Received: December 15, 2019

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Proposed Method for Minimizing Mining Tasks' Data Dimensionality

Amira M. Idrees[1]* **Wael H. Gomaa[2]**

[1]*Faculty of Computers and Information, Fayoum University, Egypt;*
[2]*Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt*
* Corresponding author's Email: ami04@fayoum.edu.eg

Abstract: Knowledge discovery techniques have heavily contributed to many fields with significant success. However, with the continuous growth of data, these techniques suffer from bottlenecks in processing these data. One of the directions to hinder this effect is reducing the data dimensionality which focuses on eliminating the attributes which have no significant effect on the discovery technique accuracy. This research proposes a novel method for reducing data dimensionality. The proposed method is based on two main pillars, the first is applying the adapted Saaty method for determining the attributes' consistency with proposing further adaptation targeting more accurate accuracy determination. The second pillar is applying the clustering techniques on the consistent attributes to eliminate the least weighted attributes in each cluster which also have the least consistent measures. The result of applying the two steps is to highlight the most significant dataset attributes. The proposed method has been successfully applied on the Gastrology dataset which attributes have been reduced from 62 attributes to 31 attributes. A set of classification techniques have been applied on the dataset to prove that the dimensionality reduction has retained the classification task accuracy, the results presented that the Random Forest algorithm had an accuracy equal 95.36% when applied to the adopted dataset.

Keywords: Data dimensionality, Weighting techniques, Saaty method, Clustering data mining, K-means clustering algorithm, Classification data mining.

1. Introduction

The Knowledge discovery field is currently one of the pillars for almost all fields. The flood of methods and techniques that apply mining techniques is continuously growing. These methods' mission is manipulating the data striving for knowledge [1]. As the data complexity keeps growing exponentially, therefore, higher efficient methods are continuously required. One of the data complexity parameters is the dimensionality of data which is getting higher [2]. Although the high dimensionality of data provides a clearer view of the data which contributes to raising the results' accuracy, however, the limit for the required dimensions is currently one of the directions under focus. The challenge of reducing data dimensionality is balancing between reducing the process complexity by reducing the high data dimensionality with retaining the required accuracy target [3].

Different fields suffer from the increase in the data volume with the high complexity and very large dimensions set. For example, in medical field [4-5], the diversity of data arises for the extensive description of the provided cases. While this description is considered mandatory for the experts in the medical field, the provided knowledge discovery methods and techniques can contribute efficiently in reducing these dimensions with high accuracy. Although it may appear that high dimensions provide more accuracy, however, it is not always the case. Situations may arise in which the high dimensionality leads to lower accuracy due to the diversity level of the data as well as its weak structure which consequently leads to losing the data identity [6].

Different techniques have been proposed for reducing the data dimensions by applying different

approaches, however, this study argues that mining techniques can provide an efficient contribution in this target. It is considered a dilemma in applying mining techniques for enhancing the effectiveness in operating and in the target for mining techniques, nevertheless, this situation provides the power of these techniques which smoothly, efficiently, and essentially contribute for all the system components including input, process, and output. Mining techniques can be merged coherently with other techniques forming a homogenous collaborative and; most important; effective environment [7]. In this study, a homogenous environment between statistical techniques and mining techniques is proposed targeting to detect the lowest set of attributes that represent the data with maintaining the data identity. The proposed study highlights the effective collaboration of these techniques through two main framework stages. The first stage includes applying the Saaty method with a proposed adaptation for more accurate determination to the attributes' consistency. The attribute consistency percentage highlights the contribution degree of this attribute to the decision discrimination. Therefore, detecting the attributes' consistency is one of the main contributing steps in highlighting the inconsistent attributes for elimination. The research argues that determining the attributes' weighting using one defined technique could lead to a biased result according to both the technique and the attributes' nature. The proposed research considers the attribute's weight is a main criterion for determining the attribute's consistency. Therefore, the proposed research includes a collaboration between different weighting techniques to determine the attributes' weighting which lead to more accurate weighting determination and consequently, more accurate detection for the attributes' consistency. The second stage of the proposed framework includes applying the clustering data mining techniques which explores the intra-relation between the attributes. These relations contribute in supporting the further attributes' elimination process targeting to the final dimensionality reduction with maintain the data accuracy.

The main strong points of the proposed research can be summarized in the following points:

The capability of utilizing the weighting techniques results in determining the attributes' consistency the research proposed an adaptation to Saaty method which depends on a set of λ values; instead of only λ_{max}; for more accurate determination of the attributes' consistency.

The research proposed that the λ values could be determined according to the weighting techniques results which highlights that the statistical techniques and machine learning techniques can be considered a two-sides coin.

Despite that the utilized techniques; Adapted Saaty method, Weighting techniques, and Clustering mining technique; have different nature, however, this situation should not hinder their effective collaboration. the proposed research approach succeeded in utilizing these techniques in a collaborative environment to reach the required dimensionality reduction target.

The proposed method could be applied on the dataset as a pre-processing step for the mining tasks in order to reduce the processing complexity and maintaining the results' accuracy with a more contribution to raise the accuracy level. Applying the proposed method for classification task on the Gastrology dataset provided high accuracy classification results which is above 95% as discussed in section 4. This result confirmed the positive contribution of the proposed method as it reached the high accuracy with the minimized dimensionality and less processing cost.

The remaining of the study describes the related work of reducing data dimensionality in section 2 and the proposed method is discussed in section 3 with highlighting the main method aspects in a formal representation. Moreover, section 4 discusses the details of applying the proposed method on a dataset and the accuracy is determined and presented, and finally, section 5 presents the conclusion with the main research highlights and the future research directions.

2. Related work

Different approaches have contributed to the field of attributes' reduction [8]. These approaches follow three directions, they are the "greedy-based, reverse- influence, and heuristics" [9]. Although the first approach is for the optimal solution, however, the extensive processing in finding the solution hinder the main aim of the reduction approach. The reverse sampling approach apparently is less complex as one of its vital parameters is the representative sample, however, according to [9], the approach complexity arises due to the large amounts of the generated samples. Finally, heuristics rely on selected seeds for starting the process which is a bottleneck in the process as seeds may belong to near or even the same zone.

Different researches followed the heuristics

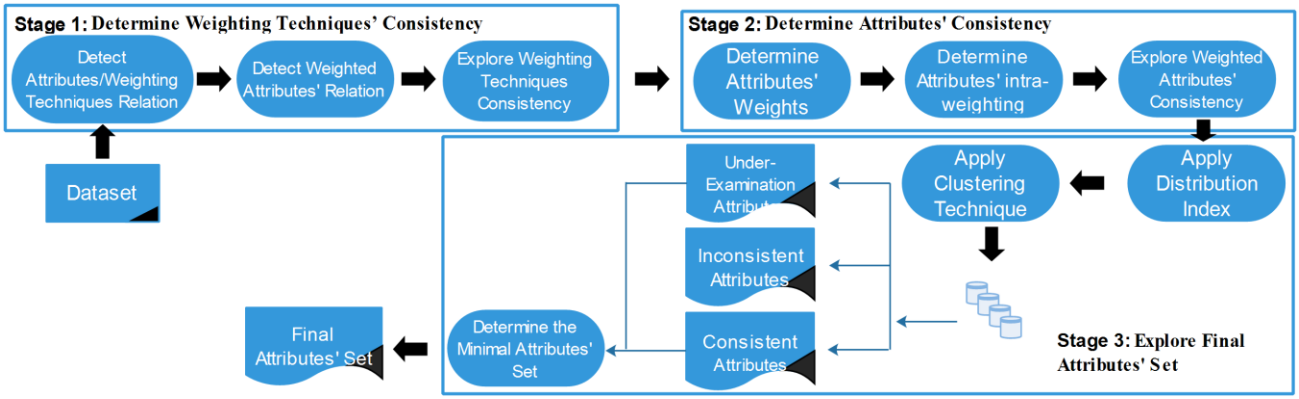

Figure. 1 The proposed method stages

approach, however, the methods proposed in these researches did not consider the attributes' inconsistency which hinders the accuracy level for the proposed methods. Moreover, data mining has contributed in different researches such as in [10], the research applied PCA for the reduction task without considering its consistency for the provided data, it was simply assumed its adequacy according to the literature review. Dimensionality reduction has been applied in different fields with a variety of perspectives. In [11], the distance measuring approach has been followed for the data reduction approach, however, the presented approach followed the grouping perspective to ensure that dimensions are smaller than the data samples. Although the current research follows the same grouping perspective but with the target of maximizing the reduction process efficiency in providing a minimal set of dimensions. The current approach has no restriction on the samples-dimensions percentage. Moreover, the presented research in [11] followed the distance perspective as a basic stage which is replaced in the current method by a proposed stage which highlights the collaborative approach in different weighting techniques targeting the coherent approaches' integration.

Another research in [12] presented a different perspective in applying data reduction for building knowledge, the research basic factor was the successful ontology determination for clustering guidance. The target of the research is promising; however, the proposed research was limited to a maximum of 20 attributes dataset with highlighting a performance issue. A recent research in [13] highlights the contribution of reducing data dimensionality for gaining higher reliability due to the eliminating the data heterogeneity. The proposed method was based on applying the Bayesian approach while neglected the attributes relationships which was stated as one of its future directions. In the current research, the attributes' relationships are the main consideration which was determined in the weighting relation between attributes. Most of the proposed researches in the field contribute in one of the reduction directions with illustrating the significant enhancement, however, in the current research, a collaborative environment between statistical approaches and mining approaches is proposed for illustrating the attributes' distribution targeting for the most suitable level of accuracy in the reduction process.

Focusing on the main utilized approaches in the current research, weighting techniques, Saaty method, and clustering. Weighting techniques have a significant contribution in features' extraction task. The research in [3] , the concept of focusing on the most effective attributes has been considered through determining the attributes' correlations which were effective, yet other more effective approaches would be considered. While in [14], the attributes' determination considered identifying the news' credibility with identifying each considered attribute, the research had no clarification for other non-considered attributes as the focus was for proving the credibility target. Other researches such as in [15] and [16] considered different fields but with the scope of identifying the effective attributes with following the attributes' semantic relations with no consideration of the individual impact for these attributes.

Moreover, clustering techniques have been extensively considered in different tasks, however, the approach of transposing the data for attributes' clustering with the aim of reducing dimensionality has not been a traditional focus.

3. The proposed method

This research aims at reducing data dimensions by eliminating the attributes that have a minimal contribution to the mining tasks. Fig. 1 illustrates the main stages in the proposed method.

Furthermore, the main research significance can be summarized in the following points:

1. As each of the weighting techniques has its main strengths and weaknesses points [17]. Therefore, the proposed method adapts applying different weighting techniques by proposing a collaborative approach for these techniques in determining the attributes' significance.
2. Evaluating the weighting techniques is performed by applying a proposed adaptation to the Saaty method for detecting the weighting technique's consistency [18].
3. Detecting the attributes' consistency is then performed using the same approach for eliminating the inconsistent attributes as the first elimination stage.
4. Considering the consistent weighting techniques, the consistent attributes' significant weighting is performed with respect to the weighting technique's consistent ratio [19].
5. The second elimination stage considered applying the clustering techniques for determining the relations between attributes and detect the highest weighted attributes in each cluster as the significant attributes [20].
6. The required method output is represented as the final set of attributes which includes the detected attributes from the clustering stage

The following steps illustrate the algorithm for the whole method stages while the following sections discuss each stage.

// Determine Weighting Techniques' Consistency

// Build Attributes/Weighting Techniques Matrix (ATT-WT)

i=1, j=1, labelD = ϕ
set WT = wt_i
repeat {
set ATT = att_j
repeat {
set label = ATT, ATT_set ATT ∩ label
apply wt_i (label)
set (AttWt (j,i))
ATT-WT (i) = ATT-WT (i) ∪ AttWt (j,i)
labelD = labelD ∪ label
next j
} until labelD = ATT
Next i
} until i=n
AttWt (j,i) = [wt_{xj},…..wt_{yj}]

// Build Weighting Techniques - Attributes' Comparison Matrix WC(i)

z=1, v=1
Repeat {
Set WT = wt_z,
Repeat {
Repeat {
x=1
if (v = x) → set wc (z) [v,x] = 1, wc (z) [x,v] = 1
else set wc (z) [v,x] = Nwt_{vx}, wc (z) [x,v] = 1/Nwt_{vx}
x++
} until x = j
V++
} until v = j
z++
} until z = i

// Build the Consistent Weighting Techniques' Set (CWT)

Set WeightingTech = | WT |
If WeightingTech >15 {
X=16
Repeat from x to WeightingTech
C_x = Forecasting (x, (1-15), (c_1-c_{15}))
}
z=1
Repeat {
Set wt = wt_z,
Set-vector $Product_z$ (wt)
Set-vector $EigenValue_z$ (wt)
Set-vector λ_z (wt)
Set-vector Factors VF_z (wt)
For each element i in VF_z {
If vf_z (i) <= 0.1
Set Cvf_z (i) = T
Else Set Cvf_z (i) = F
}
If Countif (Cvf_z (i) = T) >= 6
Set Cwt_z (wt) = T
Add wt_z to CWT
Else Set Cwt_z (wt) = F
z++
} until z = i

// Stage 2: Determine Attributes' Consistency

// Build Weighted Attributes' Matrix (WATT)

i=1, j=1, labelD = ϕ
set WT = wt_i
repeat {
set ATT = att_j
repeat {
set label = ATT, ATT_set ATT ∩ label
apply wt_i (label)
set (AttWt (j,i))
ATT-WT (i) = ATT-WT (i) ∪ AttWt (j,i)

labelD = labelD ∪ label
next j
} until labelD = ATT
Next i
} until i=n

AttWt (j,i) = [wt_{xj},…..wt_{yj}]
WATT (j) = [<wt_1, $watt_{j1}$>,…..<wt_i, $watt_{ji}$>]
// **Build Weighted Attributes' Comparison Matrix *ATTCM(i)***
∀ x, y att(x), att(y)→ AttCM(i) (x,y) = mean (x,y), AttCM(i) (y,x) = 1/mean (x,y)
// **Determine Attributes' Consistency Level Set (ATTC))**
Set attributes = | ATT |
If ATT >15 {
X=16
Repeat from x to ATT
C_x = Forecasting (x, (1-15), (c_1-c_{15}))
}
z=1
Repeat {
Set att = att_z,
Set-vector $Product_z$ (att)
Set-vector $EigenValue_z$ (att)
Set-vector λ_z (att)
Set-vector Factors VF_z (att)
For each element i in VF_z {
If vf_z (i) <= 0.1
Set Cvf_z (i) = T
Else Set Cvf_z (i) = F
}
If Countif (Cvf_z (i) = T) >= 6
Set $Catt_z$ (att) = T
Add att_z to CATT
Else Set $Catt_z$ (att) = F
z++
} until z = i
// **Stage 3: Determine Final Attributes' Set**
// Apply Silhouette index
Determine the number of clusters
//Build CLUSTATT Set
Apply Enhanced K-means Clustering Technique on CATT
Set CLUSTATT = {<att_x,…att_y>, …..
<att_g,….att_r>} where att_x, att_y, att_g, att_g ∈ CATT
//Determine the Minimal Attributes' Set
∀ c, c ∈ clusters
C-Count = |c|, added = 0
∀ x, x ∈ ATT, x ∈ c,
Countif (Cvf_z (i) = T) = |VF| → add x to FinalAtt set
CAttRem(c) = c – {x}, added++
j ∈ WT, set Avg (x) = ∑ AttWt (x, j) / |WT|
if Avg (x) = Max_average→ add x to FinalAtt set
CAttRem (c) = c – {x}, added++
Sort attributes according to both consistency λ measures count and the average weight
∀ x, x ∈ ATT, x ∈ CAttRem (c)
Add the first (num < (C-Count/2 – added)) elements in c to the FinalAtt set

3.1 Data illustration

For clarifying the method stages, the main method stakeholders could be formed in a group of sets representing the whole system illustration. The simple system illustration includes:

Input: weighting techniques and data attributes which are defined in Eqs. (1) and (2) respectively

Process: weighting, consistency, and membership

Output: the required minimal set of attributes

– The set of weighting techniques (WT):

$$WT = \{wt_1, wt_2, \ldots wt_n \mid n \in \mathbb{N}\} \quad (1)$$

– The set of attributes (ATT):

$$ATT = \{att_1, att_2, \ldots att_m \mid m \in \mathbb{N}\} \quad (2)$$

– The set of attributes with their associated weight with respect to a defined label attribute and a defined weighting technique (AttWt (j,i)) is represented in Eq. (3):

$$AttWt(j,i) = \{\langle att_x, wt_x\rangle \mid x, j \in \{1,2,\ldots m\}, i \in \{1,2,\ldots n\}, x \neq j\} \quad (3)$$

– The weighting technique i attributes' matrix (ATT-WT (i)) is represented in Eq. (4):

$$ATT\text{-}WT(i) = \bigcup_{k=1}^{j} AttWt(j,i) \quad (4)$$

– The set of all weighting techniques' attributes' ATT-WT matrix is represented in Eq. (5):

$$ATT\text{-}WT = \bigcup_{k=1}^{i} ATT\text{-}WT(i) \quad (5)$$

– The set of consistent weighting techniques (WTC) is represented in Eq. (6):

$$WTC = \{wt_x, \ldots wt_y \mid x,y \in \{1,2,\ldots n\}\} \quad (6)$$

– The set of consistent attributes (ATTC) is represented in Eq. (7):

$$ATTC = \{att_f, \ldots att_k \mid f,k \in \{1,2,\ldots m\}\} \quad (7)$$

- The set of consistent attributes associated with their average weight (ATTCW) is represented in Eq. (8):

$$ATTCW = \{<att_f, w_f>\ldots.<att_k\ w_k> \mid f,k \in \{1,2,\ldots.m\}\} \quad (8)$$

- The set of attributes' clusters (ClustAtt) is represented in Eq. (9):

$$ClustAtt = \{ <att_e,\ldots att_q>,\ldots., <att_a,\ldots att_s> \mid e,q,a,s \in \{1,2,\ldots.m\}\} \quad (9)$$

- The minimal set of attributes (FinalAtt) is represented in Eq (10):

$$FinalAtt = \{att_g,\ldots.att_d \mid g,d \in \{1,2,\ldots.m\}\} \quad (10)$$

3.2 Stage 1: Determine weighting techniques' consistency

The research argues that the collaborative approach for considering different weighting techniques is one of the vital perspectives which leads to more accurate evaluation for weighting attributes [21]. However, some of these techniques may not be compatible with the data nature, which consequently may be misleading. Therefore, this stage aims at determining the weighting techniques' consistency targeting to eliminate the results of the inconsistent weighting techniques. The main deliverable of this stage is the set of weighting techniques which should be considered to determine the attributes' weights associated with these techniques' consistency ratio. The weighting techniques' consistency ratio is considered one of the significant factors in the dimensionality minimization process. In this research, the system utilizes the enhanced Saaty method in [18] with a proposed tuning targeting higher accuracy as will be discussed in the following sections.

3.2.1. Detect attributes/weighting techniques relation (build attributes/weighting techniques matrix (ATT-WT))

This step deliverable is a matrix representation that describe the conjunction between each attribute as a label with each weighting technique with representing the impact of other attributes on the label attribute with respect to the defined weighting technique. The ATT-WT matrix for *i* is defined in Eq. (11):

$$\text{ATT-WT (i)} = \begin{bmatrix} wt_{x1} & \cdots & wt_{y1} \\ \vdots & \ddots & \vdots \\ wt_{xj} & \cdots & wt_{yj} \end{bmatrix}$$

where $x \neq y$, $x,y \neq j$ (11)

3.2.2. Detect weighted attributes relation (build weighting techniques - attributes' comparison matrix *WC(i))*

The deliverable of this step is a matrix for each weighting technique *(WC(i))*. The matrix of the weighting technique *z* illustrates the impact of attribute *x* on the attribute *y* represented in the determined weight of *x* by *z* when *y* was set to be a label attribute. Building the weighting technique matrix is performed as follows:

First, after applying the weighting technique *z* on the data with setting *y* as a label, each attribute *x* will have an associated weight wt_y. then normalizing these weights is performed aiming to transform the weights into a representative preference. The range for the normalization process is 1 for the weight from 0.0 to 0.33, 2 for the weight from 0.33 to 0.66, 3 for the weight from 0.66 to 1, and 4 for the weight which is higher than 1. Finally, the matrix *(WC(i))* is constructed by setting the element wc_{xy} is the impact of *x* on *y* = the normalized weight *Nwt(x,y)*, then wc_{yx} will be equal *1/Nwt*. The comparison matrix for the weighting technique *i* representation is defined in Eq. (12):

$$\boldsymbol{WC_i}\text{ (i)} = \begin{bmatrix} Nwt_{11} & 1/Nwt_{12} & \cdots & \mathbf{1}/Nwt_{1j} \\ \vdots & & & \\ Nwt_{1x} & Nwt_{2x} & \ddots & \vdots \\ Nwt_{1j} & Nwt_{2j} & \cdots & Nwt_{jj} \end{bmatrix}$$

where $x \neq y$, x, $y \neq j$ (12)

3.2.3. Explore weighting techniques consistency (build the consistent weighting techniques' set (CWT))

The deliverable of this step is a decision for each weighting technique to be either consistent or not consistent. Following the consistency determination method proposed in [18] which presents an adapted Saaty method for determining customers' requirements, however, the current research proposes an adaptation for the method presented in [18] based on the variation in the data nature in both researches. Although the research in [18] neglected the consideration of λ_{min} as it is considered the least important to the customers' requirements and would lead to less effective consideration to these requirements, however, the current research considers determining the consistent weighting

techniques as well as the consistent attributes targeting to eliminate the inconsistent weighting techniques and attributes' sets. These targets require the determination of all the attributes' boundaries; therefore, it is vital to consider the boundaries' aspects which provide more accurate examination to the required consistency. These aspects including the minimum, maximum, mean, median, range, upper inter-quartile range, lower inter-quartile range, mean of upper inter-quartile range and the mean value, and mean of lower inter-quartile range and the mean value. The primary condition to consider the weighting technique/attribute as consistent is to have a consistency ratio below the threshold; which is equal 0.19 according to [22]; for at least six of these aspects. It is worth mentioning that the constant used in determining the consistency ratio follows the provided constants distribution in [23], however, Olson in [23] provides constants to 15 elements only which limits the number of attributes. Therefore, a forecasting task is applied for providing further distribution to cover a higher number of elements which represents a larger set of attributes.
Finally, the deliverable of stage 1 is a set of consistent weighting techniques which is defined in Eq. (13). As a result of the consistency, these techniques will contribute in the attributes' weighting.

$$CWT = \{wt_a, \dots wt_b\} \text{ where } CWT \subseteq WT \text{ , } Cwt_a(wt) = Cwt_b(wt) = T \quad (13)$$

3.3 Stage 2: Determine attributes' consistency

Stage 2 is the basic stage for dimensionality reduction target which deliver the final set of attributes for being considered in the mining tasks. Stage 2 starts with determining the features' weighting by each of the consistent weighting techniques that are members in CWT set. Stage 2 follows the same procedure in stage 1 with slightly alternates. The following subsections highlight the differences in the applied process.

3.3.1. Determine attributes' weight (build weighted attributes' matrix (WATT))

Following the same process in Sect. 3.2.1, the weighted attributes' matrices are developed for all attributes. Therefore, each attribute *i* has an associated weight set *Watt (i)*. *Watt (i)* members are vectors of two elements; weighting technique *j* and associated attribute *i* weight.

This step deliverable is a matrix representation that describe the conjunction between each attribute as a label with each weighting technique with representing the impact of other attributes on the label attribute with respect to the defined weighting technique. The ATT-WT matrix for *i* is defined in Eq. (14):

$$WATT(j) = [<wt_1, watt_{j1}>, \dots <wt_i, watt_{ji}>] \quad (14)$$

3.3.2. Determine attributes' intra-weighting (build weighted attributes' comparison matrix *ATTCM(i))*

The deliverable of this step is a matrix for each attribute *(AttCM(j))*. The elements of the matrix ***ATTCM(i)*** are deduced by illustrating the mean of each two weighting measures x, y as a relation, then AttCM(i) (x,y) = mean (x,y) and AttCM(i) (y,x) = 1/mean (x,y). Finally, values normalization is applied following the same rang that is clarified in section 3.2.2. The comparison matrix for the attribute *j* representation is defined in Eq. (15):

$$\boldsymbol{ATTCM}(i) = \begin{bmatrix} Nwatt(i)_{11} & 1/Nwatt(i)_{12} & \cdots & 1/Nwatt(i)_{1j} \\ \vdots & & & \\ Nwatt(i)_{1x} & Nwatt(i)_{2x} & \ddots & \vdots \\ Nwatt(i)_{1j} & Nwatt(i)_{2j} & \cdots & Nwatt(i)_{jj} \end{bmatrix}$$
$$\text{where } x \neq y,\ x, y \neq j \quad (15)$$

3.3.3. Explore weighted attributes' consistency level (determine attributes' consistency level set (ATTC))

The deliverable of this step is a decision for each attribute to be either consistent or not. The decision for each attribute is deduced following the same perspective in section 3.2.3 in adapting the proposed method by [18]. As mentioned earlier, despite the research in [18] had the acceptable justification to only consider three distributions of λ, however, following the same approach strictly would hinder alternative perspective of the attribute distribution according to the difference in the main environment between this research and [18]. Therefore, the authors claim that considering other distributions provide more illustration for the attributes which consequently leads to more accurate decision for the attributes' consistency. considering different attribute's distributions highlight a clear illustration for the attribute consistency. These perspectives including calculating the minimum, maximum, mean, median, range, upper inter-quartile range, lower inter-quartile range, mean of upper inter-quartile range and the mean value, and mean of lower inter-quartile range and the mean value of λ. Finally, the same consistency threshold is below or equal 0.19. Finally, the deliverable of stage 1 is a set

of consistent attributes associated with their average weight (defined in Eq. (16)) which was calculated through the consistent weighting techniques. As a result of the consistency, these attributes will contribute in the mining tasks.

$$CATT = \{<att_a, w_a>,\ldots <att_b, w_b>\} \quad \text{where } CATT \subseteq ATT \tag{16}$$

3.4 Stage 3: Determine final attributes' set

As the research aim is to detect the most effective attributes for reducing the data dimensionality, therefore, an additional stage is proposed. In stage 3, the research argues that clustering approach can well contributes in reducing the data dimensionality. As cluster members are described by their homogenous relationship which illustrates their similar behavior, therefore, applying clustering approach on the dataset after eliminating inconsistent attributes leads to a set of clusters, each consistent attribute belongs to one or more of these clusters.

3.4.1. Apply the distribution index

One of the critical steps is determining the most suitable number of clusters for the best attributes' distribution. Therefore, the first step is applying a distribution index to determine the suitable number of clusters. the most suitable number of clusters is determined according to the results of applying Silhouette index [24].

3.4.2. Determine the attributes' clusters (build CLUSTATT Set)

In this step, Applying the clustering technique is performed to determine the attributes clusters' set (CLUSTATT). It includes a set of vectors; each vector represents one of the clusters' members as in Eq. (17):

$$CLUSTATT = \{<att_x,\ldots att_y>, \ldots.. <att_g,\ldots.att_r>\} \quad \text{where } att_x, att_y, att_g, att_g \in ATT \tag{17}$$

In each cluster, a subset of the attributes is selected as representative attributes for the class members. the strategy of selecting the representative attributes in each cluster follows the following conditions:

The attributes that are consistent for all λ measures are included in the final attributes' set.

The attribute that has the highest average weight calculated from the consistent weighting measures is included in the final attributes' set.

The attributes that have the lowest average weight calculated from the consistent weighting measures with the lowest number of consistent λ measures (6 measures) are excluded from the final attributes' set.

The remaining attributes in each cluster moves to the next step for more examination.

3.4.3. Determine the minimal attributes' set

The remaining attributes are ordered according to its consistent λ measures count and the average weight which is calculated by the consistent weighting measures.

The highest rank of these attributes representing 50% of them are determined to be included in the final attributes' set. The percentage is determined to ensure that more than 50% of each cluster's attributes are selected to avoid data characterization loss

4. Experimental case study

The experimental study includes applying the proposed method on a real dataset for proving the proposed method effectiveness. The case study in this research applied the proposed method on a Gastrology dataset. Gastrology dataset includes data for 950 records which are represented by 62 attributes covering 11 Gastrology types. For more knowledge, the Gastrology types and the associated analysis is presented in [25]. Table 1 includes the set of diseases as well as the set of attributes of the dataset. Moreover, following the proposed method

Table 1. The set of gastrology types and attributes

Gastrology Types
Toxic Hepatitis, Cholecystitis, Hepatitis C, Hepatitis A, Peptic ulcers, Obstructive jaundice, Hepatitis B, Cirrhosis, Gilbert, Hepatitis, Fatty liver, Normal
Attributes
Age, Sex, Urine Color, Aspect, Reaction Urine, Specific Gravity, Albumin, Sugar, Acetone, Bile Salt, Bilirubin, Urobilinogen, Pus, Rbcs, Epithelial Cells, Crystals, Amorphous Elements, Casts, Bilharzial Ova, Consistency, Odor, Mucus, Blood, White Cells , Red Cells, Fats, Starch, Muscle fibers, Ova and/or Parasites, Cyst, Haemoglobin, Haematocrit (P.C.V.), M.C.V., M.C.H., Platelets, Total Leukocyte count, Basophils, Eosinophils, Segmented, Lymphocytes, Monocytes, First hour, Second hour, C.R.P, Total bilirubin, Direct bilirubin, Alkaline phosphatase, SGPT, SGOT, Total protein, Albumin liver, A/G Ratio

Table 2. λ measures for information gain weighting technique

		λ	C.I	C.R	C/NC
λ max		3.189	0.094	0.163	NC
λ median		3.125	0.062	0.108	C
λ mean		3.136	0.068	0.117	NC
λ min		3.095	0.047	0.082	C
range		3.000	0.000	0.000	C
Standard Deviation		0.048			
Inter Quartile Range	**UpperQ**	3.280	0.140	0.241	NC
	LowerQ	2.992	0.004	0.007	C
Mean (Range,UpperQ)		3.140	0.070	0.121	NC
Mean (Range,LowerQ)		2.996	0.002	0.004	C
Mean (Range, Mean)		3.047	0.024	0.041	C

Table 3. Weighting techniques' consistency status

PCA	IG	IGR	SVM	GI	CS	DEV	CRL
NC	C	C	C	NC	NC	C	C

stages, a part of the results is illustrated which provides a clarification for the experimental results.

Stage 1 considers a set of weighting techniques, the experimental study considered eight of the weighting techniques, they are Principle Component Analysis (PCA), Information Gain (IG), Information Gain Ratio (IGR), Support Vector Machine (SVM), Gini Index (GI), Chi-Square (CS), Deviation (DEV), and Correlation (CRL). It is a fact that each technique has its nature and compatibility with the dataset characteristics, therefore, the previously described stage has been applied to determine the weighting technique consistency. The information gain technique has been highlighted as an example in Tables 2 and 3.

Table 2 illustrates the results for the required λ determents which provide an individual decision of the technique's consistency status. The results show that information gain has been considered consistent for 7 out of the 10 measures, therefore, it is considered a consistent weighting measure. Then, Table 3 illustrates the consistency status for all the eight weighting techniques as C refers to consistent while NC refers to inconsistent. The results show that the consistent weighting techniques which will be considered to determine the features' weight are information gain, information gain ratio, support vector machine, deviation, and correlation techniques.

After determining the consistent weighting measures, the following stage is to determine the consistent features. The required steps are applied on the dataset attributes, Bilirubin attribute results has been highlighted as an example for the attributes' set. The weighting for the Bilirubin attribute has been determined by each of the consistent techniques. The results have been transposed to virtually consider each of the weighting techniques as an attribute. Then, the percentages' matrix is constructed by several iterations. In each iteration, one of the weighting techniques is considered the reference and the relatedness between the reference technique and other techniques. Tables 4 and 5 illustrate the results which eventually measure the Bilirubin attribute consistency. Table 4 present the percentages' matrix, the product values (P), the matrix of judgment, and the Eigenvector values (EV). Then, Table 5 presents the λ determents which provide an individual decision of the attribute's consistency status. The results show that Bilirubin has been considered consistent for 6 out of the 10 measures, therefore, it is considered a consistent attribute.

After Applying the same process to all attributes, 45 out of the 62 attributes are considered consistent. The following set of attributes are considered consistent and will be considered in the following stage, while the remaining attributes are eliminated as they are determined to be inconsistent.

Consistent Attributes Set = {Red Cells, Fats, Staff, Vegetables, Starch, Muscle fibers, White Cells, Red Cells Two, White Cells Two, Ova and/or Parasites, Blood, Consistency, Colour, Odour, Mucus, Amorphous Elements, Crystals, Epithelial Cells, Moncoytes, A/G Ratio, Albumin, Volume, Casts, Bilharzial Ova, Reaction Urine, Colour urine, Aspect, Sugar, Acetone, Bile Salt, Bilirubin, Urobilinogen, Specific Gravity, Age, Total Leucocytic count, Eosinophils, Lymphocytes, M.C.H., Segmented, M.C.V., Haematocrit (P.C.V.), Basophils, Red Cell Count, Total Prot, Haemo}.

The next phase targets to further minimize the attributes' set by considering clustering the attributes. According to the proposed approach, only "*Haemo*" attribute is considered in the final attributes set as it is accepted by all the **λ** measures. Moreover, *"Total Prot, Red Cell Count, Basophils, Haematocrit (P.C.V.), M.C.V., Segmented*, Albumin, and *M.C.H."* attributes have the highest weight ranging from 0.8 to 0.7, these attributes are also considered elements in the final attributes' set.

According to the selected attributes, the remaining attributes are *Red Cells, Fats, Staff, Vegetables, Starch, Muscle fibers, White Cells, White Cells Two, Ova and/or Parasites, Blood, Consistency, Colour, Odour, Mucus, Amorphous Elements, Crystals, Epithelial Cells, Moncoytes, A/G Ratio, Volume, Bilharzial Ova, Reaction Urine,*

Table 4. Eigen vector for bilirubin attribute

	SVM	CRL	IGR	DEV	IG	*P*	*EV*
SVM	1.00	1.38	0.16	1.00	0.11	0.02	0.08
C	0.72	1.00	0.22	0.72	0.15	0.01	0.05
IGR	0.16	0.22	1.00	0.16	1.53	0.05	0.26
DEV	1.00	1.38	0.16	1.00	0.11	0.05	0.25
IG	9.43	0.15	0.66	0.11	1.00	0.07	0.70

Table 5. λ measures for bilirubin attribute

		L	C.I	C.R	C/NC
Lmax		6.789	0.447	0.399	NC
Lmean		4.449	0.138	0.123	C
Lmedian		5.424	0.106	0.095	C
Lmin		2.055	0.736	0.658	NC
Range		4.735	0.066	0.059	C
Standard Deviation		1.300			
Inter-Quartile range	**5.7489**	0.187	0.187	0.167	C
	3.1489	0.463	0.463	0.413	NC
Mean (rang,UpperQ)		5.242	0.061	0.061	C
Mean (rang,LowerQ)		3.942	0.265	0.265	NC
Mean (range, mean)		4.592	0.102	0.102	C

Figure. 2 Silhouette index distribution for six clusters

Colour urine, Aspect, Sugar, Acetone, Bile Salt, Bilirubin, Urobilinogen, Specific Gravity, Age, Total Leucocytic count, Eosinophils, Lymphocytes, Platelets.

The enhanced k-means Clustering technique [20] has been applied on the remaining 39 attributes. According to the Silhouette index distribution, the best distribution was for 6 clusters. Fig. 2 presents the Silhouette index distribution for the six clusters.

Following the proposed approach, eliminating the attributes with the least λ measures and least average weight from each cluster with maintaining at least 50% of the total attributes for each cluster, the following are the final attributes' set which will be further considered for characterizing the dataset for the classification task, they are Haemo , Total Prot, Red Cell Count, Basophils, Haematocrit (P.C.V.), M.C.V., Segmented, Albumin, and M.C.H, A/G Ratio, Specific Gravity , Moncoytes , Mucus , Aspect, Urobilinog , Bilirubin , Bile Salt , Colour urine , Reaction Urine , Bilharzial ova, Volume, Ova and/or Parasites , White Cell , White Cell II , Eosinophils , Epithelial cells , Crystals , Amorphous Elements, Lymphocyte , Platelets , Total Leucocytic count.

The final step is applied as an evaluation step. A set of classification algorithms are applied on the dataset after the dimensionality minimization process, and the classification evaluation results are illustrated in Table 6. The results present that the maximum classification accuracy is 95.36% by *Random Forest* algorithm which is a satisfying result as a primary stage. Further analysis will be performed for enhancing these results. For closure, the experiment presents that the proposed approach is satisfying in dimensionality reduction with maintaining the data mining techniques' accuracy. The dataset attributes have been reduced from 62 attributes to 31 attributes by eliminating the inconsistent attributes as well as clustering the attributes and eliminating the least weighted attributes in each cluster with maintaining at least 50% of each of the cluster's members.

As a final discussion, the results have been compared with previous research. In [11], the proposed dimensionality reduction approach has been evaluated through applying a set of classification techniques on different datasets. The results in [11] confirmed the random forest advancement by 95% accuracy with a slight less than the current research (95.36%). In addition, other classification techniques have provided lower accuracy percentages such as Decision Tree and support vector machine with 50% and 91% in [11] while it is 73.26% and 94.32% in the current research respectively. Moreover, in [10], a classification task has been applied on two datasets which highlighted the slight accuracy reduction

Table 6. Classification algorithms evaluation results

Classifier	Correctly Classified	Incorrectly Classified	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error
BayezNet	762 (80.21%)	188 (19.78%)	0.7516	0.037	0.1577
NaiveBayes	737 (77.57%)	213 (22.4211)	0.7183	0.0401	0.1758
Logistic	738 (77.68%)	212 (22.31%)	0.7163	0.0494	0.1622
Multilayer Perceptron	714 (75.15%)	236 (24.84%)	0.6821	0.048	0.1763
Simple Logistic	747 (78.63%)	203 (21.36%)	0.7278	0.0517	0.1602
SMO	637 (67.05%)	313 (32.94)	0.5469	0.1414	0.2601
IBK	708 (74.52%)	242 (25.47%)	0.6788	0.044	0.2047
KStar	773 (81.36%)	177 (18.63%)	0.7641	0.0318	0.1513
Attribute Selected	759 (79.84%)	191 (20.10%)	0.7495	0.0376	0.1698
Bagging	772 (81.26%)	178 (18.73%)	0.7656	0.045	0.1467
Classification via Regression	767 (80.73%)	183 (19.26%)	0.7582	0.0496	0.1505
Filtered Classifier	745 (78.42%)	205 (21.57%)	0.7299	0.0439	0.1689
Iterative Classifier Optimizer	822 (86.52%)	128 (13.47%)	0.7778	0.0419	0.1453
LogitBoost	786 (82.73%)	164 (17.26%)	0.7841	0.0413	0.1443
MultiClass Classifier	724 (76.21%)	226 (23.78%)	0.6947	0.07	0.1755
Random SubSpace	751 (79.05%)	199 (20.94%)	0.7336	0.0603	0.1589
Decision Tree	696 (73.26%)	254 (26.73%)	0.6609	0.0734	0.1784
JRIP	750 (78.94%)	200 (21.05%)	0.7354	0.0463	0.1734
PART	739 (77.78%)	221 (22.21%)	0.7255	0.0388	0.1815
Hoeffding Tree	734 (77.26%)	216 (22.73%)	0.7155	0.0399	0.1761
J48	754 (79.36%)	196 (20.63%)	0.7424	0.0379	0.1739
LMT	755 (79.47%)	195 (20.52%)	0.7443	0.0391	0.1664
RandomForest	**906 (95.36%)**	**44 (4.63%)**	**0.8092**	**0.0476**	**0.1386**
Randomtree	725 (76.31%)	255 (23.68%)	0.7058	0.0395	0.1987
RepTree	756 (79.57%)	194 (20.42%)	0.7457	0.0456	0.1615
SVM	896 (94.32%)	53 (5.68%)	0.7995	0.0465	0.1398

using the proposed technique in [10]. Moreover, in [12], a dimensionality reduction approach has been proposed based on clustering the attributes, however, the proposed approach did not consider the attributes' consistency with estimating equal contribution to all attributes which is not accurate decision. In addition, the research lacked proposing a solution for estimating the accurate number of clusters.

On the other hand, the system performance has been monitored to prove the contribution of the proposed method. As Random Forest classification technique has proved to provide the highest accuracy, therefore, applying the Random Forest classification technique performance has been monitored in two situations. The algorithm has been applied on the original dataset and on the dataset after dimensionality reduction. The algorithm has been applied on the system in a series of requests representing the system load. The performance has been compared according to the response time and the latency measures which highlighted the better performance for the updated dataset after reducing the number of attributes as shown in Figs. 3 to 6.

5. Conclusion

The data generation is a continuous process that could be illustrated as a two-faces situation. From one side, the growth of data is positively contributing to the information revolution. From the other side, the data growth has led to new requirements for maintaining the balance between using the generated data and reducing the processing effort. This research aimed at contributing to the required balance by proposing a novel method to reduce the data dimensionality in order to reduce the process effort with the same results' accuracy level. The proposed method was based on two main pillars in homogenous cooperation between intelligently

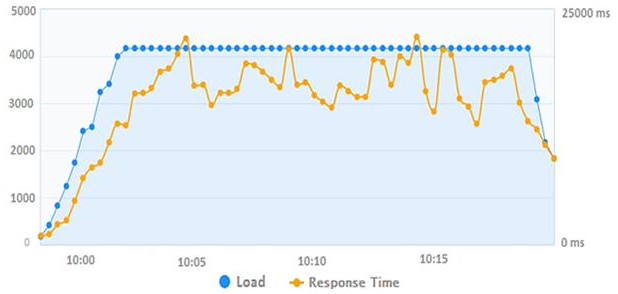

Figure. 3 Response time using the original dataset

Figure. 4 Response time using the updated dataset

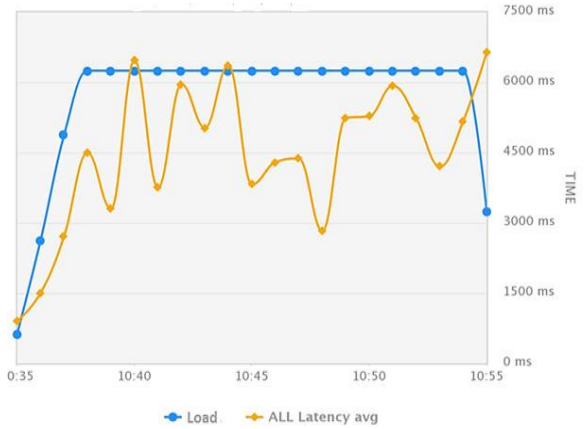

Figure. 5 Latency using the original dataset

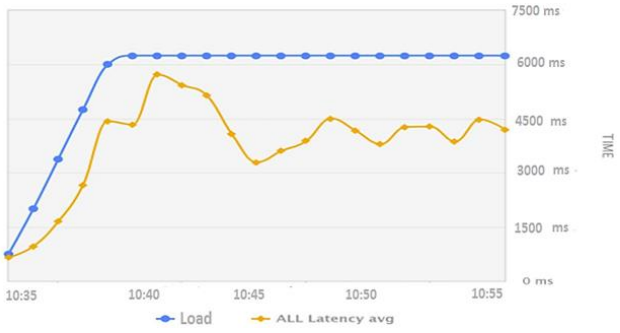

Figure. 6 Latency using the updated dataset

adapting a statistical method with applying knowledge discovery techniques. The first stage was proposing an adaptation to the Saaty method for raising the results' accuracy in determining the attributes' consistency and eliminate the inconsistent attributes. While the second stage was applying the clustering task for further intra-elimination for each cluster. The proposed method is successfully applied on the Gastrology dataset which included 62 attributes, the first stage eliminated 13 of these attributes as they were determined to be inconsistent. Then 9 of the remaining attributes are considered the most significant and have been considered in the final attributes' set. Then the second stage determined six clusters and further eliminated eight attributes to finally have 31 attributes as the final set. Moreover, determining the number of clusters was based on applying the Silhouette index to ensure the best attributes' distribution. To confirm the applicability for the proposed method, classification algorithms have been applied on the dataset after the dimensionality reduction process and the results' accuracy has been determined. The results showed that the classification accuracy reached above 95 % for the Random Forest technique which is a promising percentage.

Different future directions could be proposed including applying the proposed method on different datasets with a variety of nature to ensure its generality. Other knowledge discovery techniques could be investigated for higher contribution. Finally, a routing design approach could be further proposed which provide the suitable techniques for the dataset according to defined parameters including the number of attributes, their type, the dataset size, and the dataset complexity.

Acknowledgments

The authors would like to express their sincerest gratitude to Prof. Ayman E. Khedr who was leading all the research stages and provided insight and invaluably expertise that greatly assisted the research.

References

[1] A. E. Khedr and J. N. Kok, "Adopting Knowledge Discovery in Databases for Customer Relationship Management in Egyptian Public Banks", In: *Proc. of IFIP World Computer Congress, TC 12*, pp. 201-208, 2006.

[2] A. E. Khedr, A. M. Idrees, and A. Elseddawy, "Enhancing Iterative Dichotomiser 3 algorithm for classificat decision tree", *WIREs Data Mining and Knowledge Discovery*, Vol. 6, pp. 70-79, 2016.

[3] A. E. Khedr, A. M. Idrees, A. E.-F. Hegazy, and S. El-Shewy, "A proposed configurable approach for recommendation systems via data

mining techniques", *Enterprise Information Systems*, Vol. 12, No. 2, pp. 1-22, 2017.

[4] A. E. Khedr, A. Darwish, A. Z. Ghalwash, and M. A. Osman, "Computer-Aided Early Detection Diagnosis System of Breast Cancer with Fuzzy Clustering Means Approach", *International Journal of Cancer Research*, Vol. 48, No. 2, pp. 1257-1252, 2014.

[5] A. Khedr, S. Kholeif, and F. Saad, "An Integrated Business Intelligence Framework for Healthcare Analytics", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 7, No. 5, pp. 263-270, 2017.

[6] M. A. Osman, A. Darwish, A. E. Khedr, A. Z. Ghalwash, and A. E. Hassanien, "Enhanced breast cancer diagnosis system using fuzzy clustering means approach in digital mammography", *Handbook of Research on Machine Learning Innovations and Trends*, IGI Global, pp. 925-941, 2017.

[7] N. Sultan, A. E. I. A. M. Khedr, and S. Kholeif, "Data Mining Approach for Detecting Key Performance Indicators", *Journal of Artificial Intelligence*, Vol. 10, No. 2, pp. 59-65, 2017.

[8] A. E. Khedr, A. Khalil, and M. A. Osman, "Enhanced Liver Tumor Diagnosis Using Data Mining and Computed Tomography (CT)", In: *Proc. of the International Conference on Computing Technology and Information Management (ICCTIM)*, pp. 196-217, 2014.

[9] J. Tang, R. Zhang, P. Wang, Z. Zhao, L. Fan, and X. Liu, "A discrete shuffled frog-leaping algorithm to identify influential nodes for influence maximization in social networks", *Knowledge-Based Systems*, Vol. 187, pp. 1-12, 2020.

[10] K. K. Vasan and B. Surendiran, "Dimensionality reduction using Principal Component Analysis for network intrusion detection", *Perspectives in Science*, Vol. 8, pp. 510-512, 2016.

[11] R. Krishnana, V. A. Samaranayakeb, and S. Jagannathana, "A Multi-step Nonlinear Dimension-reduction Approach with Applications to Bigdata", In: *Proc. of INNS Conference on Big Data and Deep Learning 2018*, pp. 81-88, 2018.

[12] D. Giradi and A. Holzinger, "Dimensionality reduction for exploratory data analysis in daily medical research", *Advanced Data Analytics in Health, Part of the Smart Innovation, Systems and Technologies*, Vol. 93, pp. 3-20, 2018.

[13] H. Tian, W. Sheng, H. Shen, and C. Wang, "Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets", *Knowledge-Based Systems*, Vol. 187, 104828, 2020.

[14] A. M. Idrees, F. K. Alsheref, and A. I. ElSeddawy, "A Proposed Model for Detecting Facebook News' Credibility", *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 10, No. 7, pp. 311-316, 2019.

[15] E. Afify, A. Sharaf Eldin, A. E. Khedr, and F. K. Alsheref, "User-Generated Content (UGC) Credibility on Social Media Using Sentiment Classification", *FCI-H Informatics Bulletin*, Vol. 1, No. 1, pp. 1-19, 2019.

[16] A. E. Khedr, A. M. Idrees, and F. K. Alsheref, "A Proposed Framework to Explore Semantic Relations for Learning Process Management", *International Journal of e-Collaboration*, Vol. 15, No. 4, pp. 46-70, 2019.

[17] A. B. El Seddawy, T. Sultan, and A. E. Khedr, "A Proposed Data Mining Technique to Improve Decision Support System in an Uncertain Situation", *International Journal of Engineering Research and Development*, Vol. 8, No. 7, pp. 56-61, 2013.

[18] A. M. Idrees, A. I. ElSeddawy, and M. O. Zeidan, "Knowledge Discovery based Framework for Enhancing the House of Quality", *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 10, No. 7, pp. 324-331, 2019.

[19] M. M. Reda, Y. Helmy, A. E. Khedr, and A. Abdo, "Intelligent Decision Framework to Explore and Control Infection of Hepatitis C Virus", In: *Proc. of the International Conference on Advanced Machine Learning Technologies and Applications*, Vol. 723, pp.

264-274, 2018.

[20] A. E. Khedr, A. I. El Seddawy, and A. M. Idrees, "Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS", *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, Vol. 2, No. 6, pp. 111-118, 2014.

[21] A. E. Khedr, "Business Intelligence framework to support Chronic Liver Disease Treatment", *International Journal of Computers & Technology*, Vol. 4, No. 2, pp. 307-312, 2013.

[22] T. L. Saaty, "The Analytic Hierarchy Process", *McGraw-Hill*, New York, 1980.

[23] D. L. Olson, "The Analytic Hierarchy Process", *Decision Aids for Selection Problems*, Springer, New York, NY, pp. 49-68, 1996.

[24] A. Starczewski and A. Krzyżak, "Performance Evaluation of the Silhouette Index", In: *Proc. of the International Conference on Artificial Intelligence and Soft Computing*, *Part of the Lecture Notes in Computer Science*, Vol. 9120, pp. 49-58, 2015.

[25] D. H. Alpers, A. N. Kalloo, N. Kaplowitz, C. Owyang, and D. W. Powell, *Textbook of gastroenterology*, John Wiley & Sons, 2011.