Precision in Action: COVID-19 Detection with Generalized Linear Classifier and Two Step-AS Algorithm

Ahmed Hamza Osman [1†] , Hani Moetque Aljahdali, Sultan Menwer Altarrazi[2††] , and Ahmad A Alzahrani[3††]

[1†] Department of Information System
Faculty of Computing and Information Technology at Rabigh, King Abdulaziz University
Jeddah, Saudi Arabia
[2††] Department of Computer Science
Faculty of Computing and Information Technology at Rabigh, King Abdulaziz University
Jeddah, Saudi Arabia
[3††] Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University,
Jeddah, Saudi Arabia

Abstract
This research introduces a computer-aided intelligence model designed to automatically identify positive instances of COVID-19 for routine medical applications. The model, built on the Generalized Linear architecture, employs the TwoStep-AS clustering method with diverse filter relatives, abstraction, and weight-sharing properties to automatically identify distinctive features in chest X-ray images. Unlike the conventional transformational learning approach, our model underwent training both before and after clustering. The dataset was subjected to a compilation process that involved subdividing samples and categories into multiple sub-samples and subgroups. New cluster labels were then assigned to each cluster, treating each subject cluster as a distinct category. Discriminant features extracted from this process were used to train the Generalized Linear model, which was subsequently applied to classify instances. The TwoStep-AS clustering method underwent modification by pre-aggregating the dataset before employing the Generalized Linear model to identify COVID-19 cases from chest X-ray findings. Tests were conducted using the COVID-19 public radiology database guaranteed the correctness of the results. The suggested model demonstrated an impressive accuracy of 90.6%, establishing it as a highly efficient, cost-effective, and rapid intelligence tool for the detection of Coronavirus infections.

Keywords
Generalized Liner model, Covid-19, Two Step-AS, Clustering, X-ray images

1. Introduction

Covid-19 constitutes a diverse domestic of diseases capable of infecting humans and causing severe illnesses[1]. The current pandemic stems from a novel animal-borne illness, indicating that humans have not previously encountered this virus, and it has transitioned from animals to humans[3]. Given its novelty, there is a lack of inherent immunity among people, which distinguishes it from other viruses and contributes to its potential for widespread or local epidemics[4]. In this context, an epidemic is known as an eruption of a communicable disease significantly increasing humanity and illness over a superior geographic area, while an epidemic is a disease spreading rapidly within a short timeframe.

Previous instances include the SARS virus in 2002, which pretentious 8,096 persons and claimed over 770 lives, and the Middle East respiratory syndrome Covid-19 (MERS-CoV) in 2012, ensuing in 858 fatalities and 2,494 infections [5]. The ongoing battle against COVID-19, which began spreading in December 2019, has led to a global health crisis. COVID-19 primarily spreads through indirect or direct contact with diseased persons, breathing drops, or airborne conduction [6]. Primary symptoms include a high-temperature, dry cough, and trouble breathing, potentially progressing to severe breathing distress or organ failure, and in extreme cases, death [7, 8].

The rapid and sustained spread of COVID-19 poses challenges in our ability to effectively combat the virus, given the limited capacity of healthcare professionals and resources. This necessitates the development of tools such as contact tracing applications, statistical visualizations, dashboards, machine-learning methods, and other AI models to aid healthcare professionals in managing the pandemic.

Manuscript received December 5, 2024
Manuscript revised December 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.12.14

However, the proposed method has a limitation since it is confined to the chest X-ray dataset, and other medical datasets could be employed for COVID-19 identification. The subsequent sections of this paper are organized as follows. Section two reviews pertinent studies in this field. Section three outlines the intended proposed system. Section four deliberates on the strategy and technique. Section five scrutinizes the experimental results and the dataset. Section six encompasses a recap of the findings, discussion, and analysis, while section seven encapsulates the research summary and delineates future avenues for exploration.

2. Related works

Machine-learning has proven to be extremely effective in a wide range of picture combination processing tasks, including image-analysis [10, 11], image-segmentation [9], and image-classification [12]. Image categorization requires extracting significant features from images using descriptors, image instants [13], and SIFT [14]. These collected features are then used in prediction tasks by utilizing prediction devices such as support vector machine [15]. Traditional image fusion approaches, however, have intrinsic drawbacks, such as reduced image quality, increased noise in the final fused output image, and impracticality for real-time applications where images may blur. Color distortion and spectrum degeneration have also been reported in color photographs. In contrast to manually built features, deep neural network-based system techniques [16] show improved performance in image categorization based on extracted attributes. Several attempts have been made, leveraging machine learning techniques to categorize chest X-ray images in COVID-19 patient groups or normal cases. For instance, a Convolutional Neural Network (CNN) model was developed for spontaneous COVID-19 diagnosis from chest X-ray images, achieving a claimed classification accuracy of 96.78% using the MobileNet architecture [17]. Another study by Simi Larley [18] employed a transfer learning strategy, with reported accuracy rates of 97% and 87% for InceptionV3 and Inception-ResNetV2, respectively. The utilization of orthogonal moments, particularly orthogonal quaternion harmonic transformation moments, has proven effective in various pattern recognition and image processing applications [19] [20]. Recent research focused on developing an artificial intelligence-based programmable tomography analysis tool for monitoring COVID-19 progression, using 3D volume assessment to generate a "Corona Score" [22].

A study by Rasheed et al. [23] explored medical and technological aspects in combating the COVID-19 pandemic, offering valuable insights for virologists, infectious disease researchers, and policymakers. This study delved into the use of different technological tools and various artificial intelligence methods to aid in the pandemic, including predictive diagnostic machine learning techniques, such as deep learning.

Sethy and Behera [24] used X-ray images in combination with different CNN models and a support vector machine (SVM) for feature identification, highlighting the ResNet50 classifier combined with the SVM model as the most effective. Several recent COVID-19 studies incorporated a variety of CT image deep learning models in their analyses [25].

State-of-the-art techniques, drawing on deep learning approaches and utilizing chest X-ray images, have been developed based on research studies [11, 24, 26-30]. While machine learning approaches depend heavily on knowledge for information selection and extraction, they exhibit limited performance compared to deep learning methods. The advantages of machine learning methods, such as making the most of unstructured data, eliminating the need for engineered features, providing superior performance, reducing costs, and eliminating the need for data labeling, have led to their widespread use in automatically extracting crucial characteristics from items of interest for appropriate categorization. Notably, Apostolopoulos and Bessiana achieved a 97.8% accuracy in COVID-19 categorization with the VGG19 architecture [26], and Ozturk et al. demonstrated an 87% accuracy in categorizing coronavirus, pneumonia no-findings, and findings [11]. Sethy and colleagues developed a classification system for positive and negative coronavirus patients [24]. However, distinguishing coronavirus-caused pneumonia patients from other viral-induced pneumonia cases is crucial to prevent misdiagnosis, given the differing therapeutic approaches required for coronavirus disease. Various studies have suggested pulmonic chest infection categorization using deep learning methods [31, 32]. The current focus of research involves identifying COVID-19 patients with different pulmonary illnesses, such as edema, fibrosis, and effusion.

3. Proposed Method

This section presents a detailed discussion of the components and procedures used to create the suggested solution. Several critical phases are involved in recognizing COVID-19 from chest X-ray images: dataset collecting, data pre-processing, dataset categorization, training of models, model evaluation and analysis, and model validation and enhancement. Figure 1 depicts the system design for COVID-19 detection using TwoStep-AS and GL (TGL) and its components.

The initial step involves gathering and organizing the dataset required for training and model validation. To ensure consistency, the collected data undergoes transformations, scaling, and normalization. Subsequently, all data is categorized based on the model's classification scheme. Following that, models are trained and verified using exactly the same dataset and context as previously used models. Finally, for both the training and testing processes, the trained models are evaluated using accuracy metrics and the receiver

operational characteristic curve. Figure 1 depicts the TGL System's framework.

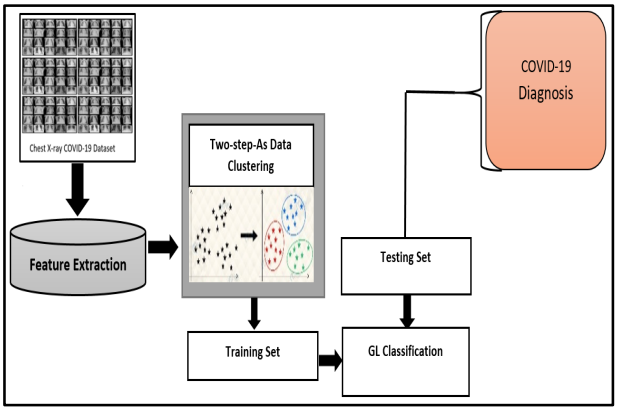

Fig. 1. TGL Classifer

Figure 1 depicts the fundamental structure of the TGL Diagnosis System. The proposed methodology comprises three stages: addressing the imbalance in the raw dataset and feature extraction, clustering instances based on their proximity to case characteristics using the TwoStep-AS algorithm, and performing diagnosis using a GL classifier during both the learning and testing phases. The proposed approach is designed to classify X-ray images into categories such as Pneumonia, COVID-19, or Non-COVID-19. The subsequent sections delve into the details of dataset modeling and the suggested TGL modeling.

A. COVID-19 Dataset

This study employs publicly accessible image repositories [33] to conduct its investigation. The dataset comprises typical chest X-ray images representing three distinct cases: normal, pneumonia, and COVID-19. Dr. Joseph Cohen, the curator of a GitHub repository containing annotated chest X-ray and CT scan images related to COVID-19, Acute Respiratory Distress Syndrome (ARDS), SARS, and MERS, meticulously captured and documented chest X-ray images from individuals infected with COVID-19. This compilation encompasses 250 confirmed chest X-ray images of individuals with COVID-19 viral infections. The Kaggle repository was utilized to procure chest X-rays from both healthy individuals and patients diagnosed with bacterial and viral pneumonia.

The application of AI-based X-ray screening for COVID-19 proves effective in both symptomatic and asymptomatic cases. However, distinguishing COVID-19 from other lower respiratory disorders that may exhibit similar features in X-ray imaging poses a unique challenge for algorithm developers. Dr. Cohen, affiliated with John Hopkins Hospital, generously contributed data in the form of JPG X-ray images, leading to the formation of a dataset sourced from Kaggle Chest X-rays datasets. This dataset facilitated a comparative analysis among healthy individuals, patients with bacterial pneumonia, and those with COVID-19 viral infection pneumonia. The collection incorporates chest images obtained from pneumonia patients admitted to hospitals. Dr. Cohen [33] established a COVID-19 X-ray image repository using publicly available images, consistently updating it with contributions from experts. Presently, the database encompasses 127 diagnostic X-ray images of COVID-19.

The National Institutes of Health Chest X-Ray Dataset (NIH) stands as another pivotal dataset in relation to COVID-19. Comprising 112,120 X-ray images with disease classifications from 30,805 distinct individuals, this dataset was constructed using Natural Language Processing to extract illness categories from corresponding radiological reports. Approximately 90% of the labels are deemed accurate, rendering them suitable for deployment in unsupervised learning scenarios.

B. Feature Extraction Process

- Statistical Feature

Upon closer examination of the X-ray images, it becomes evident that the predominant visual element is likely the excellent texture and statistical combinations. In recent years, many researchers have increasingly utilized textural and statistical characteristics to address classification challenges, and this trend is anticipated to persist. The appeal of such utilization lies in its simplicity compared to the labor-intensive process of software engineering, which demands an in-depth understanding of issue classes and methods for designing handcrafted descriptors. This function is unnecessary. While unmanufactured descriptors have some advantages, it is important to recognize that handcrafted descriptors have unique properties that can be extremely useful for a wide range of classification tasks. In this scenario, for example, the advantages of employing handmade features outweigh the disadvantages because these methods are more potent because they often operate in a more predictable manner to capture patterns connected with an issue. When handcrafted elements are used instead of unpolished ones, a more exact explanation of the patterns formed by the handmade characteristics of the photographs is more likely. Despite the fact that efforts have turned toward the usage of these two groups in feature extraction, this was not always the case. As a result, we can test the two independently and then combine the data of numerous experimental groups to arrive at a final result. In this way, we can use the complementarity among the descriptors' approaches, as demonstrated in [34, 35], to keep them from making the same mistakes while performing a given classification task. This section briefly discusses the adjectives used to describe the work. Specific texture descriptors were chosen to perform well in general applications or, more specifically, in medical picture analysis applications. The Gray-Level Co-Occurrence Matrix (GLCM) approach was used for the Texture features group. Sebastian et al. [36] defined GLCM as a matrix-based method frequently utilized in texture investigations to build linkages between pixels. When two pixels are nearby, the distance between them and the angle between their respective axes are used to calculate the relationship. As a result, the GLCM parameters are the

space's size and angles. The texture of a picture is quantified using GLCM functions, which determine the frequency of occurrence of pixel pairs with different values and in a specified spatial relationship. The GLCM generates a matrix of paired pixels with varying values and in a certain spatial relationship, from which statistical measurements are extracted. As previously stated, statistical measures of texture filter functions, as well as spatial connections of pixels in an image, were determined to be insufficient for delivering information on shape in texture features. Second-order statistics are used to build the GLCM feature set. To compute the reflection, the overall average of degrees of likeness between pixel pairs in various ways (homogeneity, uniformity, etc.) can be used. Pixel separation is one of the most crucial characteristics influencing the GLCM's discriminating abilities. When examining distance 1, the connection between pixel values is reflected (i.e., short-term neighborhood connectivity), and the change in distance value represents the change in the number of matching pixels.

- GLCM Features

GLCM provides functions that correctly capture the adjacency connection among pixels in the texture image, as described in statistical and structural texture approaches [36]. The equations used to extract characteristics from co-occurrence matrices are chosen based on the qualities to be noticed. We chose the four most relevant Haralick texture elements for future analysis based on the attributes of the X-ray imaging collection, which include correlation, homogeneity, energy, and contrast. Osman et al. [37] elaborate and offer the formulas for computing the statistical and GLCM features.

C. TwoStep-AS Cluster Algorithm

Numerous researchers, including [38-40], have employed the TwoStep Clustering algorithm across diverse domains. In their work, Najjar, A. et al. [38] applied an exploratory analytics approach to evaluate healthcare data based on the insights from the Smyth research[39]. The approach utilized a TwoStep Clustering methodology for heterogeneous finite mixture models, encompassing a joint mix of multinomial distribution and Gaussian for both categorical and numerical inputs. A hidden Markov model was incorporated for orders of categorical input. Deneshkumar et al. [40] proposed a technique for identifying outliers and determining the impact factor in diabetes patients, employing a TwoStep Clustering algorithm alongside other data mining methods. This study aims to uncover natural clusters within a knowledge collection through an exploratory technique known as TwoStep-AS Cluster, utilizing an algorithm that boasts several advantageous characteristics distinguishing it from conventional clustering methods.

- Handling Categorized or Continuous Variables: Utilizing a joint multinomial-normal distribution when variables are considered independent of one another.
- Automatic Selection of the Number of Clusters: Employing an optimization method to automatically determine the optimum number of clusters by comparing values of a model-choice criterion across various clustering solutions.
- Scalability: Constructing a Cluster Feature (CF) tree in the TwoStep-AS method to summarize entries in each cluster, facilitating the examination of large data files.

Industries like retail and consumer goods commonly apply clustering methods to analyze consumer data, tailoring marketing and product development strategies to specific consumer segments. The TwoStep-AS method incorporates log-likelihood distance, employing a pre-clustering process using the CF tree. This tree is traversed to determine the closest leaf entry for each record, updating the CF tree accordingly. The clustering stage then organizes the sub-clusters into the appropriate number of clusters using an agglomerative hierarchical approach. The log-likelihood distance measures the relationship between two clusters, utilizing probability functions based on variable values, considering categorical variables as multinomial and continuous variables as regularly distributed. The distance between clusters I and j is expressed as [42]:

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>} \quad (1)$$

Where

$$\xi_s = -N_v \left(\sum_{k=1}^{K^A} \frac{1}{2} \log(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2) + \sum_{k=1}^{K^B} \hat{E}_{vk} \right) \quad (2)$$

$$\hat{E}_{vk} = -\sum_{k=1}^{L_k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v} \quad (3)$$

The formulations encompass the following parameters:

K^A denotes the range category number of input features.

K^B is the symbolic number category of the input features.

L_k is the type number for the *k*th symbolic feature.

N_{vis} is the number of instances in cluster v.

N_{vkl} is the number of instances in cluster v that is similar to the *l*th type of the *k*th symbolic feature.

$\hat{\sigma}_k^2$ is the probable variance of the *k*th continuous feature for all instances.

$\hat{\sigma}_{vk}^2$ is the probable variance of the *k*th continuous feature for instances in the *v*th cluster.

<i,j> is an index representing the cluster molded by merging clusters i and j.

The distance between clusters i and j would be accurately equal to the reduction in log likelihood once the two clusters are joint if $\hat{\sigma}_k^2$ is ignored in the expression for ξv, and $\hat{\sigma}_{vk}^2$ is disregarded in the expression for v. Including this term helps avoid the issue created by $\hat{\sigma}_k^2$ =0, which would render the natural logarithm indeterminate. The technique comprises two phases. The first phase automatically defines the number of clusters, while the second phase computes Schwarz's Bayesian Information Criterion (BIC) for each cluster number within a given range. This indicator is then employed to determine an initial estimation for the number of clusters in the second phase.

$$BIC(J) = -2\sum_{j=1}^{J} \xi_j + m_J \log(N) \quad (4)$$

$$m_J = J\left\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right\} \quad (5)$$

$$\begin{cases} rv.binom(N, \frac{f_k}{N} & k = 1 \\ rv.binom\left(N - \sum_{i=1}^{k-1} f^*_{mi}, \frac{f_k}{N - \sum_{i=1}^{k-1} f_i}\right) & otherwise \end{cases} \quad (6)$$

The TwoStep Clustering method distinguishes itself from traditional clustering techniques through various advantageous features. Firstly, it accommodates both discrete and continuous variables as clustering inputs, expanding its applicability. Secondly, the TwoStep Clustering method demands fewer memory resources and exhibits faster calculations. Thirdly, it employs statistics as a distance index for clustering, simultaneously facilitating the automatic reorganization of data with an optimal number of clusters. Due to these attributes, the TwoStep Clustering technique is selected and explored for integration with the GL algorithm.

D. Generalized Liner Classifier

Generalized Linear Classifier is a classification model derived from the principles of Generalized Linear Models, adapted to handle categorical response variables. It provides a flexible framework for modeling relationships between predictors and categorical outcomes, making it applicable to a wide range of classification tasks.

It appears there might be a slight misspelling in your request. If you're referring to the "Generalized Linear Classifier," typically it's known as the "Generalized Linear Model (GLM)" or "Generalized Linear Regression (GLR)." A Generalized Linear Model is a flexible statistical framework that generalizes classical linear regression to accommodate various types of response variables and error distributions. The Generalized Linear Model extends the classical linear regression model to handle situations where the response variable is not normally distributed or when the relationship between variables is not linear. It consists of three main components:

The response variable, Y, is assumed to follow a probability distribution from the exponential family (e.g., Gaussian, binomial, Poisson).

The linear predictor, Xβ, where X represents the predictor variables and β is the vector of coefficients.

A link function, g(μ), connects the expected value of the response variable to the linear predictor. It specifies how the mean, μ, is related to the linear predictor. Common link functions include the identity, logit, and log.

The general form of a GLM is:

$$g(\mu)=X\beta \quad (7)$$

where:

g() is the link function.

μ is the expected value of the response variable.

X is the matrix of predictor variables.

β is the vector of coefficients.

Components of the Generalized Linear Classifier:

If you specifically meant a Generalized Linear Classifier (GLC), it could refer to a classifier based on the principles of Generalized Linear Models but adapted for classification tasks.

The suggested approach for diagnosing Covid-19 involves the integration of a Generalized Linear (GL) method with the Two-Step-AS algorithm. This method emulates human reasoning by considering multiple perspectives before arriving at a final decision. The unanimous decision to adopt this approach stems from the necessity for a high level of confidence in the real-world implementation of the research. This is particularly crucial due to the segregation of Covid-19 patients into bio-classes and the utilization of various decision-making fusion techniques, all extensively discussed in the paper. The primary objective is to elevate the learning process by grouping together Covid-19 samples with similar patterns. This grouping reduces complexity, leading to enhanced accuracy in diagnostic interpretation and, consequently, in the diagnosis itself.

4. Experimental Design and Dataset

This section delineates the manner in which the proposed Generalized Linear (GL) method was assessed and how the experiment was carried out utilizing the GL technique. According to the methodology we advocate, the existing X-ray data is enhanced by include well balanced coronavirus images. The goal of this section is to demonstrate the negative impact of imbalanced distributions on raw dataset performance. It is important to highlight that the TwoStep-AS-GL has been adjusted to perform training using the best available method parameters. This paper offers a Covid-19 diagnosis prediction strategy based on a hybrid

TwoStep-AS clustering algorithm and GL method. The goal is to improve classification diagnostic precision, reduce misdiagnosis mistakes, and boost classification accuracy. As a result, a new strategy that mixtures supervised and unsupervised learning techniques to generate a integrated instructional model is established. The TwoStep-AS clustering data structure was thoroughly investigated for X-ray chest imaging feature extraction utilizing the GL classification structure. The GL classifier was used to predict positive occurrences of Covid-19, pneumonia, and cases not discovered when the cluster findings were utilized as inputs to the classification model. The TGL model is used to investigate the effects of the qualifying procedure, taking into account the huge number of instances linked to the X-ray chest data. Two separate situations were used to detect and categorize COVID-19 in X-ray images. To begin, the TGL technique was trained to classify the X-ray pictures as COVID-19, No-Finding, or Pneumonia. In addition, two courses were trained in the TGL model: COVID-19 classes and No-Findings groups. The suggested model's output was tested for difficulties involving triple and binary categorization. The random images from this batch were utilized to assure balanced findings using a collection of chest X-ray images provided by Wang et al. [43], which comprises both normal and pneumonia images. After data balancing, the formed groups were used to identify each group separately using diagnostic cluster studies.

5. Results Discussion and Analysis

By implementing the TwoStep-AS Cluster, the performance of the NN classification algorithm can be elevated. This improvement is attributed to the fact that continuous features often exhibit enhanced performance when discretized [43]. Yang & Webb [44] utilized discretization as a technique to address continuous features in machine learning methods, enhancing the efficiency of data processing and optimizing inductive learning algorithms. The TwoStep-AS cluster, initially developed by Chiu et al.[45], is specifically designed to handle extensive datasets. Integrated into the statistical software SPSS, it serves as a clustering algorithm capable of managing both continuous and categorical data [46, 47]. Table 1 outlines the specifications of the TwoStep-AS model.

TABLE I. SPECIFICATIONS OF THE TWOSTEP-AS ALGORITHM

Minimum Number of Regular Clusters	2
Maximum Number of Regular Clusters	15
Feature Importance Method	Information Criterion
Information Criterion	Bayesian Information Criterion (BIC)
Distance Measure	Log Likelihood

As illustrated in Table 1, upon inputting a processed dataset, which refers to a quantified dataset, the system utilizes TwoStep-AS to generate a class label from the processed data. This class label comprises two labels, namely cluster-1 and cluster-2, grouped together. Subsequent to the definition of each class label, the prior probability of each class label is determined for NN calculation, a necessary step in the NN calculation process. Table 1 indicates that the TwoStep-AS algorithm generates a minimum of 2 and a maximum of 15 regular clusters. The TwoStep-AS algorithms utilize the BIC Method for Feature Importance designation and the Log Likelihood measure as the Distance Measure. The quality of the TwoStep-AS model's clustering evaluation is presented in Table 2.

TABLE II. QUALITY OF THE TWOSTEP-AS ALGORITHM.

Cluster-No	Number of Records	Goodness	Importance
Cluster1	985	0.89	1.00
Cluster2	416	-0.25	1.00

Table 2 illustrates the quality of the TwoStep-AS model, taking into account the number of records, goodness, and record importance. The goodness serves as a metric for cluster cohesion and separation. Cluster-1 has 985 records with a goodness of 0.89 and a record importance score of 1.00, while Cluster-2 has 416 records with a goodness of -0.25 and a record importance score of 1.00. The overall model goodness is measured by the Average Silhouette Coefficient, resulting in a value of 0.76 (interpreted as Good, on a scale from -1 to 1 where -1 to 0.2 is Poor, 0.2 to 0.5 is Fair, and 0.5 to 1 is Good). Additionally, importance is gauged as a measure of cluster cohesion, categorized as Poor (0 to 0.2), Fair (0.2 to 0.6), or Good (0.6 to 1).

In conducting an experimental study, a dataset related to COVID-19 was obtained for the purpose of data exploration. As mentioned earlier, the researchers utilized a tenfold cross-validation technique for both training and testing the dataset in their study. Additionally, a cross-dataset experiment was conducted, wherein the GL classifier was employed both independently and in conjunction with TwoStep-AS clustering results to assess the enhanced outcomes of the hybrid approach. The results of the cross-validation procedure were computed using equation (8) to yield the following diagnostic:

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN}+\text{FP})+(\text{TP}+\text{FN})} \times 100 \quad (8)$$

The number of COVID-19 cases correctly classified is referred to as the True Positive (TP). The number of COVID-19 instances identified erroneously is indicated by False Positive (FP). True Negative (TN) refers to the number of non-COVID-19 and pneumonia cases that were misclassified.

The number of non-COVID-19 and pneumonia cases identified erroneously is represented by False Negative (FN).

A chest X-ray dataset was evaluated to determine if patients were non-COVID-19, pneumonia, or COVID-19. The hybrid strategy was used to train and evaluate the dataset by merging TwoStep-AS and GL. Using the TwoStep-AS clustering algorithm, the dataset was then automatically separated into two clusters, each with a different amount of occurrences. In this study, the major goal of clustering is to identify patterns and structures from chest X-ray data by grouping samples with similar patterns. This minimizes the study's complexity and improves diagnostic interpretation accuracy. Tab.5 displays the results of the training and testing operations on the dataset, displaying a set of outcomes created by the Ensemble GL classifier method without clustering and with clustering using the TwoStep-AS algorithm.

TABLE III. FEATURE CHARACTERISTICS

Feature	N	Min	Max	Mean	Std. Deviation
Target	1125	1	3	2.33	.667
F1	1125	58	38747	182.88	1150.987
F2	1125	25	13462	66.25	399.881
F3	1125	4	33	5.06	.845
F4	1125	1	1	.92	.005
F5	1125	3	14	3.37	.322

The continuous variables in the COVID-19 dataset, as analyzed by the GL classifier, offer valuable insights into the dataset's characteristics. In Table 3, we observe the results for the dependent variable "Class" and the covariates (F1 to F5). For the dependent variable "Class," representing instances classified into categories (potentially non-COVID-19, pneumonia, and COVID-19), the dataset consists of 1125 observations. The statistics for "Class" include a minimum of 1, a maximum of 3, a mean of 2.33, and a standard deviation of 0.667. These statistics provide a concise overview of the continuous covariates, including their range, mean, and standard deviation. The wide range observed in variables like F1 and F2 suggests significant variability, while F3 appears to have a relatively stable range based on its mean of 5.06. Higher standard deviations, particularly in F1, indicate greater variability among data points.

We understanding that the distribution and characteristics of these continuous variables is pivotal for assessing their impact on classification outcomes. Further analyses, such as correlation assessments and evaluations of feature importance, can offer deeper insights into the relationships between these variables and the ultimate classification results. This foundational understanding sets the stage for more in-depth investigations into the dataset's dynamics.

TABLE IV. GOODNESS OF FIT

Measurement	Value	df	Value /df
Deviance	89.720	1118	.080
Scaled Deviance	1125.000	1118	-
Pearson Chi-Square	89.720	1118	.080
Scaled Pearson Chi-Square	1125.000	1118	
Log Likelihood	-173.833	-	-
Akaike's Information Criterion (AIC)	363.666	-	-
Finite Sample Corrected AIC (AICC)	363.795	-	-
Bayesian Information Criterion (BIC)	403.871	-	-
Consistent AIC (CAIC)	411.871	-	-

The results obtained from classifying the COVID-19 dataset using the GL classifier are outlined in the table 4, encompassing various goodness-of-fit metrics and information criteria. Specifically, the deviance is reported as 89.720, with Degrees of Freedom (df) being 1118, resulting in a Value/df ratio of 0.080. Deviance acts as an indicator of the model's fit, where lower values signify a better alignment with the data. In this context, the achieved deviance value is relatively low, indicating a favorable fit. Linking to the subsequent metric, the Scaled Deviance Value is recorded as 1125.000, with df being 1118. Scaled deviance, similar to deviance, assesses goodness of fit while considering the scale of the response variable. The relatively elevated value suggests a possibility for enhancing the model fit. Moving on, the Pearson Chi-Square is documented with a value of 89.720, df of 1118, and a Value/df ratio of 0.080. Similar to deviance, the Pearson Chi-Square evaluates the concordance between observed and expected values. A low value/df ratio is indicative of a favorable fit. Transitioning to the Log Likelihood, it is indicated by a value of -173.833. This metric, representing the logarithm of the likelihood function, seeks higher values for improved fit. The negative value is aligned with the logarithmic nature of the measurement. Next, Akaike's Information Criterion (AIC) is reported with a value of 363.666. AIC aims to strike a balance between fit and model complexity, where lower values suggest a favorable trade-off between fit and simplicity. Similarly, the Finite Sample Corrected AIC (AICC) is documented with a value

of 363.795. AICC, adjusted for small sample sizes, parallels AIC, and lower values are considered desirable for effective model evaluation. Moving on to the Bayesian Information Criterion (BIC), it is presented with a value of 403.871. Similar to AIC, BIC penalizes model complexity, and lower values indicate a superior model. BIC imposes a more stringent penalty for complexity. Finally, Consistent AIC (CAIC) is recorded with a value of 411.871. CAIC, which takes into account both fit and complexity, favors lower values for enhanced model performance.

The collective metrics imply that the GL classifier model exhibits a reasonable fit to the COVID-19 dataset. Nevertheless, there exists potential for improvement, and further refinement of the model or exploration of alternative approaches is worth considering.

The Omnibus Test provides strong evidence that the GL classifier, when applied to the COVID-19 dataset with the specified predictor variables, offers a statistically significant improvement in fit over an intercept-only model. This supports the validity and utility of the model in capturing and explaining the patterns in the data related to the classification of COVID-19 cases. With the Omnibus Test measurement, the Likelihood Ratio Chi-Square has been assesses whether there is a significant difference between the fitted model (GL classifier with predictor variables) and an intercept-only model (a model with no predictors).

The high value of 1932.649 and a very low p-value (0.0005) indicate that there is a significant difference between the two models. In other words, the inclusion of predictor variables in the GL model significantly improves its fit compared to a model with no predictors. The Omnibus Test supports the notion that the GL classifier, incorporating the specified predictor variables, is a statistically better fit for the COVID-19 dataset than a model without predictors. The GL classifier, as configured with the listed predictor variables, is deemed useful for explaining and predicting the variability in the dependent variable (Class), as evidenced by the significant Likelihood Ratio Chi-Square. The estimation parameters is presented in Table 5.

TABLE V. PARAMETER ESTIMATES

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-36.972	17.6338	-71.533	-2.410	4.396	1	.036
F1	.001	.0004	-5.010E-5	.002	3.381	1	.066
F2	.004	.0010	.002	.006	15.752	1	.000
F3	-11.110	3.8236	-18.604	-3.616	8.442	1	.004
F4	25.107	26.1258	-26.099	76.313	.924	1	.337
F5	21.576	8.2198	5.465	37.686	6.890	1	.009
[AS-TwoStep=Cluster-1]	-1.179	.0170	-1.213	-1.146	4790.478	1	.000
[AS-TwoStep=Cluster-2]	0[a]	.	.	.	.	.	.
(Scale)	.080[b]	.0034	.073	.087	.	.	.

The "Parameter Estimates" section furnishes essential details concerning the estimated coefficients, standard errors, confidence intervals, and hypothesis tests for each variable in the GL classifier model applied to the COVID-19 dataset. The Estimate (B) is noted as -36.972, with a Std. Error of 17.6338, a 95% Wald Confidence Interval of (-71.533, -2.410), a Wald Chi-Square of 4.396, df: 1, and a significance level (Sig.) of 0.036. Linking to the interpretation of the intercept, which represents the estimated log odds of the reference category (Class 1), the estimate of -36.972 suggests a significant negative association with the dependent variable. The confidence interval excluding zero indicates statistical significance. Moving to F1, with an Estimate of 0.001, Std. Error of 0.0004, a 95% Wald Confidence Interval of (-5.010E-5, 0.002), a Wald Chi-Square of 3.381, df: 1, and Sig.: 0.066, the coefficient implies a small positive effect, with a p-value suggesting marginal significance. The Transitioning to F2, where the Estimate is 0.004, Std. Error is 0.0010, a 95% Wald Confidence Interval of (0.002, 0.006), a Wald Chi-Square of 15.752, df: 1, and Sig.: 0.000, the positive coefficient and statistical significance (Sig. = 0.000) indicate a strong positive impact on the log odds. By examining F3 with an Estimate of -11.110, Std. Error of 3.8236, a 95% Wald Confidence Interval of (-18.604, -3.616), a Wald Chi-Square of 8.442, df: 1, and Sig.: 0.004, the negative coefficient signifies an association with lower log odds, and the p-value (Sig. = 0.004) indicates statistical significance. For F4, with an Estimate of 25.107, Std. Error of 26.1258, a 95% Wald Confidence Interval of (-26.099, 76.313), a Wald Chi-Square of 0.924, df: 1, and Sig.: 0.337, the positive coefficient is not statistically significant, as the p-value is 0.337. Considering F5, which scores an Estimate of 21.576, Std. Error of 8.2198, a 95% Wald Confidence Interval of (5.465, 37.686), a Wald Chi-Square of 6.890, df: 1, and Sig.: 0.009, the positive coefficient is statistically significant (Sig. = 0.009), suggesting a positive association. The AS-TwoStep=Cluster-1 obtained an Estimate of -1.179, Std. Error of 0.0170, a 95% Wald Confidence Interval of (-1.213, -1.146), a Wald Chi-Square of 4790.478, df: 1, and Sig.: 0.000. The TwoStep-AS clustering variable (AS-TwoStep) for Cluster-1 is highly significant (Sig. = 0.000), indicating its crucial role in the model. AS-TwoStep=Cluster-2 is recorded with an Estimate of 0 (set to zero because this

parameter is redundant). The Scale feature achieved an Estimate of 0.080, Std. Error of 0.0034, a 95% Wald Confidence Interval of (0.073, 0.087). The scale parameter provides information about the dispersion of the errors. We conclude that the intercept, F2, F3, F5, and AS-TwoStep variables exhibit significant associations with the dependent variable (Class), influencing the odds of COVID-19 classification. F1, F4, and the redundant Cluster-2 variable are not statistically significant contributors to the model. The positive coefficient for F2 suggests an increase in the odds of COVID-19 classification, while the negative coefficient for F3 implies a decrease. F5 exhibits a positive association, indicating increased odds. The AS-TwoStep clustering variable for Cluster-1 strongly influences the model, affirming the effectiveness of the TwoStep-AS clustering method in COVID-19 classification.

Another investigation, utilizing the NIH dataset, was carried out to scrutinize instances as either COVID-19 or non-COVID-19. The TGL model, coupled with several classifier methods, was utilized in the training and evaluation process to showcase the efficiency of the suggested model. Additionally, the accuracy of classification using the hybrid technique is documented in Table.6.

TABLE VI. PERFORMANCE ON THE TGL AND OTHER CLASSIFIERS METHODS.

Method	Accuracy
ANN	0.60
Support Vector Machine	0.65
Bayesian Network	0.69
C51-classifier	0.74
TGL Model	0.89

Figure 2 illustrates a comparison between the TGL model and currently employed methods. The proposed TGL method demonstrated a notable accuracy score of 0.90.6 in its application.

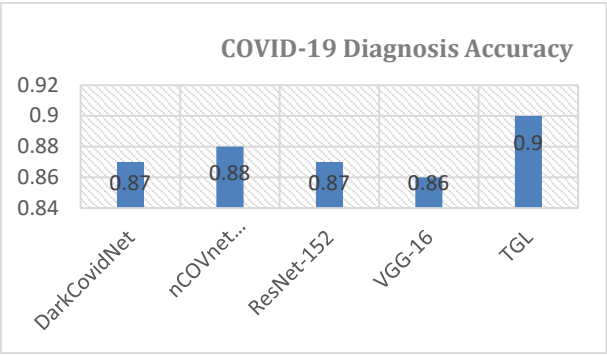

Fig. 2. Comparison between the TGL and other based methods

6. Conclusion and Future Directions

The focus of this study is on developing a novel GL method based on the TwoStep-AS clustering model (TGL) for detecting coronavirus and pneumonia cases. Clustering techniques play a crucial role in various domains that involve extensive datasets, aiming to unveil concealed patterns within the data. However, traditional clustering algorithms face challenges in effectively handling datasets containing both numerical and categorical attributes, common in real-world data. We demonstrated that the TwoStep-AS technique, known for its simplicity and automatic determination of the optimal number of clusters, can effectively address this issue.

In the initial phase of diagnosis using TGL, clinical cases undergo categorization into pneumonia, COVID-19, and normal cases. During the subsequent stage, given that Covid-19 stems from a virus, instances are further segregated into three categories: positive COVID-19, pneumonia, and negative COVID-19 (normal). The aim of the TGL method is to furnish a swift, systematic, and dependable computer-assisted solution for characterizing Covid-19 cases in patients undergoing preliminary screening with a chest X-ray scan upon admission to hospitals.

Comprehensive assessments have been carried out to showcase the effectiveness of the proposed approach, employing both learning and testing processes, and a tenfold cross-validation methodology has been utilized to illustrate the efficacy of TGL. Additionally, various tests have been executed to underscore the superiority of TGL in pinpointing Covid-19 cases compared to other cutting-edge methods for Covid-19 detection. Subsequent initiatives may involve harnessing advanced CNN techniques and diverse data mining models to refine the precision of detecting positive Covid-19 cases from chest X-ray and CT-scan images. Contemplation might also be given to adjusting the dimensions of the provided images, and the integration of machine learning-based image segmentation could further enhance performance. Furthermore, the exploration of optimized methods based on regression and classification algorithms will be undertaken to augment the predictive capability of the approach in diagnosing COVID-19.

Acknowledgments

This work was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (GCV19-6-1441). The author, therefore, gratefully acknowledge the technical and financial support from the DSR.

REFERENCES

[1] [1] J. Cui, F. Li, and Z.-L. Shi, "Origin and evolution of pathogenic coronaviruses," Nature Reviews Microbiology, vol. 17, pp. 181-192, 2019.

[2] [2] R. Tiwari, K. Dhama, K. Sharun, M. Iqbal Yatoo, Y. S. Malik, R. Singh, et al., "COVID-19: animals, veterinary and zoonotic links," Veterinary Quarterly, vol. 40, pp. 169-182, 2020.

[3] [3] B. Caballero, P. Finglas, and F. Toldrá, Encyclopedia of food and health: Academic Press, 2015.

[4] [4] C. Orbann, L. Sattenspiel, E. Miller, and J. Dimka, "Defining epidemics in computer simulation models: How do definitions influence conclusions?," Epidemics, vol. 19, pp. 24-32, 2017.

[5] [5] A. M. Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. Osterhaus, and R. A. Fouchier, "Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia," New England Journal of Medicine, vol. 367, pp. 1814-1820, 2012.

[6] [6] M. Moriyama, W. J. Hugentobler, and A. Iwasaki, "Seasonality of respiratory viral infections," Annual review of virology, vol. 7, pp. 83-101, 2020.

[7] [7] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," Journal of autoimmunity, vol. 109, p. 102433, 2020.

[8] [8] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, et al., "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," The lancet, vol. 395, pp. 1054-1062, 2020.

[9] [9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, pp. 1157-1182, 2003.

[10] [10] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The'big data'revolution in healthcare: Accelerating value and innovation," 2016.

[11] [11] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," Computers in biology and medicine, vol. 121, p. 103792, 2020.

[12] [12] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, et al., "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study," The Lancet infectious diseases, vol. 20, pp. 425-434, 2020.

[13] [13] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, et al., "Coronavirus disease 2019 (COVID-19): a perspective from China," Radiology, vol. 296, pp. E15-E25, 2020.

[14] [14] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "UCI machine learning repository-heart disease data set," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 1988.

[15] [15] J. P. Kanne, B. P. Little, J. H. Chung, B. M. Elicker, and L. H. Ketai, "Essentials for radiologists on COVID-19: an update—radiology scientific expert panel," ed: Radiological Society of North America, 2020.

[16] [16] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, "Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing," Radiology, vol. 296, pp. E41-E45, 2020.

[17] [17] E. Y. Lee, M.-Y. Ng, and P.-L. Khong, "COVID-19 pneumonia: what has CT taught us?," The Lancet Infectious Diseases, vol. 20, pp. 384-385, 2020.

[18] [18] F. Pan, T. Ye, P. Sun, S. Gui, B. Liang, L. Li, et al., "Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia," Radiology, 2020.

[19] [19] C. Long, H. Xu, Q. Shen, X. Zhang, B. Fan, C. Wang, et al., "Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?," European journal of radiology, vol. 126, p. 108961, 2020.

[20] [20] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, et al., "Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection," Radiology, p. 200463, 2020.

[21] [21] W. Kong and P. P. Agarwal, "Chest imaging appearance of COVID-19 infection," Radiology: Cardiothoracic Imaging, vol. 2, p. e200028, 2020.

[22] [22] K. McIntosh, M. S. Hirsch, and A. Bloom, "Coronavirus disease 2019 (COVID-19)," UpToDate Hirsch MS Bloom, vol. 5, 2020.

[23] [23] J. Rasheed, A. Jamil, A. A. Hameed, U. Aftab, J. Aftab, S. A. Shah, et al., "A survey on artificial intelligence approaches in supporting frontline workers and decision makers for COVID-19 pandemic," Chaos, Solitons & Fractals, p. 110337, 2020.

[24] [24] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," 2020.

[25] [25] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, et al., "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021.

[26] [26] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," Physical and Engineering Sciences in Medicine, vol. 43, pp. 635-640, 2020.

[27] [27] S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo, et al., "Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging," Frontiers in medicine, vol. 7, p. 427, 2020.

[28] [28] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet," Chaos, Solitons & Fractals, vol. 138, p. 109944, 2020.

[29] [29] S. Albahli, "A deep neural network to distinguish covid-19 from other chest diseases using x-ray images," Current medical imaging, vol. 17, pp. 109-119, 2021.

[30] [30] J. Civit-Masot, F. Luna-Perejón, M. Domínguez Morales, and A. Civit, "Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images," Applied Sciences, vol. 10, p. 4640, 2020.

[31] [31] R. H. Abiyev and M. K. S. Ma'aitah, "Deep convolutional neural networks for chest diseases detection," Journal of healthcare engineering, vol. 2018, 2018.

[32] [32] Z. Tariq, S. K. Shah, and Y. Lee, "Lung disease classification using deep convolutional neural network," in 2019 IEEE international conference on bioinformatics and biomedicine (BIBM), 2019, pp. 732-735.

[33] [33] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," arXiv preprint arXiv:2006.11988, 2020.

[34] [34] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," Pattern Recognition, vol. 71, pp. 158-172, 2017.

[35] [35] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," Applied soft computing, vol. 52, pp. 28-38, 2017.

[36] [36] B. Sebastian V, A. Unnikrishnan, and K. Balakrishnan, "Gray level co-occurrence matrices: generalisation and some new features," arXiv preprint arXiv:1205.4831, 2012.

[37] [37] A. H. Osman, H. M. Aljahdali, S. M. Altarrazi, and A. Ahmed, "SOM-LWL method for identification of COVID-19 on chest X-rays," PloS one, vol. 16, p. e0247176, 2021.

[38] [38] A. Najjar, C. Gagné, and D. Reinharz, "Two-step heterogeneous finite mixture model clustering for mining healthcare databases," in 2015 IEEE international conference on data mining, 2015, pp. 931-936.

[39] [39] P. Smyth, "Probabilistic model-based clustering of multivariate and sequential data," in Proceedings of the Seventh International Workshop on AI and Statistics, 1999, pp. 299-304.

[40] [40] V. Deneshkumar, K. Senthamaraikannan, and M. Manikandan, "Identification of outliers in medical diagnostic system using data mining techniques," International Journal of Statistics and Applications, vol. 4, pp. 241-248, 2014.

[41] [41] L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis vol. 344: John Wiley & Sons, 2009.

[42] [42] J. Bacher, K. Wenzig, and M. Vogler, "SPSS TwoStep Cluster-a first evaluation," 2004.

[43] [43] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in Machine learning proceedings 1995, ed: Elsevier, 1995, pp. 194-202.

[44] [44] Y. Yang and G. I. Webb, "A comparative study of discretization methods for naive-bayes classifiers," in Proceedings of PKAW, 2002.

[45] [45] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris, "A robust and scalable clustering algorithm for mixed type attributes in large database environment," in Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, 2001, pp. 263-268.

[46] [46] C. Michailidou, P. Maheras, A. Arseni-Papadimititriou, F. Kolyva-Machera, and C. Anagnostopoulou, "A study of weather types at Athens and Thessaloniki and their relationship to circulation types for the cold-wet period, part I: two-step cluster analysis," Theoretical and applied climatology, vol. 97, pp. 163-177, 2009.

[47] [47] S. Satish and S. Bharadhwaj, "Information search behaviour among new car buyers: A two-step cluster analysis," IIMB Management Review, vol. 22, pp. 5-15, 2010.