IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 June 2024, accepted 19 July 2024, date of publication 30 July 2024, date of current version 18 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435910

RESEARCH ARTICLE

Advances in Federated Learning: Combining Local Preprocessing With Adaptive Uncertainty Symmetry to Reduce Irrelevant Features and Address Imbalanced Data

ZAHRAA KHDUAIR TAHA[1,2], JOHNNY KOH SIAW PAW[3], YAW CHONG TAK[3], TIONG SIEH KIONG[3], (Senior Member, IEEE), KUMARAN KADIRGAMA[4,5], FOO BENEDICT[6], TAN JIAN DING[7], KHARUDIN ALI[8], AND AZHER M. ABED[9]

[1]College of Graduate Studies (COGS), Universiti Tenaga Nasional (The Energy University), Kajang 43000, Malaysia
[2]Department of Network, Faculty of Engineering, Al-Iraqia University, Baghdad, Iraq
[3]Institute of Sustainable Energy, Universiti Tenaga Nasional (The Energy University), Kajang 43000, Malaysia
[4]Faculty of Mechanical and Automotive Engineering Technology, Universiti Malaysia Pahang, Pekan, Pahang 26600, Malaysia
[5]College of Engineering, Almaaqal University, Basra 61003, Iraq
[6]Enhance Track Sdn. Bhd., Puchong, Selangor 47120, Malaysia
[7]School of Electrical Engineering and Artificial Intelligence, Xiamen University Malaysia, Sepang, Selangor 43900, Malaysia
[8]Faculty of Engineering Technology (Electrical and Automation), University College TATI, Kemaman, Terengganu 24000, Malaysia
[9]College of Engineering and Technologies, Al-Mustaqbal University, Babylon 51001, Iraq

Corresponding authors: Johnny Koh Siaw Paw (johnnykoh@uniten.edu.my), Tan Jian Ding (jianding.tan@xmu.edu.my), and Yaw Chong Tak (chongty@uniten.edu.my)

This work was supported by Tenaga Nasional Berhad (TNB) and Universiti Tenaga Nasional (UNITEN) through the BOLD Refresh Publication Fund for providing all out-laboratory support under Grant J510050002-IC-6 BOLDREFRESH2025-Centre of Excellence.

ABSTRACT Federated learning is increasingly being considered for sensor-driven human activity recognition, offering advantages in terms of privacy and scalability compared to centralized methods. However, challenges such as feature selection and client imbalanced data persist. In this study, FLP-DS2MOTE-USA is suggested, a system that integrates federated local preprocessing, adaptive thresholding based on uncertainty symmetry, and a density- sensitive synthetic minority over-sampling approach. Each client preprocesses data locally and employs DS2MOTE for class balancing. On the server side, adaptive thresholding based on uncertainty symmetry is utilized to identify the optimal client for training the global mode. Evaluation on two distinct datasets—Human Activity Recognition with Smartphones and Human Activity Recognition (OpenPose) —reveals that our model outperforms FedAvg, FedSgd, FedSmote, and FedNova, achieving accuracies of 90.57% and 96.58%, respectively. In addition, FLP-DS2MOTE-USA minimizes update size and network overhead on the Human Activity Recognition with Smartphones, while achieving improvements on the OpenPose dataset. Overall, the proposed method not only addresses issues of imbalanced data but also reduces computational complexity via streamlined local preprocessing, and server-side mechanisms ensure client privacy. It outperforms traditional federated learning techniques in both accuracy and efficiency.

INDEX TERMS Federated learning, local preprocessing, imbalance data, uncertainty symmetry.

I. INTRODUCTION

As the Internet of Things (IoT) grows, the use of sensors to gather data across various sectors, like healthcare and smart infrastructure, is on the rise. One important application of this technology is sensor-driven human activity recognition (HAR), which uses sensors to collect data and analyze human movements in real environments. This technology can enhance patient monitoring, soldier surveillance, and tracking of the elderly, among other uses [1]. This data is then analyzed

The associate editor coordinating the review of this manuscript and approving it for publication was Daniel Augusto Ribeiro Chaves.

by big data analytics or artificial intelligence (AI) to offer predictive insights or help in decision-making processes [2].

Historically, the vast amounts of data produced by these numerous devices have been handled through a centralized model, where all the information is sent to and processed on a single server. This traditional approach, while straightforward, comes with its own set of challenges. It can lead to high costs associated with transferring data and can cause congestion in network traffic. Moreover, it demands the central server to have substantial computational and storage capacities, which in turn, ramps up operational expenses.

Federated Learning (FL) is an innovative learning model that addresses these issues by distributing the process across various devices, offering a more efficient and cost-effective solution [3]. FL fosters a cooperative environment where multiple clients work together to build a global model, while keeping their data confidential and stored on their own devices. A primary strategy employed in FL is Federated Averaging, or FedAvg for short. In this process, each participant trains a local model using their specific data and then sends just the model's parameters, rather than the data itself, to a central server. This server merges these parameters to develop a unified model.

Distinguishing FL from conventional centralized methods is its utilization of the computing resources available on each client's device during the training period. However, the approach faces several challenges that can affect its efficiency, including how data is prepared, the uneven distribution of data across devices, differences in computing capabilities, network reliability, and the varying sizes of the datasets held by each client [1]. Addressing these challenges is critical for the successful deployment of FL systems.

Our research aims to address several key open research questions (ORQs) in the context of FL for HAR:

1. How can data imbalance be effectively managed in FL systems?
    - Data imbalance within and across clients poses significant challenges to the efficiency and accuracy of FL models. We introduce a modified SMOTE algorithm to create synthetic data points for underrepresented classes, addressing intra-class imbalance.
2. What are the optimal feature selection and dimensionality reduction techniques for FL?
    - Effective feature selection and dimensionality reduction are crucial for improving model performance. We employ Chi-square and Linear Discriminant Analysis (LDA) to identify pertinent features and simplify data structures.
3. How can the variability in client datasets be effectively managed?
    - Variability in the size and distribution of client datasets affects the global model's performance. We propose using symmetric uncertainty with adjustable thresholds for client selection to manage dataset variability.

FL relies heavily on the first data processing at each client, which helps to standardize heterogeneous datasets and improve the performance of the model. This step ensures every participant adds valuable data to the overall model, improving both its efficiency and accuracy.

When algorithms pull out too many features, it becomes essential to pick the most relevant ones and reduce complexity through feature selection and dimensionality reduction. These techniques refine the data on a client-by-client basis, focusing on key features to make FL models simpler yet still precise, which helps streamline training both locally and globally [4]. Another challenge in FL comes from the differences in data distribution and size across clients. These differences can appear in several forms: some clients may have an imbalance within their data (e.g., 80% of one class and 20% of another), between clients (one client's data might be mostly one class, while another's could be predominantly another class), and in the size of the datasets each client contributes (imagine one client adds 2000 samples, while others contribute only 500 or 50 samples). As a result, the global model on the server will be different from the ideal model, causing slow convergence, longer training times, reducing the efficiency and scalability of FL frameworks. Meanwhile, the discrepancies between clients and the varied sizes of their datasets might lead to a federated model that's biased or struggles to perform consistently across different scenarios [2], [4], [5].

A main goal of FL is to achieve the highest possible accuracy as quickly as can be done [7]. However, this goal can be impeded by challenges like the preprocessing of data by individual clients and the imbalance of data across different clients, which can prolong the training phase and compromise the accuracy of the model [1], [2], [5]. To address these issues, there's a significant amount of research dedicated to finding ways to lessen the negative effects of these challenges on FL, focusing on improving local data preprocessing [8] and tackling various kinds of data imbalances [8], [9].

In machine learning, local preprocessing methods like dimensionality reduction, feature extraction, and feature selection are essential. They not only help in reducing computational burden by discarding unnecessary features but also enhance the accuracy of predictions by focusing on crucial variables and minimize the risk of overfitting by simplifying the model's complexity. Yet, existing research often isolates these techniques, not always achieving optimal outcomes. For instance, some studies might exclusively apply feature extraction algorithms [11], while others prioritize feature selection [1], [7], and often, these singular strategies fall short of reaching the intended performance metrics.

Moreover, the challenge of data imbalance is pivotal in machine learning and becomes particularly complex in FL. The currently available strategies for managing imbalanced data usually lack in efficiency and flexibility in FL scenarios. The literature [1], [9], [11], [12] often focuses on a specific type of data imbalance, missing the opportunity to provide comprehensive solutions for the diverse imbalance problems.

Therefore, there's a pressing need for more research focused on developing approaches that can effectively address the unique complications of data imbalances and local preprocessing within FL environments

Despite numerous studies aimed at preventing a decline in learning efficiency due to uneven class distribution among clients, as far as we're aware, there isn't an integrated approach that addresses three key elements. These are: 1) leveling the class imbalance within individual clients, 2) alleviating class imbalance and data size variations between different clients, and 3) employing feature selection and dimensionality reduction in FL to substantially boost model performance by simplifying the data structure and focusing on essential features.

This article introduces a groundbreaking methodology to combat the prevalent problem of data imbalance in FL systems. The approach leverages a modified version of the SMOTE algorithm to create synthetic data points for classes that are underrepresented, thus solving intra-class imbalance. To cater to the variability in client datasets, both in terms of size and class distribution, the method employs an innovative use of symmetric uncertainty combined with adaptive thresholds for choosing relevant clients. In addition, the paper enriches its strategy by adopting chi-square technique and Linear Discriminant Analysis for optimized feature selection. Collectively, these techniques establish a robust and equitable framework that elevates the performance of client-based models in FL. This research is poised to become a cornerstone in the fast-developing domain of FL.

The core achievements of our research can be summarized as follows:

1. The innovative framework FLP-DS2MOTE-USA is designed to address several challenges unique to the FL environment. This framework is configured to address issues such as local data preprocessing, unbalanced data, model convergence, and communication overhead.
2. Introduced the Chi-LDA method for efficient extraction and identification of pertinent features within datasets.
3. Presented a revised version of D2SMOTE to achieve class equilibrium across various clients.
4. Implemented an approach using symmetry of uncertainty with an adjustable threshold for choosing participating clients to train and contribute optimal models to develop a strong global model.
5. Verified the performance of the suggested framework by conducting comprehensive trials using the UCI HAR and OpenPose HAR datasets.

II. RELATED WORK

In this section, the related work is organized into two parts. First, existing research utilizing federated local preprocessing techniques is reviewed, emphasizing methods for federated feature selection. Following this, the latest strategies for managing imbalanced data are surveyed through a range of re-balancing mechanisms.

A. FEDERATED LOCAL PREPROCESSING

1) FEDERATED FEATURE SELECTION

Feature selection is a crucial preprocessing procedurein data mining that simplifies the dataset by removing redundant attributes. These methods shorten the time required for training classification models while improving data presentation and comprehensibility [14]. Cassará et al. [4] have introduced a Federated Feature Selection (FFS) technique that allows Autonomous Vehicles (AVs) to collaborate in filtering sensor data, eliminating irrelevant attributes without having to share raw data. This method is built on two main components: 1) A novel algorithm employed by AVs to identify important features, leveraging a principle known as Mutual Information, and 2) A new aggregation function that is executed at the edge computing layer. Xiao et al. [11] introduced a model named a Perceptive Extraction Network (PEN), aimed at personalized data analysis. PEN is essentially comprised of two key components: a feature network and a relation network. The feature network, built around a convolutional block, is tasked with identifying specific local characteristics in data pertaining to human activities. Meanwhile, the relation network combines Long Short-Term Memory (LSTM) with attention mechanisms to capture hidden patterns within the data. Rui Z. et al. [8] utilized a Gini-impurity-based feature selection system for eHealth applications in Vertical Federated Learning. It's versatile across multiple machine learning models. The πSS−FS protocol employs a quick secret sharing approach, making it nearly as fast as unencrypted methods. Qin and Kondo [15] suggested a system for intrusion detection, which utilizes FL and feature selection techniques. First, a greedy algorithm identifies key features that improve the system's ability to detect different kinds of attacks. Then, based on these chosen features, the FL server generates several global models.

2) FEDERATED DIMENSIONALITY REDUCTION

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while preserving as much information as possible. This can be done for a variety of reasons, including to reduce model complexity, improve the performance of learning algorithms, and facilitate data visualization. Little attention has been paid to how FL pipelines can incorporate dimensionality reductions performed by collaborators. Distributed ML tasks allow dimensionality reduction to be performed centrally before distributing the data to different computational cores, but this is not possible in FL workflows because the data is not shared. El Ouadrhiri et al. suggested using Hensel's Lemma to reduce the size of the dataset without losing any information. The proposed method achieved an accuracy of 97% using only 25% of the original dataset [16]. Cheung et al. introduced a new technique known as federated principal component analysis, tailored for vertical partitioning across clients. This method focused on reducing the dimensions of datasets distributed among various clients, enabling the extraction of significant information for subsequent data analysis [17]. Huang W. and Barnard A. have innovatively designed and demonstrated a new dimensionality reduction scheme, FedRed, within the FL pipeline, specifically aimed at preparing heterogeneous datasets for collaborative learning. This approach not only facilitates the integration of diverse data sources but also enhances the efficiency of the learning process by ensuring faster convergence and greater model adaptability across varied data characteristics [18].

B. INTRA-CLASS IMBALANCE

This kind of imbalance indicates that a client's class distribution—the amount of data distributed among classes—differs from the uniform distribution. For example, Client 1's dataset contains 90 emails labeled as "Not Spam" and only 10 labeled as "Spam," indicating a significant imbalance in favor of the "Not Spam" class [3].

Abdellatif et al. [19] have used a refined strategy for allocating Edge Users (EUs) and resources in the context of a hierarchical FL system. This method leverages several edge servers to reduce the computational and communication overhead involved in transferring data between the EUs and the central server. To function as a classification tool, Tabassum et al. [12] utilized a Generative Adversarial Network (GAN) across various IoT devices. The network is trained using augmented local datasets. The system's effectiveness is evaluated by comparing its accuracy and learning speed with those of existing collaborative models designed for detecting security breaches. Khan et al. [13] proposed a model that adjusts the influence of local models according to their accuracy for each category, aiming to improve predictions of thermal comfort and data sharing efficiency. This brief summary effectively communicates the core idea of their strategy. Yang et al. [10] suggested a framework that chooses a group of clients with the least class imbalance, inspired by the multi-arm bandit concept. The proposed algorithm could significantly enhance the overall model's learning speed and decision-making quality. Li et al. [20] have developed a secure data handling system for brain imaging research across multiple sites. This system gives priority to privacy while facilitating the sharing and analysis of model parameters. It is specifically designed to classify individuals as either having Autism Spectrum Disorder (ASD) or being neurotypical, by analyzing patterns in brain communication. Cui et al. [21] developed a new framework that enhances the accuracy and speed of prediction models. The first stage begins with analyzing the dataset of each client, followed by clustering to group clients based on this analysis. Each client then trains data locally with Bidirectional Long Short-Term Memory (LSTM) algorithms. Wu et al. [22] proposed the Dynamic Synthetic Images for Federated Learning (DSIFL) method to integrate data from different local instances. The primary process of the framework involves generating a varying number of synthetic images locally using the existing model to address class imbalance. Abbas et al. [23] developed a new model that utilizes FL with a Context Aggregator to address loss factors and class imbalance within a single client. The model features Context Aggregators based on validation loss to capture and mitigate loss, and another Context Aggregator specifically designed to address class imbalance. Wu et al. [24] suggested a new technique used in home monitoring to track HAR. A Generative Convolutional Autoencoder (GCAE) is utilized to solve imbalance among classes.

C. INTER-CLASS IMBALANCE

Inter-client class imbalance happens when the variety of classes is not the same across clients, resulting in each client having a distinctive class distribution compared to the rest. Feki et al. [25] suggested a FL system using VGG16 and ResNet50 deep neural networks for identifying COVID-19 through chest X-ray images. This decentralized setup enables medical experts globally to access valuable shared medical data while preserving patient privacy. Duan et al. [26] implemented the Astraea framework aims to tackle data imbalances in FL by employing two key approaches: 1) Using Z-score-based data augmentation to dynamically adjust and downsample the data, which helps in reducing global data imbalances, and 2) Introducing a Mediator component that adjusts the training timelines for clients according to the Kullback-Leibler divergence (KLD) of their respective data sets, in order to balance local discrepancies.

D. SIZE OF DATASET

The quantity of samples from various clients in FL can have a big impact on model performance. More data means that clients' models are more robust and dependable than those of clients with smaller sample sizes. As a result, models trained on larger datasets are regarded with more confidence [27]. To tackle the issue of varying numbers of samples among clients, Gong et al. [6] introduced a client clustering technique based on weighted voting, which automatically assigns each client to the appropriate cluster. Xu et al. [28] used a feature-regularized training method to curb local overfitting issues and harmonize parameter variations across clients, which assists in compiling a unified global feature extractor. Chen et al. [29] introduced the FedUC algorithm, which regulates the update uploads in FL by employing a client scheduling strategy that considers weight divergence, update size, and loss metrics. To counteract the effects of non-independent identical distribution, the approach employs image augmentation to equalize local client data. Yang et al. presented a new approach called the Modality-Collaborative Activity Recognition Network (MCARN), designed to operate in a FL environment. This method identifies modality-dependent features that are discriminative for activity recognition and achieves successful

TABLE 1. An overview of research studies on FL for federated local preprocessing and data imbalance.

Ref.	Objective	Methodology	Dataset	Key Findings	Contribution	Related To
[5]	Minimum achievable subset of features with	Mutual information cross entropy	MAV dataset WESAD dataset	High accuracy	N/A	Feature selection
[12]	A secure FL system for HAR	Perceptive extraction network long short- term memory (LSTM) and attention mechanism	WISDM UCI_HAR OPPORTUNITY PAMAP2	Better F1-score	Improve feature extraction increase security	Feature selection
[9]	Minimize the communication data	Gini-impurity	Fetal Health dataset DDos dataset	Improved with a 27% boost in accuracy	Improve privacy-protective method	Feature selection
[16]	Boost the accuracy of identifying issues.	Greedy algorithm	NSL-KDD dataset	25.7% better accuracy	N/A	Feature selection
[17]	Increase privacy protection while managing data efficiency in FL environments	Hensel's Lemma	MNIST	Accuracy of 97%, with only 25% of the original dataset size.	Provides more protection than the standard method	Dimensionality reduction
[18]	Study the unsupervised FL under the vertically partitioned dataset setting	Federated principal component analysis	UCI Machine Learning Repository	Improve accuracy	computational and communication efficiency	Dimensionality reduction
[19]	Mitigating the data limitation	Modified principle components analysis	Metallic nanoparticles dataset	Faster convergence	N/A	Dimensionality reduction
[20]	Minimize communication overhead	Hierarchical FL	Heartbeat dataset electrocardiogra m (ECG)	4–6% increase in the accuracy 75–85% reduction in the communication	New algorithm for aggregation	Intra client class imbalance
[13]	Secure IoT and Edge devices from internal and external attacks.	GAN	KDD-CUP99 NSL-KDD UNSW-NB15	N/A	N/A	Intra client class imbalance
[14]	Create a balanced, fair global model that includes contributions from all nodes and classes.	Class Precision-Weighted Aggregation technique	Thermal comfort database	Higher accuracy of 82.85% and lower communication costs by 25%	Devised a strategy for fair model aggregation in FL, addressing imbalanced non-iid data.	Intra client class imbalance
[11]	Select the set of clients with the most balanced class distribution.	Multi-arm bandit based algorithm	CIFAR10	Optimize the global model's convergence performance.	N/A	Intra client class imbalance
[21]	Enhancing the performance of deep learning processes	Domain adaptation	Autism Brain Imaging Data Exchange dataset	N/A	Maintain privacy	Intra client class imbalance
[22]	Improved computational performance	SARIMA-based clustering	Carbon emissions	MAE dropped by 63.32%, MSE decreased by 79.27%, and CS rose by 73.17%.	Implemented practical tests and validation in the industry sector	Intra client class imbalance
[23]	Merge information from varied local institutions	Dynamically Synthetic Images for FL	Covid-19 datasets	Reach higher accuracy	Dynamically produces synthetic images based on misclassified local data	Intra client class imbalance

TABLE 1. ***(Continued.)*** **An overview of research studies on FL for federated local preprocessing and data imbalance.**

[24]	Solve imbalance class	Context Aggregator	Covid-19	Exceeds the performance of FedAvg	The research outlines an approach based on Context Aggregators for privacy-preserving collaborative learning	Intra-client class imbalance
[25]	Achieve accurate and personalized health monitoring	Generative convolutional autoencoder	HAR dataset	10% accuracy	N/A	Intra-client class imbalance
[26]	Detection of COVID-19 from Chest X-ray images	VGG16 ResNet50	Covid-19	N/A	A decentralized system is recommended to allow medical professionals to utilize confidential data efficiently, without compromising patient privacy.	Inter-client class imbalance
[27]	Prevent the bias of training caused by imbalanced data distribution	Z-score-based data augmentation	EMNIST CINIC-10	Shows+4.39 and +6.51 percent improvement of top-1 accuracy communication traffic of Astraea is reduced by 75%	N/A	Inter-client class imbalance
[7]	Improve model accuracy and reduce communication overhead	Client clustering, weighted voting	CIFAR10 MNIST FMNIST	Boosting model accuracy by 9.24% and decreasing communication overhead by 4.67%.	N/A	Data size across clients
[29]	Securing excellent results in task-specific models while lowering the likelihood of overfitting during local representation learning	Feature-regularized	Fashion-MNIST and CIFAR-10 datasets	2-5% improvement in average accuracy can be obtained with slight extra communication overhead	N/A	Data size across clients
[30]	Equalize client datasets by employing image augmentation techniques	Client scheduling algorithm	Covid-19	Improve model accuracy and reduce costs associated with wireless communications.	Compresses gradients exceeding the established limits to enhance communication efficiency.	Data size across clients
[31]	Learn common features capture imbalance	Angular margin adjustment	HAR	Improve performance	Creating opportunities for practical applications in the real world	Data size across clients
[2]	Modify the allocation of data among diverse scenarios where the distribution is neither independent nor identical.	FLY-SMOTE	Hotels D.S Bank D.S Compass D.S Adult census income D.S	Improves the balance accuracy without compromising the model's accuracy	Revise SMOTE by synthesizing new data points from a random subset of k examples instead of the full minority class.	Intra/inter client class imbalance
[3]	Improved performance over existing federated learning algorithms	Data sampling client selection	CIFAR-10 MNIST	20% higher accuracy	Improving accuracy in non-iid scenarios	Intra/inter client class imbalance

MSE: Mean absolute error, MSE: mean square error, CS: convergence speed, HAR: Human activity recognition, non-iid: non-independent identical distribution

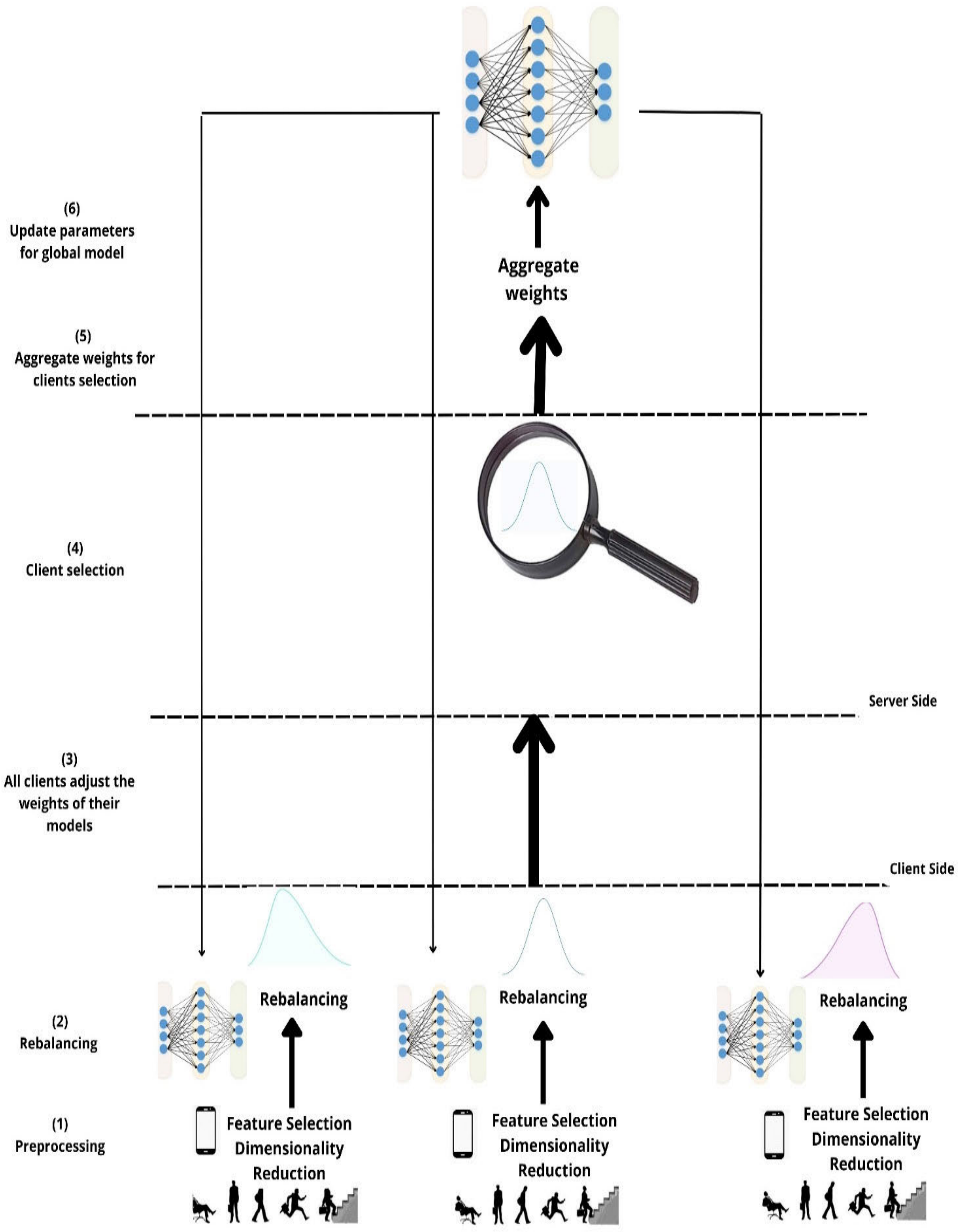

(6)
Update parameters for global model
Aggregate weights
(5)
Aggregate weights for clients selection
(4)
Client selection
Server Side
(3)
All clients adjust the weights of their models
Client Side
(2)
Rebalancing
Rebalancing
Rebalancing
Rebalancing
(1)
Preprocessing
Feature Selection Dimensionality Reduction
Feature Selection Dimensionality Reduction
Feature Selection Dimensionality Reduction

FIGURE 1. General framework of FLP-DS2MOTE-USA.

learning even when some clients are underrepresented compared to others [30].

In summary, few researchers have focused on important local preprocessing tasks such as feature selection and dimensionality reduction within the context of FL. This gap is especially important considering that FL clients often run on different devices and have different resource limitations. Moreover, there is a notable absence of comprehensive studies that address all categories of imbalanced data. According to the existing literature, there are limited studies proposing models that can also address these two types of challenges [2], [3], as shown in Table 1. Therefore, dedicated efforts are needed to develop frameworks that can effectively deal with these complexities in FL environments.

III. METHDOLOGY

An innovative framework named FLP-DS2MOTE-USA has been created to build upon current FL methodologies. This framework integrates Federated Local preprocessing (FLP), the Density-Sensitive Synthetic Minority Over-Sampling Technique (DS2MOTE), and Uncertainty Symmetric Adaptive algorithms. This approach autonomously sifts through raw data to extract and select features that are most beneficial for the learning task at hand. Moreover, it tackles the challenge of imbalanced data commonly found in real-world FL settings. Consequently, FLP-DS2MOTE-USA excels in both computational and communication efficiency, offering a well-rounded solution for practical FL deployments. Figure 1 outlines the procedural steps of FLP-DS2MOTE-USA, which shares similarities with the FedAvg framework [5]. The workflow consists of three key stages: (1) The master server distributes the overarching model to each of the client nodes; (2) Individual clients proceed to adapt this global model using their specific local data; and (3) The server then consolidates these client-specific model updates into a unified model. In a designated section, Federated Local Preprocessing (FLP) is unveiled as a method for pinpointing key features at each local data collection point, which in turn conserves both computational and communication efforts. Next, the DS2MOTE approach is suggested as a remedy for datasets where one class is disproportionately represented. The objective of this technique is to balance the class distribution as closely as possible to uniform, all without the necessity of sharing data. Finally, the server uses an adaptive uncertainty symmetry threshold to select clients that exhibit data distributions nearing uniformity. The flowchart of the methodology is presented in Appendix B.

A. CLIENT SIDE: ADDRESSING INTRA-CLIENT IMBALANCED DATA THROUGH FEDERATED LOCAL PREPROCESSING AND LOCAL REBALANCING

Two primary approaches are used by the FLP-DS2MOTE-USA framework to deal with imbalances in client data. First, by refining data features, Federated Local Preprocessing (FLP) is used to indirectly eliminate imbalanced data. Next, it directly addresses this issue by balancing class representation using the DS2MOTE algorithm.

1) FEDERATED LOCAL PREPROCESSING

In Federated Local Preprocessing (FLP), clients preprocess their data locally to emphasize the most influential features for classification tasks. This process consists of three main steps: feature selection using chi-square statistics, dimensionality reduction using Linear Discriminant Analysis (LDA), and validation of feature effectiveness using a Random Forest (RF) classifier.

Step 1: Future selection using Chi-square

In the first step, clients start by using chi-square statistics [26], [27] to determine which features in their data have the greatest influence and retain those, removing the less beneficial ones. This stage makes sure that the elements that are most important for data classification are the ones that are highlighted. The chi-square statistics for each feature in dataset $\boldsymbol{Di}$ of the $\boldsymbol{i - th}$ client is given by Equation (1)

$$\boldsymbol{X_{ij}^2 = \sum\nolimits_{j=1}^{ki} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}} \tag{1}$$

$\boldsymbol{X_{ij}^2}$: is the chi-square statistic for feature $\boldsymbol{j}$ in the dataset of client $\boldsymbol{i}$.

$\boldsymbol{O_{ij}}$ is the observed frequency of category $\boldsymbol{j}$ for a particular feature.

$\boldsymbol{E_{ij}}$ is the observed frequency of category $\boldsymbol{j}$ if there were $\boldsymbol{n}$ association (independent) between feature and class

ki is number of categories for feature in dataset Di.

Each client sorts the features based on their chi-square in descending order. High values indicate a strong association with class

The feature selection process

$$F_i = sort\left(X_i^2, order = desending\right)$$

F_i represent the indices of the top 50 features selected by client based on the chi-square statistic.

Each client has a feature matrix X_i represent the data associated with top 50 features selected from chi-square statistics. Let's denote this matrix as

$$X_i \in R^{m_i \times 50}$$

m_i is the number of samples in dataset of client i.

X_i contains data for 50 features selected based on their chi-square statistic.

Step 2: Dimensionality Reduction using Liner Discriminator analysis (LDA)

After the initial phase of data processing, Linear Discriminant Analysis (LDA) [33] is implemented to further refine the dataset by reducing the feature space to the five most discriminative features. The strength of LDA lies in its ability to enhance the distinctions among different data groups, focusing on the most critical and unique attributes, thereby indirectly addressing disparities among classes. Despite this refinement, there remains the potential for some classes to be underrepresented.

LDA focuses on finding a projection that maximizes the ratio of the between-class scatter to the within-class scatter. For each client, define Equations (2) and (3):

Within -class scatter matrix ($S_{w,i}$):

$$S_{w,i} = \sum\nolimits_{c=1}^{C} N_{c,i}(\mu_{c,i} - \mu_i)(\mu_{c,i} - \mu_i)^T \tag{2}$$

Between -class scatter matrix ($S_{B,i}$)

$$S_{B,i} = \sum\nolimits_{c=1}^{C} N_{c,i}(\mu_{c,i} - \mu_i)(\mu_{c,i} - \mu_i)^T \tag{3}$$

where

$\mu_{c,i}$ is the mean vector of class c in client i's dataset $D_{c,i}$.

$N_{c,i}$ is the number of samples of class c at client i.

μ_i is the overall mean of the dataset at client i.

The goal of LDA is to find the project matrix with the maximize the ratio of the determine of the between-class scatter matrix to the within class scatter matrix. The formula is expressed as follow (Equation (4))

$$W_i = argmax_w \frac{\left|W^T S_{B,i} W\right|}{\left|W^T S_{w,i} W\right|} \tag{4}$$

The column of W_i are the eigen vectors corresponding to the largest eigenvalues of $S_{w,i}^{-1} S_{B,i}$. Typically, you choose as many eigen vectors as there are classes minus one.

Using the project matrix W_i, the data X_i is projected in to lower-dimensional space Y_i as define with equation (5):

$$Y_i = X_i W_i \tag{5}$$

where $Y_i \in R^{mi \times (c-1)}$ is the projected data matrix with dimension reduce to number of classes minus one.

Step 3: Validation using random forest

TO ensure the effectiveness of the selected features, each client applies a Random Forest (RF) classifier to the reduced dataset Y_i. The performance is monitored through the classifier's accuracy, confirming the relevance and robustness of the selected features.

2) FEDERATED LOCAL REBALANCING USING THE DENSITY-SENSITIVE SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

Following the initial phase of data processing, the procedure to rectify class imbalances begins with the application of a technique called DS2MOTE. This approach includes creating synthetic data points to reinforce categories that are not adequately represented. It assesses both the amount of data that is lacking and how this data is spread throughout the dataset, aiming to efficiently re-establish balance.

DS2MOTE extends the principles of the Synthetic Minority Over-sampling Technique (SMOTE), aimed at remedying class imbalances through the generation of new data points. It specifically targets minority classes, enriching them by interpolating synthetic points between existing samples. This approach proves invaluable in scenarios where some classes are notably sparse in instances. By plotting new instances along the vectors joining neighboring data points, DS2MOTE introduces additional samples to these underrepresented classes. SMOTE distinguishes itself from traditional augmentation methods by operating in the feature space rather than the data space, employing the k-nearest neighbors (KNN) algorithm to do so. It calculates the nearest neighbors for each point in the minority class, then synthesizes new samples within the feature space. This strategy enhances the dataset's diversity and addresses the challenge of class imbalance. (Equation 6):

$$\forall Y_i \in D_{min};\ return\{KNN\} \tag{6}$$

The implementation of this technique facilitates a more balanced class distribution within the dataset, thereby reducing the risk of biases that can occur from an overrepresentation of the majority class. The Density-Sensitive Synthetic Minority Over-Sampling Technique (DS2MOTE) is an adaptation of the original SMOTE algorithm, incorporating an analysis of the areas where classes overlap. It not only considers the density of the minority class but also the spatial relationship between instances of the minority and majority classes when creating synthetic samples. This approach ensures that synthetic instances are added in zones where there is a higher likelihood of misclassification, significantly improving the model's ability to accurately differentiate between classes that have regions of overlap. The procedure for DS2MOTE is shown below:

1. First, the underrepresented class was identified and given the symbol $Y_{minority}$.

2. The density of each sample in the $Y_{minority}$ by dataset is then calculated using the average distance to the k-nearest neighbors (K_{nn}). After that, these densities are stored in an array known as $D_{density}$.

3. The goal is to identify and record the member(s) from the $Y_{maijority}$ for each sample. In essence we are looking to locate and document the member(s) from the $Y_{maijority}$ to every member in the $Y_{minority}[i]$ marking this information as $[i]$.

4. Depending on the density of the sample, the appropriate quantity of Y_{synth} is created.

5. To create synthetic follow these steps; for each instance, in the group randomly pick a closest match and label it as $Y_{nearest}$. Compute the difference in vectors between the minority sample and its nearest neighbor then adjust it by a value "r", between 0 and 1. To generate a sample combine this adjusted vector with the original instance to yield a new synthetic data point $Y_{synth}\,[j]$

6. The formula for generating the synthetic sample $X_{synth}\,[j]$ is expressed as follows using Equation (7):

$$Y_{synth}\,[j] = Y_{minority}\,[i] + r(Y_{nearest} - Y_{minority}[i]) \tag{7}$$

7. Continue steps 4 and 5 until the quantity difference between the classes is eliminated.

r is a random value between 0 and 1, chosen for each compositional to ensure diversity when generating synthetic data.

Algorithm 1 Federate Local Preprocessing and Local Rebalancing

```
Require: The Global model from central server
Ensure: The Client model trained on its local dataset
$K$ : total number of clients
$F$ : total number of featuers
For $i$ to $K$
    Load local dataset $D_i$
    X=[]
    F=[]
    # Perform chi-square statistic for feature selection
    For j=1 to F
    $\mathrm{X}_{\mathrm{ij}}^{2} = \sum_{\mathbf{j=1}}^{\mathbf{ki}} \frac{(\mathbf{O_{ij}-E_{ij}})^{2}}{\mathbf{E_{ij}}}$
    X.append$\mathrm{X}_{\mathrm{ij}}^{2}$
    End For
    $\mathrm{F}_i = \mathrm{sort}(\mathrm{X}_{\mathrm{ij}}^{2}, \mathrm{order} = \mathrm{densending})$
    $\mathrm{F}_{i,Top50} = \mathrm{F}_i[1:50]$
    $\mathrm{X}_i = \mathrm{D}_i[:, \mathrm{F}_{i,Top50}]$
    $\mathrm{W_i} = \mathrm{argmax_w} \frac{|\mathrm{W^T S_{B,i} W}|}{|\mathrm{W^T S_{w,i} W}|}$
    $\mathrm{Y_i} = \mathrm{X_i W_i}$
    //Rebalance
    If $\mathrm{Y_i}$ is not balance then
      Identify minority class in $\mathrm{Y_i}$ as $\mathrm{Y_{minority}}$
      Calculate $\mathrm{D_{density}}$ for $\mathrm{Y_{minority}}$ using the K-n nearest neighbors
      Determine the number of synthetic samples w $\mathrm{N_{synth}}$ needed
      $\mathrm{Y_{senth}} = []$
        For $\mathrm{y_i}$ in $\mathrm{Y_{minority}}$
          Find $\mathrm{Y_{nearest}}$, the k-nearest neighbors for $\mathrm{y_i}$ i in $\mathrm{Y_{minority}}$
            For j = 1 to $\mathrm{N_{synth}}$
              r = rand(0, 1)
              $\mathrm{Y_{synth}}\left[\mathrm{j}\right] = \mathrm{Y_{minority}}\,[\mathrm{i}] + \mathrm{r}* (\mathrm{Y_{nearest}} - \mathrm{Y_{minority}}[\mathrm{i}])$
            $\mathrm{Y_{synth}}$.append($\mathrm{Y_{synth}}$[j])
              End for
          End for
          $\mathrm{Y_{train}}$.append($\mathrm{Y_{synth}}$)
      End If
Local Model Z Train (Model, $\mathrm{Y_{train}}$)
Return Local Model
End For
```

i : index used to point to a specific sample within the subset $Y_{minority}$. This subset contains the samples from underrepresented class in the dataset post-LDA features. Each i is unique to a particular sample in $Y_{minority}$.

j : is used to reference synthetic samples generated for each element $Y_{minority}[i]$. This index is specifically utilized within a loop that creates these samples based on density calculations and the extent of class imbalance.

All the processes happening on the client side are explained in Algorithm 1.

B. SERVER SIDE: ADDRESSING INTER-CLIENT IMBALANCE

To tackle the challenge of client-specific imbalances, the framework introduces an innovative adaptive client selection strategy based on the concept of uncertainty symmetry (SU) as shown in algorithm 2.

1) ADDRESSING INTER-CLIENT IMBALANCE THROUGH SYMMETRIC UNCERTANTY WITH ADAPTIVE THERSHOULD

Server-side processing has a vital role in FL, particularly when the user's data does not follow an independent and identically distributed (iid) pattern. This process is critical in selecting the suitable users to participate. Hence, an essential tool in this selection process is known as Symmetry Uncertainty (SU), a metric that measures the similarities or differences in data distributions among clients. The formula for calculating SU is as follows in Equation (8):

$$SU\ (X, Y) = 2\frac{IG(X/Y)}{H\ (X)\, H(Y)} \tag{8}$$

Hence, in this specific context, $IG(X/Y)$ denotes the amount of information gain related to feature X, which is treated as an attribute that is unrelated to the class attribute Y. $H\ (Y)$ signifies the entropy of feature X, while $H(Y)$ represents the entropy of feature [34].

Symmetry Uncertainty at the server is used to choose clients whose data distribution is close or similar to a normal distribution. The goal of this process is to reduce the imbalance in class distribution and manage different data sizes. This step is completed by asking each client to send a brief overview of their class data to the server, presented as n_k.

This process ensures the privacy of client's data by excluding data points even if the n_k summary may not exactly match the clients actual class distribution.

The server utilizes an adaptive threshold to select the proper clients for participating in the last version of the global model. This process, in addition to the utilization of Symmetry Uncertainty (SU), is discussed in Section C.

2) SAFEGUARDING DATA PRIVACY AND AVOIDING DATA LEAKS

It is well known that protecting data privacy is at the top priority in the FLP-DS2MOTE-USA framework. There is a thorough set of measures to protect the data from being leak among the clients themselves or between clients and the server. The initial Federated Local Preprocessing (FLP) phase is essential to keep sensitive details or raw data of the client in personal device. Instead, data sent to the server are simply updates on the model, in particular adjustments to the weights after local training. This configuration could prevent any efforts to recreate or infer the original data from these changes. Moreover, while selecting the clients, the server does not receive individual data points rather it receives broad summaries about class distributions. These overviews are developed carefully to contribute to model improvement with protecting against any unauthorized attempts to access

Algorithm 2 Server-Side Client Selection Using Symmetry Uncertainty With Adaptive Threshold

```
Require: Data summaries from all clients
Ensure: Selection of suitable clients for FL process
Input:
        $K$: total number of clients
        N: set containing class distribution summaries from each client $\{n_1, n_2, \ldots, n_K\}$
        T: predefined quantile threshold for selection
Output:
        S: subset of selected clients
SU = []
S = []
For $k = 1 to K$
        $SU\,[k] = 2\frac{IG(X/Y)}{H(X)H(Y)}$
        SU.appendSU[j]
End For
$\tau_{SU} = Q(T, SU)$
For $k = 1$ to $K$
if $SU[k] < \tau_{SU}$
S.appendS[j]
Return S
```

sensitive data. Adopting these strategies helps the FLP-DS2MOTE-USA framework to ensure the privacy data of all parties in the FL cycle, creating a strong barrier against any potential leaks of clients' data.

3) ADPTIVE CLIENT SELECTION

The server uses a flexible threshold for the Symmetry Uncertainty (SU) metric to dynamically pick a group of clients that help achieve a more even global model. This method is aimed at reducing the imbalance between clients by choosing those with data distributions that are either already balanced or that enhance the diversity of the overall dataset necessary for the global model. The use of an adaptive threshold means the selection criteria can be adjusted based on the changing levels of imbalance and diversity among the clients as the learning progresses, ensuring the process remains responsive and effective throughout.

The adaptive there should is calculated using Equation (9):

$$\tau_{SU} = Q(T; SU_1, SU_2, \ldots, SU_K) \tag{9}$$

where:

Q is the quantile function.

SU_K is the symmetric uncertainty for the K client, indicates whether their data distribution is balanced or unbalanced.

T is a constant and a predefined selection used to choose the clients.

due to this formulation, the quantile evaluation of the symmetric uncertainty of all participating clients can be used to dynamically modify the threshold τ_{SU} in each selection cycle. Through doing this, it guarantees that the selection criteria will continue to be responsive and sensitive to any changes in the data's features over time, improving the learning process' effectiveness and relevance in a variety of data areas and scenarios.

IV. IMPLEMENTATION AND EVALUATION

The proposed methodology was evaluated using two distinct datasets: the Human Activity Recognition with Smartphones (UCI HAR) dataset and the HAR OpenPose. These experiments were performed on a computer equipped with an Intel(R) Core (TM) i7-10750H CPU and 24GB of RAM. For each dataset, the data was randomly split, allocating 70% for training and the remaining 30% for testing.

A. HUMAN ACTIVITY RECOGNITION WITH SMART PHONES DATASET (UCI HAR)

The HAR database was constructed from 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. This device captured data on both 3-axial linear acceleration and 3-axial angular velocity at a consistent rate of 50Hz. To ensure accurate labeling, the activities were captured on video. Subsequently, the gathered data was randomly split, allocating 70% for training purposes and the remaining 30% for testing. Six different activities are covered by the 561 unique features in the UCI HAR dataset.

B. HUMAN ACTIVITY RECOGNITION DATASET (OPENPOSE)

The dataset is produced by applying OpenPose to videos featuring people performing a range of activities, such as standing, walking, squatting, and jumping. OpenPose outputs the x and y coordinates of key points on the human body, like the nose and the right elbow, which are then used as input variables for a model designed to recognize activities. The dataset is organized into 37 columns, with 36 of them being input variables representing the x and y coordinates of these body keypoints, and the remaining column ("class") being the target label for prediction. The objective is to construct a dependable model capable of accurately identifying the activity being performed.

C. MODEL NETWORK

The neural network is designed with one input layer, three hidden layers, and a single output layer. The size of the input layer corresponds to the number of components derived from LDA. The three hidden layers have neuron counts of 256, 128, and 64, in that order. Batch normalization and dropout are applied following each layer, with dropout rates of 0.2 for the first two hidden layers and 0.1 for the third. ReLU is the activation function chosen for all hidden layers. The output layer uses a linear activation function for label classification. A learning rate of 0.001 is set, and the model is optimized using Stochastic Gradient Descent (SGD). Random search was utilized in the experiments to determine the hyperparameters of the model, such as batch size, number of epochs, and dropout rate, etc.

D. PERFORMANCE EVALUATION

Although accuracy is commonly the go-to metric for classification problems, indicating the proportion of correctly predicted instances out of the total, it can be misleading when dealing with imbalanced datasets. In such scenarios, the model may yield a high accuracy rate while being biased towards the more prevalent class and ignoring the less common one. To address this, a range of alternative evaluation metrics like accuracy, precision, recall, F1-score, mean square error, Matthews Correlation Coefficient (MCC) [35], confusion matrix are used. These additional metrics offer a more nuanced assessment of the model's effectiveness.

Accuracy (Equation 10): Accuracy is calculated as the number of correct predictions divided by the total number of predictions.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ of\ predictions} \tag{10}$$

Precision (Equation 11): Precision is the ratio of true positive predictions to the sum of true positive and false positive predictions.

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

Recall (Equation 12): Recall is the ratio of true positive predictions to the sum of true positive and false negative predictions.

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

F1-score (Equation 13): F1-score is the harmonic mean of precision and recall.

$$F1-score = 2\frac{Precision * Recall}{Precision + Recall} \tag{13}$$

Matthews Correlation Coefficient (MCC) (Equation 14): MCC is a measure of the quality of binary classifications, taking into account true and false positives and negatives.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{14}$$

False Positive Rate (FPR) (Equation 15): FPR is the ratio of false positive predictions to the sum of false positive and false negative predictions.

$$FPR = \frac{FP}{FP+FN} \tag{15}$$

TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

V. RESULT AND DISCUSSION

The FLP-DS2MOTE-USA algorithm significantly outperforms FedAvg, FedSgd, FedSmote, and FedNova in terms of efficiency when evaluated on the first dataset as shown in figure 2 (a). Its efficiency is demonstrated by reduced sizes of individual updates and lower total network overhead. Specifically, FLP-DS2MOTE-USA has an individual update size of 175896, which is considerably less than the 745240 reported for both FedSgd and FedAvg, and the 14904800 recorded for FedSmote and FedNova. Furthermore, the network overhead for FLP-DS2MOTE-USA is only 43974000, much lower than the overheads reported for FedSgd, FedAvg, FedSmote, and FedNova as depicted in the graph for the UCI HAR dataset. The notable differences in update size and network overhead clearly demonstrate the enhanced efficiency and optimization of the FLP-DS2MOTE-USA algorithm.

When assessing performance with OpenPose HAR in figure 2 (b), FLP-DS2MOTE-USA markedly surpasses FedAvg, FedSgd, FedSmote, and FedNova in terms of efficiency. It features considerably lower update sizes and reduced network overhead, as illustrated in the data visualization. These results underscore the efficiency and efficacy of FLP-DS2MOTE-USA in HAR scenarios

In summary, the FLP-DS2MOTE-USA algorithm significantly reduces computation time, network overhead, and system complexity. The implementation of Chi-square and Linear Discriminant Analysis for data preprocessing significantly enhances the computational efficiency. This is particularly advantageous for IoT devices operating in FL scenarios that handle datasets with a vast array of features. By condensing the feature set, we reduce data processing requirements, which translates to lower computational load and energy usage. Additionally, our approach of targeted synthetic data generation and selective model updates minimizes unnecessary computation, optimizing energy efficiency. These strategies ensure that resource-constrained devices, such as those used in IoT applications, can operate effectively with our framework. However, it's important to acknowledge that while our framework demonstrates substantial advantages for complex datasets, its benefits are comparatively reduced for smaller datasets with fewer features, where simpler methods may suffice.

In Figure 3 (a), the efficacy of five distinct algorithms—FLP-DS2MOTE-USA, FedAvg, FedSgd, FedSmote, and FedNova—is assessed on the UCI HAR dataset using metrics like Accuracy, Recall, Precision, F1 Score, and MCC. The results highlight FLP-DS2MOTE-USA's dominance, particularly in Precision and MCC, showcasing its superior capability in diminishing classification mistakes and accurately detecting class instances within HAR.

For OpenPose HAR, FLP-DS2MOTE-USA continues its exemplary performance, closely mirroring the score achieved in UCI HAR, as shown in Figure 3(b).

The FLP-DS2MOTE-USA technique is highly effective for HAR tasks, particularly with imbalanced datasets.

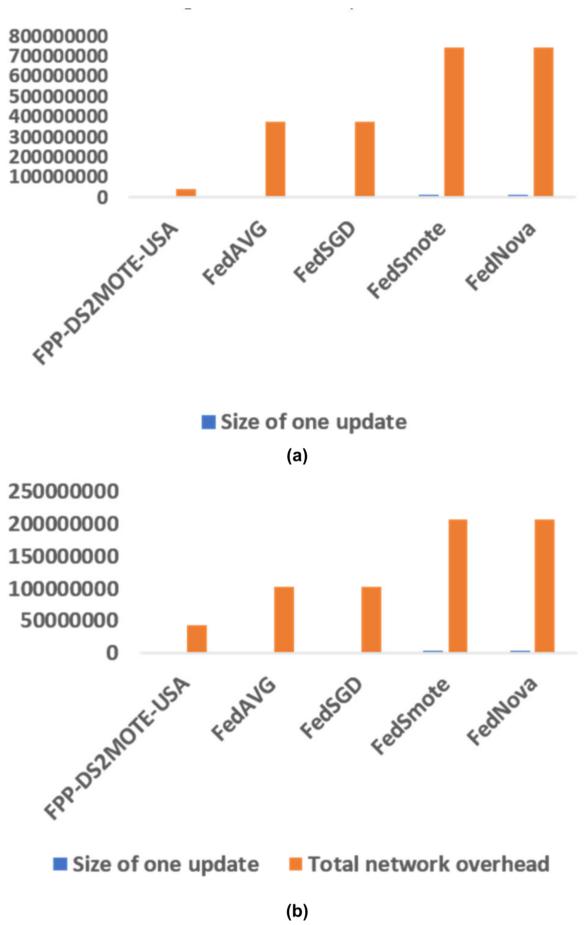


FIGURE 2. Comparison of the update size and network overhead between the suggested framework and the state-of-the-art FL algorithm (a) UCI HAR and (b) OpenPose HAR.

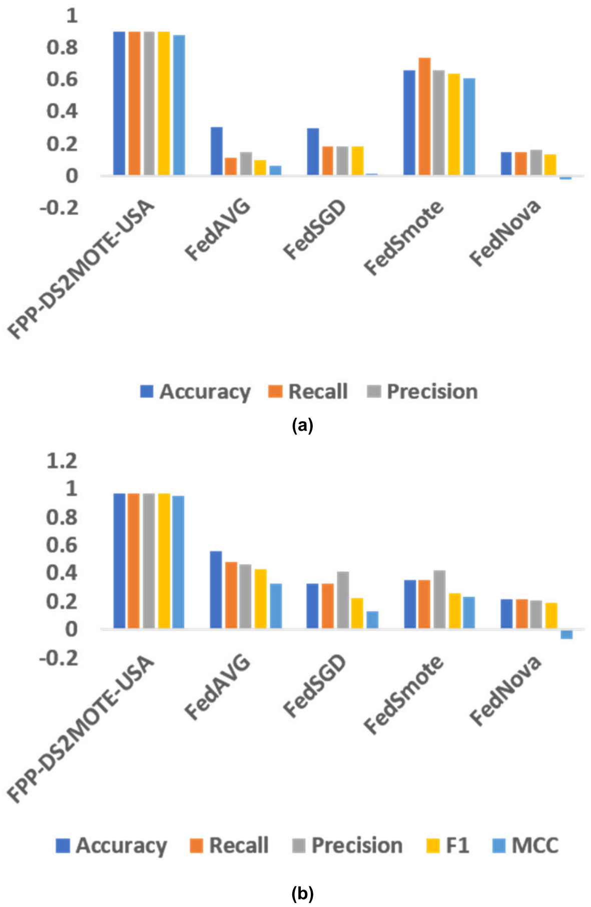


FIGURE 3. Performance metrics (Accuracy, Recall, Precision, F1, MCC) comparison of FLP-DS2MOTE-USA, FedAvg, FedSgd, FedSmote, and FedNova on UCI and OpenPose HAR Dataset.

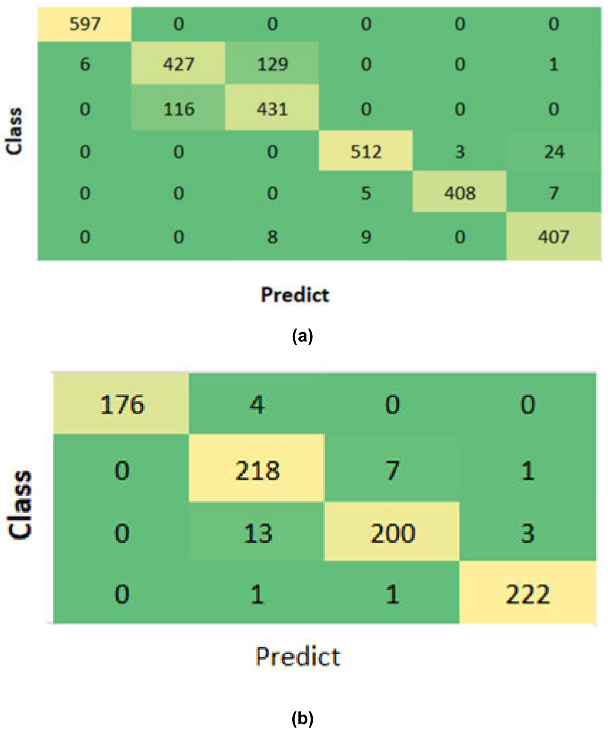
Its accuracy and unbiased classification make it suitable for real-world applications.

The lower scores of FedAvg, FedSGD, FedSMOTE, and FedNova can be attributed to their inability to adequately address the issue of data imbalance within the dataset. FedAvg and FedSGD are basic federated learning algorithms that do not incorporate any mechanisms to handle class imbalance. FedSMOTE and FedNova cannot address all types of imbalanced data as mentioned previously. As a result, their models tend to be biased towards the majority class, leading to lower performance metrics such as Precision and MCC, especially in datasets where certain activities are underrepresented.

The framework FLP-DS2MOTE-USA shows its effectiveness in dealing with class imbalance, as evidenced by the balanced results shown in the confusion matrix in Figure 4. On the other hand, the confusion matrices for the four algorithms (listed in the Appendix) reveal various misclassifications, suggesting skewed data distributions and potential challenges in handling complex feature spaces.

In FL, increasing the number of clients significantly impacts communication costs, computational demands, and learning phases. Figure 5 illustrate that as the number of clients grows, the average time per round and the total time for 50 rounds rise, reflecting heightened communication overhead and greater computational load. The UCI HAR dataset shows steeper increases compared to OpenPose, indicating higher complexity. These factors collectively result in longer training periods and delayed model convergence, highlighting the need to balance the number of clients with system efficiency to optimize performance and resource utilization. This indirectly affects the accuracy and loss of the model because the increase in the number of clients causes increased noise in the global model update, as shown in Figure 6.

Examining the five algorithms via the provided bar chart, it's clear that the proposed method surpasses both FedAvg and FedSgd. These two methods struggle because they fail to properly manage imbalances within or between classes. Conversely, while FedSmote and FedNova attempt to rectify imbalances by generating synthetic samples, their approach is insufficient. They neglect crucial factors like the distribution density of the minority classes and the closeness of minority to majority classes. Such oversights can lead to additional complications, ultimately diminishing their performance relative to the Proposed method. Figure 7 distinctly demonstrate the Proposed method's exceptional ability to achieve low



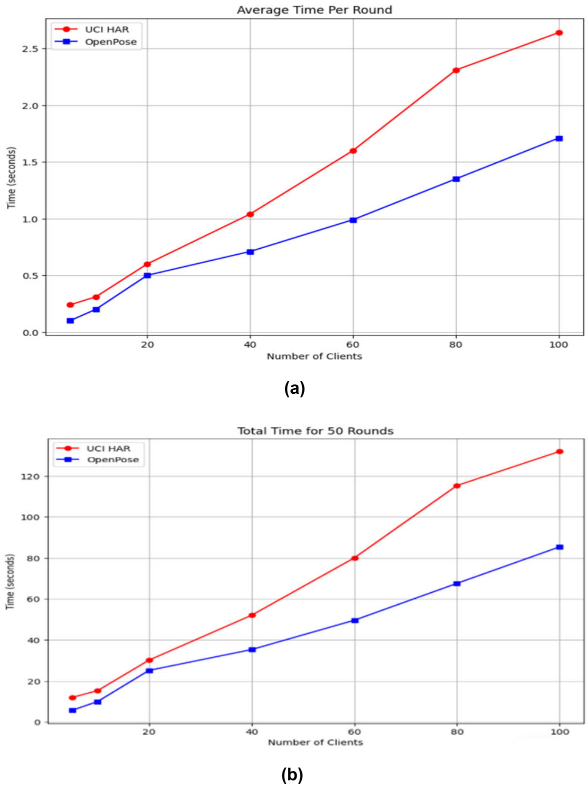
597	0	0	0	0	0
6	427	129	0	0	1
0	116	431	0	0	0
0	0	0	512	3	24
0	0	0	5	408	7
0	0	8	9	0	407

(a)

176	4	0	0
0	218	7	1
0	13	200	3
0	1	1	222

(b)

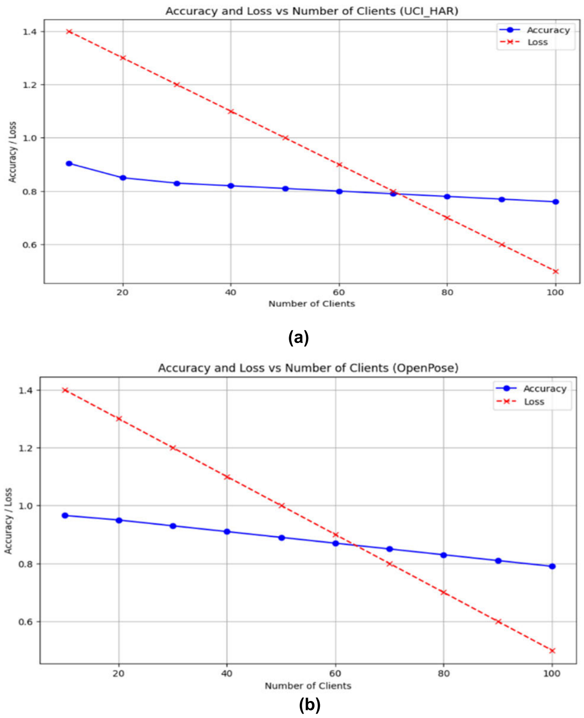
FIGURE 4. Comparison confusion matrix for FLP-DS2MOTE-USA: (a) UCI HAR and (b) OpenPose HAR.



(a)

(b)

FIGURE 5. The relationship between the average time per round and the total time with the number of customers (a) UCI HAR and (b) OpenPose HAR.



(a)

(b)

FIGURE 6. Variation in accuracy and loss with increasing number of clients: A comparative study for (a) UCI HAR and (b) OpenPose.

False Positive Rates across various classes, highlighting its robustness in handling imbalances.

For the UCI HAR dataset, the FLP-DS2MOTE-USA framework's proposed method stands out with an accuracy of 90.03%, surpassing other techniques such as SVD-LDA (88.187%), SVD (59.773%), Chi-PCA (62.10%), Chi-ICA (58.03%), and Chi-sparse PCA (55.24%) (Table 2). This underscores the proposed method's efficiency in exploiting the dataset's features to achieve superior performance. The SVD-LDA combination, improving upon the performance of SVD alone, highlights the limited efficacy of SVD when used in isolation. Incorporating LDA with SVD, however, significantly enhances the algorithm's capacity for dimensionality reduction while preserving key discriminative attributes, emphasizing the value of integrating various feature selection and reduction strategies for maximum effectiveness. Conversely, the HAR (OpenPose) dataset presents an alternative scenario, with the proposed method attaining a 96.58% accuracy, closely tailed by SVD-LDA at 96.22%, and followed by Chi-sparse PCA at 93.99%, Chi-PCA at 90.57%, SVD at 86.55%, and Chi-ICA at 95.17% (Table 2). The distinct nature and characteristics of this dataset may make it better suited to these particular feature selection and reduction approaches. Despite SVD's reduced efficiency on its own, its combination with LDA significantly bolsters the algorithm's performance, indicating that the right mix of SVD and LDA can greatly improve the model's accuracy and classification capabilities. This showcases the proposed method's versatility and effectiveness across various datasets, further solidifying its status as a robust and adaptable solution.

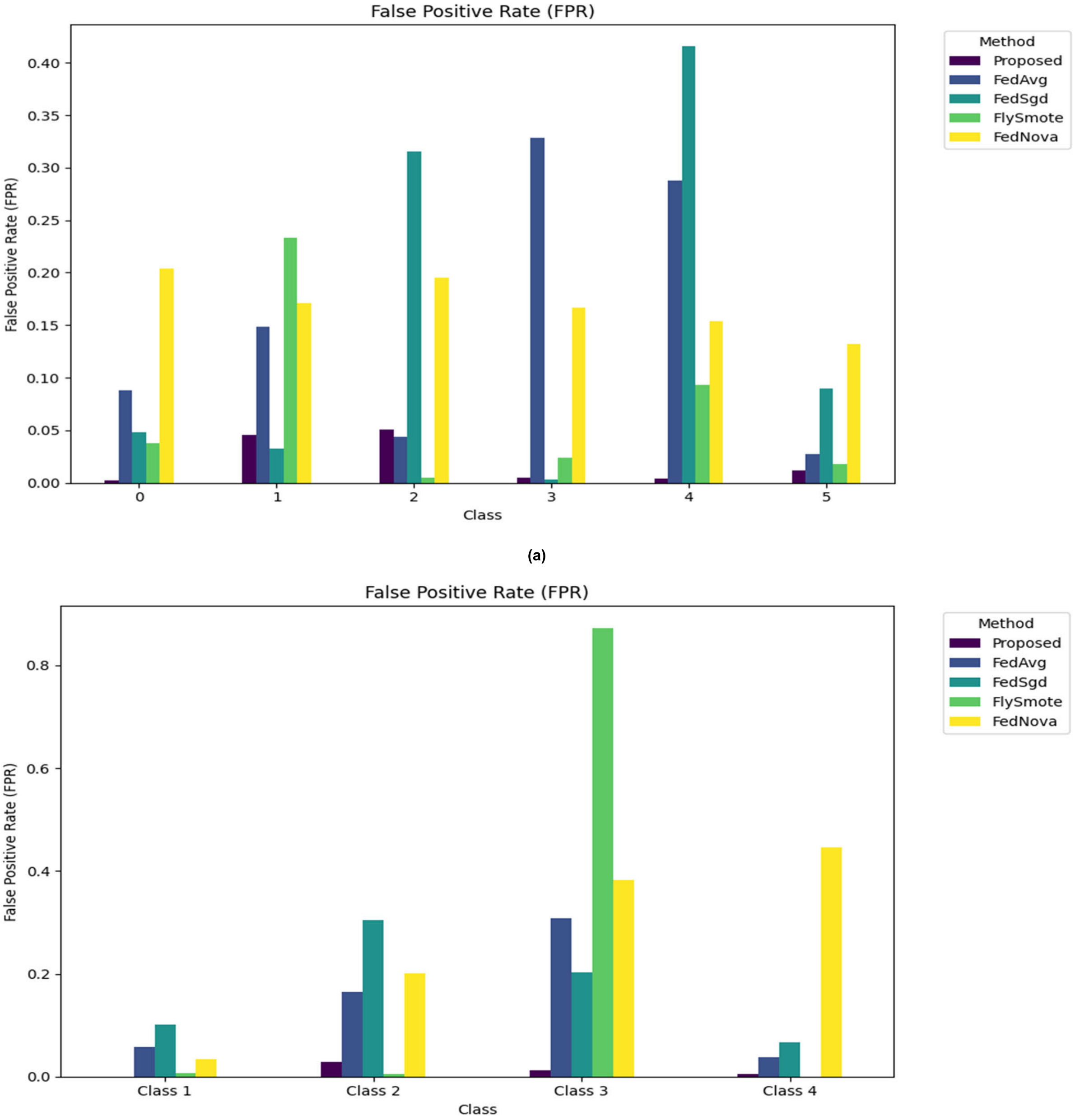


FIGURE 7. Heatmap comparison of proposed approaches with state-of-the-art fl algorithms for (a) UCI HAR and (b) OpenPose HAR .

VI. REAL-WORLD APPLICATION

The FLP-DS2MOTE-USA algorithm marks a considerable breakthrough in Human Activity Recognition (HAR), establishing unprecedented performance standards vital for real-world applications. Based on our research, this algorithm achieves reductions in individual update sizes and network overhead by over 88% and 78%, respectively, surpassing existing federated learning methods such as FedAvg, FedSgd, FedSmote, and FedNova. These significant improvements make the FLP-DS2MOTE-USA particularly advantageous for IoT environments, where it is essential to limit both computational burden and network traffic. Industries like healthcare and smart home technology are poised to gain enormously from adopting this more accurate, energy-efficient, and effective activity recognition technology. Moreover, the algorithm contributes to enhanced operational efficiency and promotes environmental sustainability within tech-driven sectors by significantly cutting down on energy usage and operational expenses.

VII. ENHANCED PERFORMANCE ANALYSIS OF FLP-DS2MOTE-USA

The inherent variability of non-iid. data across clients presents significant challenges to model efficacy in FL. Our

TABLE 2. The FLP-DS2MOTE-USA framework reveals significant differences in accuracy when various feature selection and reduction methods are applied to two HAR datasets.

Dataset	Chi-LDA	SVD-LDA	SVD	Chi-PCA	Chi-ICA	Chi-sparce PCA
UCI HAR	90.03	88.187	59.773	62.10	58.03	55.24
HAR (OpenPose)	96.58	96.22	86.55	95.17	90.57	93.99

innovative FLP-DS2MOTE-USA framework effectively navigates these issues, setting it apart from traditional methods such as FedAvg, FedSgd, FedSMOTE, and FedNova. These conventional approaches often fail to fully account for the profound effects of non-iid. data, consequently affecting their overall effectiveness. Conversely, our framework's federated local preprocessing (FLP) actively addresses these variances at their origin. This preprocessing not only enhances the quality of feature extraction but also indirectly ameliorates class imbalances, creating a more equitable and representative dataset for subsequent model training. This proactive stance against non-iid data variances is a fundamental reason for our model's enhanced performance.

Within our framework, we introduce an optimized version of the SMOTE algorithm, termed DS2MOTE, which is specifically adapted for federated contexts. DS2MOTE diverges from traditional models by accurately assessing the density between majority and minority classes to avert overfitting—a common shortfall in approaches like FedAvg and FedSgd. This deliberate recalibration ensures our model improvements are substantial, concentrating on essential features and reducing unnecessary complexity. This method effectively tackles the overfitting problems that FL models often suffer from.

VIII. CONCLUSION

This paper proposed FLP-DS2MOTE-USA framework as an approach to overcome the challenges of dataset imbalance and localized data preprocessing. Two different datasets: UCI_HAR and OpenPose HAR have been used to test this model. The model has been compared with other algorithms such as FedAvg, FedSgd, FedSmote, and FedNova, the model achieved impressive accuracies of 90.57% and 96.58%, respectively. Particularly, managing model combination in scenarios plagued by data imbalance and maintaining robust generalization had been performed by demonstrated exceptional courage of framework. Moreover, the experiments shed light on selecting optimal learning architectures for accurate HAR.

This research addressed several key open questions in the context of FL for HAR. The issue of data imbalance was managed by introducing the DS2MOTE to create synthetic data points for underrepresented classes. This approach balanced the data within client datasets, leading to significant improvements in model performance and accuracy. For optimal feature selection and dimensionality reduction, Chi-square and Linear Discriminant Analysis (LDA) were employed, helping identify pertinent features and simplify data structures, thus enhancing the efficiency and accuracy of the FL models. Additionally, symmetric uncertainty with adjustable thresholds was used for client selection, managing variability in the size and distribution of client datasets, ensuring balanced contributions from clients and improving the robustness and performance of the federated model.

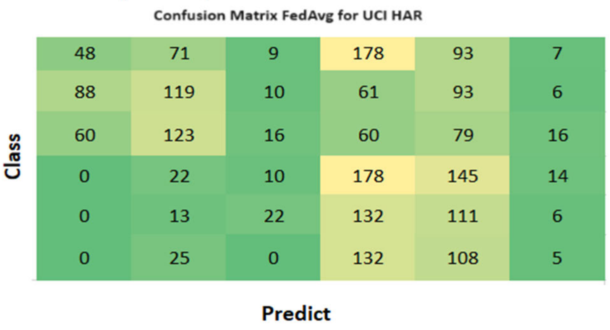
The commitment to privacy protection is a key aspect of the FLP-DS2MOTE-USA model, that ensuring the sensitive client data remains securely on the local device. This approach preserves dataset balance in a FL context, also protects data privacy and promoting effective learning.

In light of the challenges of dataset imbalance and local data preprocessing within FL environments, the development of FLP-DS2MOTE-USA represents a crucial step forward. The framework not only increased accuracy but also maintained strict privacy standards, making it particularly useful in sensitive applications. However, exploring FL is an ongoing journey with many open research questions that need to be addressed to improve and enhance the applicability of such frameworks. Future research will focus on expanding the model's adaptability to handle different forms of non-independent matching distributions, such as feature skewness and label skewness, which are common in real-world scenarios. Additionally, investigating scalability and computational efficiency will be crucial to ensure that FL can be practically implemented across more diverse datasets.

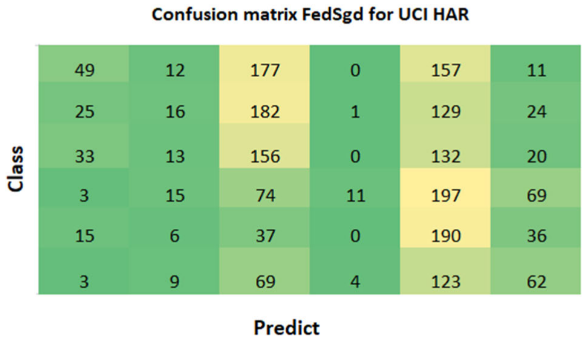
By addressing these critical questions, future developments can provide more robust, efficient, and practically applicable solutions, further advancing the field of federated learning for human activity recognition.

APPENDIX A

The confusion matrix of FedAvg, FedSgd FedSmote and FedNova when applied to the UCI_HAR and OpenPose HAR data sets respectively.



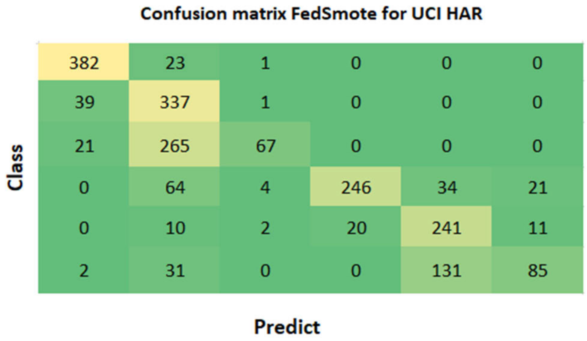
48	71	9	178	93	7
88	119	10	61	93	6
60	123	16	60	79	16
0	22	10	178	145	14
0	13	22	132	111	6
0	25	0	132	108	5



Confusion matrix FedSgd for UCI HAR

49	12	177	0	157	11
25	16	182	1	129	24
33	13	156	0	132	20
3	15	74	11	197	69
15	6	37	0	190	36
3	9	69	4	123	62

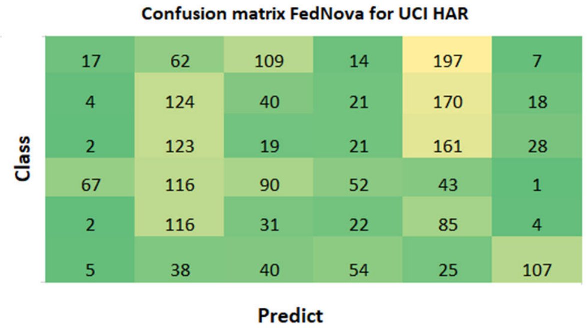
Class
Predict



Confusion matrix FedSmote for UCI HAR

382	23	1	0	0	0
39	337	1	0	0	0
21	265	67	0	0	0
0	64	4	246	34	21
0	10	2	20	241	11
2	31	0	0	131	85

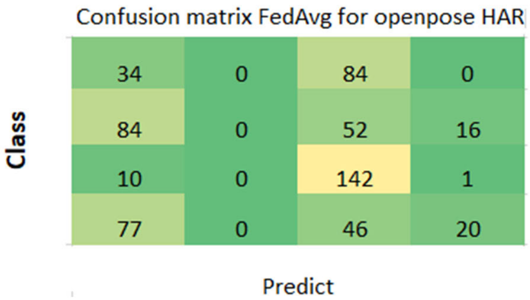
Class
Predict



Confusion matrix FedNova for UCI HAR

17	62	109	14	197	7
4	124	40	21	170	18
2	123	19	21	161	28
67	116	90	52	43	1
2	116	31	22	85	4
5	38	40	54	25	107

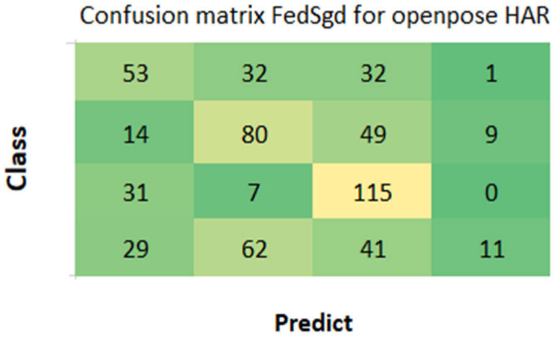
Class
Predict



Confusion matrix FedAvg for openpose HAR

34	0	84	0
84	0	52	16
10	0	142	1
77	0	46	20

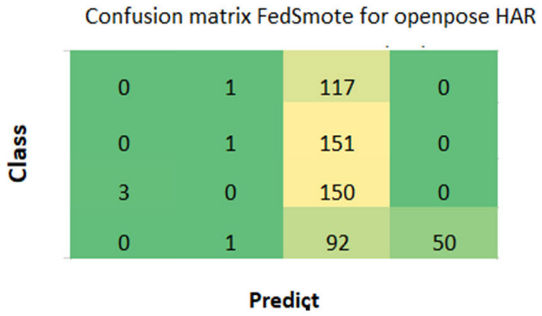
Class
Predict



Confusion matrix FedSgd for openpose HAR

53	32	32	1
14	80	49	9
31	7	115	0
29	62	41	11

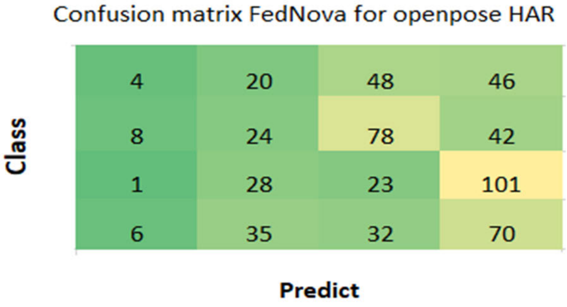
Class
Predict



Confusion matrix FedSmote for openpose HAR

0	1	117	0
0	1	151	0
3	0	150	0
0	1	92	50

Class
Predict



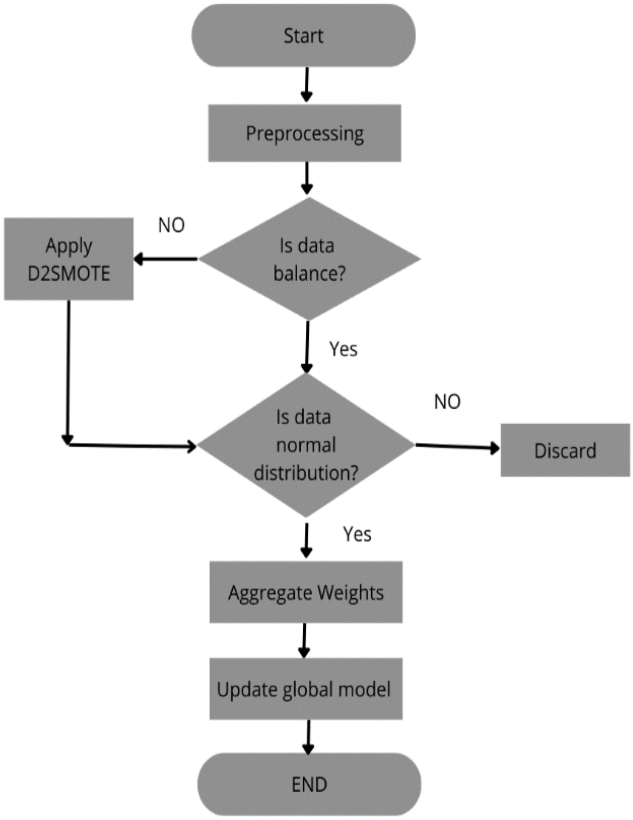
Confusion matrix FedNova for openpose HAR

4	20	48	46
8	24	78	42
1	28	23	101
6	35	32	70

Class
Predict

APPENDIX B

The flowchart details the preprocessing steps for balancing and normalizing data before updating the global model.



REFERENCES

[1] R. Presotto, G. Civitarese, and C. Bettini, ''Federated clustering and semi-supervised learning: A new partnership for personalized human activity recognition,'' *Pervas. Mobile Comput.*, vol. 88, Jan. 2023, Art. no. 101726, doi: 10.1016/j.pmcj.2022.101726.

[2] R. Younis and M. Fisichella, ''FLY-SMOTE: Re-balancing the non-IID IoT edge devices data in federated learning system,'' *IEEE Access*, vol. 10, pp. 65092–65102, 2022, doi: 10.1109/ACCESS.2022.3184309.

[3] M. Seol and T. Kim, ''Performance enhancement in federated learning by reducing class imbalance of non-IID data,'' *Sensors*, vol. 23, no. 3, p. 1152, Jan. 2023, doi: 10.3390/s23031152.

[4] P. Cassará, A. Gotta, and L. Valerio, ''Federated feature selection for cyber-physical systems of systems,'' *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9937–9950, Sep. 2022, doi: 10.1109/TVT.2022.3178612.

[5] Z. K. Taha, C. T. Yaw, S. P. Koh, S. K. Tiong, K. Kadirgama, F. Benedict, J. D. Tan, and Y. A. Balasubramaniam, ''A survey of federated learning from data perspective in the healthcare domain: Challenges, methods, and future directions,'' *IEEE Access*, vol. 11, pp. 45711–45735, 2023, doi: 10.1109/ACCESS.2023.3267964.

[6] B. Gong, T. Xing, Z. Liu, W. Xi, and X. Chen, ''Adaptive client clustering for efficient federated learning over non-IID and imbalanced data,'' *IEEE Trans. Big Data*, early access, Apr. 19, 2022, doi: 10.1109/TBDATA.2022.3167994.

[7] T. Li, A. Kumar Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, ''Federated optimization in heterogeneous networks,'' 2018, *arXiv:1812.06127*.

[8] R. Zhang, H. Li, M. Hao, H. Chen, and Y. Zhang, ''Secure feature selection for vertical federated learning in eHealth systems,'' in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 1257–1262, doi: 10.1109/ICC45855.2022.9838917.

[9] M. S. Nashwan and S. Shahid, ''Symmetrical uncertainty and random forest for the evaluation of gridded precipitation and temperature data,'' *Atmos. Res.*, vol. 230, Dec. 2019, Art. no. 104632, doi: 10.1016/j.atmosres.2019.104632.

[10] M. Yang, X. Wang, H. Zhu, H. Wang, and H. Qian, ''Federated learning with class imbalance reduction,'' in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 2174–2178, doi: 10.23919/EUSIPCO54536.2021.9616052.

[11] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, and B. Zhao, ''A federated learning system with enhanced feature extraction for human activity recognition,'' *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107338, doi: 10.1016/j.knosys.2021.107338.

[12] A. Tabassum, A. Erbad, W. Lebda, A. Mohamed, and M. Guizani, ''FEDGAN-IDS: Privacy-preserving IDS using GAN and federated learning,'' *Comput. Commun.*, vol. 192, pp. 299–310, Aug. 2022, doi: 10.1016/j.comcom.2022.06.015.

[13] A. N. Khan, A. Rizwan, R. Ahmad, Q. W. Khan, S. Lim, and D. H. Kim, ''A precision-centric approach to overcoming data imbalance and non-IIDness in federated learning,'' *Internet Things*, vol. 23, Oct. 2023, Art. no. 100890, doi: 10.1016/j.iot.2023.100890.

[14] A. M. Alhassan and W. M. N. Wan Zainon, ''Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis,'' *IEEE Access*, vol. 9, pp. 87310–87317, 2021, doi: 10.1109/ACCESS.2021.3088613.

[15] Y. Qin and M. Kondo, ''Federated learning-based network intrusion detection with a feature selection approach,'' in *Proc. Int. Conf. Electr., Commun., Comput. Eng. (ICECCE)*, Jun. 2021, pp. 1–6, doi: 10.1109/ICECCE52056.2021.9514222.

[16] A. E. Ouadrhiri, A. Abdelhadi, and P. H. Phung, ''Hensel's compression-based dimensionality reduction approach for privacy protection in federated learning,'' in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2023, pp. 298–303, doi: 10.1109/ICNC57223.2023.10074197.

[17] Y.-M. Cheung, J. Jiang, F. Yu, and J. Lou, ''Vertical federated principal component analysis and its kernel extension on feature-wise distributed data,'' 2022, *arXiv:2203.01752*.

[18] W. Huang and A. S. Barnard, ''Federated data processing and learning for collaboration in the physical sciences,'' *Mach. Learn. Sci. Technol.*, vol. 3, no. 4, Dec. 2022, Art. no. 045023, doi: 10.1088/2632-2153/aca87c.

[19] A. A. Abdellatif, N. Mhaisen, A. Mohamed, A. Erbad, M. Guizani, Z. Dawy, and W. Nasreddine, ''Communication-efficient hierarchical federated learning for IoT heterogeneous systems with imbalanced data,'' *Future Gener. Comput. Syst.*, vol. 128, pp. 406–419, Mar. 2022, doi: 10.1016/j.future.2021.10.016.

[20] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, ''Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results,'' *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101765, doi: 10.1016/j.media.2020.101765.

[21] T. Cui, Y. Shi, B. Lv, R. Ding, and X. Li, ''Federated learning with SARIMA-based clustering for carbon emission prediction,'' *J. Cleaner Prod.*, vol. 426, Nov. 2023, Art. no. 139069, doi: 10.1016/j.jclepro.2023.139069.

[22] J. C.-H. Wu, H.-W. Yu, T.-H. Tsai, and H. H.-S. Lu, ''Dynamically synthetic images for federated learning of medical images,'' *Comput. Methods Programs Biomed.*, vol. 242, Dec. 2023, Art. no. 107845, doi: 10.1016/j.cmpb.2023.107845.

[23] Q. Abbas, K. M. Malik, A. K. J. Saudagar, and M. B. Khan, ''Context-aggregator: An approach of loss- and class imbalance-aware aggregation in federated learning,'' *Comput. Biol. Med.*, vol. 163, Sep. 2023, Art. no. 107167, doi: 10.1016/j.compbiomed.2023.107167.

[24] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, ''FedHome: Cloud-edge based personalized federated learning for in-home health monitoring,'' *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2818–2832, Aug. 2022, doi: 10.1109/TMC.2020.3045266.

[25] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad, ''Federated learning for COVID-19 screening from chest X-ray images,'' *Appl. Soft Comput.*, vol. 106, Jul. 2021, Art. no. 107330, doi: 10.1016/j.asoc.2021.107330.

[26] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, ''Self-balancing federated learning with global imbalanced data in mobile systems,'' *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 1, pp. 59–71, Jan. 2021, doi: 10.1109/TPDS.2020.3009406.

[27] U. Michieli and M. Ozay, ''Are all users treated fairly in federated learning systems?'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2318–2322, doi: 10.1109/CVPRW53098.2021.00263.

[28] J. Xu, Y. Yan, and S.-L. Huang, ''FedPer++: Toward improved personalized federated learning on heterogeneous and imbalanced data,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8, doi: 10.1109/IJCNN55064.2022.9892585.

[29] S. Chen, Z. Jie, G. Wang, K.-C. Li, J. Yang, and X. Liu, ''A new federated learning-based wireless communication and client scheduling solution for combating COVID-19,'' *Comput. Commun.*, vol. 206, pp. 101–109, Jun. 2023, doi: 10.1016/j.comcom.2023.04.023.

[30] X. Yang, B. Xiong, Y. Huang, and C. Xu, ''Cross-modal federated human activity recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5345–5361, Aug. 2024, doi: 10.1109/tpami.2024.3367412.

[31] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, ''A chi-square statistics based feature selection,'' in *Proc. IEEE 9th Int. Conf. Softw. Eng. Serv. Sci.*, Nov. 2018, pp. 160–163.

[32] S. D. Bolboacă, L. Jäntschi, A. F. Sestraş, R. E. Sestraş, and D. C. Pamfil, ''Pearson-Fisher chi-square statistic revisited,'' *Information*, vol. 2, no. 3, pp. 528–545, Sep. 2011, doi: 10.3390/info2030528.

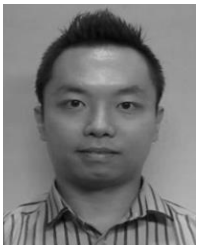
[33] H. Zhao, Z. Wang, and F. Nie, ''A new formulation of linear discriminant analysis for robust dimensionality reduction,'' *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 629–640, Apr. 2019, doi: 10.1109/TKDE.2018.2842023.

[34] S. I. Ali and W. Shahzad, ''A feature subset selection method based on symmetric uncertainty and ant colony optimization,'' in *Proc. Int. Conf. Emerg. Technol.*, Oct. 2012, pp. 1–6, doi: 10.1109/ICET.2012.6375420.

[35] D. Chicco, M. J. Warrens, and G. Jurman, ''The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and brier score in binary classification assessment,'' *IEEE Access*, vol. 9, pp. 78368–78381, 2021, doi: 10.1109/ACCESS.2021.3084050.

ZAHRAA KHDUAIR TAHA received the B.Sc. degree in electrical engineering and the M.Sc. degree in electronics and communication engineering from the University of Baghdad, Iraq, in 2008 and 2013, respectively. Since 2013, she has been a Lecturer with Al-Iraqia University. She is the author of more than 26 articles. Her current research interests include machine learning, deep learning, pattern recognition, the Internet of Things, data mining, and medical image processing.

JOHNNY KOH SIAW PAW received the bachelor's (Hons.), M.Sc., and Ph.D. degrees in electrical and electronic engineering from Universiti Putra Malaysia, in 2000, 2002, and 2008, respectively. He is currently a Professor with the Institute of Sustainable Energy, Universiti Tenaga Nasional. His research interests include machine intelligence, automation technology, and renewable energy.



YAW CHONG TAK received the bachelor's and master's degrees (Hons.) in electrical and electronics engineering and the Ph.D. degree in artificial neural networks from Universiti Tenaga Nasional (UNITEN), Malaysia, in 2008, 2012, and 2019, respectively. His research interests include artificial neural networks and renewable energy. He is currently a Postdoctoral Researcher with the Institute of Sustainable Energy, UNITEN.

TIONG SIEH KIONG (Senior Member, IEEE) received the B.Eng.(Hons.), MSc., and Ph.D. degrees in electrical, electronic and system engineering from the Nasional University of Malaysia (UKM), in 1997, 2000, and 2006, respectively. He is currently a Professor with the College of Engineering. He is also the Director of the Institute of Sustainable Energy (ISE), Universiti Tenaga Nasional. His research interests include renewable energy, artificial intelligence, data analytics, microcontroller systems, and communication systems. He is a Professional Engineer registered with the Board of Engineers Malaysia (BEM).

KUMARAN KADIRGAMA is currently a Professor and a Research Fellow of the Advanced Nano Coolant Lubricant Laboratory (ANCoL), Automotive Engineering Research Group, Faculty of Mechanical and Automotive Engineering Technology, University Malaysia Pahang (UMP). In research, he is involved in the supervision of postgraduate students at the master's and Ph.D. levels and has mentored numerous students to graduation. He has also published and presented various technical articles and journals at an international and national level. He has an H-index of 31 with 4199 citations. He has received grants totaling RM 6.02 million from various agencies and institutions. He has won gold medals in the International Invention, Innovation and Technology Exhibition (ITEX), Seoul International Invention Fair (SIIF), South Korea, and British Invention Show (BIS). He is a Professional Engineer registered with the Board of Engineers Malaysia (BEM) and a Chartered Engineer (U.K.) with the Institution of Mechanical Engineers (IMechE).

FOO BENEDICT received the two bachelor's degree in mechanical engineering from the University of Southern Queensland, Australia, and the Ph.D. degree in engineering from Universiti Malaysia Pahang. He also received a Graduate Certificate in mechanical engineering from the University of Southern Queensland. He is currently the Managing Director and the Founder of Enhance Track Sdn. Bhd. He has been managing the company for 17 years since its establishment, in 2005. As the Managing Director of Enhance Track Sdn. Bhd., he has vast experience in many industries, such as oil and gas, renewable energy, education, and laboratory and testing equipment. In this role, he has also personally been involved with all the research and development projects that the company has undertaken and improving the outcome for clients in Malaysia, Australia, and the Middle East. He is also active in associations and professional bodies, such as the Associated Chinese Chamber of Commerce and Industries of Malaysia (ACCCIM) and the Selangor and Kuala Lumpur Foundry and Engineering Association (SFEIA). He is also a Professional Technologist registered with the Malaysia Board of Technologists (MBOT).

TAN JIAN DING received the Ph.D. degree from the University of Malaya, Malaysia. He is currently an Assistant Professor with the School of Electrical Engineering and Artificial Intelligence, Xiamen University, Malaysia. He is a principal investigator of several ongoing research grants and projects. His research interests include electronics, artificial intelligence, renewable energy, robotics, control engineering, the Internet of Things, and soft computing. He is a registered member of the Board of Engineering Malaysia and the Institute of Engineering, Malaysia.

KHARUDIN ALI received the Diploma degree in industrial automation (mechatronics) and the bachelor's degree (Hons.) in engineering technology (mechatronics) from the TATI University College, in 2005 and 2012, respectively, the M.Sc. degree in mechatronic engineering from the University Malaysia Pahang, in 2017, and the Ph.D. degree from University Tenaga Nasional (UNITEN), Malaysia, in 2020. He is a Professional Technologies, in 2018. He has published a number of 33 papers in ISI and Scopes with Impact Journal and International Conferences and supervised around 85 graduate and two postgraduate (master's) students. He also publishes one book in Eddy current inspection. His current research interests include non-destructive testing, sensor system design, embedded systems, the IoT, and artificial intelligence. He has vast experience in industrial, teaching, and training, since 2004, in industrial includes ESCATEC Mechatronics, and TATI University College. The author also receives a few awards, including the Best Paper Award at the ICON 2013 Conference and IGRAD 2018 Conference and the Excellent Worker Award from TATI University College, in 2012 and 2018.

AZHER M. ABED received the Ph.D. degree in renewable energy from Universiti Kebangsaan Malaysia (UKM). He is currently a Mechanical Engineer. He serves as the Dean of the College of Engineering and Technologies. His remarkable academic journey encompasses over 200 publications in peer-reviewed journals, showcasing expertise in renewable energy, heat transfer, and fluid flows, thereby significantly advancing these pivotal fields. His dedication to practical solutions is evident in his patented solar-assisted cooling system. Through research, publications, and inventions, he continues to drive advancements and promote sustainability in the realm of renewable energy and mechanical engineering. His noteworthy contributions have earned him recognition, including an Energy Efficiency Award from the World Society of Sustainable Energy Technologies (WSSET) for his work on absorption cooling systems.

• • •