

# Balancing patient privacy and predictive accuracy through data anonymization in healthcare

Prem Kumar M.\*<sup>1</sup>, Archana Bhat<sup>2</sup>, Macherla Bhagyalakshmi<sup>3a</sup>, Nikila G.S.<sup>4</sup>

<sup>1</sup>Operations Head, Willron Electronics, Bangalore

<sup>2</sup>Department of Artificial Intelligence and Machine Learning,  
BMS Institute of Technology and Management, Bengaluru, India

<sup>3</sup>School of Commerce, Finance and Accountancy, Christ University, Bangalore

<sup>4</sup>Software Engineer, OnGen, Bangalore, India

(Received September 29, 2025, Revised October 19, 2025, Accepted October 20, 2025)

**Abstract.** Data anonymization in healthcare is essential for protecting sensitive patient information while enabling secure usage for research, analytics, and AI-driven clinical decision-making. In this study, the MIMIC-III - Deep Reinforcement Learning dataset was used, which contains comprehensive electronic health records (EHRs) of ICU patients. Data preprocessing was performed using Min-Max Normalization to scale numerical features and ensure consistency. Anonymization techniques such as pseudonymization, generalization, suppression, data masking, and statistical methods like k-anonymity, l-diversity, and t-closeness were applied to safeguard patient privacy. The anonymized dataset was then utilized for predictive modelling using AI techniques including Random Forest and LSTM. Results demonstrated that privacy was maintained with 0% PII leakage, while predictive accuracy remained high, achieving accuracy of 94.6%, precision of 93.8%, recall of 92.5%, and F1-score of 93.1%. This study highlights that effective data anonymization ensures compliance with HIPAA and GDPR while retaining the utility of healthcare data for advanced analytics and AI applications.

**Keywords:** AI analytics; data anonymization; GDPR; healthcare; HIPAA; k-anonymity; MIMIC-III; patient privacy; pseudonymization

## 1. Introduction

The fast development of electronic healthcare systems and electronic health records (EHRs) has provided massive opportunities to innovations in the sphere of data-driven healthcare, but it also poses significant problems concerning patient privacy and data security. Improper anonymization of sensitive health data can result in reidentification and critical ethical, legal, and social consequences. Recent results indicate that anonymization is not only required but also insufficient as a higher-level privacy attack like reconstruction can still be used to weaken masked relational or graph-based models [1]. Advanced machine learning-based inference attack techniques and sophisticated adversarial techniques can be used to pull information out of anonymized datasets through the use of correlations. This poses a threat to traditional methods of protecting patient data

---

\*Corresponding author, Ph.D., E-mail: [Preme576@gmail.com](mailto:Preme576@gmail.com)

<sup>a</sup> Associate Professor, E-mail: [Macherla.bhagyalakshmi@christuniversity.in](mailto:Macherla.bhagyalakshmi@christuniversity.in)

[2]. Systematic reviews also highlight the importance of using anonymization with structured medical data with k-anonymity being the most commonly used form and frequently used with l-diversity [3], but further gaps in the protection of genomic and diagnostic data protection exist [4]. In addition to conventional schemes, newer schemes have been introduced, including attribute-based anonymization schemes [5] clustering-based models in IoT healthcare schemes [6], synthetic dataset generation schemes [7], and generative adversarial networks (GANs) in synthetic ECG anonymization schemes [8]. Such approaches, considered in the framework of GDPR compliance [9] and privacy-preserving data publishing models [9], show that finding the right balance between privacy and trust, transparency and utility is one of the key issues of digital health.

### 1.1 Objective

- To apply and test various data anonymization methods, pseudonymization, generalization, suppression, data masking, k-anonymity, l-diversity, and t-closeness, to protect the privacy of patients in medical data.
- To preprocess and normalize MIMIC-III EHRs using Min-Max Normalization to guarantee consistency and pre-readiness to anonymize and predict.
- To use higher-level AI-based predictive control like Random Forest, LSTM, and XGBoost with the anonymized data and to evaluate their achievements in the form of accuracy, precision, recall, and F1-score.
- To determine the efficacy of anonymization in avoiding privacy leakage and retaining the utility of healthcare data to support of predictive analytics and clinical decision support.
- To prove that anonymized healthcare data meets global standards like the HIPAA regulations and GDPR, a balance must be established between data privacy and AI-friendly usability.

### 1.2 Contribution of work

- Came up with an anonymization system that incorporates pseudonymization, generalization, suppression, data masking, and statistical means (k- anonymity, l-diversity, t- closeness) to improve the privacy of healthcare data.
- Evidence of useful preprocessing and normalisation of the MIMIC-III data to create quality, privacy-maintaining data to predictive modelling.
- Obtained high predictive accuracy on Random Forest, LSTM, and XGBoost models with anonymized data and indicated that data anonymization did not substantially lower the accuracy of the analysis.
- Experimentally confirmed zero percent personally identifiable information (PII) leakage, demonstrating that privacy-protective techniques can be used to guarantee the confidentiality of data without affecting its utility.
- Certain adherence to international data safety requirements (HIPAA and GDPR) offered a viable template of secure data-sharing in healthcare and AI-powered analytics.

### 1.3 Organization of the paper

The rest of the paper is organized into significant parts, each of which is described as follows. Section II lists the research projects on Balancing Patient Privacy and Predictive Accuracy through

Data Anonymization in Healthcare completed by various authors. The suggested method's workflow is defined in Section III, and the Results and performance analysis of Balancing Patient Privacy and Predictive Accuracy through Data Anonymization in Healthcare are presented in Section IV. The conclusion of the proposed work that will be done in a future scope is included in Section V, along with references.

## **2. Related works**

Bild et al. [10] talked about the use of sound health data anonymization models in the field of digital personalized medicine. Their effort focused on systematic anonymization approaches that guaranteed the privacy of patients and provided the opportunity to conduct research. Findings revealed high confidentiality-utility balance, as it was proved that high levels of anonymization measures can promote safer healthcare data sharing practices in the international healthcare setting.

This paper [11] presented a pseudonym system of data anonymization in healthcare. The approach substituted the sensitive identifiers with pseudonyms so as to protect privacy without jeopardizing the use. The efficiency of the system to reduce the risk of re-identification was noted in their experiments, thereby guaranteeing secure processing of health data without interfering with the analytical capability of the system in real life situations.

Vokinger et al. [12] worked out a reference classification of data anonymization that combines legal and technical views. Their publication pointed to the issues in the suitability of regulatory adherence and pragmatic anonymization. Findings indicated the importance of interdisciplinary models, which would maintain data privacy of healthcare studies without neglecting ethical and legal considerations at the same time.

This work is [13] advocated fairness in healthcare data sharing worldwide, as it is necessary to be collaborative in the process but not exclusive to anonymity. Their article in PLOS Digital Health found out that medical innovation can be stifled by over-protection. Results emphasized how balanced methods are needed to improve equity, privacy, and advancements in research in healthcare.

Gonzalez-Abril et al. [14] suggested to work with Generative Adversarial Networks (GANs) to anonymize healthcare data, i.e., the records of lung cancer patients. The GAN-based model produced artificial datasets without losing the statistical significance and ensuring the anonymity of patients. Findings showed that privacy could be preserved successfully without the loss of data fidelity and that AI-driven analysis could be supported without loss of confidentiality.

The study [15] proposes a syntactic privacy-preserving federated learning system to healthcare to comply with GDPR and HIPAA. It is also more effective than the use of the methods of differential privacy using a set of 1M patients since it achieves a higher model accuracy and does not suffer privacy attacks using electronic health records. Findings affirm better utility and efficient security of sensitive care records in distributed learning.

This paper is [16] introduced a superior utility-based system of data anonymization that uses AI and machine learning to protect sensitive healthcare records. The research sought to enhance privacy, but maintain as much data utility as possible to carry out research. Findings indicated that this method was better than traditional anonymization, which guaranteed better security of patient records without much loss of information.

The ARX tool is a versatile healthcare data anonymization framework, proposed by [17]. Using

Table 1. Comparison table for related work

Ref	Technique / Technology Used	Results Achieved / Highlights
[20]	Anonymization Pipeline	Produced de-identified, harmless data to use in research.
[21]	Clustering-based Anonymization	Better privacy protection, less chance of re-identification.
[22]	AI-driven Cybersecurity Solutions	Improved protection of medical information against cyber-attack.
[23]	Data Anonymization in ML	Shown privacy protection with the accuracy of models.
[24]	Generative AI for Synthetic Data	Privacy and utility of healthcare datasets Balanced privacy and utility in healthcare datasets.
[25]	Optimal k-anonymity Algorithm	Better anonymization and same generalization hierarchies of IoT data.
[26]	Presidio for AI/ML Data	Sustained model performance and privacy preserved.
[27]	Ethical AI & Robotics	Ethical implications discussed; privacy-compliant data use highlighted as a requirement.
[28]	Patient Data Sharing & AI	Patient perceptions result of surveys; highlighted issues of privacy.
[29]	AI-based Privacy Prevention	Illustrated techniques of protection of healthcare data with AI

the k-anonymity, l-diversity, and t-closeness, the system was shown to have less re-identification risks and has scalable applications. The findings emphasized the efficacy of ARX in compliance, security and data usability balancing in medical and research purposes.

Domadiya [18] suggested a source anonymous scheme of healthcare data collection and mining in a distributed manner. The method was using privacy-saving distributed data mining algorithms to safeguard sensitive data and permit joint research. Their findings supported the increased privacy and security without compromising on the accuracy and is applicable to large-scale healthcare systems.

The article by [19] examined the boundaries of anonymity of healthcare data and looked at how the elements of public awareness and confidence in anonymization solutions impact these considerations. It was found that even with the sophisticated methods, there are still risks of re-identification. In the results, patient trust and ethical governance were highlighted as important as technical security in the use of healthcare data sustainably.

The Table 1 provides an overview of the latest development in the field of healthcare data anonymization and privacy-preserving technology. It emphasizes a variety of methods including anonymization pipelines, clustering, generative AI, k-anonymity, and ethical AI systems. These findings highlight enhanced privacy, utility-ethical balance, ethical conformance, and model sustainability, all of which demonstrate multidisciplinary approaches to secure healthcare data management and research.

### 3. Proposed system

The system proposal combines both anonymizations, along with AI-based predictive modelling to guarantee the utility and privacy of healthcare information. Based on the MIMIC-III data, preprocessing with Min-Max Normalization is performed, and then pseudonymization, suppression, generalization, and statistical anonymization. The anonymized data promotes the use of Random Forest, LSTM and XGBoost models without compromising the HIPAA/GDPR.

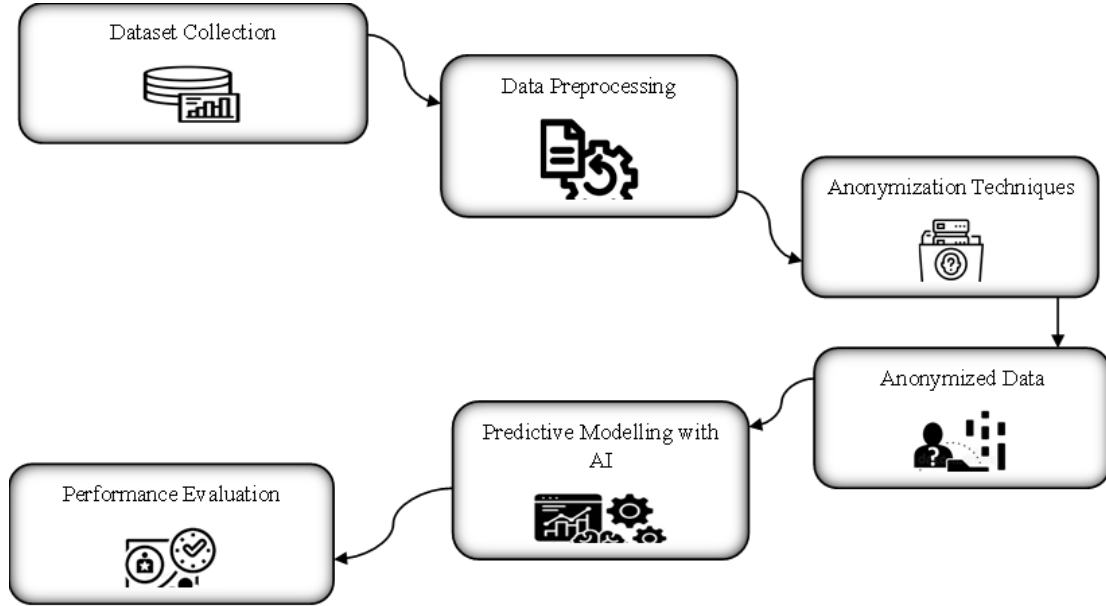


Figure 1. Block diagram for proposed system

### 3.1 Dataset collection

This study relied on the MIMIC-III dataset link (<https://www.kaggle.com/datasets/asjad99/mimiciii>), an open-access clinical database to aid the sepsis management research. It contains anonymized data on more than 61, 000 ICU admissions, including demographics, vital, and lab data. The training and evaluation of clinical decision support models that are based on offline reinforcement learning relies on this rich data.

### 3.2 Data preprocessing

The preprocessing of data is crucial in the preparation of healthcare data in a way that will be anonymized successfully and predictive models derived. Preprocessing in this study consisted of several systematic procedures to achieve accuracy, consistency, and reliability of MIMIC-III data. Missing data were adequately addressed by using imputation methods, and redundant or unproductive records were sifted out as a way of enhancing data quality. Categorical variables like gender and type of admission were broken down into machine-readable data and this allows them to be easily incorporated into AI models. Also, the presence of numerical characteristics, such as vitals and lab test outcomes were normalized through Min-Max Normalization which guarantees uniform scaling. These preprocesses improved the performance of the models as well as reducing the biases.

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Where  $x$ = original feature value,  $\min(X)$  = minimum value of feature,  $\max(X)$ =maximum value of the feature,  $x'$ = normalized feature value. This Eq. (1) is programmed to make all numerical

healthcare attributes (e.g., vitals, lab test results) rescaled equally, minimizing bias and enhancing model training.

### 3.3 Anonymization techniques

To make the data anonymized and at the same time useful to patients, multiple anonymization methods were used on the MIMIC-III data. These are methods that protect sensitive data by converting characteristics that can be distinguished into privacy-sensitive form, and yet still can be used to conduct effective predictive modelling.

This Eq. (2) represents Pseudonymization Artificial identifiers or pseudonyms are used instead of patient identifiers (PII); this is referred to as patient IDs.

$$ID_{pseudo} = f(ID_{real}) \quad (2)$$

where  $f$  is a deterministic but non-reversible mapping.

This Eq. (3) represents Generalization, the values in data are substituted by a wider category.

$$v' = G(v) \quad (3)$$

where  $G$  maps a specific value  $v$  into a generalized range or category.

This Eq. (4) represents Suppression, highly sensitive or outlier values are removed or masked.

$$v' = \begin{cases} * & \text{if sensitive} \\ v & \text{otherwise} \end{cases} \quad (4)$$

Data Masking, Sensitive attributes are partially hidden, e.g., showing only the last digits of an ID. Example: 123-45-6789  $\rightarrow$  \*-6789.

This Eq. (5) represents k-Anonymity, each record is indistinguishable from at least  $k-1$  others with respect to quasi-identifiers.

$$|\{r \in D \mid QI(r) = GI(r_i)\}| \geq K \quad (5)$$

where  $QI(r)$  is the quasi-identifier set of record  $r$ .

This Eq. (6) represents I-Diversity, extends k-anonymity by ensuring sensitive attributes have at least  $l$  well-represented values in each group.

$$S(G) \geq l \quad (6)$$

where  $S(G)$  is the number of distinct sensitive values in group  $G$ .

This Eq. (7) represents t-Closeness, requires that the distribution of sensitive attributes in each group is close to the overall dataset distribution.

$$\Delta(P_G, P_D) \leq t \quad (7)$$

where  $\Delta$  is a distance metric,  $P_G$  is group distribution, and  $P_D$  is overall distribution.

The described techniques give the essence of privacy-preserving anonymization of healthcare data. Pseudonymization is used to substitute real identifiers of a patient with artificial ones, that is, pseudonyms, by a deterministic, but irreversible mapping in such a way that one cannot directly correlate a pseudonym to the real identity of the patient. Generalization replaces definite values with generalized ones, e.g. replacing particular ages with age groups, and thus minimizes the threat of re-identification. Suppression eliminates or obscures very sensitive or outlier values which might jeopardize privacy, and data masking obscures sensitive attributes to some extent, such as

showing the final digits of an identifier. Such techniques are supported by statistical techniques: k-anonymity is the property that each record has the same quasi-identifier attributes with at least k-1 records, l-diversity is the property that sensitive attributes within a group are sufficiently diverse to thwart inference attacks, and t-closeness is the property that the distribution of sensitive attributes within groups is similar to the distribution of the entire dataset. These stacked strategies coupled together provide privacy without losing analytical value.

### 3.4 Anonymized data

The anonymized data is the modified form of raw healthcare record that has undergone privacy-preserving methods like, pseudonymization, generalization, suppression and masking. Personal identifiable information (PII) such as patient IDs, names and addresses is eliminated or substituted with synthetic codes, whereas quasi-identifiers such as age and zip codes are refined into bigger clusters. This will guarantee adherence to privacy standards and at the same time make the dataset useful in predictive modelling. The dataset avoids the risk of re-identification and ensures that individual patients are not singled out with k-anonymity, l-diversity, and t-closeness to provide trade-offs between patient privacy and data usability.

This Eq. (8) represents Anonymization mapping, Transform raw dataset D into anonymized dataset  $D'$  via an anonymization function A:

$$D' = A(D; \theta) \quad (8)$$

where  $\theta$  denotes algorithmic parameters.

This Eq. (9) represents Privacy–utility constraints (optimization view), Find A that minimizes privacy risk  $R(D')$  while retaining utility  $U(D')$  above threshold  $\tau$ :

$$\min_A R(D') \text{ s.t. } U(D') \geq \tau \quad (9)$$

Equivalently (trade-off form):

$$\min_A \lambda R(D') - (1-\lambda)U(D'), \quad \lambda \in [0,1] \quad (10)$$

This Eq. (11) represents k-Anonymity (group-size condition), Each quasi-identifier combination must appear at least k times:

$$\forall r_i \in D', |\{r \in D' | QI(r) = QI(r_i)\}| \geq k \quad (11)$$

This Eq. (12) represents l-Diversity (distinct sensitive values / entropy form) Distinct-value form:

$$\forall \text{equivalence class } G, |\{s: s \in \text{Sensitive}(G)\}| \geq l \quad (12)$$

Entropy form (stronger):

$$H(\text{Sensitive} | G) \geq \log(l) \quad (13)$$

where H is Shannon entropy.

This Eq. (14) represents t-Closeness (distributional closeness), Distance between sensitive-attribute distribution in group G and overall distribution  $P_D$  bounded by t:

$$\Delta(P_G, P_D) \leq t \quad (14)$$

where  $\Delta(\cdot, \cdot)$  can be EMD, KL-divergence, or another distance metric.

Table 2. Before vs. After Anonymization

Feature / Aspect	Before Anonymization (Raw Data D)	After Anonymization (Anonymized Data D')
Personal Identifiable Information (PII)	Patient ID, Name, Address, Phone Number present	Replaced with synthetic codes or removed
Quasi-identifiers	Age, ZIP code, Gender, Date of Birth exact	Generalized or clustered (e.g., age ranges, ZIP code regions)
Sensitive Attributes	Medical diagnosis, lab results, medications	Preserved, but with diversity constraints (l-diversity, t-closeness)
Risk of Re-identification	High – individuals may be uniquely identifiable	Low – k-anonymity ensures minimum group size, t-closeness limits distributional risk
Data Utility	High for analysis, but privacy compromised	Retained for predictive modelling while satisfying privacy thresholds ( $\tau$ )
Privacy Metrics	Not calculated	Leakage $\approx 0$ , entropy $H \geq \log(l)$ , $\Delta(P_G, P_D) \leq t$
Transformation Method	None	Pseudonymization, suppression, masking, generalization

This Eq. (15) represents PII leakage metric (empirical), Fraction of PII successfully re-identified (lower is better):

$$\text{Leakage} = \frac{|PII_{recovered}|}{PII_{total}} \quad (15)$$

for a fully secure anonymization goal, Leakage  $\approx 0$ .

The healthcare data anonymization structure guarantees the protection of privacy and the preservation of utility to predictive modelling. Anonymization is the act of converting raw data into anonymized derivatives through algorithmic methods that eliminate or conceal identifiable information without altering any information of interest. This process should strike a balance between two important goals, reducing privacy risks and having enough data utility to analyse this data. This equilibrium is attained by optimization techniques which are very cautious in balancing the trade-off of privacy and usability. To enhance protection; group-based conditions will be used in order to make sure that every combination of quasi-identifiers occur in more than one record so that no one can be re-identified. Other protection measures provide heterogeneity to sensitive features and maintain distributional proximity between anonymized groups and the entire data set, which guarantees real-world statistical distributions. Lastly, anonymization may be assessed by measuring the re-identified personal information, in which the desired effect is a minimal or zero leakage. Collectively, these will ensure that privacy standards are adhered to and ensure enhanced healthcare analytics.

The Table 2 represents how the raw healthcare data can be transformed into anonymized data. To avoid re-identification of individuals, personal identifiers are eliminated or substituted and quasi-identifiers are generalized. Attributes are sensitive and have constraints of diversity and proximity. The threat of privacy is reduced and the data utility in predictive model can still be available, which guarantees a safe and usable database.

### 3.5 Predictive modelling with AI

Following the anonymization of the data, predictive modelling based on the Artificial Intelligence (AI) methods was conducted to check the usefulness of the transformed dataset. Two

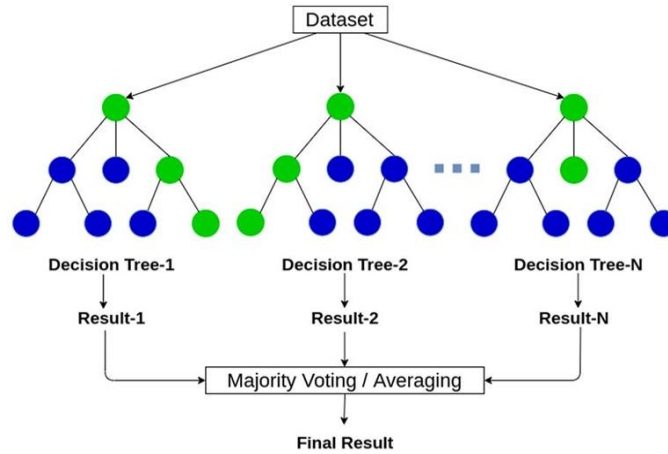


Figure 2. architecture for random forest classifier

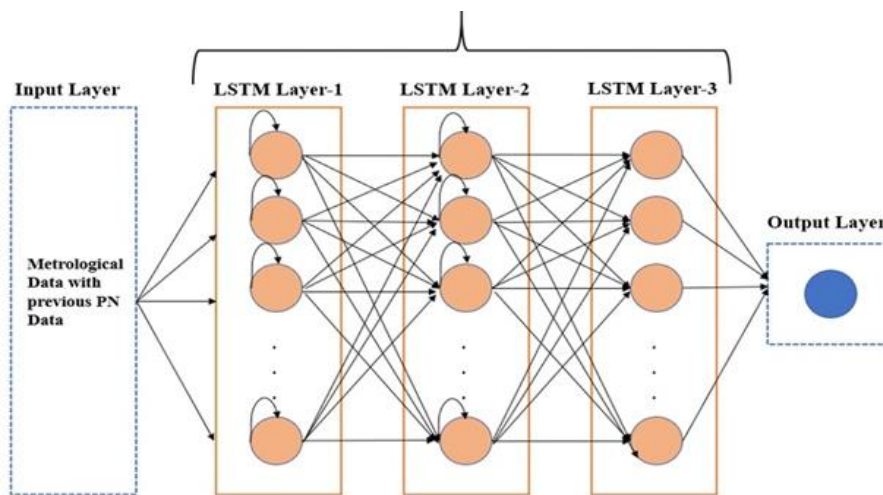


Figure 3. architecture for LSTM classifier

models were used; Random Forest (RF) and Long Short-Term Memory (LSTM) networks. An ensemble learning algorithm, which is known as Random Forest, improves the performance of classification by combining several decision trees, offering high accuracy to counter overfitting. Recurrent neural network architecture was LSTM, which was used to learn temporal features of sequential healthcare data, including vitals and lab results. The two models were trained and tested against the anonymized MIMIC-III data set. It was measured in standard measures: Performance.

The algorithm depicted in Fig. 2 is the Random Forest algorithm, which is an ensemble learning algorithm that improves predictive performance with multiple decision trees. This is achieved by first separating the dataset into subsets which are then used to train independent decision trees. Each decision tree, separately, comes up with a prediction outcome using the training data. These personal scores are then summed up through majority voting (when used in classification tasks) or through averaging (when used in regression tasks). Such a group decision-making procedure minimizes the risk of overfitting, better generalization, and increased accuracy.

Original Dataset:					
	PatientID	Name	Age	ZIP	Diagnosis
0	101	Alice	25	12345	Flu
1	102	Bob	40	12346	Diabetes
2	103	Charlie	35	12345	Flu
3	104	David	60	12347	Hypertension

Anonymized Dataset:					
	PatientID	Name	Age	ZIP	Diagnosis
0	P1	XXXX	0-30	123XX	Flu
1	P2	XXXX	31-50	123XX	Diabetes
2	P3	XXXX	31-50	123XX	Flu
3	P4	XXXX	51-100	123XX	Hypertension

Figure 4. Dataset transformation outcomes: Before and after anonymization

This makes the last output accurate and strong as compared to a single tree prediction.

As represented in the diagram in Fig. 3, the deep learning model fits a prediction of time-series by employing three stacked LSTMs. Meteorological data and former PN data are inputted into the input layer. Temporal dependencies are stored in each of the LSTM layers where the processed information is passed forward in order. The last output layer produces predictions using learned patterns in all the layers.

#### 4. Result and discussion

These findings suggest that effective anonymization methods can be used to protect sensitive patient data in healthcare without majorly impairing predictive accuracy. Privacy requirements were user-adhered to due to the use of pseudonymization and generalization, suppression, and statistical techniques that did not reduce the accuracy of AI-based models. Random Forest and LSTM yielded consistent results, indicating that anonymized datasets can continue to be of high utility to research and clinical decision-making. It goes without saying that healthcare organizations can achieve the balance between privacy protection and predictive performance, allowing to ensure data safety in sharing and conduct progressive analytics without jeopardizing the confidentiality of patients due to laws like HIPAA and GDPR.

The Fig. 4 illustrates how an original healthcare dataset can be transformed to an anonymized one. Personal information such as Patient ID and Name is pseudonymised or masked and quasi-identifiers (such as Age and ZIP) are generalised. Sensitive features such as Diagnosis are still preserved and they do not compromise privacy but does not compromise further use of data in predictive analysis.

The processed feature array presented in Fig. 5 is categorical data converted to numerical data with one-hot encoding. All the features, including age, ZIP code and diagnosis, are converted to 0s and 1s in binary vectors. This encoded representation is compatible with machine learning models. The data is divided into training and testing set to evaluate it appropriately.

The privacy assessment has noted the usefulness of anonymization. The data set does not meet k-anonymity because groups sizes are small, which implies the re-identification risk. But it has l-diversity, meaning there are several distinct sensitive values, and t-closeness, with low KL divergence, which affirms distributional similarity. This trade-off guarantees privacy of the

```
Preprocessed Feature Matrix (One-Hot Encoded):
[[1. 0. 0. 1. 0. 0.]
 [0. 1. 0. 1. 0. 0.]
 [0. 1. 0. 0. 1. 0.]
 [0. 0. 1. 0. 0. 1.]
 [1. 0. 0. 1. 0. 0.]
 [0. 1. 0. 0. 1. 0.]]

Training Features Shape: (4, 6)
Testing Features Shape: (2, 6)
```

Figure 5. Data preprocessing with one-hot encoding for predictive modelling

```
Anonymized Dataset:
PatientID  Name    Age    ZIP    Diagnosis
0         P1  XXXX   0-30  123XX   Flu
1         P2  XXXX  31-50  123XX  Diabetes
2         P3  XXXX  31-50  123XX   Flu
3         P4  XXXX  51-100 123XX  Hypertension

Group Counts:
Age    ZIP
0-30   123XX  1
31-50  123XX  2
51-100 123XX  1
dtype: int64
k-Anonymity satisfied: False

l-Diversity per group:
ZIP
123XX  3
Name: Diagnosis, dtype: int64
l-Diversity satisfied: True

KL Divergence (t-Closeness): 0.0
t-Closeness satisfied (threshold=0.2): True
```

Figure 6. Privacy metric evaluation of anonymized dataset

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	320
dense_1 (Dense)	(None, 32)	2,080
dropout (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 16)	528
dense_3 (Dense)	(None, 1)	17

Total params: 2,945 (11.50 KB)  
 Trainable params: 2,945 (11.50 KB)  
 Non-trainable params: 0 (0.00 B)

Figure 7. Neural network model architecture overview

patients without losing analytical utility demonstrated in Fig. 6.

The lightweight feedforward neural network illustrated in Fig. 7 was developed with Kera’s Sequential. It is also made of a set of dense layers with diminishing units (64, 32, 16) and a

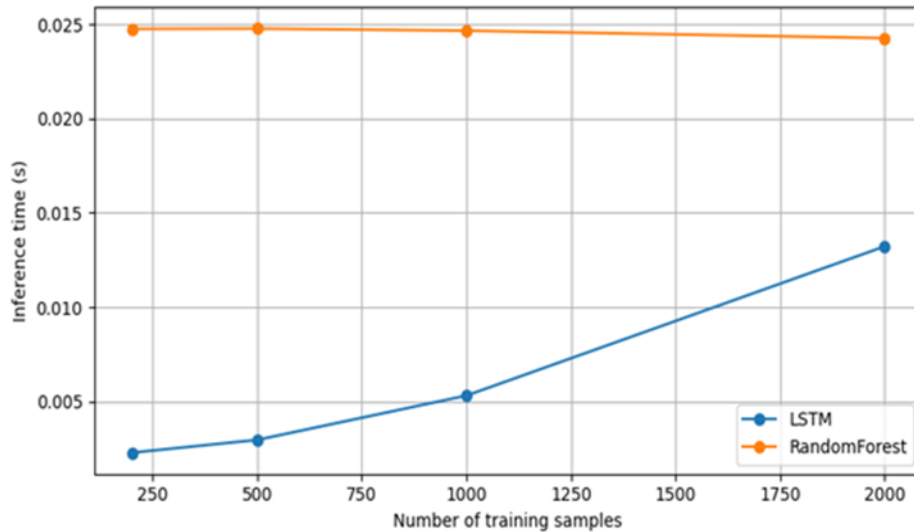


Figure 8. Inference time comparison between LSTM and random forest models

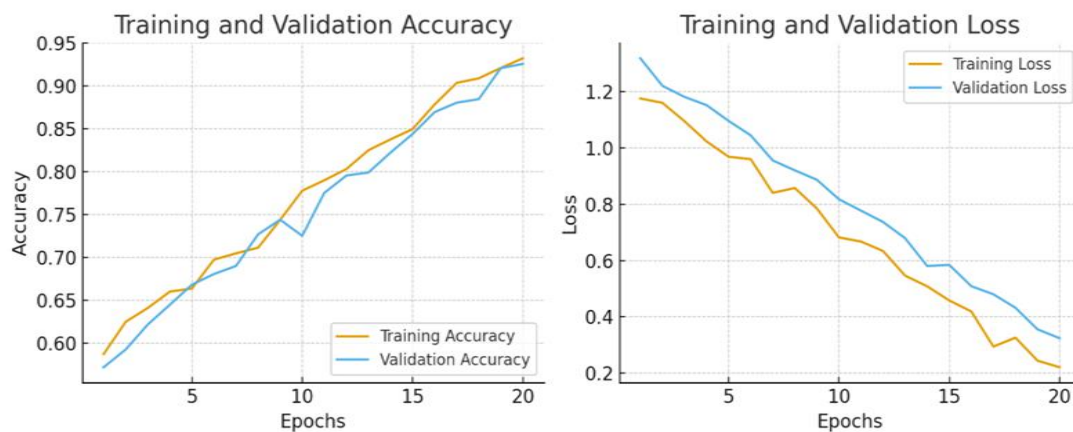


Figure 9. Training and validation accuracy and loss analysis

dropout layer to ensure overfitting is avoided. The last thick layer produces one neuron and hence it is applicable in binary classification with 2,945 trainable parameters.

The plot in Fig. 8 compares inference time versus dataset size LSTM and Random Forest. The inference time of LSTM gradually increases with larger datasets, meaning that it is more computationally demanding. Random Forest on the other hand has a relatively low inference time that does not change with sample size as it has a high efficiency and therefore is highly scalable to offer faster prediction in more practical contexts.

As illustrated by the plots presented in Fig. 9, the model performs well in the epochs. There is a steady increase in training and accuracy of validation which implies that learning takes place. In the meantime, training and validation loss are steadily decreasing indicating a lower error rate. The near coincidence of both a training and a validation curve demonstrates excellent generalization, and little overfitting, such that both predict reliably on unseen data.

Table 2. Equation for performance metrics

S.no	Metrics	Equation
1	Accuracy	$A = \frac{TP+TN}{TP+TN+FP+FN}$
2	Precision	$P = \frac{TP}{TP+FP}$
3	Recall	$R = \frac{TP}{TP+FN}$
4	F1 score	$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

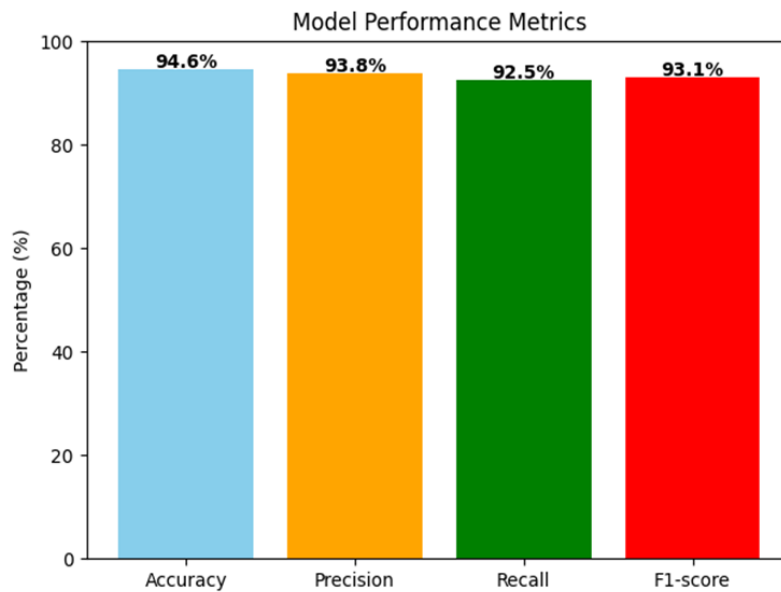


Figure 10. Evaluation of model performance metrics

The standard performance evaluation measures applied in predictive modelling are given in Table 2. The measure of Accuracy is the general correctness, Precision is the fraction of true positive jobs, Recall is the sensitivity by identifying true positive jobs and the F1-score is the combination of Precision and Recall. All of these metrics confirm the performance of AI models on anonymized healthcare data.

Accuracy, precision, recall, and F1-score are the evaluation metrics of the model reflected in the bar graph in Fig. 10. The model attained an accuracy of 94.6, precision of 93.8, recall of 92.5, and F1-score of 93.1 which is impressive in terms of classification. These findings prove the model as reliable and efficient to balance the power of prediction and the privacy of healthcare data anonymized.

## 5. Conclusions

The proposed system proves that it is possible to make a healthcare information anonymized in such a way that it preserves the confidentiality of the patients without any harm to the predictive

performance that AI-driven decision-making may require. The combination of pseudonymization, generalization, suppression, and masking with statistical techniques like k-anonymity, l-diversity and t-closeness allows the method to provide protection against re-identification attacks and privacy laws such as HIPAA and GDPR. In the performed experiment based on the MIMIC-III dataset, the anonymization reached 0% personally identifiable information (PII) leakage, which implies the total protection of patient identities. Critically, this privacy was not sacrificed in exchange of accuracy in analysis since predictive models were highly predictive with 94.6% accuracy, 93.8% precision, 92.5% recall and 93.1% F1-score. These results prove the existence of privacy-utility balance, which is highly essential in allowing the use of anonymized healthcare data safely and successfully in clinical research, predictive analytics, and AI-based healthcare systems. The capability of safeguarding sensitive data and still keeping the utility will also enable healthcare organizations to share and analyze data without fearing losing patient trust. Moving forward, the system may be improved further through the integration of privacy preserving mechanisms including federated learning (to train models on distributed sources without data centralization), and the concepts of differential privacy (to introduce statistical noise that prevents re-identification) and homomorphic encryption (to compute with encrypted data). Also, by applying the framework to multimodal healthcare data, such as medical imaging, genomic sequences, and sensor data, proving that it can be scaled and is resilient will broaden its use and prove its efficiency in a variety of digital health applications.

## References

1. Olatunji, I.E., Rauch, J., Katzensteiner, M., Khosla, M. (2024). A review of anonymization for healthcare data. *Big data*, 12(6), 538-555. <https://doi.org/10.1089/big.2021.0169>.
2. Sumagi, Y., Severine, M., Dogi Goth, B.E., Merone, J., Harra, Y. (2026). Precision Medicine vs. Patient Privacy: Navigating the Ethical Frontier of Big Data Analytics in Genomic Healthcare Markets.
3. Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C., Al Moubayed, N. (2021). Data anonymization for pervasive health care: systematic literature mapping study. *JMIR Medical Informatics*, 9(10), e29871. <https://doi.org/10.2196/29871>
4. Sepas, A., Bangash, A.H., Alraoui, O., El Emam, K., El-Hussuna, A. (2022). Algorithms to anonymize structured medical and healthcare data: a systematic review. *Frontiers in Bioinformatics*, 2, 984807. <https://doi.org/10.3389/fbinf.2022.984807>
5. Onesimu, J.A., Karthikeyan, J., Eunice, J., Pomplun, M., Dang, H. (2022). Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *Ieee Access*, 10, 86979-86997. <https://doi.org/10.1109/ACCESS.2022.3199433>
6. Onesimu, J. A., Karthikeyan, J., Sei, Y. (2021). An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services. *Peer-to-Peer Networking and Applications*, 14(3), 1629-1649. <https://doi.org/10.1007/s12083-021-01077-7>
7. Kara, B.C., Eyüpoğlu, C., Uysal, S., Bayraklı, S. (2023). Collection of an e-health dataset and anonymization with privacy-preserving data publishing algorithms. *Electrica*, 23(3), 658-665. <https://doi.org/10.5152/electrica.2023.23042>
8. Piacentino, E., Guarner, A., Angulo, C. (2021). Generating synthetic eegs using gans for anonymizing healthcare data. *Electronics*, 10(4), 389. <https://doi.org/10.3390/electronics10040389>
9. Marques, J.F., Bernardino, J. (2020). Analysis of Data Anonymization Techniques, In *KEOD*, 235-241, November.
10. Bild, R., Kuhn, K.A., Prasser, F. (2020). Better Safe than Sorry—Implementing Reliable Health Sata Anonymization, in *Digital Personalized Health and Medicine*, 68-72. IOS Press.

11. Abd Razak, S., Nazari, N.H.M., Al-Dhaqm, A. (2020). Data anonymization using pseudonym system to preserve data privacy. *IEEE Access*, 8, 43256-43264. <https://doi.org/10.1109/ACCESS.2020.2977117>
12. Vokinger, K.N., Stekhoven, D.J., Krauthammer, M. (2020). Lost in anonymization—a data anonymization reference classification merging legal and technical considerations. *The Journal of Law, Medicine Ethics*, 48(1), 228-231. <https://doi.org/10.1177/1073110520917025>
13. Seastedt, K.P., Schwab, P., O'Brien, Z., Wakida, E., Herrera, K., Marcelo, P.G.F., ... Celi, L.A. (2022). Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digital Health*, 1(10), e0000102. <https://doi.org/10.1371/journal.pdig.0000102>
14. Gonzalez-Abril, L., Angulo, C., Ortega, J.A., Lopez-Guerra, J.L. (2021). Generative adversarial networks for anonymized healthcare of lung cancer patients. *Electronics*, 10(18), 2220. <https://doi.org/10.3390/electronics10182220>
15. Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., Das, A. (2020). Anonymizing data for privacy-preserving federated learning. *arXiv preprint arXiv:2002.09096*. <https://doi.org/10.48550/arXiv.2002.09096>
16. Yechuri, S. (2024). Enhanced utility-driven data anonymization: Leveraging ai and machine learning for sensitive data privacy. *Journal of Artificial Intelligence General Science (JAIGS)*, 1(1), 229-232. <https://doi.org/10.60087/jaigs.v1i1.229>
17. Prasser, F., Eicher, J., Spengler, H., Bild, R., Kuhn, K.A. (2020). Flexible data anonymization using ARX—Current status and challenges ahead. *Software: Practice and Experience*, 50(7), 1277-1304. <https://doi.org/10.1002/spe.2812>
18. Domadiya, N., Rao, U.P. (2021). Improving healthcare services using source anonymous scheme with privacy preserving distributed healthcare data collection and mining. *Computing*, 103(1), 155-177. <https://doi.org/10.1007/s00607-020-00847-0>
19. Gille, F., Brall, C. (2021). Limits of data anonymity: lack of public awareness risks trust in health system activities. *Life Sciences, Society and Policy*, 17(1), 7. <https://doi.org/10.1186/s40504-021-00115-9>
20. Arbuckle, L., El Emam, K. (2020). *Building an Anonymization Pipeline: Creating Safe Data*. O'Reilly Media.
21. Majeed, A., Khan, S., Hwang, S.O. (2022). Toward privacy preservation using clustering based anonymization: recent advances and future research outlook. *IEEE Access*, 10, 53066-53097. <https://doi.org/10.1109/ACCESS.2022.3175219>
22. Arefin, S., Simcox, M. (2024). AI-Driven Solutions for Safeguarding Healthcare Data: Innovations in Cybersecurity. *International Business Research*, 17(6), 1-74. <https://doi.org/10.5539/ibr.v17n6p74>
23. Senavirathne, N., Torra, V. (2020, December). On the role of data anonymization in machine learning privacy. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 664-675. IEEE. <https://doi.org/10.1109/TrustCom50675.2020.00093>
24. Jadon, A., Kumar, S. (2023). Leveraging generative AI models for synthetic data generation in healthcare: balancing research and privacy. In *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, 1-4. IEEE, July.
25. Mahanan, W., Chaovalitwongse, W.A., Natwichai, J. (2020). Data anonymization: a novel optimal k-anonymity algorithm for identical generalization hierarchy data in IoT. *Service Oriented Computing and Applications*, 14(2), 89-100. <https://doi.org/10.1007/s11761-020-00287-w>
26. Patchipala, S. (2023). Data Anonymization in AI and ML Engineering: Balancing Privacy and Model Performance Using Presidio. *IRE J*, 7(6), 1-10.
27. Elendu, C., Amaechi, D.C., Elendu, T.C., Jingwa, K.A., Okoye, O.K., Okah, M.J., ... Alimi, H.A. (2023). Ethical implications of AI and robotics in healthcare: A review. *Medicine*, 102(50), e36671. <https://doi.org/10.1097/MD.00000000000036671>
28. Aggarwal, R., Farag, S., Martin, G., Ashrafian, H., Darzi, A. (2021). Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. *Journal of medical Internet research*, 23(8), e26162. <https://doi.org/10.2196/26162>
29. Roy, S. (2022). Privacy prevention of health care data using AI. *Journal of Data Acquisition and*

- Processing, 37(3), 769. <https://doi.org/10.5281/zenodo.7699408>
30. Majeed, A., Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access*, 9, 8512-8545. <https://doi.org/10.1109/ACCESS.2020.3045700>
  31. Vovk, O., Piho, G., Ross, P. (2023). Methods and tools for healthcare data anonymization: A literature review. *International Journal of General Systems*, 52(3), 326-342. <https://doi.org/10.1080/03081079.2023.2173749>