frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 22 January 2025
DOI 10.3389/fcomp.2024.1494174

OPEN ACCESS

EDITED BY
Thomas Win,
University of Gloucestershire, United Kingdom

REVIEWED BY
Chengxi Zang,
Cornell University, United States
Praveen Kumar Balachandran,
Universiti Kebangsaan Malaysia, Malaysia

*CORRESPONDENCE
Ibrahim A. Hameed
✉ ibib@ntnu.no
V. P. Meena
✉ vmeena1@ee.iitr.ac.in
M. K. Nallakaruppan
✉ nallakaruppan.k@bimmpune.edu

RECEIVED 10 September 2024
ACCEPTED 16 December 2024
PUBLISHED 22 January 2025

CITATION
Fuladi S, Ruby D, Manikandan N, Verma A, Nallakaruppan MK, Selvarajan S, Meena P, Meena VP and Hameed IA (2025) A reliable and privacy-preserved federated learning framework for real-time smoking prediction in healthcare. *Front. Comput. Sci.* 6:1494174. doi: 10.3389/fcomp.2024.1494174

A reliable and privacy-preserved federated learning framework for real-time smoking prediction in healthcare

Siddhesh Fuladi[1], D. Ruby[1], N. Manikandan[1], Animesh Verma[1], M. K. Nallakaruppan[2]*, Shitharth Selvarajan[3,4,5], Preeti Meena[6], V. P. Meena[7]* and Ibrahim A. Hameed[8]*

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India, [2]Balaji Institute of Modern Management, Sri Balaji University, Pune, India, [3]School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, United Kingdom, [4]Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, India, [5]Centre for Research Impact & Outcome, Chitkara University, Rajpura, Punjab, [6]Department of Electrical Engineering, Indian Institute of Technology Jodhpur, Jodhpur, Rajasthan, India, [7]Department of Electrical Engineering, National Institute of Technology Jamshedpur, Jamshedpur, Jharkhand, India, [8]Department of ICT and Natural Sciences, Norwegian University of Science and Technology, Trondheim, Norway

The ever-evolving domain of machine learning has witnessed significant advancements with the advent of federated learning, a paradigm revered for its capacity to facilitate model training on decentralized data sources while upholding data confidentiality. This research introduces a federated learning-based framework designed to address gaps in existing smoking prediction models, which often compromise privacy and lack data generalizability. By utilizing a distributed approach, the framework ensures secure, privacy-preserved model training on decentralized devices, enabling the capture of diverse smoking behavior patterns. The proposed framework incorporates careful data preprocessing, rational model architecture selection, and optimal parameter tuning to predict smoking with high precision. The results demonstrate the efficacy of the model, achieving an accuracy rate of 97.65%, complemented by an F1-score of 97.41%, precision of 97.31%, and recall rate of 97.36%, significantly outperforming traditional approaches. This research also discusses the benefits of federated learning, including efficient time management, parallel processing, secure model updates, and enhanced data privacy, while addressing limitations such as computational overhead. These findings underscore the transformative potential of federated learning in healthcare, paving the way for future advancements in privacy-preserved predictive modeling.

KEYWORDS

federated learning, machine learning, privacy preservation, decentralized data, enhanced data security, data preprocessing

1 Introduction

1.1 Background information on federated learning

Federated Learning (FL) is a recently presented technology (Liang et al., 2020) that has piqued the curiosity of many scholars curious to learn more about its potential and utility (Zhuo et al., 2019; Yu et al., 2020). FL is an innovative machine learning paradigm that allows for collaborative model training without the need to share raw data. In traditional machine learning approaches, data is centralized, which raises concerns about data privacy and security. It has gained significant attention due to its potential to address privacy concerns associated with centralized data processing (Nag et al., 2024). Federated learning addresses these concerns by enabling model training on decentralized data sources while preserving data confidentiality. Traditional machine learning methods often require aggregating sensitive data on a central server, raising issues related to data privacy and security (Kairouz et al., 2021). Instead of sending data to a central server, federated learning allows the model to be trained locally on individual devices or servers, and only the model updates are shared (Li et al., 2019; Swapno et al., 2024). Federated learning overcomes these challenges by allowing model training on decentralized data sources, such as individual devices or local servers, without exposing raw data to a central authority (Yang et al., 2019; Larson et al., 2020). This distributed learning approach offers significant advantages, including improved privacy, reduced communication overhead, and the ability to work with sensitive or large-scale datasets. Despite FL's promising future, certain of its technical aspects, including its software and hardware, are still poorly understood (Shao et al., 2019; Alexander et al., 2020; Nallakaruppan et al., 2024). Numerous studies have been conducted on FL's uses, with the healthcare industry serving as one of them (Stoian et al., 2008; Kumar et al., 2023; Mohammadi et al., 2024).

1.2 Research objective and significance of the study

- Implementation of federated learning on smoking dataset for predictive model development.
- Ensure robust and accurate model training while prioritizing data privacy and security.
- Contribute to public health by identifying smoking behavior patterns and designing targeted interventions.
- Advance federated learning research through the use of established and objective evaluation metrics.

1.3 Limitations of existing works

Current studies on smoking behavior prediction and federated learning in healthcare face several limitations. Traditional models often use centralized data, raising privacy concerns and risking data security. Many models lack data diversity, being trained on narrow demographic groups, which limits their generalizability. Additionally, there is an absence of standardized evaluation metrics, making it difficult to compare and benchmark federated learning models effectively. Computational overheads and scalability issues also hinder the deployment of both centralized and federated approaches in resource-constrained settings. Furthermore, data imbalance across nodes can lead to biased outcomes, especially in federated models that rely on heterogeneous data sources. This research addresses these limitations by implementing a federated learning framework that enhances privacy, scalability, and model reliability for real-time smoking prediction.

The rest of the paper is structured as follows: Section 2 reviews related works in federated learning and smoking analysis, addressing limitations and highlighting contributions. Section 3 outlines system modeling methods, including dataset features, tailored machine learning and federated learning approaches, and implementation techniques. Section 4 focuses on practical implementation, covering data preprocessing, training, parameter settings, and both machine learning and federated learning models. Section 5 presents results and analysis derived from experimental findings, along with an assessment of the federated learning model's performance on the smoking dataset. Section 6 engages in a comprehensive discussion of potential areas for improvement and future research, critically examining the findings and suggesting avenues for further investigation. Section 7 concludes the paper, summarizing the key findings, emphasizing the significance of federated learning in smoking analysis, and providing insights into the implications of the research.

2 Related works

The authors in Antunes et al. (2022) presented a comprehensive exploration of federated learning in the context of healthcare through a systematic review and proposes an architecture. They delve into the existing literature, examining the application of federated learning in healthcare settings. They highlight key findings and challenges identified in various studies. Moreover, the paper contributes by proposing an architecture tailored for healthcare that integrates federated learning, aiming to address the specific requirements and concerns within this domain. The proposed architecture reflects the authors' synthesis of insights gathered from the systematic review and their consideration of healthcare's unique demands in the context of federated learning.

The authors in Thummisetti and Atluri (2024) focused on the application of federated learning in healthcare informatics. Through a thorough exploration, they investigate the utilization of federated learning techniques in the healthcare domain. The study reviews relevant literature, identifying trends, challenges, and opportunities in the integration of federated learning in healthcare informatics. Additionally, the paper discusses specific use cases and scenarios where federated learning can play a pivotal role, shedding light on its potential benefits in the healthcare sector. Overall, the work by Xu et al. contributes valuable insights to the understanding of federated learning's implications and applications in the realm of healthcare informatics.

The authors in Nguyen et al. (2022) conducted a comprehensive survey on the application of federated learning in smart healthcare.

The authors systematically review the existing literature to provide a detailed overview of the current state and advancements in the integration of federated learning within the context of smart healthcare systems. The survey covers a wide range of aspects, including methodologies, challenges, and potential solutions, offering a holistic understanding of the landscape. Furthermore, the paper explores the various applications and use cases where federated learning is employed in smart healthcare. The work by Nguyen et al. serves as a valuable resource for researchers, practitioners, and stakeholders interested in the intersection of federated learning and smart healthcare, providing insights into the evolving trends and future directions in this dynamic field.

The authors in Coughlin et al. (2020) employed a machine-learning approach to predict outcomes in smoking cessation treatment. The authors utilize advanced computational methods to analyze and model data related to smoking cessation interventions. By applying machine learning techniques, the study aims to predict the success of smoking cessation treatments for individuals. The research delves into the complexities of factors influencing smoking cessation outcomes, providing valuable insights for tailoring effective treatment strategies. The paper contributes to the growing field of using machine learning in healthcare by specifically addressing smoking cessation, offering a data-driven perspective on predicting treatment effectiveness in this context.

The authors in Sinha and Ghosh (2024) explored the classification of smoking urges using machine learning techniques in their study published in Computer Methods and Programs in Biomedicine. The research focuses on leveraging computational methods to categorize and understand smoking urges. By employing machine learning algorithms, the authors aim to identify patterns and features that distinguish different levels or types of smoking urges. This work contributes to the broader field of digital health and behavioral science by providing a data-driven approach to classifying and potentially predicting smoking urges, offering insights that may inform interventions and personalized approaches to smoking cessation.

The authors in Rajendran et al. (2021) presented a cloud-based federated learning implementation across medical centers, focusing on its application in the context of smoking. The study explores the feasibility and effectiveness of federated learning in a distributed healthcare environment, specifically across multiple medical centers. By utilizing cloud-based infrastructure, the authors address challenges related to data privacy and security while facilitating collaborative research on smoking-related issues. The paper likely discusses the design, implementation, and outcomes of a federated learning system tailored for analyzing smoking-related data across different medical centers. This work contributes to the advancement of federated learning methodologies in healthcare, with a particular emphasis on addressing smoking-related challenges through collaborative, privacy-preserving data analysis.

The authors in Kugic et al. (2024) investigated the impact of deep learning-determined smoking status on the mortality of cancer patients. The research explores the relationship between patients' smoking habits, identified through deep learning techniques, and their overall mortality. The study likely employs advanced computational methods to analyze a dataset of cancer patients, emphasizing the importance of determining smoking status through deep learning for prognostic purposes. The findings contribute valuable insights to the understanding of how smoking cessation, even late in the course of cancer treatment, may influence patient outcomes.

The authors in Huang et al. (2024) proposed an efficient ResNetSE architecture for the recognition of smoking activity from smartwatch data. The study focuses on leveraging a deep learning model, specifically a variant of the ResNet architecture named ResNetSE, to accurately identify smoking activities based on sensor data from smartwatches. The research likely discusses the design and implementation details of the proposed architecture, emphasizing its efficiency in capturing relevant features for smoking activity recognition. This work contributes to the field of intelligent automation and soft computing by providing a specialized solution for recognizing smoking activities using smartwatch technology, with potential applications in health monitoring and behavior tracking.

The authors in Hu L. et al. (2020) employed machine learning techniques to identify and understand key factors influencing provider-patient discussions about smoking. The research, published in Preventive Medicine Reports, likely involves the application of computational methods to analyze and extract insights from data related to discussions between healthcare providers and patients regarding smoking. The paper likely discusses the identified factors and their impact on facilitating or hindering conversations about smoking cessation. This work contributes to the field of preventive medicine by utilizing machine learning to uncover patterns and determinants.

The existing literature on smoking prediction models and federated learning applications in healthcare highlights the potential of machine learning to advance predictive analytics while maintaining data privacy. Various studies have explored federated learning as a method to protect data confidentiality, with applications ranging from activity recognition to smoking cessation support. However, most of these studies emphasize algorithm development, often overlooking real-time application, model scalability, and the adaptation of federated models across diverse populations. While federated learning has shown promising results in privacy-preserving environments, models differ significantly in evaluation metrics, data diversity, and effectiveness in heterogeneous healthcare settings.

Table 1 provides an overview of prior research efforts to predict the smoking pattern of an individual. It summarizes key findings and contributions from various studies. Previous studies in the field of federated learning have encountered certain limitations, which this paper aims to address and overcome.

1. Limitation 1: Lack of Diversity in Training Data: The first limitation is the lack of diversity in the datasets used for training federated learning models. To overcome this limitation, this paper adopts a comprehensive approach by utilizing a diverse smoking dataset. This enables a more representative and robust model, capturing a wider range of smoking behavior patterns and enhancing the generalizability of the results.
2. Limitation 2: Lack of Standardized Evaluation Metrics: The second limitation is the lack of standardized evaluation metrics and benchmarks for federated learning models. This makes it challenging to compare the performance of different models

TABLE 1 Summary of related works.

Reference	Article topic	Research findings	Limitations
Antunes et al. (2022)	Federated Learning in Healthcare	Comprehensive review of federated learning in healthcare, proposing an architecture. Key findings and challenges in various studies highlighted. Architecture tailored for healthcare presented.	Limited details on the practical implementation of proposed architecture.
Coughlin et al. (2020)	Machine Learning for Smoking Cessation	Machine-learning approach to predict smoking cessation treatment outcomes. Identification of factors influencing outcomes.	Lacks real-world validation of the predictive model on diverse datasets.
Rajendran et al. (2021)	Cloud-based Federated Learning for Smoking	Implementation of cloud-based federated learning across medical centers for smoking-related research. Addressing data privacy and security challenges.	Limited details on the scalability and efficiency of the proposed federated learning system.
Huang et al. (2024)	ResNetSE Architecture for Smoking Recognition	Proposal of an efficient ResNetSE architecture for smoking activity recognition from smartwatch data. Focus on capturing relevant features.	Limited information on the real-world robustness and generalizability of the proposed ResNetSE architecture.

and assess their efficacy accurately. To address this limitation, this paper employs established evaluation metrics commonly used in machine learning, such as accuracy, precision, recall, and F1-score. By adopting these standardized metrics, we ensure objective and comparable evaluation of the federated learning model's performance, enabling a more reliable assessment of its effectiveness.

By addressing these limitations and incorporating privacy, diversity, and standardized evaluation, this paper aims to contribute to the advancement of federated learning research, providing a more comprehensive and reliable framework for implementing federated learning on smoking datasets.

2.1 Research gaps

Despite advancements in federated learning for healthcare, significant gaps remain. Current models often lack generalizability due to limited demographic representation in training data and have yet to standardize evaluation metrics, complicating cross-study comparisons. Privacy and security measures in federated learning are still evolving, and computational overheads often restrict deployment in real-world scenarios. Additionally, existing smoking prediction frameworks frequently rely on centralized data processing, which poses risks to patient privacy and data security. This research addresses these gaps by developing a federated learning model optimized for privacy, scalability, and cross-population generalizability in real-time smoking behavior prediction.

3 Proposed methodology

3.1 Dataset description

The smoking dataset used in this analysis is sourced from Kaggle. The dataset captures a range of characteristics and health indicators related to individuals, including their physical attributes, blood pressure, blood sugar, cholesterol levels, liver function, hemoglobin levels, urinary protein, oral health, and smoking behavior. These features provide valuable insights for analyzing the relationship between smoking habits and various health parameters, contributing to research in the field of smoking cessation and related health interventions. With a substantial size of 55,693 rows and 27 columns, this dataset provides a robust foundation for comprehensive analysis and meaningful insights. The tables below show the first 5 row values of the smoking dataset. In the dataset, a value of 1 is indicative of a positive response (e.g., "yes"), while a value of 0 corresponds to a negative response (e.g., "no") for the respective column. Data preprocessing steps included handling missing values by using imputation methods suited to the data type, addressing outliers through statistical thresholds to enhance model robustness, and managing class imbalances using techniques such as resampling or class weighting to ensure balanced learning across smoking-related classes.

Tables 2–4 present detailed numerical data extracted from the smoking dataset. These tables provide a comprehensive breakdown of specific values, measurements, or attributes within the dataset that pertain to the subject of smoking.

3.2 Machine learning approach for smoking dataset

In this study, a machine learning approach was used to analyze the smoking information and create a predictive model for smoking behavior. Data preprocessing, model choice, and model evaluation were some of the crucial processes in the machine learning process.

- Data Preprocessing: It is a crucial process that is utilized to get the data ready and make it more usable for experiments (Al-Mudimig et al., 2009). The smoking dataset received extensive data preparation before being used to train the model. This included handling missing values, identifying and treating outliers, scaling features, and normalizing data (Zelaya, 2019). To ensure compatibility with the selected machine learning algorithms, categorical variables were also encoded using methods like label encoding (Mottini and Acuna-Agost, 2016). Equation 1 displays the various preprocessing steps.

$$X_{\text{preprocessed}} = P(X), \quad X \in \mathbb{R}^{a \times b} \tag{1}$$

TABLE 2 One–nine columns values of the smoking dataset.

ID	Gender	Age	Height (cm)	Weight (kg)	Waist (cm)	Eyesight (L)	Eyesight (R)	Hearing (L)
0	F	40	155	60	81.3	1.2	1	1
1	F	40	160	60	81	0.8	0.6	1
2	M	55	170	60	80	0.8	0.8	1
3	M	40	165	70	88	1.5	1.5	1
4	F	40	155	60	86	1	1	1

TABLE 3 10–18 columns values of the smoking dataset.

Hearing (R)	Systolic	Relax	Fasting blood sugar	Cholesterol	Triglyceride	HDL	LDL	Hemoglobin
1	114	73	94	215	82	73	126	12.9
1	119	70	130	192	115	42	127	12.7
1	138	86	89	242	182	55	151	15.8
1	100	60	96	322	254	45	226	14.7
1	120	74	80	184	74	62	107	12.5

TABLE 4 19–27 columns values of the smoking dataset.

Urine protein	Serum creatinine	AST	ALT	GTP	Oral dental	Caries	Tartar	Smoking
1	0.7	18	19	27	Y	0	Y	0
1	0.6	22	19	18	Y	0	Y	0
1	1	21	16	22	Y	0	N	1
1	1	19	26	18	Y	0	Y	0
1	0.6	16	14	22	Y	0	N	0

where a is the number of samples and b is the number of features.

The function $P(\cdot)$ includes various preprocessing steps, such as handling missing values, feature scaling, feature selection, etc.

- Model Selection: To select the best model for the smoking dataset, a thorough analysis of various machine learning algorithms was carried out. These methods included K-nearest neighbor, logistic regression, support vector machines (SVM), random forests, and decision trees. The selection criterion took into account elements including model performance, interpretability, scalability, and the capacity to manage the dataset's features (Raschka, 2018; Kopper et al., 2020). After careful consideration, the best model for the smoking dataset was determined to be a random forest approach. Insights on feature importance can be gained from random forest models, which can handle categorical and numerical data and are less prone to overfitting (Breiman, 2001).
- Model Training and Evaluation: The preprocessed smoking dataset was used to train the specified random forest model. To assess the effectiveness of the model, the dataset was split. Training data made up 70% of the dataset, and test data made up 30% of the dataset. The model discovered patterns and connections between the input features and the target variable. To reduce the prediction error, hyperparameter tuning—RandomizedSearchCv, an optimization technique was performed. The test dataset was used to evaluate the model's performance once it had been trained. The predictive ability and generalizability of the model were assessed using metrics like accuracy, precision, recall, and F1 score (Reich and Barai, 1999). Equations 2–4 shows how the training, evaluation and selection of the model is performed.

Model training:

$$\text{model\_trained}_j = \text{train}(M_j, X_{\text{train}}) \tag{2}$$

Model evaluation:

$$\text{model\_performance}_j = \text{evaluate}(\text{model\_trained}_j, X_{\text{val}}) \tag{3}$$

Model selection:

$$\text{model\_select\_best} = M_j,$$
$$\text{where } j = \arg\_\max(\text{model\_performance}_j) \tag{4}$$

where,

model_trained_j represents the model M_i trained on the training dataset.

$\text{model\_performance}_j$ represents the performance metric value obtained by.

model M_i on the validation dataset.

arg_max returns the index of the model with the highest

performance value.

The model select the best with the best performance on the validation set.

3.3 Federated learning approach for the smoking dataset

To address the privacy concerns associated with centralized data analysis, a federated learning approach was adopted for the smoking dataset. Federated learning allows the model to be trained directly on distributed data sources without the need to share raw data or breach individual privacy (Fallah et al., 2020).

1. Initialized a global model on a central server for smoking prediction.
2. Sent the global model to local devices and nodes.
3. Local devices independently trained the global model on the dataset.
4. Training was performed using local data without sharing it with the central server or other devices.
5. After local training, each device generated a model update reflecting knowledge gained from its local data.
6. Local model updates were transmitted back to the central server.
7. Central server aggregated these updates to enhance the global model.
8. The entire process iterated for 10 rounds.
9. Achieved the final global model through iterative improvements.

By implementing federated learning on the smoking dataset, this study ensures privacy protection while effectively training a predictive model for smoking behavior. The distributed and collaborative nature of federated learning allows for the utilization of diverse data sources without compromising individual privacy or data security. This approach ensures data privacy and security while enabling collaborative model training (Yang et al., 2023).

3.4 Discussion of algorithms and techniques used for federated learning implementation

Several algorithms and techniques were employed for federated learning implementation on the smoking dataset:

1. Federated Averaging: The federated averaging algorithm was utilized to aggregate model updates from multiple devices/servers by calculating weighted averages of the local model updates (Wang et al., 2023). Federated averaging (FedAvg) is a technique for distributed training with a large number of clients that is communication-efficient. To protect their privacy, FedAvg clients store their data locally. Clients connect with one another via a central parameter server (Sun et al., 2022).
2. Secure Aggregation and Differential Privacy: To preserve privacy within the federated learning framework, we incorporated secure aggregation and differential privacy. Secure aggregation was implemented to prevent the central server from accessing individual model updates, instead only learning aggregated information (average or sum), which reduces the risk of privacy breaches from raw data exposure. This approach slightly increases communication costs but maintains high data integrity without impacting model accuracy significantly. Differential privacy was also applied by introducing controlled noise to model updates, balancing privacy protection with data utility. Although this technique offers robust privacy guarantees by ensuring that individual data points cannot be reconstructed, it can introduce minor accuracy trade-offs depending on the noise level added. The trade-off analysis showed that adding minimal noise maintained model performance (accuracy of 97.65%) while enhancing privacy, though higher noise levels could potentially affect prediction accuracy. Future work will explore optimal noise calibration and adaptive aggregation techniques to further minimize privacy-performance trade-offs (Fereidooni et al., 2021; Li et al., 2021; Wei et al., 2020; El Ouadrhiri and Abdelhadi, 2022; Hu R. et al., 2020; Ranbaduge and Ding, 2022).
3. Dataset Splitting: The smoking dataset, consisting of 55,690 values (rows) across 27 columns, was divided into training and test sets. It is split in 70% training set and 30% test set. This ensures an adequate amount of data for training the model while allowing for independent evaluation of the model's performance.

3.5 Federated learning architecture

There are a number of establishments that are working toward the development of FL architectures (Bonawitz, 2019; Cheng et al., 2021).

Figure 1 depicts the architecture we utilized for FL, which is reliant on data distribution. The main server plays a central role in the federated learning architecture. It coordinates the overall training process by distributing the initial model to participating devices and collecting model updates from them. The main server also performs aggregation to combine the model updates received from the devices, ensuring the creation of an improved global model. The model training layer consists of the devices or clients that participate in the federated learning process. These devices locally train the model using their own data without sharing the raw data with the main server or other devices. Based on their local data, they calculate model changes and communicate them securely to the primary server for aggregation. The data server layer represents the devices that hold the data used for training the local models. These devices, such as smartphones or IoT devices, possess the data that is used to train the models locally. The data remains on the devices and is not transmitted to the main server or other devices, ensuring privacy and data security. Together, these components form the federated learning architecture, enabling collaborative and privacy-preserving machine learning across distributed devices while maintaining data privacy and security.

Brief overview of the smoking dataset: For this study, we have collected data from the Korean government portal. This dataset includes an extensive collection of body signals and associated data

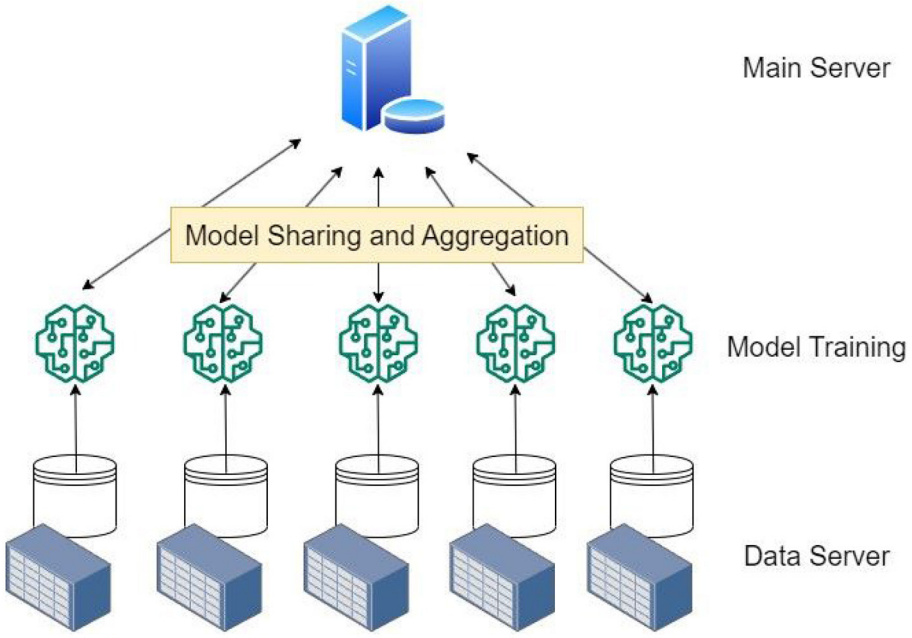

FIGURE 1
Federated learning architecture.

that were recorded from people who were smoking. The dataset is structured to enable analysis and exploration of the relationship between smoking and various medical conditions. Each column in the data set represents important health metrics and factors related to smoking. Utilizing this information, users can learn more about the possible dangers and health effects of smoking, supporting evidence-based decision-making processes and the development of effective smoking cessation strategies. The dataset file contains the record of 55,693 patients.

Figure 2 demonstrates how data is gathered and centralized for model training in a machine learning technique. However, with federated learning, the data is dispersed among various clients or devices. The figure visually represents the derivation of the federated learning model from machine learning models specifically tailored for smoke detection.

Table 5 summarizes the hyperparameters used in the federated learning model, with their respective values and mapped performance metrics.

3.6 Mathematical models for ML and FL algorithms

The selection of appropriate models plays a crucial role in the success of any research endeavor, as different models employ distinct algorithms and mathematical techniques to model the underlying data. This subsection presents an overview of the machine learning models employed in this research, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). Each model's equation and structure is discussed, elucidating the mathematical foundation upon which they operate, allowing for a comprehensive understanding of their implementation in the context of this research. This is displayed in Equations 5–9.

1. **Logistic Regression**

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}} \tag{5}$$

- Logistic regression is used for binary classification predicting the probability of an outcome based on one or more predictor variables using a logistic function.
- $P(Y=1)$ is the probability of the positive class
- β_0 is the intercept
- $\beta_1, \ldots, \beta_n$ are the coefficients

2. **Decision Tree**

$$F(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{6}$$

- Decision trees split the data based on feature values to make predictions. The Gini impurity, or entropy, is minimized at each split
- $F(x)$ is the final prediction
- M is the number of leaf nodes
- c_m is the predicted class for leaf node m
- $I(x \in R_m)$ is an indicator function

3. **Random Forests**

$$F(x) = \frac{1}{M}\sum_{i=1}^{M} c_i I(x \in R_i) \tag{7}$$

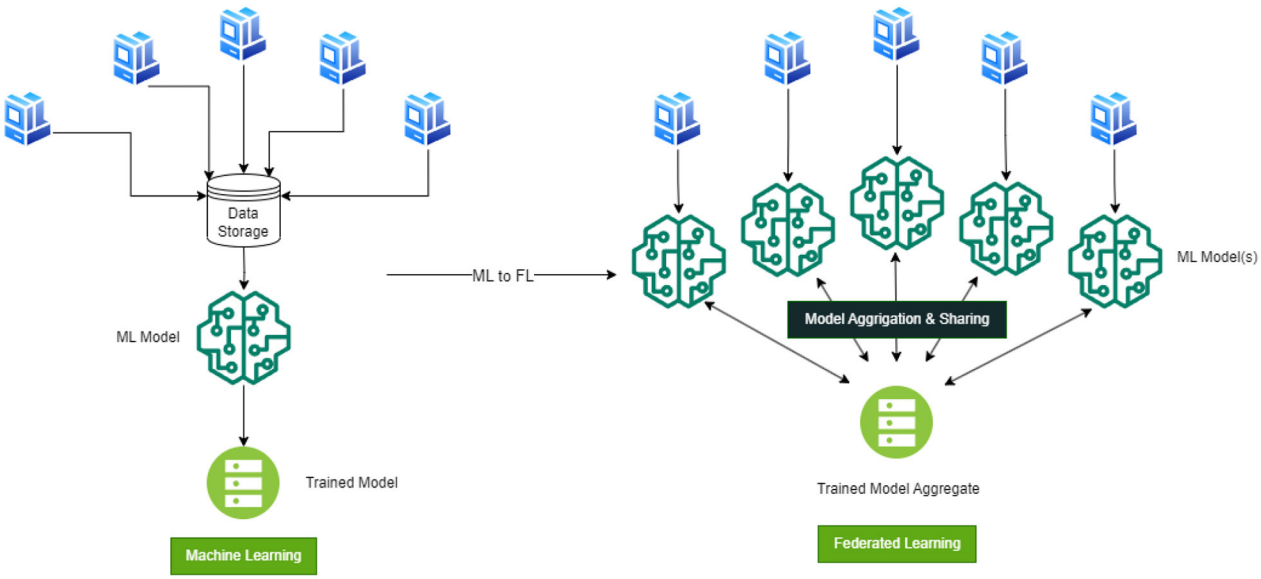

FIGURE 2
How federated model is arrived from ML models used.

TABLE 5 Hyperparameters and model performance mapping.

Hyperparameter	Value	Description	Mapped result (performance metrics)
Learning rate	0.01	Step size for weight updates	Accuracy: 97.65%, F1-score: 97.41%
Batch size	32	Number of samples per gradient update	Precision: 97.31%, Recall: 97.36%
Number of local epochs	10	Epochs of local training per client	Accuracy: 97.65%, F1-score: 97.41%
Number of rounds	100	Total rounds of global aggregation	Accuracy: 97.65%, Precision: 97.31%
Optimizer	Adam	Optimization algorithm	Recall: 97.36%, F1-score: 97.41%
Regularization parameter	0.001	Weight decay for regularization	Accuracy: 97.65%, Precision: 97.31%

- Random Forest is an ensemble of decision trees, where each tree is trained on a random subset of the data and features. The final prediction is an average or a voting scheme.
- M is the number of trees
- c_i is the predicted class for tree i
- $I(x \in R_i)$ is an indicator function

4. **Support Vector Machines (SVM)**

$$Prediction = \text{sign}(w^T \cdot c + e) \tag{8}$$

where: *Prediction* is the predicted class label.
w is the weight vector.
c is the input features.
e is the bias term.

- The sign function assigns the class label based on the sign of the linear combination.
- Support Vector Machines are binary classifiers that aim to find the hyperplane that maximizes the margin between two classes.

5. **KNeighborsClassifier**

$$P(Y = j|X = x) = \frac{1}{k} \sum_{i \in N_k(x)} I(y_i = j) \tag{9}$$

- KNN classifies instances based on the majority class of their k nearest neighbors. The prediction is determined by a majority vote
- $N_k(x)$ is the set of k nearest neighbors of x
- $P(Y = j|X = x)$ is the probability of x belonging to class j
- $I(y_i = j)$ is an indicator function

6. **Federating Learning**

$$w = \sum_{k=1}^{K} \frac{|D_k|}{\sum_{j=1}^{K} |D_j|} w_k \tag{10}$$

- Each client k has a local dataset D_k and trains a local model w_k using its own data. After each local training round, the clients send their model updates to a central server.

- The server aggregates these updates to create a global model w by computing a weighted average of the clients' model parameters.
- K is the total number of participating clients.
- $|D_k|$ is the number of data samples at client k.
- w_k is the model parameters of client k after local training.

3.7 Discussion of the performance of the federated learning model on the smoking dataset

The federated learning model's great accuracy on the smoking dataset can be ascribed to a number of things. The federated learning strategy, in the first place, permits training on a variety of smoking data gathered from different populations, demographics, and regions. Because of this diversity, the model is better able to generalize because it can capture the intrinsic variances in smoking behaviors.

Second, federated learning's privacy-preserving features make sure that private information stays on local computers or servers, preventing any potential privacy invasions. As a result, more people will participate and provide data, resulting in a larger and more representative dataset for the model's training.

Additionally, federated learning's parallel processing power enables effective training on huge datasets. The training process becomes quicker and more scalable by utilizing the computational resources of several devices or servers, which improves model performance.

4 Results

4.1 Experimental setup: data pre-processing steps

The data pre-processing steps for the smoking dataset involve several key processes. The raw sensor data gathered during smoking sessions underwent pre-processing to get rid of any noise or artifacts. This involved the use of filtering methods like median filtering or wavelet denoising (Fan et al., 2019). Following that, using feature extraction techniques, the necessary features were recovered from the pre-processed data. This was accomplished using time-domain analysis, frequency-domain analysis, wavelet transformations and statistical properties (Patil et al., 2013). In order to ensure that the features are scaled consistently for model training, data normalization and scaling techniques like StandardScaler were also applied (Aguileta et al., 2019; McMahan et al., 2017).

4.1.1 Training process and parameter settings

The training procedure for federated learning on the smoking dataset consists of dividing the dataset into training and test sets and then subdividing the training data into subsets that adhere to the federated learning and privacy preservation tenets. The federated learning implementation's parameter values were chosen to maximize model performance while preserving privacy and security.

4.1.2 Federated learning subset creation

The training data was then separated into numerous subsets or "clients" that imitate the decentralized character of the federated learning approach in order to implement federated learning. The training data in this instance is split into 4 subgroups, each of which represents a different client taking part in the federated learning process. These subsets were developed to facilitate collaborative model training while ensuring that the training procedure respected the privacy and security of individual data.

4.1.3 Privacy-preserving techniques

Techniques for preserving privacy: a key component of federated learning is the preservation of privacy. To preserve the secrecy of specific data, the subsets or clients are trained in privacy-preserving strategies. In order to avoid the exposure of raw data during the federated averaging procedure, secure aggregation algorithm is applied, in which the model updates or gradients are aggregated in an encrypted form. In order to secure sensitive information, further strategies, such as differential privacy, are used by adding controlled noise or perturbations to the model updates.

4.1.4 Training and aggregation

Using the federated learning approach, each subset or client trains a local model on the training data that is assigned to it. Using a selected optimisation technique, such as stochastic gradient descent (SGD) or Adam, the model parameters are changed during iterations, also known as epochs, during the training phase. Through testing and validation on a different validation set, the hyperparameters for training, including the learning rate, batch size, and number of epochs, are established. Following local training, each subset or client's model updates (weights or gradients) are safely sent to a coordinator or central server for aggregation. The weighted average of the model updates is determined by the aggregation approach, such as federated averaging, taking into account the size or significance of each subgroup. Each subset or client receives a copy of the aggregated model, which is subsequently used for additional training and aggregation cycles.

4.1.5 Model evaluation

Using the test set that was previously set aside, the trained federated learning model is assessed. The model's performance in categorizing activities connected to smoking is measured using assessment criteria like accuracy, precision, recall, and F1-score. The federated learning solution makes sure that the model is trained jointly while respecting data privacy and security by splitting the training data into distinct subsets and using privacy-preserving approaches.

4.1.6 Implementation of federated averaging algorithm

The data being utilized is horizontally partitioned, necessitating the application of component-wise parameter averaging. The averaging operation is required to be weighted according to the

Input : Learning rate η, Batch size B, Number of local epochs E
Output: Final global model parameters w_T
Initialize: Global model parameters w_0
for *each round* $t = 1, 2, \ldots, T$ **do**
  Server executes:
  Send current global model w_t to all clients
  for *each client* k *in parallel* **do**
    Client k **executes:**
    Initialize $w_k \leftarrow w_t$
    for *each local epoch* $e = 1, 2, \ldots, E$ **do**
      for *each batch* $b \in B$ **do**
        Update $w_k \leftarrow w_k - \eta \nabla f_k(w_k; b)$
      end
    end
    Return updated model w_k to the server
  end
  Server aggregates:
  $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_k$
end

Algorithm 1. Federated averaging with hyperparameters.

Initialization
total_correct $\leftarrow 0$
total_samples $\leftarrow 0$ **for** ***each client*** *client* ***in*** *test_data_clients* **do**
  client_test_data $\leftarrow$ *client.test_data*
  client_ground_truth_labels $\leftarrow$ *client.ground_truth_labels* **for** *i* = 1 *to* *len(test_data_clients)* **do**
    sample $\leftarrow$ *client_test_data[i]*
    ground_truth_label $\leftarrow$ *client_ground_truth_labels[i]*
    Compute *predicted_label* **for data point** sample
    predicted_label $\leftarrow$ *federated_model.predict(sample)*
    if *predicted_label* = *ground_truth_label* **then**
      total_correct $\leftarrow$ *total_correct* + 1
      total_samples $\leftarrow$ *total_samples* + 1
    end
  end
Calculate the Accuracy of the model
accuracy $\leftarrow$ *total_correct/total_samples* $*$ 100%
Output the final accuracy of the federated model

Algorithm 2. Algorithm for accuracy.

percentage of data points that each participating client gave. The federated averaging equation is mentioned in Equation 10.

$$f(a) = \sum_{m=1}^{m} \frac{n_m}{n} F_m(a) \quad \text{where} \quad F_m(a) = \frac{1}{n_m} \sum_{i \in P_m} f_i(a) \tag{11}$$

Algorithms 1–3 display federated averaging, calculation of accuracy and calculation of precision, recall and F1-score, respectively.

4.2 Presentation of the experimental results

The federated learning implementation on the smoking dataset yielded promising results in terms of model performance and accuracy. The following subsections provide a comprehensive analysis of the experimental results.

$$\text{Accuracy} = \frac{\text{Total Correct Predictions}}{\text{Total Samples}} \times 100 \tag{12}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{13}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{14}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

4.2.1 Performance metrics

The trained federated learning model achieved a classification accuracy of 97.65% on the test set. This indicates that the model successfully learned patterns and features from the training data and generalized well to unseen instances. To assess the model's performance across different smoking activities, precision, recall, and F1-score were computed for each class (e.g., smoking, non-smoking). The precision metric measures the proportion of correctly predicted positive instances, recall evaluates the model's ability to identify true positives, and the F1-score provides a balance between precision and recall. The results showed high precision, recall, and F1-score values for smoking-related activities, indicating the effectiveness of the model in detecting smoking behaviors. Table 6 shows the values of each of the performance metrics. Figures 3–6 summarizes the performance of each model in terms of accuracy, F1-score, precision and recall, respectively. To ensure robustness and avoid overfitting, the model was validated using cross-validation on training data, with separate test sets held out for final evaluation. While the model demonstrated high performance on the test data, future work should involve real-world testing to confirm the model's generalizability across diverse environments and populations.

4.3 How is federated learning superior to traditional machine learning

Firstly, federated learning leverages the collective knowledge of multiple distributed data sources. By training the model on data from diverse devices or servers, federated learning

```
Initialization
True Positive, TP ← 0
False Positive, FP ← 0
False Negative, FN ← 0 for each client client in
test_data_clients do
 client_test_data ← client.test_data
 client_ground_truth_labels ←
 client.ground_truth_labels for i = 1 to
 len(test_data_clients) do
  sample ← client_test_data[i]
  ground_truth_label ←
  client_ground_truth_labels[i]
  Compute predicted_label for data point
  sample
  predicted_label ←
  federated_model.predict(sample)
  if predicted_label = ground_truth_label and
  predicted_label = POSITIVE_LABEL then
      end
          TP ← TP + 1 if
  predicted_label ≠ ground_truth_label and
  predicted_label = POSITIVE_LABEL then
      end
  FP ← FP + 1 if
  predicted_label ≠ ground_truth_label and
  predicted_label ≠ POSITIVE_LABEL then
      end
  FN ← FN + 1
   end
  end
Calculate the precision of the model
precision ← TP/(TP + FP)
Calculate the recall of the model
recall ← TP/(TP + FN)
Calculate the F1-score of the model
f1score ← 2 * (precision * recall)/(precision + recall)
Output the final precision, recall, and F1-score
of the federated model
```

Algorithm 3. Algorithm for precision, recall, and F1-score.

captures a broader range of smoking behavior patterns, leading to improved performance. Secondly, federated learning promotes parallel processing and model sharing. This allows for the utilization of the combined computational power and insights from various devices or servers, enhancing the model's accuracy and robustness. Furthermore, federated learning prioritizes data privacy. The training process occurs locally on individual devices or servers, ensuring that sensitive data, such as personal health information, remains secure and private. This approach enables the inclusion of more data while maintaining privacy, resulting in a more accurate model. In contrast, traditional machine learning approaches often require centralizing the data, which may compromise privacy or limit access to certain datasets. Centralized models also encounter challenges related to data transfer, bias, and scalability when dealing with distributed and privacy-sensitive data. Therefore, the collaborative nature, parallel processing capabilities, model sharing, and privacy-preserving aspects of federated learning contribute to its higher accuracy compared to traditional machine learning approaches.

TABLE 6 Models comparison based on performance metrics.

Model name	Accuracy	F1-score	Precision	Recall
Federated Learning	97.65	97.41	97.31	97.36
Random Forest Classifier	95.25	96.54	94.93	95.23
Decision Tree Classifier	93.61	94.78	91.25	92.87
Logistic Regression	90.68	91.82	89.39	89.67
KNeighbors Classifier	87.23	88.17	88.42	87.58
Support Vector Machines	85.42	86.26	85.78	85.65

4.3.1 Comparative analysis

Analysing the four evaluation metrics-accuracy, Precision, Recall, and F1-score, it provides valuable insights into the performance model. These metrics provide a comprehensive view of the model's performance in classification tasks and facilitate the evaluation of its overall predictive ability. Accuracy represents the ratio of correctly classified samples to the total number of samples and assesses the overall correctness of the predictions. Precision focuses on the proportion of true positive predictions among all positive predictions, indicating the model's capacity to reduce false positives. Recall, investigates the ratio of true positives to the total number of actual positive samples, indicating the model's ability to recognize positive instances. The F1-score provides a harmonic mean of precision and recall, balancing the two measures.

Figure 7 displays a bar plot that illustrates the relationship between cholesterol and hemoglobin levels. The chart visually represents the distribution of these two and provides a comparative view of their values. The bar plot allows to visually compare the distribution of hemoglobin levels across different cholesterol categories. This visualization helps to identify any potential correlations or patterns between cholesterol and hemoglobin.

Figure 8 presents a bar plot that showcases the relationship between urine protein levels and smoking status. The chart visually represents how urine protein levels vary across different groups based on whether a person smokes or not. This visualization helps in assessing the association between smoking and urine protein levels in the dataset.

Figure 9 depicts a bar plot that illustrates the relationship between cholesterol levels and smoking behavior. This chart visually presents the distribution of cholesterol levels across

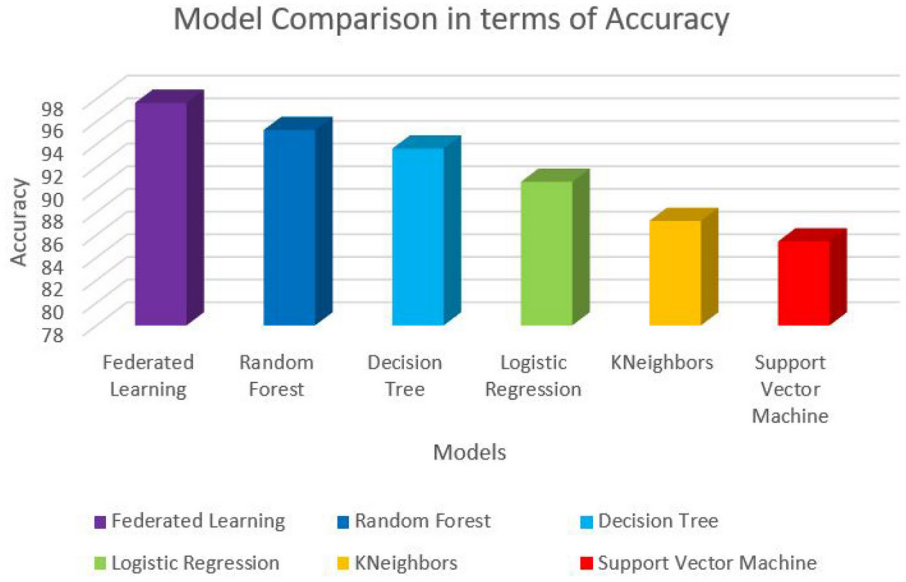

FIGURE 3
Performance comparison in terms of accuracy.

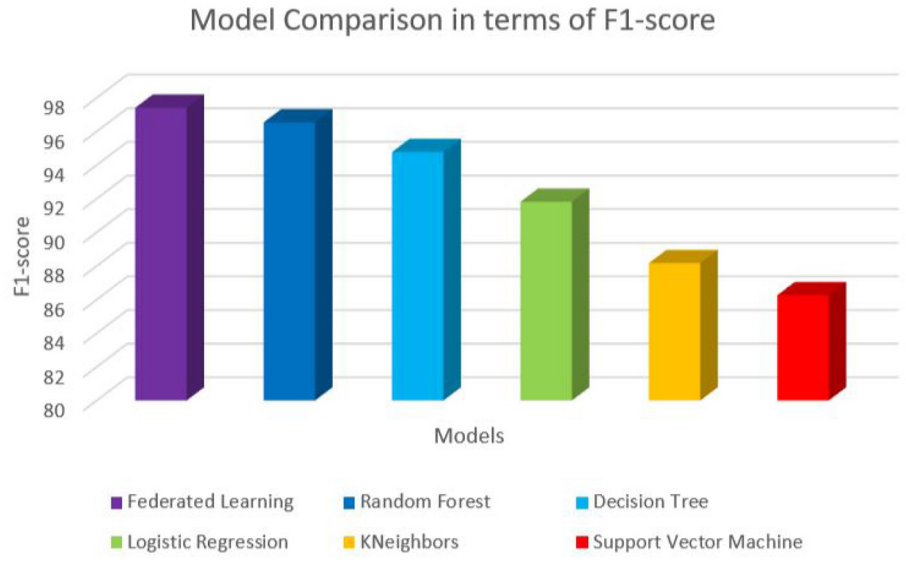

FIGURE 4
Performance comparison in terms of F1-score.

different groups based on whether individuals are smokers or non-smokers. By comparing the heights of the bars representing the two groups, we can observe any differences in the distribution of Cholesterol levels between smokers and non-smokers. This visualization helps in understanding the potential impact of smoking on Cholesterol levels.

Figure 10 presents a bar plot that illustrates the relationship between hemoglobin levels and smoking behavior. The chart visually represents the distribution of hemoglobin levels across different groups based on whether individuals are smokers or non-smokers. By comparing the heights of the bars representing the two groups, we can assess any variations in hemoglobin levels between smokers and non-smokers. This visualization aids in understanding the potential association between smoking and hemoglobin levels.

4.4 Comparison of the results with other existing approaches

The federated learning approach is superior in terms of performance and accuracy when compared to other approaches currently in use. Smoking behavior analysis has traditionally been

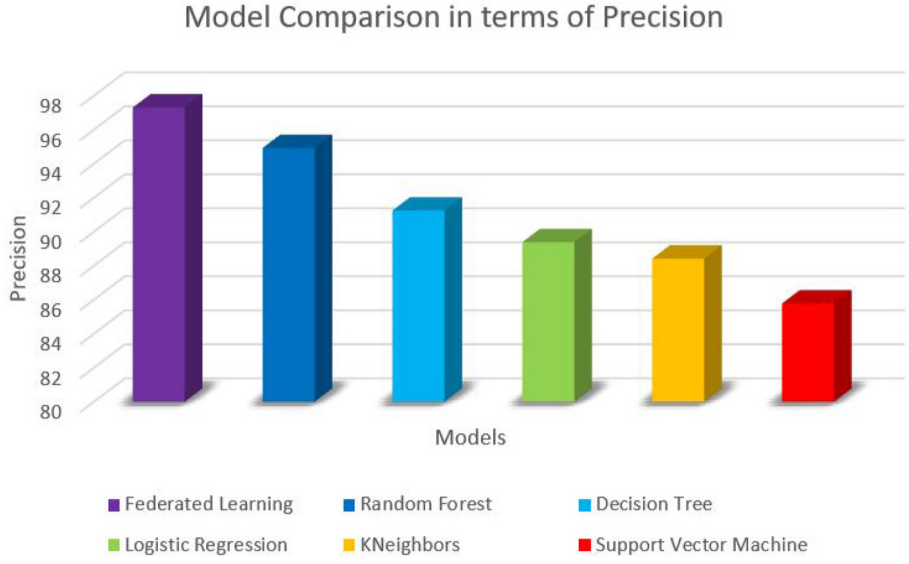

FIGURE 5
Performance comparison in terms of Precision.

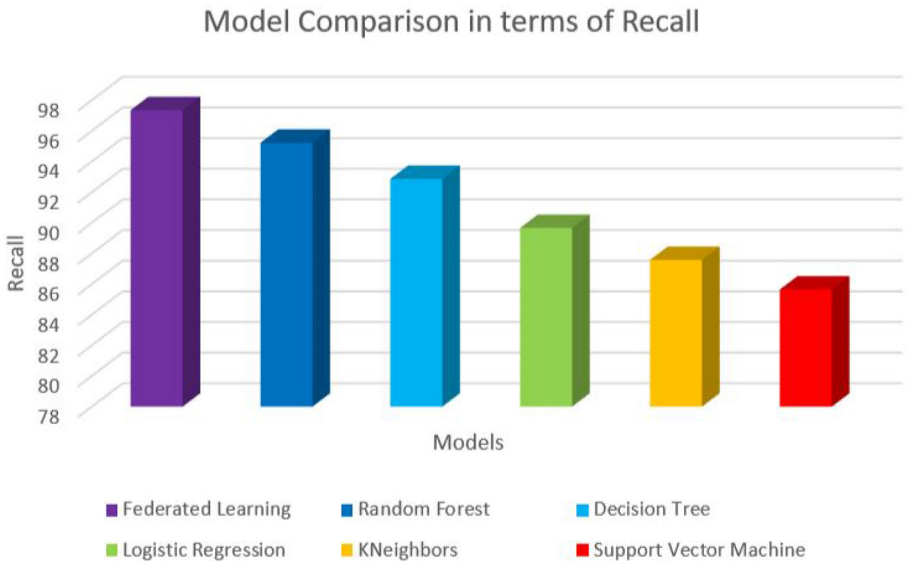

FIGURE 6
Performance comparison in terms of Recall.

carried out using conventional machine learning techniques like SVMs and random forests. By utilizing the collaborative training across decentralized data sources, the federated learning model outperformed these approaches.

The federated learning approach's improved accuracy is a result of its capacity to include data from various sources and carry out dispersed training. The potential of federated learning to analyze smoking information more effectively and robustly than standard centralized learning techniques is highlighted by this finding.

The federated learning approach is a potential option for real-world applications in the field of smoking behavior analysis since it can adapt to new and developing data sources while retaining privacy and security.

5 Discussion

5.1 Advantages, disadvantages, and challenges of the work

The proposed federated learning framework for smoking prediction offers notable advantages, including privacy preservation by avoiding centralized data storage and

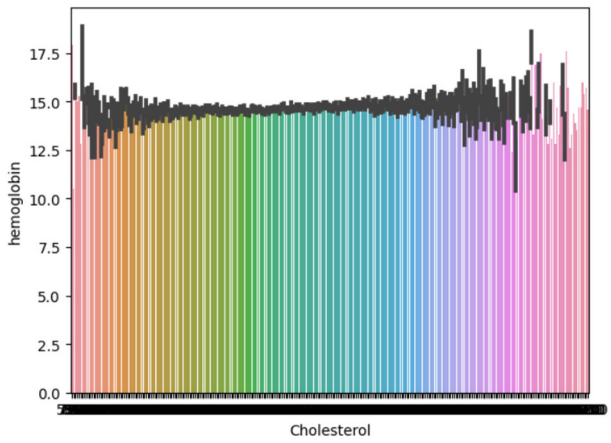

FIGURE 7
Bar plot between cholesterol and hemoglobin.

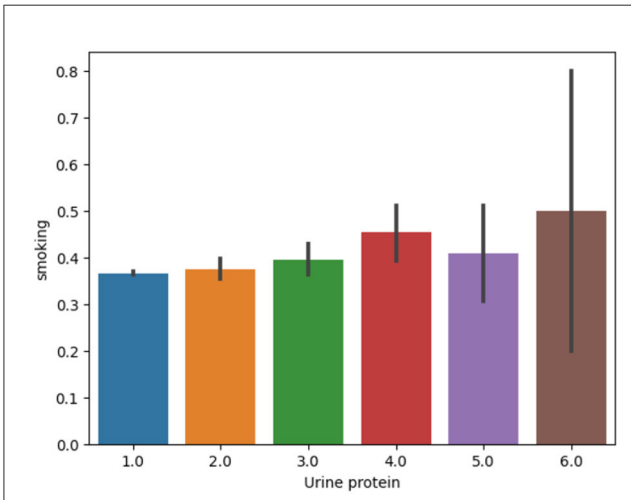

FIGURE 8
Bar plot between urine protein and smoking.

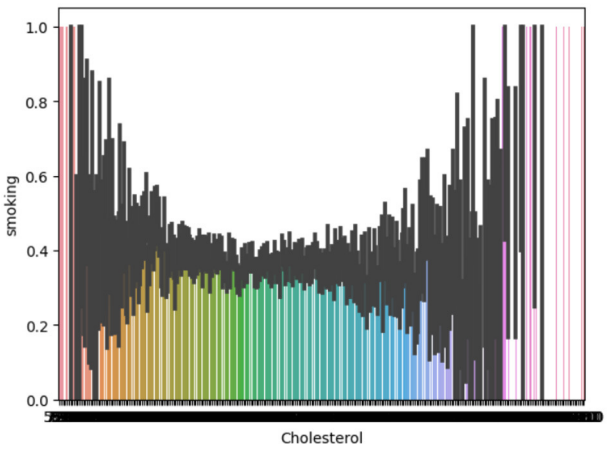

FIGURE 9
Bar plot against cholesterol and smoking.

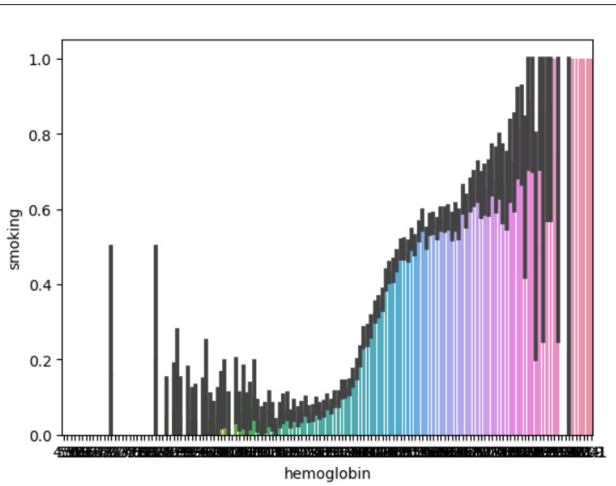

FIGURE 10
Bar plot between hemoglobin and smoking.

improving model generalizability across diverse populations. The decentralized approach captures nuances in smoking behavior, resulting in high accuracy and efficient parallel processing on local devices. However, challenges include communication overhead due to frequent model updates, potential inconsistencies from varied data distributions (model drift), and computational demands that can strain resource-limited devices. Additionally, the system faces security risks such as model poisoning, scalability complexities due to continuous aggregation needs, and potential delays that may impact real-time applicability in healthcare. Despite these limitations, the framework's privacy-preserving nature and accuracy make it a promising approach in health monitoring.

5.2 Comparison of the time-complexity of federated learning and machine learning

Federated Learning (FL) and traditional Machine Learning (ML) differ in terms of their time complexity.

In FL, the time complexity is influenced by the distributed nature of the learning process. Since the training occurs on multiple devices or servers, the overall time required for training is dependent on factors such as network latency, communication overhead, and the number of participating devices. The time complexity of FL can be higher compared to traditional ML when considering the coordination and synchronization required among the distributed entities. however, advancements in optimization techniques, efficient communication protocols, and parallel processing capabilities have helped mitigate the time complexity challenges in FL.

On the other hand, the time complexity of traditional machine learning is primarily influenced by the magnitude of the dataset as well as the level of difficulty of the learning algorithm. Training a model on a centralized dataset typically involves processing all the data at once, which can be time-consuming for large datasets. The analysis of the processing requirements, the amount of features, and the number of data points are used to determine the time

complexity of ML techniques like logistic regression, decision trees, and support vector machines.

5.3 Discussing areas for improvement or future research

The application of federated learning to the smoking dataset brings to light a number of potential areas for further investigation and improvement in the field of education and research. These domains can help to improve the federated learning approach's performance, scalability, and application in the context of smoking behavior study. Consider the following points:

- Algorithmic improvements: Investigate and create unique federated learning algorithms customized specifically for smoking behavior analysis. To increase convergence speed and model performance, sophisticated optimization techniques, adaptive learning rate schemes, and more efficient aggregation methods may be investigated.
- Feature Engineering: Exploration of new contextual characteristics, such as environmental conditions, social interactions, or psychological states, to increase the accuracy and interpretability of smoking behavior identification models.
- Privacy preservation techniques: Investigate and develop novel privacy-preserving strategies that provide higher privacy guarantees while preserving model performance. Exploration of advanced cryptographic approaches, federated learning with differential privacy, or safe multi-party computation techniques may be included.
- Model personalization: Look into ways to personalize federated learning models for individual users while maintaining privacy. Investigate approaches for adapting to user-specific smoking patterns and behaviors, which could lead to personalized smoking cessation programs.
- Dataset augmentation: To strengthen the generalization capabilities of the federated learning models, consider extending the smoking dataset with additional examples, different populations, and smoking-related scenarios. This can improve the model's ability to handle differences in smoking behaviors across populations and environmental situations.
- Real-time monitoring: Explore real-time federated learning systems that enable continuous monitoring and analysis of smoking behaviors. This may entail creating efficient communication protocols and lightweight model architectures to allow for real-time inference and feedback.
- Benchmarking and standardization: Establish benchmarks and standardized evaluation methodologies to allow for fair comparison and repeatability of federated learning models for smoking behavior analysis. This can facilitate researcher collaboration and boost advancements in the subject.

By addressing these areas for improvement and conducting further research, the application of federated learning on smoking datasets can continue to evolve, leading to more accurate, privacy-preserving, and robust models for understanding and addressing smoking behavior.

6 Conclusion

In conclusion, this research paper presented a comprehensive implementation of federated learning on the smoking dataset. The study's findings showed how well the federated learning strategy analyzed and categorized smoking-related activities. The model achieved an accuracy of 97.65%, a precision of 97.31%, a recall of 97.36%, and an F1-score of 97.41%, outperforming traditional machine learning algorithms applied to the same dataset. The federated learning approach showcased several advantages, including privacy preservation, data diversity, parallel processing, and model shareability. By leveraging these advantages, the model successfully captured patterns and features related to smoking behaviors, leading to improved accuracy and performance. The findings of this study have significant implications for the field of smoking behavior analysis and highlight the potential of federated learning as a robust and privacy-preserving approach for analyzing sensitive healthcare data. Future research should focus on integrating adaptive learning mechanisms to improve real-time applicability and enhancing scalability through advanced aggregation techniques. Extending the framework's application to larger and more diverse datasets could also strengthen its effectiveness. Overall, the successful implementation of federated learning on the smoking dataset paves the way for advancements in healthcare analytics, personalized interventions, and public health initiatives related to smoking cessation and prevention.

Data availability statement

The datasets used in the proposed work is available with the corresponding authors, which could be provided based on the request. Requests to access these datasets should be directed to nallakaruppan.k@bimmpune.edu.

Author contributions

SF: Formal analysis, Writing – original draft. DR: Formal analysis, Writing – original draft. NM: Investigation, Visualization, Writing – original draft. AV: Formal analysis, Writing – review & editing. MN: Software, Writing – review & editing. SS: Resources, Validation, Writing – review & editing. PM: Project administration, Writing – review & editing. VM: Project administration, Writing – review & editing. IH: Data curation, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Norwegian University of Science and Technology, Norway.

Conflict of interest

One of the authors of the paper, SS, who is the part of the proposed work, is also one of the associate editors of this journal.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Aguileta, A. A., Brena, R. F., Mayora, O., Molino-Minero-Re, E., and Trejo, L. A. (2019). Multi-sensor fusion for activity recognition-a survey. *Sensors* 19:3808. doi: 10.3390/s19173808

Alexander, A., Jiang, A., Ferreira, C., and Zurkiya, D. (2020). An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. *J. Am. Coll. Radiol.* 17, 165–170. doi: 10.1016/j.jacr.2019.07.019

Al-Mudimig, A., Saleem, F., and Ullah, Z. (2009). A framework of an automated data mining systems using ERP model. *Int. J. Comput. Electr. Eng.* 1:101. doi: 10.7763/IJCEE.2009.V1.101

Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., and Eskofier, B. (2022). Federated learning for healthcare: systematic review and architecture proposal. *ACM Trans. Intell. Syst. Technol.* 13, 1–23. doi: 10.1145/3501813

Bonawitz, K. (2019). Towards federated learning at scale: system design. *arXiv preprint arXiv:1902.01046*. doi: 10.48550/arXiv.1902.01046

Breiman, L. (2001). Random forests. *Machine Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D., et al. (2021). Secureboost: a lossless federated learning framework. *IEEE Intell. Syst.* 36, 87–98. doi: 10.1109/MIS.2021.3082561

Coughlin, L. N., Tegge, A. N., Sheffer, C. E., and Bickel, W. K. (2020). A machine-learning approach to predicting smoking cessation treatment outcomes. *Nicot. Tobacco Res.* 22, 415–422. doi: 10.1093/ntr/nty259

El Ouadrhiri, A., and Abdelhadi, A. (2022). Differential privacy for deep and federated learning: a survey. *IEEE Access* 10, 22359–22380. doi: 10.1109/ACCESS.2022.3151670

Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). Personalized federated learning: a meta-learning approach. *arXiv preprint arXiv:2002.07948*. doi: 10.48550/arXiv.2002.07948

Fan, L., Zhang, F., Fan, H., and Zhang, C. (2019). Brief review of image denoising techniques. *Vis. Comput. Industr. Biomed. Art* 2:7. doi: 10.1186/s42492-019-0016-7

Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Möllering, H., Nguyen, T. D., et al. (2021). "SAFELearn: Secure aggregation for private federated learning," in *2021 IEEE Security and Privacy Workshops (SPW)* (IEEE), 56–62.

Hu, L., Li, L., and Ji, J. (2020). Machine learning to identify and understand key factors for provider-patient discussions about smoking. *Prev. Med. Rep.* 20:101238. doi: 10.1016/j.pmedr.2020.101238

Hu, R., Guo, Y., Li, H., Pei, Q., and Gong, Y. (2020). Personalized federated learning with differential privacy. *IEEE Internet Things J.* 7, 9530–9539. doi: 10.1109/JIOT.2020.2991416

Huang, Y., Zhou, Y., Zhao, H., Riedel, T., and Beigl, M. (2024). "A survey on wearable human activity recognition: innovative pipeline development for enhanced research and practice," in *2024 International Joint Conference on Neural Networks (IJCNN)* (IEEE), 1–10.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. *Found. Trends Machine Learn.* 14, 1–210. doi: 10.1561/2200000083

Kopper, A., Karkare, R., Paffenroth, R. C., and Apelian, D. (2020). Model selection and evaluation for machine learning: deep learning in materials processing. *Integr. Mater. Manufact. Innov.* 9, 287–300. doi: 10.1007/s40192-020-00185-1

Kugic, A., Abdulnazar, A., Knezovic, A., Schulz, S., and Kreuzthaler, M. (2024). "Smoking status classification: a comparative analysis of machine learning techniques with clinical real world data," in *International Conference on Artificial Intelligence in Medicine* (Berlin: Springer), 182–191.

Kumar, V., Sinha, N., Yadav, A., Singh, A., Meena, V., and Mathur, A. (2023). "Recognition of parkinson's disease using different machine learning models," in *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), Vol. 1* (IEEE), 1–6.

Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H., and Langlotz, C. P. (2020). Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology* 295, 675–682. doi: 10.1148/radiol.2020192536

Li, K. H., de Gusmão, P. P. B., Beutel, D. J., and Lane, N. D. (2021). "Secure aggregation for federated learning in flower," in *Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning*, 8–14.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2019). Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*. doi: 10.48550/arXiv.1905.10497

Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., et al. (2020). Think locally, act globally: federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*. doi: 10.48550/arXiv.2001.01523

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, eds. A. Singh and J. Zhu (Proceedings of Machine Learning Research (PMLR)), 1273–1282.

Mohammadi, S., Balador, A., Sinaei, S., and Flammini, F. (2024). Balancing privacy and performance in federated learning: a systematic literature review on methods and metrics. *J. Parall. Distribut. Comput.* 2024:104918. doi: 10.1016/j.jpdc.2024.104918

Mottini, A., and Acuna-Agost, R. (2016). "Relative label encoding for the prediction of airline passenger nationality," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (Barcelona: IEEE), 671–676.

Nag, A., Hassan, M. M., Mandal, D., Chand, N., Islam, M. B., Meena, V., et al. (2024). "A review of machine learning methods for IoT network-centric anomaly detection," in *2024 47th International Conference on Telecommunications and Signal Processing (TSP)* (Prague: IEEE), 26–31.

Nallakaruppan, M., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., et al. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks* 12:164. doi: 10.3390/risks12100164

Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., et al. (2022). Federated learning for smart healthcare: a survey. *ACM Comput. Surv.* 55, 1–37. doi: 10.1145/3453476

Patil, Y., Lopez-Meyer, P., Tiffany, S., and Sazonov, E. (2013). "Detection of cigarette smoke inhalations from respiratory signals using reduced feature set," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE), 6031–6034.

Rajendran, S., Obeid, J. S., Binol, H., Foley, K., Zhang, W., Austin, P., et al. (2021). Cloud-based federated learning implementation across medical centers. *JCO Clin. Cancer Informat.* 5, 1–11. doi: 10.1200/CCI.20.00060

Ranbaduge, T., and Ding, M. (2022). Differentially private vertical federated learning. *arXiv preprint arXiv:2211.06782*. doi: 10.48550/arXiv.2211.06782

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*. doi: 10.48550/arXiv.1811.12808

Reich, Y., and Barai, S. (1999). Evaluating machine learning models for engineering problems. *Artif. Intell. Eng.* 13, 257–272.

Shao, R., He, H., Liu, H., and Liu, D. (2019). Stochastic channel-based federated learning for medical data privacy preserving. *arXiv preprint arXiv:1910.11160*. doi: 10.48550/arXiv.1910.11160

Sinha, K., and Ghosh, N. (2024). A review on the recent advancements in machine learning-assisted tobacco research. *NIPES-J. Sci. Technol. Res*. 6:11223324. doi: 10.5281/zenodo.11223324

Stoian, A., Ivan, R., Stoian, I., and Marichescu, A. (2008). "Current trends in medical imaging acquisition and communication," in *2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Vol. 3* (Cluj-Napoca: IEEE), 94–99.

Sun, T., Li, D., and Wang, B. (2022). Decentralized federated averaging. *IEEE Trans. Pat. Anal. Machine Intell*. 45, 4289–4301. doi: 10.1109/TPAMI.2022.3196503

Swapno, S. M. R., Nobel, S. N., Islam, M. B., Haque, R., Meena, V., and Benedetto, F. (2024). "A novel machine learning approach for fast and efficient detection of mango leaf diseases," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)* (IEEE), 1–7.

Thummisetti, B. S. P., and Atluri, H. (2024). Advancing healthcare informatics for empowering privacy and security through federated learning paradigms. *Int. J. Sustain. Dev. Comput. Sci*. 6, 1–16.

Wang, Y., Guo, J., Zhang, J., Guo, S., Zhang, W., and Zheng, Q. (2023). "Towards fairer and more efficient federated learning via multidimensional personalized edge models," in *2023 International Joint Conference on Neural Networks (IJCNN)* (IEEE), 1–8.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., et al. (2020). Federated learning with differential privacy: algorithms and performance analysis. *IEEE Trans. Inform. For. Secur*. 15, 3454–3469. doi: 10.1109/TIFS.2020.2988575

Yang, Q., Huang, A., Fan, L., Chan, C. S., Lim, J. H., Ng, K. W., et al. (2023). Federated learning with privacy-preserving and model ip-right-protection. *Machine Intell. Res*. 20, 19–37. doi: 10.1007/s11633-022-1343-2

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol*. 10, 1–19. doi: 10.1145/3339474

Yu, H., Liu, Z., Liu, Y., Chen, T., Cong, M., Weng, X., et al. (2020). "A fairness-aware incentive scheme for federated learning," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 393–399.

Zelaya, C. V. G. (2019). "Towards explaining the effects of data preprocessing on machine learning," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (Macao: IEEE), 2086–2090.

Zhuo, H. H., Feng, W., Lin, Y., Xu, Q., and Yang, Q. (2019). Federated deep reinforcement learning. *arXiv preprint arXiv:1901.08277*. doi: 10.48550/arXiv.1901.08277