexts, they face considerable difficulties when used with large, intricate datasets, which are typical of big data situations. Given the exponential expansion and complexity of modern datasets across several fields, this limitation presents a significant obstacle. In order to bridge this gap, our work implements and experiments with the technique in a large-scale data environment, this advances XAI. With its distinct method of using clustering techniques to generate explanations, the model naturally fits in with the complexity of big data [29]. Because clustering can efficiently divide large datasets into manageable groupings, allowing for more accurate and computationally effective explanations, such alignment is crucial [30].

2.1. Contribution

This section explains the framework that was created to address the current business issue. With the application of log assessment system generated in collaboration with domain experts, multiple gigabytes of streaming log and corporate data are automatically examined to determine pertinent elements. On the basis of it, we create a pipeline for machine learning that is both automatic and interpretable to structure the user requirements and determine the likelihood of escalation. For the consumer, a well-developed decision support system provides an explanation of the historical data and forecast likelihood based on features that have been collected from log and enterprise data. From the perspective of the consumer, handling data sources has significant advantages. We can discover characteristics linked to prediction based on which features best explain a forecast.

3. Methodology

3.1. Dataset construction

This section shows how the dataset is structured for the arrangement of the experiment is explained in the first Algorithm. We will dive deep into the procedure in the following sections. This work considers five datasets namely susy, HTRU, Gamma, diabetes and avila. The hyper-parameters k are 30, 20, 20, 22 and 26 respectively.

3.2. Labelling

Let I represent the whole set of user's data. The labeling strategy for a single example client $i \in I$ utilizing a sliding window strategy is shown. Since progression decisions are taken weekly, the step size is set at one week. As suggested by domain specialists, we fixed the window length to ten steps. Various values were also assessed, but no enhancement was found. A feature vector x_i, t_{pred}, where t_{pred} is the final week in a window which is retrieved from this window. Let $T_{i,esc}$ be the collection of all client I escalation flags and let $T_{i,esc}$ represent a particular moment in time when a customer I escalation flag occurred.

Two steps are chosen as the prediction interval. The label (y_i, t_{pred}) for this collection is assigned to 1 if there was actually an escalation $t_{i,esc} \in T_{i,esc}$ in the predicted range. The 4 actions that present after an escalation $T_{i,esc}$ is referred to as an infected interval. Since researchers presume that there exists a particular concentration on consumers for whom a current escalation happened, domain experts selected this value. All samples that include weeks from the infected time duration in the sliding window are not included. We continuously execute this process for every end user for a set period of 104 steps which in the present instance is equal to two years. This permits us to replicate our framework's actual effectiveness for a whole year. Instances for every consumer with full information are incorporated in the target dataset, $D_{tpred} = (x_{tpred}, y_{tpred})$. Fig 1 depicts the final distributions. The total amount of escalations $|\left|y_{t_{pred}}\right||$ is relatively consistent as time passes, as the consumers count (y_{tpred}) increases. This happens due to the number of consumers that can be focused on each week is restricted owing to a lack of service resources. Lastly, as an industry benchmark dataset, we make available to the investigation community an anonymized form of D.

3.3. Feature representation

This process is necessary to extract the required features from large-scale dataset and to reduce the computational complexity. Some features may cause noise to enter computed features. Therefore, we agreed to aggregate features per week in discussion with domain experts. This work compute characteristics for an interval of 10 weeks for every client in I and forecast weeks y_{tpred}.

Log Information: It is not possible to study the machine log data in its raw format. Rather, to identify particular event sequences identified by domain specialists, we employ a log evaluation methodology.

The features that were selected have distinct interpretations that are connected to particular system errors that impact clients' everyday work schedules. These characteristics include things like system delay, UI pop-ups, UI freeze, and abort of execution. We also determine whether a system's software (SW) has been updated.

Business Information: The accessible corporate data can be divided into 2 related groups: sales information and client service tokens, as explained. The quantity and expense of restored parts are considered as sales features. Based on accessibility in the various ticketing structures, features obtained from ticket information include the amount of open tickets, the rated degree of severity, the age of the earliest open ticket, and the frequently visited site for each client. It is possible to extract these attributes globally.

3.4. Time Complexity

As Algorithm 1 contains 3 nested loops that rely on the consumers count $|I|$, the amount of time steps $|W_{exp}|$ to take into account, and the time steps counts of the monitoring window $|W_{obs}|$, its worst case time complexity is $O(n^3)$. We simply need to update fresh data every week, this particular study is not concerned with time complexity.

3.5. Clustering and prediction

The clustering phase is the focus of the framework evaluation's initial stage. First, we will examine two clustering techniques on the various datasets that spark.ml provides. After the optimal configuration has been determined, the investigators can assess the quality measurements suggested in the preparation phase. The following clustering techniques are employed:

K-means: This clustering technique divides a set of n data points into $k(\geq\ n)$ groupings to reduce the sum of squares within the cluster. It enables the initialization of this algorithm in 2 modes:

(i) K-Means represents the normal K-Means approach where centers (k) are selected arbitrarily and then, during each iteration, the approach allocates every case to the nearest center and recalculates the centers, until the process of convergence.

ii) A parallelized version of the K-Means++ technique, KMeans // computes the initial k centers based on the data allocation to identify the most favorable clusters in less iterations.

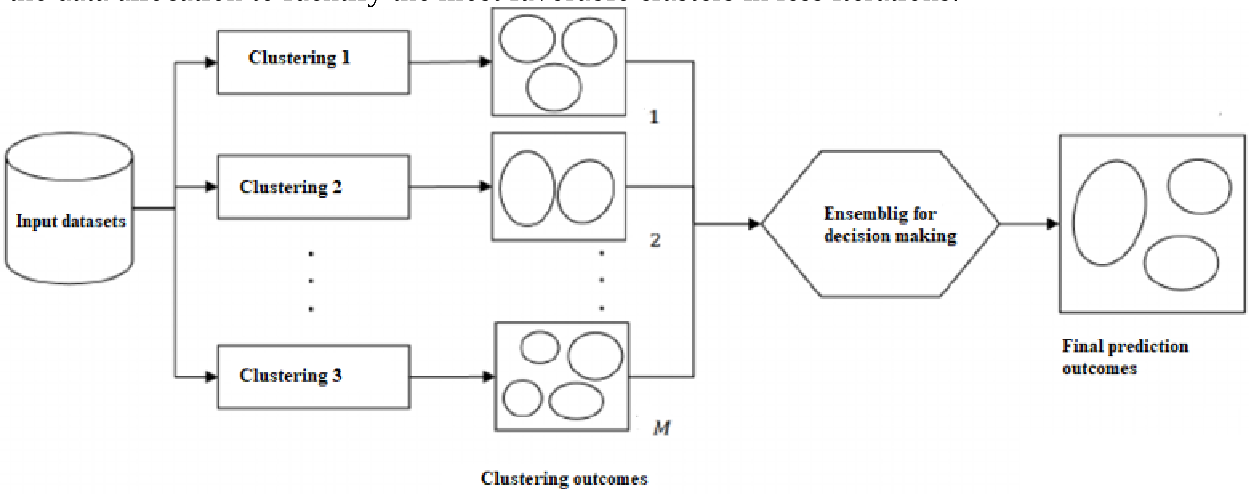

Fig 1: Interesting k-means

iii) Intersecting K-Means: This technique integrates the components of divergent hierarchical clustering (also known as top-down clustering) and K-Means clustering. As a replacement for partitioning the extracted dataset into k equal clusters in all iterations, the intersecting K-Means approach gradually partition a single cluster into 2 sub-clusters during each intersecting stage which can be carried out employing K-Means, until an optimal number of k clusters are attained. For both clustering techniques, the total number of clusters (k) must be specified. To improve the outcomes, we sought the optimal transaction between k and the price. In the last scenario, the inertia value is the average squared space among every instance and its closest centroid. An algorithm that performs better has a lower inertia value. The technique involves visualizing the inertia as the total quantity of clusters improves and repeatedly executing the algorithm while increasing the cluster counts $(k\ \in\ [2,40])$. The evaluation was conducted using the initial data set sizes, and thus the inertia value is solely reliant on the dataset length. The technique is obviously the same which previously has occurrences far greater inertia values compared to the other datasets.

3.6. Ensembling model

The following are some advantages of ensemble decision tree techniques:

- Consumers can learn which of the specified features are "correlating" with escalations or opinions of clients by using the estimated feature importance.
- The model output from ensemble approaches is a probability that can be read as the sentiment of the customer (probability for escalation). For better troubleshooting, we can give the consumer the importance of each parameter for all forecasts because every combination of time point (week) in a window and intended feature is represented as a single input factor. We use XGBoost (XGB) and Random Forest (RF) as our decision tree ensemble approaches.

Two ensemble learning methods that are applicable to both classification and regression are Random Forest and XGBoost. In this instance, an issue of binary classification is of significance to researchers. A group of weak classifiers $\{C_i\}$ that each receive the identical input x_{tpred} and produce the predicted class $C_i(x_{tpred}) \in \{0, 1\}$ is generally referred as an ensemble learning approach. The following is the definition of the ensemble technique's probability result:

$$\hat{y}_{t_{pred}} = \frac{\sum_{n=1}^{N} \hat{y}_{t_{pred,n}}}{N} \in R_{[0,1]} \quad (1)$$

By using a bagging (bootstrap integration) technique, the decision trees for RF are produced separately and concurrently. The Gini impurity criteria is the objective function that needs to be reduced in this case. This indicates that there are two steps involved in creating each decision tree:

1) Bootstrapping: Here, the source dataset $D_{train} = (x_j, y_j)$ are sampled individually for every base classifier C_i on data points and features with $j \in \{1, \ldots, m\}$. The data points in D are sampled independent and equally distributed into a subset $D_{train_i} = \left(x_j, y_j\right)_{j \in J_i}$, where $J_i \subset \{1, \ldots, m\}$, to put it another way. Additionally, if x_j composed the features $F = \{f_k : k \in \{1, \ldots, n\}\}$, then x_j^i has features from $F^i \subset F$. This is because the feature space is sampled and independent and equally distributed

2) Aggregating: Calculating the average or, in this instance, selecting the class by majority vote. The probability output is what we are interested in in this instance. To create a powerful classifier, Gradient Boosting also combines numerous weak classifiers. Unlike bagging, decision tree structures are constructed sequentially rather than concurrently, and the outcomes are aggregated as they are generated. In this instance, we used the gradient boosting library XGBoost. The predicted class (in this case, 0 or 1) and the probability the model gives each prediction are output by the model in both scenarios. The anticipated consumer response is determined by using the probability that the model allocates to class 1. Every input x_{tpred} has a value $\hat{y}_{t_{pred}} \in R_{[0,1]}$. The combination of DTs generates a prediction according to the majority vote as shown in Fig 2.

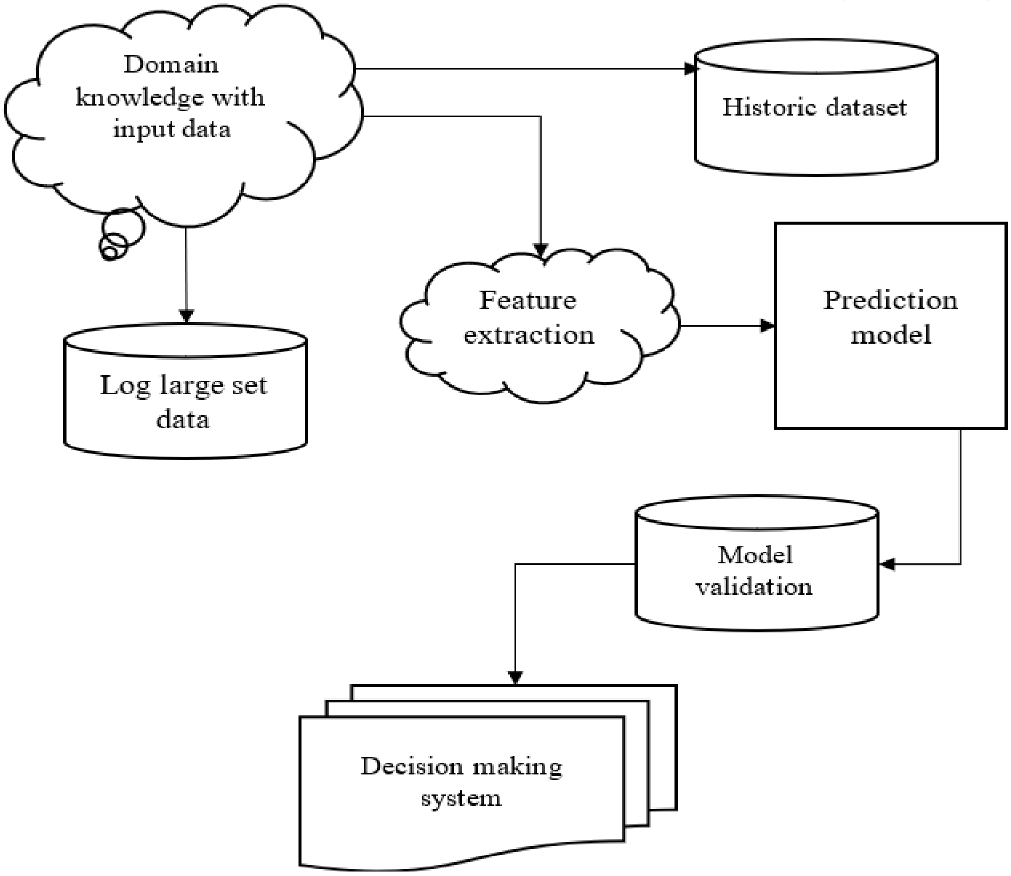

Fig 2: High-level data analytical representation

Using the imblearn framework, we solve the unequal class issue by employing either arbitrary under-sampling of the majority class or arbitrary oversampling of the minority class, i.e. SMOTE. In the selection method, we regard the sampling strategy as a hyper-parameter. We used two distinct methods for data fusion. To train a single classifier for "early" fusion (M1, M2), we basically stacked enterprise and log (x_{log}) features. We trained first ensemble classification system based on x_{ent} and next based on x_{log} for "late" fusion. For the very last prediction, each ensemble classifier's result was subsequently input into a logistic regression layer. One of the two ensemble classifiers is XGB or RF. Furthermore, we attempted to train a classifier just using x_{ent} or x_{log}.

3.7. Network tuning

Finding the ideal value of k for the two methods of clustering we have just described is the first step in our analysis. To improve the effectiveness of the clustering procedure, we pre-processed the data using a Min-Max scaler before running these algorithms. As the algorithms are based on the same K-Means technique, it is remarkable that they differ very little in terms of inertia. Results from K-Means Random and integrated k-means are strikingly comparable which is to be expected as their main difference is in the initialization stage. The proposed clustering, on the other hand, continuously performs worse in comparison, as seen in all datasets. Another important consideration is the execution time; cluster computation time plays a major role in the explanation phase and should be kept within tolerable time limits. The execution timings for the previously stated methods are shown in Fig 3 to Fig 6. Although there were little variations in the inertia values, the execution durations of K-Means and the proposed clustering show a sharp discrepancy. The sequential structure makes it less practical in real-world situations because it takes significantly longer to execute than K-Means.

3.8. Clustering Quality

According to the information in the pre-processing stage, a clustering outcome should have the following qualities in order to be helpful in the explanation procedure:

(i) High coverage expects the clusters to include as diverse examples as feasible;

(ii) High purity, to ensure the reliability of the resulting explanations which depend on the clustering outcomes; and

(iii) Low overlap among clusters specifying distinct classes. We combined all of these attributes into a single quality metric in order to count them all, as indicated below:

$$quality = w_p.purity + w_c.coverage + w_o.(1 - overlap) \qquad (2)$$

Where, the weights for overlap, coverage, and purity are denoted by w_o, w_c, and w_p, accordingly. If the program favors one property above the others for any reason, these weights can be helpful. To help with noise suppression, it is recommended, for instance, to select a low weight w_c for the coverage term if the dataset has an excessive amount of noise. We set $w_p = w_c = w_o = 1/3$ and assumed identical contributions for my investigations. Although the clusters' explanation is calculated in the actual space, we used the actual instances this time rather than preparing the information with a scalar to assess the metric utilizing the optimal k that was acquired in the preceding section. The outcomes for the various algorithms are displayed in Tab 1 to Tab 5. The quality assessment examination shows that the techniques for clustering reliably perform well across the majority of metrics. One important finding is that for every algorithm, the coverage (C) metric consistently registers the datasets. It is a built-in feature of the K −Means approach which guarantees full coverage by ensuring that all instances are included in the clustering procedure. With the provided dataset, the intersecting K-Means technique typically performs worse than its competitors when considering overall quality (Q). Intersecting K-Means operates better in this particular case because of a slightly higher purity (P) value. In this case, purity indicates how homogeneous the clusters are, indicating that, while though Intersecting K-means is generally less efficient, it can, in some circumstances, produce superior homogeneity. It is imperative, nevertheless, to strike a compromise between these quality indicators and execution time. The other types of K-means variations outperform the intersecting K-means method in terms of computing performance, even though it occasionally offers somewhat higher purity. Given the significant difference in execution durations, with intersecting K-means being noticeably requires more computational time, the decision to prefer the least time-effective K-means techniques for all datasets seems appropriate. This choice is predicated on the idea of attaining best performance as a whole, which takes into account both the algorithms' practical viability in terms of time and computational restrictions as well as the quality of clustering.

Algorithm 1:

```
Set dataset = $\emptyset$;
  for all users $i \in I$ do
    for $t_{pred} = t_0$ do
      if users exists then    //$t_{pred} - 10$
        for $t = t_{pred} - n + 1: t_{pred}$ do
          extract the data features → $X_{i,t,log}$
            extract the domain knowledge features → $x_{i,t,ent}$
        $x_{i,t_{pred}} \rightarrow \left(x_{i,t,log,}x_{i,t,ent}\right);\ \ for\ all\ \in \{t_{pred} - n + 1, \dots, t_{pred}\}$
        if $t_i \in T_{i,esc}\ \in \left\{t_{pred} + 1, t_{pred} + 2\right\}$ then
          $y_{i,t_{pred}} \rightarrow 1$
        else
          $y_{i,t_{pred}} \rightarrow 0$
        $D \rightarrow D_{i,t_{pred}} = (x_{i,t_{pred},}y_{i,t_{pred}})$
  for setting flag for user data
    $D \rightarrow D$   //discard
//Training and validation
for $t_{pred} = t_o$ do
  $D_{training} \rightarrow D_{prediction}: t_{pred};$
  $D_{training^*} \rightarrow D_{t_{prediction}}: t_{pred};$
  $D_{validation} \rightarrow D_{validation}: t_{pred};$
  $D_{testing^*} \rightarrow D_{t_{prediction}}:$
Model selection with $D_{training}$ and $D_{validation}$ on average training with dataset hyper-parameters;
Testing model;
Compute performance metrics.
```

4. Numerical results and discussion

This study relies on the instance-space transformation which demands a proximity function to be estimated. This function depends on calculating the gap between two cases, generates an integer that represents how close or comparable two data points are to one another. We evaluated our methodology with respect to changes in proximity and distance functions. Specifically, we looked at the effectiveness of the following proximity functions as well as the Euclidean, Cosine distance, Minkowski, and Chebyshev.

$$P_1(x_1,x_2) = \frac{1}{1+\delta(x_1,x_2)} \tag{3}$$

$$P_2(x_1,x_2) = e^{\delta(x_1,x_2)} \tag{4}$$

$$P_3(x_1,x_2) = -\,\delta(x_1,x_2) \tag{5}$$

$$P_4(x_1,x_2) = 1 - \frac{\delta(x_1,x_2) - \min\,(\delta(x_1,x_2))}{\max\big(\delta(x_1,x_2)\big) - \min(\delta(x_1,x_2))} \tag{6}$$

$$P_5(x_1,x_2) = \max(\delta(x_1,x_2)) - \delta(x_1,x_2) \tag{7}$$

Here, x_1 and x_2 are two distinct data points; $\min(\delta(x_1,x_2))$ and $\max(\delta(x_1,x_2))$ reflect the least and greatest distances recorded in the entire dataset, respectively, while $\delta(x_1,x_2)$ is a standard distance function. Eq. 3 to Eq. 7 explain the ideal proximity functions, where the proximity shows a linear connection with the distance as shown by the results in Tab 1 to Tab 5. While Euclidean or Minkowski distances are usually the most successful arrangements, cosine distance generally performs better than other approaches. Two primary criteria will be used to calculate the methodology's evaluation:

1. The technique's effectiveness will be assessed in regard to whether the explanation procedure is impacted by transformation. In particular, we gathered the linear models' altered coefficients of determination (R^2) that were learned on both the original and modified areas. In statistical evaluation, modified R^2 is a value that represents the amount of explained diversity in the data set that is used to evaluate how effectively an algorithm explains and identifies future outcomes. Unlike R^2, modified R^2 is not influenced by the number of factors that the linear framework was used to fit. In order to explain occurrences that were not previously seen by the cluster algorithm, the datasets were divided into two sets, and no confidence intervals were calculated.
2. To see the differences between a conventional setting and a large-scale data engine while dealing with enormous volumes of data, the execution times of the basic Python program and Spark will be compared in order to assess the methodology's performance.

4.1. Efficacy

The modified R^2value provides a suggestive indicator for the explained variability in the data set is employed to evaluate the technique's effectiveness.

$$modified\ R^2 = 1 - \frac{n-1}{n-k-1}.\frac{RSS}{TSS} \tag{8}$$

Here, n represents the total amount of samples in the neighborhood (initially declared as 1000), k represents the feature count on which the framework has been fitted and RSS and TSS represents total sum of squares and residual sum of squares respectively. A comprehensive overview of the performance measures in the actual space and the converted space is shown. The average adjusted R^2 values across different datasets together with their corresponding 95% CIs, are used to quantify this contrast.

Table 1: Quality evaluation

Dataset	**C**	**P**	**O**	**Q**
A	1.00	0.89	0.98	0.96
B	1.00	0.82	0.97	0.93
C	1.00	0.79	0.96	0.92
D	1.00	0.98	0.98	0.98
E	1.00	0.75	0.98	0.91

Table 2: Quality evaluation with conventional k-means

Dataset	C	P	O	Q
A	1.00	0.84	0.97	0.95
B	1.00	0.83	0.97	0.93
C	1.00	0.76	0.99	0.92
D	1.00	0.97	0.99	0.98
E	1.00	0.76	0.97	0.92

Table 3: Quality evaluation with proposed intersecting cluster

Dataset	C	P	O	Q
A	1.00	0.85	0.98	0.96
B	1.00	0.81	0.98	0.94
C	1.00	0.82	0.99	0.94
D	1.00	0.97	0.99	0.98
E	1.00	0.74	0.98	0.93

Table 4: Proposed model evaluation with existing approaches

Dataset	Lime	Castle	CEIK
A	0.59 ± 0.015	0.61 ± 0.015	0.71 ± 0.016
B	0.70 ± 0.058	0.72 ± 0.047	0.82 ± 0.047
C	0.57 ± 0.083	0.60 ± 0.082	0.70 ± 0.082
D	0.59 ± 0.065	0.63 ± 0.036	0.73 ± 0.036
E	0.63 ± 0.071	0.62 ± 0.065	0.71 ± 0.064

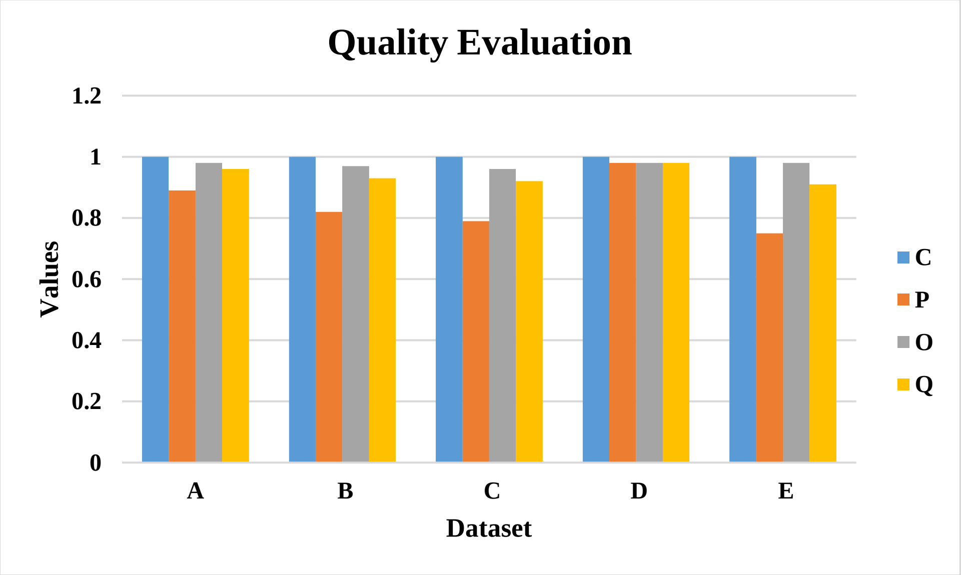

Fig 3: Quality evaluation

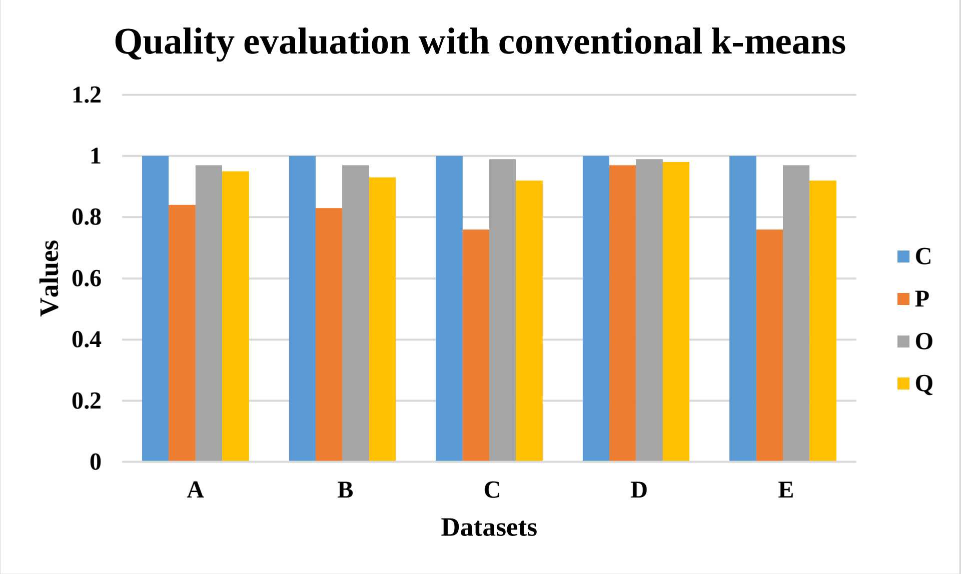

Fig 4: Quality evaluation with conventional k-means

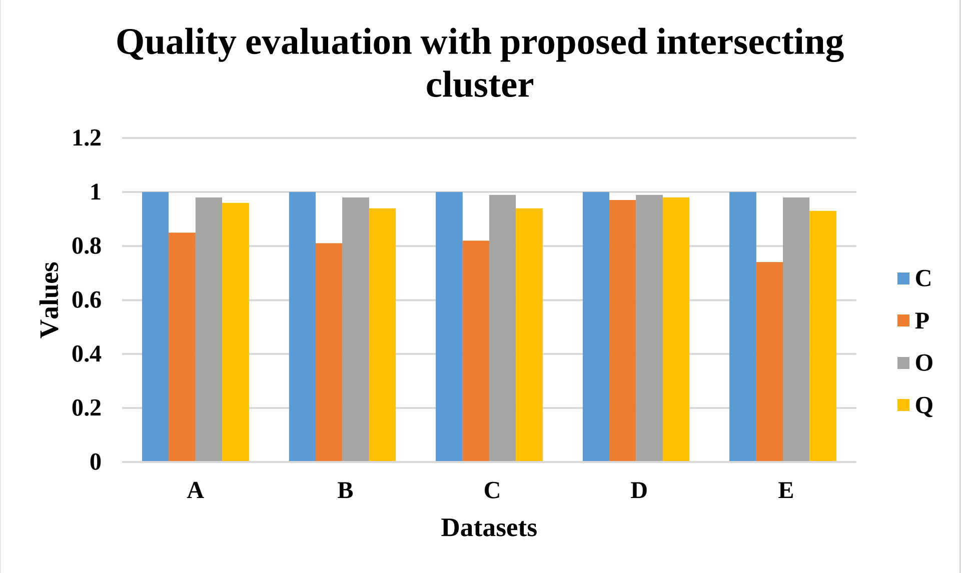

Fig 5: Quality evaluation with proposed intersecting k-means

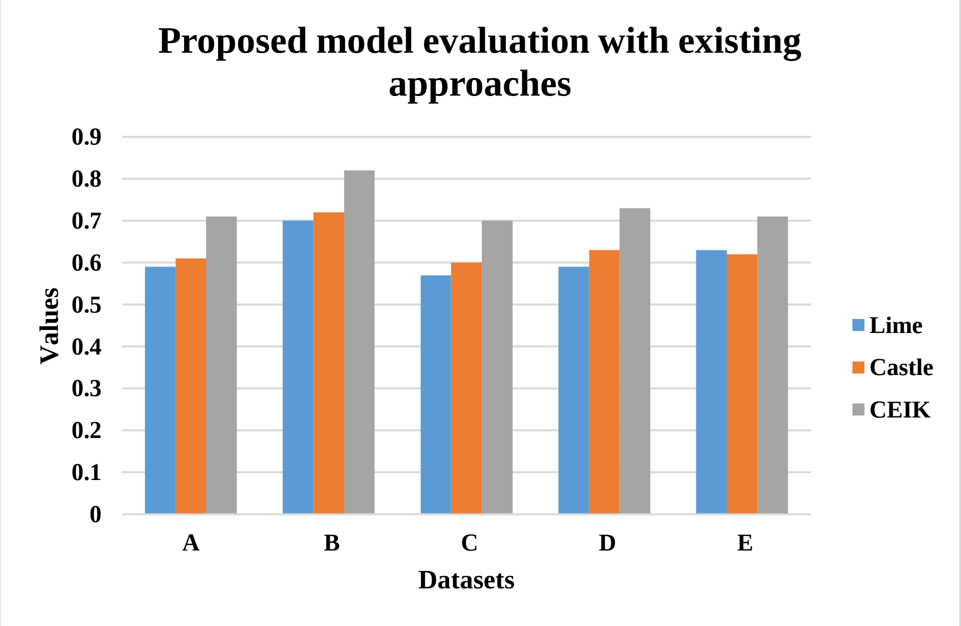

Fig 6: Proposed model evaluation with existing approaches

Table 5: Proximity and distance analysis

ataset	Proximity	Euclidean	Minkowski	Chebyshev	Cosine	Mean
Bank	P_1	0.23	0.28	0.26	0.39	0.29
	P_2	0.22	0.22	0.16	0.58	0.30
	P_3	0.68	0.69	0.42	0.70	0.62
	P_4	0.33	0.34	0.25	0.65	0.40
	P_5	0.69	0.69	0.40	0.69	0.62
	Mean	0.44	0.44	0.30	0.60	
Titanic	P_1	0.28	0.44	0.16	0.45	0.30
	P_2	0.15	0.28	0.14	0.43	0.22
	P_3	0.52	0.15	0.17	0.52	0.43
	P_4	0.40	0.40	0.13	0.43	0.35
	P_5	0.5	0.5	0.18	0.50	0.43
	Mean	0.35	0.35	0.16	0.45	
Diabetes	P_1	0.13	0.13	0.22	0.43	0.35
	P_2	0.08	0.08	0.15	0.44	0.29
	P_3	0.58	0.58	0.15	0.56	0.59
	P_4	0.45	0.46	0.32	0.55	0.54
	P_5	0.58	0.58	0.14	0.54	0.62
	Mean	0.36	0.38	0.19	0.51	
Magic	P_1	-0.045	-0.045	0.007	0.25	0.24
	P_2	-0.062	-0.061	-0.042	0.36	0.19
	P_3	0.583	0.58	0.197	0.50	0.45
	P_4	0.45	0.43	0.12	0.53	0.45
	P_5	0.57	0.58	0.1	0.51	0.46
	Mean	0.29	0.30	0.027	0.43	
Spambase	P_1	-0.04	-0.045	0.007	0.25	0.04
	P_2	-0.06	-0.061	-0.042	0.36	0.05
	P_3	0.5	0.58	0.19	0.50	0.46
	P_4	0.44	0.44	0.12	0.53	0.38
	P_5	0.5	0.59	0.22	0.51	0.46
	Mean	0.30	0.30	0.1	0.43	
Digits	P_1	0.07	0.07	0.02	0.28	0.11
	P_2	-0.001	-0.02	0.03	0.42	0.11
	P_3	0.5	0.56	0.09	0.50	0.43
	P_4	0.4	0.45	0.07	0.49	0.36
	P_5	0.5	0.56	0.08	0.51	0.44
	Mean	0.3	0.33	0.06	0.45	

A thorough examination of the data shows that an excellent decision of pivot points in proposed model has no negative impact on its performance in comparison to other approaches. The comparatively close modified R^2 values for both approaches across many datasets make this clear. In the Avila dataset, for example, proposed model modified R^2 value (0.610 ± 0.016) is marginally greater than existing approach (0.592± 0.015), indicating that the transformation in proposed model maintains its explanatory power. To completely comprehend the ramifications, it is imperative to investigate these findings further. The range that we may anticipate the genuine modified R^2 score to fall inside with a probability of 95% is shown by the confidence intervals. A more accurate estimate is indicated by a smaller interval. The interval width in the Diabetes dataset, for instance, is 0.058 for existing and 0.047 for proposed model, suggesting a more accurate estimation in the converted space.

4.2. Efficiency

By comparing the computational duration of the basic proposed methodology with its adaption for the search engine, this section provides an extensive evaluation of the computational efficiency. Because of the further computational processes that are inherent to proposed model namely, the initial clustering stage and successive data conversion, which increase complication and computation time to the workflow, this research does not include comparison with existing model. The search engine has a sub-linear pattern of growth in computation time across different dataset lengths, according to the data shown in Fig 5. This implies that as data size grows, the proposed design will be more scalable.

On the other hand, proposed model independent implementation exhibits an almost exponential development with execution times rising sharply with larger datasets. With distributed computing characteristics, the processing effectiveness of huge datasets is increased by utilizing simultaneous, processing across numerous nodes, are responsible for the sub-linear pattern seen with search engine. This is especially visible from the analysis of the raw execution times shows the relative increase rates of processing time for the two algorithm variants. For sustainable data processing systems, these results highlight the significance of improving algorithmic implementations, particularly when working with massive amounts of data can profit from the computing models. The findings specifies that the larger datasets, where computational expenditure are better controlled and minimized, are better suited for the implementation.

5. Conclusion

Designing a revolutionary XAI approach over a large-scale data architecture was the objective of this research. More specifically, Clustering Ensemble with Intersecting k-means (CEIK) is a new, XAI methodology that offers a thorough, comprehensive explanation for the classifier's predictions. Its effectiveness is in taking advantage of the model's local and global cluster behavior, which not only produces the rule explanation but also offers opposing directions for the prediction. The overhead associated with the clustering step is closely tied to the dataset dimension and the dimensionality curse is significantly impacts the conversion analysis represent two of the primary shortcomings of the actual algorithm. Actually, the first problem is solved with specially tuned clustering strategies in a distributed setting. Since the transformation is fully parallelizable, it may be computed in a distributed framework with ease and efficiency. The following is a summary of the key conclusions drawn from our Clustering Ensemble with Intersecting k-means (CEIK) findings:

- The sub-linear growth in execution time of our CEIK suggests improved handling of big datasets. With the proposed clustering process, we have attained excellent coverage and integrity.
- The proposed CEIK has shorter ranges of certainty and an adjusted R^2 that is comparable to existing methods, suggesting more accurate estimate in transformed region. In general, the distance operates better than other distance measures.
- Intersecting K-Means reveals slightly higher purity. The methodology will be extended in future works to cope with multimodal data (such as text and images), naturally altering the notions of neighborhood and clusters in the process. Furthermore, we wish to confirm that our work is applicable to many types of issues (e.g., regression, text, or image production).
- Finally, in order to examine the framework's usefulness in practice, we would prefer to concentrate on particular application domains (such as finance and social network research) rather than general-purpose datasets. However, the shortcoming in handling big data can be resolved with the adoption of hybridized deep learning approaches.

REFERENCES

[1] Hochstein, D. Rangarajan, N. Mehta, and D. Kocher, ‘‘An industry/academic perspective on customer success management,’’ J. Service Res., vol. 23, no. 1, pp. 3–7, 2020.

[2] Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, ‘‘Machinery health prognostics: A systematic review from data acquisition to RUL prediction,’’ Mech. Syst. Signal Process., vol. 104, pp. 799–834, May 2018.

[3] Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, ‘‘Applications of machine learning to machine fault diagnosis: A review and roadmap,’’ Mech. Syst. Signal Process., vol. 138, Apr. 2020, Art. no. 106587.

[4] Sipos, D. Fradkin, F. Moerchen, and Z. Wang, ‘‘Log-based predictive maintenance,’’ in Proc. 20th Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD), Aug. 2014, pp. 1867–1876.

[5] Calabrese, M. Cimmino, F. Fiume, M. Manfrin, L. Romeo, S. Ceccacci, M. Paolanti, G. Toscano, G. Ciandrini, A. Carrotta, M. Mengoni, E. Frontoni, and D. Kapetis, ‘‘SOPHIA: An event-based IoT and machine learning architecture for predictive maintenance in industry 4.0,’’ Information, vol. 11, no. 4, p. 202, Apr. 2020.

[6] Firouzi F, Farahani B, Marinšek A. The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT). Inf Syst. 2022;107:101840.

[7] Strouse D, McKee K, Botvinick M, Hughes E, Everett R. Collaborating with humans without human data. Adv Neural Inf Process Syst. 2021;34:14502–15.

[8] Zhou X, Chai C, Li G, Sun J. Database meets artificial intelligence: a survey. IEEE Trans Knowl Data Eng. 2022;34(3):1096–116.

[9] Jiao L, Zhang R, Liu F, Yang S, Hou B, Li L, Tang X. New generation deep learning for video object detection: a survey. IEEE Trans Neural Netw Learn Syst. 2022;33(8):3195–215.

[10] Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng. 2022;34(1):50– 70.

[11] Li Z, Li S, Bamasag OO, Alhothali A, Luo X. Diversified regularization enhanced training for effective manipulator calibration. IEEE Trans Neural Netw Learn Syst. 2023;34(11):8778

[12] Bharati S, Mondal MRH, Podder P. A review on eXplainable Artificial Intelligence for healthcare: why, how, and when? IEEE Trans Artif Intell. 2023.

[13] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Guidotti R, Del Ser J, Díaz-Rodríguez N, Herrera F. eXplainable Artificial Intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion. 2023;99:101805

[14] Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N. A survey on XAI and natural language explanations. Inf Process Manage. 2023;60(1):103111.

[15] Viswan V, Shaffi N, Mahmud M, Subramanian K, Hajamohideen F. eXplainable Artificial Intelligence in Alzheimer’s disease classification: a systematic review. Cogn Comput. 2023

[16] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 1135–44

[17] Schwalbe G, Finzel B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min Knowl Disc. 2023

[18] Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, Ranjan R. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv. 2023

[19] Gunning D, Aha D. DARPA’s Explainable Artificial Intelligence (XAI) Program. AI Mag. 2019;40(2):44–58

[20] Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. Benchmarking and survey of explanation methods for black box models. Data Min Knowl Disc. 2023

[21] Gunning D, Aha D. DARPA’s Explainable Artificial Intelligence (XAI) Program. AI Mag. 2019;40(2):44–58

[22] Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local rule-based explanations of black box decision systems. arXiv:1805.10820 [Preprint]. 2018

[23] La Gatta V, Moscato V, Postiglione M, Sperlì G. CASTLE: ClusterAided Space Transformation for Local Explanations. Expert Syst Appl. 2021;179:115045

[24] Rjoub G, Bentahar J, Abdel Wahab O, Mizouni R, Song A, Cohen R, Otrok H, Mourad A. A survey on eXplainable Artificial Intelligence for cybersecurity. IEEE Trans Netw Serv Manage. 2023;20(4):5115–40.

[25] Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif Intell Rev. 2023;56(6):5261–315.

[26] Lamy J-B, Sekar B, Guezennec G, Bouaud J, Séroussi B. eXplainable Artificial Intelligence for breast cancer: a visual case-based reasoning approach. Artif Intell Med. 2019;94:42–53.

[27] Chen L, Gao Y, Zheng B, Jensen CS, Yang H, Yang K. Pivotbased metric indexing. Proc VLDB Endow. 2017;10(10):1058–69.

[28] Guidotti R, Monreale A, Giannotti F, Pedreschi D, Ruggieri S, Turini F. Factual and counterfactual explanations for black box decision making. IEEE Intell Syst. 2019

[29] Jia Y, Bailey J, Ramamohanarao K, Leckie C, Ma X. Exploiting patterns to explain individual predictions. Knowl Inf Syst. 2019.

[30] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17). Red Hook: Curran Associates Inc.; 2017. pp. 4768–77.