Journal of Business and Social Sciences (ISSN: 2350-8876)
Research Department, Hetauda School of Management and Social Sciences (HSMSS), Hetauda

A GANs Based Data Synthetic Technique for Enhancement of Prediction Accuracy of Mushroom Classification

Narayan Sapkota
Department of Information Technology,
Hetauda School of Management and Social Science, Hetauda, Nepal,
narayan.sapkota47@gmail.com

Cite this paper:

Sapkota, N. (2023). A GANs based data synthetic technique for enhancement of prediction accuracy of mushroom classification. *Journal of Business and Social Sciences, 5*(1), 21 – 36.

https://doi.org/10.3126/jbss.v5i1.72443

Abstract

In the landscape of machine learning and data-driven decision-making, limited data availability often undermines classification model accuracy. This study pioneers a solution by leveraging Copula Generative Adversarial Networks (Copula GANs) to generate high-fidelity synthetic data, with a focus on mushroom classification. Copula GANs replicate original dataset characteristics, as confirmed by thorough Category Coverage and TV Complement evaluations that validate its ability to accurately emulate category distributions and multivariate dependencies. To substantiate the practical impact, a mushroom classification task employs a decision tree. Results showcase notable accuracy enhancement through the integration of synthetic data with real data. Fine-tuning Copula GAN parameters, exploring feature interpretability, extending the technique to diverse domains, and merging with traditional data augmentation methods are promising future avenues. In essence, this study pioneers Copula GAN-generated synthetic data as a novel solution to data scarcity. The outcomes highlight the efficacy of synthetic data augmentation, advancing the potential of machine learning models across real-world applications.

Keywords: GAN, data synthesis, Copula GAN, mushroom classification, machine learning

Introduction

Mushrooms are the fleshy, edible fruit bodies of several species of fungi that belong to the Basidiomycetes class. They often grow on the surface of the soil or plant-derived materials like straw and wood. Since mushrooms lack chlorophyll, they are not autotrophs, yet they may still be able to obtain the nutrition needed for their growth by degrading complex substrates with their enzymes(Devika & Karegowda, 2021). Out of the estimated 1500,000 species in the world, the number of mushroom species that have been discovered so far is less than 69,000 (Wibowo et al.,

2018). To date, it is identified that 7000 species of mushrooms are edible (Bhatt, 2016)but the remaining are not easily distinguishable as edible or non-edible.Because mushrooms are an essential source of protein and contain several medicinal components, including ones that can even treat cancer, consumption of mushrooms has been rising globally over the past ten years (Kaushik & Choudhury, 2022). A few mushroom species are extremely dangerous, therefore not all of the mushrooms found in nature may be eaten. Generally, family of Agaricus and Lepiota easily found in wild open-area with various shapes, colours and characteristics are poisonous (Wibowo et al., 2018). It has been discovered that certain poisonous mushrooms resemble edible mushrooms quite a bit by physical appearance, therefore, they only are differentiated by one or two distinct traits, such as cap form, gill colour, odour, etc. (Wagner et al., 2021).

Consequently, a lot of researchers are working to develop trustworthy ways to identify if a mushroom is poisonous or not. Traditionally, a method involves boiling mushrooms with rice in a pot, where a change in rice colour indicates the presence of a poisonous mushroom(Ketwongsa et al., 2022). Another traditional approach for identifying poisonous mushrooms involves using a silver spoon to stir a pot of boiling mushrooms; if the spoon changes from silver to black, it is considered an indicator of toxicity. The aforementioned methods, however, are not exact and reliable since certain deadly mushrooms do not react to them and lack standardization and objective quantification, leading to inconsistencies in results between different practitioners. This subjectivity reduces the reproducibility and reliability of the classification process. Furthermore, traditional methods are not adaptable to the evolving understanding of mushroom toxicity or the discovery of new species. As the scientific knowledge about mushrooms expands, the shortcomings of these methods become more pronounced. Their inability to accommodate new information limits their applicability in a rapidly changing field. These methods are prone to false positives and false negatives, leading to misclassification of mushrooms.

The circle under the cap, multi-coloured scales on the cap, and the presence of vivid are the primary physical traits of deadly mushrooms (Ketwongsa et al., 2022). As a result, the intensive study of the precise, reliable and robust methods for the classification of edible and non-edible mushrooms is inevitable. With the development of artificial intelligence last decades, several machine learning and deep learning-based prediction models have been developed to increase classification accuracy. Although these deep learning models have significantly improved in terms of performance, a lot depends on the quality and quantity of the data utilized to train the model. If there is insufficient data or a lot of noise in the data, the prediction accuracy will drastically decrease due to inaccurate learning (Moon et al., 2020). On the other hand, gathering enough good-quality data is costly and time-consuming. Synthetic data can be of considerably greater quality than actual data since they are created rather than gathered or measured(Chatterjee & Byun, 2023). Furthermore, privacy restrictions can be used to ensure that the synthetic data does not expose any significant information, such the clinical records of patients. Along with this, creating diverse datasets for testing and analysis, and data augmentation are additional benefits of synthetic data.

To address the challenges of limited and imbalanced datasets, researchers have turned to Generative Adversarial Networks (GANs) as a promising solution for tabular data synthesis. GANs

are a class of deep learning models that involve a generator network and a discriminator network engaged in a competitive learning process. The generator network aims to produce synthetic data samples that resemble the real data, while the discriminator network attempts to distinguish between real and synthetic data. Through this adversarial training, GANs have demonstrated remarkable capabilities in generating high-quality synthetic data across various domains. This study is concentrated on creating synthetic tabular data using CopulaGAN(CopulaGAN Model — SDV 0.18.0 Documentation, n.d.), TGAN(Xu & Veeramachaneni, 2018), and CTGAN(Xu et al., 2019) which addressed the following research section.

The summary of the contribution follows:

- Addressing Data Limitations: Limited and imbalanced datasets have been a challenge in mushroom classification. This research aim to address this issue by employing a synthetic tabular data generation techniques using CopulaGAN(CopulaGAN Model — SDV 0.18.0 Documentation, n.d.), TGAN(Xu & Veeramachaneni, 2018), and CTGAN (Xu et al., 2019)for improving the classification accuracy of mushrooms.
- Enhanced Classification Accuracy: This study focuses on the application of GANs, specifically CopulaGAN, TGAN, and CTGAN, to create synthetic tabular data. By combining these synthetic datasets with the original data, the study aims to improve the accuracy of mushroom classification algorithms.
- The comparative analysis of various supervised machine learning algorithms for the classification by using the synthetic data, original data, and combined data.

The paper's organization is divided into five sections: The related works, methods for (tabular) synthetic data generation, the existing research on using GANs to synthesize tabular data and the various methods used to do so are also covered in Section 2. Section 3 Provides a summary of the proposed methodology. Analysis and explanation of the results obtained from experiments are discussed in Section 4. The conclusion and additional recommendations for future works are described in Section 5.

Related Work

Mushroom Classification using Deep learning and Machine Learning

The paper by (Sunita & Bishan, 2015) focuses on the use of classification techniques for analyzing mushroom data sets using WEKA. Several methods namely naive Bayes, Bayes net, and ZeroR are used to identify mushrooms and used accuracy, mean absolute error, and kappa statistic for the evaluation of the performance of the classification techniques. The Bayes net outperformed the other algorithms with the highest accuracy, and lowest mean absolute error. Specifically, a fascinating finding arises in the context of mushroom classification, indicating that greater training set sizes lead to enhanced model accuracy.

The paper(Verma & Dutta, 2018) compared three classification algorithms(ANN, Adaptive Neuro-fuzzy inference System (ANFIS), and Naive Bayes) to classify the mushrooms. The performance of the methods is evaluated using accuracy, MAE, and kappa statistics. Their study revealed ANFIS as the superior method, surpassing others in accuracy (99.8769%), MAE (0.0008), kappa statistics (0.9980), with ANN ranking second (accuracy:96.738%, MAE: 0.0338, kappa statistics: (0.9338). This paper also distinguished that as in (Sunita & Bishan, 2015), accuracy increased as the training size increased. In (Ottom et al., 2019) applied several methods neural networks (NN), support vector machines (SVM), decision trees, and k Nearest Neighbors (KNN) to the image dataset of mushroom for classification task. The findings demonstrate that the most accurate method (accuracy is 94%) for identifying mushroom images is KNN and NN's performance (59% accuracy) is not satisfactory due to the problem of data insufficiency during training phase. The (Chitayae & Sunyoto, 2020) compared KNN and Decision Tree methods and the decision tree algorithm has the highest degree of accuracy (91.93%). Further, the paper (Kousalya et al., 2022)used four classification methods such as Naive Bayes, Decision Tree (C4.5), SVM, and Logistic Regression and the decision tree algorithm has the highest degree of accuracy (93.34%) and is faster than the other algorithms. Also, in (Paudel & Bhatta, 2022)the performance of two Reduced Error Pruning (REP) Tree and Random Forest tree-based classification algorithms are compared and Random Forest method beats the REP Tree technique with a value of 100% for accuracy, precision, and recall.

In the research work carried out by (Wagner et al., 2021), natural language processing was used for the creation of primary data that contains 173 species from 23 families as well as secondary data was also generated. The secondary data is employed as pilot data for the classification and various machine learning algorithms, including naive Bayes, logistic regression, linear discriminant analysis (LDA), and random forests (RF), have been evaluated and the RF provided the best results with a five-fold Cross-Validation accuracy and F2-score of 1.0. Furthermore, the pilot data yielded conclusive outcomes, indicating non-linear separability. This finding underscores the opportunity for the use of alternative methods to generate the synthetic data for classification, suggesting the synthetic data has the potential to enhance classification accuracy. The study by (Ketwongsa et al., 2022) used convolutional neural networks (CNN) and region convolutional neural networks (R-CNN)to distinguish between five dominant types of mushrooms and 98.50% and 95.50% accuracy is achieved respectively. In (Alkronz et al., 2019)a multi-layer ANN model is employed to determine if a mushroom is edible or toxic and only 99.25% accuracy is achieved. This due to insufficient training data and the features present in the dataset. Also, in the paper by (Moon et al., 2020), conditional tabular GAN (CTGAN) is used to solve the problem of data shortage for the electric load. The data used for the training is a mixture of generated data and real data. Their results were very outstanding and concluded that CTGAN is one of the effective ways to produce synthetic data.

Having thoroughly reviewed the existing literature on mushroom classification utilizing various machine learning techniques, we now pivot to our proposed research endeavor. Building upon the insights gained from the limitations and trends highlighted in the prior studies, our research seeks

to contribute to the field by leveraging Generative Adversarial Networks (GANs) for the synthesis of enhanced training datasets. As elucidated by the works of (Sunita & Bishan, 2015), (Verma & Dutta, 2018), (Ottom et al., 2019), and others, challenges such as data insufficiency, imbalanced class distributions, and the need for effective feature representation have been recurrent impediments in achieving higher prediction accuracy. Drawing inspiration from recent advancements in GANs for tabular data synthesis (Bourou et al., 2021), we aim to harness the power of GANs to augment our training data with synthetically generated samples that capture the intricate distributions present in real-world tabular data. This innovative approach holds the potential to address the aforementioned limitations, enabling us to achieve a substantial enhancement in the accuracy and robustness of our mushroom classification model. In the subsequent sections, we will delve into the specifics of our GAN-based data synthesis technique, its design, implementation, and the anticipated benefits it brings to the realm of mushroom classification.

GANs for Data Synthesis

GANs is a breakthrough concept in the field of artificial intelligence developed by Ian Goodfellow and his colleagues in 2014(Goodfellow et al., 2014) to create new data samples that resemble a given dataset. A GAN consists of two main components: generator G and the discriminator D. The generator learns to produce realistic data, and the data it generates serves as negative examples for the discriminator. The discriminator learns to differentiate between the fake data generated by the generator and real data, penalizing the generator for creating implausible results. During training, the generator initially generates obviously fake data, prompting the discriminator to quickly detect its falseness. As training progresses, the generator improves, gradually creating output that can deceive the discriminator. Ultimately, if the generator is successful, the discriminator struggles to distinguish between real and fake data, classifying some fake data as real, and its accuracy drops. Both the generator and the discriminator are neural networks, with the generator's output directly connected to the discriminator's input. Through backpropagation, the discriminator's classification informs the generator's weight updates, allowing it to refine its output based on the feedback received from the discriminator. This adversarial nature of GANs leads to a tug-of-war between the generator and discriminator, pushing them to improve iteratively. When the generator becomes skilled enough to generate data that can deceive the discriminator into misclassifying fake data as real, it indicates that the generator has learned to produce realistic samples.

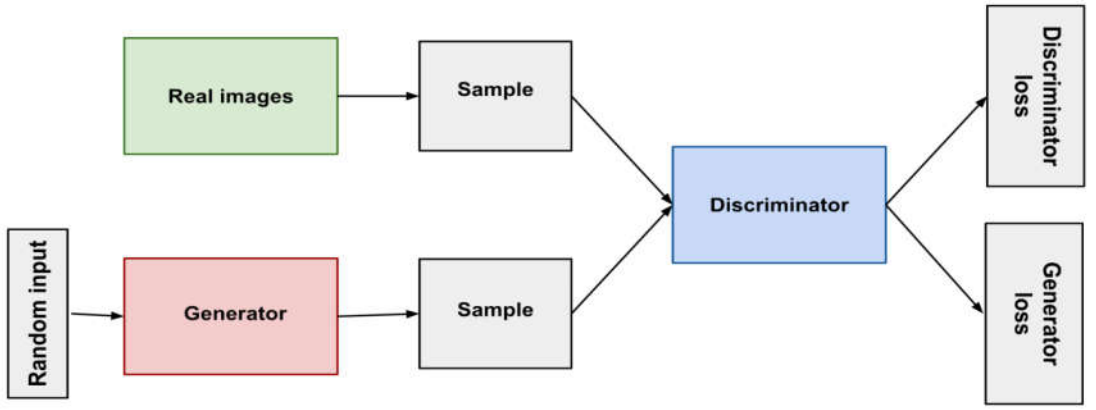

Figure 1 GAN architecture

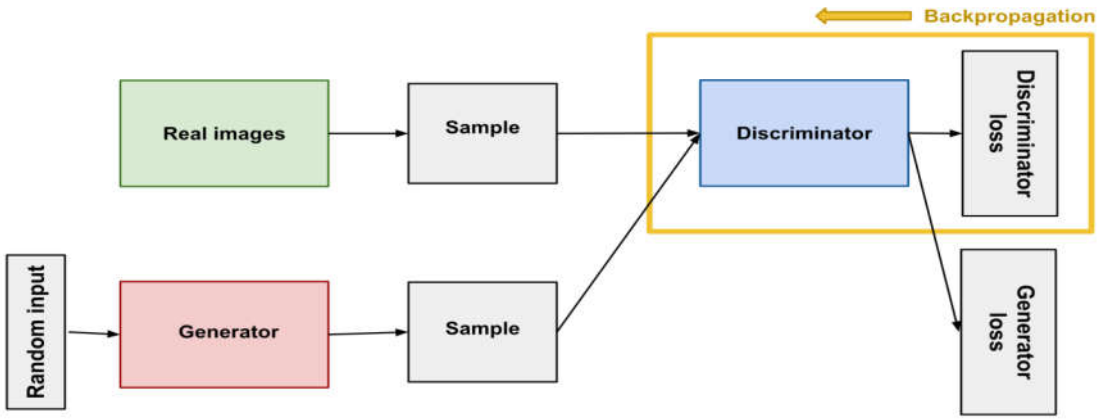

Figure 2 Backpropagation in discriminator training

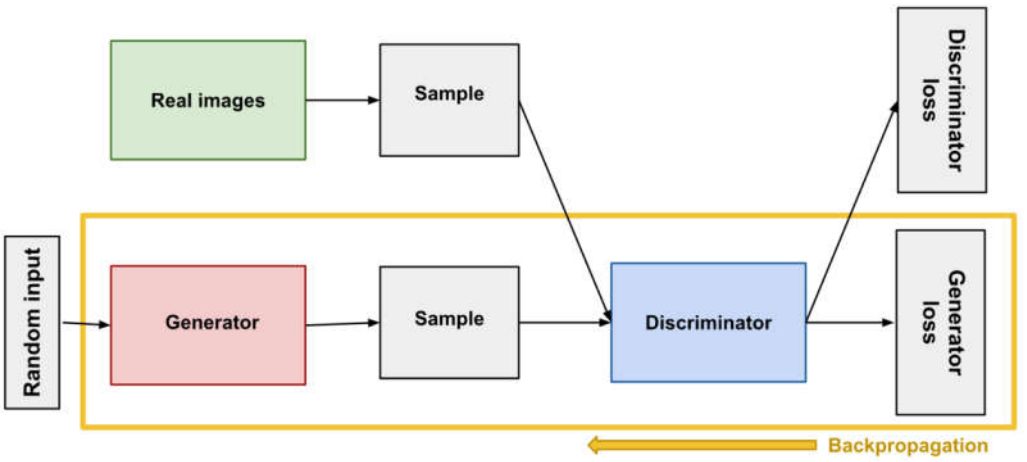

Figure 3 Backpropagation in generator training

The GAN's generative component, denoted as G, grasps the underlying data distribution p(g) in the genuine data space x. By incorporating an input noise variable, G creates novel adversarial instances G(z) designed to mirror the x distribution. The training of Generator G revolves around maximizing the likelihood that the Discriminator D accurately identifies generated examples as authentic, while D's training centers on discerning whether a given sample originates from the real data or has been produced by Generator G. The mathematical formulation of the Vanilla GAN is rooted in the cross-entropy comparison between the actual and generated distributions, and it is expressed as follows(Bourou et al., 2021):

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

For those looking for further understanding into GANs, we suggest consulting the original paper by(Goodfellow et al., 2014).

GANs for Tabular Data Synthesis

Deep neural networks frequently exhibit inferior performance in contrast to more conventional machine learning techniques such as methods based on decision trees when confronted with tabular data(Borisov et al., 2022). Nonetheless, the reasons behind why deep learning struggles to attain equivalent predictive excellence as observed in other domains like image classification, computer vision, and natural language processing often remain ambiguous. According to (Borisov et al., 2022) the four potential major reasons for aforementioned problem are :

A. ***Low Quality Training Data:*** The quality of data poses a prevalent concern in real-world tabular datasets. These datasets commonly exhibit several issues, such as missing values, outliers, data that is incorrect or inconsistent, imbalance in class distribution due to costly nature of data collection, and they also tend to be relatively small in size compared to the high-dimensional feature vectors derived from the data. While these hurdles impact all machine learning algorithms, a majority of contemporary decision tree-based algorithms possess the capability to internally manage missing values and address variations in variable ranges, achieved by identifying suitable approximations and determining split points.

B. ***Missing or Complex Irregular Spatial Dependencies:*** Spatial correlation is frequently absent among variables within tabular datasets, and the interconnections between features often exhibit intricate and irregular patterns. When dealing with tabular data, it becomes necessary to establish the structure and associations among its features through learning from the ground up. Consequently, the inherent biases employed by well-known models designed for uniform data types, like convolutional neural networks, prove inadequate for effectively representing this particular data category.

C. ***Dependency on Pre-processing:*** A key advantage of deep learning on homogeneous data is that it includes an implicit representation learning step, so only a minimal amount of pre-processing or explicit feature construction is required. However, for tabular data and deep neural networks the performance may strongly depend on the selected pre-processing strategy. Handling the categorical features remains particularly challenging and can easily lead to a very sparse feature matrix (e.g., by using a one-hot encoding scheme) or

introduce a synthetic ordering of previously unordered values (e.g., by using an ordinal encoding scheme). Lastly, pre-processing methods for deep neural networks may lead to information loss, leading to a reduction in predictive performance.

D. ***Importance of Single Features:*** While typically changing the class of an image requires a coordinated change in many features, i.e., pixels, the smallest possible change of a categorical (or binary) feature can entirely flip a prediction on tabular data. In contrast to deep neural networks, decision-tree algorithms can handle varying feature importance exceptionally well by selecting a single feature and appropriate threshold (i.e., splitting) values and "ignoring" the rest of the data sample. Individual weight regularization may mitigate this challenge and motivate more work in this direction.

GAN models have demonstrated significant potential in generating synthetic images and text. Recently, researchers have been exploring the application of GANs for generating tabular data due to their ability to effectively model data distributions, which traditional statistical techniques may struggle with. The process involves creating a synthetic table, Tsyn, from an existing table, Treal, consisting of both a training set, Ttrain, and a test set, Ttest. The GAN model is trained on Ttrain, where the data generator G learns the data distribution for each column in the table T and uses this knowledge to generate synthetic data for Tsyn(Bourou et al., 2021).

A successful data generator G for tabular data needs to address various challenges inherent in real-world tabular data. Notably, T can contain mixed data types, including numerical and categorical columns. The numerical columns can have either discrete or continuous values, requiring the Generator G to learn and generate a mix of data types simultaneously. Additionally, the shape distribution of each column can vary, often following non-Gaussian and multimodal patterns, which can cause vanishing gradient problems when applying min-max transformations.In the context of categorical columns in real-world tabular data, an imbalance problem frequently arises, with some classes having significantly more instances than others. This imbalance can lead to mode collapse and inadequate training of the minor classes. Moreover, the presence of sparse one-hot-encoded vectors can cause issues during the training procedure of the Discriminator D, as it may rely on the distribution's rareness rather than the realness of the values to distinguish real from fake data. To overcome these challenges, innovative techniques and tailored approaches are required to develop an effective and robust GAN model for generating high-quality synthetic tabular data.

Methodology

Dataset Description

The dataset utilized in this study is publicly accessible and has been sourced from The Audubon Society Field Guide to North American Mushrooms. This dataset was contributed by Jeff Schlimmer, a contributor associated with the University of California, Irvine (UCI). For those interested in exploring the dataset further, it is available for access at the following URL: https://archive.ics.uci.edu/ml/datasets.html.

Within this dataset, one can explore into a rich collection of 8124 distinct data points, each offering a unique glimpse into the world of mushrooms. These data set are meticulously organized and characterized by 22 different variables. The focus of this dataset centers on a fascinating array of 23 distinct species of fungi, all of which belong to the *Agaricus* and *Lepiota* families. These species, known for their diverse characteristics and intriguing attributes, fall under the category of nominal data type. Mushroom classification datasets used for distinguishing between edible and non-edible mushroom species possess several characteristic features and attributes. Understanding these dataset characteristics is essential for effectively applying data synthesis techniques using GANs. In below table, the key characteristics of datasets are presented.

Feature	**Meaning**	**Representation in Datasets**
Cap Shape	Shape of mushroom cap	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
Cap Surface	Texture of mushroom cap surface	fibrous=f, grooves=g, scaly=y, smooth=s
Cap Color	Color of mushroom cap	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
Bruises	Presence of bruises when touched	bruises=t, no=f
Odor	Scent of the mushroom	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
Gill Attachment	Attachment of gills to cap	attached=a, descending=d, free=f, notched=n
Gill Spacing	Spacing between gills	close=c, crowded=w, distant=d
Gill Size	Size of gills	broad=b, narrow=n
Gill Color	Color of gills	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
Stalk Shape	Shape of the stalk	enlarging=e, tapering=t
Stalk Root	Type of stalk root	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
Stalk Surface Above Ring	Texture of stalk surface above the ring	fibrous=f, scaly=y, silky=k, smooth=s
Stalk Surface Below Ring	Texture of stalk surface below the ring	fibrous=f, scaly=y, silky=k, smooth=s
Stalk Color Above Ring	Color of stalk above the ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
Stalk Color Below Ring	Color of stalk below the ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
Veil Type	Type of veil covering the gills	partial=p, universal=u
Veil Color	Color of the veil	brown=n, orange=o, white=w, yellow=y

Feature	Meaning	Representation in Datasets
Ring Number	Number of rings on the stalk	none=n, one=o, two=t
Ring Type	Type of ring on the stalk	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
Spore Print Color	Color of spore print	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
Population	Density of mushroom sightings	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
Habitat	Environment where mushrooms are found	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

Understanding these characteristic features of mushroom classification datasets is essential for synthesizing realistic and representative synthetic data using GANs. Addressing class imbalance and accurately modelling the complex relationships between the features contribute to the generation of high-quality synthetic mushroom datasets, which can enhance the effectiveness of classification algorithms and promote accurate edible and non-edible mushroom identification.

Tabular Data Generation UsingCopulaGAN

The CopulaGAN(*CopulaGAN Model — SDV 0.18.0 Documentation*, n.d.) model, a modified version of CTGAN available in the SDV open-source library, employs a transformation method based on the Cumulative Distribution Function (CDF) through GaussianCopula. Specifically, CopulaGAN leverages these variations of CTGAN to facilitate data learning. Copulas, rooted in probability theory, depict the associations among random variables. Throughout training, CopulaGAN strives to grasp the data characteristics and structure of the training dataset. Non-numeric and missing data are converted using Reversible Data Transformation (RDT), leading to a completely numerical representation that enables the model to comprehend the probability distributions for each column in the table. Moreover, CopulaGAN endeavors to capture the relationships between the various columns within the table.

Synthetic Data Evaluation

Category Coverage: This measurement assesses the extent to which a synthetic column encompasses all potential categories found within areal column, while disregarding any missing values. The process involves two steps: initially, it determines the count of distinct categories, denoted as "c," existing within the genuine column "r." Subsequently, it tallies the number of these categories that appear in the synthetic column "s." The outcome is a ratio representing the portion of actual categories that have been replicated in the synthetic dataset.

$$\text{score}=\frac{c_s}{c_r}$$

TV Complement: This measurement calculates the resemblance between an authentic column and a fabricated one concerning their shapes, specifically focusing on the marginal distribution or one-dimensional histogram. Designed for discrete, categorical data, this assessment employs the Total

Variation Distance (TVD) to quantify the dissimilarity between the genuine and generated columns. To achieve this, it initiates by determining the occurrence frequency for each category value, subsequently converting it into a probability representation. The TVD metric then gauges disparities in probabilities, as depicted in the provided formula.

$$\delta(R,S) = \frac{1}{2}\sum_{\omega\in\Omega}|R_\omega - S_\omega|$$

Here, ω describes all the possible categories in a column, Ω. Meanwhile, R and S refer to the real and synthetic frequencies for those categories. The TVComplement returns 1-TVD so that a higher score means higher quality.

Decision Tree as a Learning Algorithms

A decision tree is a versatile machine learning algorithm tailored for effectively handling categorical features, which are discrete and qualitative in nature. This algorithm constructs a hierarchical tree-like structure of decisions by recursively partitioning the dataset based on feature values. It begins at the root node and progressively splits the data into branches according to specific feature thresholds. These divisions lead to subgroups, ultimately culminating in terminal nodes where categorical labels or values are assigned. Decision trees excel in scenarios involving categorical features due to their inherent ability to capture complex relationships and interactions between discrete variables. Their interpretability makes them valuable for understanding the decision-making process. Each internal node of the tree represents a decision based on a categorical attribute, while each leaf node corresponds to a class or an outcome. The algorithm is capable of handling nonlinearity and interactions among categorical attributes, making it a valuable tool in tasks such as customer segmentation, recommendation systems, and fraud detection, where understanding intricate categorical patterns is crucial for accurate predictions.

Evaluation of Performance for Machine Learning Models

Accuracy: Accuracy is a widely used performance metric that measures the correctness of a machine learning model's predictions. It calculates the ratio of correctly predicted instances to the total instances in a dataset. While intuitive and easy to interpret, accuracy may be misleading when dealing with imbalanced datasets, where one class has significantly more samples than the other. In such cases, a high accuracy can result from the model simply predicting the majority class. It's essential to consider additional metrics, especially for imbalanced scenarios.

ROC and AUC: The Receiver Operating Characteristic (ROC) curve is a graphical tool that evaluates a model's classification ability across various threshold settings. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity). The Area Under the Curve (AUC) quantifies the overall performance of the ROC curve. AUC ranges from 0 to 1, with a higher value indicating better discrimination between classes. ROC and AUC are valuable for binary classification tasks, helping to assess the model's capability to differentiate between positive and negative instances, regardless of the chosen decision threshold. AUC provides a single scalar value that captures the model's performance across different threshold levels, making it a robust metric for model comparison and selection.

Result and Discussion

The experiments are designed for the investigation of performance, reliability, and validity of Copula GAN for the purpose of generation of synthetic mushroom data. As the main focus of this study involved creating synthetic data through Copula GANs and evaluate the reliability and validity of the data generated. For the aforementioned task, 1500 epochs are used to generate data that closely mirrored the characteristics and distribution of the original dataset. We have generated the synthetic data of the same size as the original dataset because we can compare the distribution of original dataset and to avoid the biasness. A graphical representation of the generator and discriminator loss during training further shed light on the GAN learning dynamics. The plotted loss graph shown in Figure 4 provides a visual understanding of the adversarial competition between the generator and discriminator. The decreasing generator loss and increasing discriminator loss over the training epochs signify a converging process. This convergence suggests that the GAN architecture successfully navigated the intricate task of generating data that aligns more closely with the original distribution. The average data quality of feature ***cap_color*** is the lowest and ***veil-type*** is the highest as can be seen from Figure 5. The quality score of 92 % indicates Copula GANs excel at capturing intricate relationships between multiple categorical variables, making them a promising option for generating data.

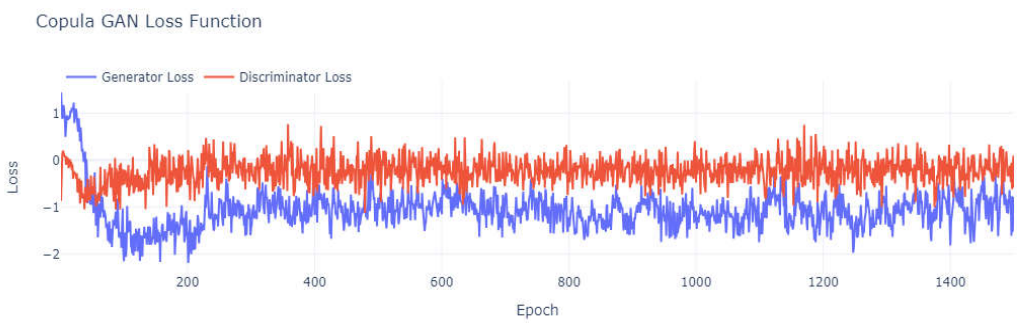

Figure 4 Generator vs Discriminator Loss

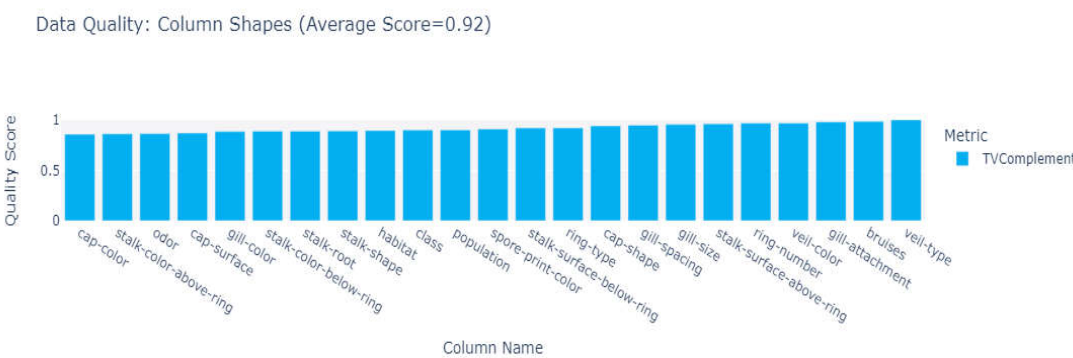

Figure 5 Data Quality of Each Feature

Further, additionally, we used two important measures, Category Coverage and TV Complement, to assess the quality of the generated synthetic data. The achieved average scores of 0.99 and 0.90, respectively, validate the efficacy of our synthetic data. These high scores indicate that the synthetic dataset effectively covers the categories present in the original data and captures the underlying distribution accurately. Accurately replicating the data's natural traits is crucial to ensure that the generated data is valuable for improving classification models.

We used a decision tree model with the Gini impurity criterion to measure the effect of the generated synthetic data on classification accuracy. The model was tested on three distinct datasets: real data only, synthetic data only, and a combined dataset comprising real and synthetic data. The obtained accuracy scores underscore the potential of synthetic data augmentation which can be seen from Figure 6. While the real data achieved an accuracy of 0.99 and the synthetic data achieved 1.0, the combined dataset yielded a remarkable accuracy of 1.0. This outcome accentuates the utility of incorporating high-quality synthetic samples in enhancing the performance of classification models.

Data Type used	**Accuracy**
Real	0.99
Synthetic	**1**
Real + Synthetic	**1**

Figure 6 Accuracy for each type of dataset

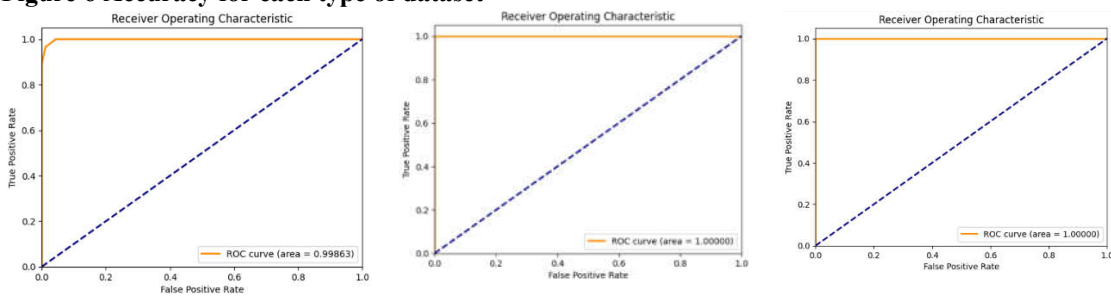

Figure 7 ROC Curve a. Real data b. Synthetic data c. Real+ synthetic data

Conclusion

In this study, we have demonstrated the efficacy of a Copula GAN-based synthetic data technique to enhance the accuracy of mushroom classification. The successful generation of synthetic data that closely replicates the characteristics of the original dataset highlights the potential of generative modelling in addressing the challenges posed by limited data availability. The remarkable evaluation scores for Category Coverage and TV Complement further validate the quality and representativeness of the generated synthetic data.

Our exploration extended to the classification realm, where we employed a decision tree model to assess the impact of synthetic data augmentation. The substantial increase in accuracy achieved

with the combined dataset underscores the practical utility of our approach. The convergence of the generator and discriminator loss during training adds an insightful dimension, shedding light on the internal dynamics of the GAN architecture.

As we look ahead, several avenues for future research beckon. Firstly, investigating the interpretability of the features generated by the GAN and their contribution to model performance could unravel valuable insights. Moreover, extending our approach to diverse domains beyond mushroom classification holds promise. Exploring the applicability of Copula GAN-generated synthetic data in other complex classification tasks could pave the way for novel insights and improvements in various fields. Furthermore, the potential of hybrid approaches, combining GAN-generated data with traditional data augmentation techniques, remains to be explored. Such a fusion could harness the strengths of both methodologies to further enhance classification outcomes. In conclusion, our study introduces a robust approach for data enhancement using Copula GAN-generated synthetic data, with promising implications for classification tasks. The groundwork laid herein opens the door to a realm of possibilities for further advancement and innovation in data synthesis and classification methodologies.

References

Alkronz, E. S., Moghayer, K. A., Meimeh, M., Gazzaz, M., Abu-Nasser, B. S., & Abu-Naser, S. S. (2019). Prediction of whether mushroom is edible or poisonous using back-propagation neural network. *International Journal of Academic and Applied Research (IJAAR)*, *3*(2). 1-8

Bhatt, R. (2016). Wild edible mushrooms from high elevations in the Garhwal Himalaya-I. *Current Research in Environmental & Applied Mycology*, *6*(2), 118–131. https://doi.org/10.5943/cream/6/2/6

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/TNNLS.2022.3229161

Bourou, S., El Saer, A., Velivassaki, T.-H., Voulkidis, A., & Zahariadis, T. (2021). A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. *Information*, *12*(9), 375. https://doi.org/10.3390/info12090375

Chatterjee, S., & Byun, Y.-C. (2023). A Synthetic Data Generation Technique for Enhancement of Prediction Accuracy of Electric Vehicles Demand. *Sensors*, *23*(2), 594.

Chitayae, N., & Sunyoto, A. (2020). Performance Comparison of Mushroom Types Classification Using K-Nearest Neighbor Method and Decision Tree Method. *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 308–313.

CopulaGAN Model—SDV 0.18.0 documentation. (n.d.). Retrieved August 6, 2023, from https://sdv.dev/SDV/user_guides/single_table/copulagan.html

Devika, G., & Karegowda, A. G. (2021). Identification of edible and non-edible mushroom through convolution neural network. *3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021)*, 312–321.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 2672

Kaushik, S., & Choudhury, T. (2022). Mushroom Classification Using MI, PCA, and MIPCA Techniques. In V. Skala, T. P. Singh, T. Choudhury, R. Tomar, & Md. Abul Bashar (Eds.), *Machine Intelligence and Data Science Applications* (pp. 679–693). Springer Nature Singapore.

Ketwongsa, W., Boonlue, S., & Kokaew, U. (2022). A New Deep Learning Model for the Classification of Poisonous and Edible Mushrooms Based on Improved AlexNet Convolutional Neural Network. *Applied Sciences*, *12*(7), 3409. https://doi.org/10.3390/app12073409

Kousalya, K., Krishnakumar, B., Boomika, S., Dharati, N., & Hemavathy, N. (2022). Edible Mushroom Identification Using Machine Learning. *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 1–7.

Moon, J., Jung, S., Park, S., & Hwang, E. (2020). Conditional Tabular GAN-Based Two-Stage Data Generation Scheme for Short-Term Load Forecasting. *IEEE Access*, *8*, 205327–205339. https://doi.org/10.1109/ACCESS.2020.3037063

Ottom, M. A., Alawad, N. A., & Nahar, K. M. (2019). Classification of mushroom fungi using machine learning techniques. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(5), 2378–2385.

Paudel, N., & Bhatta, J. (2022). Mushroom Classification using Random Forest and REP Tree Classifiers. *Nepal Journal of Mathematical Sciences*, *3*(1), 111–116.

Sunita, B., & Bishan, D. (2015). Mushroom classification using data mining techniques. *International Journal of Pharma and Bio Sciences*, *6*(1), 1170.

Verma, S. K., & Dutta, M. (2018). Mushroom classification using ANN and ANFIS algorithm. *IOSR Journal of Engineering (IOSRJEN)*, *8*(01), 94–100.

Wagner, D., Heider, D., & Hattab, G. (2021). Mushroom data creation, curation, and simulation to support classification tasks. *Scientific Reports*, *11*(1), 8134. https://doi.org/10.1038/s41598-021-87602-3

Wibowo, A., Rahayu, Y., Riyanto, A., & Hidayatulloh, T. (2018). Classification algorithm for edible mushroom identification. *2018 International Conference on Information and Communications Technology (ICOIACT)*, 250–253. https://doi.org/10.1109/ICOIACT.2018.8350746

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, *32*.

Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. *ArXiv Preprint ArXiv:1811.11264.*