IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

DIFFERENTIATION AND TAGGING OF REAL VS SYNTHETIC DATA

Abhay Pramod, Harish Charan U, Sidharth M, Benil R, Deepa R
Student
Sri Manakula Vinayagar Engineering College

Abstract—This approach addresses data security and authenticity in vital areas like healthcare by combining data watermarking and synthetic data synthesis. Real data is frequently hard to come by because of privacy laws and expenses. We create synthetic datasets that preserve the statistical characteristics of actual data in order to get around this and facilitate efficient model training. But synthetic data raises questions about misuse and dependability. We incorporate undetectable watermarks into genuine data to guarantee authenticity and enable distinct separation from fake data. This maintains confidence and openness while preventing unlawful use. The method works especially well in medical imaging, where diagnosis depends on data integrity. Watermarks improve the security of datasets and guarantee adherence to privacy regulations. Through the use of watermarking and synthetic data, this system reduces data scarcity while confirming authenticity, promoting safe machine.

Keywords – Synthetic Data Generation, Real Data Classification, Watermark Embedding, Neural Networks, Generative Adversarial Networks (GANs), Machine Learning (ML), dual-model approach using Incremental Ensemble Learning (IEL), Secure Data Data Authenticity.

I. INTRODUCTION

Data is the backbone of machine learning, directly influencing model performance, reliability, and generalization. Traditionally, models rely on real-world datasets collected from various sources, including medical imaging, cybersecurity, and financial transactions. However, acquiring high-quality real data is often constrained by privacy regulations, ethical concerns, high costs, and limited availability. Model development is hampered by this shortage, especially in domains where data-driven choices have a substantial influence on people's quality of life. One effective way to address these issues is through the creation of synthetic data. Synthetic datasets can duplicate the statistical characteristics of real data by utilizing Generative Adversarial Networks (GANs) and other machine learning techniques, offering a substitute for AI.

While this approach enhances model scalability and addresses data shortages, it also raises concerns regarding authenticity, reliability, and potential misuse. As synthetic data becomes more sophisticated, differentiating it from real data is critical to ensuring AI models remain trustworthy and unbiased.

In medical imaging, for example, AI-driven diagnostic tools rely heavily on training datasets for accurate predictions. If synthetic data is misused or mixed with real data without proper authentication, diagnostic outcomes may be compromised, leading to incorrect treatment decisions. Similarly, in cybersecurity, adversarial actors could exploit synthetic data to create misleading patterns, impacting security threat detection systems. The lack of transparency in data origins makes it imperative to establish mechanisms for verifying authenticity and traceability.

This study introduces an integrated system combining synthetic data generation with image watermarking to address these challenges. Unlike conventional data verification methods, this approach embeds robust, imperceptible watermarks into real datasets, ensuring authenticity and preventing unauthorized modifications. By explicitly marking synthetic data, the system enhances transparency, making it easier to distinguish between real and artificially generated datasets.

The suggested framework guarantees adherence to privacy laws and moral AI practices in addition to enhancing data security. The incorporation of

watermarking techniques guarantees dataset integrity, reducing risks associated with misinformation and bias. Beyond authentication, the system supports secure data processing, making it particularly relevant in domains where precision and trust are paramount, such as healthcare, financial analytics, and cybersecurity. The ability to trace data sources fosters confidence in AI applications usage in critical fields.

II. RELATED WORK

[1] Synthetic Data in Health Care: A Narrative Review - Gonzales, S. S., et al.

His narrative review delves into the rapidly developing topic of healthcare synthetic data. It covers a range of synthetic data creation techniques, such as machine learning methods like Generative Adversarial Networks (GANs) and statistical approaches. The potential of synthetic data to improve data accessibility for research and development is highlighted in the report while addressing privacy concerns inherent in real patient data. Gives a thorough review of the uses of synthetic data in healthcare, emphasizing the possible advantages for research facilitation and data accessibility. May lack detailed technical evaluations of specific synthetic data generation methods.

[2] Ensemble Learning for Disease Prediction: A Review - Mahajan, P., et al.

The existing research on the use of ensemble machine learning techniques—bagging, boosting, and stacking—for disease prediction is examined in detail in this review. It provides a complete evaluation of the literature on disease prediction models utilizing various ensemble classes and offers an overview of the ensemble technique. Offers an in-depth analysis of ensemble learning techniques applied to disease prediction, discussing various methods like bagging, boosting, and stacking. Focuses primarily on disease prediction, potentially limiting insights into other healthcare applications.

[3] Deep Ensemble Learning Approaches in Healthcare to Enhance the Prediction and Diagnosing Performance: The Workflows, Deployments, and Surveys on the Statistical, Image-Based, and Sequential Datasets - D.-K., et al.

This paper explores deep ensemble learning methods across various data types in healthcare, providing insights into workflows and deployment strategies. It discusses the application of deep learning models in ensemble frameworks to improve predictive performance in healthcare settings. Explores deep ensemble learning methods across various data types in healthcare, providing insights into workflows and deployment strategies. May be complex for readers without a deep learning background.

[4] Methods for Generating and Evaluating Synthetic Longitudinal Patient Data: A Systematic Review- Perkonoja, K., et al.

This systematic review focuses on synthetic data generation for longitudinal patient records, discussing various methods and evaluation techniques. It provides a comprehensive analysis of existing approaches to creating synthetic longitudinal data in healthcare. Focuses on synthetic data generation for longitudinal patient records, discussing various methods and evaluation techniques. Concentrates on longitudinal data, which may not be applicable to all healthcare data types.

[5] A Review of Generative Adversarial Networks for Electronic Health Records: Applications, Evaluation Measures, and Data Sources - Ghosheh, G., et al.

This paper provides a detailed examination of GAN applications in Electronic Health Records (EHRs), including evaluation metrics and data sources. It discusses how GANs can be utilized to generate synthetic EHR data for various applications. Provides a detailed examination of GAN applications in EHRs, including evaluation metrics and data sources. Primarily focuses on GANs, potentially overlooking other synthetic data generation methods.

[6] Deep Neural Network-Based Ensemble Learning Algorithms for the Healthcare System (Diagnosis of Chronic Diseases) - Abdollahi, J., et al.

This study discusses the application of deep neural network ensemble methods in diagnosing chronic diseases, highlighting improved accuracy. It explores how combining multiple neural network models can enhance diagnostic performance. Discusses the application of deep neural network ensemble methods in diagnosing chronic diseases, highlighting improved accuracy. May not cover non-chronic disease applications extensively.

[7] Synthetic Data Generation in Healthcare: A Scoping Review of Reviews on Domains, Motivations, and Future Applications - Rujas, M., et al.

This scoping review analyzes various healthcare domains where synthetic data is applied, discussing motivations and future uses. It gives a summary of previous reviews on the creation of synthetic data in the medical field. Analyzes various healthcare domains where synthetic data is applied, discussing motivations and future uses. Being a scoping review, it may lack detailed methodological insights.

III. LITERATURE REVIEW OF RELATED WORKS

SL.NO	TITLE	AUTHOR	YEAR	ADVANTAGE	DISADVANTAGE
1	Synthetic Data in Health Care: A Narrative Review	Aldren Gonzales, Guruprabha Guruswamy, Scott R. Smith	2022	Gives a thorough review of the uses of synthetic data in healthcare, emphasizing the possible advantages for research facilitation and data accessibility.	May lack detailed technical evaluations of specific synthetic data generation methods.
2	Ensemble Learning for Disease Prediction: A Review	Palak Mahajan, Shahadat Uddin, Farshid Hajati, Mohammad Ali Moni	2023	The study highlights the importance of static code analysis tools in enhancing code quality and provides practical guidelines for educators.	The study did not find acquire impacts from collaboration mode or teaching method on code quality, limiting insights on these factors.
3	Deep Ensemble Learning Approaches in Healthcare to Enhance the Prediction and Diagnosing Performance: The Workflows, Deployments, and Surveys on the Statistical, Image-Based, and Sequential Datasets	Duc-Khanh Nguyen, Chung-Hsien Lan, Chien-Lung Chan	2021	Explores deep ensemble learning methods across various data types in healthcare, providing insights into workflows and deployment strategies.	May be complex for readers without a deep learning background.
4	Methods for Generating and Evaluating Synthetic Longitudinal Patient Data: A Systematic Review	Katariina Perkonoja, Kari Auranen, Joni Virta	2023	Focuses on synthetic data generation for longitudinal patient records, discussing various methods and evaluation techniques.	Concentrates on longitudinal data, which may not be applicable to all healthcare data types.
5	A Review of Generative Adversarial Networks for Electronic Health Records: Applications, Evaluation Measures, and Data Sources	Ghadeer Ghosheh, Jin Li, Tingting Zhu	2022	Provides a detailed examination of GAN applications in EHRs, including evaluation metrics and data sources.	Primarily focuses on GANs, potentially overlooking other synthetic data generation methods.
6	Deep Neural Network-Based Ensemble Learning Algorithms for the Healthcare System (Diagnosis of Chronic Diseases)	Jafar Abdollahi, Babak Nouri-Moghaddam, Mehdi Ghazanfari	2021	Discusses the application of deep neural network ensemble methods in diagnosing chronic diseases, highlighting improved accuracy.	May not cover non-chronic disease applications extensively.

7	Synthetic Data Generation in Healthcare: A Scoping Review of Reviews on Domains, Motivations, and Future Applications	Miguel Rujas, Rodrigo Martín Gómez del Moral Herranz, Giuseppe Fico, Beatriz Merino-Barbancho	2024	Analyzes various healthcare domains where synthetic data is applied, discussing motivations and future uses.	Being a scoping review, it may lack detailed methodological insights.

IV. PROPOSED SYSTEM

In machine learning applications, the suggested technique improves the efficacy of synthetic data, especially in the healthcare industry. We combine several synthetic data production methods, such as Copula-based synthesis and Generative Adversarial Networks (GANs), to overcome the shortcomings of current systems and produce high-quality datasets that closely mimic real-world data. By improving the availability of diverse medical datasets, the system supports robust AI model training while maintaining privacy compliance.

To evaluate the usability and reliability of synthetic datasets, we employ Incremental Ensemble Learning models, such as Adaptive Random Forest classifiers, Softmax Regressor, and K-Nearest Neighbors (KNN). We combine several synthetic data production methods, such as Copula-based synthesis and Generative Adversarial Networks (GANs), to overcome the shortcomings of current systems and produce high-quality datasets that closely mimic real-world data.

To ensure data security and authenticity, we introduce an innovative watermarking mechanism for real patient reports derived from healthcare data. This approach embeds imperceptible watermarks in real medical reports to enhance integrity and prevent unauthorized modifications. Synthetic datasets remain unwatermarked to support research and simulation use cases.

Deployed as a scalable framework, the system combines synthetic data classification, patient report generation, and watermarking to improve data reliability. In healthcare analytics and medical imaging, the incorporation of cutting-edge machine learning methods with security protocols encourages moral AI practices.

A . DATA COLLECTION

Data collection in this study involves acquiring both real and synthetic datasets to support machine learning applications in lung cancer risk factor analysis. Real data is obtained from publicly available medical databases, hospital records, and research repositories. The dataset includes patient demographics, clinical test results, medical history, and lifestyle factors, ensuring a diverse and representative sample. Preprocessing steps such as data cleaning, anonymization, and feature selection are performed to enhance data quality and ensure compliance with privacy regulations. Pearson's correlation coefficient is utilized to assess feature relationships and maintain data integrity.

B . PRE-PROCESSING:

In order to guarantee the quality, consistency, and dependability of the real and synthetic datasets utilized in this study, data preparation is an essential step. The collected real-world data undergoes several preprocessing techniques, including data cleaning, normalization, and feature selection.

Missing values are handled through imputation techniques, while duplicate and inconsistent records are removed to maintain data integrity.

C. MODEL TRAINING

In this study, two Incremental Ensemble Learning (IEL) models are developed to evaluate the classification performance of synthetic versus real data. These models incorporate adaptive learning techniques to handle evolving datasets, ensuring robustness and accuracy in lung cancer risk factor analysis.

The first IEL model consists of an Adaptive Random Forest classifier, which dynamically updates its decision trees based on new incoming data, making it well-suited for continuous learning. It is complemented by a Softmax Regressor, which ensures efficient multi-class classification by assigning probabilities to different categories. The second IEL model integrates K-Nearest Neighbors (KNN) with Adaptive Random Forest to improve classification performance by leveraging distance-based learning along with ensemble techniques.

These models are created to address the challenges of real and synthetic data differentiation while improving classification accuracy. Traditional static models often fail to adapt to new patterns, whereas IEL models continuously refine their learning, making them ideal for healthcare applications where data distributions may evolve over time. Additionally, their ability to process both real and synthetic datasets enables a comprehensive evaluation of how well synthetic data can replicate real-world distributions, ultimately contributing to secure and reliable ML-driven medical research.

A. SECURITY FEATURES

This study uses watermarking as a security method to guarantee the integrity and authenticity of actual healthcare data while guarding against abuse or unauthorized changes. Since synthetic datasets are generated for research and simulation purposes, only real patient reports derived from authentic healthcare data are embedded with watermarks. This approach helps differentiate genuine medical documents from synthetic ones, reducing the risk of fraudulent use.

A robust, imperceptible watermarking technique is applied to real patient reports, embedding unique identifiers such as encrypted patient-specific metadata, hospital codes, or timestamps. The watermark remains invisible to users but can be extracted and verified when needed, ensuring data authenticity. Additionally, watermarking protects against unauthorized tampering, as any modification to the report will alter the embedded watermark, signaling potential data breaches.

B. RESULT AND DISCUSSION

The effectiveness of Incremental Ensemble Learning (IEL) models trained on both synthetic and real datasets is used to assess the suggested system. To evaluate how well synthetic data replicates real-world distributions, the models' classification accuracy, precision, recall, and F1-score are examined. Pearson's correlation coefficient is used to measure the similarity between real and synthetic datasets, revealing a strong positive correlation, especially for GAN-generated data, which demonstrates superior performance in capturing complex patterns.

Among the IEL models, the combination of Adaptive Random Forest and Softmax Regressor achieves higher classification accuracy compared to the Adaptive Random Forest and KNN model, indicating that probabilistic classification enhances the differentiation of data categories. The findings validate the high fidelity of synthetic data in lung cancer risk factor analysis by demonstrating that models trained on GAN-generated synthetic data perform similarly to those trained on real data.

CONCLUSION

This study presents a novel approach to addressing the growing demand for high-quality datasets in machine learning, particularly in healthcare applications such as lung cancer risk factor analysis. The suggested approach successfully assesses the classification performance of synthetic versus actual data by combining Incremental Ensemble Learning (IEL) techniques with synthetic data generation methods, such as Copula-based models and Generative Adversarial Networks (GANs). The results demonstrate that GAN-generated synthetic data closely mirrors real-world distributions, achieving high classification accuracy and proving its applicability in medical research.

REFERENCE

[1] Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. IEEE International Conference on Data Science and Advanced Analytics (DSAA). This paper introduces the Synthetic Data Vault (SDV) for generating synthetic data and compares its effectiveness with real data for model training.

[2] Yoshua Bengio et al. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence. This is a foundational paper on data representations and synthetic data applications in representation learning

[3] Goodfellow, I., et al. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS). [4] Frid-Adar, M., et al. (2018). GAN-based synthetic Abro, G.E.M., & Shaikh, S.A. (2018). Prototyping IoT is based on smart wearable jacket design for securing the life of coal miners. 2018 International Conference on Computing, Electronics & Communications Engineering

[4] Gang Liu, Shifang Cai, Ce Wang, "Speech Emotion Recognition based on Emotion Perception" published in the year of 2023.

[5] Bińkowski, M., et al. (2018). Demystifying MMD GANs. International Conference on Learning Representations (ICLR). This paper evaluates metrics for assessing the quality of synthetic data generated by GANs, which can be useful in comparing real vs. synthetic data.

[6] Bińkowski, M., et al. (2018). Demystifying MMD GANs. International Conference on Learning Representations (ICLR). This paper evaluates metrics for assessing the quality of synthetic data generated by

GANs, which can be useful in comparing real vs. synthetic data.

[7] N. A. A. Khleel and K. Nehéz, “Deep convolutional neural network model for bad code smells detection based on oversampling method,” Indones. J. Electr. Eng. Comput. Sci., vol. 26, no. 3, pp. 1725–1735, Jun. 2022, doi:

[8] J. A. Harer, L. Y. Kim, R. L. Russell, O. Ozdemir, L. R. Kosta, A. Rangamani, et al., "Automated software vulnerability detection with machine learning", *CoRR*, vol. abs/1803.04497, 2018.

[9] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, et al., "Vuldeepecker: A deep learning-based system for vulnerability detection", *CoRR*, vol. abs/1801.01681, 2018.

[10] S. Jain and A. Saha, “Rank-based univariate feature selection methods on machine learning classifiers for code smell detection,” Evol. Intell., vol. 15, no. 1, pp. 609–638, Mar. 2022, doi: 10.1007/S12065-020-00536- Z.

EGAPSO based on similarity measures,” Alexandria Eng. J., vol. 57, no. 3, pp. 1631–1642, Sep. 2018, doi: 10.1016/J.AEJ.2017.07.006.