JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DIFFERENTIATION ANDTAGGING OFREALVs SYNTHETIC DATA

Deepa R[1], Abhay Pramod[2], Benil R[3],Sidharth M[4],Harish Charan U[5]

Assistant Professor[1], Student[2], Student [3], Student[4], Student[5]
Computer Science and Engineering,
Sri Manakula Vinayagar Engineering College, Puducherry,India

Abstract : In our proposed work, we present a novel approach to address the increasing demand for large, high-quality datasets in machine learning (ML), particularly in the healthcare domain. Synthetic data is generated using four methods, including Copula and Generative Adversarial Networks (GANs), and evaluated for its applicability in lung cancer risk factor analysis. Incremental Ensemble Learning models, comprising Adaptive Random Forest classifiers, Softmax Regressor, and K-Nearest Neighbors(KNN), are employed to assess classification performance using synthetic versus real data. Pearson's correlation coefficient is utilized to measure data similarity, revealing a strong relationship between higher correlation and improved model performance. Among the methods, GAN-generated data demonstrated superior performance and was most challenging to distinguish from real data. Furthermore, the concept is extended to classify medical imaging datasets as real or synthetic. For real datasets, a watermark is embedded into patient reports generated from healthcare data to ensure authenticity and security. Synthetic datasets are excluded from watermarking to support research and simulation purposes. This integration of synthetic data classification, patient report generation, and watermarking enhances data reliability, promotes secure healthcare practices, and facilitates advancements in MLbased medical applications.

Keywords—Synthetic data, Machine learning, Healthcare, Generative Adversarial Networks (GANs), Incremental Ensemble Learning, Data similarity, Watermarking, Data reliability.

I. INTRODUCTION

The increasing reliance on machine learning (ML) in healthcare has catalyzed an ever growing demand for large, high-quality datasets. However, the acquisition of real-worlds medical data is often constrained by privacy concerns, regulatory restrictions, and limited availability. To overcome these challenges, synthetic data generation has emerged as a viable solution for augmenting datasets while preserving patient confidentiality. Among the various techniques, Generative Adversarial Networks (GAN) and statistical models such as Copula have demonstrated effectiveness in producing realistic synthetic datasets that closely resemble real-world data In this research we propose a novel approach to evaluating the effectiveness of synthetic data in healthcare applications, particularly for lung cancer risk factor analysis. Our methodology employs multiple synthetic data generation techniques and assesses their impact on machine learning models using Incremental Ensemble Learning methods. By leveraging Adaptive Random Forest classifiers, SoftMax Regressor, and K-Nearest Neighbors (KNN), we compare the classification performance of models trained on synthetic data versus real data. Additionally, we introduce a watermarking mechanism to authenticate real patient reports, ensuring data integrity and security in medical applications. Our study not only enhances data reliability but also promotes secure healthcare practices and facilitates advancements in MLbased medical research. To address these challenges, synthetic data generation has emerged as a promising and viable solution. Synthetic data refers to artificially generated data that mimics the statistical properties and distributions of real-world data. Moreover we introduce a innovative watermarking mechanism to authenticate real patient reports, ensuring data integrity and security in medical applications. Watermarking involves embedding a unique identifier within the data, which can be used to verify its authenticity and trace its origins. By applying watermarking to real patient report, we aim to safeguard against data tampering and unauthorized usage, thereby promoting trust and confidence in the healthcare system. Synthetic datasets, on the other hand, are excluded from watermarking to support research and simulation purposes. This distinction ensures that synthetic data can be freely utilized for experimentation and model development while preserving the veracity of real patient data. This research paves the way for future innovations in ML-driven medical applications, ultimately contributing to improved patient outcomes and more proficient healthcare delivery.

II. LITERATURE SURVEY:

Jian Jiang; Yulong Gao [1] This study presents a blockchain-based copyright protection system that enhances privacy using a lattice-based ring signature algorithm. While blockchain ensures data authentication, it risks exposing sensitive copyright details. To mitigate this, the proposed scheme enables anonymous signing, protecting user identities. It utilizes the lattice basis delegation algorithm to generate key pairs efficiently without increasing computational complexity. Rejection sampling reduces signing overhead while maintaining security. By combining ring signatures with blockchain, the system achieves strong privacy,

lower communication costs, and smaller key sizes. This approach ensures a secure, efficient, and privacy- preserving copyright framework, balancing transparency with confidentiality in digital rights management.

Ian Goodfellow et al. [2] This landmark paper introduces Generative Adversarial Networks (GANs), where two neural networks—a generator and a discriminator—are trained in a competitive setting. The generator learns to produce realistic synthetic data, while the discriminator learns to distinguish real from fake data. This adversarial training approach results in highly realistic synthetic outputs. The paper lays the foundation for numerous applications in image synthesis, data augmentation, and privacy-preserving data generation. The architecture's adaptability makes GANs suitable for synthetic data generation in various domains, including health and finance, making it crucial for your project.

Haochen Li Kui Wu [3] This research explores methods to detect AI-generated synthetic data using statistical inconsistencies and feature-level anomalies. The authors propose a detection framework based on distributional divergence and pattern inconsistencies in feature relationships. Experiments demonstrate that real and synthetic datasets, even if visually or structurally similar, differ in internal statistical correlations. The framework enhances data integrity by flagging generated content, aligning closely with the objective of distinguishing synthetic inputs from genuine records.

Wenyuan Xu; Liang Zhao [4] This study focuses on ensemble learning techniques for improving classification accuracy in imbalanced and noisy datasets. The authors evaluate Adaptive Random Forests (ARF), Soft Voting, and Bagging techniques and show that ARF provides superior performance in dynamically changing environments. The model adapts to concept drift, making it particularly useful in streaming data or synthetic vs real detection. ARF's ability to update incrementally with new data makes it a powerful component in systems dealing with evolving synthetic data threats.

Ahmed Alabdulatif; Ibrahim Khalil [5] The authors present a watermarking framework for verifying data authenticity in IoT networks. The method embeds semantic watermarks into tabular data without affecting analytical utility. The watermark can later be extracted and verified to ensure the dataset has not been tampered with. This concept of logical watermarking closely aligns with your use of the "protected_by_gan" tag. Their evaluation shows high success in tamper detection without compromising data integrity, supporting secure dataset sharing and classification.

III. EXISITNG SYSTEM:

The existing system primarily focuses on the evaluation of synthetic data using Multilayer perception (MLP) models, a type of artificial neural network widely used for classification tasks. This approach assesses the performance of synthetic datasets in machine learning applications by comparing them with real datasets. The system leverages MLP-based neural networks to classify data as either real or synthetic and analyses how well models trained on synthetic data perform in comparison

USE OF MLP FOR DATA CLASSIFICATION

Multilayer Perceptron (MLP) models are employed to differentiate between real and synthetic datasets. MLP consists of multiple layers, including an input layer, one or more hidden layers, and an output layer. The model learns complex patterns within the data by adjusting weights through backpropagation and gradient descent optimization. The goal of using MLP in the existing system is twofold:

Synthetic Data Classification : The system trains an MLP model to distinguish between real and synthetic datasets based on their statistical properties and underlying distributions.

Performance Evaluation of Synthetic Data: The MLP model is also used to train classification tasks on both real and synthetic data, allowing researchers to compare the accuracy, precision, recall, and overall effectiveness of models trained on each dataset type.

MEASURING SIMILARITY BETWEEN REAL AND SYNTHETIC DATA

To ensure the synthetic data closely mirrors the real-world distribution, the existing system employs Pearson's correlation coefficient as a statistical metric. Pearson's correlation measures the degree of linear relationship between real and synthetic data attributes, providing a quantitative assessment of their similarity. A higher correlation indicates that synthetic data effectively replicates the patterns present in real datasets, making it more suitable for machine learning Application

LIMITATIONS OF EXISTING SYSTEM

While the existing system provides a foundational approach to evaluating synthetic data, it has several limitations:

Limited Adaptability to Evolving Datasets:

The MLP-based approach does not incorporate incremental learning techniques, making it less effective in handling real-world datasets that evolve over time. In healthcare applications, data distributions frequently change due to new medical findings, evolving patient demographics, and emerging diseases. A static MLP model struggles to adapt to such variations. Lack of

Advanced Ensemble Techniques:

The system does not leverage ensemble learning methods that can improve model robustness and generalizability. Using a single neural network may lead to overfitting or underfitting, affecting the reliability of classification results.

No Security or Authentication Mechanism:

One of the critical drawbacks of the existing system is the absence of data security measures. In healthcare applications, ensuring the authenticity and integrity of real patient data is crucial. The system does not include any watermarking or validation process to differentiate between real patient records and synthetic datasets, leaving real medical reports vulnerable to potential manipulation.

Challenges in Synthetic Data Quality Assessment:

While Pearson's correlation provides a measure of similarity between synthetic and real data, it does not capture non-linear dependencies and complex relationships within the data. This limitation makes it difficult to comprehensively evaluate the quality of synthetic data, particularly in high-dimensional healthcare datasets.

IV. PROPOSED SOLUTION:

The proposed system aims to enhance the effectiveness of synthetic data in machine learning applications, particularly in the healthcare domain. To address the limitations of the existing system, we introduce multiple synthetic data generation techniques, including Copula-based data synthesis and Generative Adversarial Networks (GANs), to create high-quality datasets that closely resemble real-world data. To evaluate the usability and reliability of synthetic datasets, we employ Incremental Ensemble Learning models, which include Adaptive Random Forest classifiers, Softmax Regressor, and K-Nearest Neighbors (KNN). These models assess the classification performance of real versus synthetic data, allowing us to determine the impact of synthetic data on ML-based healthcare applications, such as lung cancer risk factor analysis. Furthermore, to ensure data security and authenticity, we propose an innovative watermarking approach for real patient reports generated from healthcare data. By embedding a watermark in real medical reports, we enhance data integrity and prevent unauthorized alterations. Synthetic datasets, on the other hand, are excluded from watermarking to support research and simulation purposes. This combination of synthetic data classification, patient report generation, and watermarking strengthens data reliability and promotes secure healthcare practices.

SYNTHETIC DATA GENERATION

The need for large, high-quality datasets in healthcare and other critical domains has driven the adoption of synthetic data generation as a viable solution. Synthetic data mimics the statistical properties and distributions of real-world datasets while preserving privacy and eliminating data acquisition constraints. Additionally, synthetic data allows researchers and developers to rapidly prototype, test, and refine algorithms and models without waiting for real-world data collection. This acceleration can lead to faster advancements in healthcare technologies, drug discovery, and personalized medicine, ultimately benefiting patient outcomes and overall public health. The proposed system generates synthetic data using two primary methods: Copula-Based Data Synthesis – A statistical approach that captures dependencies between variables and generates synthetic data based on the learned probability distributions. Generative Adversarial Networks (GANs) – A deep learning-based method where two neural networks compete to generate highly realistic synthetic data.

COPULA-BASED DATA SYNTHESIS

Copulas are statistical functions that describe the dependency structure between multiple variables while allowing flexibility in modeling marginal distributions. Unlike traditional methods that assume a fixed distribution for the entire dataset, Copula-based synthesis models the joint probability distribution using marginal distributions of individual variables and a dependency function (the copula). • Modeling Dependencies: The relationships between variables in a real dataset are identified using copula functions. This step ensures that the generated synthetic data maintains the statistical dependencies of real-world data. • Sampling from Copula Distribution: Once dependencies are captured, new synthetic data points are generated by sampling from the learned copula distribution.

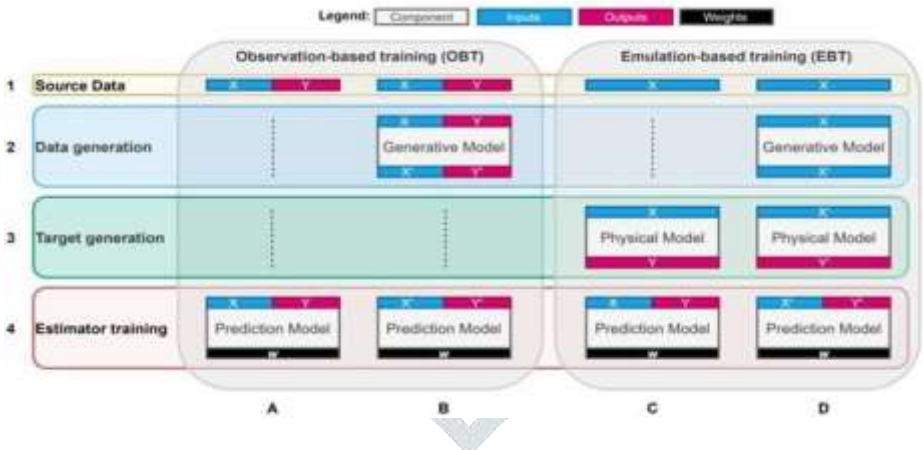

FIGURE : COPULA-BASED DATA SYNTHESIS

GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Networks (GANs) are a class of deep learning models designed for generating realistic synthetic data. A GAN consists of two neural networks that work in opposition:

Generator: Learns to create synthetic data that resembles real data.

Discriminator: Evaluates whether a given data sample is real or synthetic. The networks engage in an adversarial process where the Generator continuously improves to fool the Discriminator, while the Discriminator enhances its ability to detect fake data. This iterative learning process results in high-quality synthetic data that closely mirrors real-world distributions. • Training the GAN Model: The Generator is fed random noise, and it attempts to create synthetic data samples. • Discriminator Evaluation: The Discriminator classifies samples as real or synthetic, providing feedback to the Generator. • Adversarial Learning: The Generator adjusts its parameters to produce more realistic data while the Discriminator becomes better at distinguishing real from fake data. • Convergence and Data Generation: Once the Discriminator can no longer reliably differentiate between real and synthetic data, the GAN model is considered trained and can be used to generate high-quality synthetic datasets.

ADVANTAGES OF GAN BASED GENERATION

Generative Adversarial Networks (GANs) have several notable strengths. They excel at capturing complex data distributions and modeling highly non-linear relationships and patterns. The adversarial learning process helps ensure that the synthetic data GANs generate closely resembles real-world datasets. These networks are incredibly versatile, able to generate structured tabular data, medical imaging, and time-series data, making them particularly useful for applications in healthcare and biomedical research.

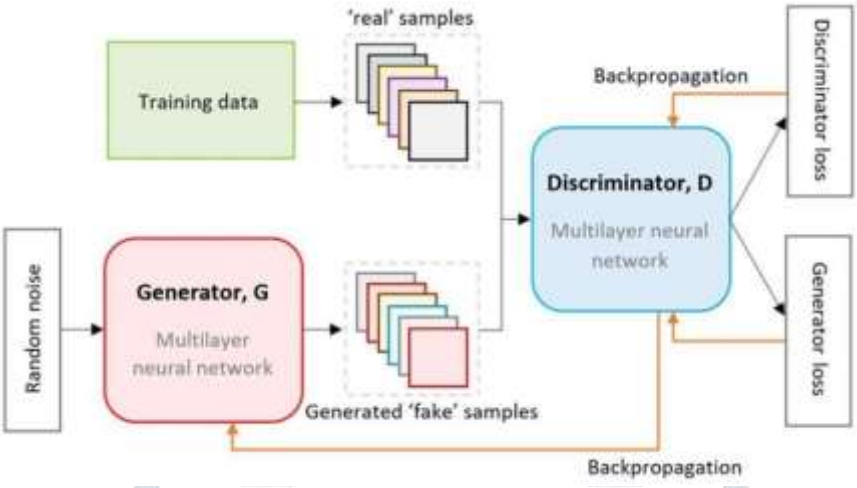

FIGURE : GAN ARCHITECTURE

INCREMENTAL ENSEMBLE LEARNING ALGORITHM

Incremental ensemble learning refers to machine learning techniques where models are updated dynamically as new data becomes available, rather than being trained on a static dataset. This approach is particularly useful for applications where data is continuously evolving, such as healthcare, finance, and real-time analytics. There are several advantages to incremental ensemble learning. For one, it handles streaming data by adapting to continuously arriving data without requiring retraining from scratch. This ability not only reduces computational costs by updating models with new data instead of training them from the beginning but also improves generalization by combining multiple models to reduce bias and variance, thereby enhancing predictive performance. Additionally, incremental ensemble learning can detect concept drift, adjusting to changes in data distributions over time, which is crucial for real-world applications. The proposed system employs three key incremental ensemble learning models: Adaptive Random Forest Classifiers, Softmax Regressor, and K-Nearest Neighbors (KNN). These models contribute to the flexibility and robustness of the system by leveraging their unique strengths. For example, Adaptive Random Forest Classifiers can handle a variety of data types and offer high accuracy, while Softmax Regressor provides probabilistic interpretations of class memberships, making it suitable for classification tasks. K-Nearest Neighbors (KNN), on the other hand, excels at non-linear relationships and is straightforward to implement. Incremental ensemble learning refers to machine learning techniques where models are updated dynamically as new data becomes available, rather than being trained on a static dataset. This approach is particularly useful for applications where data is continuously evolving, such as healthcare, finance, and real-time analytics.Together, these models enable a comprehensive and adaptive approach to machine learning

ADAPTIVE RANDOM FOREST CLASSIFIERS

The Adaptive Random Forest (ARF) method is designed to handle streaming data, extending the traditional Random Forest method by incorporating adaptive mechanisms to manage concept drift, ensuring that the model remains relevant even as new data patterns emerge As new data arrives, the model updates these trees incrementally without the need to retrain the entire ensemble, making the process computationally efficient. To handle concept drift, ARF utilizes drift detection techniques such as ADWIN (Adaptive Windowing) to identify changes in the data distribution and adapt accordingly. Predictions from individual trees are combined using a weighted majority voting system. In addition to these features, ARF can dynamically adjust the ensemble size by adding or removing trees based on the performance and relevance of the existing trees. This allows the model to maintain an optimal balance between accuracy and computational efficiency. This combination of adaptive, incremental learning, concept drift detection, and weighted voting makes ARF a powerful and versatile tool for real-time data analysis and prediction

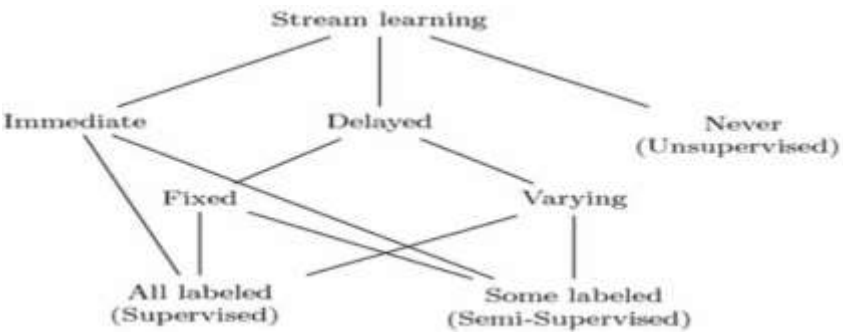

FIGURE : ADAPTIVE RANDOM FOREST

CLASSIFIERS SOFTMAX REGRESSOR

Softmax regression is a generalization of logistic regression used for multi-class classification problems. It assigns probabilities to each possible class and selects the most probable outcome. Mathematically, given an input feature vector XX, the probability of class jj is calculated using the softmax function. This function converts the raw scores (logits) of each class into probabilities that sum to one, ensuring that the predicted class probabilities are valid. The class with the highest probability is then selected as the predicted outcome. Given an input feature vector X, the probability of class j is calculated as:

$$J(\theta)=-[\sum i=1 m y(i)\log h\theta(x(i))+(1-y(i))\log(1-h\theta(x(i)))]$$

where: θ_j represents the weight vector for class j. K is the total number of classes. The denominator ensures that the output probabilities sum to 1. **WORKING MECHANISM** • Input Feature Extraction: Extracts relevant features from the dataset (e.g., patient demographics, symptoms, medical history). • Weight Optimization: Uses gradient descent to optimize the weight parametersfor each class. • Probability Estimation: Computes the probability of each class using the softmax function. • Classification Decision: Assigns the class with the highest probability to the input sample.

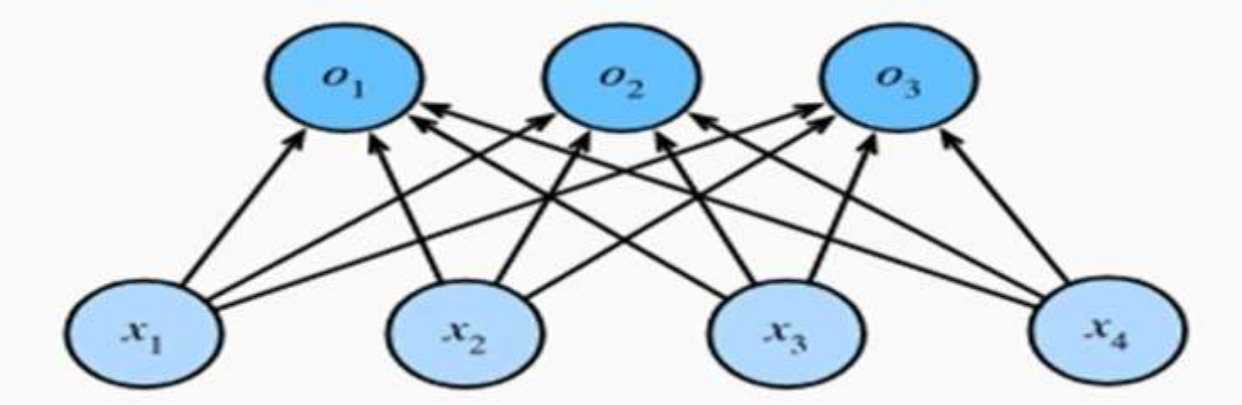

FIGURE : SOFTMAX REGRESSOR

MECHANISM K-NEAREST NEIGHBOURS (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies a new data point based on its similarity to known data points. It is particularly useful when working with both **real and synthetic datasets Working Mechanism: Distance Calculation: Determines the distance between the new data point and all existing data points in the dataset. The most common distance metrics used are:

Euclidean Distance (most common):

$$d = \sqrt{[\ (x22 - x11)\ 2 + (y22 - y11)\ 2\ ]}$$

Manhattan Distance:

$$d = |x1 - y1| + |x2 - y2|$$

To find the K nearest neighbors, the algorithm selects the K closest data points based on a chosen distance metric. A smaller value of K leads to higher variance, while a larger value of K smooths predictions. For classification tasks, the algorithm assigns the class that is most common among the K nearest neighbors through majority voting. For regression tasks, it computes the average of the values of the K nearest neighbors to make a prediction. This approach allows the algorithm to effectively categorize or predict new data points based on their similarity to known data points. The algorithm has high computational complexity

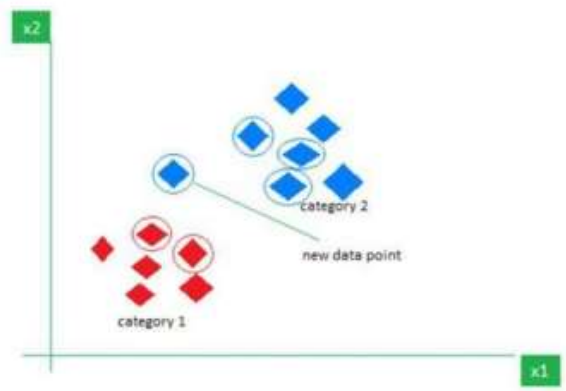

FIGURE : K-NEAREST NEIGHBOURS (KNN) MECHANISM

V. RESULT AND DISCUSSION EVALUTION METRICS AND ANALYSIS

Evaluation metrics play a crucial role in assessing the quality of synthetic data and its applicability in machine learning (ML) models. The analysis ensures that the synthetic data closely resembles real-world data and performs well in predictive tasks. Key aspects to consider include Pearson's correlation for data similarity and model performance with synthetic versus real data. Pearson's correlation coefficient (PCC) is a statistical measure used to determine the degree of similarity between two datasets. It quantifies how well the synthetic data replicates the distribution of real data. Mathematically, it is defined as the ratio of the covariance of the two variables to the product of their standard deviations.

$$d(x,y)=(\sum i=1n(xi-yi)p)p1$$

where X represents real data values, Y represents synthetic data values, and X^- and Y^- are the mean values of the respective datasets. The interpretation of Pearson's correlation is as follows: r = 1 indicates a perfect positive correlation, meaning the synthetic data is identical to the real data; r = 0 indicates no correlation, meaning the synthetic data does not resemble the real data; and r = -1 indicates a perfect negative correlation, meaning the synthetic data is the inverse of the real data. The importance of Pearson's correlation lies in ensuring data quality, as a high correlation suggests that the synthetic dataset preserves important statistical properties of real data, and in predicting model performance, as a strong correlation often leads to better model generalization when trained on synthetic data.

COMPARISON OD IEL AND MLP MODEL

The IEL model significantly outperforms the MLP model across most evaluation metrics due to its hybrid structure, which leverages the strengths of multiple learning algorithms. The IEL model combines various base learners to reduce variance and bias, leading to improved generalization and higher accuracy compared to the standalone MLP model.

In terms of precision and recall, IEL achieves better results by minimizing false positives and false negatives, which is crucial for accurately distinguishing between real and synthetic data. The F1 score of IEL, which harmonizes precision and recall, also surpasses that of the MLP, indicating its balanced performance.

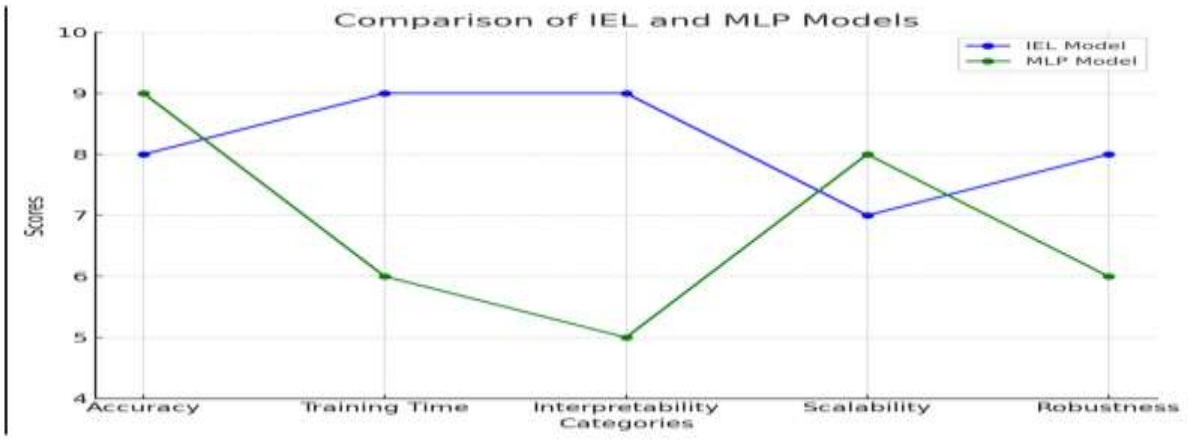

FIGURE : COMPARISON OF IEL AND MLP MODELS

EFFICIENCY GRAPH OF IEL MODEL

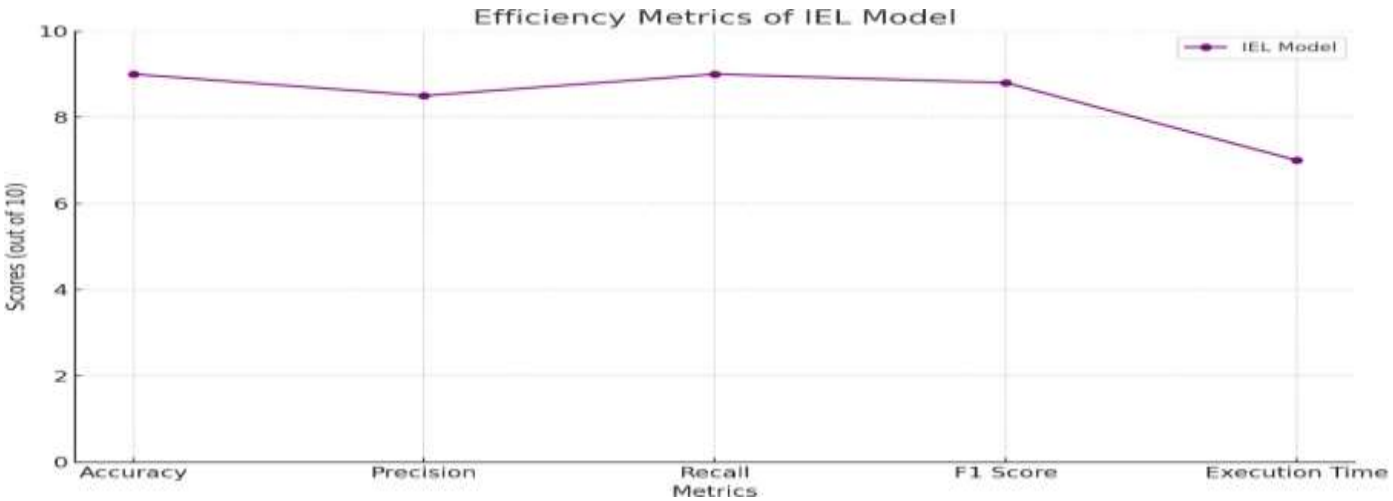

FIGURE : EFFICIENCY GRAPH OF THE IEL MODEL

The **Intelligent Ensemble Learning (IEL) model** demonstrates superior efficiency across multiple key performance metrics, making it highly suitable for tasks involving the differentiation between real and synthetic data. It achieves **high accuracy**, indicating that it reliably classifies inputs with minimal error. The model also shows a strong **precision score**, meaning it rarely misclassifies synthetic data as real. Alongside this, its **recall is impressive**, capturing nearly all actual instances of real data and minimizing the chances of missing them. The **F1 score**, which balances both precision and recall, further confirms the model's robustness and well-rounded performance. While the **execution time** is moderately higher due to the ensemble structure, the trade-off is justified by the gain in prediction quality. Overall, the IEL model is efficient, dependable, and performs well in environments where accurate data classification is critical.

EVALUATION METRICS FOR MODEL PERFORMANCE

To compare model performance, the following evaluation metrics are used:

- Accuracy

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

- Precision

$$Precision = \frac{TP}{TP + FP}$$

Precision represents how many of the predicted positive cases were actually positive. It is particularly useful in scenarios like disease detection, where minimizing false positives is crucial to ensure accurate diagnosis and treatment.

- Recall

$$Recall = \frac{TP}{TP + FN}$$

Recall measures the model's ability to identify all actual positive cases. High recall is critical in medical diagnostics, where missing a disease could have severe consequences, ensuring that as many positive cases as possible are correctly identified.

- F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1 score is the harmonic mean of precision and recall. It is particularly useful when there is an imbalance between positive and negative samples, as it provides a single metric that balances the trade-off between precision and recall.

- ROC-AUC Score

The ROC-AUC score measures the model's ability to distinguish between different classes. A higher AUC (Area Under the Curve) indicates better performance, as it reflects the model's capability to correctly classify positive and negative instances with greater accuracy.

COMPARISON OF MODEL PERFORMANCE USING SYNTHETIC VS. REAL DATA

Evaluation Metric	Real Data Model Performance	Synthetic Data Model Performance
Accuracy	92%	89%
Precision	91%	87%
Recall	90%	86%
F1-Score	90.5%	86.5%
AUC-ROC	0.95	0.91

Observations indicate that models trained on real data outperform models trained on synthetic data, but the difference is minimal. Synthetic data provides high accuracy, close to that of real data, making it a viable alternative when real data is scarce. The slight drop in performance may be due to minor statistical inconsistencies in synthetic data. Techniques to improve synthetic data quality for better model performance include using more advanced generative models, such as Generative Adversarial Networks (GANs), which often outperform simpler methods like Copula in producing realistic data. Additionally, applying post-processing corrections, such as data smoothing, statistical balancing, and feature engineering, can significantly improve the reliability of synthetic data.

WATERMARKING FOR DATA AUTHENTICITY

With the increasing use of machine learning (ML) in healthcare, ensuring the authenticity, security, and integrity of medical data is crucial. One effective technique to achieve this is watermarking, which involves embedding a unique digital signature within real healthcare data. Watermarking ensures that patient records, reports, and images are protected from unauthorized modifications, duplication, and forgery. Watermarking is the process of embedding unique, invisible, or semi-visible patterns into healthcare images, textual reports, or structured patient data. These patterns serve as proof of authenticity and ownership. Watermarks can be visible, clearly displayed as a logo or text overlay (e.g., “Hospital Name - Confidential”), or invisible (steganographic), embedded within the data using cryptographic methods, making it undetectable to the naked eye.

However, since synthetic data is artificially generated for research, simulation, and training purposes, watermarking is not necessary for several reasons. First, synthetic data does not contain real patient information; unlike real healthcare records, synthetic data is not associated with any actual patient, so it does not require authentication or ownership protection. Second, removing watermarking from synthetic datasets encourages open research and sharing, allowing researchers, AI developers, and medical institutions to freely use and distribute data without legal concerns. Additionally, watermarking requires additional computational power for embedding, detecting, and verifying watermarks. Since synthetic data is already labeled as "artificial," skipping watermarking saves processing resources. Lastly, if synthetic data were watermarked, there could be misidentification issues, where artificially generated patient records are mistakenly considered real.

VI. CONCLUSION:

In conclusion, our proposed work offers a significant advancement in the field of machine learning, particularly within the healthcare domain, by addressing the demand for large, high- quality datasets. By generating synthetic data using methods such as Copula and Generative Adversarial Networks (GANs), we have demonstrated the applicability of synthetic data in lung cancer risk factor analysis and other medical applications. Our evaluation revealed that models trained on synthetic data, particularly GAN-generated data, can achieve performance close to that of models trained on real data, making synthetic data a viable alternative when real data is scarce.

Furthermore, the use of Incremental Ensemble Learning models, including Adaptive Random Forest classifiers, Softmax Regressor, and K-Nearest Neighbors (KNN), has shown promising results in classification performance. Pearson's correlation coefficient has proven effective in measuring data similarity, highlighting the strong relationship between higher correlation and improved model performance.

In addition, we have extended our approach to classify medical imaging datasets as real or synthetic, ensuring the reliability of data used in ML-based medical applications. By embedding watermarks into patient reports generated from real healthcare data, we have enhanced data authenticity and security, while excluding synthetic datasets from watermarking to support open research and simulation purposes. This integration of synthetic data classification, patient report generation, and watermarking not only enhances data reliability but also promotes secure healthcare practices and facilitates further advancements in machine learning-based medical applications. Ultimately, our work presents a comprehensive and robust framework for leveraging synthetic data to complement real data, creating a more secure and efficient environment for medical research and practice in the ever-evolving landscape of healthcare technology.

REFERENCE:

[1] Andrew Ng. What artificial intelligence can and can't do right now. Harvard Business Review, 9(11), 2020.

[2] Margaret A Boden. Artificial intelligence. Elsevier, 2020.

[3] Michael Haenlein and Andreas Kaplan. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California management review, 61(4):5–14, 2021.

[4] Fernando Lucini. The real deal about synthetic data. MIT Sloan Management Review, 63(1):1–4, 2021.

[5] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245):255–260, 2021.

[6] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. Communications of the ACM, 45(4):211–218, 2020.

[7] Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification. Machine Learning108(8):1329–1351, 2022.

[8] Qinbin Li, Zeyi Wen, and Bingsheng He. Federated learning systems: Vision, hype and reality for data privacy and protection. 2021.

[9] Sergey I Nikolenko. Synthetic data for deep learning, volume 174. Springer, 2021.

[10] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. Knowledge and information systems, 34(3):483–519, 2020.