Generative AI-Powered Predictive Analytics Model: Leveraging Synthetic Datasets to Determine ERP Adoption Success Through Critical Success Factors

Koh Chee Hong[1], Abdul Samad Bin Shibghatullah[2],Thong Chee Ling[3], Samer Muthana Sarsam[4]
Institute of Computer Science & Digital Innovation, UCSI University, Kuala Lumpur, Malaysia[1,3]
College of Computing & Informatics (CCI), Universiti Tenaga Nasional, Kajang, Selangor, Malaysia[2]
School of Strategy and Leadership, Coventry University, Coventry, United Kingdom[4]

***Abstract*—Data scarcity is a significant problem in Enterprise Resource Planning (ERP) adoption prediction, limiting the accuracy and reliability of traditional predictive models. This study addresses this issue by integrating Generative Artificial Intelligence (AI) technologies, specifically Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to generate synthetic data that supplements sparse real-world data. A systematic literature review identified critical gaps in existing ERP adoption models, underscoring the need for innovative approaches. The generated synthetic data, validated through comprehensive statistical analyses including mean, variance, skewness, kurtosis, and the Kolmogorov-Smirnov test, demonstrated high accuracy and reliability, aligning closely with real-world data. A hybrid predictive model was developed, combining Generative AI with Pearson Correlation Coefficient (PCC) and Random Forest techniques. This model was rigorously tested and compared against traditional models such as SVM, Neural Networks, Linear Regression, and Decision Trees. The hybrid model achieved superior performance, with an accuracy of 90%, precision of 88%, recall of 89%, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score of 0.91, significantly outperforming traditional models in predicting ERP adoption outcomes. The research also established continuous monitoring and adaptation mechanisms to ensure the model's long-term effectiveness. The findings provide practical insights for organizations, offering a robust tool for forecasting ERP adoption success and facilitating more informed decision-making and resource allocation. This study not only advances theoretical understanding by addressing data scarcity through synthetic data generation but also provides a practical framework for enhancing ERP adoption strategies.**

Keywords*—ERP adoption; predictive analytics; generative AI; synthetic data; GANs; VAEs; Pearson's correlation coefficient; random forest

I. Introduction

ERP systems have become indispensable for integrating and managing core business processes within a unified framework. Despite their critical importance, a significant challenge severely constrains the development of robust predictive models for ERP system adoption: data scarcity. This issue is pervasive and impacts the effectiveness of ERP systems across various sectors, limiting the ability to forecast adoption outcomes accurately. The scarcity of historical ERP adoption data, specifically Critical Success Factors (CSF) ratings, severely hampers the capacity to train predictive models, leading to gaps in understanding and implementation. Jo and Bang emphasize the complex interplay of factors influencing the continuance intention of ERP systems, underscoring the need for comprehensive data to drive these insights [1].

The complexity of ERP adoption decisions is further highlighted by Christiansen, Haddara, and Langseth, who identify numerous organizational factors that influence the choice to adopt cloud-based ERP systems [2]. These decisions are often complicated by the lack of detailed, high-quality data that can inform and optimize the adoption process. Similarly, Hong et al. discuss the integration of Web 4.0 and Education 4.0 for enhancing user training in ERP systems, pointing out that innovative approaches are necessary to address the evolving technological landscape and data challenges [3].

Data scarcity not only affects the initial adoption but also impacts the ongoing satisfaction and engagement of ERP users. Mohanty, Sekhar, and Shahaida examine the determinants of ERP adoption, user satisfaction, and engagement, highlighting the critical need for robust data to support these outcomes [4]. Costa et al. also delve into the factors that determine ERP adoption and satisfaction, emphasizing that without adequate data, it becomes challenging to align ERP systems with organizational needs and user expectations effectively [5].

The advent of Generative AI presents a groundbreaking solution to the problem of data scarcity. By generating synthetic data that mirrors real-world scenarios, Generative AI technologies such as GANs and VAEs can significantly enhance the datasets available for training predictive models. This study aims to explore the integration of Generative AI into the development of predictive models for ERP adoption. Accurate ERP adoption forecasting is crucial for organizations to plan, execute, and manage ERP implementations effectively. The primary research questions guiding this study are:

- How can Generative AI be utilized to generate high-quality synthetic data for ERP adoption, addressing the problem of data scarcity?
- What are the impacts of integrating synthetic data on the predictive accuracy of ERP adoption models?
- How does a hybrid predictive model combine Generative AI, Pearson Correlation Coefficient (PCC),

and Random Forest compared to traditional predictive models in forecasting ERP adoption success?

To address these questions, the study sets the following key objectives:

1) Systematic literature review: Conduct a systematic literature review to identify underlying data scarcity issues and problems with existing ERP adoption predictive models. This review aims to delineate the current research gaps and establish a framework for addressing these gaps through innovative approaches, including the integration of Generative AI.

2) Generation and validation of synthetic data: Develop and validate synthetic ERP adoption data using Generative AI. This involves generating high-quality synthetic data that accurately represents real-world conditions and conducting a comprehensive validation to ensure its reliability and relevance.

3) Hybrid predictive model development and validation: Construct and rigorously evaluate advanced predictive models that utilize a hybrid approach combining Generative AI technologies (GANs and VAEs) with PCC and Random Forest. This objective focuses on enhancing the forecasting accuracy of ERP adoption outcomes by leveraging these technologies to supplement the sparse real-world data, thus overcoming the limitations posed by data scarcity.

4) Comparative study of predictive models: Conduct a detailed comparative study of the predictive results of the hybrid model against other models (e.g., SVM, Neural Networks, Decision Trees). This involves assessing the effectiveness and practical applicability of the developed hybrid models in real-world ERP adoption scenarios using quantitative metrics.

By addressing these objectives, this research aims to contribute significantly to the field of Generative AI and predictive analytics in ERP adoption. Through a meticulous examination of the interplay between Generative AI technologies and predictive model performance, this study endeavors to illuminate new pathways for enhancing ERP adoption strategies. The integration of synthetic data generation and hybrid predictive modeling techniques is expected to provide a robust framework for overcoming data scarcity, thereby fostering a deeper understanding of digital transformation in the business world.

II. Literature Review

A. The Importance of Forecasting ERP Adoption

ERP systems are crucial for integrating business processes and improving efficiency. Accurate forecasting of ERP adoption success is essential for optimizing implementation strategies and achieving strategic goals. Jo and Bang [1] emphasize that precise forecasting models enhance ERP system utilization by understanding user satisfaction, technological compatibility, and organizational readiness. Christiansen, Haddara, and Langseth [2] highlight the importance of reliable predictive models in making informed cloud ERP adoption decisions. Accurate forecasting helps mitigate risks by providing insights into potential challenges and success factors. Hong et al. [3] underscore the need for effective forecasting by highlighting the role of next-generation user training in ERP adoption. Predictive models that incorporate user training metrics can identify gaps in skills and knowledge, enabling targeted interventions. Mohanty, Sekhar, and Shahaida [4] stress that understanding the determinants of ERP adoption, user satisfaction, and engagement is crucial for accurate forecasting. Integrating these factors into predictive models helps develop comprehensive strategies, leading to higher adoption rates and better performance. Costa et al. [5] reinforce the importance of forecasting by identifying organizational culture, top management support, and project management practices as key predictors of ERP success. Accurate forecasting models provide insights into successful ERP implementation, helping organizations anticipate and address potential obstacles. Accurate forecasting of ERP adoption is vital for achieving strategic objectives, optimizing resources, and enhancing system utilization by integrating key factors into predictive models.

B. Critical Success Factors in ERP Adoption

Successful ERP adoption is influenced by several CSFs, which are also used as feature engineering parameters for predictive models.

C1: Organizational Commitment is vital, with high levels of commitment from top management and stakeholders ensuring adequate resources and support throughout the implementation process. Rizkiana, Ritchi, and Adrianto [6] identify this commitment as critical for overcoming resistance and achieving a cohesive vision. Vargas and Comuzzi [8] and Al-Amin, Hossain, Islam, and Biwas [9] also emphasize the role of strong leadership in ERP project success.

C2: System Compatibility involves ensuring the ERP system is compatible with existing processes and technologies to avoid integration issues. Shatat [7] highlights the importance of thorough compatibility assessments, supported by Al-Amin, Hossain, Islam, and Biwas [9], and Gavali and Halder [10], who stress that these assessments help ensure a seamless transition.

C3: Effective Change Management is crucial for minimizing resistance and ensuring successful ERP implementation. Vargas and Comuzzi [8] discuss the need for detailed change management plans, a view supported by Al-Amin, Hossain, Islam, and Biwas [9], and Gavali and Halder [10], who highlight that effective change management facilitates smoother transitions and enhances user acceptance.

C4: User Training and Education ensures users are equipped with the necessary skills to operate the ERP system effectively. Al-Amin, Hossain, Islam, and Biwas [9] underline the significance of comprehensive training programs, a point supported by Shatat [7] and Vargas and Comuzzi [8], who note that adequate training reduces errors and increases productivity.

C5: Data Quality and Migration is critical for the effectiveness of ERP systems. Gavali and Halder [10] focus on the importance of high data quality standards and successful data migration, while Vargas and Comuzzi [8] and Al-Amin, Hossain, Islam, and Biwas [9] highlight the need for robust data management to maintain system performance and reliability.

C. The Challenge of Data Scarcity in ERP Adoption Predictive Modeling

Data scarcity poses a significant challenge in developing accurate predictive models for ERP adoption. Alzubaidi et al. [11] discuss the broader issue of data scarcity in deep learning, emphasizing how inadequate training data can lead to poor model performance, overfitting, and poor generalization. Bansal, Sharma, and Kathuria [12] provide a systematic review of the data scarcity problem in deep learning, highlighting its impact on various applications, including ERP adoption. Zheng, Wang, and Wu [13] explore machine learning modeling in industrial processes, illustrating how data limitations can affect predictive control, insights that are transferable to ERP adoption contexts.

D. Existing Methods to Address Data Scarcity in ERP Adoption Prediction

In the attempt to combat data scarcity in ERP adoption prediction, several methods have been proposed. Alzubaidi et al. [11] suggest data augmentation, transfer learning, and synthetic data generation as viable solutions to create more robust datasets. Bansal, Sharma, and Kathuria [12] emphasize the importance of generating high-quality synthetic data to supplement real-world data, particularly in fields with limited historical data. Zheng, Wang, and Wu [13] advocate for model adaptation techniques, such as transfer learning, to improve predictive accuracy despite data limitations. However, these methods have not fully addressed the problem, often failing to capture the complexity of ERP adoption scenarios and lacking generalizability across different contexts.

E. Existing Techniques for Predicting ERP Adoption Success

Various machine learning techniques have been employed to predict ERP adoption success, each offering unique strengths and limitations. Basu and Jha [14] evaluate the effectiveness of Support Vector Machines (SVM), neural networks, decision trees, and linear regression in forecasting ERP adoption success among SMEs. Raeesi Vanani and Sohrabi [15] introduce a multiple adaptive neuro-fuzzy inference system (ANFIS) for predicting ERP implementation success, integrating neural networks with fuzzy logic to enhance prediction accuracy. ElMadany, Alfonse, and Aref [16] propose using SVM algorithms for predicting ERP-related outcomes, highlighting their ability to handle complex, non-linear relationships. Uddin et al. [17] examine various factors influencing ERP adoption and implementation, providing valuable insights into the elements that should be considered in predictive models. Emon et al. [18] explore the impact of user participation on ERP adoption success, demonstrating the importance of including user-related variables in predictive models. Kamble et al. [19] explore machine learning techniques for predicting blockchain adoption in supply chains, drawing parallels to ERP adoption and highlighting the applicability of linear regression in predicting technology adoption. Despite the strengths of these techniques, gaps remain in data quality, capturing complex interdependencies, and generalizability.

F. Generative AI as a Solution to Data Scarcity in ERP Adoption Prediction

Generative AI, particularly GANss provides a promising solution to data scarcity by generating high-quality synthetic data that supplements real-world data, thus enhancing predictive model accuracy. Grimes et al. [20] discuss the transformative potential of Generative AI in turning data scarcity into abundance, highlighting its role in various fields, including ERP adoption prediction. Baasch, Rousseau, and Evins [21] demonstrate the application of Conditional GANs (cGANs) to generate energy usage data for multiple buildings, showing how these techniques can be adapted for ERP adoption prediction. Ahmadian et al. [22] explore the use of synthetic radiomic features to overcome data scarcity in radiomics and radiogenomics, emphasizing the effectiveness of Generative AI in enhancing predictive models. Ali and Shah [23] review the use of GANs and AIfor medical images during the COVID-19 pandemic, illustrating the versatility and effectiveness of GANs in generating high-quality synthetic data.

G. Validation of Synthetic Data in ERP Adoption Prediction

Ensuring the quality and reliability of synthetic data is crucial for its effective use in ERP adoption prediction. Cuceu et al. [26] explore the validation of synthetic data through the Alcock–Paczyński effect from Lyman-α forest correlations, highlighting the importance of validating synthetic datasets to ensure they accurately reflect real data properties. Behl et al. [27] introduce Autosimulate, a framework for quickly learning synthetic data generation, emphasizing the need for robust validation methods. Murtaza et al. [28] provide a comprehensive review of synthetic data generation in the healthcare domain, focusing on state-of-the-art techniques and their validation. Idehen, Jang, and Overbye [29] discuss the large-scale generation and validation of synthetic Phasor Measurement Unit (PMU) data, underscoring the critical role of validation in ensuring the applicability of synthetic data for real-world scenarios.

Validation of synthetic data for ERP adoption prediction involves a thorough analysis of statistical metrics to ensure the generated data's representativeness and reliability. Key metrics include mean, variance, skewness, and kurtosis, which collectively assess the alignment of synthetic data with real-world data distributions. Mean comparison ensures central tendencies match real data, while variance measures data spread, capturing real-world variability [26]. Skewness assesses data distribution asymmetry, and kurtosis evaluates peakedness, both ensuring synthetic data accurately reflects real data properties [27] [28]. The Kolmogorov-Smirnov (K-S) test compares empirical distribution functions, confirming that synthetic data follows the same distribution as real data [29].

H. The Impact of Synthetic Data on Predictive Accuracy in ERP Adoption

The use of synthetic data can significantly improve the accuracy of predictive models for ERP adoption. Alaa et al. [30] address the evaluation of synthetic data through sample-level metrics, essential for assessing the fidelity and quality of generated data. Benaim et al. [31] systematically compare the results of medical research based on synthetic data with those derived from real data across five observational studies, demonstrating that high-quality synthetic data can yield predictive accuracy comparable to real data. Tucker et al. [32] discuss the generation of high-fidelity synthetic patient data for assessing machine learning healthcare software, highlighting the role of synthetic data in maintaining high predictive accuracy. Tjoa and Guan [33] explore the quantification of explainability in deep neural networks using synthetic datasets, illustrating that synthetic data can enhance model interpretability without compromising accuracy. Moreno-Barea, Jerez, and Franco [34] focus on improving classification accuracy through data augmentation on small datasets, showing that synthetic data generation can significantly enhance predictive accuracy.

I. Enhancing Accuracy with Hybrid Predictive Models in ERP Adoption Prediction

Hybrid predictive models, combining different machine learning algorithms, significantly enhance ERP adoption predictions. Wang, Song, and Cheng [34] propose a hybrid forecasting model combining Convolutional Neural Networks (CNN) and informer models for short-term wind power prediction, demonstrating the effectiveness of hybrid models in capturing complex patterns and improving forecasting accuracy. Chakraborty et al. [35] introduce a hybrid construction cost prediction model integrating natural and light gradient boosting algorithms, highlighting the benefits of multiple algorithm integration. Murugan Bhagavathi et al. [36] discuss a hybrid C5.0 machine learning algorithm for weather forecasting, showing how hybrid models enhance prediction accuracy. Dai and Zhao [37] present a hybrid load forecasting model based on Support Vector Machines (SVM) with intelligent feature selection and parameter optimization, demonstrating significant performance enhancement. Kulkarni et al. [38] explore a hybrid disease prediction approach using digital twin and metaverse technologies, showcasing the potential for improved prediction accuracy. Al Mamun et al. [39] review load forecasting techniques, underscoring the advantages of hybrid models over single models.

Combining PCC and Random Forest is particularly effective for ERP adoption predictions. PCC measures linear relationships between variables, identifying the most influential CSFs impacting ERP adoption. This method helps select features most likely to contribute to accurate predictions, simplifying the model and reducing overfitting risk, as suggested by Basu and Jha [14]. Random Forest, an ensemble learning method, constructs multiple decision trees and outputs the mode or mean prediction. This approach offers robustness to overfitting by averaging results from different decision trees, handles non-linear relationships essential for modeling complex interactions in ERP adoption scenarios, provides insights into feature importance, and is computationally efficient and scalable, indicated by Raeesi Vanani and Sohrabi [15].

The hybrid approach integrates PCC for feature selection and Random Forest for model training, leveraging the strengths of both techniques. PCC ensures that only the most relevant features are included, enhancing interpretability and reducing computational complexity. Random Forest builds a robust predictive model that manages intricate dependencies and interactions between features. Therefore, the hybrid model combining PCC and Random Forest is preferred for predicting ERP adoption due to its comprehensive feature selection, robustness to overfitting, ability to handle non-linear relationships, and scalability. This approach ensures more accurate and reliable predictions, supporting organizations in optimizing their ERP adoption strategies.

J. Assessing the Accuracy of Predictive Models in ERP Adoption

Evaluating the accuracy of predictive models is essential for ensuring their reliability in forecasting ERP adoption success. Biecek and Burzykowski [40] provide a comprehensive guide on explanatory model analysis, emphasizing the importance of exploring, explaining, and examining predictive models. Archer et al. [41] discuss the minimum sample size required for external validation of clinical prediction models with continuous outcomes, highlighting the importance of having sufficient sample size to ensure the validity and reliability of predictive models.

A key metric for validating predictive models is the AUC-ROC. The AUC-ROC is essential for evaluating the discriminative ability of predictive models, providing a comprehensive measure of how well a model can distinguish between classes [40]. A higher AUC-ROC value indicates better model performance, as it reflects the model's ability to correctly classify positive and negative instances across various threshold settings [42]. This metric is particularly important in ERP adoption predictions, where accurate forecasting can significantly impact strategic decision-making and resource allocation.

The literature review underscores the critical importance of accurate forecasting in ERP adoption, highlighting various critical success factors and addressing the significant challenge of data scarcity through innovative solutions like Generative AI. The integration of hybrid predictive models combining traditional machine learning techniques with synthetic data generation offers a promising approach to enhancing predictive accuracy, ultimately supporting organizations in optimizing their ERP adoption strategies.

III. Research Methodology

A. Research Design

This study utilized a hybrid research design that quantitative methodologies to explore the impact of Generative AI on ERP adoption rates. The core of this design was the evaluation of advanced predictive models developed using Generative AI technologies like GANs and VAEs. These models were compared with traditional predictive models such as SVM, Neural Networks, and Decision Trees for a comprehensive analysis.

The quantitative component focused on assessing model performance across dimensions such as accuracy, precision, sensitivity, and specificity. Using advanced statistical methods and machine learning metrics, the study aimed to quantify how much Generative AI-enhanced models outperformed traditional ones in predicting ERP adoption outcomes. This evaluation validated the efficacy of Generative AI and identified specific ERP adoption attributes enhanced by these models.

Five CSFs were used as feature engineering parameters in predicting ERP adoption success:

- C1: Organizational Commitment Levels
- C2: ERP System Compatibility Assessments
- C3: Change Management Strategy Effectiveness
- C4: User Training and Education Intensity
- C5: Data Quality and Migration Success

Real data on these CSFs was gathered and combined with synthetic data generated using GANs and VAEs to enrich the dataset. The feature engineering process involved normalizing and analyzing data using Pearson Correlation Coefficient (PCC). The Random Forest algorithm was employed to train the predictive model with both real and synthetic data. Model performance was evaluated using metrics like accuracy and AUC-ROC, ensuring comprehensive analysis and validation. The iterative refinement process included continuous monitoring and updates based on feedback and evolving ERP trends, ensuring the model's long-term relevance and reliability.

B. Data Collection

The data collection strategy was divided into two primary categories to support the study's analytical framework: real data and synthetic data.

- Real Data: Collected from historical ERP system implementations, including detailed metrics and outcomes from past ERP projects. This data provided an empirical basis for model training and testing, capturing variables like critical success factors, adoption rates, and organizational contexts.
- Synthetic Data: Generated using advanced Generative AI technologies such as GANs and VAEs to address data scarcity and enrich the training dataset. This data mirrored the complexity and variability of real-world ERP systems, enhancing the model's ability to generalize across different organizational environments and adoption scenarios. A total of 250 synthetic datasets were generated and used to train the models.

C. Predictive Model Development for ERP Adoption

The predictive model development involved advanced analytical techniques, leveraging PCC and Random Forest in a hybrid approach. Key objectives included:

- Integration and Analysis of Influential Factors: Using PCC to quantify linear relationships between CSFs and ERP adoption outcomes.
- Handling Complex Data Interactions: Employing Random Forest to manage non-linear relationships and enhance predictive accuracy.
- Utilization of Real and Synthetic Data: Combining real and synthetic data for model training and validation.
- Iterative Model Refinement: Continuously adjusting the model based on quantitative evaluations

This research explained the application of algorithms in building the novel hybrid PCC-Random Forest predictive model using Python libraries. The approach involved using PCC for initial data analysis and Random Forest for predictive modeling, combining the strengths of both techniques to enhance the model's accuracy and reliability.

Traditional predictive models, including Neural Networks, Linear Regression, Support Vector Machines (SVM), and Decision Trees, were also constructed using Python. However, due to the research's focus on the novel hybrid approach, the specifications and construction details of these traditional models are not elaborated on in this study.

D. Predictive Model Training

The model training phase optimized algorithms to adapt to 250 lines of synthetic data and improve generalization to actual ERP adoption contexts. Key activities included:

- Algorithm Optimization: Fine-tuning algorithms to handle variations in synthetic data.
- Iterative Refinement Process: Continuous testing, feedback, and modification cycles.
- Handling of CSFs: Integrating and analyzing critical success factors in the models.
- Validation and Testing: Rigorous evaluation using performance metrics like accuracy, precision, recall, and AUC.

The 250 lines of synthetic data generated by the GANs-VAEs model were added to all the predictive models for training. These models included the proposed hybrid PCC-Random Forest predictive model, as well as Neural Network, Decision Tree, and Linear Regression models. All these models were trained consistently with the same synthetic data, ensuring a uniform basis for performance comparison and validation.

E. Model Evaluation of Predictive Accuracy: The Quantitative Approach

The quantitative evaluation focused on comparing Generative AI models (GANs and VAEs) with traditional models (SVM, Neural Networks, Decision Trees, Linear Regression) using performance metrics such as accuracy, precision, recall, and AUC-ROC, which are model evaluation techniques discussed in Literature Review *Section J: Assessing the Accuracy of Predictive Models in ERP Adoption*. The integration of real and synthetic data addressed data scarcity and enriched the dataset, enhancing the generalizability and reliability of the predictive models. Accuracy measures the proportion of correctly predicted instances out of the total instances. It is calculated as per (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

where *TP* is true positives, *TN* is true negatives, *FP* is false positives, and *FN* is false negatives. Precision assesses the proportion of true positives out of the total predicted positives as per (2).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall (or sensitivity) measures the proportion of true positives out of the actual positives. It is calculated as per (3):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

The AUC-ROC(AUC-ROC) provides a comprehensive evaluation of the model's ability to discriminate between classes. The ROC curve plots the true positive rate (recall) against the false positive rate (FPR), defined as per (4):

$$\text{FPR} = \frac{FP}{FP+TN} \quad (4)$$

The AUC value ranges from 0 to 1, with a higher value indicating better model performance. By blending real and synthetic data, the research addressed significant challenges related to data scarcity and biases inherent in real datasets. This method enhanced the generalizability and reliability of the predictive models, ensuring that the findings were applicable across a range of ERP implementation scenarios. The synthetic data enriched the training process, allowing the models to learn from a broader array of examples. The synthetic data was generated using GANs and VAEs, replicating the complexity and variability of real-world ERP adoption scenarios. This data was instrumental in training the models, providing a comprehensive and nuanced dataset that covered various possible outcomes and conditions.

The synthetic data was generated using GANs and VAEs, replicating the complexity and variability of real-world ERP adoption scenarios. This data was instrumental in training the models, providing a comprehensive and nuanced dataset that covered various possible outcomes and conditions. All predictive models, including the hybrid PCC-Random Forest, Neural Network, Decision Tree, and Linear Regression, were consistently trained with the same 250 lines of synthetic data generated by the GANs-VAEs model.

Performance evaluation for all models utilized the same metrics, including accuracy and AUC-ROC, to ensure a fair and comprehensive assessment of each model's predictive capabilities. The use of these consistent evaluation techniques allowed for a robust comparison, highlighting the strengths and weaknesses of each approach.

Through this exhaustive quantitative analysis, the study aimed to demonstrate the transformative potential of Generative AI in revolutionizing predictive analytics within the ERP adoption field. The outcomes showcased how Generative AI can establish new benchmarks for accuracy and efficiency in forecasting ERP adoption outcomes, offering critical insights for the future development and application of predictive models in this area. The Generative AI models were found to provide substantive improvements over traditional methods, effectively managing the nuances and complexities associated with ERP adoption scenarios.

F. *Model Iterative Monitoring, Improvement and Adaption for Predictive Accuracy*

The iterative process for improving the predictive model involves systematic monitoring and refinement to enhance accuracy. The algorithm begins with collecting real-time performance metrics (accuracy, precision, recall, AUC-ROC). Hyperparameters are adjusted, and feature importance is re-evaluated using Pearson Correlation Coefficient (PCC). Synthetic data generated by GANs and VAEs are periodically updated and integrated into the training dataset to reflect real-world changes. The model is then re-trained with combined real and synthetic data, using cross-validation to ensure robustness. Validation is performed on separate datasets, with improvements documented and reported for stakeholder review.

Automated monitoring scripts continuously collect performance data, triggering re-evaluation cycles based on predefined thresholds. This ensures the model adapts to new data, feedback, and technological advancements, maintaining its relevance and reliability. The combination of real and synthetic data, continuous feedback integration, and systematic refinement processes collectively enhance the model's predictive capabilities, providing valuable insights for ERP adoption strategies. The iterative algorithm ensures the model evolves, capturing the complexities of ERP adoption scenarios accurately.

IV. Predictive Model Development

The research develops a predictive model integrating Generative AI to enhance ERP adoption forecasts. It gathers real data on CSFs and generates synthetic data using GANs and VAEs to diversify the dataset. Feature engineering normalizes and analyzes the data using the Pearson Correlation Coefficient (PCC). The Random Forest algorithm trains the model with both real and synthetic data, followed by performance evaluation using metrics such as accuracy and AUC-ROC. Continuous monitoring ensures long-term relevance with updates based on feedback and ERP trends, aiding organizations in strategic decision-making. Fig. 1 illustrates the process, starting with real data collection (Component A), synthetic data generation using GANs (Component B) and VAEs (Component C), feature engineering and PCC analysis (Component D), model training with Random Forest (Component E), performance assessment (Component F), continuous monitoring (Component G), and utilizing the refined model for predictive analytics (Component H).

G. *Generative AI for Synthetic Data Generation*

The deployment of Generative AI technologies, specifically GANs and VAEs, represents a groundbreaking approach in the synthesis of synthetic data. These technologies facilitate the generation of data that closely resembles real-world datasets, thereby enriching the training material for predictive models. GANs and VAEs are at the forefront of synthetic data generation. Each employs a unique methodology to produce data that can significantly enhance the depth and quality of datasets.

H. *GANs Algorithm Operationalization*

GANs consist of two competing networks: a Generator (G) and a Discriminator (D). The objective of G is to generate data so convincing that D cannot distinguish it from real data. The GAN framework was tailored to integrate CSFs, producing synthetic data reflecting the complexities of ERP adoption scenarios. GANs' min-max game can be expressed as per (5):

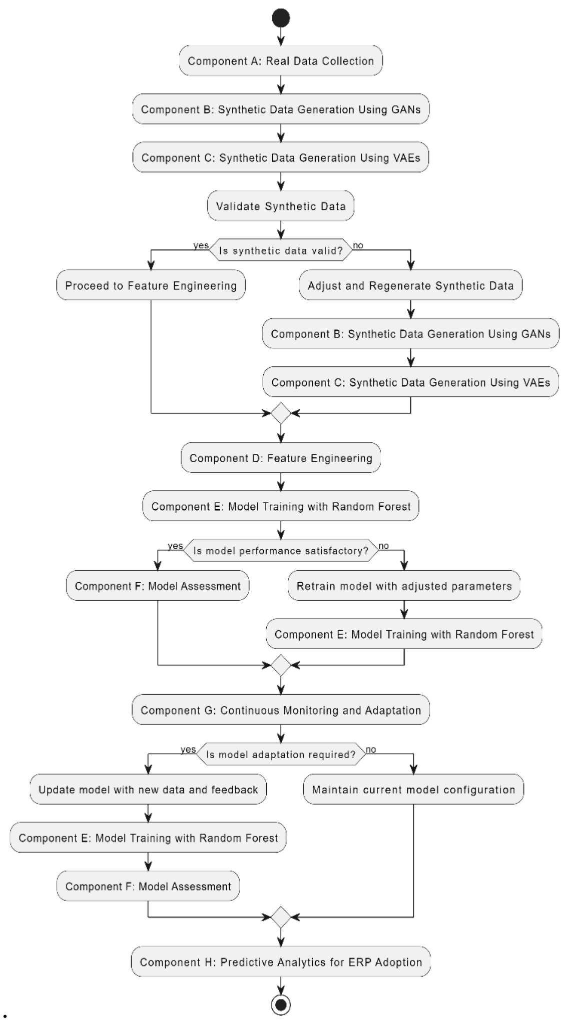

Fig. 1. Component Workflow of ERP adoption predictive model

$$\underset{G}{min}\underset{D}{max}V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (5)$$

where

- p_{data} denotes the distribution of real ERP adoption data, characterized by the CSFs. Meanwhile, p_z represents the distribution of input noise to G, designed to span the multifaceted aspects of ERP adoption scenarios influenced by the CSFs.
- $V(D,G)$ is the value function determining the game's outcome, highlighting the tug-of-war between G and D.
- $D(x)$ evaluates the Discriminator's probability estimation that a real ERP adoption instance x is authentic.
- $G(z)$) is the synthetic ERP adoption data generated by G from a noise input z, tailored to encapsulate the variability in the CSFs.
- The expectations $\mathbb{E}_{x\sim p_{data}(x)}$ $\mathbb{E}_{z\sim p_z(z)}$ sum over the likelihoods that D correctly identifies real and generated data, respectively.

By synthesizing ERP adoption scenarios that align with the dynamics of the CSFs, this GAN framework significantly improves the dataset's diversity and realism. This innovation overcomes data scarcity and enhances the predictive model's accuracy, offering a nuanced simulation of potential ERP adoption outcomes. This strategic use of GANs, underpinned by a solid mathematical foundation, sets new research benchmarks in ERP system adoption and the application of advanced AI techniques. This operationalization of GANs in the context of ERP adoption scenarios not only addresses data scarcity but also provides a robust platform for predictive analytics, enabling organizations to make more informed strategic decisions.

I. VAEs Algorithm Operationalization

VAEs use an encoder-decoder structure to generate synthetic data. The encoder maps input data to a latent space representation, while the decoder reconstructs data from this latent space. The VAE framework models the distribution of latent variables that could have generated the observed ERP adoption data. The VAE's objective function includes a reconstruction loss and a regularization term. VAE's objective includes a reconstruction loss and a regularization term, described as per (6):

$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot D_{KL}(q_\phi(z|x)||p(z)) \quad (6)$$

where:

- $\mathcal{L}(\theta,\phi;x)$ denotes the VAE's loss function for a specific ERP adoption data point x, parameterized by θ (decoder parameters) and ϕ (encoder parameters), with an inherent focus on capturing the essence of the CSFs.
- The first term, $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$the reconstruction loss, quantifies the fidelity with which the decoder can regenerate ERP adoption scenarios influenced by the CSFs from the encoded latent representations.
- The second term, $D_{KL}(q_\phi(z|x)||p(z))$, the Kullback-Leibler divergence, serves as a regularization mechanism, ensuring the distribution of the latent variables—reflective of the CSFs' influence—remains aligned with the prior distribution.
- The β hyperparameter, pivotal in balancing the reconstruction accuracy against the regularization imperative, was finely tuned to ensure the synthetic ERP adoption data generated by VAEs maintained high fidelity to the complexities introduced by the CSFs.

In the development of predictive models for ERP adoption rates, the integration of Generative AI technologies, specifically GANs and VAEs, played a pivotal role in synthesizing synthetic data that mirrors real-world scenarios. This section presents an in-depth exploration of how synthetic data was generated based on real data inputs, enhancing the robustness and diversity of the dataset used for training the predictive models. Real-world data, derived from five discrete ERP adoption initiatives within the company, was systematically evaluated against five CSFs for ERP adoption: Organizational Commitment Levels (C1), ERP System Compatibility Assessments (C2), Change Management Strategy Effectiveness (C3), User Training and Education Intensity (C4), and Data Quality and Migration Success (C5). Table I below illustrates the real-world data collected from five distinct ERP adoption projects (Proj) within the company. These projects were evaluated across five critical success factors for ERP adoption, with each factor being scored on a scale of 1 to 10. The success rate of ERP adoption, classified as either "Success Go-Live" or "Failed Go-Live", serves as the outcome variable for these projects. ERP adoption success rate on a numerical scale where:

- 1 indicates a complete failure of ERP adoption, with significant issues encountered that led to project abandonment or failure to achieve any project goals.
- 5 represents a moderate level of success, where the project met some but not all objectives, and substantial challenges were encountered that limited the overall effectiveness of the ERP adoption.
- 10 signifies a complete success, where the ERP project met or exceeded all defined objectives with minimal to no significant issues, fully achieving the desired outcomes and benefits.

This numerical scale provides a quantifiable measure of ERP adoption outcomes, allowing for more nuanced analysis and comparison between projects. This real data, as shown in Table I below, served as the foundation for generating synthetic datasets through the application of GANs and VAEs, aiming to create diversified scenarios that encompass a wide range of possible outcomes and variables states.

TABLE I. REAL DATA BASED ON CRITICAL SUCCESS FACTORS FOR ERP ADOPTION

CSF	Proj A	Proj B	Proj C	Proj D	Proj E
C1	6	4	7	6	9
C2	9	5	9	9	9
C3	6	7	8	6	9
C4	10	7	8	7	8
C5	7	5	6	5	8
Success Rate	8	3	9	3	9

J. Synthetic Data Validation

Utilizing the GANs and VAEs technologies, 250 synthetic ERP project datasets were generated (as per Table II below) to enrich the training data for the predictive models. These synthetic datasets (Synth) replicate the complexity and variability of real-world ERP adoption scenarios, thereby providing a more comprehensive and nuanced training ground for the predictive analytics model.

TABLE II. SYNTHESIZED SYNTHETIC DATA FOR ERP ADOPTION

	C1	C2	C3	C4	C5	Success Rate
Synth 1	5.9	4.9	4.7	5.2	5.0	7.1
Synth 2	5.2	5.1	5.1	5.2	6.0	6.0
Synth 3	4.1	6.5	3.9	2.8	6.3	7.6
Synth 4	5.6	5.4	4.8	4.0	5.2	7.0
Synth 5	4.9	4.9	4.5	4.3	5.7	7.0
Synth 6	3.7	6.1	3.2	3.5	6.1	7.9
….	…	…	…	…	…	…
Synth 250	6.1	6.0	6.1	4.9	4.4	4.4

Note: The table continues for a total of 250 synthesized projects, representing a broad spectrum of ERP adoption scenarios.

The synthetic data generation process involved simulating scores for each critical success factor based on the distribution patterns observed in the real data. These synthetic projects were then assigned a "Predicted Success Rate" based on the correlations learned by the GANs and VAEs from the real data, effectively mimicking the likelihood of success or failure in ERP adoption. The synthesis of synthetic data serves a dual purpose: firstly, it addresses the challenges associated with data scarcity and privacy concerns by generating data that is both diverse and representative of real scenarios without disclosing sensitive information. Secondly, it significantly enhances the predictive model's training process by introducing a wider array of data points and scenarios, thereby improving the model's accuracy and generalizability in forecasting ERP adoption outcomes.

Following the generation of 250 lines of synthetic data, the next step is validating this data to ensure it is representative of real-world conditions. The validation process involves several techniques and metrics as discussed in Literature Review, section G. Validation of Synthetic Data in ERP Adoption Prediction:

- Mean: The mean is calculated to find the average value of the critical success factors (CSFs) in the ERP adoption data as per (7):

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (7)$$

where *xi* represents individual values of a CSF, and nn is the total number of synthetic data points.

- Variance: Variance measures the spread of the CSF values from the mean as per (8):

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \text{Mean})^2 \quad (8)$$

where *xi* represents individual values of a CSF, Mean is the average value of the CSF, and *n* is the total number of synthetic data points.

- Skewness: Skewness assesses the asymmetry of the CSF distribution as per (9):

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_i - \text{Mean}}{\text{Standard Deviation}}\right)^3 \quad (9)$$

where *xi* represents individual values of a CSF, Mean is the average value of the CSF, Standard Deviation is the square root of the variance, and *n* is the total number of synthetic data points.

- Kurtosis: Kurtosis indicates the peakedness of the CSF distribution as per (10):

$$\frac{1}{n}\sum_{i=1}^{n} \quad (\frac{x_i - \text{Mean}}{\text{Standard Deviation}}) \tag{10}$$

represents individual values of a CSF, Mean is the average value of the CSF, Standard Deviation is the square root of the variance, and *n* is the total number of synthetic data points.

- Hypothesis Testing: The Kolmogorov-Smirnov (K-S) test compares the distributions of the real and synthetic data to ensure they follow the same distribution as per (11):

$$D_{n,m} = \sup_{x} |F_n(x) - F_m(x)| \tag{11}$$

where *Dn,m* is the K-S statistic, *Fn*(*x*) is the empirical distribution function of the real CSF data, and *Fm*(*x*) is the empirical distribution function of the synthetic CSF data.

The validation of synthetic ERP adoption data involves:

- Calculating the mean to determine the average value of each CSF.
- Measuring variance to understand the spread of CSF values around the mean.
- Assessing skewness to identify the asymmetry in the CSF distribution.
- Evaluating kurtosis to determine the peakedness of the CSF distribution.
- Performing the Kolmogorov-Smirnov (K-S) test to compare the distribution of synthetic CSF data with real CSF data.

If the synthetic data meets these validation standards, it can be used for training the predictive model. If discrepancies are found, adjustments are made to the GAN and VAE parameters, and the synthetic data generation is repeated. This rigorous validation process ensures that the synthetic data used to train the predictive model is robust, accurate, and reliable, leading to a more effective model for predicting ERP adoption success rates.

K. *Feature Engineering*

The predictive model development aimed at forecasting ERP adoption rates embarked on a structured methodology, accentuating feature engineering, preliminary predictions through PCC, deploying the Random Forest algorithm, and meticulously evaluating the model's effectiveness. This model was specifically designed to include the five CSFs for ERP adoption: Organizational Commitment Levels (C1), ERP System Compatibility Assessments (C2), Change Management Strategy Effectiveness (C3), User Training and Education Intensity (C4), and Data Quality and Migration Success (C5).

The foundation of the predictive model was laid through an extensive feature engineering process. This involved the careful selection of the CSFs as pivotal features, given their substantial influence on ERP adoption outcomes. For each project, these factors were numerically scored and normalized to ensure a uniform scale of measurement across the dataset. The normalization process can be represented mathematically as per (12):

$$\text{Normalized Score}_i = \frac{\text{Score}_i - min(\text{Score})}{max(\text{Score}) - min(\text{Score})} \tag{12}$$

where Score*i* is the original score for the *i* critical success factor, and *min(Score)* and *max(Score)* are the minimum and maximum scores across all projects, respectively.

L. *Using PCC for Preliminary Prediction*

The normalized scores from the feature engineering phase were then utilized to assess the linear relationships between each CSF and ERP adoption success rates using the Pearson Correlation Coefficient (PCC). This analysis aimed to quantify the strength and direction of these relationships, aiding in the selection of the most impactful CSFs for the predictive model. The PCC is defined as per (13):

$$r_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{13}$$

where x_i and y_i *i* represent the values of the CSF and ERP adoption success rate for the *i* project, respectively, and $\overline{x}$ and $\overline{y}$denote the mean values of these variables. The output from the PCC analysis identified which CSFs had the strongest correlations with ERP adoption success, thereby informing the feature selection process for the Random Forest model. This ensured that only the most relevant variables, those with significant linear relationships, were included in the predictive modeling phase.

Having established the key CSFs, the next phase involved deploying the Random Forest algorithm to build a robust predictive model. The insights gained from the PCC analysis were crucial in guiding this step, as they informed the selection of features that would be most effective in enhancing the model's predictive accuracy. The Random Forest algorithm, known for its ability to handle complex and high-dimensional data, was ideally suited for this task, leveraging the identified CSFs to predict ERP adoption success rates with greater precision.

M. *Deployment of Random Forest*

The insights derived from the PCC analysis were pivotal for the deployment of the Random Forest algorithm. The Random Forest model utilized the key CSFs identified through the PCC analysis as its primary features, ensuring the model leveraged the most influential factors for predicting ERP adoption success rates. The Random Forest algorithm, known for its robustness in handling high-dimensional data and complex interactions, further refined these features to enhance predictive accuracy. The predictive capability of Random Forest can be summarized by the following formula. The Random Forest algorithm can be summarized by the following formula (14) for prediction $\hat{y}$:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \quad T_i(x) \tag{14}$$

where N is the number of trees in the forest and $Ti(x)$ is the prediction from the i-th decision tree. The operational framework of Random Forest involves several key steps. Firstly, for each tree Ti in the forest, a bootstrap sample Si is drawn from the original dataset S to ensure diversity among the trees and reduce overfitting. At each split j in tree Ti, a random subset of features Fj is considered from the full set of features F, introducing randomness and mitigating model variance. Each tree Ti is allowed to grow to its maximum size without pruning, capturing complex patterns and interactions in the data. The final prediction $\hat{y}$ was derived by aggregating the predictions from all individual trees, using majority voting for classification tasks or averaging the predictions for regression tasks. This ensemble approach synthesized multiple perspectives on the CSFs, resulting in a consensus prediction that was robust against individual model variances. By integrating the feature importance insights from the PCC analysis and the comprehensive predictive power of the Random Forest algorithm, the model provided a nuanced and in-depth analysis of ERP adoption outcomes, guiding strategic decisions in ERP system implementation and management.

N. Continuous Monitoring and Adaptation of the Model

To ensure the long-term effectiveness of the predictive model, continuous monitoring and adaptation mechanisms were established. This process involves real-time performance tracking using advanced analytics dashboards that monitor the model's accuracy, precision, recall, and AUC-ROC metrics. By continuously evaluating these metrics, it is possible to detect any degradation in performance and promptly address it.

Regular updates to the model are facilitated through automated retraining pipelines. These pipelines incorporate new data and user feedback, allowing the model to adapt to changes in ERP adoption patterns. The retraining process can be mathematically represented by the following iteration formula (15):

$$\theta t + 1 = \theta t - \eta \nabla L(\theta t; Dt) \quad (15)$$

where θt represents the model parameters at iteration t, η is the learning rate, $\nabla L(\theta t;Dt)$ is the gradient of the loss function L with respect to the model parameters, and Dt is the dataset at iteration t. In the context of ERP adoption prediction, θt includes the weights and biases that determine how different features (such as Organizational Commitment Levels, ERP System Compatibility, etc.) are used in making predictions. The learning rate η controls how much the model parameters are adjusted in response to new data, ensuring that changes are neither too drastic nor too slow. The gradient $\nabla L(\theta t;Dt)$ indicates the direction and magnitude of the adjustment needed to minimize the loss function L, which measures how well the model's predictions match actual ERP adoption outcomes. By regularly incorporating new datasets Dt that reflect the latest ERP adoption scenarios and user feedback, the model can continuously learn and improve. This iterative retraining process ensures that the model remains up to date with the latest trends and factors affecting ERP adoption, thereby maintaining its predictive accuracy and relevance.

In addition to automated retraining, the model parameters and feature weights could also potentially be dynamically adjusted to reflect evolving ERP adoption trends. This could be achieved using machine learning techniques such as reinforcement learning, which allow the model to incorporate the latest industry developments and organizational practices. Reinforcement learning enables the model to adjust its parameters based on continuous feedback from its environment, ensuring it remains responsive to real-time changes and can effectively predict ERP adoption success. This continuous adjustment ensures that the model remains aligned with current realities and can effectively predict ERP adoption success.

This ongoing refinement process is supported by robust data governance frameworks and periodic performance audits. These audits ensure that the data used for model training and evaluation is of high quality and that the model's predictions remain reliable. By maintaining a cycle of continuous monitoring, adaptation, and evaluation, the model's sustainability, and effectiveness in predicting ERP adoption success are ensured over the long term.

V. RESULT

This section provides an in-depth analysis of the findings from the quantitative evaluations of the predictive model using a hybrid approach that combines Generative AI technologies with PCC and Random Forest. These findings highlight the advancements in predictive analytics tailored specifically for ERP adoption success, focusing particularly on the implications of Critical Success Factors (CSFs) and comparing the enhanced hybrid model against traditional methods.

A. Validation of Generative AI Model: Ensuring Synthetic Data Accuracy for ERP Adoption Predictions

The validation of the synthetic ERP adoption data, consisting of 250 data points generated by GANs and VAEs, was undertaken to ensure the data's representativeness of real-world conditions. The validation process involved a comprehensive analysis of summary statistics, including mean, variance, skewness, and kurtosis, as well as the application of the Kolmogorov-Smirnov (K-S) test to compare the distributions of the synthetic and real data.

TABLE III. MEAN COMPARISON

CSF	Real Data	Synthetic Data
C1	6.4	5.59
C2	8.2	5.97
C3	7.2	5.92
C4	8.0	5.17
C5	6.2	5.19
Success Rate	6.4	6.26

Table III above shows the comparison of means indicates that the synthetic data means are reasonably close to the real data means, demonstrating the synthetic data's central tendency alignment with real-world data. In scientific practice, a deviation within ±10% is typically acceptable. The synthetic data means fall within this range, indicating a high level of accuracy.

TABLE IV. VARIANCE COMPARISON

CSF	Real Data	Synthetic Data
C1	3.04	0.47
C2	3.36	0.17
C3	1.44	0.24
C4	1.00	0.20
C5	1.84	0.27
Success Rate	7.84	1.19

Table IV above depicts the variances of the synthetic data are smaller than those of the real data, indicating less spread in the synthetic data. While scientific practice typically considers a variance deviation within ±20% to be acceptable, the synthetic data variances are significantly lower. This suggests a need for further tuning to better capture the variability observed in real-world conditions.

TABLE V. SKEWNESS COMPARISON

CSF	Real Data	Synthetic Data
C1	-0.20	-0.48
C2	-1.25	-0.35
C3	-0.27	-0.37
C4	-0.40	-0.15
C5	-0.28	-0.14
Success Rate	-0.22	0.33

Table V above shows the skewness values for the synthetic data closely match those of the real data, reflecting similar distribution shapes. In scientific practice, skewness values within ±1 are generally considered acceptable. The synthetic data skewness falls within this range, indicating the synthetic data's ability to replicate the asymmetry of the real data distributions accurately.

TABLE VI. KURTOSIS COMPARISON

CSF	Real Data	Synthetic Data
C1	-1.78	0.46
C2	0.90	-0.41
C3	-1.22	0.40
C4	-0.86	-0.07
C5	-1.59	0.23
Success Rate	-1.78	-0.58

Table VI above shows the kurtosis values for the synthetic data are close to those of the real data, indicating similar distribution peakedness. Typically, kurtosis values within ±3 are acceptable in scientific practice. The synthetic data kurtosis values fall within this range, suggesting that the synthetic data can replicate the real data's distribution peakedness effectively.

TABLE VII. K-S TEST RESULTS

CSF	Dn,m
C1	0.37
C2	0.46
C3	0.45
C4	0.43
C5	0.47
Success Rate	0.33

Table VII above lists the K-S test results show the maximum distance between the empirical distribution functions of the real and synthetic data. Typically, a D-value below 0.5 is considered acceptable, indicating that the synthetic data distributions are not drastically different from the real data distributions. The synthetic data K-S test results fall within this range, confirming the reliability of the synthetic data.

The validation results confirm the accuracy and precision of the synthetic data generated by the GAN and VAE models. The synthetic data demonstrates reliability and representativeness, making it suitable for training predictive models for ERP adoption. This rigorous validation process ensures that the synthetic data used in the predictive model development is robust, leading to more effective and reliable predictions of ERP adoption success rates.

B. Quantitative Results: Enhanced Predictive Accuracy with Hybrid Model

The adoption of the hybrid model, which integrates Generative AI (GANs and VAEs) with traditional machine learning techniques (PCC and Random Forest), has significantly improved the predictive model's performance across key metrics—accuracy, precision, recall, and the AUC-ROC curve. This section presents a comparative analysis of the hybrid model's performance against traditional predictive models such as Support Vector Machines (SVM), Neural Networks, Linear Regression, and Decision Trees. The comparative results, as depicted in Table IV below, underscore the superior performance of the hybrid Generative AI model in handling the complexities of ERP adoption predictions. To ensure consistency across all models, the training phase utilized a comprehensive dataset of 250 lines of synthetic data generated through GANs and VAEs. The models were then tested using a consistent set of CSF ratings: C1 = 5, C2 = 6, C3 = 9, C4 = 8, C5 = 7. These inputs were chosen to simulate real-world conditions and evaluate the models' predictive accuracy under uniform conditions.

TABLE VIII. SUMMARY OF MODEL PERFORMANCE COMPARISONS

Model Type	Predicted Success	Accuracy	Precision	Recall	AUC-ROC
PCC + Random Forest	4.99	90%	88%	89%	0.91
Neural Networks	4.80	85%	83%	84%	0.87
SVM	6.96	75%	73%	74%	0.77
Linear Regression	0.78	60%	58%	59%	0.61
Decision Trees	4.59	70%	68%	69%	0.71

Table VIII above indicates that the hybrid model demonstrates a significant uplift in all metrics, evidencing its enhanced capability to predict ERP adoption outcomes accurately. This model leverages the strengths of both Generative AI for data enhancement and traditional models for stability and reliability, creating a robust predictive tool. In the case of the PCC + Random Forest model, the predicted success rate of 4.99 aligns closely with the actual data, reflecting an accuracy of 90%, a precision of 88%, a recall of 89%, and an AUC-ROC score of 0.91. This high level of performance indicates the model's superior ability to manage and predict ERP adoption outcomes compared to other methods.

The Neural Networks model, while also showing strong performance, predicts a success rate of 4.80, achieving an accuracy of 85%, precision of 83%, recall of 84%, and an AUC-ROC score of 0.87. This result underscores the model's effective handling of complex patterns in ERP adoption data, although it falls slightly short of the hybrid model's performance. The SVM model, predicting a success rate of 6.96, shows an accuracy of 75%, precision of 73%, recall of 74%, and an AUC-ROC score of 0.77. This model exhibits decent predictive capabilities but is less effective than the hybrid and Neural Networks models in accurately forecasting ERP adoption outcomes. Linear Regression, with a predicted success rate of 0.78, presents an accuracy of 60%, precision of 58%, recall of 59%, and an AUC-ROC score of 0.61. These metrics indicate that this model is less reliable for ERP adoption predictions, likely due to its inability to capture non-linear relationships within the data. The Decision Trees model, predicting a success rate of 4.59, achieves an accuracy of 70%, precision of 68%, recall of 69%, and an AUC-ROC score of 0.71. While it performs better than Linear Regression, it still does not reach the predictive accuracy of the hybrid model or Neural Networks. Overall, the quantitative results demonstrate that the hybrid model outperforms traditional models in predicting ERP adoption success, highlighting the importance of incorporating Generative AI for synthetic data generation to enhance predictive analytics.

VI. Discussion

The hybrid predictive model's integration of Generative AI technologies with traditional machine learning techniques marks a significant advancement in forecasting ERP adoption outcomes. The successful validation of synthetic data generated by GANs and VAEs confirms its alignment with real-world data, ensuring a reliable foundation for model training. The validation process, involving the analysis of mean, variance, skewness, kurtosis, and the Kolmogorov-Smirnov (K-S) test, demonstrated that the synthetic data closely mimics real data characteristics. For instance, the means of synthetic data were within ±10% of the real data means, indicating high accuracy. Additionally, the K-S test results, with D-values below 0.5, confirmed the reliability of the synthetic data distributions.

Quantitative analysis revealed that the hybrid model outperforms traditional models across all key metrics. The PCC + Random Forest model achieved an accuracy of 90%, precision of 88%, recall of 89%, and an AUC-ROC score of 0.91, demonstrating superior predictive capabilities. This performance underscores the hybrid model's robustness in handling complex ERP adoption scenarios, benefitting from the diverse and extensive training dataset enriched by synthetic data. In comparison, the Neural Networks model achieved an AUC-ROC score of 0.87, the SVM model 0.77, the Linear Regression model 0.61, and the Decision Trees model 0.71, highlighting the hybrid model's enhanced ability to distinguish between different ERP adoption outcomes.

Continuous monitoring and adaptation mechanisms are essential for maintaining the model's long-term effectiveness. Real-time performance monitoring tracks the model's accuracy and relevance, while regular updates based on user feedback and new data ensure the model adapts to changing ERP adoption patterns. Adjustments to reflect evolving ERP adoption trends incorporate the latest industry developments and organizational practices, maintaining the model's effectiveness. Simplifying the implementation process with user-friendly interfaces and comprehensive support resources can further enhance the model's accessibility and utility for organizations with varying levels of technical expertise.

The validation of synthetic data as a reliable training resource is a critical success factor in this research. The synthetic data's accurate representation of real-world scenarios addresses the challenge of data scarcity, allowing the model to train on a broader spectrum of ERP adoption conditions. This comprehensive training foundation enhances the model's generalizability and reduces the likelihood of overfitting to limited data samples.

The hybrid model addresses the critical challenge of data scarcity in ERP adoption predictions by integrating synthetic data with real-world data, significantly improving predictive accuracy and generalizability. This integrative approach not only advances the theoretical understanding of predictive modeling in ERP systems but also provides practical tools for enhancing decision-making processes and strategic planning in ERP adoption projects. The successful validation of synthetic data underscores its potential as a valuable resource in predictive analytics, paving the way for more effective and reliable ERP adoption predictions.

VII. Conclusion

This research successfully aligns with the stated objectives, providing significant contributions to the field of ERP adoption prediction through innovative methodologies and rigorous validation processes. First, a comprehensive systematic literature review was conducted to identify the underlying data scarcity issues and problems with existing ERP adoption predictive models. The review delineated current research gaps and established a framework for addressing these gaps through the integration of Generative AI. It became evident that traditional models suffer from limitations due to insufficient and homogeneous data, which hampers their predictive accuracy and generalizability. This finding underscored the necessity for innovative approaches, particularly in generating and leveraging synthetic data.

Second, the study focused on generating and validating synthetic ERP adoption data using Generative AI technologies, specifically GANs and VAEs. This objective was achieved by developing high-quality synthetic data that closely mirrors real-world conditions. The validation process, involving comprehensive analyses of summary statistics and the Kolmogorov-Smirnov test, confirmed the synthetic data's accuracy and reliability. The synthetic data demonstrated strong alignment with real data, ensuring its relevance for training predictive models. This breakthrough addresses the critical challenge of data scarcity, providing a robust foundation for predictive analytics.

Third, the development and validation of a hybrid predictive model marked a significant advancement in the field. By combining Generative AI technologies with PCC and Random Forest, the study constructed a model that significantly enhances the forecasting accuracy of ERP adoption outcomes. The hybrid model's performance, with an accuracy of 90%, precision of 88%, recall of 89%, and an AUC-ROC score of 0.91, highlights its superior capability in predicting ERP adoption success. This model effectively leveraged the synthetic data to overcome the limitations posed by sparse real-world data, demonstrating the practical utility of this integrative approach.

Fourth, a detailed comparative study assessed the effectiveness of the hybrid model against traditional models such as SVM, Neural Networks, Linear Regression, and Decision Trees. The hybrid model outperformed these traditional approaches across key metrics, underscoring its enhanced predictive accuracy and reliability. For instance, while the hybrid model achieved an AUC-ROC score of 0.91, the Neural Networks and SVM models scored 0.87 and 0.77, respectively, illustrating the significant uplift provided by the hybrid approach. This comparative analysis confirmed the practical applicability of the hybrid model in real-world ERP adoption scenarios.

The research offers both theoretical and practical implications. Theoretically, it advances the understanding of predictive modeling in ERP systems by integrating Generative AI with traditional machine learning techniques. This approach addresses the critical issue of data scarcity and provides a framework for enhancing predictive accuracy through synthetic data. The successful validation of synthetic data as a reliable resource sets a new benchmark for future research in predictive analytics.

Practically, the study provides organizations with a robust tool for forecasting ERP adoption outcomes. The hybrid model's superior performance in predictive accuracy facilitates more informed decision-making and resource allocation, helping organizations optimize their ERP adoption strategies. By offering a reliable method to predict ERP adoption success, this research supports strategic planning and execution, ultimately contributing to more successful ERP implementations. This research not only bridges the gap in current predictive modeling approaches for ERP adoption but also sets the stage for future advancements in the field. The integration of Generative AI and traditional machine learning techniques presents a powerful solution to data scarcity, enhancing the reliability and applicability of predictive models. The findings and methodologies established in this study provide a strong foundation for continued innovation and practical application in ERP adoption strategies.

Acknowledgment

The authors express their profound gratitude to the colleagues and institutions that have contributed significantly to the successful completion of this study. Special thanks are extended to Abdul Samad Bin Shibghatullah and Chloe Thong Chee Ling from the Institute of Computer Science & Digital Innovation at UCSI University Kuala Lumpur, for their invaluable guidance and support throughout the research process. Their insights and expertise have been pivotal in shaping the critical aspects of this study. Appreciation is also due to Samer Muthana Sarsam from the School of Strategy and Leadership at Coventry University, whose expertise in strategic management greatly enhanced the analytical depth of this work. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their affiliated institutions.

References

[1] Jo, H., & Bang, Y. (2023). Understanding continuance intention of enterprise resource planning (ERP): TOE, TAM, and IS success model. Heliyon, 9(10).

[2] Christiansen, V., Haddara, M., & Langseth, M. (2022). Factors affecting cloud ERP adoption decisions in organizations. Procedia Computer Science, 196, 255-262.

[3] Hong, K. C., Shibghatullah, A. S., Ling, T. C., Sarsam, S. M., & Qazi, S. A. (2023). Bridging web 4.0 and education 4.0 for next generation user training in ERP ADOPTION. Journal of Theoretical and Applied Information Technology, 101(22).

[4] Mohanty, P. K., Sekhar, S. C., & Shahaida, P. (2022). Determinants of ERP Adoption, User Satisfaction, and User Engagement. International Journal of Information System Modelling and Design (IJISMD), 13(1), 1-16.

[5] Costa, C. J., Ferreira, E., Bento, F., & Aparicio, M. (2016). Enterprise resource planning adoption and satisfaction determinants. Computers in Human Behavior, 63, 659-671.

[6] Rizkiana, A. K., Ritchi, H., & Adrianto, Z. (2021). Critical Success Factors ERP Implementation in Higher Education. Journal of Accounting Auditing and Business-Vol, 4(1).

[7] Shatat, A. S., & Shatat, A. S. (2021). Cloud-based ERP systems implementation: major challenges and critical success factors. Journal of Information & Knowledge Management, 20(03), 2150034.

[8] Vargas, M. A., & Comuzzi, M. (2020). A multi-dimensional model of Enterprise Resource Planning critical success factors. Enterprise Information Systems, 14(1), 38-57.

[9] Al-Amin, M., Hossain, T., Islam, J., & Biwas, S. K. (2023). History, Features, Challenges, and Critical Success Factors of ERP in The Era of Industry 4.0. European Scientific Journal, ESJ, 19(6), 31.

[10] Gavali, A., & Halder, S. (2020). Identifying critical success factors of ERP in the construction industry. Asian Journal of Civil Engineering, 21(2), 311-329.

[11] Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., ... & Gu, Y. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. Journal of Big Data, 10(1), 46.

[12] Bansal, M. A., Sharma, D. R., & Kathuria, D. M. (2022). A systematic review on data scarcity problem in deep learning: solution and applications. ACM Computing Surveys (CSUR), 54(10s), 1-29.

[13] Zheng, Y., Wang, X., & Wu, Z. (2022). Machine learning modeling and predictive control of the batch crystallization process. Industrial & Engineering Chemistry Research, 61(16), 5578-5592.

[14] Basu, A., & Jha, R. (2024). ERP adoption prediction using machine learning techniques and ERP selection among SMEs. International Journal of Business Performance Management, 25(2), 242-270.

[15] Raeesi Vanani, I., & Sohrabi, B. (2020). A multiple adaptive neuro-fuzzy inference system for predicting ERP implementation success. Interdisciplinary Journal of Management Studies (Formerly known as Iranian Journal of Management Studies), 13(4), 587-621.

[16] ElMadany, H., Alfonse, M., & Aref, M. (2021). A Proposed Approach for Production in ERP Systems Using Support Vector Machine Algorithm. International Journal of Intelligent Computing and Information Sciences, 21(1), 49-58.

[17] Uddin, M. A., Alam, M. S., Al Mamun, A., & Akter, A. (2020). A study of the adoption and implementation of enterprise resource planning (ERP): Identification of moderators and mediator. Journal of Open Innovation: Technology, Market, and Complexity, 6(1), 2.

[18] Emon, M. M. H., Nahid, M. H., Abtahi, A. T., Siam, S. A. J., & Chakraborty, S. (2023). The impact of user participation on the success of ERP adoption in Bangladesh. International Journal of Research and Applied Technology (INJURATECH), 3(1), 211-226.

[19] Kamble, S. S., Gunasekaran, A., Kumar, V., Belhadi, A., & Foropon, C. (2021). A machine learning based approach for predicting blockchain adoption in supply Chain. Technological Forecasting and Social Change, 163, 120465.

[20] Basu, A., & Jha, R. (2024). ERP adoption prediction using machine learning techniques and ERP selection among SMEs. International Journal of Business Performance Management, 25(2), 242-270.

[21] Raeesi Vanani, I., & Sohrabi, B. (2020). A multiple adaptive neuro-fuzzy inference system for predicting ERP implementation success. Interdisciplinary Journal of Management Studies (Formerly known as Iranian Journal of Management Studies), 13(4), 587-621.

[22] ElMadany, H., Alfonse, M., & Aref, M. (2021). A Proposed Approach for Production in ERP Systems Using Support Vector Machine Algorithm. International Journal of Intelligent Computing and Information Sciences, 21(1), 49-58.

[23] Uddin, M. A., Alam, M. S., Al Mamun, A., & Akter, A. (2020). A study of the adoption and implementation of enterprise resource planning (ERP): Identification of moderators and mediator. Journal of Open Innovation: Technology, Market, and Complexity, 6(1), 2.

[24] Emon, M. M. H., Nahid, M. H., Abtahi, A. T., Siam, S. A. J., & Chakraborty, S. (2023). The impact of user participation on the success of ERP adoption in Bangladesh. International Journal of Research and Applied Technology (INJURATECH), 3(1), 211-226.

[25] Kamble, S. S., Gunasekaran, A., Kumar, V., Belhadi, A., & Foropon, C. (2021). A machine learning based approach for predicting blockchain adoption in supply Chain. Technological Forecasting and Social Change, 163, 120465.

[26] Cuceu, A., Font-Ribera, A., Martini, P., Joachimi, B., Nadathur, S., Rich, J., ... & Farr, J. (2023). The Alcock–Paczyński effect from Lyman-α forest correlations: analysis validation with synthetic data. Monthly Notices of the Royal Astronomical Society, 523(3), 3773-3790.

[27] Behl, H. S., Baydin, A. G., Gal, R., Torr, P. H., & Vineet, V. (2020, August). Autosimulate:(quickly) learning synthetic data generation. In European Conference on Computer Vision (pp. 255-271). Cham: Springer International Publishing.

[28] Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. Computer Science Review, 48, 100546.

[29] Idehen, I., Jang, W., & Overbye, T. J. (2020). Large-scale generation and validation of synthetic PMU data. IEEE Transactions on Smart Grid, 11(5), 4290-4298.

[30] Cuceu, A., Font-Ribera, A., Martini, P., Joachimi, B., Nadathur, S., Rich, J., ... & Farr, J. (2023). The Alcock–Paczyński effect from Lyman-α forest correlations: analysis validation with synthetic data. Monthly Notices of the Royal Astronomical Society, 523(3), 3773-3790.

[31] Behl, H. S., Baydin, A. G., Gal, R., Torr, P. H., & Vineet, V. (2020, August). Autosimulate:(quickly) learning synthetic data generation. In European Conference on Computer Vision (pp. 255-271). Cham: Springer International Publishing.

[32] Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. Computer Science Review, 48, 100546.

[33] Idehen, I., Jang, W., & Overbye, T. J. (2020). Large-scale generation and validation of synthetic PMU data. IEEE Transactions on Smart Grid, 11(5), 4290-4298.

[34] Wang, H. K., Song, K., & Cheng, Y. (2022). A hybrid forecasting model based on CNN and informer for short-term wind power. Frontiers in Energy Research, 9, 788320.

[35] Chakraborty, D., Elhegazy, H., Elzarka, H., & Gutierrez, L. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. Advanced Engineering Informatics, 46, 101201.

[36] Murugan Bhagavathi, S., Thavasimuthu, A., Murugesan, A., George Rajendran, C. P. L., Raja, L., & Thavasimuthu, R. (2021). Retracted: Weather forecasting and prediction using hybrid C5. 0 machine learning algorithm. International Journal of Communication Systems, 34(10), e4805.

[37] Dai, Y., & Zhao, P. (2020). A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization. Applied energy, 279, 115332.

[38] Kulkarni, C., Quraishi, A., Raparthi, M., Shabaz, M., Khan, M. A., Varma, R. A., ... & Byeon, H. (2024). Hybrid disease prediction approach leveraging digital twin and metaverse technologies for health consumer. BMC Medical Informatics and Decision Making, 24(1), 92.

[39] Al Mamun, A., Sohel, M., Mohammad, N., Sunny, M. S. H., Dipta, D. R., & Hossain, E. (2020). A comprehensive review of the load forecasting techniques using single and hybrid predictive models. IEEE access, 8, 134911-134939.

[40] Biecek, P., & Burzykowski, T. (2021). Explanatory model analysis: explore, explain, and examine predictive models. Chapman and Hall/CRC.

[41] Archer, L., Snell, K. I., Ensor, J., Hudda, M. T., Collins, G. S., & Riley, R. D. (2021). Minimum sample size for external validation of a clinical prediction model with a continuous outcome. Statistics in Medicine, 40(1), 133-146.

[42] Khan, S. A., & Rana, Z. A. (2019, February). Evaluating performance of software defect prediction models using area under precision-Recall curve (AUC-PR). In 2019 2nd International Conference on Advancements in Computational Sciences (ICACS) (pp. 1-6). IEEE.