International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

Generative AI for Synthetic Medical Data Generation and Privacy Protection

Jenifer.M[1],Jayasree.M[2],Vani sree.A[3]
Assistant professor[1], Department of Computer Application , Student[2,3] ,Student of Computer Applications.
Sri Krishna Arts and Science College Kuniyamuthur-Coimbatore[1,2,3]

***Abstract*—Healthcare data is essential for developing accurate machine learning models, but privacy concerns often limit its accessibility. This study proposes a Generative AI-based approach for generating synthetic healthcare data that preserves the statistical characteristics of real patient data while protecting sensitive information. The methodology involves data preprocessing, synthetic data generation using a generative model, and evaluation through statistical similarity and machine learning performance metrics. The generated synthetic dataset is compared with the original dataset to assess data quality, utility, and privacy preservation. The results demonstrate that synthetic data can effectively support medical analytics without exposing confidential patient information. This approach has the potential to facilitate secure data sharing, improve AI model development, and promote privacy-preserving healthcare research.**

Keywords*—Synthetic Healthcare Data, Generative Artificial Intelligence (Generative AI), Conditional Tabular Generative Adversarial Network (CTGAN), Privacy Preservation, Healthcare Analytics, Machine Learning, Synthetic Data Generation, Electronic Health Records (EHRs), Data Security, Medical Data Analytics.

I. Introduction

The rapid advancement of digital healthcare technologies has transformed the healthcare industry by generating vast amounts of medical data from diverse sources, including Electronic Health Records (EHRs), wearable health monitoring devices, medical imaging systems, laboratory reports, genomic databases, and Internet of Medical Things (IoMT) devices. These large-scale healthcare datasets have become essential for developing Artificial Intelligence (AI) and Machine Learning (ML) models that support disease prediction, early diagnosis, clinical decision-making, treatment planning, personalized medicine, and healthcare resource management. AI-driven healthcare systems have significantly improved the quality, efficiency, and accuracy of medical services by enabling healthcare professionals to make informed decisions based on data-driven insights. Furthermore, the increasing adoption of digital health technologies has accelerated medical research, enhanced patient care, and contributed to the development of innovative healthcare solutions. However, despite these advancements, healthcare data contains highly sensitive personal and clinical information, making patient privacy, data security, and ethical data management major concerns. Healthcare organizations must comply with strict privacy regulations that govern the collection, storage, and sharing of patient information. Consequently, access to high-quality healthcare datasets for research and AI model development is often restricted. Traditional privacy-preserving methods, such as data anonymization and de-identification, provide a certain level of protection but may still be vulnerable to re-identification attacks, particularly when dealing with large, high-dimensional datasets. These limitations create significant challenges for researchers, healthcare institutions, and developers seeking to build accurate and reliable AI models while ensuring patient confidentiality. In addition, limited data availability, class imbalance, and insufficient records for rare diseases further reduce the effectiveness of machine learning models, highlighting the need for alternative approaches to secure healthcare data sharing.[1],[5]

Synthetic healthcare data generation using Generative AI has emerged as an effective and innovative solution to address these challenges by creating artificial datasets that closely resemble the statistical properties and patterns of real healthcare data without exposing sensitive patient information. Advanced generative models, including Generative Adversarial Networks (GANs), Conditional Tabular Generative Adversarial Networks (CTGANs), and Variational Autoencoders (VAEs), have demonstrated remarkable capability in learning complex relationships among healthcare attributes and generating realistic synthetic records suitable for research and AI applications. Unlike conventional privacy-preserving techniques, synthetic data generation produces entirely new patient records rather than modifying existing ones, thereby significantly reducing the risk of patient identification while maintaining the usefulness of the data. Synthetic datasets can be utilized for machine learning model training, testing, validation, healthcare analytics, educational purposes, and collaborative medical research without compromising patient confidentiality. Furthermore, synthetic data helps overcome challenges such as limited data availability, class imbalance, restricted access to rare disease records, and institutional barriers to data sharing. This study proposes a CTGAN-based framework for generating privacy-preserving synthetic healthcare data and evaluates its effectiveness using appropriate measures of statistical similarity, data utility, and privacy preservation. The proposed approach aims to demonstrate that synthetic healthcare data can serve as a reliable alternative to real-world medical datasets, enabling secure data sharing, supporting the development of trustworthy AI models, and promoting innovation in healthcare analytics while maintaining the highest standards of patient privacy and ethical data usage.[6][7]

II. Literature Review

A. Existing Studies

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

The increasing demand for privacy-preserving healthcare analytics has encouraged researchers to explore synthetic healthcare data generation using Generative AI. Several studies have demonstrated that synthetic data can effectively replicate the statistical properties of real healthcare datasets while protecting sensitive patient information. Deep learning models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Conditional Tabular GAN (CTGAN) have been widely adopted for generating realistic synthetic data from electronic health records, medical images, and clinical datasets. These techniques have shown promising results in improving data accessibility, supporting AI model development, and enabling secure data sharing among healthcare organizations.

Researchers have also evaluated synthetic datasets using various statistical and machine learning metrics to ensure data quality, privacy preservation, and practical utility. Existing studies report that synthetic healthcare data can reduce data scarcity, improve model performance, and support research in disease prediction, personalized medicine, and clinical decision-making. Despite these advancements, maintaining an optimal balance between data realism and privacy protection remains a significant challenge.[3][1][8]

B. *Research Gap*

Although significant progress has been made in synthetic healthcare data generation using Generative AI, several research challenges remain. Many existing studies primarily focus on generating realistic synthetic data but provide limited analysis of the balance between data quality, privacy preservation, and practical usability. In addition, the performance of generative models varies depending on the characteristics of the healthcare dataset, making it difficult to identify a single approach that consistently produces high-quality synthetic data across different medical applications.

Furthermore, there is a need for efficient and reproducible frameworks that can generate synthetic healthcare data while maintaining the statistical properties of the original dataset and ensuring patient confidentiality. Existing methods often lack comprehensive evaluation using multiple metrics to assess data similarity, utility, and privacy. Therefore, this study proposes a CTGAN-based approach to generate privacy-preserving synthetic healthcare data and evaluates its effectiveness in supporting secure medical analytics, machine learning model development, and healthcare research. The proposed framework aims to provide a reliable solution for enabling safe data sharing while preserving the accuracy and usefulness of healthcare information.[7][9][4]

III. METHODOLOGY

This study adopts a systematic methodology to generate privacy-preserving synthetic healthcare data using the Conditional Tabular Generative Adversarial Network (CTGAN). The proposed framework is designed to produce realistic synthetic healthcare records that preserve the statistical characteristics of the original dataset while ensuring that sensitive patient information remains confidential. The methodology consists of several sequential stages, including dataset collection, data preprocessing, CTGAN model training, synthetic data generation, and evaluation of the generated data. Each stage plays an important role in ensuring the quality, reliability, and privacy of the synthetic dataset.[5][8]

Initially, a publicly available healthcare dataset is selected and analyzed to understand its structure, attributes, and data distribution. The dataset then undergoes preprocessing to improve its quality by handling missing values, removing duplicate records, encoding categorical variables, and normalizing numerical features. These preprocessing steps ensure that the dataset is suitable for training the CTGAN model and help improve the overall performance of synthetic data generation. After preprocessing, the CTGAN model is trained to learn the complex relationships and statistical patterns present in the original healthcare dataset. Once the training process is completed, the model generates synthetic healthcare records that closely resemble the original data without reproducing actual patient information.

The generated synthetic dataset is then evaluated using multiple performance measures to determine its effectiveness. Statistical similarity is analyzed to compare the distributions of the original and synthetic datasets, while data utility is assessed to verify whether the generated data can support machine learning and healthcare analytics. Privacy preservation is also evaluated to ensure that the synthetic records do not expose confidential patient information or allow re-identification of individuals. Based on these evaluations, the generated synthetic data is analyzed for its suitability in medical research, AI model development, and secure healthcare data sharing. The proposed methodology provides a structured and reliable framework for generating high-quality synthetic healthcare data that supports innovation in healthcare analytics while maintaining patient privacy and data security.[3][10]

A. *Dataset*

The study utilizes a publicly available healthcare dataset for evaluating the proposed CTGAN-based synthetic data generation framework. The dataset contains real-world hospital records that include patient demographic details, admission information, medical conditions, laboratory test results, medications, and treatment outcomes. Since the dataset includes both numerical and categorical attributes, it is well-suited for generative AI-based synthetic data generation in healthcare applications.

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

The selected dataset is analyzed to understand its structure, feature types, and distribution patterns before applying any preprocessing techniques. This step helps in identifying missing values, duplicate records, and inconsistent data entries that may affect model performance. After analysis, necessary preprocessing operations are performed to clean and standardize the dataset so that it can be effectively used for training the CTGAN model.

The cleaned dataset is then used as input for the CTGAN model to learn the underlying statistical relationships between different healthcare attributes. Once trained, the model generates synthetic healthcare records that preserve the original data characteristics without exposing sensitive patient information. These synthetic records are further used for evaluation in terms of statistical similarity, data utility, and privacy preservation.[2][7]

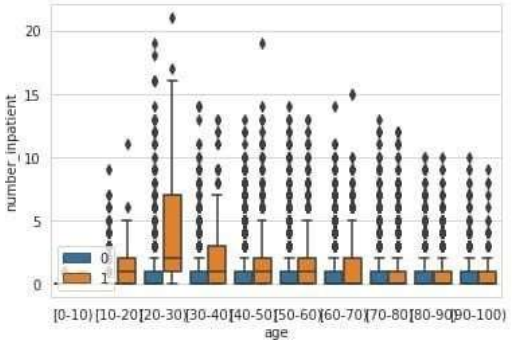

FIG :1

B. Data Preprocessing

Data preprocessing is an essential step in the proposed methodology, as the quality of the input data directly affects the performance of the CTGAN model. The original healthcare dataset may contain missing values, duplicate records, inconsistent data formats, and categorical attributes that require transformation before model training. Therefore, preprocessing is performed to improve data quality, ensure consistency, and prepare the dataset for synthetic data generation.

Initially, missing values are identified and handled using appropriate techniques to minimize information loss. Duplicate records are removed to eliminate redundancy and improve dataset reliability. Categorical variables, such as gender, admission type, and diagnosis, are converted into numerical representations using encoding techniques, while numerical attributes are normalized to maintain a consistent scale across all features. The preprocessed dataset is then validated to ensure that it is clean, complete, and suitable for training the CTGAN model. These preprocessing steps enhance the model's ability to learn the underlying statistical patterns of the healthcare data and generate realistic, high-quality synthetic records.

The preprocessing steps include:

- Data Cleaning: Remove missing values and incorrect or inconsistent entries.
- Duplicate Removal: Eliminate duplicate records to improve data quality.
- Categorical Encoding: Convert categorical features into numerical values for model training.
- Feature Scaling: Normalize numerical attributes to maintain consistency
- Data Validation: Verify the processed dataset before training the CTGAN model[8][1]

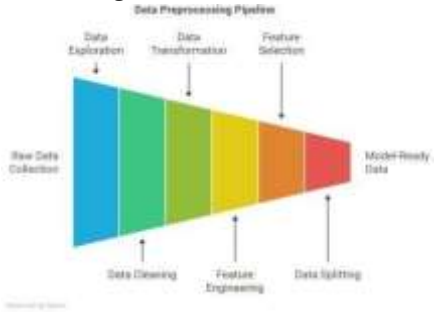

FIG:2

C. Algorithm

The proposed framework employs the Conditional Tabular Generative Adversarial Network (CTGAN) algorithm to generate realistic and privacy-preserving synthetic healthcare data. CTGAN is a deep learning-based generative model specifically designed for tabular datasets containing both numerical and categorical features. It consists of two neural networks: a Generator and a Discriminator, which are trained simultaneously through an adversarial learning process. The Generator creates synthetic healthcare records by learning the statistical patterns and relationships within the original dataset, while the Discriminator distinguishes between real and synthetic records. During training, both networks continuously improve their performance until the generated data closely resembles the characteristics of the original healthcare dataset.

CTGAN consists of two neural networks: a Generator and a Discriminator. The Generator creates synthetic healthcare records by learning the statistical distribution and relationships among the attributes in the original dataset. The Discriminator receives both real and synthetic data and attempts to distinguish between them. During training, the Generator continuously improves its ability to produce realistic synthetic records, while the Discriminator becomes better at identifying generated data. This adversarial learning process continues until the generated synthetic data closely resembles the original healthcare dataset.

CTGAN is particularly suitable for healthcare applications

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

because it effectively handles mixed data types and imbalanced datasets, which are common in medical records. Its conditional generation mechanism enables the model to preserve the distribution of categorical variables while maintaining meaningful relationships among different healthcare attributes. As a result, the generated synthetic data retains high data quality and utility without revealing sensitive patient information. The synthetic dataset is then evaluated using statistical similarity, data utility, and privacy preservation metrics to verify its reliability. The generated data can be safely used for healthcare analytics, machine learning model development, medical research, and educational purposes, providing a secure alternative to real patient data while ensuring compliance with data privacy requirements.[3][8][5]

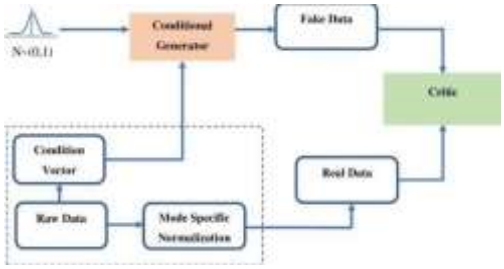

FIG :3

D. Synthetic Data Generation Using Generative AI

Synthetic data generation is the central component of the proposed framework, where the trained CTGAN model is employed to generate artificial healthcare records that closely replicate the statistical distribution of the original dataset. After successful training, the model learns complex, high-dimensional relationships among healthcare attributes such as patient demographics, diagnosis codes, laboratory test results, medication records, and clinical outcomes. Instead of reproducing exact instances from the training data, the model generates new and unique samples that maintain the underlying structure, correlations, and probability distributions of the original dataset. This ensures that the synthetic data is realistic, diverse, and statistically consistent while eliminating the risk of exposing sensitive patient information.

The CTGAN-based generation process operates through a competitive learning mechanism between the generator and discriminator networks. The generator produces synthetic samples by sampling from a learned latent space, while the discriminator evaluates whether a given sample belongs to the real dataset or is synthetically generated. This adversarial process continues iteratively, allowing both networks to improve over time until the generator produces highly realistic synthetic records that are nearly indistinguishable from real data. Additionally, the conditional sampling mechanism in CTGAN ensures balanced generation across categorical variables, which is particularly important in healthcare datasets where class imbalance is common (e.g., rare diseases or underrepresented patient groups).

Another important aspect of the synthetic generation process is its ability to maintain inter-feature dependencies. In healthcare data, variables are often strongly correlated, such as the relationship between age and disease prevalence, medication type and diagnosis, or laboratory results and treatment outcomes. CTGAN preserves these dependencies effectively, ensuring that the generated dataset remains meaningful and suitable for downstream applications. This makes the synthetic data highly useful for training machine learning models without introducing bias caused by missing or imbalanced real-world data.

Finally, the generated synthetic dataset undergoes validation to ensure that it meets the required standards of fidelity, utility, and privacy. Fidelity ensures that the statistical distribution of synthetic data aligns closely with real data, utility ensures that machine learning models trained on synthetic data perform comparably to those trained on real data, and privacy ensures that no individual record can be traced back to real patients. The resulting dataset can then be safely used for healthcare analytics, predictive modeling, academic research, and AI system development. This approach provides a scalable and secure alternative to real patient data, enabling innovation in healthcare while maintaining strict privacy compliance.[2][9][10]

FIG:4

IV. System Architecture

The proposed system architecture illustrates the complete workflow for generating privacy-preserving synthetic healthcare data using the CTGAN model. The architecture is designed to transform raw healthcare data into high-quality synthetic data through a structured pipeline that ensures data quality, statistical consistency, and patient privacy protection. The system integrates data preprocessing, deep learning-based generative modeling, and evaluation techniques to produce reliable synthetic datasets suitable for healthcare analytics and machine learning applications.

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

The architecture begins with the input healthcare dataset, which may contain patient demographic information, clinical records, laboratory results, diagnosis details, and treatment information. This raw data is first passed through a data preprocessing module, where missing values are handled, duplicate records are removed, categorical variables are encoded, and numerical features are normalized. This step

ensures that the dataset is clean, consistent, and suitable for training the CTGAN model.

After preprocessing, the processed dataset is fed into the CTGAN training module, where the generator and discriminator networks are trained in an adversarial manner. The generator learns the underlying distribution of the healthcare data, while the discriminator distinguishes between real and synthetic samples. Through continuous optimization, the generator becomes capable of producing realistic synthetic healthcare records that preserve the statistical properties of the original dataset.[8][4][11]

Once training is complete, the synthetic data generation module produces new artificial healthcare records. These records maintain the relationships between attributes while ensuring that no real patient information is reproduced. The generated dataset is then passed to the evaluation module, where it is assessed using statistical similarity, data utility, and privacy preservation metrics. This ensures that the synthetic data is both realistic and safe for use. The system consists of the following stages:

- Data Collection: A publicly available healthcare dataset is collected, containing patient demographic information, medical history, laboratory test results, medications, and clinical outcomes.
- Data Preprocessing: The collected dataset is cleaned by removing missing values and duplicate records. Categorical features are encoded into numerical values, and numerical attributes are normalized to improve data quality and prepare the data for model training.
- CTGAN Model Training: The preprocessed dataset is provided as input to the CTGAN model. The model learns the underlying statistical distributions and relationships among different healthcare attributes through an adversarial training process.
- Synthetic Data Generation: Once the model is trained, it generates synthetic healthcare records that closely resemble the statistical properties of the original dataset while ensuring that no real patient's information is reproduced.
- Data Evaluation: The generated synthetic dataset is evaluated using statistical similarity, data utility, and privacy preservation metrics to verify its quality, usefulness, and confidentiality.
- Healthcare Analytics: The validated synthetic dataset can be used for machine learning model development, medical research, predictive analytics, education, and secure data sharing without compromising patient privacy.

FIG: 5

V. Results and Discussion

The proposed CTGAN-based framework was evaluated by generating synthetic healthcare data from the preprocessed dataset and comparing it with the original dataset. The generated synthetic data was analyzed to determine how closely it matches the statistical distribution and structural characteristics of the real healthcare data. The evaluation focused on key parameters such as statistical similarity, data utility, and privacy preservation.

The results indicate that the synthetic dataset closely resembles the original dataset in terms of feature distributions and inter-variable relationships. Important healthcare attributes such as patient demographics, diagnosis patterns, and treatment-related information show consistent trends between real and synthetic data. This confirms that the CTGAN model has effectively learned the underlying data distribution and is capable of generating realistic synthetic records.[5][9]

Regarding privacy preservation, the synthetic data does not directly replicate real patient records, thereby reducing the risk of sensitive information leakage. Since the data is artificially generated, it ensures compliance with privacy requirements and supports safe data sharing for research and medical applications.

Overall, the results confirm that the proposed approach successfully generates high-quality synthetic healthcare data that balances realism, utility, and privacy, making it suitable for healthcare analytics and AI-based applications.

The performance of the generated synthetic data is

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

analyzed using multiple evaluation metrics. Statistical similarity is assessed by comparing the distribution of numerical and categorical features in both the original and synthetic datasets. Data utility is evaluated by examining whether the synthetic dataset can support machine learning tasks with performance comparable to that of the original data. Privacy preservation is analyzed to ensure that the generated records do not reveal the identity or confidential information of individual patients. Overall, the findings indicate that the proposed CTGAN model successfully balances data realism and privacy, making the generated synthetic dataset a reliable alternative for secure healthcare research and AI-based medical applications.[1][5][11]

Discussion

- Statistical Similarity: The synthetic dataset closely follows the distribution and characteristics of the original healthcare data, indicating that the model has successfully learned the underlying data patterns.
- Data Utility: The generated synthetic data is suitable for training and testing machine learning models without significantly affecting predictive performance.
- Privacy Preservation: Since the generated records are artificial and do not directly correspond to real patients, the risk of exposing confidential information is significantly reduced.
- Performance Analysis: The CTGAN model effectively handles both numerical and categorical healthcare features, producing realistic synthetic data with high quality and consistency.
- Key Findings: The proposed approach demonstrates that Generative AI can generate reliable, privacy-preserving synthetic healthcare data, supporting secure data sharing, healthcare analytics, and future AI-driven medical research. try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

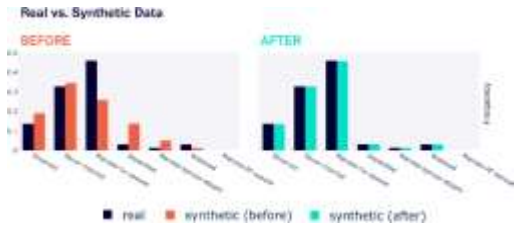

FIG:6

VI. Conclusion

This study proposed a Generative AI-based framework for generating privacy-preserving synthetic healthcare data using the Conditional Tabular Generative Adversarial Network (CTGAN). The approach focuses on transforming sensitive healthcare datasets into synthetic datasets that preserve the statistical properties, patterns, and relationships present in the original data while ensuring that no real patient information is exposed. The methodology includes key stages such as data preprocessing, CTGAN model training, synthetic data generation, and evaluation using metrics like statistical similarity, data utility, and privacy preservation.

The results of this study demonstrate that CTGAN is effective in generating high-quality synthetic healthcare data that closely resembles real-world data distributions. The generated synthetic dataset maintains important feature relationships and can be used for machine learning model training, healthcare analytics, and medical research without compromising patient confidentiality. This makes it a suitable alternative to real healthcare data, especially in scenarios where data access is restricted due to privacy regulations.[4][8]

Overall, the proposed framework highlights the potential of Generative AI in addressing critical challenges in healthcare data sharing, such as data scarcity, privacy concerns, and regulatory limitations. By enabling the generation of realistic yet secure synthetic datasets, this approach supports the development of reliable AI-based healthcare systems and promotes safer data utilization in medical research and analytics.[1][9]

References

[1] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data Using Conditional GAN," Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.
[2] I. J. Goodfellow et al., "Generative Adversarial Networks," Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014.
[3] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," International Conference on Learning Representations (ICLR), 2014.
[4] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," The New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.
[5] "Ethics and Governance of Artificial Intelligence for Health," Geneva, Switzerland, 2021.
[6] "General Data Protection Regulation (GDPR)," Regulation (EU) 2016/679, 2016.
[7] "Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," 1996.
[8] F. Chollet, Deep Learning with Python, 2nd ed. Shelter Island, NY, USA: Manning Publications, 2021.
[9] C. C. Aggarwal, Neural Networks and Deep Learning. Cham, Switzerland: Springer, 2018.
[10] "Diabetes 130-US Hospitals Dataset for Diabetes Readmission

Prediction.

[11] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, pp. 399–410, 2016.