A Comparison of Sampling Methods for Dealing with Imbalanced Wearable Sensor Data in Human Activity Recognition using Deep Learning

Mariam El Ghazi*, Noura Aknin
Information Technology and Modeling, Systems Research Unit, Abdelmalek Essaadi University, Tetouan, Morocco

***Abstract*—Human Activity Recognition (HAR) holds significant implications across diverse domains, including healthcare, sports analytics, and human-computer interaction. Deep learning models demonstrate great potential in HAR, but performance is often hindered by imbalanced datasets. This study investigates the impact of class imbalance on deep learning models in HAR and conducts a comprehensive comparative analysis of various sampling techniques to mitigate this issue. The experimentation involves the PAMAP2 dataset, encompassing data collected from wearable sensors. The research includes four primary experiments. Initially, a performance baseline is established by training four deep-learning models on the imbalanced dataset. Subsequently, Synthetic Minority Over-sampling Technique (SMOTE), random under-sampling, and a hybrid sampling approach are employed to rebalance the dataset. In each experiment, Bayesian optimization is employed for hyperparameter tuning, optimizing model performance. The findings underscore the paramount importance of dataset balance, resulting in substantial improvements across critical performance metrics such as accuracy, F1 score, precision, and recall. Notably, the hybrid sampling technique, combining SMOTE and Random Undersampling, emerges as the most effective method, surpassing other approaches. This research contributes significantly to advancing the field of HAR, highlighting the necessity of addressing class imbalance in deep learning models. Furthermore, the results offer practical insights for the development of HAR systems, enhancing accuracy and reliability in real-world applications. Future works will explore alternative public datasets, more complex deep learning models, and diverse sampling techniques to further elevate the capabilities of HAR systems.**

Keywords*—Human activity recognition (HAR); class imbalance; sampling methods; wearable sensors; deep learning; synthetic minority over-sampling technique (SMOTE); random undersampling; PAMAP2 dataset; bayesian optimization

I. Introduction

Human Activity Recognition (HAR) is a multidisciplinary field focused on the automated identification and categorization of human activities, primarily relying on data collected from diverse sensors. Its applications extend into critical domains, particularly Sports or healthcare [1]. The automatic detection and classification of human activities can significantly improve the quality of life for elderly individuals and dependents, enhancing their safety, well-being, and independence [2].HAR systems play a vital role in smart home environments by providing context-aware services to residents, monitoring their activities, and alerting caregivers in case of any abnormal situations[3].

Deep learning models have revolutionized HAR due to their capacity to process and analyze sensor data effectively. Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) are two prominent deep learning architectures that excel at learning complex patterns and temporal dependencies from sensor data. These models have demonstrated remarkable performance in HAR, making them the focal point of this study [4] [5]. While deep learning models have shown promise in HAR, one significant challenge is posed by imbalanced datasets [6]. In many real-world scenarios, certain activities or classes are more frequent than others in the data, creating an imbalance. This imbalance can adversely affect the performance and accuracy of HAR models, as they may become biased towards the majority class, leading to poor recognition of minority activities [7].

In summary, Human Activity Recognition (HAR) holds vital implications for various domains. Deep learning models, such as LSTM and CNN, enhance HAR by effectively processing sensor data. The use of sampling techniques to address class imbalance significantly boosts model performance. This research underscores the importance of balanced datasets in HAR and provides practical insights for real-world applications.

The primary contributions of this study are as follows:

- The profound impact of class imbalance on the performance of deep learning models in HAR is investigated, and a range of sampling techniques designed to alleviate this issue is introduced and rigorously evaluated, offering valuable insights into enhancing model performance within imbalanced datasets.
- Three distinct sampling techniques are evaluated: SMOTE Random Undersampling, and a hybrid approach combining both methods.
- A detailed comparative analysis of the efficacy of these sampling methods in enhancing learning from imbalanced human activity data through deep machine learning algorithms is provided. Specifically, a test is conducted on Vanilla LSTM, 2 Stacked LSTM, 3 Stacked LSTM, and the Hybrid CNN-LSTM model.

The findings consistently demonstrate that the hybrid sampling techniques consistently outperform state-of-the-art models across critical performance metrics, including accuracy, precision, recall, and F1 score.

The paper is organized into distinct sections to effectively present the research findings. Section II presents the Related Works, reviewing prior research in the field related to the problem. Section III elaborates on the Materials and Methods employed in the experimental approach. Section IV presents the outcomes of the experiments and engages in a thorough discussion of the findings. Finally, in Section V, the Conclusion presents the key findings, and future directions of sensor -based HAR using deep learning models.

II. Related Work

In the field of Human Activity Recognition (HAR), addressing imbalanced data presents a significant challenge, a common issue observed in various public datasets, including Opportunity [8], WISDM V1.1[9], SPHERE [10] and PAMAP2 [11]. Imbalanced data can profoundly affect the performance of deep learning models utilized in HAR tasks. To tackle this challenge, several studies have explored the integration of deep learning models with sampling methods specifically designed for Human Activity Recognition based on sensor data.

Jeong et al. (2022) conducted a comprehensive study focusing on the influence of undersampling and oversampling techniques for classifying physical activities using an imbalanced accelerometer dataset. Their findings proposed that ensemble learning, coupled with well-defined feature sets and undersampling, exhibits robustness in the classification of physical activities within imbalanced datasets. This approach proves particularly effective in real-world scenarios, where imbalanced class distributions are commonplace. Furthermore, the study underscored the superiority of ensemble learning over other machine learning and deep learning models in handling small datasets with subject variability [12].

Hamad et al. (2020) evaluated the efficacy of imbalanced data handling methods in the context of deep learning applied to smart home environments. Leveraging a CNN LSTM model and a dataset comprising daily living activities collected from two real intelligent homes, their research demonstrated a significant performance improvement by applying the SMOTE oversampling method. This enhancement resulted in a notable increase in accuracy (from 0.60-0.62 to 0.71-0.73) when compared to training on the original imbalanced data [13].

Alani et al. (2020) delved into the classification of imbalanced multi-modal sensor data for HAR within smart home environments, using deep learning techniques in conjunction with oversampling (specifically, SMOTE) and undersampling methods. The results unequivocally favored the SMOTE method over undersampling in effectively addressing imbalanced data challenges within HAR tasks using the SPHERE dataset [14].

Alharbi et al. (2022) made significant contributions by investigating the effectiveness of oversampling methods, such as SMOTE and its hybrid variations, in improving the classification of minority classes in diverse datasets. For instance, on the PAMAP2 dataset, the MLP achieved an F1 score of 0.7185 using the SMOTE sampling method, compared to its baseline score of 0.7473. [7].

In addition to the aforementioned studies, recent research has showcased the potential of deep learning models for HAR:

- Wan et al. (2020) introduced deep models for real-time HAR using smartphones, including CNN and LSTM models, achieving high accuracies of 91.00% and 85.86%, respectively on the PAMAP 2 Dataset [15].
- Xu et al. (2022) proposed several methods, including classical CNN, LSTM, and Inception-LSTM with attention mechanisms, achieving F1 scores ranging from 0.8949 to 0.9513 [16].
- Tehrani et al. (2023) utilized a deep multi-layer Bi-LSTM model for sensor-based HAR, obtaining promising results with F1-score, Precision, Recall, and Accuracy all reaching 93.41% [17].
- Thakur et al. (2022) demonstrated that a hybrid model combining CNN and LSTM with an autoencoder for dimensionality reduction achieved an impressive F1 score of 0.9446 and an accuracy of 94.33% for HAR [4].
- Challa et al. (2022) proposed a multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data, achieving an impressive accuracy of 94.29% [18].

Table I summarizes the performance of various deep learning models on the PAMAP2 dataset using different sampling methods for sensor-based HAR.

These studies collectively emphasize the positive impact of oversampling techniques, particularly SMOTE, in enhancing model performance when compared to training on imbalanced datasets. These insights lay the foundation for this research, which aims to build upon this foundation and further investigate the efficacy of sampling methods in improving the performance of deep learning models for HAR on the PAMAP2 dataset.

In the field of HAR, a significant gap in existing research has been found. There hasn't been enough focus on how different sampling techniques affect the performance of deep learning models in HAR. While some studies have tackled imbalanced data in HAR, they often overlook the critical role that sampling methods play. This gap highlights the need for a more thorough investigation into how sampling techniques and deep learning intersect in HAR. That's where the research steps in. The commitment is to address this gap by thoroughly studying how various sampling methods impact the performance of deep learning models in real-world HAR scenarios. The goal is to provide a clearer picture of how sampling methods and deep learning models work together, ultimately improving the accuracy and reliability of activity recognition in sensor-based applications.

TABLE I. PREVIOUS STUDIES PERFORMANCE ON PAMAP2 DATASET USING DEEP LEARNING MODELS AND SAMPLING METHODS FOR SENSOR BASED HAR

Study	year	Dataset	Classification method	Sampling method	Accuracy	F1 score	Precision	Recall
[14]	2020	SPHERE	CNN	NONE	0.7030	-	-	-
[14]	2020	SPHERE	LSTM	NONE	0.6598	-	-	-
[14]	2020	SPHERE	CNN-LSTM	NONE	0.6829	-	-	-
[14]	2020	SPHERE	CNN	SMOTE	0.9355	-	-	-
[14]	2020	SPHERE	LSTM	SMOTE	0.9298	-	-	-
[14]	2020	SPHERE	CNN-LSTM	SMOTE	0.9367	-	-	-
[14]	2020	SPHERE	CNN	UNDERSAMPLING	0.2937	-	-	-
[14]	2020	SPHERE	LSTM	UNDERSAMPLING	0.3794	-	-	-
[14]	2020	SPHERE	CNN-LSTM	UNDERSAMPLING	0.3085	-	-	-
[15]	2020	PAMAP2	LSTM	NONE	0.8580	0.8534	0.8651	0.8467
[16]	2022	PAMAP2	LSTM	NONE	0.8920	0.8949	0.8969	0.8928
[17]	2023	PAMAP2	Bi-LSTM	NONE	0.9341	0.9341	0.9341	0.9347
[4]	2022	PAMAP2	convLSTM AE	NONE	0.9433	0.9446	-	-
[18]	2022	PAMAP2	CNN-BiLSTM	NONE	0.9429	-	-	-
[7]	2022	PAMAP2	MLP	SMOTE	--	0.7473	0.7769	0.7493

III. MATERIAL AND METHODS

In this research, the impact of class imbalance on HAR using wearable sensor data and deep learning models was investigated. To address this issue, three sampling methods were thoroughly examined: SMOTE, Random Undersampling, and a hybrid combination of the aforementioned techniques. The study involved the training of four deep learning models, including Vanilla LSTM, 2 Stacked LSTM, 3 Stacked LSTM, and Hybrid CNN-LSTM, on the PAMAP2 dataset. Through rigorous experimentation and evaluation, the aim was to identify the most effective sampling approach to improve model performance and generalization in HAR. The findings are expected to contribute valuable insights towards enhancing the accuracy and reliability of HAR systems deployed in real-world scenarios.

A. PAMAP2 Dataset

The PAMAP2 dataset [11], which stands for "Physical Activity Monitoring using a Multipurpose Sensor" holds a prominent role in the realm of Human Activity Recognition (HAR) research. Its comprehensive data collection approach, diverse participant demographic, and meticulous data organization make it a valuable resource for the research community.

Here are some key characteristics of the PAMAP2 dataset, as extracted from the dataset documentation [11] :

- Participant Diversity: One noteworthy aspect of the PAMAP2 dataset is the diversity of its participant pool. This dataset comprises data contributed by both genders, with a broad age range spanning from 23 to 32 years. Moreover, it includes individuals with varying physical characteristics, such as weights ranging from 65 to 95 kilograms and heights spanning between 168 and 194 centimeters. This demographic diversity empowers researchers to develop activity recognition models applicable to a broad spectrum of individuals.

- Data Collection: Researchers collected the dataset using a range of wearable sensors, including those worn on the wrist, chest, and ankle. These sensors operated at a high sampling rate of 100Hz, enabling the capture of an extensive volume of data, essential for detailed analysis of activities and movements (see Table II).

- Activity Variety: The PAMAP2 dataset offers an array of data related to various physical activities. It encompasses 12 distinct activity types, with detailed descriptions provided in Table III. This categorization serves as a valuable reference for activity labeling and model development.

- Data Format: The dataset is thoughtfully organized, with raw data from all sensors synchronized and labeled, and then consolidated into a single data file per participant and session. These files are presented in text format (.dat), simplifying structured data manipulation and analysis.

TABLE II. PAMAP2 DATASET DESCRIPTION

Dataset	Labels	Sampling Rate	Windows Size	Overlap	# Subjects
PAMAP2	12	100 Hz	1s	50%	9

Table III provides an overview of the data distribution within the PAMAP2 dataset, highlighting a significant class imbalance among different activity labels in both the training and testing datasets. This issue is further underscored in Fig. 1, where it becomes evident that the "Rope jumping" activity exhibits notably fewer instances compared to other activities. This skewed data distribution can exert a substantial impact on the performance of deep learning models. Consequently, it becomes imperative to implement suitable sampling strategies to guarantee the reliability and accuracy of results.

TABLE III. DATA DISTRIBUTION PER ACTIVITY IN THE PAMAP2 DATASET

Class id	Activity label	# Instances Training Set 70%	# Instances Testing Set 70%
0	Lying	100298	42633
1	Sitting	58380	25358
2	Standing	70165	29808
3	Walking	86117	36789
4	Running	30100	12950
5	Cycling	63755	27585
6	Nordic Walking	78154	33678
7	Ascending stairs	41442	17872
8	Descending stairs	32800	14030
9	Vacuum cleaning	60703	26256
10	Ironing	87885	37343
11	Rope jumping	10889	4564

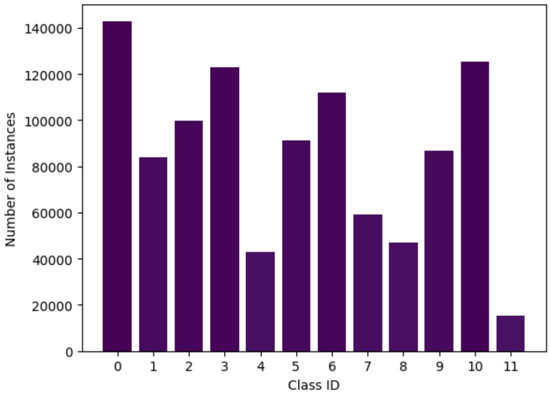


Fig. 1. Data distribution by activity in PAMAP2 dataset.

B. Deep Learning Models

1) Long short-term memory (LSTM): LSTM networks belong to the category of recurrent neural networks (RNNs) and hold significance in time series applications, particularly HAR, that involve the classification of activities based on sensor data, such as accelerometers and gyroscope readings from smartphones. The strength of LSTM networks in HAR lies in their capability to capture and model long-term dependencies present within the sensor data [19].

2) Hybrid deep learning model (CNN-LSTM): This research harnesses the power of hybrid models, specifically the integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This combination, as evidenced by several studies [14][4], holds promise for achieving high performance in HAR tasks. The rationale behind selecting this hybrid model is compelling: CNN excels at capturing spatial relationships within data, while LSTM is adept at modeling temporal dependencies. This combination allows us to leverage the strengths of both architectures [20]. One notable advantage of the hybrid model is that CNN accelerates the feature extraction process, enhancing training efficiency. This synergy between CNN and LSTM contributes to the model's overall effectiveness in recognizing human activities based on sensor data.

3) Deep learning models configurations: In the following deep learning configuration for multiclass HAR classification, various layers play distinct roles. These include the LSTM layer, dropout layer, dense layer with Softmax activation for probability estimation, convolutional (Conv1D) layer, and max pooling layers. Each layer plays a specific role in a deep learning architecture for multiclass classification. The LSTM layer captures sequential dependencies in the data, making it suitable for time series or sequential data. The dropout layer helps prevent overfitting by randomly deactivating a fraction of neurons during training, enhancing the model's generalization. The dense layer, often found in the final stage, produces class scores. The softmax activation function applied to these logits converts them into class probabilities. The convolutional (Conv1D) layer extracts spatial features from the input data. Max pooling layers reduce the spatial dimensions while retaining essential information, aiding in feature selection and computational efficiency. Combined, these layers enable the deep learning model to process, understand, and classify data efficiently and accurately.

In this study, several configurations of deep learning models for HAR are explored. These configurations include:

a) Vanilla LSTM: This straightforward LSTM setup consists of a single hidden layer of LSTM units and an output layer for prediction. It has proven its effectiveness in various small sequence prediction tasks [21].

Fig. 2 illustrates the architecture of the Vanilla LSTM model. It provides a visual representation of the model's structure, showcasing the flow of data through its layers, including the LSTM layer, dropout layers, and dense layers, ultimately leading to the output layer for activity classification.

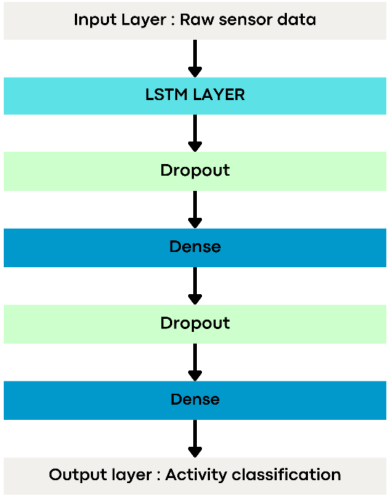

Input Layer : Raw sensor data
LSTM LAYER
Dropout
Dense
Dropout
Dense
Output layer : Activity classification

Fig. 2. Structure of vanilla LSTM model.

b) 2-Stacked LSTM: Variants of the Stacked LSTM model, featuring two hidden layers, are also investigated in this study. Emerging from research findings, Stacked LSTM networks exhibit improved recognition efficiency by iteratively extracting temporal features [21].

Fig. 3 provides an overview of the 2-Stacked LSTM model's architecture. This model is specifically designed for Human Activity Recognition (HAR) and excels in capturing intricate temporal patterns within sensor data. It consists of two LSTM layers with 64 units each, enabling the understanding and modeling of complex temporal relationships. Dropout layers are strategically placed to prevent overfitting during training. The model then utilizes two Dense layers, with 96 and 12 units, for feature extraction and final classification. Overall, the 2-stacked LSTM model's structure is optimized for accurate and robust activity recognition in HAR applications.

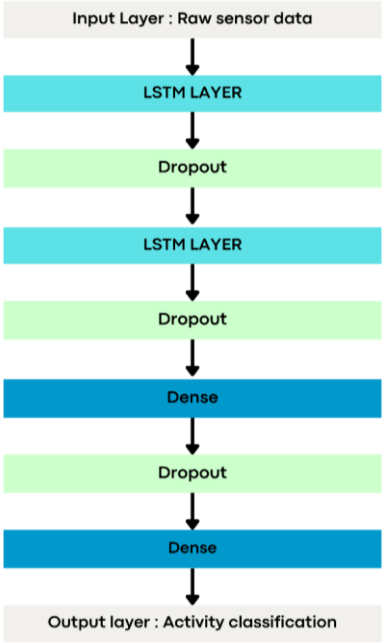

Input Layer : Raw sensor data
LSTM LAYER
Dropout
LSTM LAYER
Dropout
Dense
Dropout
Dense
Output layer : Activity classification

Fig. 3. Structure of 2-stacked LSTM model.

c) 3-Stacked LSTM: Similar to the 2-Stacked LSTM but with an additional layer of LSTM units, this configuration aims to further enhance the model's capacity for temporal feature extraction [21].

Fig. 4 provides an overview of the 3-Stacked LSTM model's architecture, designed for Human Activity Recognition (HAR). This model excels at capturing intricate temporal patterns within sensor data. It comprises three LSTM layers, each with 32 units, to model complex temporal relationships. Dropout layers are integrated to prevent overfitting during training. The model also includes two dense layers with 64 and 12 units, respectively, for feature extraction and final activity classification. In summary, the 3-Stacked LSTM model is engineered to achieve robust and accurate activity recognition in HAR scenarios by effectively handling temporal data dependencies and ensuring generalization through dropout mechanisms.

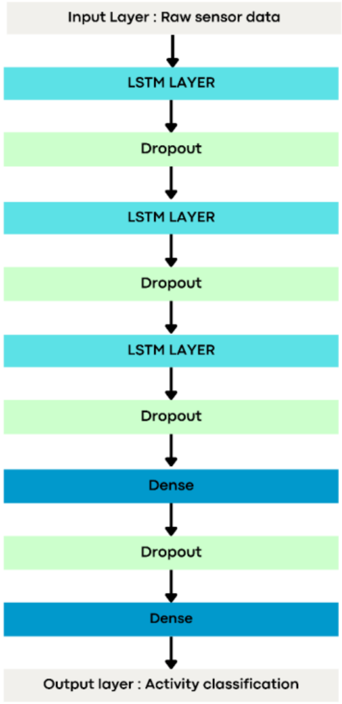


Fig. 4. Structure of 3-stacked LSTM model.

d) Hybrid Model (CNN-LSTM): The CNN-LSTM model, a hybrid architecture that seamlessly combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) layers.

Fig. 5 outlines the Hybrid CNN-LSTM model for HAR. The architecture starts with a CNN layer followed by dropout and max pooling for feature extraction. Subsequently, an LSTM layer captures temporal patterns with dropout for regularization. The final dense layer performs activity classification. This design effectively handles spatial and temporal aspects of sensor data, ensuring robust activity recognition in HAR.

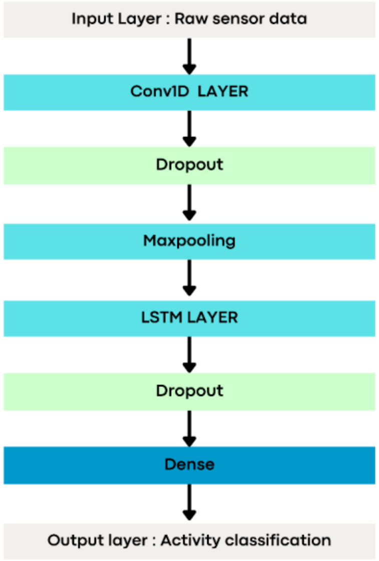

Input Layer : Raw sensor data
Conv1D LAYER
Dropout
Maxpooling
LSTM LAYER
Dropout
Dense
Output layer : Activity classification

Fig. 5. Structure of hybrid CNN-LSTM model.

C. Sampling Techniques

To tackle the challenge of imbalanced data, three different sampling techniques were applied:

1) SMOTE (Synthetic minority over-sampling technique): This Sampling technique serves as an effective tool for addressing imbalanced datasets in the realm of machine learning. Its function involves creating synthetic data points for the underrepresented class by bridging the gap between existing samples. In the context of sensor-based Human Activity Recognition (HAR) using deep learning, SMOTE plays a vital role in enhancing the classification accuracy of models like Multi-Layer Perceptrons (MLPs) [7].

Deep learning models require a high amount of data and are very sensitive to the imbalanced class problem. This is where SMOTE steps in, generating artificial samples for the minority class, thereby balancing the dataset and significantly improving the classification accuracy of these deep learning models [14].

2) Random undersampling: This Sampling method addresses imbalanced datasets by randomly removing samples from the majority class to achieve balance. In sensor-based Human Activity Recognition (HAR) with deep learning, is employed to boost deep learning model classification accuracy [14]. Deep learning models require a high amount of data and are sensitive to class imbalances. Thus, Random Undersampling eliminates samples from the majority class, balancing the dataset and improving classification accuracy [14]. However, this method can lead to the loss of critical information from the majority class, potentially impacting the model's classification accuracy [7]. Hence, it is crucial to carefully select the samples for removal to prevent the loss of vital information.

3) Hybrid sampling: Hybrid sampling is a technique used to deal with imbalanced datasets in machine learning. It involves combining oversampling and undersampling methods to balance the dataset. This method generates synthetic samples for the minority class using SMOTE and randomly removes samples from the majority class using Random Undersampling. The combination of these two methods helps to balance the dataset and improve the classification accuracy of deep learning models [7][14]. Hybrid sampling is particularly effective in sensor-based Human Activity Recognition (HAR) when combined with deep learning models. This technique successfully addresses the challenge of imbalanced classes while simultaneously mitigating the risk of losing valuable information from the majority class that can occur with random undersampling alone. By generating synthetic samples for the minority class through SMOTE, hybrid sampling ensures a well-represented minority class in the dataset. This balanced dataset significantly enhances the classification accuracy of deep learning models, while also promoting data diversity [14].

Table A1 in Appendix A provides a comprehensive overview of the dataset instances before and after the application of various sampling methods. The table allows for a clear visualization of how each sampling technique impacts the dataset composition.

D. Hyperparameter Tuning with Bayesian Optimization

The Model Hyperparameters are crucial in deep learning, shaping training algorithms and model performance. Bayesian optimization offers an effective means to optimize these parameters, particularly in complex, function-based problems lacking simple analytical solutions. To apply Bayesian optimization to time series and sensor-based Human Activity Recognition (HAR) using LSTM models, the following steps can be followed:

Step 1: Define the hyperparameter search space.

Step 2: Specify the objective function to evaluate model performance.

Step 3: Initialize the Bayesian optimization algorithm with hyperparameter values.

Step 4: Iteratively use the algorithm to suggest hyperparameters for evaluation.

Step 5: Continue until predefined convergence criteria are met, like a set number of iterations or desired performance levels.

E. Evaluation Metrics

In the experiment, various evaluation metrics were used to assess the HAR model's performance. These metrics included accuracy, F1 score, precision, recall, and the confusion matrix. These evaluation metrics determine the performance of a model on a dataset. The most common metric is the confusion matrix which is a two-dimension table of class labels; one represents the current class and the other represents the predicted one. Accuracy is the most used one to evaluate model classification. It defines a ratio of correct predictions and overall predictions. The accuracy can be a good measure when the dataset class is balanced. Otherwise, this metric is not appropriate for evaluation. In the case of imbalanced datasets, other metrics are used such as precision, recall, f-measure, and specificity. Table IV presents the definition of all these metrics [22].

Understanding these performance metrics requires knowledge of four fundamental terms used in their measurement: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

TABLE IV. PERFORMANCE METRICS

Metric	Formula	Definition
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	the ratio of correct predictions and overall predictions
Precision	$\frac{tp}{tp+fp}$	the ratio of correct predictions to the total predicted
Recall of sensitivity	$\frac{tp}{tp+fn}$	the ratio of correct predictions to the samples in the actual class
Specificity	$\frac{tn}{tn+fp}$	The ratio of actual class 0 to the correctly predicted 0
F1 score / F-measure	$\frac{2(recall*precision)}{recall+precision}$	The weighted average of precision and Recall if the data is imbalanced

IV. Experiments and Results

A. Experimental Design

This research aims to investigate the impact of data balancing techniques on the performance of deep learning models for Human Activity Recognition (HAR) by addressing the following research questions:

1) How does class imbalance affect the performance of deep learning models in Human Activity Recognition (HAR) when applied to wearable sensor data?

2) What are the comparative effects of different sampling techniques, such as SMOTE, Random Undersampling, and Hybrid Sampling, in addressing the class imbalance in wearable sensor data for HAR?

3) What role does hyperparameter tuning play in improving the accuracy and performance of deep learning models for HAR, particularly in the context of imbalanced datasets?

4) Which combination of sampling technique and hyperparameter tuning strategy yields the most significant performance improvements in HAR using deep learning models for imbalanced wearable sensor data?

The hypothesis guiding this study is that balancing the dataset will result in enhanced classification accuracy in HAR using deep learning models. The experiments were carried out using the PAMAP2 dataset collected from wearable sensors, encompassing wrist, chest, and ankle devices. Four deep learning models were employed: Vanilla LSTM, 2-Stacked LSTM, 3-Stacked LSTM, and CNN-LSTM.

B. Experimental Setup

The conducted experiments are performed on an NVIDIA GPU V100 using the Google Collaboratory Pro+ platform. The four models' hyperparameters were optimized through Bayesian Hyperparameter Optimization, utilizing the Keras Tuner library[23]. The experiment setup is detailed in Table V.

TABLE V. EXPERIMENTAL SETUP

Platform	Google Colab Pro+
GPU	NVIDIA GPU V100
RAM	15 GB
Tenserflow version	2.12.0
Keras Version	2.12.0
Keras Tuner Version	1.3.5

C. Experiment Pipeline

To evaluate the models' performance on the PAMAP2 dataset, a comprehensive experiment pipeline was executed. This pipeline is composed of multiple stages, each playing a vital role in the experiments (see Fig. 6):

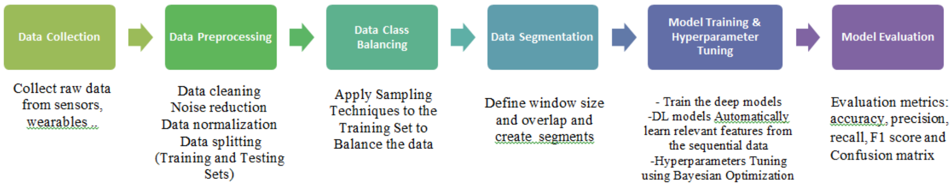

Data Collection
Data Preprocessing
Data Class Balancing
Data Segmentation
Model Training & Hyperparameter Tuning
Model Evaluation
Collect raw data from sensors, wearables ..
Data cleaning
Noise reduction
Data normalization
Data splitting
(Training and Testing Sets)
Apply Sampling Techniques to the Training Set to Balance the data
Define window size and overlap and create segments
- Train the deep models
-DL models Automatically learn relevant features from the sequential data
-Hyperparameters Tuning using Bayesian Optimization
Evaluation metrics: accuracy, precision, recall, F1 score and Confusion matrix

Fig. 6. Experiments pipeline.

1) Data collection: Initially, the raw sensor data from wearable devices were collected.

2) Data preprocessing: The dataset goes through a preprocessing phase, involving actions like data cleaning, noise reduction, and normalization.

During this stage, the raw sensor data from wearable devices is readied for the proposed model. The subject-specific files containing activity records are consolidated into one data frame. To adhere to PAMAP2 guidelines, invalid orientation columns are removed, and transient activity rows are dropped. Non-numeric data is transformed into numeric form, and missing values are interpolated to ensure data integrity. Scaling is applied to normalize input features, ensuring data uniformity. Labels are encoded and converted into categorical variables, a critical step for activity classification during model training.

The data is then split into training and testing sets, with 70% allocated for training and 30% for testing. Data is segmented into overlapping windows, with a window size of 1 second and a 50% overlap. This segmentation process creates segments and associated labels for both training and testing. The segments and labels are reshaped to align with the LSTM model's input format. The experiment validates the shape of training and testing segments before moving on to model training and evaluation phases.

3) Data class balancing: In the data balancing stage, three different sampling techniques were applied to tackle class imbalance in the experiments: SMOTE, Random Undersampling, and a hybrid approach. SMOTE was used to generate synthetic instances for minority classes, Random Undersampling involved reducing instances in the majority class, and the hybrid approach combined both methods. The objective was to create balanced datasets to enhance model training. These sampling techniques were exclusively applied to the training set to ensure class balance for improved model performance.

4) Data segmentation: Data were segmented into overlapping windows, following a window size of one second with a 50% overlap. This facilitated the data's suitability for deep learning models.

5) Training and hyperparameter optimization: Deep learning models performed feature extraction automatically to identify relevant patterns in the segmented data.

The four models were trained using Bayesian Optimization to fine-tune hyperparameters for optimal model performance. The Keras Tuner library is utilized to search for the best hyperparameters.

The models are fine-tuned by adjusting several critical hyperparameters: the LSTM units, which determine the number of LSTM units in each LSTM layer, are explored within the range of 64 to 256, with a step size of 32. Similarly, the dense units, specifying the number of units in the dense layer, are considered within the range of 32 to 128, with a step size of 32. The batch size hyperparameter, significant for model training, is chosen from the options of 32, 64, or 128. The learning rate, influencing the optimizer's learning rate, is selected from values like 1e-3, 1e-4, or 1e-5. Furthermore, the dropout rate, responsible for controlling the dropout applied after each LSTM layer and the dense layer, varies from 0.1 to 0.5, with a step size of 0.1. The optimizer hyperparameter allows the choice of ADAM or RMSprop as the optimizer used to compile the model. The number of epochs in this experiments ranges from 50 to 100.This extensive exploration and fine-tuning of the models ultimately result in enhanced accuracy and robust performance for HAR tasks. All these hyperparameters are summarized in Table VI.

TABLE VI. HYPERPARAMETER RANGES FOR BAYESIAN OPTIMIZATION

Hyperparameter	Search Space
Lstm Units	[32, 64, 96, 128]
Dense Units	[32, 64, 96, 128]
Dropout Rate	[0.1, 0.2, 0.3, 0.4, 0.5]
Optimizer	['adam', 'rmsprop']
Learning Rate	[1e-2, 1e-3, 1e-4]
Batch Size	[32, 64, 128]
Epochs	[50, 51, ..., 100]

6) *Model evaluation*: To evaluate the models performance on the PAMPA2 dataset. The evaluation metrics were used including accuracy, precision, recall, F1-score and confusion matrix. These metrics were compared against those reported in previous literature studies conducted on the same dataset, enabling a comprehensive assessment of the proposed model's effectiveness and advancements in HAR.

Four experiments were conducted:

- Experiment 1: Train and test the four models on an imbalanced dataset.
- Experiment 2: Train and test the four models on the balanced dataset with SMOTE.
- Experiment 3: Train and test the four models on the balanced dataset with Random Undersampling.
- Experiment 4 : Train and test the four models on the balanced dataset with hybrid Sampling(SMOTE & Random Undersampling).

D. Experiments Results

1) *Experiment 1: Baseline*: In Experiment 1, the baseline was established to compare the effects of various data balancing techniques.

TABLE VII. THE SUMMARIZED HYPERPARAMETERS OF THE FOUR MODELS FOUND BY KERAS TUNER ON IMBALANCED DATA

Hyper parameter	Vanilla LSTM	2Stacked LSTM	3 Stacked LSTM	CNN LSTM
Lstm Units	32	64	96	CNN units:128 Lstm units : 64
Dense Units	32	96	128	-
Dropout Rate	0.1	0.3	0.2	0.4
Optimizer	RMSPROP	ADAM	ADAM	ADAM
Learning Rate	0.001	0.001	0.001	0.001
Batch Size	32	32	128	128
Epochs	82	78	78	65

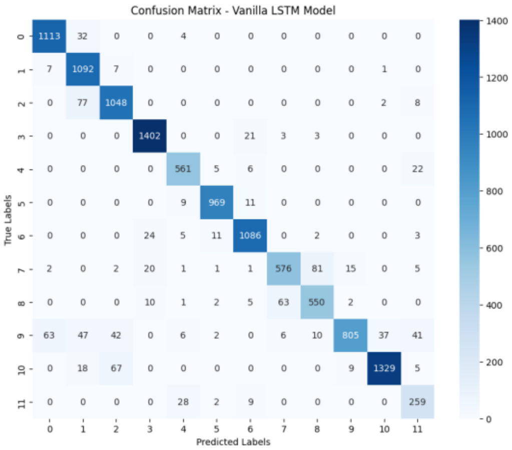
Subsequently, the four deep learning models were trained on the preprocessed imbalanced dataset. The optimization of hyperparameters for these models was carried out using Keras Tuner Bayesian optimization. The best hyperparameters of each model are summarized in Table VII.

Table VIII presents the results of Experiment 1, showcasing the performance metrics of the models, which include accuracy, precision, recall, and F1-score. These metrics were measured to establish a baseline for comparison.

TABLE VIII. RESULTS OF EXPERIMENT 1 ON IMBALANCED DATA

Metrics	Vanilla LSTM	2 Stacked LSTM	3 Stacked LSTM	CNN LSTM
Accuracy	0.9257	**0.9531**	0.9232	0.9308
F1 Score	0.9250	**0.9529**	0.9232	0.9297
Precision	0.9281	**0.9536**	0.9268	0.9341
Recall	0.9257	**0.9531**	0.9232	0.9308

Fig. 7 to Fig. 10 depicts the confusion matrices for the Vanilla LSTM model, 2-Stacked LSTM, 3-Stacked LSTM, and CNN-LSTM, respectively, on the imbalanced dataset.



	0	1	2	3	4	5	6	7	8	9	10	11
0	1113	32	0	0	4	0	0	0	0	0	0	0
1	7	1092	7	0	0	0	0	0	0	0	1	0
2	0	77	1048	0	0	0	0	0	0	0	2	8
3	0	0	0	1402	0	0	21	3	3	0	0	0
4	0	0	0	0	561	5	6	0	0	0	0	22
5	0	0	0	0	9	969	11	0	0	0	0	0
6	0	0	0	24	5	11	1086	0	2	0	0	3
7	2	0	2	20	1	1	1	576	81	15	0	5
8	0	0	0	10	1	2	5	63	550	2	0	0
9	63	47	42	0	6	2	0	6	10	805	37	41
10	0	18	67	0	0	0	0	0	0	9	1329	5
11	0	0	0	0	28	2	9	0	0	0	0	259

Fig. 7. Confusion matrix of vanilla LSTM model on imbalanced data.

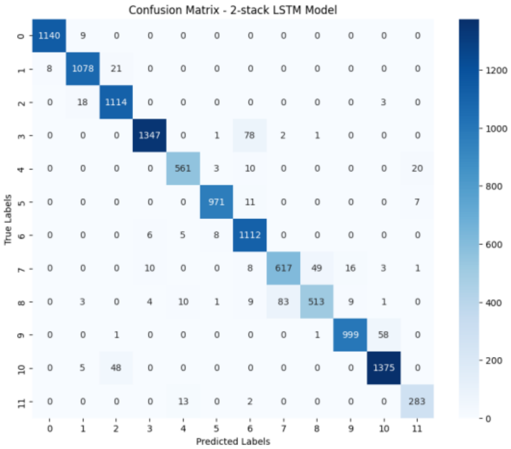


Fig. 8. Confusion matrix of 2 stacked LSTM model on imbalanced data.

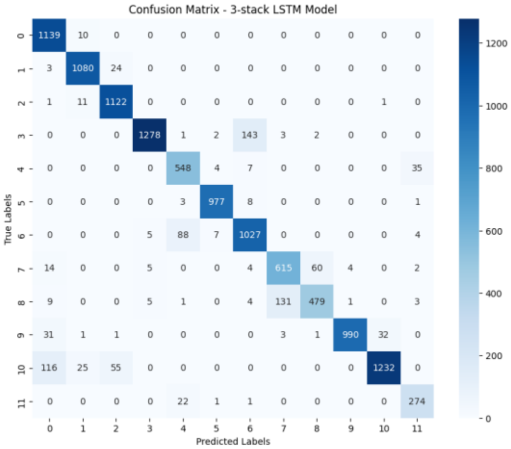


Fig. 9. Confusion matrix of 3 stacked LSTM model on imbalanced data.

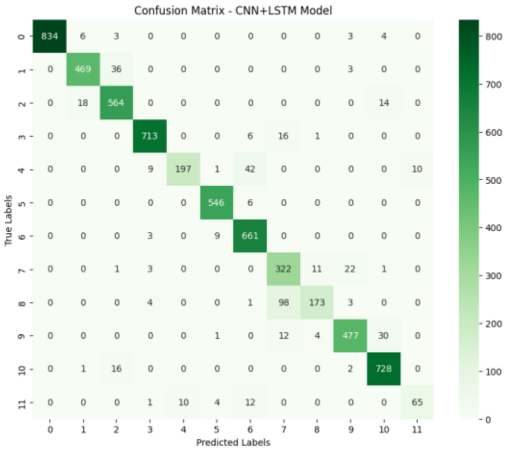


Fig. 10. Confusion matrix of CNN LSTM model on imbalanced data.

2) Experiment 2: Balancing data with SMOTE: In this experiment, the evaluation of model performance was conducted when trained on a dataset balanced using the Synthetic Minority Over-sampling Technique (SMOTE). The four models underwent training on the SMOTE-balanced dataset, and the search for the best hyperparameters for each model was facilitated by Keras Tuner, as shown in Table VIII .

Performance metrics achieved in this experiment were observed and reported in Table IX, with a comparison to those from Experiment 1. Fig. 11 to Fig. 14 depicts the confusion matrices for the Vanilla LSTM model, 2-Stacked LSTM, 3-Stacked LSTM, and CNN-LSTM, respectively, on the balanced data with SMOTE (see Table X).

TABLE IX. THE SUMMARIZED HYPERPARAMETERS OF THE FOUR MODELS FOUND BY KERAS TUNER ON BALANCED DATA WITH SMOTE

Hyper parameter	**Vanilla LSTM**	**2 Stacked LSTM**	**3 Stacked LSTM**	**CNN LSTM**
LSTM Units	96	96	64	CNN UNITS:128 LSTM UNITS:32
Dense Units	64	32	128	32
Dropout Rate	0.1	0.4	0.4	0.3
Optimizer	RMSPROP	ADAM	RMSPROP	ADAM
Learning Rate	0.01	0.001	0.01	0.001
Batch Size	32	32	32	64
Epochs	77	91	58	94

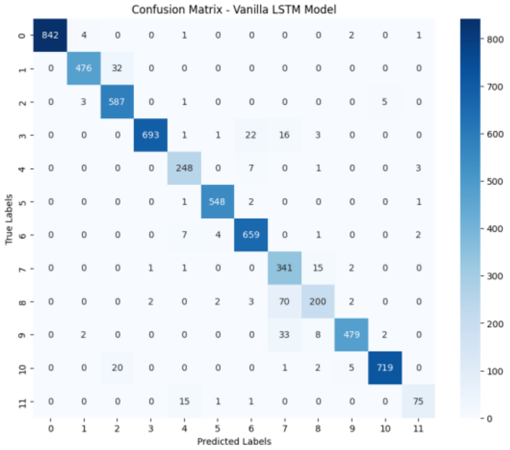


Fig. 11. Confusion matrix of vanilla LSTM on balanced data with SMOTE.

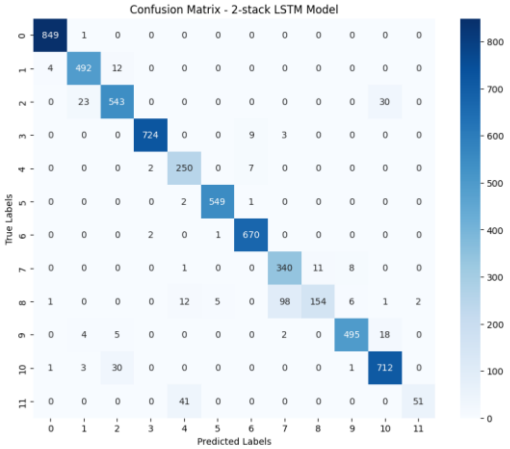


Fig. 12. Confusion matrix of 2 stack LSTM on balanced data with SMOTE.

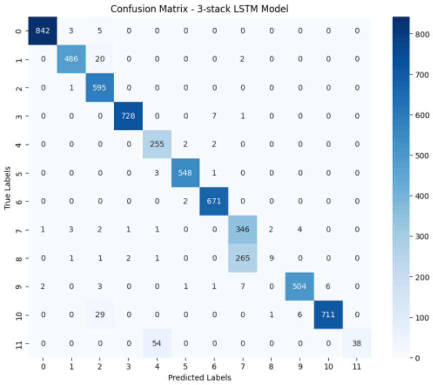


Fig. 13. Confusion matrix of 3 stacked LSTM on balanced data with SMOTE.

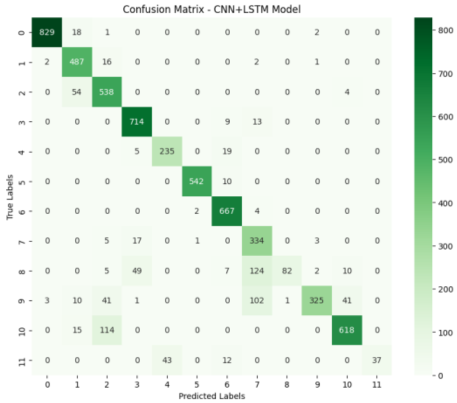


Fig. 14. Confusion matrix of CNN LSTM on balanced data with SMOTE.

3) *Experiment 3: Random undersampling*: Experiment 3 entailed the assessment of the models' performance when trained on a dataset balanced through Random Undersampling. The four models underwent training on the randomly undersampled Training set. The search for the best hyperparameters for each model was conducted using Keras Tuner, as indicated in Table XI and Table XII shows the Experiment 3on balanced data with random undersampling.

Performance metrics achieved in this experiment were observed and reported in Table XIII, with a comparison to those from Experiment 1. Fig. 15 to Fig. 18 illustrate the confusion matrices for the Vanilla LSTM model, 2-Stacked LSTM, 3-Stacked LSTM, and CNN-LSTM, respectively, on the balanced data achieved through Random Undersampling.

4) *Experiment 4: Hybrid sampling*: In Experiment 4, the examination of the models' performance was carried out when trained on a dataset balanced using hybrid sampling, combining SMOTE and random undersampling. The four models underwent training on the hybrid-sampled dataset. The search for the best hyperparameters for each model was conducted using Keras Tuner, as indicated in Table XIII.

Performance metrics from this experiment were documented in Table XIV and compared with the results from Experiment 1. Fig. 19 to Fig. 22 illustrate the confusion matrices for the Vanilla LSTM model, 2-Stacked LSTM, 3-Stacked LSTM, and CNN-LSTM, respectively, on the balanced data achieved through hybrid Sampling.

TABLE X. RESULTS OF EXPERIMENT 2 ON BALANCED DATA WITH SMOTE

	Imbalanced data				**Balanced data With Smote**			
	Vanilla LSTM	***2 Stacked LSTM***	***3 Stacked LSTM***	***CNN LSTM***	***Vanilla LSTM***	***2 Stacked LSTM***	***3 Stacked LSTM***	***CNN LSTM***
Accuracy	0.9257	0.9531	0.9232	0.9308	**0.9499**	0.9438	**0.9282**	0.8756
F1 Score	0.9250	0.9529	0.9232	0.9297	**0.9503**	0.9415	**0.9129**	0.8687
Precision	0.9281	0.9536	0.9268	0.9341	**0.9537**	0.9470	**0.9386**	0.8975
Recall	0.9257	0.9531	0.9232	0.9308	**0.9499**	0.9438	**0.9282**	0.8756

TABLE XI. THE SUMMARIZED HYPERPARAMETERS OF THE FOUR MODELS FOUND BY KERAS TUNER ON BALANCED DATA WITH RANDOM UNDERSAMPLING

Hyper Parameters	**Vanilla LSTM**	**2 Stacked LSTM**	**3 stacked LSTM**	**CNN LSTM**
Lstm Units	32	96	128	CNN UNITS:128 LSTM UNITS:32
Dense Units	96	64	32	--
Dropout Rate	0.2	0.2	0.3	0.2
Optimizer	RMSPROP	RMSPROP	RMSPROP	RMSPROP
Learning Rate	0.01	0.01	0.01	0.0001
Batch Size	64	64	128	64
Epochs	75	62	62	73

TABLE XII. RESULTS OF EXPERIMENT 3 ON BALANCED DATA WITH RANDOM UNDERSAMPLING

	Imbalanced Data				**Balanced Data With Random Undersampling**			
	Vanilla LSTM	***2 Stacked LSTM***	***3 Stacked LSTM***	***CNN LSTM***	***Vanilla LSTM***	***2 Stacked LSTM***	***3 Stacked LSTM***	***CNN LSTM***
Accuracy	**0.9257**	**0.9531**	**0.9232**	**0.9308**	0.7295	0.6361	0.2953	0.3706
F1 Score	**0.9250**	**0.9529**	**0.9232**	**0.9297**	0.6946	0.6070	0.2132	0.3194
Precision	**0.9281**	**0.9536**	**0.9268**	**0.9341**	0.6812	0.6113	0.3058	0.3767
Recall	**0.9257**	**0.9531**	**0.9232**	**0.9308**	0.7295	0.6361	0.2953	0.3706

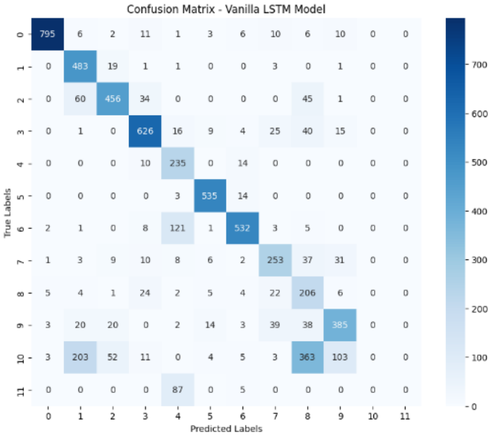


Fig. 15. Confusion matrix of Vanilla LSTM on balanced data with random undersampling.

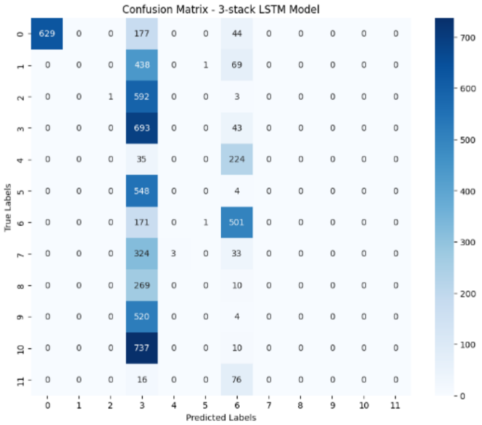


Fig. 17. Confusion matrix of 3 stacked LSTM on balanced data with random undersampling.

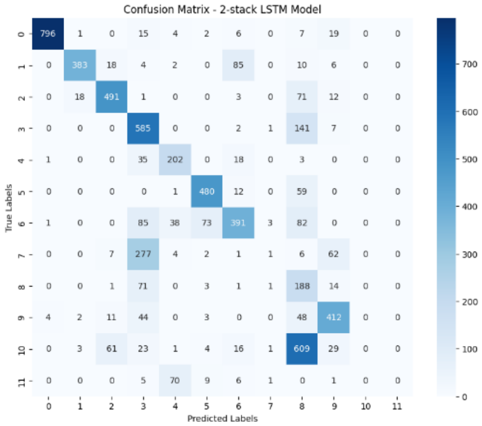


Fig. 16. Confusion matrix of 2 stacked LSTM on balanced data with random undersampling.

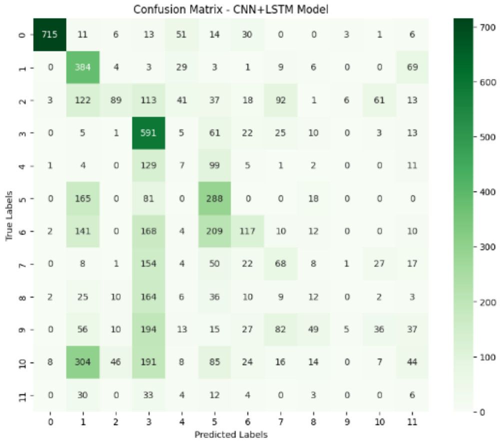


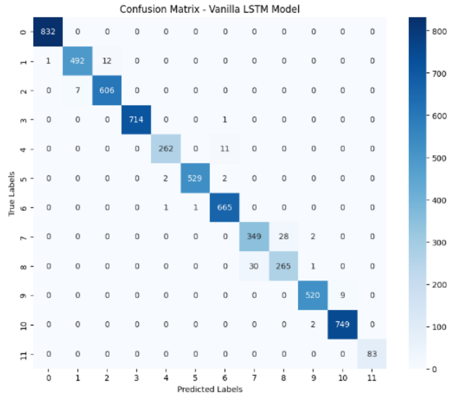
Fig. 18. Confusion matrix of CNN-LSTM on balanced data with random undersampling.

TABLE XIII. THE SUMMARIZED HYPERPARAMETERS OF THE FOUR MODELS FOUND BY KERAS TUNER ON BALANCED DATA WITH HYBRID SAMPLING

Hyper parameter	**Vanilla LSTM**	**2 Stacked LSTM**	**3 Stacked LSTM**	**CNN LSTM**
Lstm Units	32	64	96	CNN UNITS:96 LSTM UNITS:96
Dense Units	64	128	32	--
Dropout Rate	0.3	0.3	0.2	0.3
Optimizer	ADAM	RMSPROP	ADAM	ADAM
Learning Rate	0.001	0.01	0.001	0.001
Batch Size	128	64	32	32
Epochs	56	78	81	93

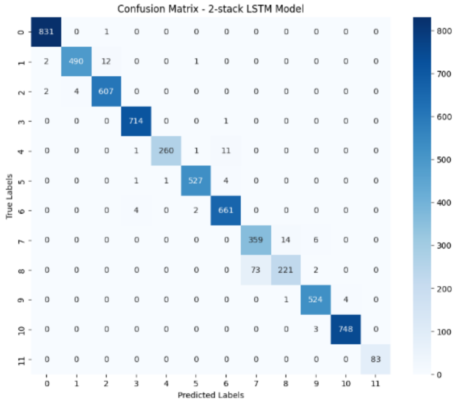
TABLE XIV. RESULTS OF EXPERIMENT 2 ON BALANCED DATA WITH HYBRID SAMPLING

	Imbalanced data				**After hybrid undersampling**			
	Vanilla LSTM	***2 Stacked LSTM***	***3 Stacked LSTM***	***CNN LSTM***	***Vanilla LSTM***	***2 Stacked LSTM***	***3 Stacked LSTM***	***CNN LSTM***
Accuracy	0.9257	0.9531	0.9232	0.9308	**0.9821**	**0.9755**	**0.9828**	**0.9351**
F1 Score	0.9250	0.9529	0.9232	0.9297	**0.9821**	**0.9752**	**0.9828**	**0.9342**
Precision	0.9281	0.9536	0.9268	0.9341	**0.9822**	**0.9764**	**0.9828**	**0.9350**
Recall	0.9257	0.9531	0.9232	0.9308	**0.9822**	**0.9755**	**0.9828**	**0.9351**



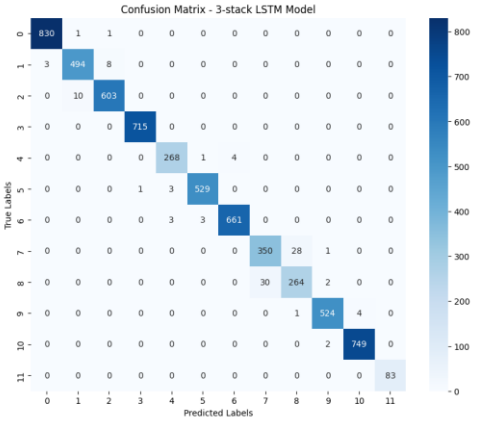
	0	1	2	3	4	5	6	7	8	9	10	11
0	832	0	0	0	0	0	0	0	0	0	0	0
1	1	492	12	0	0	0	0	0	0	0	0	0
2	0	7	606	0	0	0	0	0	0	0	0	0
3	0	0	0	714	0	0	1	0	0	0	0	0
4	0	0	0	0	262	0	11	0	0	0	0	0
5	0	0	0	0	2	529	2	0	0	0	0	0
6	0	0	0	0	1	1	665	0	0	0	0	0
7	0	0	0	0	0	0	0	349	28	2	0	0
8	0	0	0	0	0	0	0	30	265	1	0	0
9	0	0	0	0	0	0	0	0	0	520	9	0
10	0	0	0	0	0	0	0	0	0	2	749	0
11	0	0	0	0	0	0	0	0	0	0	0	83

Fig. 19. Confusion matrix of Vanilla LSTM on balanced data with hybrid sampling.



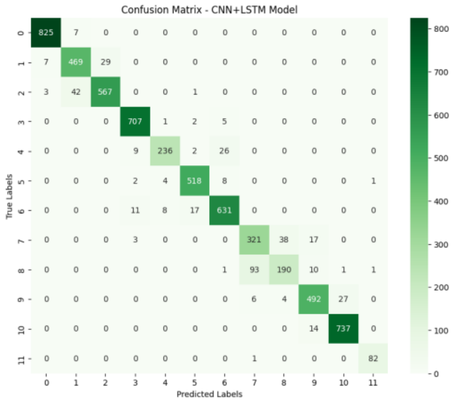
	0	1	2	3	4	5	6	7	8	9	10	11
0	831	0	1	0	0	0	0	0	0	0	0	0
1	2	490	12	0	0	1	0	0	0	0	0	0
2	2	4	607	0	0	0	0	0	0	0	0	0
3	0	0	0	714	0	0	1	0	0	0	0	0
4	0	0	0	1	260	1	11	0	0	0	0	0
5	0	0	0	1	1	527	4	0	0	0	0	0
6	0	0	0	4	0	2	661	0	0	0	0	0
7	0	0	0	0	0	0	0	359	14	6	0	0
8	0	0	0	0	0	0	0	73	221	2	0	0
9	0	0	0	0	0	0	0	0	1	524	4	0
10	0	0	0	0	0	0	0	0	0	3	748	0
11	0	0	0	0	0	0	0	0	0	0	0	83

Fig. 20. Confusion matrix of 2 stacked LSTM on balanced data with hybrid sampling.



	0	1	2	3	4	5	6	7	8	9	10	11
0	830	1	1	0	0	0	0	0	0	0	0	0
1	3	494	8	0	0	0	0	0	0	0	0	0
2	0	10	603	0	0	0	0	0	0	0	0	0
3	0	0	0	715	0	0	0	0	0	0	0	0
4	0	0	0	0	268	1	4	0	0	0	0	0
5	0	0	0	1	3	529	0	0	0	0	0	0
6	0	0	0	0	3	3	661	0	0	0	0	0
7	0	0	0	0	0	0	0	350	28	1	0	0
8	0	0	0	0	0	0	0	30	264	2	0	0
9	0	0	0	0	0	0	0	0	1	524	4	0
10	0	0	0	0	0	0	0	0	0	2	749	0
11	0	0	0	0	0	0	0	0	0	0	0	83

Fig. 21. Confusion matrix of 3 stacked LSTM on balanced data with hybrid sampling.



	0	1	2	3	4	5	6	7	8	9	10	11
0	825	7	0	0	0	0	0	0	0	0	0	0
1	7	469	29	0	0	0	0	0	0	0	0	0
2	3	42	567	0	0	1	0	0	0	0	0	0
3	0	0	0	707	1	2	5	0	0	0	0	0
4	0	0	0	9	236	2	26	0	0	0	0	0
5	0	0	0	2	4	518	8	0	0	0	0	1
6	0	0	0	11	8	17	631	0	0	0	0	0
7	0	0	0	3	0	0	0	321	38	17	0	0
8	0	0	0	0	0	0	1	93	190	10	1	1
9	0	0	0	0	0	0	0	6	4	492	27	0
10	0	0	0	0	0	0	0	0	0	14	737	0
11	0	0	0	0	0	0	0	1	0	0	0	82

Fig. 22. Confusion matrix of CNN-LSTM on balanced data with hybrid sampling.

E. Comparative Results Analysis

In this research paper, a comparative study was conducted employing four distinct deep learning models: Vanilla LSTM, 2 Stacked LSTM, 3 Stacked LSTM, and hybrid CNN-LSTM. The study aimed to address the challenge of class imbalance in Human Activity Recognition (HAR) through the utilization of three sampling techniques: SMOTE, Random Undersampling, and a novel Hybrid Sampling approach. The performance of these models was evaluated based on key metrics, including accuracy, F1 score, precision, and recall.

Accuracy Comparison of Deep Learning Models on Different Sampling Techniques (illustrated in Fig. 23):

- For models trained on imbalanced data, the 2 Stacked LSTM model exhibited the highest accuracy, achieving 0.9531. It was closely followed by the Vanilla LSTM model with an accuracy of 0.9257.
- When using SMOTE to balance the data, the Vanilla LSTM model performed remarkably well, with an accuracy of 0.9499. The 2 Stacked LSTM also showed strong performance with an accuracy of 0.9438.
- For Hybrid Sampling, the models reached even higher accuracy. The 2 Stacked LSTM achieved an accuracy of 0.9755, and the Vanilla LSTM excelled further with an impressive accuracy of 0.9821. The 3 Stacked LSTM model with Hybrid Sampling exhibited the most remarkable performance, achieving an accuracy of 0.9828.

F1-score Comparison of Deep Learning Models on Different Sampling Techniques (see Fig. 24):

- In terms of F1 score, similar trends were observed. The 2 Stacked LSTM model performed exceptionally well across all sampling techniques, reaching an F1 score of 0.9529 for imbalanced data and 0.9415 for data balanced with SMOTE.

- The models with Hybrid Sampling outperformed the others in F1 score. The 2 Stacked LSTM model achieved an F1 score of 0.9752, and the Vanilla LSTM excelled with an impressive F1 score of 0.9821. The 3 Stacked LSTM model with Hybrid Sampling exhibited the most remarkable performance, with an F1 score of 0.9828.

Precision Comparison of Deep Learning Models on Different Sampling Techniques (see Fig. 25):

- Precision results followed a similar pattern. The 2 Stacked LSTM model consistently showed high precision across all sampling techniques, with values ranging from 0.9536 to 0.9470.
- When using Hybrid Sampling, precision levels were remarkably high, with the models achieving precision values ranging from 0.9764 to 0.9537.
- The 3 Stacked LSTM model with Hybrid Sampling exhibited the most exceptional performance, with a precision of 0.9828.

Recall Comparison of Deep Learning Models on Different Sampling Techniques (see Fig. 26):

- Recall rates were also in line with accuracy and F1 score trends. The 2 Stacked LSTM model exhibited high recall, especially with Hybrid Sampling, where it reached a recall rate of 0.9755.
- The Vanilla LSTM model also performed well, achieving recall rates ranging from 0.9438 to 0.9499.
- The 3 Stacked LSTM model with Hybrid Sampling showed the most impressive result, with a Recall of 0.9828.

In summary, this study conclusively demonstrates the efficacy of hybrid sampling techniques in effectively addressing class imbalance challenges in HAR. The proposed models consistently achieve good results, especially the 3 Stacked LSTM, surpassing other models in terms of accuracy, precision, recall, and F1 scores. This underscores the crucial importance of balancing data for better-performing deep models. The comparative plots in Fig. 23 to Fig. 26 provide a visual representation of these findings.

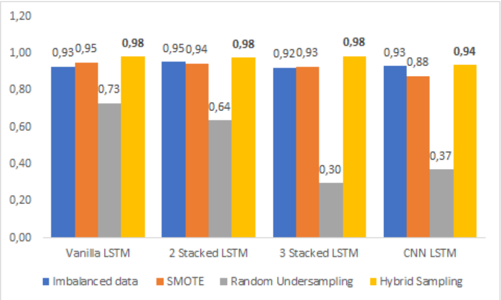


Fig. 23. Accuracy comparison of deep learning models on different sampling techniques.

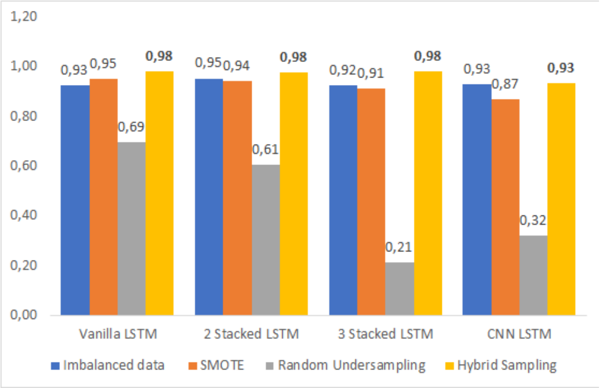


Fig. 24. F1 score comparison of deep learning models on different sampling techniques.

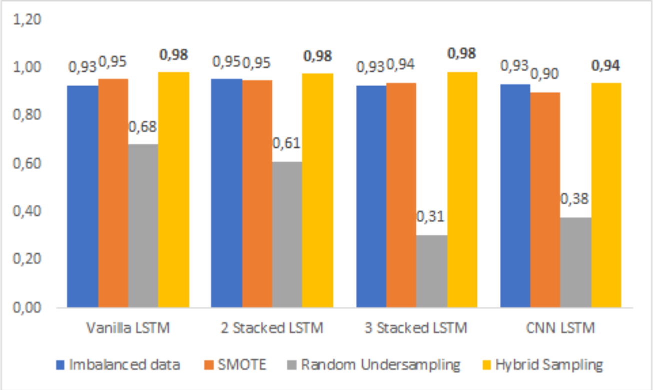


Fig. 25. Precision comparison of deep learning models on different sampling techniques.

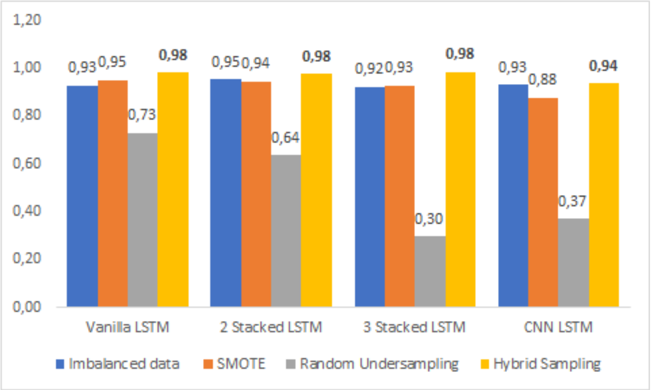


Fig. 26. Recall comparison of deep learning models on different sampling techniques.

Comparison with Previous Studies:

Previous research has extensively explored diverse deep-learning models for Human Activity Recognition (HAR) using the PAMPA2 dataset. As demonstrated in Table XV, these prior studies have yielded impressive outcomes. In 2022, an exemplary convLSTM Autoencoder (AE) model exhibited remarkable accuracy, recording a value of 0.9433, along with an F1 score of 0.9446 [4]. Similarly, in 2023, a Bi-LSTM model demonstrated commendable performance, achieving a high accuracy of 0.9341 and an F1 score of 0.9341, complemented by notable precision and recall values [17].

TABLE XV. COMPARISON WITH PREVIOUS WORKS

Study	year	Dataset	Classification method	Accuracy	F1 score	Precision	Recall
[15]	2020	PAMAP2	LSTM	0.8580	0.8534	0.8651	0.8467
[15]	2020	PAMAP2	LSTM	0.8580	0.8534	0.8651	0.8467
[16]	2022	PAMAP2	LSTM	0.8920	0.8949	0.8969	0.8928
[4]	2022	PAMAP2	convLSTM AE	**0.9433**	**0.9446**	-	-
[18]	2022	PAMAP2	CNN-BiLSTM	0.9429	-	-	-
[16]	2022	PAMAP2	LSTM	0.8920	0.8949	0.8969	0.8928
[18]	2022	PAMAP2	CNN-BiLSTM	0.9429	-	-	-
[7]	2022	PAMAP2	MLP with SMOTE	--	0.7473	0.7769	0.7493
[17]	2023	PAMAP2	Bi-LSTM	0.9341	0.9341	0.9341	0.9347
This Study	2023	PAMAP2	Vanilla LSTM on imbalanced data	0.9257	0.9250	0.9281	0.9257
	2023	**PAMAP2**	**2 Stacked lstm on imbalanced data**	**0.9531**	**0.9529**	**0.9536**	**0.9531**
	2023	PAMAP2	3 Stacked LSTM on imbalanced data	0.9232	0.9232	0.9268	0.9232
	2023	PAMAP2	CNN-LSTM on imbalanced data	0.9308	0.9297	0.9341	0.9308
	2023	**PAMAP2**	**Vanilla LSTM with SMOTE**	**0.9499**	**0.9503**	**0.9537**	**0.9499**
	2023	**PAMAP2**	**2 Stacked LSTM with SMOTE**	**0.9438**	0.9415	**0.9470**	**0.9438**
	2023	PAMAP2	3 Stacked LSTM with SMOTE	0.9282	0.9129	0.9386	0.9282
	2023	PAMAP2	CNN-LSTM with SMOTE	0.8756	0.8687	0.8975	0.8756
	2023	PAMAP2	Vanilla LSTM with Random Undersampling	0.7295	0.6946	0.6812	0.7295
	2023	PAMAP2	2 Stacked LSTM with Random Undersampling	0.6361	0.6070	0.6113	0.6361
	2023	PAMAP2	3 Stacked LSTM with Random Undersampling	0.2953	0.2132	0.3058	0.2953
	2023	PAMAP2	CNN-LSTM with Random Undersampling	0.3706	0.3194	0.3767	0.3706
	2023	PAMAP2	CNN-LSTM with Hybrid Sampling	0.9351	0.9342	0.9350	0.9351
	2023	**PAMAP2**	**2 Stacked LSTM with Hybrid Sampling**	**0.9755**	**0.9752**	**0.9764**	**0.9755**
	2023	**PAMAP2**	**Vanilla LSTM with Hybrid Sampling**	**0.9821**	**0.9821**	**0.9822**	**0.9822**
	2023	**PAMAP2**	**3 Stacked LSTM with Hybrid Sampling**	**0.9828**	**0.9828**	**0.9828**	**0.9828**

Finally, our study demonstrates the effectiveness of Hybrid Sampling techniques in addressing class imbalance in HAR, leading to higher accuracy, precision, recall, and F1 scores. These models consistently outperformed the best-performing models from previous research, underscoring their potential to significantly enhance the accuracy and reliability of HAR systems and demonstrating the importance of tackling the imbalanced data problem.

V. DISCUSSION

Prior studies such as [6], [24] have highlighted the lack of works that address and investigate the impact of the class imbalance problem in human activity recognition. This present study fills this gap by comparing three sampling approaches, SMOTE, Random Undersampling, and Hybrid sampling to reduce the class imbalance and substantially improve human activity recognition (HAR) performance.

In this section, a comprehensive discussion of the experimental findings and their implications for the field of HAR using deep learning models is presented. The consideration encompasses the following key aspects: the impact of class imbalance, the effectiveness of sampling techniques, and the significance of hyperparameter tuning.

1) Hyperparameter tuning enhances model adaptability and performance: In all the experiments, hyperparameter tuning was applied in each scenario, proving to be a highly beneficial approach. The optimization of hyperparameters for each experiment ensured that the deep learning models were tailored to perform optimally under specific conditions. This adaptability is crucial in real-world applications where data characteristics and sampling techniques may vary. Moreover, hyperparameter tuning significantly contributed to the fairness of this comparative analysis. It prevented any model from having an unfair advantage due to suboptimal hyperparameters, ensuring a more equitable evaluation of different sampling techniques.

Overall, the inclusion of hyperparameter tuning in this experimental design serves as a robust foundation for meaningful comparisons and insights into HAR.

2) Addressing class imbalance with sampling techniques: The experiments aimed to investigate the impact of different sampling techniques on the performance of deep learning models in HAR. To address this, four experiments were conducted, each involving variations in data preprocessing and sampling, and each of them incorporated hyperparameter tuning.

The results clearly demonstrate the notable impact of sampling techniques on model performance, further enhanced by hyperparameter tuning.

In Experiment 2, following the application of SMOTE and Hyperparameter Tuning, substantial improvements in accuracy, F1 score, precision, and recall were observed across all models. This underscores the effectiveness of SMOTE in addressing the class imbalance issue, especially when combined with optimal hyperparameters. The balanced dataset led to enhanced recognition efficiency, with significant gains in accuracy and F1 score.

In Experiment 3, involving Random under-sampling and Hyperparameter Tuning, the models exhibited decreased performance compared to the baseline.

In Experiment 4, employing hybrid sampling and hyperparameter tuning, remarkable results were achieved. By combining the strengths of SMOTE and Random Undersampling with fine-tuned hyperparameters, high accuracy and F1 scores were achieved, surpassing the baseline. This confirms the potential of hybrid sampling as a powerful technique for enhancing model performance, especially when hyperparameters are tuned effectively.

Hybrid sampling demonstrates its effectiveness in balancing data by leveraging the strengths of both oversampling (SMOTE) and undersampling (Random Undersampling) techniques. It begins by oversampling the minority class, increasing its representation, and then follows with undersampling the majority class to reduce redundancy. This approach enhances model performance, mitigates overfitting, and ensures that deep learning models are exposed to a more representative and diverse distribution of data. Consequently, these factors contribute to improved generalization, enabling models to make more accurate predictions. It is the combination of these advantages that positions hybrid sampling as an outperforming technique compared to other sampling methods.

3) Model performance and generalization: The findings suggest that deep learning models trained on balanced datasets exhibit improved performance compared to those trained on imbalanced data. This result highlights the significance of addressing class imbalance in HAR applications. Furthermore, these models demonstrated robust generalization capabilities, indicating their potential for real-world deployment.

4) Practical implications: The practical implications of this research extend to various applications, including healthcare, fitness tracking, and human-computer interaction. By improving the accuracy and reliability of HAR systems through both sampling techniques and hyperparameter tuning, this work contributes to enhancing user experiences and promoting healthier lifestyles.

5) Limitations and future work: It's important to acknowledge the limitations of this study. The choice of datasets, model architectures, and hyperparameters may impact the generalizability of the findings. Future research could explore additional datasets, and more complex model architectures, and further investigate hyperparameter tuning techniques. Additionally, the real-world deployment of HAR systems should consider challenges related to sensor placement, data privacy, and user variability.

In conclusion, this study emphasizes the critical role of both sampling techniques and hyperparameter tuning in improving the performance of deep learning models for HAR. SMOTE and hybrid sampling methods, when coupled with effective hyperparameter tuning, demonstrate their effectiveness in addressing class imbalance. The achievement of enhanced accuracy and F1 scores through these combined techniques paves the way for more reliable and efficient HAR systems with broader applications.

VI. Conclusion

In this extensive study on Human Activity Recognition (HAR) using deep learning models and wearable sensor data, the goal was to enhance the accuracy and reliability of HAR systems, which are crucial in healthcare and sports analytics. The challenge of imbalanced datasets in HAR was addressed by exploring different sampling techniques: Synthetic Minority Over-sampling Technique (SMOTE), random undersampling, and hybrid sampling (a combination of SMOTE and random undersampling). These techniques were tested with various deep learning models, including Vanilla LSTM, 2 Stacked LSTM, 3 Stacked LSTM, and Hybrid CNN-LSTM. The findings showed significant improvements in model performance when using sampling techniques to balance the data. SMOTE and hybrid sampling were particularly effective in countering class imbalance, leading to notable enhancements in model accuracy, precision, recall, and the F1 score. The importance of hyperparameter tuning, involving adjustments to specific model settings, was also highlighted. By fine-tuning these parameters, even better model performance was achieved, emphasizing the critical connection between data preprocessing and parameter configuration. As wearable sensors become more prevalent, this research contributes to the creation of systems that can better understand and interpret human actions in various real-world scenarios. Future work will involve experiments with more diverse public datasets, the exploration of more complex deep learning models, and the investigation of additional sampling techniques to further advance the field of Human Activity Recognition

References

[1] S. S. Zhang et al., "Deep Learning in Human Activity Recognition withWearable Sensors: A Review on Advances," Sensors, vol. 22, no. 4, p. 1476, Feb. 2022, doi: 10.3390/s22041476.

[2] M. Gochoo, F. Alnajjar, T. H. Tan, and S. Khalid, "Towards privacy - preserved aging in place: A systematic review," Sensors, vol. 21, no. 9, p. 3082, Apr. 2021, doi: 10.3390/s21093082.

[3] A. Kristoffersson and M. Lindén, "A systematic review on the use of wearable body sensors for health monitoring: A qualitative synthesis," Sensors (Switzerland), vol. 20, no. 5, p. 1502, Mar. 2020, doi: 10.3390/s20051502.

[4] D. Thakur, S. Biswas, E. S. L. L. Ho, and S. Chattopadhyay, "ConvAE-LSTM: Convolutional Autoencoder Long Short-Term Memory Network

for Smartphone-Based Human Activity Recognition," IEEE Access, vol. 10, pp. 4137–4156, Jun. 2022, doi: 10.1109/ACCESS.2022.3140373.

[5] J. Yu, A. de Antonio, and E. Villalba-Mora, "Deep Learning (CNN, RNN) Applications for Smart Homes: A Systematic Review," Computers, vol. 11, no. 2, pp. 1–32, Feb. 2022, doi: 10.3390/computers11020026.

[6] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Comput. Surv., vol. 52, no. 4, 2019, doi: 10.1145/3343440.

[7] F. Alharbi, L. Ouarbya, and J. A. Ward, "Comparing Sampling Strategies for Tackling Imbalanced Data in Human Activity Recognition," Sensors, vol. 22, no. 4, pp. 1–20, 2022, doi: 10.3390/s22041373.

[8] D. Roggen et al., "Collecting complex activity datasets in highly rich networked sensor environments," INSS 2010 - 7th Int. Conf. Networked Sens. Syst., pp. 233–240, 2010, doi: 10.1109/INSS.2010.5573462.

[9] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity Recognition using Cell Phone Accelerometers," 2010.

[10] A. A. A. A. Alani, G. Cosma, and A. Taherkhani, "Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning," in Proceedings of the International Joint Conference on Neural Networks, 2020. doi: 10.1109/IJCNN48605.2020.9207697.

[11] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in ACM Int. Conf. Proceeding Ser, 2012.

[12] D. H. Jeong, S. E. Kim, W. H. Choi, and S. H. Ahn, "A Comparative Study on the Influence of Undersampling and Oversampling Techniques for the Classification of Physical Activities Using an Imbalanced Accelerometer Dataset," Healthc., vol. 10, no. 7, 2022, doi: 10.3390/healthcare10071255.

[13] R. A. Hamad, M. Kimura, and J. Lundström, "Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments," SN Comput. Sci., vol. 1, no. 4, pp. 1–10, 2020, doi: 10.1007/s42979-020-00211-1.

[14] A. A. Alani, G. Cosma, and A. Taherkhani, "Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning," in Proceedings of the International Joint Conference on Neural Networks, 2020. doi: 10.1109/IJCNN48605.2020.9207697.

[15] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep Learning Models for Real-time Human Activity Recognition with Smartphones," Mob. Networks Appl., vol. 25, no. 2, pp. 743–755, Apr. 2020, doi: 10.1007/s11036-019-01445-x.

[16] Y. Xu and L. Zhao, "Inception-LSTM Human Motion Recognition with Channel Attention Mechanism," Comput. Math. Methods Med., vol. 2022, 2022, doi: 10.1155/2022/9173504.

[17] A. Tehrani, M. Yadollahzadeh-Tabari, A. Zehtab-Salmasi, and R. Enayatifar, "Wearable Sensor-Based Human Activity Recognition System Employing Bi-LSTM Algorithm," Comput. J., no. April, 2023, doi: 10.1093/comjnl/bxad035.

[18] S. K. Challa, A. Kumar, and V. B. Semwal, "A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data," Vis. Comput., vol. 38, no. 12, pp. 4095–4109, 2022, doi: 10.1007/s00371-021-02283-3.

[19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[20] C. Liu, L. Zhang, Z. Liu, K. Liu, X. Li, and Y. Liu, "Lasagna: Towards deep hierarchical understanding and searching over mobile sensing data," Proc. Annu. Int. Conf. Mob. Comput. Networking, MOBICOM, vol. 0, no. 1, pp. 334–347, Oct. 2016, doi: 10.1145/2973750.2973752.

[21] S. Mekruksavanich and A. Jitpattanakul, "Lstm networks using smartphone data for sensor-based human activity recognition in smart homes," Sensors, vol. 21, no. 5, pp. 1–25, 2021, doi: 10.3390/s21051636.

[22] N. S. Khan and M. S. Ghani, A Survey of Deep Learning Based Models for Human Activity Recognition, vol. 120, no. 2. Springer US, 2021. doi: 10.1007/s11277-021-08525-w.

[23] "KerasTuner." https://keras.io/keras_tuner/ (accessed Jul. 23, 2023).

[24] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," ACM Comput. Surv., vol. 54, no. 4, Jul. 2021, doi: 10.1145/3447744.

APPENDIX A

TABLE A1: DATASET INSTANCES BEFORE AND AFTER APPLYING SAMPLING METHODS

Activity ID	Class ID	# Instances in training set of the imbalanced data	# Instances in the testing set	# Instances training set After SMOTE	# Instances training set After Random Undersampling	# Instances training set After Hybrid Sampling
1	0	100298	42633	100298	10889	1968
2	1	58380	25358	100298	10889	1160
3	2	70165	29808	100298	10889	1436
4	3	86117	36789	100298	10889	1723
5	4	30100	12950	100298	10889	599
6	5	63755	27585	100298	10889	1249
7	6	78154	33678	100298	10889	1546
12	7	41442	17872	100298	10889	849
13	8	32800	14030	100298	10889	661
16	9	60703	26256	100298	10889	1226
17	10	87885	37343	100298	10889	1774
24	11	10889	4564	100298	10889	221