

Integrating Unsupervised and Supervised ML Models for Analysis of Synthetic Data From VAE, GAN, and Clustering of Variables

Lakshmi Prayaga, University of West Florida, USA*

 <https://orcid.org/0000-0003-4995-8298>

Krishna Devulapalli, Indian Institute of Chemical Technology, India

Chandra Prayaga, University of West Florida, USA

 <https://orcid.org/0000-0002-7534-4313>

Aaron Wade, University of West Florida, USA

 <https://orcid.org/0000-0001-6425-581X>

Gopi Shankar Reddy, University of West Florida, USA

Sri Satya Harsha Pola, University of West Florida, USA

ABSTRACT

Clustering of variables is a specialized approach for dimensionality reduction. This strategy is evaluated for data reduction with a Kaggle diabetes dataset. Since the original dataset is small, Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) are used to generate 100,000 records and tested for resemblance to the real data using standard statistical methods. VAE-data is more representative of the real data than GAN-data when analyzed using machine learning (ML) models. Applying Clustering of Variables on VAE-data yields new synthetic variables (SV). SV-data is then augmented with target variable data. Random Forest model is used on VAE and SV data. SV-data results matched VAE-data, proving the new data's quality. SV-data also provides insights into correlations and data dispersion patterns. This analysis implements a combination of Unsupervised learning (clustering of variables) and Supervised learning (classification) which is reflected in the results.

KEYWORDS

Cluster of Variables, Dimensionality Reduction, Generative Adversarial Networks, Synthetic Data, Variational Auto Encoders

INTRODUCTION

Machine learning (ML) algorithms can be broadly classified as supervised and unsupervised learning types. Supervised ML is ideal when the target variable data are available along with the feature variables data. They are used for classification and regression problems in general. When the target variable data are not available and the objective is to classify the data into natural groups, unsupervised ML models such as cluster analysis are used.

DOI: 10.4018/IJDA.343311

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Clustering is an unsupervised learning model typically used to group items or entities with similar attributes together. Clustering algorithms have been used in multiple domains, such as forecasting customer demand based on recency, frequency, and monetary characteristics (Seyedan et al., 2022), clustering of vascular risk factors (Holthuis et al., 2021), improving predictions of stock market by “using information of similar stocks, determined via clustering, compared to a prediction model that does not take into account such cluster-derived data,” (Javier, 2023), and accurately predicting spatiotemporal patterns in travel time by a joint iterative clustering and predicting algorithm (Shaji et al., 2022).

However, as the applications of ML algorithms expand in scope, and with the advent of wearable devices and other technological advances (Tufail et al., 2023), researchers confront two main challenges: (1) obtain a sufficiently large data set necessary for constructing meaningful machine learning models (L’Heureux et al., 2017) and (2) high-dimensional data sets are becoming more prevalent across multiple disciplines (Yuan, 2023) including genetics (Chi et al., 2016), organizational psychology, and neuroscience (Waldman et al., 2019). The effectiveness of machine learning models is inherently tied to the quality and quantity of data used for training. However, acquiring such data, which are both abundant and of high quality, is often scarce.

In this context, synthetic data emerge as a pivotal solution. Synthetic data offer a means to overcome the limitations of data scarcity by providing an avenue to generate data sets that possess both the required quantity and quality. By leveraging synthetic data, machine learning practitioners can enhance the robustness and reliability of their models, ensuring they are equipped to make accurate predictions in various domains and applications.

A second component required for the analysis of high-dimensional data is dimensionality reduction. Dimensionality reduction becomes necessary for meaningful data analysis. Dimensionality reduction also provides more insightful visualizations (Xia et al., 2023). It is also the case that available real data in many domains are often limited in size to build robust ML models (de Melo et al., 2022). It is in this context that this study uses 100,000 records of synthetic data (variational autoencoder (VAE)-data)) generated based on the real diabetes data set from Kaggle, which has just 768 records. Clustering of variables is performed on the VAE-data to obtain synthetic variables, which are linear combinations of the features in the VAE-data. The synthetic variables are lesser in number than the features in the VAE-data and thus result in dimensionality reduction. This newly generated synthetic variables data (SV-data) are used to train and test unsupervised clustering and supervised classification models to predict the outcome value of the diabetes condition. The quality of the new synthetic variables in terms of capturing the inherent patterns of the real data is assessed by applying ML methods to both VAE-data and SV-data. The resulting accuracies of each model as applied to the original data are compared.

CONTRIBUTION TO THE LITERATURE

Our contribution to this body of literature encompasses several novel aspects, which are discussed below. First, we introduce a unique integration of unsupervised techniques, such as clustering, with supervised methods, such as classification. This fusion not only represents an innovative approach but also signifies a comprehensive strategy aimed at enhancing classification accuracy by leveraging the strengths of both types of algorithms.

Second, our study demonstrates the quality and reliability of synthetic data generated through this integrated approach. By applying a combination of unsupervised and supervised machine learning models to both VAE and SV, we observed comparable accuracies. Specifically, utilizing a Random Forest classifier on 80% of the VAE-data for training and testing, with the remaining 20% reserved for independent testing, yielded promising results. These findings underscore the potential of synthetic data generators, such as variational autoencoders, in addressing challenges related to limited real data availability and privacy concerns.

Third, our investigation into the preservation of inherent data patterns reinforces the credibility of the generated synthetic data. The observed improvements in accuracy, obtained through the combined approach of supervised and unsupervised methodologies, further validate the fidelity of synthetic data representations. This breakthrough not only enhances the quality of synthetic data sets for training of machine learning models but also offers potential solutions for mitigating issues associated with data scarcity and privacy constraints (Tufail et al., 2023).

Fourth, the wide-ranging applications of our approach are noteworthy, particularly in domains such as healthcare, finance, and cybersecurity, where synthetic data generation is increasingly utilized (Alkhalifah et al., 2022). This methodology holds promise for scenarios involving limited data availability or high-dimensional data sets requiring sophisticated feature engineering or dimensionality reduction techniques.

Finally, our study justifies the effectiveness of the clustering of variables (Chavent et al., 2022) technique in performing dimensional reduction of the original VAE-data while maintaining results close to the original data for ML applications. This establishes the clustering of variables technique as a powerful tool for data reduction that can be seamlessly integrated into machine learning applications in place of original data.

Our research questions and hypotheses for the study are:

RESEARCH QUESTIONS

1. Is the classification accuracy similar between the original VAE-data and the SV-data derived using clustering of variables?
2. Do the synthetic variables, generated from VAE using clustering of variables, preserve the inherent patterns of correlations among the original variables?

HYPOTHESES

Hypothesis 1: The accuracy of the classification model applied to the original VAE-data will differ significantly from that of the SV-data.

Null Hypothesis 1: There are no significant differences in the accuracy of the classification model when applied to the original VAE-data compared to the SV-data.

Hypothesis 2: The newly generated synthetic variables will accurately capture the underlying patterns of correlations present among the original variables.

Null Hypothesis 2: The newly generated synthetic variables will not accurately capture the underlying patterns of correlations present among the original variables.

DATA SOURCE

The data set (Khare, 2023), used in this study is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the data set is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the data set. Several constraints were placed on the selection of these instances from a larger database. All patients here are females, at least 21 years old, and of Pima Indian heritage.

DATA SET OVERVIEW

The data set comprises 768 samples, with 34.9% having diabetes, each characterized by various attributes that play a crucial role in the analysis conducted in this project. The data set is structured with nine columns, each providing distinct information. Below is a concise description of each column:

Table 1. Variable Description of Diabetes Data Set

Column	Description
Pregnancies (int64)	Number of pregnancies
Glucose (int64)	Glucose concentration in blood
BloodPressure (int64)	Blood pressure
SkinThickness (int64)	Thickness of the skinfold
Insulin (int64)	Insulin level in the body
BMI (float64)	Body mass index
DiabetesPedigreeFunction (float64)	Hereditary risk of diabetes
Age (int64)	Age of the individual
Outcome (bool)	Presence or absence of diabetes

METHODOLOGY

In this hybrid study of integrating supervised and unsupervised machine learning models, the unsupervised ML method, clustering of variables, is performed on the VAE-data, using the ClustofVar package in R (Chavent et al., 2022), with which dimensionality reduction is achieved. This new data with the reduced dimensions are the SV-data. The package PyCaret (Ali, 2020) is then used to classify both the VAE-data and SV-data in terms of several ML methods in order to compare their respective accuracies.

Software Used for This Research

The following packages and software have been used for this research since they were open source and had the required features to complete this project.

- R
- ClustOfVar Package in R (Chavent et al., 2022) to cluster variables
- Synthetic Data Vault (Patki et al., 2016) to generate the Generative Adversarial Network (GAN) and VAE-data
- PyCaret (Moez Ali, 2020) for running the machine learning models
- ICSNP Package (Nordhausen et al., 2023) for Hotelling's T^2 test

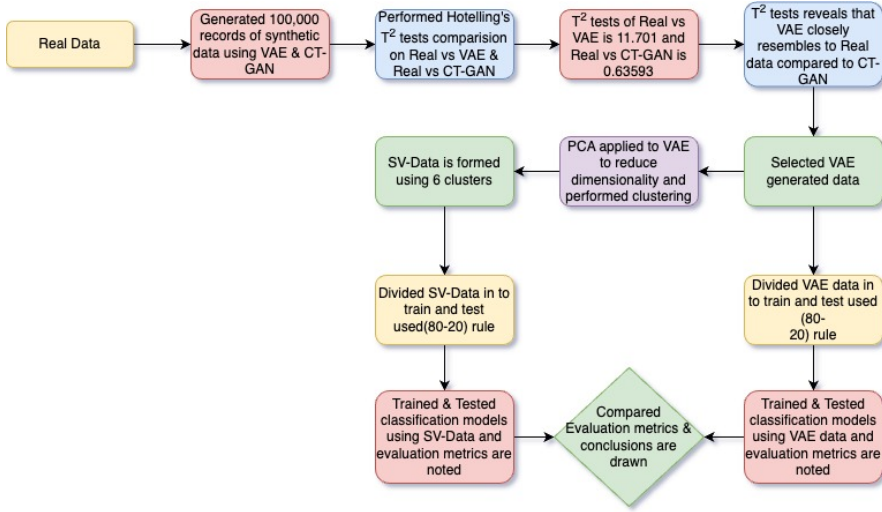
Synthetic Data Generation

GANs and VAEs are both popular techniques in the field of deep learning for generative modeling. They are used to generate new data that are similar to a given data set, making them particularly useful in tasks for image generation, data augmentation, and more. GANs focus on generating data by training a generator to produce realistic samples and a discriminator to distinguish between real and generated data. VAEs, on the other hand, generate data by encoding input data into a probabilistic latent space and then sampling from this space to produce new data. Both techniques have their strengths and are used in various applications depending on the specific requirements of the task (Sami & Mobin, 2019).

Conditional GANs and VAEs for Generating the Synthetic Data

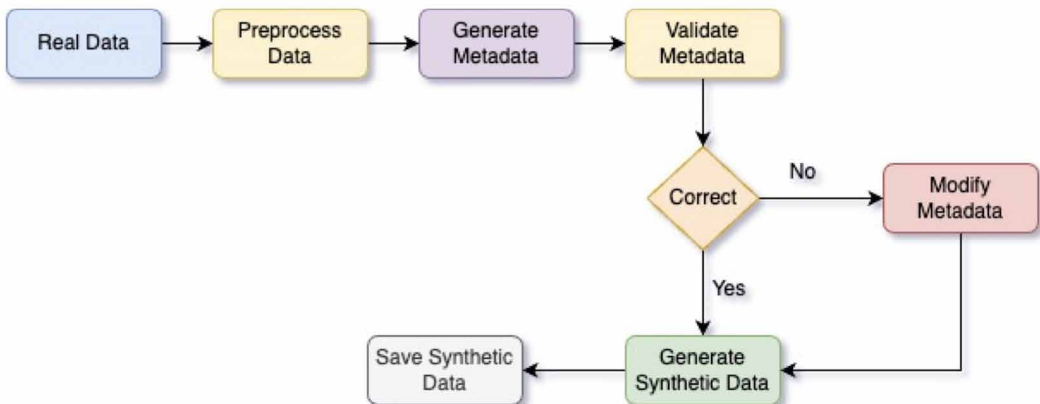
When using GANs and VAEs for synthetic data generation, the primary objective is to approximate the underlying data distribution of the real data set (Niederberger, 2012; Jordan et al., 1999). The

Figure 1. Work Flow Diagram



VAE framework incorporates the evidence lower bound (ELBO) loss function, which consists of two components: the reconstruction loss, often represented as L_{rec} , measuring the fidelity of the generated data to the real data, and the Kullback-Leibler (KL) divergence, denoted as L_{KL} , regulating the distribution of the latent space (Reizinger et al., 2022). Mathematically, the ELBO loss is defined as $L_{ELBO} = L_{rec} - \beta \cdot L_{KL}$, where β is a hyperparameter balancing the reconstruction fidelity and the divergence from the prior distribution. The aim is to minimize this loss function during training. VAEs thus aim to generate data that not only resemble the real data set but also maintain a structured latent space. Conversely, conditional tabular GAN (CT-GAN) (Skoularidou et al., 2019), although effective in generating synthetic data, might struggle with preserving the underlying structure of the data, leading to discrepancies compared to the real data set. This discrepancy is reflected in terms of statistical measures, such as distributional similarity or fidelity to the original data, where VAE typically outperforms CT-GAN, as reflected in our results. Figure 2 below shows the workflow of the current study.

Figure 2. Flowchart of Synthetic Data Generation with Synthetic Data Vault (SDV)



Data preprocessing enhances the quality of generated data (Patki et al. 2016), much like any other data creation method, and the following methods have been implemented in:

The synthetic data vault (SDV) synthetic data generation process, depicted in Figure 2, initiates with preprocessing real data. This step encompasses essential cleaning tasks such as handling null values and duplicates. Subsequently, metadata extraction occurs, capturing feature types and crucial information for subsequent phases. The extracted metadata undergo thorough validation, ensuring accuracy, with any discrepancies prompting revisions.

Upon successful validation, synthetic data generation commences. This phase employs CT-GAN and VAE models trained over 3,000 epochs. These models generate two distinct data sets, each comprising 100,000 synthetic instances. This systematic approach guarantees that the synthetic data preserves the structural and format characteristics of the real data while safeguarding privacy and integrity.

Evaluation of Synthetic Data

The quality of the generated synthetic data was tested by statistical measure (Hotelling's T^2 test) and with the help of pair plots, correlation plots, and density plots of the data sets. The results are discussed below.

Hotelling's T^2 Test to Select Between CT-GAN- and VAE-Data Sets

Hotelling's T^2 test is employed to assess significant differences between the mean vectors (multivariate means) of two multivariate data sets (Zaiontz, 2023; Schumacker, 2016, pp. 27–55; Ramasamy, 2021). In this study, the original Kaggle diabetes data set serves as the basis data set, while the CT-GAN-generated data set (CT-GAN-data) and VAE-generated data set (VAE-data) are treated as separate multivariate data sets for comparison.

The null hypothesis states that there are no differences between the population mean vectors of the two data sets. A significance level of 0.05 is chosen for hypothesis testing.

The R function Hotelling's $T^2()$ from the ICSNP (Nordhausen et al., 2023) package is utilized to calculate the Hotelling's T^2 test statistic value. If the resulting p value is less than .05, the null hypothesis is rejected, indicating significant differences between the data sets. Conversely, if the p value exceeds .05, the null hypothesis cannot be rejected, suggesting no significant differences. Based on the test results of the T^2 statistic, the CT-GAN- or VAE-data that closely resemble the original Kaggle data set are selected for further analysis.

RESULTS AND DISCUSSION FROM HOTELLING'S T^2 TEST

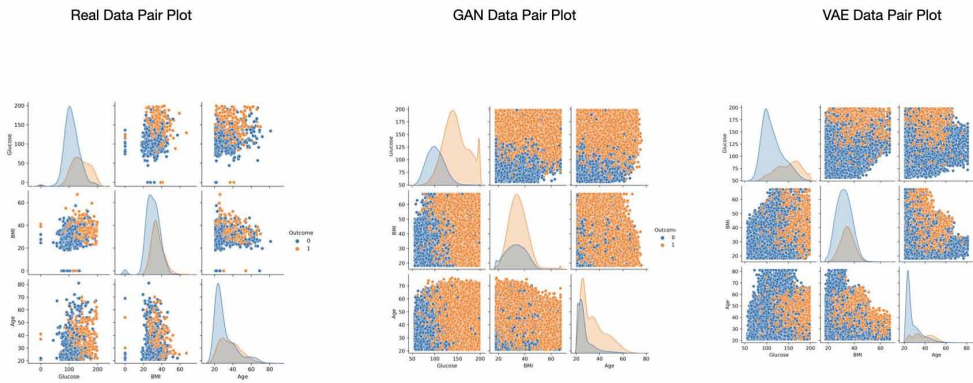
Hotelling's T^2 test is performed independently between 1) the original Kaggle diabetes data set and CT-GAN-data and 2) the original Kaggle diabetes data set and VAE-data, and the results are presented in Table 2.

In Table 2, the second column contains the two multivariate data sets used for Hotelling's T^2 test. The third column contains the T^2 test statistic value. $n1$ and $n2$ represent the sample sizes of the two data sets, respectively. The last column contains the p value associated with the T^2 statistic value.

Table 2. Hotelling's T^2 Test Results

S.No.	Type	T^2	$n1$	$n2$	p
1	Kaggle data set vs. CT-GAN-data	11.701	8	50,000	<2.2e-16
2	Kaggle data set vs. VAE-data	0.63593	8	100,000	0.7482

Figure 3. Pair Plot Analysis of Synthetic Data



Comparison Between CT-GAN-Data and the Original Kaggle Data Set

The high T^2 statistic value (11.701) and the associated p value ($<2.2e-16$) suggest a significant difference between the CT-GAN-data and the original Kaggle data set.

Therefore, we reject the null hypothesis, indicating that the CT-GAN-data significantly differs from the original Kaggle data set.

Comparison Between VAE-Data and Original Kaggle Data Set

In the case of the VAE-data, the low T^2 statistic value (0.63593) and the associated p value (0.7492) indicate no significant difference between the VAE-data and the original Kaggle data set.

Consequently, we fail to reject the null hypothesis, concluding that the VAE-data do not significantly differ from the original Kaggle data set.

Figure 3 compares the distribution and intervariable relationships of three data sets, Real Data and synthetic data generated by a CTGAN and a VAE using pair plots for the variables Age, BMI, and Glucose. The real data serve as a benchmark, showcasing inherent patterns and distribution characteristics. The CTGAN-data capture the general structure but exhibit noticeable deviations, particularly in the spread and density of points. In contrast, the VAE-data closely mirror the real data, with scatter plots indicating similar clustering and density plots reflecting the distribution shapes more accurately. This visual analysis suggests that VAE outperforms the CT-GAN in replicating the complex statistical properties of the real data set.

Figure 4 displays a side-by-side comparison of correlation matrices for Real Data, CTGAN-data, and VAE-data, each reflecting the linear relationships between several biomedical variables. The CTGAN-data Correlation Plot, while capturing the general trend of relationships in the real data, shows notable variances, such as excessive positive correlations with Outcome and other variables. On the other hand, the VAE-data Correlation Plot exhibits a higher fidelity to the Real Data Correlation Plot, with the strengths and directions of the correlations between variables like Age, BMI, and Glucose appearing more closely aligned. This comparative visualization suggests that the VAE model is more adept at replicating the complex correlation structure of the real data set, whereas the CT-GAN model, although effective, displays slight discrepancies in capturing some of the nuances of the data.

Figure 5 displays density plots for “Glucose” and “Insulin” across Real, CTGAN-, and VAE-data sets. The Real versus CTGAN comparison for “Glucose” reveals that the CTGAN’s density plot deviates with an additional peak and less smoothness, whereas the Real versus VAE comparison shows a VAE plot that is smoother and more closely aligns with the real data, despite a slight underestimation of the right tail. For “Insulin,” the CTGAN density plot suggests a broader value distribution, lacking the sharp peak present in the real data, while the VAE plot, although slightly broader at the base, more

Figure 4. Correlation Plot Analysis of Synthetic Data

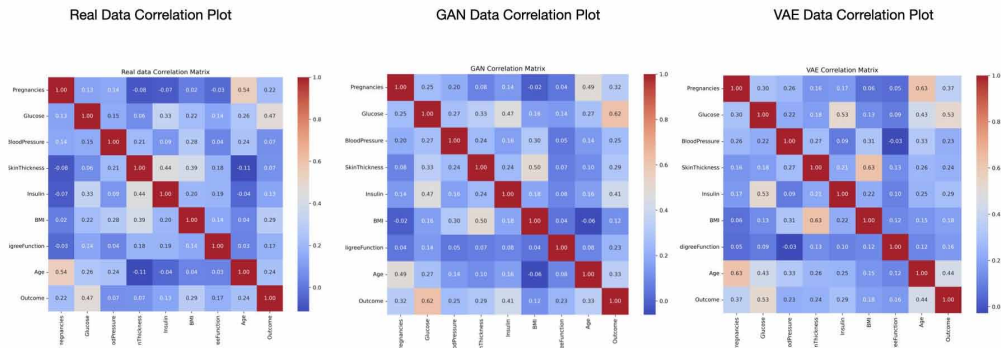
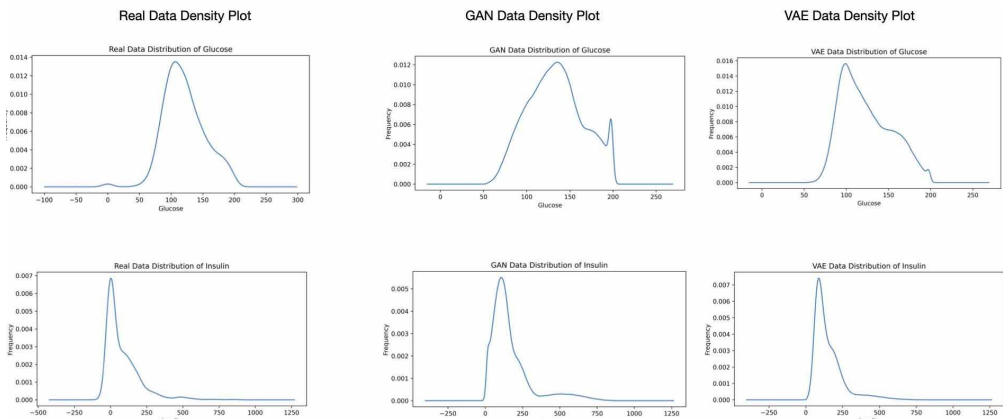


Figure 5. Density Plot Analysis



accurately captures the sharp peak, suggesting that the VAE is generally more effective at mimicking the real data distribution for both variables than the CTGAN.

Following the results of the statistical analysis presented above, VAE-data is used for further analysis, which includes combining unsupervised learning (clustering of variables) and supervised learning (comparison of classification models).

CLUSTERING OF VARIABLES (UNSUPERVISED LEARNING)

The clustering of variables is an unsupervised learning method that entails the categorization of related variables in order to detect meaningful patterns and minimize redundancy.

Data Reduction by Synthetic Variables

Principal component analysis (PCA) is used to create synthetic variables from the original data set. Let the original data set X consist of p features (columns) and n observations (rows). X is, therefore, an $n \times p$ matrix, each element of which may be denoted X_{ij} , where the row index i runs from 1 to n , and the column index j runs from 1 to p .

The method of PCA consists of constructing new synthetic variables S , which are linear combinations of the original variables (features). The synthetic variables are less in number compared

to the features, and thus a reduction of dimensionality of data is obtained. Let q be the number of synthetic variables ($q < p$). The linear combinations defining the synthetic variables may be written:

$$S_i = \sum_{j=1}^p w_{ij} X_j, i = 1, 2, \dots, q \quad (1)$$

Note that each S_i is a column vector with n elements, as is each feature X_j . The loadings w_{ij} are elements of the loading matrix W , representing the weights assigned to each original feature in the construction of the synthetic variable S_i .

The loading matrix W is obtained by first calculating pairwise correlations of the features. For two features X_i and X_j , the correlation coefficient r_{ij} is defined as:

$$r_{ij} = \frac{\sum_{k=1}^n (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\sum_{k=1}^n (X_{ki} - \bar{X}_i)^2 \sum_{k=1}^n (X_{kj} - \bar{X}_j)^2}} \quad (2)$$

Here, the average $\langle X_i \rangle$ represents the average of the i th feature, and $\langle X_j \rangle$ the average of the j th.

The matrix of pairwise correlation coefficients is a $p \times p$ matrix. This matrix is diagonalized, and its eigenvalues and eigenvectors determined. The eigenvectors corresponding to the largest q eigenvalues are then chosen to calculate the synthetic variables, since the largest eigenvalues correspond to the most correlated features.

Representing the eigenvectors by v_i , $i = 1, 2, \dots, q$, each synthetic variable S_i is calculated as the inner product of the original data matrix X with v_i :

$$S_i = X \cdot v_i$$

The elements of the eigenvector v_i are thus the loadings in Equation (1).

Various methods, such as the ClustOfVar package, have been developed specifically for the clustering of numerical variables, providing specific techniques for this purpose (Chavent et al., 2022). Additionally, the application of PCA in ClustOfVar, as demonstrated in this study, showcases the use of PCA to transform variables into principal components, contributing to the synthesis of variables within clusters (Shoji et al., 2022). Furthermore, the quality of variable clustering is crucial and emphasizes the importance of high-quality clustering to maximize the clustering criterion (Ni & Li, 2019). Strategies such as the ClustVarLV package in R offer approaches for deciphering the underlying structure of a data set through the clustering of features around latent variables, providing a comprehensive strategy for variable clustering. Vigneau et al. (2015) and Taskin et al. (2023) have used the copula-based clustering of variables technique along with the Random Forest model on the MIMIC-III Sepsis Dataset and the SMS Spam Collection Dataset and showed an improvement in CPU times and accuracy.

These approaches are thus useful for dimension reduction and variable selection. Several specific methods have been developed for the clustering of numerical variables in literature (SAS Institute Inc., 2011; Vigneau & Qannari, 2023). However, far fewer methods have been proposed concerning qualitative variables or mixtures of quantitative and qualitative variables. The R package was developed specifically for this purpose.

The scores of clusters suggested by Marie (Chaven et al., 2022) in the package is implemented in this study. ClustOfVar is a package in R which is used to cluster variables, in contrast to clustering

of observations/samples, as is done in the case of existing cluster analysis methods such as K-means clustering. The ClustOfVar package includes several methods to cluster variables, including hierarchical clustering. This package also accommodates quantitative, qualitative, and mixtures of quantitative and qualitative variables.

By using the methods of the package, the high-dimensional feature set can be grouped into a few selected clusters(K) of variables (Hummel et al., 2017). Each cluster represents a common aspect of the data, as in the case of principal component analysis. The cluster scores can be treated as new synthetic variables, and thus data reduction is achieved using these new synthetic variables (Ni & Li, 2019).

The variables are partitioned into a number of clusters based on hierarchy. Once the variables are partitioned, a dendrogram showing the hierarchy can be drawn, and the optimum number of clusters may be chosen by generating a stability plot and dispersion plots. The stability plot consists of the plot of the mean (over $B = 40$ bootstrap samples) of the adjusted Rand indices obtained. The Rand index is a measure used to evaluate the similarity between two data clusterings. It compares how pairs of data points are grouped in two different clusterings, providing a measure of their agreement. This plot reveals the stability of the partitions of the dendrogram and the optimum number of clusters to select. It also shows the dispersions of these indices over the specified number n of bootstrap replications for partition and decides the number of clusters to select.

RESULTS AND DISCUSSION OF COMBINED UNSUPERVISED AND SUPERVISED METHODS

Results and discussion of applying the ClustofVar (unsupervised) method and PyCaret (supervised) to evaluate VAE- and SV-data are presented below.

Results of Clustering of Variables (Unsupervised) Analysis on VAE-Data

Using the methods from the ClustOfVar package, we explore the associations among the eight quantitative feature variables using the correlation circle of the first two PCA dimensions. The resulting plot, depicted in Figure 6, offers a visual representation of correlated or anticorrelated variables. However, it does not offer a definitive partitioning of variables.

It is observed from Figure 6 that the variables Age and Pregnancies are well correlated. Similarly, the variables Glucose and Insulin have a good correlation. The variables BMI and SkinThickness are also well correlated.

Figure 6. Correlation Circle of the First Two PCA Dimensions

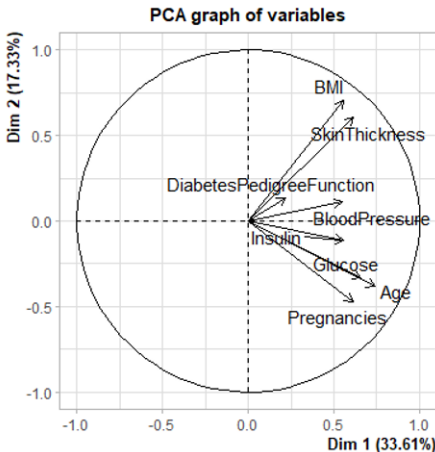
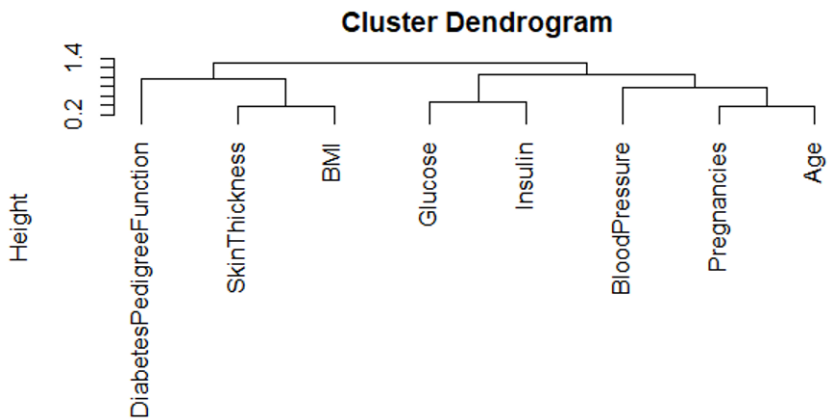


Figure 7. Cluster Dendrogram



Hierarchical clustering of variables is then performed on the variables, and the dendrogram generated is displayed in Figure 7.

Figure 7 shows the correlation between the variables in terms of R^2 . It can be observed that (1) the variables Pregnancies and Age are correlated, (2) the variables Glucose and Insulin are correlated, and (3) the variables BMI and SkinThickness are correlated.

In order to determine the optimum number of clusters to choose, we study the stability plot of the partition of the dendrogram and the dispersion plot of clusters generated by the hierarchical cluster of variables. Using the methods of the package, the stability plot and the dispersion plots are generated and are displayed in Figures 8 and 9.

Figure 8 displays the plot of the mean (over the $B = 40$ bootstrap samples) of the adjusted Rand indices. This plot clearly suggests that six clusters is the optimal number of clusters for this data set.

Figure 8 shows the dispersion of these indices over the $B = 40$ bootstrap replications for partition, and this plot also suggests retaining six clusters.

Figure 8. Stability of Partitions

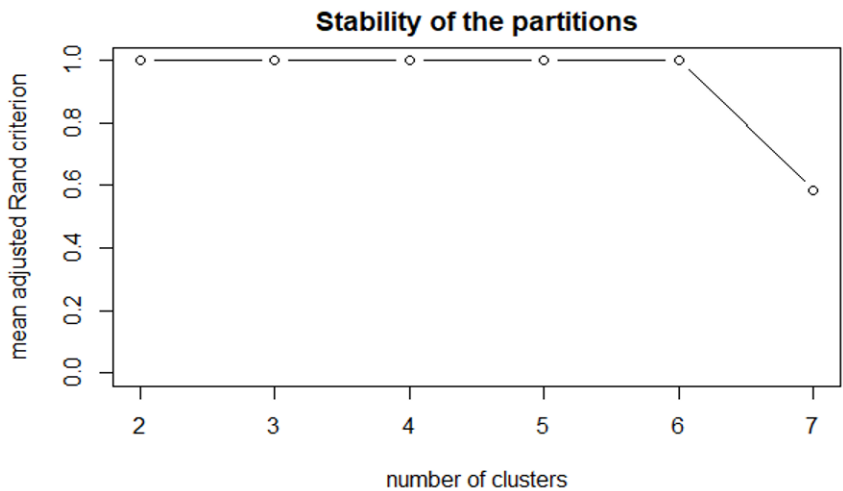
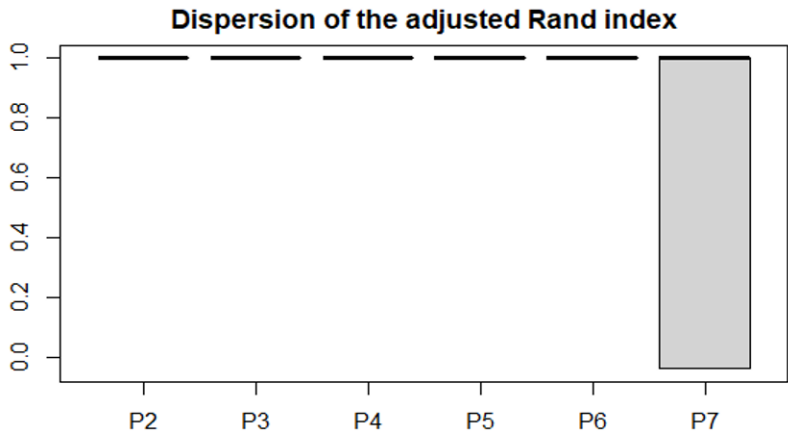


Figure 9. Boxplot of Dispersion of the Adjusted Rand Index



The function `cutree()` is used to cut the dendrogram obtained by the hierarchical clustering into partitions of ($K = 6$) clusters of the ($p = 8$) variables. The output of `cutree()` is a vector that assigns each observation to a particular cluster. The details of each cluster and the variables included in each cluster are obtained and displayed in Table 3. This table shows the variables Age and Pregnancies are allotted to Cluster 1, the variable Glucose to Cluster 2, and Blood Pressure to Cluster 3. Cluster 4 contains the two variables SkinThickness and BMI. Cluster 5 contains the single variable BMI, and Cluster 6 contains the single variable DiabetesPedigreeFunction.

The results of cluster analysis with six clusters are presented in Figure 10.

It is observed from Figure 10 that in Cluster 1, the variables Age and Pregnancies are strongly related and have high correlations with the synthetic variable 1. Similarly in Cluster 4, BMI and SkinThickness are also strongly related and have high correlations with synthetic variable 4. The remaining four variables Glucose, Insulin, Age, and BMI have formed separate clusters.

The `ClustOfVar` package also evaluates the cluster scores for all the 100,000 observations in each cluster. These 100,000 X 6 observations form the new synthetic data, and the six clusters can be treated as the new synthetic variables in the SV-data.

Table 3. Details of the Six Clusters for the Diabetes Data Set

Variable	Cluster
Pregnancies	1
Glucose	2
BloodPressure	3
SkinThickness	4
Insulin	5
BMI	4
DiabetesPedigreeFunction	6
Age	1

Figure 10. Cluster Analysis Results

```

$cluster1
      squared loading correlation
Pregnancies 0.8117859 -0.9009916
Age          0.8117859 -0.9009916

$cluster2
      squared loading      correlation
              1              1

$cluster3
      squared loading      correlation
              1              1

$cluster4
      squared loading correlation
BMI          0.812535 -0.9014072
SkinThickness 0.812535 -0.9014072

$cluster5
      squared loading      correlation
              1              1

$cluster6
      squared loading      correlation
              1              1

```

Results of Classification (Supervised) Methods on VAE- and SV-Data

PyCaret from Python was used to evaluate and compare the SV-data derived from cluster scores and the VAE-data. The PyCaret module automatically fits 14 machine learning models to the given data set and selects the best classification model based on certain metrics and criteria. Table 4 displays PyCaret results for the VAE-data. While fitting the PyCaret module, the target variable was set as Outcome, and the remaining eight variables were set as feature variables. The results are displayed in Table 4. The compare_models() function of PyCaret picked Random Forest as the best model for the VAE-data with an accuracy of 84.92%, Recall 0.6349, Precision 0.7891, and F1 Score 0.7036. Table 4 also shows that the top 10 models starting from Random Forest classifier to SVM Linear Kernel yielded accuracies above 80%.

PyCaret was also used to evaluate the SV-data, with the class variable Outcome as the target variable and the six cluster scores as feature variables. The results are displayed in Table 5. The compare_models() function identified Light Gradient Boosting Machine as the best model, with metrics: accuracy 84.16%, Recall 0.6176, Precision 0.7748, and F1 Score 0.6873. The Random Forest model also yielded similar results with the metrics: accuracy 84.01%, Recall 0.6123, Precision 0.7732, and F1 Score of 0.6833. It is noted that the metrics for the SV-data from the Light Gradient Boosting Machine model are very close to the Random Forest model metrics obtained for the VAE-data. Table 5 also shows that the top 9 models starting from Light Gradient Boosting Machine to Quadratic Discriminant Analysis yielded accuracies above 80%.

This justifies the result that the cluster of variables technique has not only performed dimensional reduction of the original VAE-data but yielded results which are very close to the results of the original data for machine learning applications. This establishes the fact that the relatively new cluster of variables technique is not only a powerful technique for dimension reduction but can also be used in machine learning applications in place of original data.

Table 4. VAE-Data PyCaret Results

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	TT (sec)
Random Forest classifier	0.8492	0.9008	0.6349	0.7891	0.7036	0.6040	0.6105	3.107
Light Gradient Boosting Machine	0.8481	0.8998	0.6395	0.7820	0.7036	0.6028	0.6084	0.334
Extra Trees classifier	0.8473	0.8994	0.6262	0.7888	0.6981	0.5977	0.6049	3.187
Gradient Boosting classifier	0.8418	0.8892	0.6356	0.7641	0.6939	0.5884	0.5930	3.412
AdaBoost classifier	0.8343	0.8791	0.6273	0.7447	0.6809	0.5701	0.5740	0.901
Logistic Regression	0.8321	0.8713	0.6119	0.7471	0.6727	0.5613	0.5664	3.835
Ridge classifier	0.8315	0.0000	0.6039	0.7499	0.6690	0.5577	0.5637	0.054
Linear Discriminant Analysis	0.8307	0.8703	0.6312	0.7317	0.6777	0.5638	0.5666	0.193
K Neighbors classifier	0.8221	0.8411	0.6122	0.7160	0.6600	0.5405	0.5436	0.824
SVM - Linear Kernel	0.8053	0.0000	0.5820	0.7045	0.6192	0.4932	0.5083	0.841
Decision Tree classifier	0.7780	0.7297	0.6190	0.6038	0.6113	0.4559	0.4560	0.213
Naive Bayes	0.7629	0.8176	0.3878	0.6292	0.4798	0.3373	0.3539	0.047
Dummy classifier	0.7180	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.056
Quadratic Discriminant Analysis	0.5981	0.6177	0.5399	0.4582	0.3748	0.1425	0.1703	0.119

Table 5. Synthetic Data PyCaret Results

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	TT (Sec)
Light Gradient Boosting Machine	0.8416	0.8901	0.6176	0.7748	0.6873	0.5831	0.5898	0.282
Extra Trees classifier	0.8409	0.8863	0.6069	0.7799	0.6826	0.5786	0.5868	2.064
Gradient Boosting classifier	0.8403	0.8856	0.6215	0.7678	0.6869	0.5813	0.5872	3.641
Random Forest classifier	0.8401	0.8862	0.6123	0.7732	0.6833	0.5783	0.5854	3.421
AdaBoost classifier	0.8314	0.8750	0.6180	0.7441	0.6739	0.5615	0.5658	0.958
Ridge classifier	0.8299	0.0000	0.6035	0.7445	0.6666	0.5541	0.5596	0.067
Linear Discriminant Analysis	0.8292	0.8683	0.6297	0.7277	0.6751	0.5600	0.5628	0.060
Naive Bayes	0.8166	0.8567	0.5731	0.7195	0.6379	0.5173	0.5233	0.058
Quadratic Discriminant Analysis	0.8088	0.8515	0.6335	0.6704	0.6513	0.5198	0.5203	0.045
Decision Tree classifier	0.7687	0.7176	0.6002	0.5881	0.5940	0.4324	0.4325	0.217
Logistic Regression	0.7181	0.4980	0.0000	0.0000	0.0000	0.0000	0.0000	0.074
Dummy classifier	0.7181	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.030
K Neighbors classifier	0.7013	0.5986	0.1860	0.4306	0.2598	0.1082	0.1231	0.310
SVM - Linear Kernel	0.6328	0.0000	0.2082	0.1043	0.0106	0.0558	0.0558	1.058

The results from the study show that hypothesis 1, stated in the Introduction, “The accuracy of the classification model applied to the original VAE-data will differ significantly from that of the SV-data.” is negated since the accuracy scores of both the VAE-data and SV-data are very close.

Null hypothesis 2, “The newly generated synthetic variables will not accurately capture the underlying patterns of correlations present among the original variables.” is negated since the correlations among variables in the VAE-data are indeed retained in the SV-data.

By negating hypothesis 1 and null hypothesis 2, this study answered the research questions 1) “Is the classification accuracy similar between the original VAE-data and the SV-data derived using clustering of variables?” and 2) “Do the synthetic variables, generated from VAE using clustering of variables, preserve the inherent patterns of correlations among the original variables?”

LIMITATIONS OF OUR STUDY

Some of the limitations of our study are that (1) synthetic data may not be able to completely capture the complex patterns to match the real data (Dankar & Ibrahim, 2021), (2) bias amplification could occur, suggesting that any bias in the real data set could be amplified in the synthetic data (Dankar & Ibrahim, 2021), and (3) evaluation metrics for testing synthetic data are challenging, leading to uncertainty about the effectiveness (Sampath et al., 2021).

Some of the limitations of the clustering of variables used in the study are (1) if the variables are highly correlated, the influence of one variable may overshadow the others, which could result in inaccurate clustering (Ezugwu et al., 2022), (2) it may be challenging to maintain the normalization and scaling patterns specific to the variable type (Ezugwu et al., 2022), and (3) there may be difficulty in interpreting the resulting clusters and domain expertise may be needed to interpret the grouping of variables (Ezugwu et al., 2022). However, utilizing variable clustering (Scrucca & Raftery, 2018) and synthetic data for training, then testing the model on real data, is an effective method to address challenges posed by high-dimensional and limited quantity data sets, as shown in our study.

CONCLUSION

An attempt is made in this study to use the cluster of variables technique to derive new synthetic variables from existing variables as a method of data reduction. For this, the diabetes data from Kaggle is utilized. Synthetic data using GAN and VAE were generated to address the challenge of the limited size of the original data. The VAE-data is closer to the real data as seen by statistical tests, and it is used for subsequent analysis.

Our experimental results demonstrate notable performance improvements when utilizing synthetic data generated by VAEs and GANs in combination with clustering of variables. Despite the inherent limitations of these techniques, the trained models consistently achieved competitive performance on real data, as displayed in Table 1. This study demonstrated that the novel combination of supervised learning (classification) and unsupervised learning (clustering of variables) methods fared well both on the real data and synthetic data. Future research directions may include exploring alternative clustering algorithms, refining synthetic data generation techniques, and investigating the scalability of our approach to larger and more diverse data sets.

ACKNOWLEDGMENT

We would like to thank iSpaceInc for partially funding this project.

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING STATEMENT

Funding was received from the University of West Florida.

We would like to thank ISpaceInc for partially funding this project.

CORRESPONDING AUTHOR

Correspondence should be addressed to Lakshmi Prayaga (lprayaga@uwf.edu)

PROCESS DATES

Received: 12/29/2023, Revision: 3/18/2024, Accepted: 3/18/2024

REFERENCES

- Ali, M. (2020). *PyCaret: An open-source, low-code machine learning library in Python* (Version 1.0.0) [Computer software]. <https://www.pycaret.org>
- Alkhalifah, T., Wang, H., & Ovcharenko, O. (2022). MLReal: Bridging the gap between training on synthetic data and real data applications in machine learning. *Artificial Intelligence in Geosciences*, 3, 101–114. doi:10.1016/j.aiig.2022.09.002
- Chavent, M., Kuentz, V., Lique, B., & Saracco, J. (2017). *ClustOfVar: Clustering of variables* (R package version 1.1) [Computer software]. <https://CRAN.R-project.org/package=ClustOfVar>
- Chi, W., Li, W. D., Wang, N., & Song, Z. (2016). Can genes play a role in explaining frequent job changes? An examination of gene-environment interaction from human capital theory. *The Journal of Applied Psychology*, 101(7), 1030–1037. doi:10.1037/apl0000093 PMID:27077527
- ClustOfVar Package in R. (2022). *Package “ClustOfVar.”* <https://cran.r-project.org/web/packages/ClustOfVar/ClustOfVar.pdf>
- Dankar, F. K., & Ibrahim, M. (2021). Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences (Basel, Switzerland)*, 11(5), 2158. doi:10.3390/app11052158
- de Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., & Hodgins, J. (2022). Next-generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, 26(2), 174–187. doi:10.1016/j.tics.2021.11.008 PMID:34955426
- Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110, 104743. doi:10.1016/j.engappai.2022.104743
- Holthuis, E. I., Visseren, F. L. J., Bots, M. L., & Peters, S. A. E.UCC-SMART Study Group. (2021). Risk factor clusters and cardiovascular disease in high-risk patients: The UCC-SMART study. *Global Heart*, 16(1), 85. doi:10.5334/gh.897 PMID:35141126
- Hummel, M., Edelmann, D., & Kopp-Schneider, A. (2017). Clustering of samples and variables with mixed-type data. *PLoS One*, 12(11), e0188274. doi:10.1371/journal.pone.0188274 PMID:29182671
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2), 183–233. doi:10.1023/A:1007665907178
- Khare, A. D. (2023, November 16). *Diabetes dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>
- L’Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine learning with Big Data: Challenges and approaches. *IEEE Access : Practical Innovations, Open Solutions*, 5, 7776–7797. <https://ieeexplore.ieee.org/document/7906512>. doi:10.1109/ACCESS.2017.2696365
- Ni, J., & Li, L. (2019). Memetic variable clustering and its application. *Mathematical Problems in Engineering*, 2019, 1–15. doi:10.1155/2019/4195318
- Niederberger, T. (2012). *Markov chain Monte Carlo methods for parameter identification in systems biology models* [Doctoral dissertation, University of Hertfordshire]. <https://core.ac.uk/download/216084033.pdf>
- Nordhausen, K., Sirkia, S., Oja, H., & Tyler, D. E. (2023). *ICSNP: Tools for multivariate nonparametrics* (R package version 1.1-2) [Computer software]. <https://cran.r-project.org/web/packages/ICSNP/ICSNP.pdf>
- Patki, N., Wedge, R., & Veeramachaneni, K. (2016, October 17–19). *The synthetic data vault* [Conference session]. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada. IEEE. <https://doi.org/10.1109/DSAA.2016.49>
- Ramasamy, K. (2021). *Hotelling’s T-square test for two independent samples*. LinkedIn. <https://www.linkedin.com/pulse/hotellings-t-square-test-two-independent-samples-kumar-ramasamy/>

Reizinger, P., Gresele, L., Brady, J., Von Kügelgen, J., Zietlow, D., Schölkopf, B., Martius, G., Brendel, W., & Besserve, M. (2022). Embrace the gap: VAEs perform independent mechanism analysis. *Advances in Neural Information Processing Systems*, 35, 12040–12057. https://proceedings.neurips.cc/paper_files/paper/2022/file/4eb91efe090f72f7cf42c69aab03fe85-Paper-Conference.pdf

Sáenz, J., Quiroga, F. M., & Bariviera, A. F. (2023). Data vs. information: Using clustering techniques to enhance stock returns forecasting. *International Review of Financial Analysis*, 88, 102657. doi:10.1016/j.irfa.2023.102657

Sami, M., & Mobin, I. (2019, March 13–15). *A comparative study on variational autoencoders and generative adversarial networks* [Conference session]. 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIIT), Yogyakarta, Indonesia. <https://doi.org/doi:10.1109/ICAIIIT.2019.8834544>

Sampath, V., Murtua, I., Aguilar Martín, J. J., & Gutierrez, A. (2021). A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of Big Data*, 8(1), 27. doi:10.1186/s40537-021-00414-0 PMID:33552840

SAS Institute Inc. (2011). *SAS/STAT software* (Version 9.2) [Computer software]. The VARCLUS procedure. <https://support.sas.com/documentation/onlinedoc/stat/930/varclus.pdf>

Schumacker, R. E. (2016). *Using R with multivariate statistics: A primer* (1st ed.). Sage Publications., <https://study.sagepub.com/multivariatewithr> doi:10.4135/9781071909768

Scrucca, L., & Raftery, A. E. (2018). Clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R. *Journal of Statistical Software*, 84(1). Advance online publication. doi:10.18637/jss.v084.i01 PMID:30450020

Seyedan, M., Mafakheri, F., & Wang, C. (2022). Cluster-based demand forecasting using Bayesian model averaging: An ensemble learning approach. *Decision Analytics Journal*, 3, 100033. doi:10.1016/j.dajour.2022.100033

Shaji, H. E., Tangirala, A. K., & Vanajakshi, L. (2022). Joint clustering and prediction approach for travel time prediction. *PLoS One*, 17(9), e0275030. doi:10.1371/journal.pone.0275030 PMID:36149882

Shoji, T., Sato, N., Fukuda, H., Muraki, Y., Kawata, K., & Akazawa, M. (2022). Clinical implication of the relationship between antimicrobial resistance and infection control activities in Japanese hospitals: A principal component analysis-based cluster analysis. *Antibiotics (Basel, Switzerland)*, 11(2), 229. doi:10.3390/antibiotics11020229 PMID:35203831

Taskin, Z. I., Yildirak, K., & Aladag, C. H. (2023). An enhanced random forest approach using CoClust clustering: MIMIC-III and SMS spam collection application. *Journal of Big Data*, 10(1), 38. doi:10.1186/s40537-023-00720-9

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics (Basel)*, 12(8), 1789. doi:10.3390/electronics12081789

Vigneau, É., Chen, M., & Qannari, E. M. (2015). Clustvarlv: An R package for the clustering of variables around latent variables. *The R Journal*, 7(2), 134–148. doi:10.32614/RJ-2015-026

Vigneau, É., & Qannari, E. M. (2023). Clustering of variables around latent components. *Communications in Statistics. Simulation and Computation*, 32(4), 1131–1150. <https://www.tandfonline.com/doi/full/10.1081/SAC-120023882>. doi:10.1081/SAC-120023882

Xia, J., Zhang, Y., Song, J., Chen, Y., Wang, Y., & Liu, S. (2023). *Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study*. arXiv. <https://arxiv.org/pdf/2110.02894.pdf>

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling tabular data using conditional GAN*. arXiv. <https://arxiv.org/pdf/1907.00503.pdf>

Yuan, S., DeRoover, K., & Van Deun, K. (2023). Simultaneous clustering and variable selection: A novel algorithm and model selection procedure. *Behavior Research Methods*, 55(5), 2157–2174. doi:10.3758/s13428-022-01795-7 PMID:36085542

Zaiontz, C. (2023). *Hotelling's T-square test for independent samples*. Real Statistics Using Excel. <https://real-statistics.com/multivariate-statistics/hotellings-t-square-statistic/hotellings-t-square-independent-samples/>

Lakshmi Prayaga is a Professor in the department of Information Technology, University of West Florida. Her research focuses on applications of data science in healthcare, sports medicine, management and training, stock markets. Topics of interest include data science and data visualizations. She has co-authored books on robotics, Android App development, beginning game programming, programming the web with ColdFusion and XHMTL, and using game programming to teach computer science concepts. She has also published numerous publications in International journals and conferences. She teaches graduate and undergraduate courses in Data Analytics, Data Visualizations, Machine learning and Script programming. She has an Ed.D. in Instructional Technology and a M.S. in Software Engineering, both from UWF an MBA from Alabama A&M University.

Krishna Devulapalli is a retired scientist from the Indian Institute of Chemical Technology, India. Currently he is a Freelance Data Scientist and is recognized as TapChief Expert in Data Science by TapChief, India. His research interests include applied statistics in multiple domains such as correlations among physico-chemical attributes of substances, healthcare analytics, BioInformatics, BioStatistics, Chemometrics, Reliability Studies, Pattern Recognition, Neural Networks, Rule Based Systems, Machine Learning etc. He has published more than 30 research papers in various journals and also presented number of papers in conferences and seminars. He has contributed some chapters in books related to Medical Statistics. He is a Member of various professional Societies like Indian Society of Medical Statistics, Computer Society of India, Indian Society of Analytical Scientists etc. He is recognized as a Guest Faculty in various organizations like Statistics Department, Osmania University, NIPER Guwahati, IICT, CSI, CMC etc.

Chandra Sekhar Prayaga is currently Professor of Physics, University of West Florida (UWF). He has a Ph.D. in Physics from the Indian Institute of Science, Bangalore, India (1975), where he was also a faculty member from 1981 to 1987. He has 45-plus years of experience in teaching physics, and has helped raise more than \$3 million in funding for research and projects involving University of West Florida faculty and students. His current research interests include optical and electronic properties of liquid crystals, Langmuir-Blodgett films, phase transitions and laser spectroscopy, physics education and data analytics, applications of Machine Learning to solve problems in physics. He mentors undergraduate student research projects, and coordinates summer camps on science and technology for middle and high school students.. He is co-author of a book, "Robotics: A Project-Based Approach," Cengage Publishers (2014).

Aaron Wade received his PhD from Florida State University in Condensed Matter Physics in 2008. He held a postdoctoral position at the University of Cincinnati in the Nano Materials Physics Group. He was hired at the University of West Florida in 2011 and is an Associate Professor of Physics. Aaron publishes and presents his work on the applications of machine learning, designing and characterizing visual research equipment, characterizing optical and electron properties of small molecule fluorophores, physics education research, and development of applied STEM+C active learning projects for 5th-12th graders.

Gopi Shankar Reddy is a master's student in Data Science at UWF and a research assistant, a graduate from IIIT Bangalore. His journey includes roles as a machine learning engineer and data analyst from 2019 to 2023. Passionate about solving real-world problems using data, he achieved a significant milestone in his career by developing a model for detecting cervical cancer and creating computer vision applications. Committed to expanding the impact of data for positive change, he thrives on pushing the boundaries of what data can achieve.

Sri Satya Harsha Pola, Mathematics and Statistics, University of West Florida (sp187@students.uwf.edu) Sri Satya Harsha Pola is a master's student in Data Science at the University of West Florida, serving as a research and teaching assistant. With a bachelor's from JNTU Hyderabad, he has showcased his skills by participating in the International Planetary Aerial Systems Challenge for designing a Mars drone. Passionate about leveraging data to drive meaningful solutions, he aims to become an accomplished data scientist, contributing innovative ideas that create positive social impact.