Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Federated Learning with Privacy-Preserving Big Data Analytics for Distributed Healthcare Systems

Shuchona Malek Orthi[1], Md Habibur Rahman[2] ✉ Kazi Bushra Siddiqa[3], Mukther Uddin[4], Sazzat Hossain[5], Mohd Abdullah Al Mamun[6] and Md Nazibullah Khan[7]

[16]*College of Business, Westcliff University, Irvine, CA 92614, USA*
[23457]*School of Business, International American University, Los Angeles, CA 90010, USA*

Corresponding Author: Md Habibur Rahman, **E-mail**: rahman@aub.ac.bd

| ABSTRACT

An architecture for privacy-preserving federated learning that can classify chest X-ray pictures of tuberculosis (TB) in decentralized healthcare settings. The suggested solution protects patient privacy and ensures regulatory compliance by facilitating the cooperative training of ML models across many healthcare organizations without necessitating direct access to private patient data. A structured data pre-processing pipeline is implemented, including initial inspection, image resizing to 224×224 pixels, normalization, and class balancing using Synthetic Minority Oversampling Technique (SMOTE) to manage non-IID distributions across federated clients. The federated setup simulates realistic clinical environments where each node holds a portion of the dataset and only shares model updates with the central server. The ResNet deep learning model is deployed as the primary classifier and its performance is evaluated against Dense Net and Squeeze Net using four key evaluation metrics: accuracy, precision, recall, and F1-score. Compared to the comparison models, the suggested ResNet model achieves better performance according to accuracy (96.7%), precision (96.8%), recall (98.0%), and F1-score (97.4%). Squeeze Net achieved a rate of 94.18% accuracy and Dense Net 94%. This framework's integration with cloud-based platforms increases its scalability and real-time applicability. It offers a secure, scalable, and high-performance solution for tuberculosis (TB) diagnosis in healthcare environments through the use of federated learning and big data analytics. The results validate its potential as a foundation for broader applications in privacy-aware medical AI systems.

| KEYWORDS

Federated Learning, Privacy Preservation, TB Chest X-ray dataset, Machine learning, Big Data Analytics, ResNet model.

| ARTICLE INFORMATION

ACCEPTED: 01 July 2025 **PUBLISHED:** 01 August 2025 **DOI:** 10.32996/jcsts.2025.7.8.31

1. Introduction

A great deal of data has been produced by electronic health records (EHRs), diagnostic imaging, mobile health applications, and wearable sensors as a result of the digitalization of healthcare systems. The statistics presented here have the ability to greatly enhance clinical decision-making, facilitate early diagnosis, and pave the way for customized treatment plans [1][2]. Privacy concerns, questions of data ownership, and issues with regulatory compliance are all magnified by the delicate nature of healthcare data. Traditional ML approaches that depend on consolidating huge datasets into one place are hindered by strict regulations like GDPR and HIPAA, which impose limits on the centralization of medical data.

The use of big data in healthcare is much esteemed both domestically and internationally [3]. To put it in perspective, certain industrialized nations' platforms are more developed than others. Despite the fact that medical data has a wealth of useful information that can revitalize fields like public health and health management, there is a growing concern about privacy breaches due to the increasing complexity and heterogeneity of healthcare data. Big data analytics, on the other hand, is a crucial tool for handling and understanding massive amounts of medical data [4][5]. Big data in medicine enables high-dimensional pattern recognition, population-level trend analysis, and predictive modeling for diseases such as diabetes, cardiovascular conditions, and cancer [6][7]. However, centralizing big data from distributed healthcare providers introduces concerns regarding data breaches, operational costs, and institutional reluctance to share proprietary or sensitive patient information [8]. It emphasized the necessity of privacy-conscious and decentralized learning systems that would be capable of utilizing big data without going against moral or legal norms.

Model training could occur on client devices, local institutional servers but within the decentralized approach of FL, the only data shared with a central aggregator would be model parameters, rather than raw data. This configuration allows shared learning with high data security [9][10]. Various ML and DL models have been adapted for FL in healthcare [11], including Logistic Regression for binary classification tasks SVM for pattern recognition, and DT for interpretable risk assessment [12]. Further sophistication uses Multilayer Perceptron (MLPs) and structured data, CNNs in the case of medical imaging, and LSTM networks to analyze sequential data like ECG or vital signs in patients [13]. Training such models in a federated setting avoids analytical value of distributions of the local data and maintains patient privacy.

The versatility of FL can lead to the training of ML algorithms to learn based on productions across many client-side data in a succession of running iterations [14]. Federated learning, in conjunction with big data analytics, can give rise to privacy-safe, scalable, collaborative intelligence at distributed healthcare settings [15]. It enables real applications like sharing of diagnostic models between different hospitals, patient monitoring using AI via the edge device, and sharing multi-institutional research data without jeopardizing sensitive patient data [16][17]. Furthermore, this connection facilitates the creation of powerful prediction models for early illness identification, individualized treatment recommendations, and hospital resource efficiency. By enabling institutions to contribute to model training while retaining data locally, federated learning not only addresses legal and ethical constraints but also enhances model generalization by learning from diverse and non-IID datasets across heterogeneous healthcare systems.

1.1 Motivation with Contribution

The exponential growth of healthcare data, driven by digital diagnostics and imaging technologies, has made data-driven decision-making increasingly important in clinical environments. However, traditional centralized machine learning approaches pose significant risks to patient privacy, violate regulatory constraints, and lack scalability across multiple medical institutions. The ongoing worldwide epidemic of TB highlights the need of developing diagnostic tools that are both smart and privacy-conscious, able to make use of dispersed data without jeopardizing personal patient information. A new paradigm called FL has recently surfaced as a viable option for collaborative model training over distributed nodes that maintain data locality. Motivated by this, their study aims to apply FL to large-scale TB chest X-ray datasets, addressing the dual challenges of diagnostic accuracy and data confidentiality in healthcare systems. This research primarily contributes to the following areas:

- Developed a secure and privacy-aware federated learning architecture tailored for distributed TB chest X-ray classification across multiple healthcare nodes, ensuring compliance with data protection standards.
- A comprehensive preprocessing pipeline is implemented, including image resizing, normalization, and SMOTE-based class balancing to handle non-IID data distribution.
- The federated environment is used to train and assess the proposed ResNet model, which is then compared to Dense Net and Squeeze Net models.
- The suggested ResNet model's efficacy in TB chest X-ray classification inside a federated learning framework, surpassing Dense Net and Squeeze Net in every important evaluation measure (accuracy, precision, recall, and F1-score) on a consistent basis.
- The architecture proves that federated learning may help with scale implementation in clinical contexts, maintain privacy of patient data, and provide good diagnostic results.

1.2 Significance and Novelty

This research proposes a novel FL framework for TB chest X-ray classification that prioritizes privacy preservation and scalability in distributed healthcare environments. Contrary to traditional centralized models, which require the exchange of sensitive patient data with a single server, the proposed solution allows collaborative model training across several institutions while maintaining secrecy. The framework incorporates client-specific preprocessing, including image resizing, normalization, and SMOTE-based balancing to address data heterogeneity and imbalance across non-IID clients. It also integrates deep convolutional models such as ResNet within a federated setting, achieving superior performance compared to Dense Net and Squeeze Net. large-scale data processing and real-time model training. This combined focus on federated deep learning, big data handling, and secure, scalable deployment makes the proposed approach a practical and novel solution for modern privacy-aware healthcare systems.

1.3 Structure of Paper

The remainder of the document is structured as follows: Subsequently, Section II examines relevant literature on healthcare federated learning. Section III presents the suggested approach. Section IV covers the main points covered, as well as the experimental findings and a comparison of the models. Section V offers a concise overview of the subject and proposes potential research directions.

2. Literature Review

The use of FL for precise, private, and scalable healthcare analytics is reviewed in this section. Table I summarizes the examined papers, emphasizing important datasets, techniques, conclusions, and future directions that are pertinent to federated healthcare systems.

Zhou et al. (2025) FL has emerged as a key enabler of privacy-preserving distributed model training in edge computing environments, crucial for service-oriented applications such as personalized healthcare Federated Learning (RCFL) framework designed to enhance privacy protection and communication efficiency in edge-based service environments. RCFL integrates a global privacy-preserving mechanism with an innovative privacy encoding strategy that minimizes privacy risks over multiple data releases while significantly reducing communication overhead. The proposed framework's theoretical analysis demonstrates its ability to maintain differential privacy across numerous interactions, ensuring robust model convergence and efficiency. Experimental results using MNIST and CIFAR-10 datasets reveal that RCFL can lower the MPLA success rate from 88.56% [18].

Rampone, Ivaniv and Rampone (2025) federated learning-based healthcare model leveraging MIMIC-III dataset achieving a high accuracy of 95% ROC-AUC while preserving patient privacy. By integrating cloud platforms like Databricks and Google Cloud Storage with scalable tools such as Apache Spark and MLlib, the model processes over 1.6 million records in just 0.16 seconds. The results highlight the feasibility of real-time, privacy-preserving analytics, making federated learning a powerful solution for secure, distributed healthcare and IoMT environments [19].

Mohanty, Dash and Tripathy (2024) privacy-preserving, decentralized learning strategies are necessary for NLP and ML problems because of increasing rules and regulations. Federated learning (FL) provides straightforward methods for multiple clients to work together to train a single global model that meets the objectives of all clients while protecting local data, offering continuous training, and building a useful platform to handle heterogeneous input from various models. a novel Improved-FLNLP and analyzes NLP workflow for smart healthcare with a special focus on FL. IFLNLP is an analytical framework for FL and NLP-based smart healthcare systems. the proposed approach, IFLNLP, yields the highest level of data privacy (98%) and cost effectiveness (95%), showing its strong privacy protection skills and better cost effectiveness as compared to NLP and FL-based system and comparatively better accuracy than the NLP-based system [20].

N, S and R (2024) Federated learning (FL) and its evolution path in healthcare. to scope a wide variety of healthcare applications in FL. Exactly what research direction is moving in interesting for research communities to guide their future course, uniquely focuses on examining numerous FL-based healthcare implementations, detailing their core methodologies and performance metrics, which, to their knowledge. Privacy-preserving collaborative distributed learning through federated learning in healthcare enhances research collaborations, thereby resulting in better-performing models [21].

Tan et al. (2023) the healthcare industry has made extensive use of ML and DL for big data analysis to guarantee computing efficiency while also protecting private information via data security and confidentiality. An optimized DSFedL framework is implemented via a data-sharing hub by assessing an accuracy-privacy loss function. This refers to the data-sharing function of the FedL model. Their optimized DSFedL has been tested on three open-source Cardiothoracic Databases (ICBHI, Cogware COVID-19, and MIT-BIH Arrhythmia) and found to be effective in simulating nonIID datasets. The results demonstrate an optimal outcome according to data security and efficiency [22].

Tiwari et al. (2023) federated learning (FL) can be better solutions for classification and privacy issues related to skin datasets. The collaboration of both in the medical field can be a significant contribution. Federated learning-based deep learning was applied to a publicly available dataset consisting of 10 different skin diseases. Various deep learning models were applied, and privacy concerns were preserved with a federated learning approach, the dataset was enhanced with several image augmentation strategies. several models were used and after observation, it was found that the Inception Net outperformed and produced a better accuracy rate of 91% [23].

Hegde, Shenoy and Venugopal (2023) ML/ DL techniques to EHR can make the healthcare system smarter. However, the main issue is that preserving the patient's privacy is paramount, Federated Learning (FL) approach is used to build a smart healthcare system and maintain the patient's privacy. This study compares the ANN and Logistic Regression (LR) in the FL environment. FL-ANN and FL-LR framework is designed with three hospitals and is executed independently with diabetes and CKD (chronic kidney disease) datasets. These frameworks are done based on the performance metrics and the FL Process Time. The outcome is that the FL-LR framework outperforms FL-ANN concerning FL-Process Time. The FL-LR with CKD dataset results in a better accuracy of 92% with FL Process Time of 10.5 seconds [24].

Several recent studies have investigated an application of FL techniques within the healthcare domain, demonstrating significant advancements in achieving privacy-preserving distributed model training while maintaining high predictive accuracy across decentralized clinical environments. A variety of models, including ensemble approaches such as RF, XGBoost, and federated neural networks, have been employed to address challenges related to data imbalance and heterogeneity inherent in electronic health records. In particular, certain frameworks leverage federated clustering and anomaly detection mechanisms to identify rare disease patterns across multiple data silos. Additionally, other approaches emphasize the effectiveness of FL-integrated deep learning models and logistic regression, further enhanced by the incorporation of differential privacy and secure aggregation protocols to support real-time, privacy-compliant medical decision-making. Despite these promising outcomes, key challenges remain, including the need to handle non-IID data distributions, improve model interpretability, and ensure scalability across diverse healthcare infrastructures. Furthermore, ongoing research highlights the importance of developing robust FL architectures that combine temporal health data analysis with explainable AI (XAI) methodologies to foster transparency, clinical trust, and ethical deployment in real-world healthcare systems.

Table 1: Comparative Analysis of Recent Studies on Federated Learning in Big Data Healthcare

Author(s)	Dataset	Methodology	Key Findings	Advantages	Limitations	Future Work
Zhou et al. (2025)	MNIST, CIFAR-10	RCFL framework with privacy encoding and communication optimization	Maintains differential privacy; reduces communication overhead; lowers MPLA success rate from 88.56%	High privacy protection; efficient communication	Limited to image datasets; not validated in real healthcare data	Apply RCFL to real-world healthcare data and more diverse edge scenarios
Rampone, Ivaniv & Rampone (2025)	MIMIC-III	FL model with Spark, MLlib, Databricks, and GCP	95% ROC-AUC; processes 1.6M records in 0.16 sec	High scalability and accuracy; cloud integration	Focuses only on structured clinical data	Extend to image/EHR integration and test in live hospital settings
Mohanty, Dash & Tripathy, (2024)	NLP in Smart Healthcare	IFLNLP: Improved FL framework for NLP	98% data privacy; 95% cost effectiveness; better accuracy than standalone NLP	Strong privacy & cost performance in NLP	Limited comparison with other FL-NLP models	Extend to multilingual health NLP and cross-institution collaboration
T N, S and R (2024)	Multiple healthcare FL studies	Systematic analysis of FL evolution in healthcare	FL enhances collaborative research and privacy-preserving model development	Comprehensive overview for future guidance	Lacks original experimentation	Deeper empirical benchmarking of proposed directions
Tan et al. (2023)	ICBHI, Coswara COVID-19, MIT-BIH	DSFedL: Optimized FL with accuracy-privacy tradeoff function	Handles non-IID data efficiently with good privacy and accuracy	Balanced tradeoff between security and performance	Simulated datasets only	Apply DSFedL to real-world clinical datasets
Tiwari et al. (2023)	Skin Disease Dataset	FL with Deep Learning + Image Augmentation	Inception Net achieved 91% accuracy; privacy preserved	Effective model and augmentation strategy	Results limited to dermatology domain	Explore other modalities like radiology and oncology
Hegde, Shenoy & Venugopal (2023)	Diabetes and CKD datasets	FL-ANN and FL-LR across three hospitals	FL-LR achieves 92% accuracy with faster training (10.5s)	Strong privacy with performance comparison	Limited to two disease types and small network	Extend to federated multi-disease diagnosis with hybrid models

3. Methodology

A privacy-preserving federated learning architecture for tuberculosis chest X-ray classification in a dispersed healthcare setting is outlined in the suggested technique, as shown in Figure 1. To start, the TB chest X-ray dataset is acquired. Then, there is a structured data preparation phase where the images are inspected first, resized to 224×224 pixels, and normalized to ensure that the intensity values of images are consistent. To address data imbalance, particularly within non-IID client distributions, SMOTE is employed. The dataset is then split into training, validation, and testing sets while preserving class ratios. Federated learning presents advantages regarding the confidentiality of patient data, as it allows keeping client-specific data at the local level and keeping global model updates synchronized in a secure manner. ResNet is used as the main model in classification, and the comparison of its performance is made with Dense Net and Squeeze Net because of its better feature extraction. The metrics applied to measure the model are F1-score, recall, accuracy and precision. In this federated architecture, the potential of deep learning to be applied to large healthcare data and retain privacy and scalability were shown, thus enabling real-time diagnostics in multi-institutional medical systems.

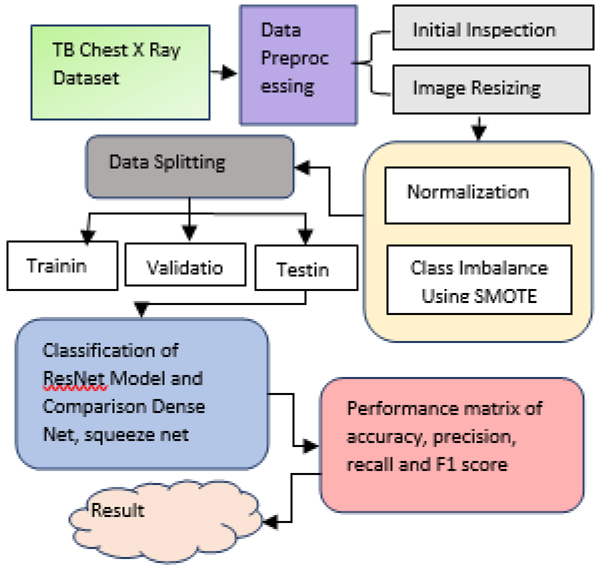

Fig. 1. Flowchart for Healthcare Using Federated Learning Models

3.1 Data Collection

The data used in this study was on publicly accessible websites like Kaggle that provided a large number of medical images that played a critical role in determining the effectiveness and scalability of the recommended models in a Big Data healthcare environment. As an illustration, the chest X-ray dataset about TB has 7000 images, 3500 with a normal result and 3500 that exhibit symptoms of tuberculosis. This enables the models to effectively distinguish healthy instances and those affected by tuberculosis. The dataset also forms a good basis to verify the model's ability to scale and perform accurate and privacy-preserving diagnoses as the size and variety of the data are quite resembles the situation in large-scale data environments frequently encountered in real-world healthcare analytics.

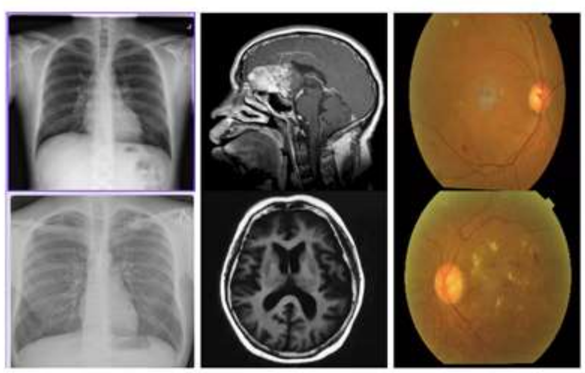

Fig. 2. Sample Image of the Dataset Utilized

Figure 2 shows several photos from the dataset to highlight how diverse it is and how it may be used in a healthcare Big Data setting:

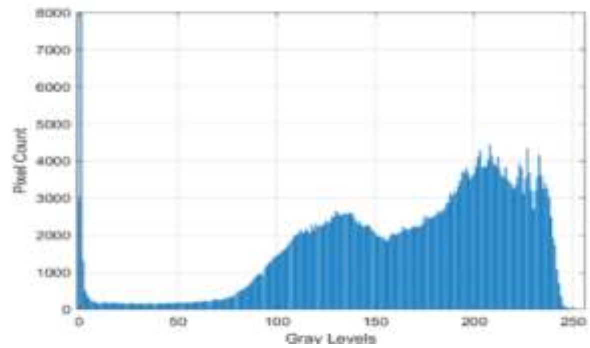

Fig. 3. Histogram for Grayscale Pixel Intensity Distribution of a Chest X-ray Image

The grayscale intensity distribution of a TB chest X-ray image, where the majority of pixels fall within the higher intensity range (150–250), indicates a brighter image with clear anatomical structures in Figure 3. Lower counts in the darker pixel ranges (0–100) suggest minimal noise or underexposed regions. This contrast-rich distribution supports effective feature extraction during deep learning-based classification. Such histogram analysis is essential in the preprocessing phase to ensure consistent image quality across distributed clients.

3.2 Data Preprocessing

Data preprocessing involved inspecting the TB Chest X-ray dataset for class distribution, image quality, and consistency. To make sure that deep learning models were given uniform input, all photos were first normalized and then scaled to 224×224 pixels. SMOTE was applied to balance the TB-positive and normal cases among the federated clients with class imbalance. The data was further partitioned into training, validation and testing to arrive at clean, balanced, and structural data that federated model training. Key steps in data preprocessing include:

- **Initial Inspection:** Initial inspection involved reviewing the dataset for class distribution, image quality, and format consistency. This step ensured there were no corrupted or mislabeled files and confirmed equal representation of TB-positive and normal cases.
- **Image Resizing:** To guarantee compatibility with common deep learning models like ResNet, the chest x-ray pictures were scaled to 224×224 pixels. This resizing step standardizes input dimensions, enabling efficient batch processing and stable model training.

3.3 Normalization

To ensure that the model converges more smoothly when training, normalization is an essential preprocessing step that uniformly adjusts the pixel values. For image data, pixel intensities (originally ranging from 0 to 255) typically normalized to a [0, 1] range using Equation (1).

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x is the original pixel value. This transformation reduces the impact of varying image intensities and helps the neural network learn more efficiently. It also ensures numerical stability in gradient-based optimization algorithms.

3.4 Class Imbalance Using SMOTE

During the preprocessing step, SMOTE was used to reduce data imbalance. In order to decrease class bias while maintaining data distribution features, SMOTE generates synthetic samples that efficiently improve the representation of minority classes. The balanced client configuration after preprocessing proved that the model could withstand a certain level of infrastructure and data heterogeneity, as shown in the SMOTE research. Both the pre- and post-SMOTE distributions of client data are shown in Figure 4.

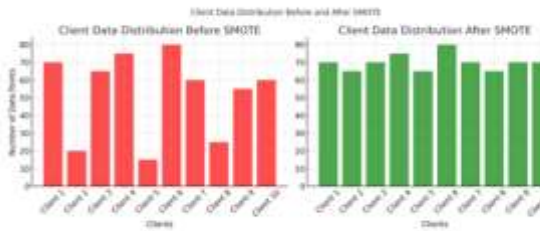

Fig. 4. Client Data Distribution Before and After SMOTE

The original biased distribution is shown in the left subplot. Clients 2, 5, and 8 had much reduced data representation. Next, it can see the balanced distribution after SMOTE preparation in the right subplot. This will guarantee that the data distribution is unaltered after SMOTE.

3.5 Data Splitting

The dataset was split into 70% for training, 15% for validation, and 15% for testing to ensure effective model evaluation. This stratified split maintains class distribution across all subsets, supporting balanced learning and unbiased performance measurement.

3.6 Classification Equation of Proposed Model Resnet in Healthcare System

ResNet's high performance is applicable to a wide range of tasks, including picture classification, picture creation, visual identification, NLP, voice recognition, and user prediction [25]. The residual unit's fundamental structure is shown in Figure 5. H(x) is the Input Value's underlying mapping after two branches, whereas F(x) is the Input Value's residual mapping following two weight layers. It is evident that by using an identity function as the shortcut link, the Residual Unit shifts the problem from fitting the relationship among H(x) and x to the relationship among F(x) and x.

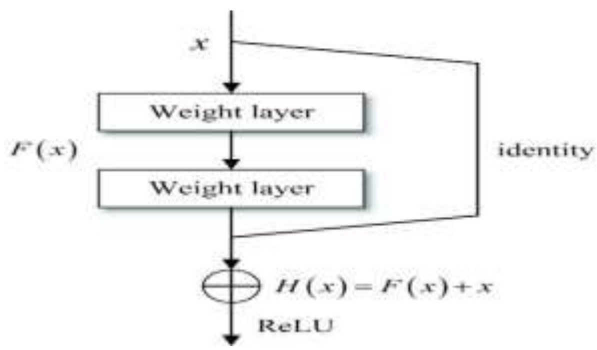

Fig. 5. ResNet Architecture

The ReLU activation function, which is either an identity function or a constant function of 0, is used as the activation function prior to the output layer of the residual unit in the residual network. In order to facilitate learning, it combined several leftover components. To begin with, let's pretend that the ReLU activation functions shown in Figure 5 are identity functions in Equation (2).

$$x_{l+1} = x_1 + F(x_1, \{W_1\}) \qquad (2)$$

where x_1 denotes the input of the l_{th} ResNet unit, and W_1denotes the weights. x_{l+1} and $F(x_1, \{W_1\})$ stand for an output that is directly forward propagated and the residual mapping that has to be learnt from the l_{th} ResNet unit, respectively

Equation (3) may then be used to describe the residual network's forward propagation process.

$$x_L = x_1 + \sum_{i=1}^{L-1} F(x_i, \{W_i\}) \qquad (3)$$

Where x_L denote the accumulated output of L-1 Connected Residual Unit.

3.7 Fedrated Learning Set Up

An employment of a Federated Learning architecture allows for the decentralization of sensitive patient data while facilitating collaborative model training across several dispersed healthcare nodes. The TB chest X-ray dataset is distributed over ten client nodes, each of which represents a distinct healthcare facility with data distributions independent of IID. Using a federated learning system, medical picture categorization may be done in a way that protects patients' privacy. Clients may train their models locally in this setup, eliminating the need to transmit raw data to a central server. A more effective and precise global model is the end product. Due to the non-IID storage of medical imaging data, each client is granted access to a randomly chosen portion of the dataset. varied medical facilities or equipment may have data with varied distributions, and thus, non-IID arrangement matches real-world realities. The model parameters are updated in the following way during local training using SGD, which updates the local model on each client k, represented as w_t^k in Equation (4).

$$w_t^k = w_{t-1}^k - \eta \nabla L_k\left(w_{t-1}^k\right) \tag{4}$$

Where η stands for the Learning Rate and $\nabla L_k\left(w_{t-1}^k\right)$ denotes a gradient of the loss function L_k for client k's local dataset.

3.8 Performance Matrix

Accuracy, precision, recall, and F1-score are some of the conventional performance measures utilized to assess the suggested federated learning model's efficacy in classifying TB chest X-rays. These measurements provide a thorough grasp of the model's diagnostic efficacy across dispersed healthcare nodes. It is also possible to examine the results of predictions using a confusion matrix, where TP stands for correctly identified tuberculosis (TB) positive cases, TN for correctly identified normal cases, FP for normal cases wrongly predicted as TB positive, and FN for TB positive cases wrongly predicted as normal. These metrics assess the model's ability to generalize across different types of client data while protecting patient privacy in a DL setting.

1. Accuracy

A popular statistic for assessing how accurate a classification model's predictions is overall is accuracy. In Equation (5) below, it is computed as the ratio of properly identified examples to all occurrences in the dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{5}$$

2. Precision

Precision is defined as the ratio of correctly categorized photos to all classified images [26]. The precision (Pr) may be expressed using Equation (6):

$$Precision = \frac{TP}{TP+FR} \times 100 \tag{6}$$

3. Recall

Recall measures how many out of all images in the database were properly classified. Equation (7) provides a concise formula that summarizes it as follows:

$$Recall = \frac{TP}{TP+FN} \times 100 \tag{7}$$

4. F1 Score

A higher F-score suggests that the system has a more effective predictive capacity, as it represents the harmonic means of recall and precision. Precision and recall alone, however, are inadequate metrics for evaluating system performance. According to Equation (8), the F-score can be determined as follows:

$$F1 - score = \frac{2 \times recall \times precision}{recall + precision} \tag{8}$$

4. Results and Discussion

The experimental results for medical image classification using DL approaches on TB chest X-ray datasets are presented in this section. Model performance is evaluated using important measures such as F1-score for binary classification tasks, recall, accuracy, and precision. This implementation was executed on a Google Colab environment using the Python programming language in a Jupiter Notebook. It made use of core Python libraries like TensorFlow, Kera's, pandas, NumPy, seaborn, and matplotlib PyTorch. The system specifications included an Intel R Xeon R CPU E5-2697 v4 @ 2.30GHz, 64 GB RAM, and a 16 GB NVIDIA GeForce GTX 1080 GPU, which were used to train the ResNet architecture. The analysis includes performance evaluation of the ResNet model through comprehensive metrics including confusion matrix analysis, ROC curve assessment, and training-validation loss curves. The following outputs provide detailed insights into the medical diagnosis results, supporting the effectiveness of the proposed approach for automated healthcare imaging systems, incorporating federated learning methodologies with privacy-preserving analytics for distributed medical networks.

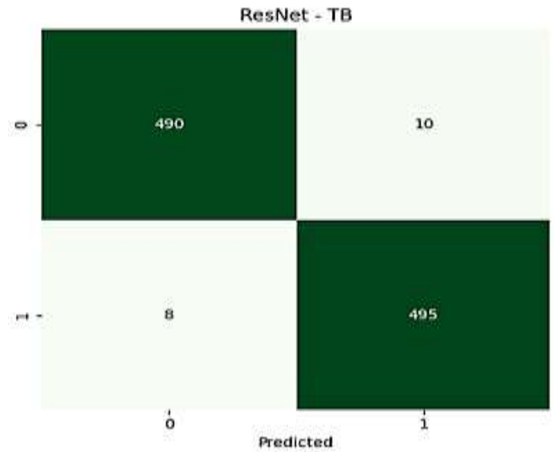

Fig. 6. Confusion Matrix of ResNet Model

This confusion matrix depicts the assessment of a ResNet model's performance in TB detection. Figure 6 shows the classification results with the y-axis displaying the actual labels and the x-axis showing the predicted labels. The model achieved high accuracy with 490 TN, 495 TP, 10 FP, and 8 FN. This ResNet-based TB detection system could assist radiologists in automated chest X-ray screening for early tuberculosis diagnosis in clinical settings. The implementation leverages big data analytics in healthcare.

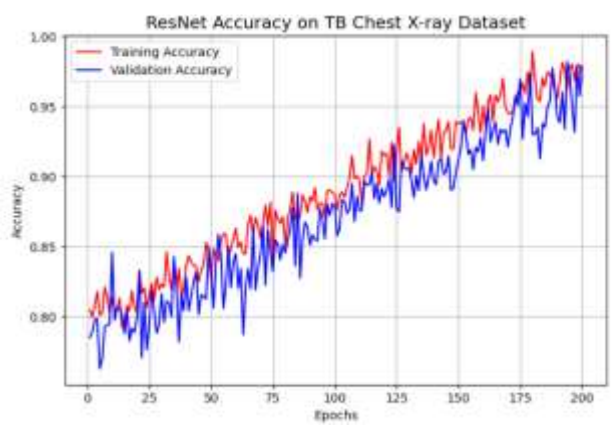

Fig. 7. Accuracy Graph of ResNet Classifier

Figure 7 illustrates the training and validation accuracy curves for a ResNet model applied to tuberculosis (TB) chest X-ray classification over 200 epochs in Figure 7. Training accuracy is shown by the red line, and validation accuracy by the blue line. Both curves demonstrate consistent improvement from approximately 80% to 98% accuracy, with convergence around epoch 150, indicating effective model learning without significant overfitting.

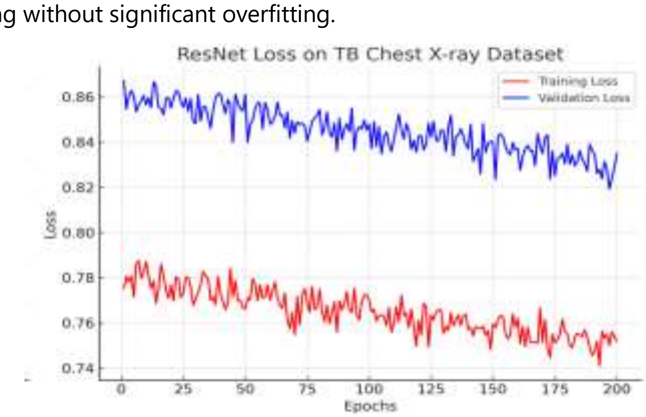

Fig. 8. Loss Graph of ResNet Classifier

Figure 8 shows the ResNet model's training and validation loss curves on the tuberculosis (TB) chest X-ray dataset with 200 epochs. Training loss is shown by the red line and validation loss by the blue line. Both curves demonstrate steady decline from initial values around 0.78-0.86 to final values of approximately 0.75-0.83, indicating consistent model optimization and convergence without overfitting throughout the training process.

Table 2: Resnet Proposed Models' Performance On Tb Chest X-Ray Dataset

Measure	ResNet
Accuracy	96.7
Precision	96.8
Recall	98.0
F1-score	97.4

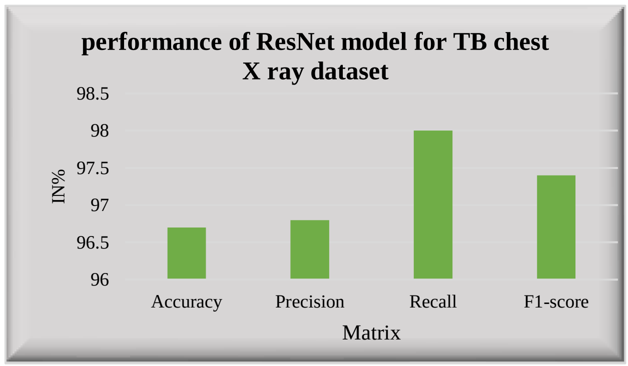

Fig. 9. Performance Metrics of Chest X-ray Dataset

Figure 9 presents Table II presents the performance metrics of the proposed ResNet model on the TB chest X-ray dataset. The model demonstrates exceptional diagnostic capabilities with an accuracy of 96.7%, precision of 96.8%, recall of 98.0%, and F1score of 97.4%. The model's strong performance in medical imaging classification tasks, as measured by these criteria, shows that it can reliably detect TB patients with few false positives and negatives. This approach enables federated learning with privacy-preserving analytics for distributed healthcare systems, allowing collaborative model training while maintaining patient data confidentiality across medical institutions.

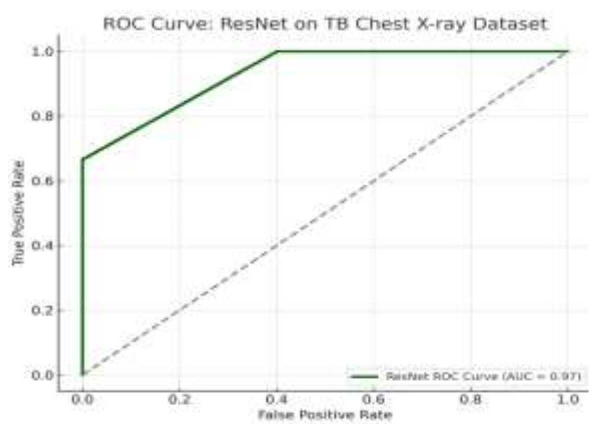

Fig. 10. ROC curve of ResNet Model

Figure 10 shows the ROC curve for a ResNet model applied to TB chest X-ray classification. The green curve demonstrates excellent diagnostic performance with an AUC of 0.97, significantly outperforming the ResNet model. The curve shows high TPR across low FPR, indicating superior discriminative ability for big data applications in healthcare, utilizing machine learning.

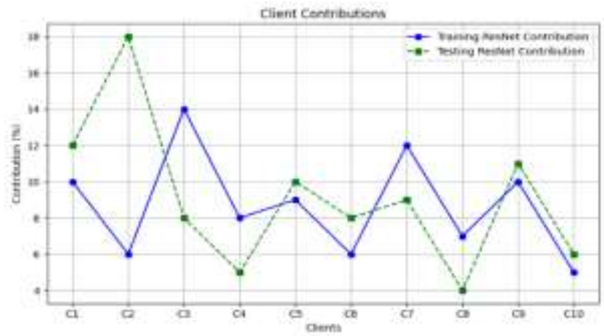

Fig. 11. Line Graph of Client Contribution in the Federated Setup with ResNet Model

A line graph comparing client-wise contribution percentages between Training and testing ResNet models within a federated learning environment in Figure 11. The plot reveals fluctuations in contributions across clients, with ResNet showing peak participation from Client 2, while Google Net exhibits more consistent but moderate contributions. This variation reflects the impact of non-IID data distribution on local model updates in distributed healthcare systems.

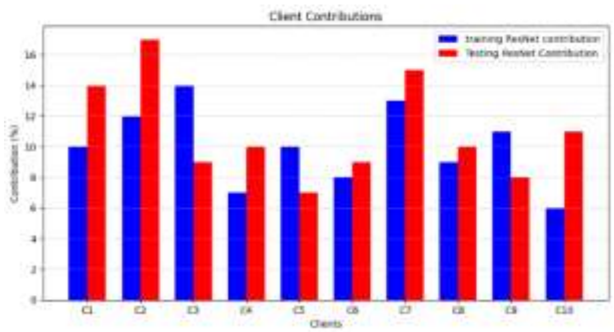

Fig. 12. Bar Chart Client Contribution in the Federated Setup with ResNet Model

Figure 12 shows the client-wise contribution percentages for training and testing ResNet models in a federated learning setup. Each bar represents the model update weight contributed by individual clients, showing noticeable variability due to non-IID data distribution that ResNet consistently achieves higher or comparable contributions, indicating more stable learning across clients in distributed healthcare environments.

The proposed approach utilizing ResNet deep learning architecture demonstrates significant advantages in detecting TB by Chest X-ray images with exceptional accuracy, precision, and recall metrics. Its ability to effectively minimize FP and FN ensures reliable diagnostic outcomes and reduced misdiagnosis risks, which is critical in clinical healthcare environments. Compared to traditional diagnostic methods, deep learning models such as ResNet exhibit superior performance through robust feature extraction, pattern recognition in medical imaging, and handling of diverse radiological presentations. These can be proved by the high ROC AUC scores, which is 0.97, indicative of excellent discriminatory power to detect TB. The innovation of the current study is that the study performs a thorough assessment of ResNet architecture on real data of TB chest X-ray and presents performance results on chest X-rays together with in-depth confusion matrices and ROC curve analysis, which are essential to prove the clinical effectiveness of this model. Moreover, the hybridization of big data and federated learning models focuses on scalable and privacy-respecting design, which challenges such as the distribution of data between different hospitals and the confidentiality of patients are difficult to address. The research validates the effectiveness of DL approaches in medical imaging and shows the practical implications of implementing advanced AI structures to make healthcare diagnoses with better precision and performance efficiency.

4.1 Discussion

Table III shows the comparative performance analysis of DL models within a FL framework of TB X-ray classification on the chest. The proposed ResNet-based model performed the best by an accuracy of 96.7, precision of 96.8, recall of 98.0 and F1 score of 97.4 within a privacy-preserving and distributed health situation. Comparatively, similar but slightly lower competitive accuracy of 94 percent accuracy 96 percent precision, 95 percent recall, and 96 percent F1score were recorded using Dense Net. Squeeze Net yielded rather humble results of 94.18% accuracy, 94.31% precision, 94.18% recall and 94.17% F1score, albeit being more lightweight. Those models were tested in a federated learning system where the training process is distributed among various healthcare nodes, which perform heavy tasks on medical imaging data. The proposed ResNet model has not only shown a high level of classification ability but also showed a significantly high level of scalability and compliance with privacy regulations, thus a model well suited to a reality-guided deployable in large data-driven distributed healthcare systems. This points to the practical good of combining deep convolutional models such as ResNet with federated learning to have high diagnostic accuracy and safe collaboration across organizations.

Table 3: Comparison Between Proposed Model And Existing Models For Federated Learning In Healthcare

Measure	ResNet	Dense net[27]	Squeeze Net[28]
Accuracy	96.7	94	94.18
precision	96.8	96	94.31
Recall	98.0	95	94.18
F1- score	97.4	96	94.17

The efficiency of the proposed ResNet-type deep learning that is introduced in a federated learning environment in the TB chest X-ray classification method proves to be more effective within a distributed healthcare setup. The accuracy of ResNet is the highest compared to other models; it surpasses Dense Net and Squeeze Net, that are utilized to form comparative baselines. The implementation of modern convolutional architectures in privacy-preserving settings helps the framework to capture complex trends in medical imaging data, thus maintaining confidentiality of the patient. The federated architecture enables decentral nodes to train the model without exchanging raw data according to privacy issues and regulatory limits in clinical settings. The robustness of ResNet is indicated by its excellent capability of processing high-resolution image features as well as heterogeneous client data. Nonetheless, it has its problems, such as handling non-IID data distributions and low-latency communication to support real-time diagnosis. Nevertheless, the proposed framework continues to present a scalable, secure and accurate solution to TB detection in multi-institutional healthcare networks in the broader context of utilizing federated learning to perform privacy-preserving big data analytics in the medical setting.

5. Conclusion and Future Work

Federated learning has already established itself as a unique learning platform because it enables edge devices to learn the model locally and use their local data to train the model at the local site. a federated learning-based solution to privacy-preserving TB chest X-ray classification in a distributed healthcare scenario. Federated learning ensures that sensitive personal patient data is distributed at the hospital or nodes, thus staying under privacy regulations, such as HIPAA. Pre-processing pipelines incorporate necessary image resizing to 224x224 pixels, normalization and SMOTE to counter class imbalance between distributed datasets that are not IID. ResNet was chosen due to its impressive feature extraction components and was trained coordinated among various clients. Its benchmarking was compared to Dense Net and Squeeze Net on the same federated configuration. The proposed ResNet model achieved the highest accuracy of 96.7%, demonstrating its robustness and effectiveness in medical image classification tasks. In contrast, Dense Net and Squeeze Net achieved accuracies of 94% and 94.18%, respectively. These results validate the superior performance of ResNet in the federated learning environment. For future work, the framework can be extended by incorporating secure aggregation techniques and differential privacy mechanisms to further enhance privacy during model updates. Additionally, the deployment of this system on real hospital networks with live patient data can be explored to validate its effectiveness in practical healthcare scenarios. The integration of multimodal data, CT scans, EHRs and optimization for low-resource edge devices are also promising directions to improve diagnostic accuracy and accessibility.

Funding: This research received no external funding

Conflicts of Interest: The author declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers

References

[1] Akter, M.S and Juthi, S. (2025). Big Data Analytics In Healthcare: Tools, Techniques, And Applications - A Systematic Review," *Innov. Eng. J*, doi: 10.70937/jnes.v2i01.51.
[2] Butt M. (2023). A Fog-Based Privacy-Preserving Federated Learning System for Smart Healthcare Applications, doi: 10.3390/electronics12194074.
[3] Dhiman G. (2022). Federated Learning Approach to Protect Healthcare Data over Big Data Scenario, *Sustainability*, doi: 10.3390/su14052500.
[4] Hasan, K. (2024). Federated Learning for Privacy-Preserving AI in Healthcare Systems, *Int. J. Sustain. Dev. Comput. Sci. Eng*.
[5] Hegde, M.G and Venugopal, K. R. (2023). A Comparative Study of Neural Network and Machine Learning on Privacy Preserving Federated Learning for Healthcare Applications, in *2023 IEEE Technology & Engineering Management Conference - Asia Pacific (TEMSCON-ASPAC)*, IEEE, Dec. 1–6. doi: 10.1109/TEMSCON-ASPAC59527.2023.10531360.
[6] Kareem, A., Liu H and Velisavljevic, V. (2023). A federated learning framework for pneumonia image detection using distributed data, *Healthc. Anal*, doi: 10.1016/j.health.2023.100204.
[7] Modalavalasa, G. (2023). Advanced Blockchain Mechanisms for Strengthening Data Security And Ensuring Privacy in Decentralized Systems, *Int. J. Recent Technol. Sci. Manag*. 89–98.
[8] Mohanty, R and Tripathy, P. K. (2024). IFLNLP: An Improved Federated Learning and Natural Language Processing-based Smart Healthcare System, in *2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/IC-SIT63503.2024.10862165.

[9] NTNRS. (2024). Privacy-preserving federated learning in healthcare: Fundamentals, state of the art and prospective research directions," in *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, IEEE, Jul., pp. 1438–1443. doi: 10.1109/AIC61668.2024.10730929.

[10] Oh W and Nadkarni, G. N. (2023). Federated Learning in Health care Using Structured Medical Data, *Adv. Kidney Dis. Heal*, doi: 10.1053/j.akdh.2022.11.007.

[11] Pandya, S. (2024). Integrating Smart IoT and AI-Enhanced Systems for Predictive Diagnostics Disease in Healthcare, *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, doi: 10.32628/CSEIT2410612406.

[12] Prajapati, N.K. (2025). Federated Learning for Privacy-Preserving Cybersecurity: A Review on Secure Threat Detection, *Int. J. Adv. Res. Sci. Commun. Technol*. 520–528, Apr. 2025, doi: 10.48175/IJARSCT-25168.

[13] Prajapati, V. (2025). Enhancing Threat Intelligence and Cyber Defense through Big Data Analytics: A Review Study, *J. Glob. Res. Math. Arch*. 1–6, 2025.

[14] Pandya, S. (2024). Predictive Modeling for Cancer Detection Based on Machine Learning Algorithms and AI in the Healthcare Sector," *TIJER - Int. Res. J*. 549–555,

[15] Prajapati, N. (2025). The Role of Machine Learning in Big Data Analytics: Tools, Techniques, and Applications, *ESP J. Eng. Technol. Adv*, doi: 10.56472/25832646/JETA-V5I2P103.

[16] Raja, R., Mukherjee I and Sarkar, B. K. (2020). A Systematic Review of Healthcare Big Data, *Sci. Program*, doi: 10.1155/2020/5471849.

[17] Rampone, G., Ivaniv T and Rampone, S. (2025). A Hybrid Federated Learning Framework for Privacy-Preserving Near-Real-Time Intrusion Detection in IoT Environments, *Electronics*, doi: 10.3390/electronics14071430.

[18] Rahman T. (2020). Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization, *IEEE Access*. 191586–191601, doi: 10.1109/ACCESS.2020.3031384.

[19] Sheeba A and Anbu, S. (2020). Health Care System in India: An Overview," in *International issues on Health Economics and Management*. 215–218.

[20] Song, C and Yang, N. (2024). Secure and Efficient Federated Learning Schemes for Healthcare Systems, *Electronics*, doi: 10.3390/electronics13132620.

[21] Singh, G and Fisichella, M. (2023). Federated Learning to Safeguard Patients' Data: A Medical Image Retrieval Case," *Big Data Cogn. Comput.*, doi: 10.3390/bdcc7010018.

[22] Teo Z.L. (2024). Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture," *Cell Reports Med.*, doi: 10.1016/j.xcrm.2024.101419.

[23] Thokala, V.S. (2024). Improving Data Security and Privacy in Web Applications: A Study of Serverless Architecture, *Tech. Int. J. Eng. Res*. 74–82

[24] Tan, T., Wang, J and Tan, Z. (2023). An Optimized Federated Learning Approach with the Data-Sharing Function to the Analysis of Cardiothoracic Time-Series Signals," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Jul. 2023, pp. 1–6. doi: 10.1109/EMBC40787.2023.10340752.

[25] Tiwari, R. G and Trivedi, N.K. (2023). Enhancing Skin Disease Classification and Privacy Preservation through Federated Learning-Based Deep Learning, in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, IEEE, Dec. 2023, pp. 1–7. doi: 10.1109/ICAIIHI57871.2023.10489335.

[26] Wang L and Alexander, C.A. (2019). Big Data Analytics in Healthcare Systems, *Int. J. Math. Eng. Manag. Sci*, doi: 10.33889/IJMEMS.2019.4.1-002.

[27] Yang, X., Ye, Q and Cai, G. (2022). PD-ResNet for Classification of Parkinson's Disease From Gait," *IEEE J. Transl. Eng. Heal. Med*. 1–11, doi: 10.1109/JTEHM.2022.3180933.

[28] Zhou, H and Xiang, Y. (2025). Robust Federated Learning for Privacy Preservation and Efficiency in Edge Computing, *IEEE Trans. Serv. Comput*, doi: 10.1109/TSC.2025.3562359.

[29] Joy, M. S. A. ., Alam, G. T. ., & Bakhsh, M. M. . (2024). Transforming QA Efficiency: Leveraging Predictive Analytics to Minimize Costs in Business-Critical Software Testing for the US Market. Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 7(01), 77–89. https://doi.org/10.60087/jaigs.v7i01.297

[30] Bakhsh, M. M. ., Joy, M. S. A. ., & Alam, G. T. . (2024). Revolutionizing BA-QA Team Dynamics: AI-Driven Collaboration Platforms for Accelerated Software Quality in the US Market. Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 7(01), 63–76. https://doi.org/10.60087/jaigs.v7i01.296

[31] Bakhsh, M. M., Alam, G. T. ., & Nadia, N. Y. . (2025). Adapting Agile Methodologies to Incorporate Digital Twins in Sprint Planning, Backlog Refinement, and QA Validation. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 4(2), 67-79. https://doi.org/10.60087/jklst.v4.n2.006

[32] Gazi Touhidul Alam; Md Ismail Jobiullah; Arannita Saha Suspee; Mohammed Majid Bakhsh; Abu Saleh Muhammad Saimon; Syed Mohammed Muhive Uddin. "Creating a Knowledge Hub: AI-Powered Learning Management Systems for BA-QA Training." Volume. 10 Issue.4, April-2025 International Journal of Innovative Science and Research Technology (IJISRT), 3111-3118, https://doi.org/10.38124/ijisrt/25apr1081

[33] Alam, G. T., Bakhsh, M. M. ., Nadia, N. Y. ., & Islam, S. A. M. . (2025). Predictive Analytics in QA Automation:: Redefining Defect Prevention for US Enterprises. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 4(2), 55-66. https://doi.org/10.60087/jklst.v4.n2.005

[34] DATA ANALYTICS AND ITS APPLICATIONS. (2025). Journal of Medical & Health Sciences Review, 2(2). https://doi.org/10.62019/4k9hag93

[35] BIG DATA ANALYTICS: APPLICATIONS, PROSPECTS AND CHALLENGES. (2025). Journal of Medical & Health Sciences Review, 2(2). https://doi.org/10.62019/a0kkys98