Fairness in Human Activity Recognition from Wearable Inertial Sensor Data

by

Shaily Roy
20166051

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
August 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Shaily

Shaily Roy
20166051

Approval

The thesis/project titled "Fairness in Human Activity Recognition from Wearable Inertial Sensor Networks" submitted by

1. Shaily Roy (20166051)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on September 7, 2023.

Examining Committee:

Examiner:
(External)

Dr. Ashis Talukder
Associate Professor
Department of Management Information Systems (MIS)
University of Dhaka

Examiner:
(Internal)

Dr. Muhammad Iqbal Hossain
Associate Professor
Department of Computer Science and Engineering
Brac University

Examiner:
(Internal)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Amitabha Chakrabarty
Professor
Department of Computer Science and Engineering
Brac University

Chairperson:

Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Human Activity Recognition (HAR) has gained significant attention in recent years due to its potential applications in various domains such as healthcare, smart environments, and human-computer interaction. As HAR technologies become more pervasive, there is a growing concern about the fairness and equity aspects associated with their deployment. In this study, we address the issue of fairness in human activity recognition models. We begin by analyzing the data and identifying an unfairness gap in our initial model, which was based on a neural network architecture known as Multi-Layer Perceptron (MLP). To quantify the unfairness, we employ two well-established fairness metrics: equalized odds (EO) gap and demographic parity (DP) gap. To bridge this gap, two approaches were proposed, incorporating fairness losses: demographic parity loss and equalized odds loss. These losses aim to eliminate gender bias in the model's predictions. To optimize the model with fairness losses, Lagrangian multiplier is employed, achieving a balance between accuracy and fairness in each training epoch. The results demonstrate the successful transformation of the initial unfair model into a fairer one, particularly addressing gender bias. The proposed approach attains a strong accuracy rate, approximately 95% for certain activities and over 75% for others. Furthermore, we compare our approach with state of art machine learning methods such as logistic regression and decision trees, further validating its effectiveness. This work contributes to advancing fairness-aware machine learning techniques for human activity recognition, promoting ethical and unbiased AI systems.

Keywords: Neural Network; Deep Learning; Human Activity Recognition; Fairness; Multi-Layer Perceptron; Demographic Pairity; Equalized Odds: Lagrangian Multiplier

Acknowledgement

Firstly, all praise to the Almighty for whom the thesis has been completed without any major interruption.
Secondly, I express my deepest gratitude to my advisor, Md. Golam Rabiul Alam, for his kind support and valuable advice throughout the course of this work. His guidance has been instrumental in shaping this endeavor, and his willingness to assist me whenever I needed help is deeply appreciated.
Lastly, I extend my heartfelt thanks to my parents. Their unwavering support and constant encouragement have been the pillars that have carried me through this journey. Without their steadfast presence, it would not have been possible to reach this point. Their kind support and prayers have paved the way for my imminent graduation.

Table of Contents

Declaration i

Approval ii

Abstract iv

Acknowledgment v

Table of Contents vi

List of Figures viii

List of Tables x

Nomenclature xi

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Aims and Objectives . . . 2
- 1.3 Contribution . . . 3

2 Literature Review 5
- 2.1 Human Activity Recognition . . . 5
- 2.2 Fairness in Machine Learning . . . 7
- 2.3 Lagrangian Multiplier Method in Constrained Optimization . . . 9
- 2.4 Related Work . . . 10

3 Methodology 13
- 3.1 Dataset and Features . . . 14
- 3.2 Data Pre-Processing . . . 15
  - 3.2.1 Data Distribution . . . 20
- 3.3 Feature Engineering: Enhancing the Dataset . . . 21
  - 3.3.1 Lag Features Creation . . . 21
  - 3.3.2 Handling Missing Data and Transformations . . . 21
  - 3.3.3 t-SNE Dimensionality Reduction . . . 21
- 3.4 Prediction Models Specification . . . 23
  - 3.4.1 Logistic Regression . . . 23
  - 3.4.2 Multi-Layer Perceptron . . . 24
- 3.5 Fairness Gap . . . 27
- 3.6 Introducing Fair Loss . . . 28

3.6.1 Equalized Odds Loss 28
3.6.2 Demographic Parity Loss 28
3.6.3 Lagrangian Method 29

4 Result Analysis **31**
4.1 Prediction Model Performance 31
4.2 Fairness Evaluation 34
4.3 Discussion 38

5 Conclusion **40**

Bibliography **45**

Appendix A **46**

Appendix B **53**

Appendix C **55**

List of Figures

2.1 HAR research and the sensing technologies used [9] 6
2.2 The general framework of human activity recognition (HAR) using machine learning approaches [22] 7

3.1 Proposed Model 13
3.2 Data Correlation 15
3.3 Activity Distribution 16
3.4 Similarity in Acceleration Data across Three Axes 16
3.5 Similarity in Gyroscope Data across Three Axes 17
3.6 Distribution of Mean Acceleration Values for Static and Dynamic Activities 18
3.7 Distribution of Mean Gyroscope Values for Static and Dynamic Activities 18
3.8 Boxplot of the normalized maximum acceleration (`acc_norm_max`) for all activities 19
3.9 Boxplot of the mean normalized acceleration (`acc_norm_mean`) for all activities 19
3.10 Activity Percentage 20
3.11 t-SNE Dimension Reduction Plot 23
3.12 Proposed MLP Model 25

4.1 Confusion Matrix of MLP Model for Valid and Test Dataset 33
4.2 Synthetic Data 37
4.3 Fairness Gap Comparison Graph for Fair and Unfair MLP Models . . 38

1 Boxplot of the normalized minimum acceleration (`acc_norm_min`) for all activities 46
2 Boxplot of the standard deviation of normalized acceleration (`acc_norm_std`) for all activities 47
3 Boxplot of the interquartile range of normalized acceleration (`acc_norm_iqr`) for all activities 47
4 Boxplot of the median absolute deviation of normalized acceleration (`acc_norm_mad`) for all activities 48
5 Boxplot of the difference in normalized acceleration (`acc_norm_diff`) for all activities 48
6 Boxplot of the minimum normalized gyroscope reading (`gyro_norm_amin`) for all activities 49
7 Boxplot of the maximum normalized gyroscope reading (`gyro_norm_amax`) for all activities 49

8 Boxplot of the mean normalized gyroscope reading (`gyro_norm_mean`) for all activities 50
9 Boxplot of the standard deviation of normalized gyroscope reading (`gyro_norm_std`) for all activities 50
10 Boxplot of the interquartile range of normalized gyroscope reading (`gyro_norm_iqr`) for all activities 51
11 Boxplot of the median absolute deviation of normalized gyroscope reading (`gyro_norm_mad`) for all activities 51
12 Boxplot of the range of normalized gyroscope reading (`gyro_norm_ran`) for all activities 52
13 Selected Features from the Dataset 54
14 Dimension Reduction using TSNE (t-Distributed Stochastic Neighbor Embedding) LOGS 54

List of Tables

3.1 Sample Data from the Dataset 14

4.1 State of Art Accuracy of Logistic Regression (LR) and Decision Tree (DT) 32
4.2 Multi-Layer Perceptron (MLP) Accuracy of Activities 32
4.3 Fairness Gaps for Different Activities 34
4.4 Fair MLP Performance Comparison using EO Loss and DP Loss for Validation Data 35
4.5 Fair MLP Performance Comparison using EO Loss and DP Loss for Test Data 35
4.6 Fair MLP with Lagrangian Multiplier Performance Comparison for Valid Data 36
4.7 Fair MLP with Lagrangian Multiplier Performance Comparison for Test Data 36
4.8 Unfair MLP Model Performance Metrics for Random Data 36
4.9 Fair MLP with Lagrangian Multiplier Performance Comparison for Randomly Generated Data 37

1 Unfair MLP Results and Fairness Gap for Each Fold 55
2 MLP Results and Fairness Gap for Each Fold using EO Loss 55
3 Fair MLP Results using DP Loss and Lagrangian Multiplier 56
4 Fair MLP Results using EO Loss and Lagrangian Multiplier 56

Nomenclature

The following list describes several symbols & abbreviations that will be later used within the body of the document

β	Beta
γ	Gamma
σ	Sigma
e	Exponential
$\sum$	Summation
α	Alpha
λ	Lambda
E	Expected Value

Chapter 1

Introduction

Machine learning algorithms make predictions by identifying patterns within training datasets and applying those patterns to new instances. However, this seemingly objective process can have biases and reinforce stereotypes, particularly when it comes to gender [1]. In this context, the importance of fairness becomes evident. Fairness in technology, specifically within the realm of human activity recognition based on gender, poses significant challenges that deserve careful consideration. It is crucial to recognize that fairness extends beyond mere technical accuracy. In healthcare, the ramifications are critical, potentially leading to misdiagnoses that disproportionately affect specific patient groups. For instance, an AI system that inaccurately assesses certain activities as less indicative of severe health issues in particular genders or age groups can result in delayed or incorrect medical interventions, jeopardizing lives. In the realm of criminal justice, biased predictive tools can perpetuate profiling and sentencing disparities, as algorithms may erroneously flag activities as more indicative of criminal behavior for specific racial or socioeconomic demographics. The consequences extend to the financial sector, where biased credit scoring algorithms can hinder access to credit and financial resources for certain demographic groups, exacerbating economic disparities. In education, unfair recognition of activities can compound existing inequalities by shaping educational opportunities and outcomes based on inaccurate assessments. Biased hiring systems, influenced by discriminatory activity recognition, undermine efforts to build diverse and inclusive workplaces. The significance of fairness is not confined to human welfare alone; it also impacts technological advancements. In autonomous vehicles, fair recognition of activities on the road is essential for road safety, as biased algorithms might underestimate the presence of pedestrians or misclassify certain activities as less risky, potentially leading to accidents and safety concerns. Moreover, the retail and advertising industries must steer clear of discriminatory practices driven by biased recognition, as this could alienate customers and result in missed opportunities for businesses. In each of these contexts, the pursuit of fairness is paramount to upholding ethical standards, fostering inclusivity, and preventing the exacerbation of societal inequalities [2]. The implications of failing to account for gender-based differences in activity recognition are profound. AI systems may erroneously classify activities as gender-specific, such as assuming that men are more likely to engage in physically active pursuits while underestimating the physical activities of women. As we delve deeper into the challenges surrounding fairness in human activity recognition based on gender, it becomes evident that we must grapple with the consequences of gender-related assumptions, the potential for biased outcomes, and the broader societal impact of algorithmic decisions. Through a deliberate exploration and implementation of fair AI practices, we can strive to ensure that technology contributes positively to society, promoting inclusivity and equality for individuals of all genders.

1.1 Motivation

The motivation behind this study stems from a deep recognition of the ethical concerns associated with biased human activity recognition models. Several researchers acknowledges that deploying such models can perpetuate discrimination, reinforce stereotypes, and lead to unfair outcomes for individuals belonging to certain demographic groups [3]. In domains such as healthcare, surveillance, and personal assistance, the consequences of biased decision-making can result in unequal access to resources and unfair treatment based on demographic attributes [4]. By striving to mitigate social disparities and promote fairness, the study aims to make a positive impact on society at large. While fairness is a paramount consideration, the study also recognizes the importance of maintaining a reasonable level of accuracy and performance in human activity recognition. Finding a balance between fairness and performance is crucial to ensure that the models are not only fair but also effective in accurately recognizing and classifying human activities.
Another thought behind the research lies in identifying a research gap in the domain of fairness in human activity recognition. While fairness issues have been extensively studied in other areas such as hiring, recommendation and lending [5], the specific context of human activity recognition requires tailored approaches to address bias and ensure fairness. By exploring the applicability of fairness metrics, proposing novel fairness losses, and integrating them into the training process, the study aims to contribute to filling this research gap and advancing the field of fairness-aware machine learning techniques in human activity recognition.
Lastly, the work acknowledges the practical implications of developing fairness-aware models that can be deployed in real-world applications. By transforming an initial unfair model into a fairer one, particularly with respect to gender bias, the study offers practical insights and solutions that can be implemented in various domains. This practicality promotes the adoption of fairness-aware machine learning models, fostering the development of more trustworthy AI systems.

1.2 Aims and Objectives

The research objectives revolve around understanding and addressing fairness issues in human activity recognition models. By achieving fairer outcomes, the aim is to promote inclusivity and minimize discrimination in automated systems. The applicability of fairness metrics such as EO gap and DP gap are explored, and a novel approach to integrate fairness losses into the training procedure is implemented. In essence, the research objectives are not confined solely to theoretical exploration, but also encompass practical methodologies to effectively tackle the fairness challenges that can arise within automated systems. Through this holistic approach, the research aims to contribute to the development of advanced recognition models that not only excel in accuracy but also set new standards for equity in automated decision-making processes.

1.3 Contribution

The contributions of this work include:

1. **Recognition of Human Activity and Development of a Neural Network Model:** The first contribution of this endeavor involves the successful recognition of human activities through the formulation and deployment of a meticulously designed Neural Network Model. This model serves as a foundational framework for subsequent analytical undertakings.

2. **Identification and Quantification of Fairness Gaps:** A pivotal facet of this work is the systematic identification and subsequent quantification of fairness disparities inherent in the initial Multi-Layer Perceptron (MLP)-based human activity recognition model. This discernment underscores the imperative need for addressing such inequities.

3. **Introduction of Fairness Losses for Mitigation:** A pioneering stride in this research is the introduction and incorporation of two widely acknowledged fairness losses, namely the demographic parity loss and the equalized odds gap. These losses are strategically integrated to rectify the previously detected fairness gaps and promote equity within the model's outcomes.

4. **Transformation of Unfair Model to Attain Fairness, Particularly in Gender Bias:** A salient outcome of this study is the transformation of an initial model characterized by inequity into a notably fairer iteration. The focus of this transformation primarily centers upon mitigating gender bias, exemplifying the potential of the proposed techniques to rectify biases within the model's predictions.

5. **Leveraging Lagrangian Multipliers for Fairness-Optimized Model:** The infusion of Lagrangian multipliers into the optimization process stands as a noteworthy achievement in this work. This augmentation facilitates the refinement of the model by concurrently optimizing it with the newly introduced fairness losses, thus culminating in predictions that exhibit commendable fairness while upholding predictive accuracy.

6. **Advancement of Fairness-Aware Machine Learning for Human Activity Recognition:** Conclusively, this work contributes to the advancement of the broader domain of fairness-aware machine learning techniques, particularly in the context of human activity recognition. The insights garnered and techniques formulated hold promise for diverse applications spanning various domains, thus auguring well for the ongoing progress in this field.

The experimentation phase of this study was conducted employing the publicly available (HUGADB) dataset, which has been meticulously created for the specific purpose of human activity analysis. Through the utilization of binary classification for each distinct activity, the model achieves a notable accuracy range of approximately 80% to 90% for different activities, thereby further underscoring the efficacy of the developed techniques within real-world scenarios.

The thesis is structured into five main chapters to ensure a coherent presentation of the research. Chapter 1 introduces the research problem, objectives, methodology, and the overall structure of the thesis. Chapter 2 delves into the literature review, examining existing work in human activity recognition and fairness in machine learning. Chapter 3 details the data collection process, pre-processing techniques, feature engineering, and the employed deep learning model. Chapter 4 presents the results of the experiments and discusses the achieved accuracy and fairness metrics. Finally, Chapter 5 concludes the thesis by summarizing the research, discussing implications, limitations, and suggesting future research directions. Additionally, an appendix provides supplementary materials and in-depth analyses.

Chapter 2

Literature Review

While several researchers have advocated for AI fairness in diverse fields of research, none of them have specifically addressed the obstacles associated with implementing AI fairness in activity recognition based on gender. This section presents an overview of existing literature relevant to fairness considerations in the context of human activity recognition, with a focus on gender-based disparities. We categorize the related work into subsections, highlighting various fairness metrics and techniques employed for mitigating biases in machine learning models. Fairness involves recognizing and addressing underlying biases and ensuring that AI systems do not inadvertently amplify societal inequalities. In the realm of human activity recognition, this means understanding how gender norms and expectations can influence data patterns and then designing models that mitigate any unintended consequences [2] [6].

2.1 Human Activity Recognition

The recognition of human activities plays a crucial role in facilitating human interaction and interpersonal relationships. Extracting this information, which offers insights into a person's identity, personality, and psychological state, is a complex endeavor. The ability of humans to discern and understand the activities of others has become a focal point of study within the fields of computer vision and machine learning. This research has led to the development of various applications, such as video surveillance systems, human-computer interaction, and robotics for analyzing human behavior, which all necessitate a robust multiple activity recognition system [7] [8].
Human activity recognition (HAR) is a complex endeavor that has been approached through two distinctive methodologies: the utilization of external sensors and wearable sensor systems [10]. In the former approach, devices are strategically positioned at specific locations, necessitating voluntary interactions from users to infer activities. Conversely, wearable sensor systems involve the direct attachment of devices to the user's body [11]. Recent research has witnessed growing interest in both depth sensor and wearable sensor technologies, as illustrated in Figure 2.1. Challenges arise from the task of detecting humans in image sequences, which demands substantial computational processing and can potentially impact real-time HAR performance [12]. Additionally, using RGB cameras raises privacy concerns, particularly among vulnerable groups such as the elderly. In contrast, depth sensors offer benefits like af-

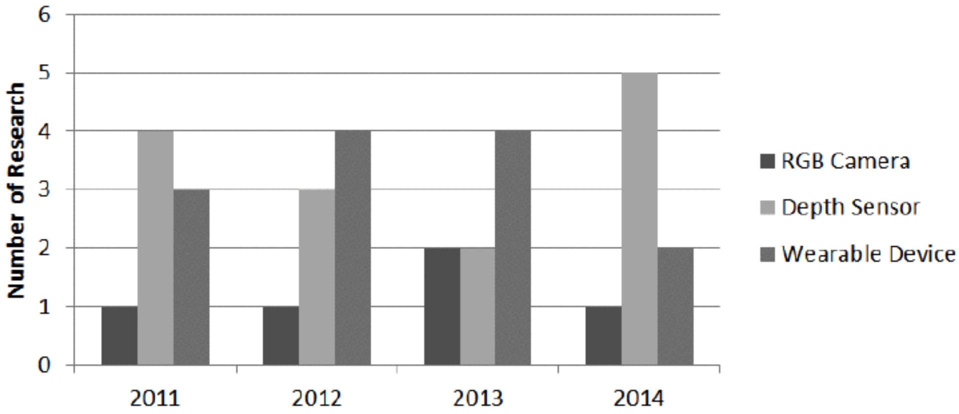

Figure 2.1: HAR research and the sensing technologies used [9]

fordability, high sampling rates, and the ability to merge visual and depth data [13]. On the other hand, wearable sensor systems address limitations like occlusion and viewpoint issues, which are often present in RGB cameras and depth sensors [14]. Wearable sensors offer the advantage of unobtrusive monitoring without disrupting daily activities [15], although maintaining recognition accuracy becomes challenging when dealing with multiple worn sensors [16]. Nevertheless, wearable-based HAR systems have demonstrated promise [17].

Recent efforts have been focused on identifying human activities using wearable sensors. Despite the range of proposed methodologies, existing solutions often operate within predefined configurations and limited sensor parameters. The Human Activity Recognition (HAR) framework serves as a means to gain insights into individual behaviors and actions [18]. This involves collecting signals from intelligent sensors or smartphones, followed by applying machine learning techniques to recognize and analyze these activities. Over the past decade, research centered around sensor-based Human Activity Recognition (HAR) has experienced substantial growth. Various methods have emerged that effectively discern human daily activities. Notably, a study by [19] conducted a comparative analysis between deep learning techniques and traditional pattern recognition methods for sensor-based activity recognition. The findings underscored the effectiveness of deep learning in accurately discerning human activities. This highlights the contemporary trend of adopting cutting-edge approaches to enhance the comprehension of activities recorded by sensors.

This exploration revolves around three fundamental aspects: the type of sensor employed, the intricate architecture of deep learning models, and the specific scenarios in which these methods prove valuable. It's important to acknowledge that there is no universal solution within the realm of machine learning methods [20]. Evidently, not all classification techniques yield optimal performance. Therefore, conducting a comprehensive and systematic exploration of each technique becomes crucial. Another study highlighted the significance of the decision tree algorithm as a state-of-the-art technique for human activity recognition using sensor data [21]. They also proposed neural network models as potential primary tools for this purpose. However, the researchers delved into challenges related to the complexities of data collection and the potential issue of overfitting during data training. Based on their comprehensive survey, the researchers concluded that no single method stands out as the ultimate solution for recognizing all types of activities. Instead, the choice of an appropriate method relies on a careful consideration of various factors specific to the application at hand. Consequently, despite the myriad of available methods, certain challenges persist and await effective resolutions. Figure 2.2 illustrates the conven-

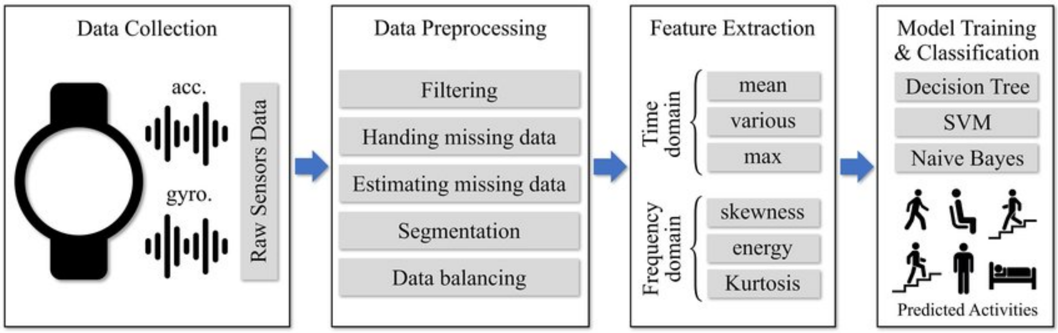

Figure 2.2: The general framework of human activity recognition (HAR) using machine learning approaches [22]

tional approach to HAR, which involves the use of machine learning techniques such as artificial neural networks and decision trees, relying on manually engineered features. However, research suggests that these methods struggle to capture intricate patterns [22]. To address this limitation, there are authors who propose a novel hybrid model, CNN-LSTM, which combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This model automates feature extraction and captures temporal dependencies, showcasing the potential of deep learning in revolutionizing HAR and leveraging the capabilities of the Internet of Things (IoT) for improved activity recognition across diverse domains.

2.2 Fairness in Machine Learning

Research findings indicate that when Machine Learning (ML) algorithms are solely optimized for accuracy improvement using historical data, they not only replicate but can also magnify existing biases or discrimination [23]. The widely accepted definition of AI-fairness states that "discrimination is deemed to exist when a model produces different decisions for two individuals who share a characteristic relevant to decision-making but differ solely in a sensitive attribute, such as gender or race." Deploying unfair AI systems can result in a disparate impact, wherein certain practices disproportionately harm individuals with specific protected characteristics over others [24]. Unfairness can stem from various factors, including biased sampling, erroneous labeling (especially when labeling is subjective), biased representation (such as incomplete or correlated features), suboptimal or insensitive optimization algorithms, shifts in population or data distribution, and a failure to account for domain-specific, legal, or ethical constraints [25]. Addressing bias and discrimination has been a longstanding pursuit within the realms of philosophy, psychology, and more recently, in the field of machine learning. Yet, the journey to combat discrimination and attain fairness begins with the foundational step of defining what fairness truly entails. While philosophical and psychological inquiries have grappled with this notion well before the advent of computer science, the absence of a universally agreed-upon definition underscores the inherent complexity associated with resolving this intricate issue [26]. The absence of a singular, universally accepted definition underscores the intricate nature of this challenge [27]. Diverse cultural perspectives and differing value systems across various societies contribute to the divergence in defining fairness. This complexity arises from the realization that a singular, universally applicable definition remains elusive due to the myriad ways in which fairness is perceived across cultures. This challenge is evident even in the field of computer science, where attempts to introduce novel fairness criteria to algorithms have predominantly originated in the Western world. Despite the extensive utilization of common datasets and problems to showcase the effectiveness of these constraints, a unanimous consensus on the most suitable constraints for specific issues has not yet been reached. In its broadest sense, fairness denotes the absence of bias or preference directed towards an individual or group based on inherent or acquired characteristics, particularly within decision-making contexts. However, the aspiration for fairness often encounters significant practical complexities. Recognizing these challenges, a multitude of fairness definitions have emerged, each tailored to address distinct forms of algorithmic bias and discriminatory practices. These definitions collectively reflect the endeavor to strike a balance between various perspectives and address the intricate interplay between societal norms, data, and technological advancements. Group fairness and individual fairness are pivotal concepts in the realm of fairness and ethics within machine learning [28]. These concepts target distinct aspects of bias and discrimination, with the shared goal of enhancing fairness in algorithmic decision-making. Each perspective confronts the intricate challenges that arise in this pursuit [29].

Group Fairness: Group fairness revolves around ensuring equitable and just treatment for diverse demographic groups through machine learning models. This approach encompasses protected attributes like gender, race, age, and more, which are often used to categorize individuals into distinct groups. The objective of group fairness is to mitigate differential treatment and outcomes among these groups, thereby curbing potential biases and discriminatory outcomes [30].
The concept of group fairness can be expressed as follows: On average, the instances from both groups should have similar labels, with the similarity weighted by the proximity of the labels of the instances [29]. Formally, this is operationalized as:

$$f_2(w, S) = \frac{1}{n_1 n_2} \sum_{(x_i,y_i)\in S_1} \sum_{(x_j,y_j)\in S_2} d(y_i, y_j)\,(w \cdot x_i - w \cdot x_j)^2 \tag{2.1}$$

Here, $f_2(w, S)$ represents the fairness metric for group fairness, where n_1 and n_2 denote the sample sizes of groups S_1 and S_2 respectively. The term $d(y_i, y_j)$ measures the proximity of labels y_i and y_j, while $w \cdot x_i$ signifies the model's prediction for instance x_i.

Individual Fairness: Individual fairness adopts a more nuanced approach, concentrating on treating individual instances impartially instead of focusing on groups. At its core, individual fairness asserts that similar instances should be treated comparably by the model. This implies that if two instances are akin in terms of relevant attributes, their predictions or outcomes should also be similar.
The definition for individual fairness, as articulated in Reference [29], is as follows: For every pair of instances $(x, y) \in S_1$ and $(x', y') \in S_2$, a model w is penalized for its disparate treatment of x and x', with the penalty weighted by a function of $|y - y'|$.

In this context, S_1 and S_2 signify different groups sampled from the population. Formally, this can be mathematically expressed as:

$$f_1(w,S) = \frac{1}{n_1 n_2} \sum_{(x_i,y_i)\in S_1} \sum_{(x_j,y_j)\in S_2} d(y_i,y_j)\,(w\cdot x_i - w\cdot x_j)^2 \tag{2.2}$$

Here, $f_1(w,S)$ denotes the fairness metric for individual fairness, with the same symbols as in the group fairness equation.
In our study, our primary focus will be on exploring the realm of group fairness. Within this context, several metrics and definitions are employed to quantify and assess group fairness, including Equalized Odds, Equal Opportunity, Demographic Parity, and Disparate Impact. Notably, our study narrows its scope to the examination of Equalized Odds and Demographic Parity, as these two metrics encompass and encapsulate the essence of the broader fairness considerations.

Demographic Parity: Demographic parity suggests that each group should receive positive outcomes at similar rates, regardless of their background. This prevents the model from favoring one group over another when making predictions. In mathematical terms, it means that the chance of getting a positive outcome should be the same for different groups, represented as [31]:

$$P(Y|A=0) = P(Y|A=1) \tag{2.3}$$

where:

$$\begin{aligned} P(Y|A=0) &: \text{Probability of positive outcome given } A=0 \\ P(Y|A=1) &: \text{Probability of positive outcome given } A=1 \end{aligned}$$

Equalized Odds: Equalized odds aims for fairness in two ways. First, it ensures that different groups have an equal chance of getting positive outcomes. Second, it aims to distribute mistakes, like false positives, equally among the groups. This prevents the model from favoring any specific group, not only in positive outcomes but also in errors. Mathematically, equalized odds requires that the positive outcome doesn't depend on the protected class, A, and the real outcome, Y, and that false positives are distributed equally among the groups [31]:

$$P(Y=1|A=0,Y=y) = P(Y=1|A=1,Y=y) \tag{2.4}$$

where:

$$\begin{aligned} P(Y=1|A=0,Y=y) &: \text{Probability of positive outcome given } A=0 \text{ and } Y=y \\ P(Y=1|A=1,Y=y) &: \text{Probability of positive outcome given } A=1 \text{ and } Y=y \\ y &: \text{Possible real outcomes, } y\in\{0,1\} \end{aligned}$$

2.3 Lagrangian Multiplier Method in Constrained Optimization

The Lagrangian multiplier method stands as a powerful technique harnessed for solving problems in constrained optimization. This approach becomes particularly

efficacious when confronted with the challenge of optimizing a function while concurrently adhering to specific constraints.
At its core, the Lagrangian multiplier method revolves around minimizing the Lagrangian function $L(x, \lambda)$ by determining optimal values for x and λ. This endeavor entails calculating partial derivatives of L with respect to both x and λ, setting them to zero, and then solving the resultant system of equations. By undergoing this process, we unveil the optimal solution that harmonizes the original objective with the imposed constraints.
In the context of the descent approach, the Lagrangian multiplier method can be leveraged iteratively [32]. An often-used iterative equation in the descent approach takes the form:

$$\lambda x_{k+1} = \lambda x_k + \alpha_k h(x_k) \tag{2.5}$$

In this equation, x_k denotes the current value of optimization variables, λ represents the Lagrange multiplier vector, $h(x_k)$ encompasses the constraint functions evaluated at x_k, and α_k stands as a step size parameter governing the magnitude of the update. Through this equation, the Lagrange multipliers dynamically adjust based on constraint violations, directing the optimization variables toward a solution that reconciles both the objective and the constraints. This iterative process of updating x and λ persists until a convergence criterion is met, ultimately yielding an optimal solution that effectively balances the interplay between the objective and the constraints.

2.4 Related Work

Prior research has witnessed the emergence of various models for activity recognition based on wearable sensors. One notable approach introduced a continuous autoencoder model designed to harness multi-label activity recognition through RFID readers discreetly affixed to the body [33]. Furthermore, another innovative wearable model leveraged RFID readers to delve into radio pattern mining, facilitating the identification of activities across diverse body-mounted locations [34]. In a study conducted by [35], the adoption of a Convolutional Neural Network (CNN) architecture ushered in automated feature learning from raw inputs in the realm of human activity recognition. This pioneering endeavor aimed to tackle the challenges stemming from limited training data size. The arsenal included techniques such as pooling, Rectified Linear Unit (ReLU) activation, and a soft-max classifier, strategically deployed to mitigate the perils of overfitting within the neural network. The overarching goal was to bolster the model's capacity for generalization. Notably, experimental evaluations revealed a marginal but discernible enhancement in classification accuracy when compared to the conventional Support Vector Machines (SVM).
The work illuminated in [36] underscores the pivotal role of the multilayer perceptron (MLP) as a robust tool adept at confronting intricate classification quandaries, particularly within domains such as human activity recognition. The distinctive attributes of MLP, characterized by its multilayer feed-forward architecture and the activation functions embedded within its hidden units, converge to minimize the discrepancy between predicted outcomes and desired network outputs [37]. These antic-

ipated outputs align with class labels within the context of classification tasks. Rigorous evaluation of MLP's performance is achieved through P-fold cross-validation. However, owing to the continuous values spanning from 0 to 1 generated by the output neurons of MLP, crafting a traditional confusion matrix poses considerable challenges. Nevertheless, the study adeptly quantifies MLP's prowess by scrutinizing the counts of accurately and erroneously classified feature vectors. The outcome unveiled an impressive overall correct classification rate of 96.2%, underscoring MLP's effectiveness in navigating the intricacies of human activity recognition.

A noteworthy application of the Lagrangian multiplier method finds resonance in the work of [38].

While direct investigations into fairness within human activity recognition are scarce, a distinctive research endeavor by [39] pivots its focus toward redressing fairness concerns in Human Activity Recognition (HAR), particularly concerning older adults. This multifaceted exploration centered on the identification and implementation of optimal bias mitigation algorithms, with a specific emphasis on addressing age and disability biases. This distinctive undertaking sets itself apart by venturing into the relatively uncharted terrain of single Wireless Sensor Network (WSN)-based multi-label diverse activity recognition tailored explicitly for the older demographic. This ambitious quest encompassed the recognition of a wide spectrum of activities, spanning both upper and lower extremities, while surmounting the challenges inherent to single-device-based recognition. The research team also undertook the pivotal task of scrutinizing age and disability biases lurking within AI-driven activity recognition, yielding invaluable insights into their existence and impact. Within the realm of fairness in machine learning classification, another scholarly pursuit underscores the necessity of accommodating additional constraints in constructing equitable systems, especially in the healthcare sector [40]. Notably, the emphasis revolves around the concept of classification parity, particularly concerning safeguarded attributes such as age, gender, and race. The research measures fairness by evaluating classification parity with respect to predictive performance, delves into the repercussions of removing safeguarded features on machine learning performance and fairness, and delineates the thresholds where models achieve relative fairness while preserving robust predictive performance.

To fortify fairness within classification, they introduce a neural network-based framework christened as FNNC. This pioneering framework aspires to uphold fairness while concurrently safeguarding high classification accuracy. The Lagrangian multiplier method is artfully deployed to seamlessly incorporate fairness constraints into the learning process. This integration entails the strategic modification of the loss function to encapsulate constraints like demographic parity and equal opportunity, thus fostering fairness. The Lagrange multipliers play a pivotal role in orchestrating a harmonious equilibrium between fairness imperatives and accuracy objectives within the loss function. This methodological synergy empowers them to optimize the model's performance while meticulously preserving the requisites of fairness. The Lagrangian multiplier approach provides a structured mechanism for seamlessly amalgamating fairness constraints with the neural network's overarching objective, ultimately culminating in the enhancement of fairness-aware classification.

$$L_{\mathrm{NN}}(h(X), A, Y) = \hat{l}_{\mathrm{CE}}(h(X), Y) + \lambda l_k(h(X), A, Y) \tag{2.6}$$

Here, L_{NN} signifies the comprehensive loss function encompassing Lagrange multipliers, $\hat{l}_{\text{CE}}$ represents the cross-entropy loss, l_k embodies the specific fairness constraint, and λ denotes the Lagrangian multiplier.

Similarly, in another remarkable study by [41], the innovative FairGBM framework takes center stage. This novel paradigm harnesses the power of Lagrangian multipliers for the purpose of fairness-constrained optimization within the domain of Gradient Boosting Decision Trees (GBDT). This pioneering approach transmutes fairness constraints into a dual optimization problem, effectively striking an equilibrium between model performance and the imperative of fairness. The FairGBM conceptual framework casts optimization as a two-player game: one player minimizes loss concerning model output ($\hat{Y}$), while the other maximizes loss with the assistance of Lagrangian multipliers. This strategic interplay culminates in heightened accuracy.

These noteworthy research endeavors collectively serve as a source of inspiration for the adoption of the Lagrangian multiplier method in augmenting fairness within our proposed model.

Chapter 3

Methodology

Our methodology, presented in Figure 3.1, unfolds by first acquiring a sensor dataset containing participant information related to gender, an essential aspect of human activity recognition. Once the data collection phase concludes, we proceed to extract pertinent features from the raw sensor data. To maintain the independence of our analysis, the dataset is then split into distinct training and testing subsets.The core of our approach involves leveraging a deep learning model to train the data until it reaches a predefined accuracy threshold for activity recognition. With this milestone achieved, our attention shifts to the critical aspect of fairness assessment. This evaluation process involves scrutinizing the behavior of the model with respect to participant gender information. The ultimate objective of this assessment is to guarantee the model's predictions and outcomes remain consistent and unbiased across diverse gender categories, upholding the principles of fairness throughout its operations. By following this systematic workflow, we not only achieve high accuracy in activity recognition but also underscore the importance of fairness in the realm of machine learning.

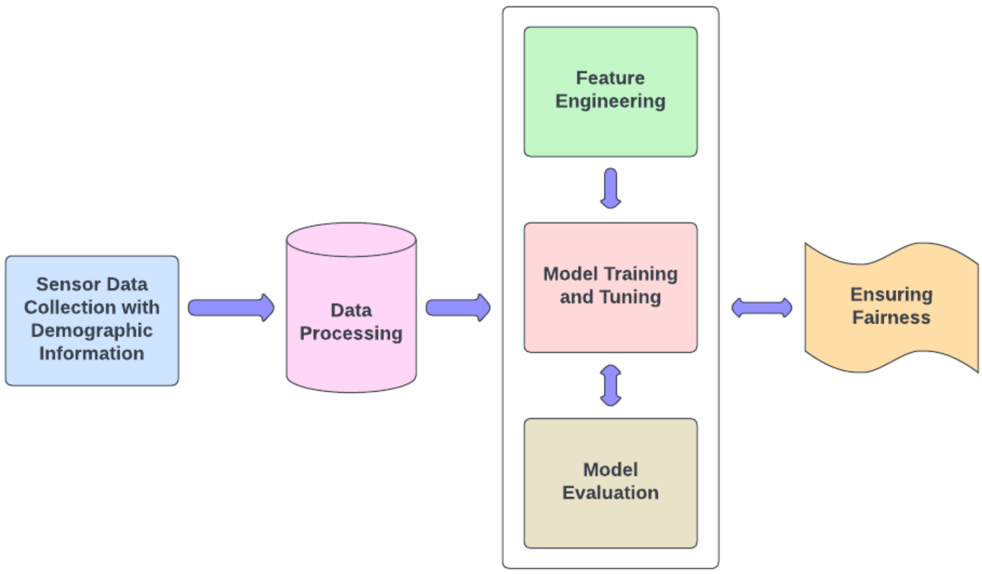

Figure 3.1: Proposed Model

3.1 Dataset and Features

The dataset is sourced from the reference [42], and it represents a valuable collection of human gait data with a specific focus on activity analysis and recognition. The dataset offers a diverse array of human activities, meticulously segmented and annotated for accurate interpretation. Among the activities included are walking, running, ascending and descending stairs, and sitting down, among others. The data collection process leverages a body sensor network, consisting of six wearable inertial sensors strategically placed on both the right and left thighs, shins, and feet. Additionally, two electromyography sensors are affixed to the quadriceps region of the front thigh to capture muscle activity [43]. The dataset encompasses a comprehensive feature set, as depicted in Table 3.1.

- `acc_rf_x`, `acc_rf_y`, `acc_rf_z`: These columns correspond to accelerometer readings along the x, y, and z axes, respectively.
- `gyro_rf_x`, `gyro_rf_y`, `gyro_rf_z`: These columns contain gyroscope readings along the x, y, and z axes, respectively.
- `act`: This column signifies the activity label associated with each recorded data instance. Activities are numerically encoded, such as values 8 and 2, which correspond to different types of activities.
- `filename`: This column provides the file path indicating the source of the data along with the participant ID.

To further enrich the dataset, gender information was incorporated by extracting relevant data from an independent file. This auxiliary file contained gender details for each participant ID, which were then linked to the corresponding IDs for subsequent fairness measurements. The dataset encompasses data from a total of 18 participants, encompassing 14 males and 4 females.

Table 3.1: Sample Data from the Dataset

acc_rf_x	acc_rf_y	acc_rf_z	gyro_rf_x	gyro_rf_y	gyro_rf_z	act	filename
-8268	-4716	12884	-10	21	-17	8	../HugaData/HuGaDB_v1_various_08_10.txt
-8284	-4728	12852	-14	22	-18	8	../HugaData/HuGaDB_v1_various_08_10.txt
-8276	-4716	12916	-8	17	-17	8	../HugaData/HuGaDB_v1_various_08_10.txt
-8228	-4764	12932	-8	20	-14	8	../HugaData/HuGaDB_v1_various_08_10.txt
-8304	-4796	13040	-7	21	-16	8	../HugaData/HuGaDB_v1_various_08_10.txt
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-25656	-3600	9588	-3666	-8927	-2948	2	../HugaData/HuGaDB_v1_running_08_02.txt
-28868	-7996	4976	-4428	-8637	-2969	2	../HugaData/HuGaDB_v1_running_08_02.txt
-28220	-9852	5400	-3544	-7292	-3608	2	../HugaData/HuGaDB_v1_running_08_02.txt
-32768	-6336	5588	-3365	-5796	-4499	2	../HugaData/HuGaDB_v1_running_08_02.txt
-32768	-952	4892	-3321	-4461	-5827	2	../HugaData/HuGaDB_v1_running_08_02.txt

This dataset's depth and diversity lay a strong foundation for investigating various aspects of human activity recognition, making it an excellent resource for experimentation and analysis in this domain.

3.2 Data Pre-Processing

Prior to embarking on data visualization, a series of crucial preprocessing steps are meticulously undertaken to ensure the integrity and reliability of the dataset. Activities are appropriately labeled as categorical variables, while other columns undergo standardization as necessary. Rigorous efforts are made to eliminate missing values, thereby mitigating the potential for data inconsistencies to creep into subsequent analyses. The initial dataset boasts a diverse array of activities, encompassing 'Bicycling,' 'Down by elevator,' 'Going down,' 'Going up,' 'Running,' 'Sitting,' 'Sitting down,' 'Sitting in car,' 'Standing,' 'Standing up,' 'Up by elevator,' and 'Walking.'
The initial step toward unveiling latent insights involves the exploration of data correlations, achieved through the application of a correlation matrix (Figure 3.12). This matrix elegantly encapsulates the correlation coefficient between pairs of features, with color gradients communicating both strength and direction of correlation. In our visualization, a prominent diagonal line is readily apparent, characterized by the deepest hue, and emblematic of a perfect positive correlation. This diagonal alignment signifies a consistent linear relationship between each feature and itself, resulting in a correlation coefficient of 1.0.
As we deviate from this diagonal, the matrix exhibits varying degrees of color intensity, signifying correlations among distinct feature pairs. The sporadic emergence of darker cells, juxtaposed with intermittent shifts in intensity, serves as a visual testament to the nuanced absence of a perfect correlation.

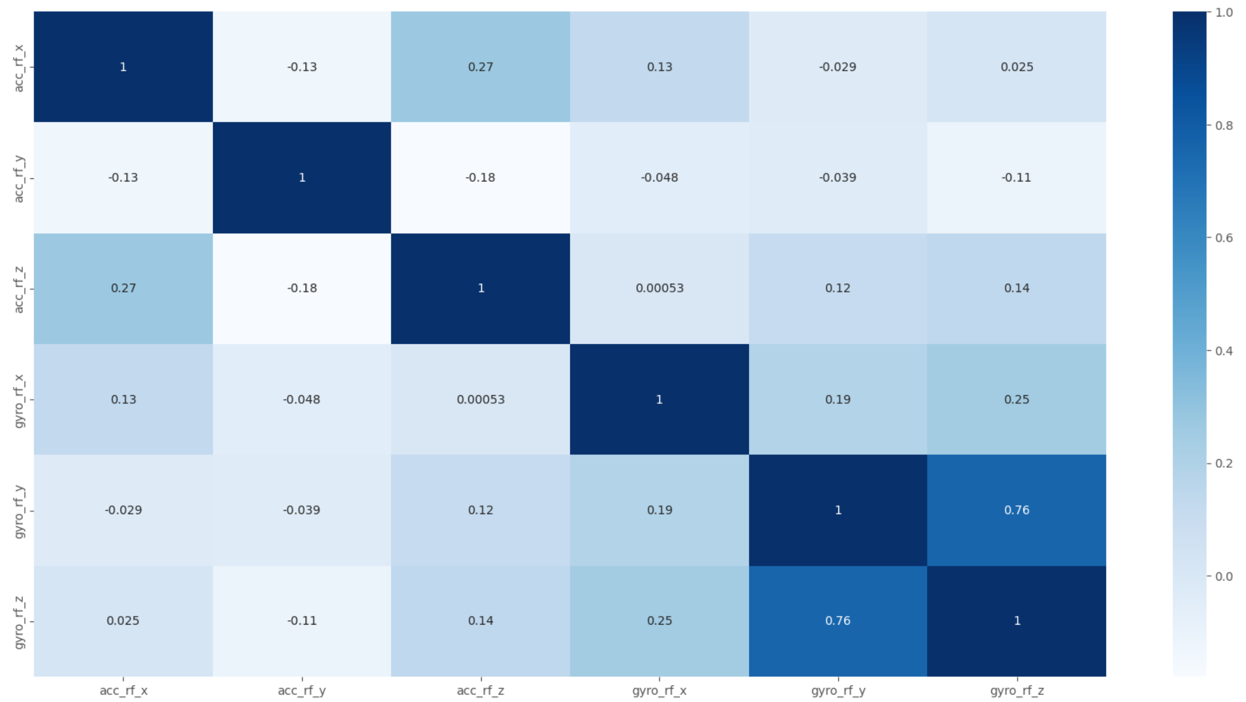

Figure 3.2: Data Correlation

After looking at correlations, we also checked how different activities are distributed in the dataset,the results of which are vividly illustrated in Figure 3.3. Notably, 'Walking' is the most common activity, making up a big 32.16% of the dataset. On the flip side, 'Sitting in a car' is the least common, only at 1.06%. 'Running' is the second most common activity. 'Sitting' and 'Standing' activities have proportions even higher than 'Walking' and 'Running.' These statistics highlight that some types of activities are not very common. This suggests the possibility of combining these less common activities to improve our ability to make accurate predictions.

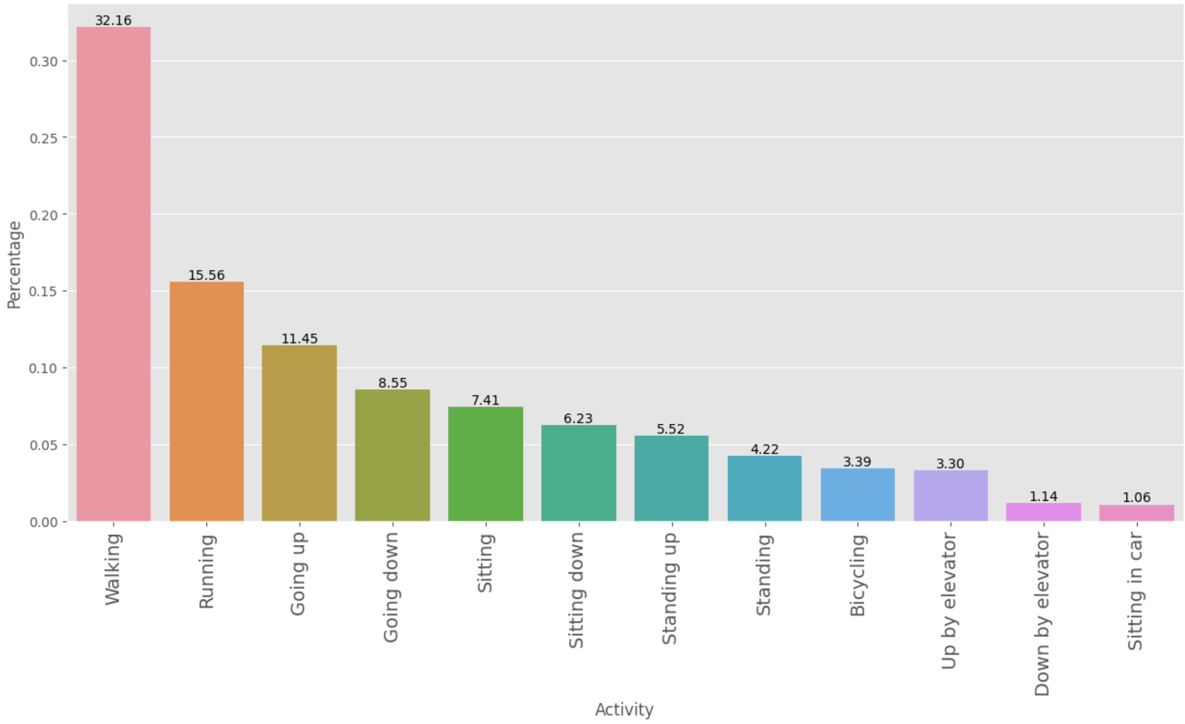

Figure 3.3: Activity Distribution

To ensure the reliability of the extracted features, we performed a comprehensive visual analysis. First, we carefully examined the consistency in acceleration data across all three axes, as shown in Figure 3.4. Then, we conducted a similar evaluation for gyroscope data along the same axes, illustrated in Figure 3.5. Surprisingly, both visual representations displayed a strong consistency, indicating that the measured attributes or activities are uniform across different orientations or directions. These sensor measurements reveal common characteristics regardless of the specific axis being observed.

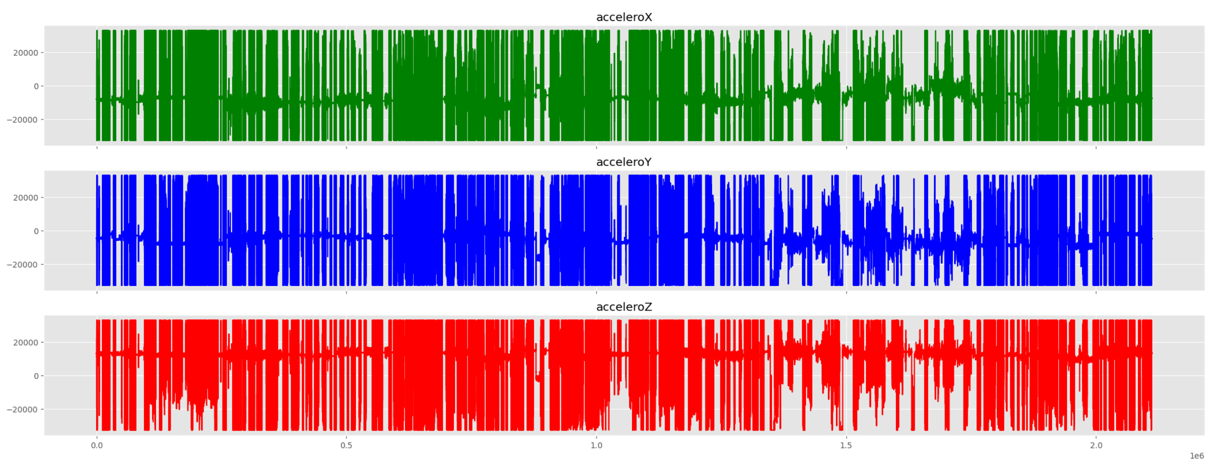

Figure 3.4: Similarity in Acceleration Data across Three Axes

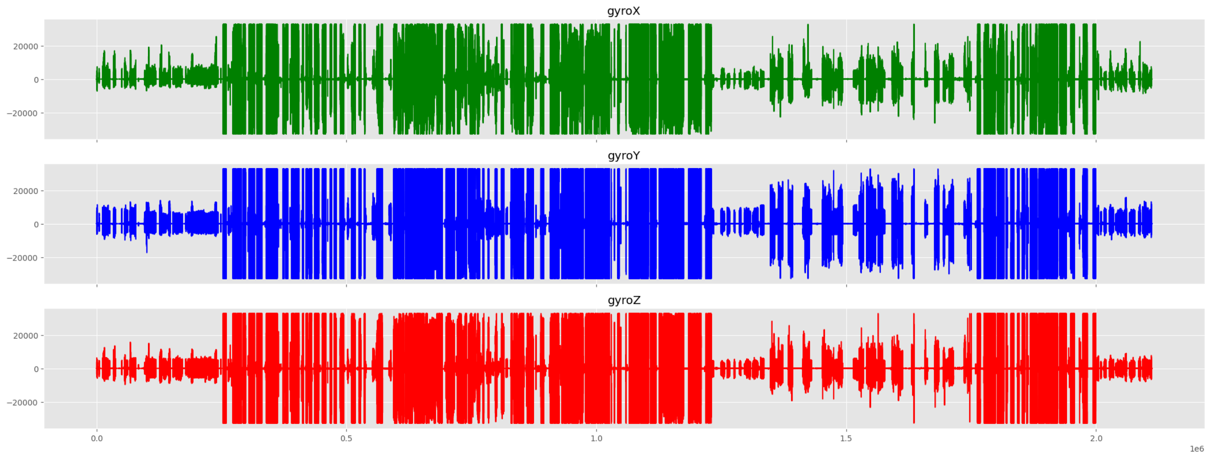

Figure 3.5: Similarity in Gyroscope Data across Three Axes

The insights drawn from our earlier analyses have now set the stage for a series of carefully planned steps to further refine, validate, and augment our dataset. This process is initiated by delving into the realm of statistical analysis, with the aim of extracting even more detailed insights that might have been lurking beneath the surface. A focus is placed on some key metrics, including normal acceleration, colatitude, and longitude, all of which are derived from the IMU data we have collected. In particular, valuable information about how forces act perpendicular to motion is revealed by normal acceleration, providing a window into the complex forces at play during different activities. On the other hand, an understanding of the tilt and rotation of objects in relation to a reference point is provided by colatitude and longitude, shedding light on the intricate spatial dynamics involved.
A range of statistical features are delved into, building on this foundation. These features encompass a variety of essential measures that help us understand the data better [44]. The boundaries are set by the minimum and maximum values, the sense of central tendency is provided by the average values, and the spread of data is indicated by the standard deviation. Insights into the distribution's shape are offered by measures like skewness and kurtosis, while details about data variability are revealed by interquartile range, median absolute deviation, and mean absolute deviation. Each of these measures contributes a piece of the puzzle, providing a deeper understanding of the data's characteristics and patterns. As this process is navigated through, new facets of the data continue to be uncovered, with each statistical measure acting like a spotlight, illuminating a different aspect of the activities being studied. Not only is our knowledge of the dataset enhanced by this exploration, but it also lays the groundwork for more sophisticated analyses down the line. It becomes a journey of discovery, where each metric and measure adds to our understanding and helps make sense of the complex world of human activities.
The dual subplots displayed in Figure 3.6 and Figure 3.7 serve the purpose of visualizing the distribution patterns of mean acceleration and gyroscope values associated with various physical activities. These activities are segregated into two categories: "Static Activities" and "Dynamic Activities," with each category presented in its own distinct subplot. In the "Static Activities" plot, mean acceleration values are analyzed for activities such as sitting, standing, and elevator-related movements, while the "Dynamic Activities" subplot illustrates mean acceleration values for more

physically active tasks like walking, running, and bicycling. These visualizations facilitate an assessment of how mean acceleration values vary across different activities. The subsequent part of the code performs a similar visualization process for mean gyroscope values, providing a comparison of their distribution among the same set of activities. These visualizations offer valuable insights into the distribution and variability of sensor readings during various physical activities, aiding in the comprehension of the dataset's underlying characteristics. Notably, it is observed that mean acceleration values are higher for walking and running compared to sitting and standing, a logical outcome given the greater physical activity involved in the former activities.

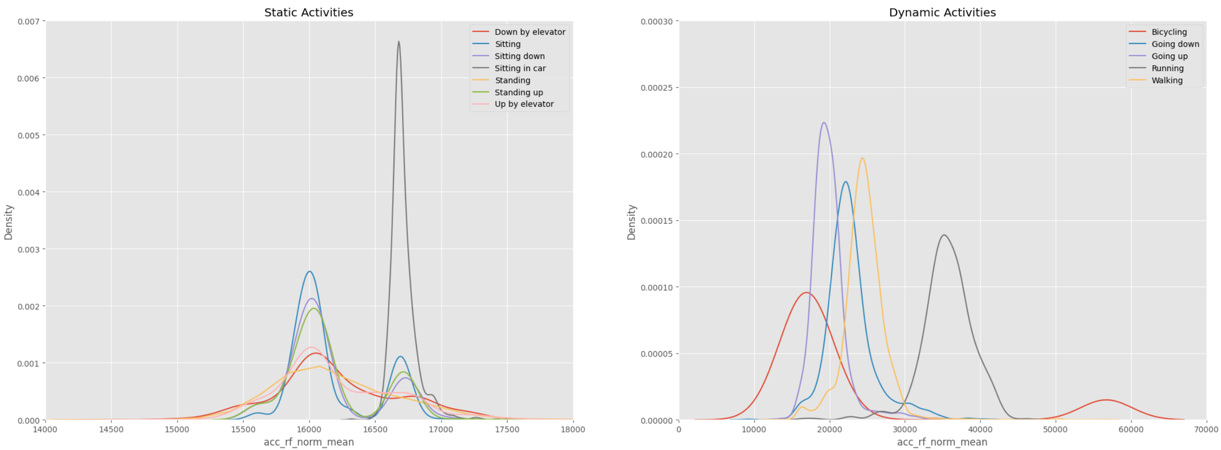

Figure 3.6: Distribution of Mean Acceleration Values for Static and Dynamic Activities

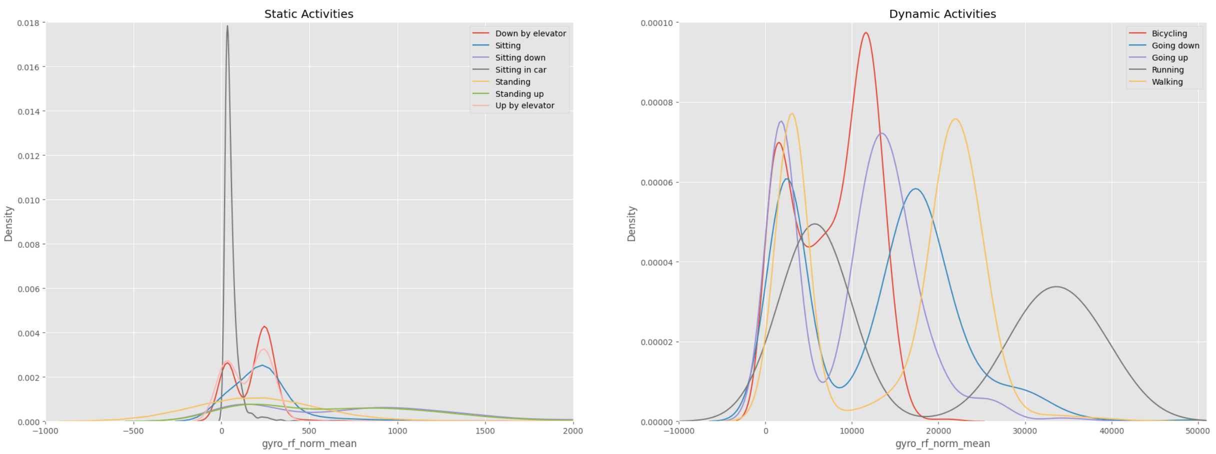

Figure 3.7: Distribution of Mean Gyroscope Values for Static and Dynamic Activities

The next phase involves the utilization of box plots to visually analyze the distribution patterns inherent in the features (Figure 3.8, Figure 3.9). These plots craft a visual narrative that unveils the distribution of data points, the extent of value ranges, and the existence of outliers.
It's worth noting that certain activities like Sitting down, Standing up and so on exhibit a lower count of outliers, indicating greater consistency in sensor readings.

This finding emphasizes that specific activities showcase more uniform and stable data, while others display increased variability or instances of unusual data points.

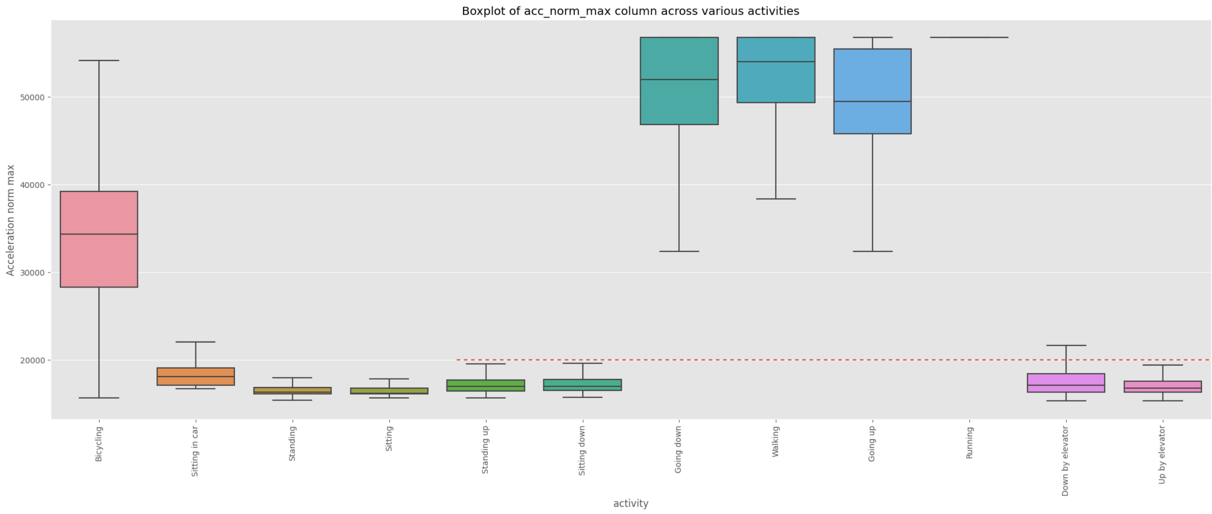

Figure 3.8: Boxplot of the normalized maximum acceleration (`acc_norm_max`) for all activities

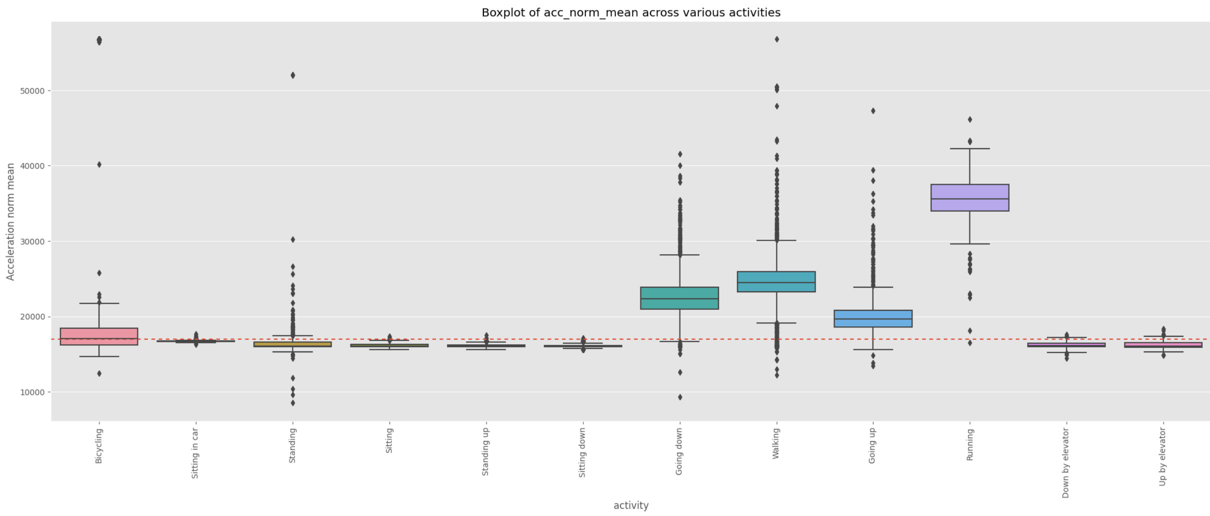

Figure 3.9: Boxplot of the mean normalized acceleration (`acc_norm_mean`) for all activities

These insightful box plots not only provide a window into the distribution of data points and the dispersion of values but also shed light on the presence of outliers that might hold valuable information. The discernible variation in the quantity of outliers across different activities triggers a strategic decision to consolidate activities that exhibit similar patterns. For instance, activities involving shifts in elevation—such as 'Going up,' 'Going down,' 'Down by elevator,' and 'Standing up'—are astutely grouped under the overarching category of 'Standing.' Similarly, activities that entail transitions between sitting and movement, vividly exemplified by 'Sitting in car' and 'Sitting down,' harmoniously find their home within the realm of 'Sitting.' We excluded bicycling data from our analysis due to the limited dataset, consisting of data from only one individual. This restricted dataset makes it impractical to assess the representativeness and fairness of the activity for broader conclusions. This pragmatic consolidation initiative serves as a pivotal maneuver to enhance the simplicity of subsequent analyses, thus expediting the discernment of profound trends and patterns that might otherwise be obscured in the intricate web of diversity. It's noteworthy that this process of consolidation bears the dual advantage of bolstering the statistical robustness of ensuing analyses. By creating broader categories that encompass a more substantial volume of data points, we forge a leveled playing field for the comparison of different activities. This, in turn, fortifies the statistical resilience of subsequent fairness analyses, enabling us to draw more confident conclusions about the treatment of various activities in the context of fairness.

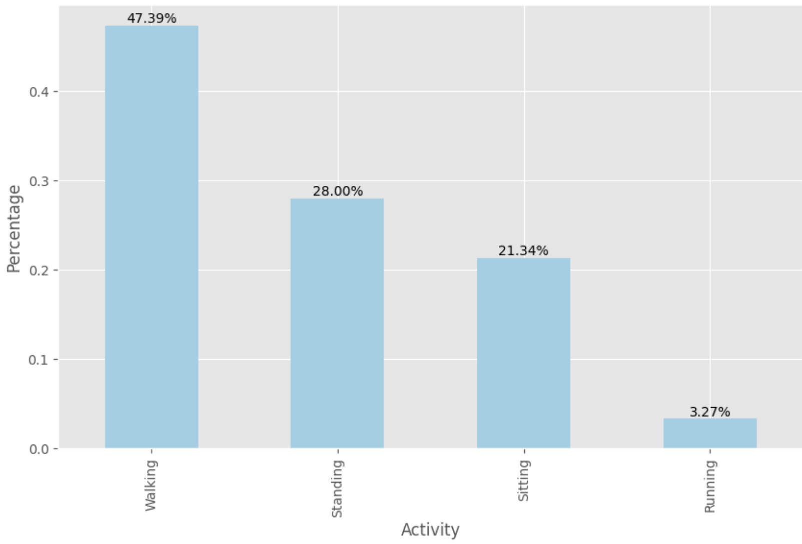

Figure 3.10: Activity Percentage

3.2.1 Data Distribution

To establish a robust and representative dataset for subsequent fairness analysis, we are following a meticulous approach. Initially, we implement a three-fold partition, randomly segregating data into train and test sets while guaranteeing the inclusion of one male and one female participant in each disjoint test set. Next, we divide the remaining data into training and validation sets in an 80:20 ratio. The validation set contains data from the same people as the training set, while the test set is completely new. This setup allows us to evaluate the model's performance on both known and unknown data. This distribution strategy is essential for fair analysis. By including data from familiar users in the validation set and new, unseen data in the training set, we can assess the model's effectiveness in various scenarios. This approach is crucial for making the model fair and reliable for both existing and new users.

3.3 Feature Engineering: Enhancing the Dataset

In the process of feature engineering, we have adopted a comprehensive approach to enhance the dataset's representational capabilities. This involves a series of steps aimed at creating new features that capture underlying patterns and relationships in the data.

3.3.1 Lag Features Creation

One of the initial steps involves integrating lagged features, a technique that enriches the temporal information captured in the DataFrame. This is achieved by shifting the values of existing columns to previous positions. Specifically, we create new columns named column_lag_1 and column_lag_2, where the original column values are displaced by one and two positions backward, respectively:

$$\text{column\_lag\_1} = \text{X[column].shift(1)} \tag{3.1}$$

$$\text{column\_lag\_2} = \text{X[column].shift(2)} \tag{3.2}$$

Incorporating lagged features proves especially advantageous in situations where historical data plays a pivotal role in comprehending and forecasting underlying patterns.

3.3.2 Handling Missing Data and Transformations

To ensure the dataset's completeness, we employ a strategy to handle missing values. Any missing values within the DataFrame are replaced with zeros, ensuring that no information is lost during subsequent analyses.
Furthermore, to account for skewed data distributions and to capture potential nonlinear relationships, we apply logarithmic and absolute value transformations. We create new columns named column_logabs for each original column. These new columns contain values obtained by taking the natural logarithm of the absolute values of the original data and then adding 1:

$$\text{column\_logabs} = \ln(|\text{X[column]}| + 1) \tag{3.3}$$

These transformed features can help to balance the influence of extreme values and enhance the dataset's suitability for modeling.
The culmination of these procedures has resulted in a revised DataFrame of dimensions $(16038, 792)$, contrasting the initial shape of $(16038, 132)$. More details about a segment of this modified dataset can be found in Appendix 5.

3.3.3 t-SNE Dimensionality Reduction

Then we use t-SNE, or t-Distributed Stochastic Neighbor Embedding which is a powerful dimensionality reduction algorithm frequently employed in data analysis and visualization tasks. According to the experiment of Figure 14, during the initial 250 iterations, the KL divergence value of 73.369057 reflects the divergence between the high-dimensional probability distribution P and the initially approximated low-dimensional distribution Q. This early divergence value is relatively

high due to the exaggeration factor applied during these initial iterations. As the algorithm progresses over 1000 iterations, the KL divergence drops to 1.153913, indicating a substantial reduction in the mismatch between the two distributions. This demonstrates the success of t-SNE in mapping the high-dimensional data into a lower-dimensional space where pairwise similarities are more faithfully preserved.
t-SNE is utilized to transform high-dimensional data into a lower-dimensional representation, often two or three dimensions, while preserving the underlying structure and relationships within the data [45]. The t-SNE objective function is defined as follows [46]:

$$p_{ij} = \frac{e^{\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2\right)}}{\sum_{k \neq i} e^{\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2\right)}} \tag{3.4}$$

p_{ij} represents the conditional probability that point $\mathbf{x}_i$ would select point $\mathbf{x}_j$ as its neighbor based on their similarity. The numerator calculates the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ using the Gaussian kernel function with variance σ_i^2. The denominator normalizes the sum of conditional probabilities over all other points.

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}} \tag{3.5}$$

here, q_{ij} is the conditional probability that point $\mathbf{y}_i$ would select point $\mathbf{y}_j$ as its neighbor in the lower-dimensional space.

$$\mathrm{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.6}$$

The equation above represents the Kullback-Leibler (KL) divergence $\mathrm{KL}(P||Q)$ between the distributions P and Q in the high-dimensional space. Minimizing this divergence in the lower-dimensional space helps preserve relationships between data points during dimensionality reduction. The equation sums over all pairs of data points, calculating the divergence between p_{ij} and q_{ij}.
The t-SNE results, as depicted in Figure 3.11, illustrate the transformation of 16,038 data samples into a reduced-dimensional space. In this context, the x-axis and y-axis of the scatter plot correspond to the two dimensions of the lower-dimensional space. Each point on the plot symbolizes an original data point, with its placement determined by its coordinates in this transformed space. A discernible observation emerges: the data points associated with various activities, except for sitting, exhibit distinct clustering. This distinctness suggests that the dataset's activities are well-separated, with minimal overlap or mixing, underscoring the effective distribution of data across most activities.

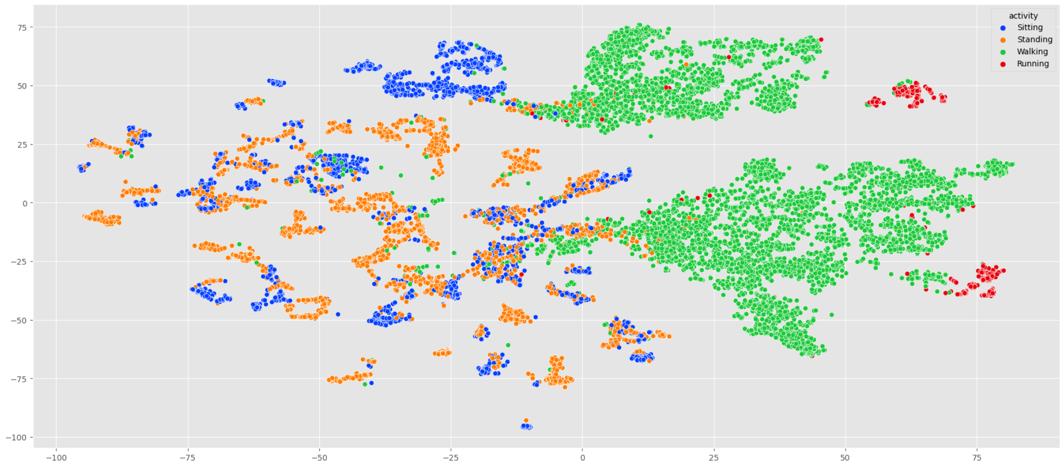

Figure 3.11: t-SNE Dimension Reduction Plot

3.4 Prediction Models Specification

3.4.1 Logistic Regression

Logistic Regression is a statistical method used to predict the likelihood of an event with two possible outcomes based on one or more input variables [47]. When applied to forecasting activities using sensor data, it proves to be a valuable tool. In this context, let's symbolize the activity as Y, where $Y = 1$ signifies that the activity is happening, and $Y = 0$ means the activity is not occurring. We also have input variables, represented as $X_1, X_2, \ldots, X_p$, which can stand for various features extracted from the sensor data.
The logistic regression model takes these input variables, combines them linearly, and then uses the logistic function to estimate the probability of the activity taking place:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}} \tag{3.7}$$

Here's what these symbols mean:
- $P(Y = 1)$ represents the predicted probability of the activity occurring. - e is the base of the natural logarithm (≈ 2.71828). - $\beta_0, \beta_1, \ldots, \beta_p$ are coefficients associated with the input variables $X_0, X_1, \ldots, X_p$.
The logistic function $\frac{1}{1+e^{-z}}$ maps any real number z to a range between 0 and 1, making it suitable for predicting probabilities.
In the context of activity prediction, we estimate the values of coefficients $\beta_0, \beta_1, \ldots, \beta_p$ from training data using techniques such as maximum likelihood estimation. Once these coefficients are determined, the model can be used to predict the probability of an activity based on new sensor data [48]:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}} \tag{3.8}$$

If the calculated $P(Y = 1)$ is greater than a specific threshold (e.g., 0.5), we predict that the activity is occurring; otherwise, we predict that it is not happening.

Now, let's discuss the Decision Tree algorithm, which constructs a tree structure by selecting the most informative features at each branching point, called decision nodes [49]. This algorithm aims to minimize data impurity or maximize information gain with each split, resulting in optimal data separation.
The decision process in a Decision Tree can be represented as follows:

$$\text{Predicted Activity} = \begin{cases} \text{Activity A} & \text{if } X_i < T_i \\ \text{Activity B} & \text{if } X_i \geq T_i \end{cases} \tag{3.9}$$

Here's what these terms mean:
- X_i denotes the value of a specific input variable. - T_i represents the threshold value for the split on X_i.
This decision-making process continues recursively until a leaf node is reached.
Constructing a Decision Tree involves finding the best splits and threshold values that maximize the separation of activities. This is typically achieved using criteria like Gini impurity or information gain [50].
Once the Decision Tree is built, it can be utilized to predict activity for new sensor data by following the decision path down the tree until a leaf node is reached. The predicted activity associated with that leaf node is considered the final prediction.

3.4.2 Multi-Layer Perceptron

The Multilayer Perceptron (MLP) is a fundamental artificial neural network architecture used for solving various machine learning tasks, including activity prediction based on sensor data. MLPs are particularly effective at capturing complex patterns and relationships in data [51].
An MLP consists of multiple layers of interconnected nodes, where each node (neuron) performs a weighted sum of its input values and applies an activation function to produce an output. The architecture typically includes an input layer, one or more hidden layers, and an output layer. The computations in an MLP can be described as follows:
1. Input Layer: The input layer receives the sensor data as input features.
2. Hidden Layers: Each neuron in the hidden layers computes a weighted sum of its inputs and applies an activation function. The output of each neuron becomes the input for the neurons in the next layer.
3. Output Layer: The output layer produces the final prediction for the activity. The output is often passed through an activation function suitable for the task, such as the sigmoid function for binary classification.
Mathematically, the computation in an MLP can be represented as follows [52]:
For a neuron j in hidden layer h:

$$a_j^{(h)} = \sum_{i=1}^{p} w_{ij}^{(h)} \cdot X_i + b_j^{(h)} \tag{3.10}$$

$$z_j^{(h)} = \sigma(a_j^{(h)}) \tag{3.11}$$

Where:

- $a_j^{(h)}$ is the weighted sum of inputs for neuron j in hidden layer h.

- $w_{ij}^{(h)}$ is the weight connecting input X_i to neuron j in hidden layer h.
- $b_j^{(h)}$ is the bias for neuron j in hidden layer h.
- σ is the activation function.
- $z_j^{(h)}$ is the output of neuron j in hidden layer h.

The output of the MLP is obtained by propagating the inputs through each layer until the final output is computed. The predicted activity is based on the output of the output layer.

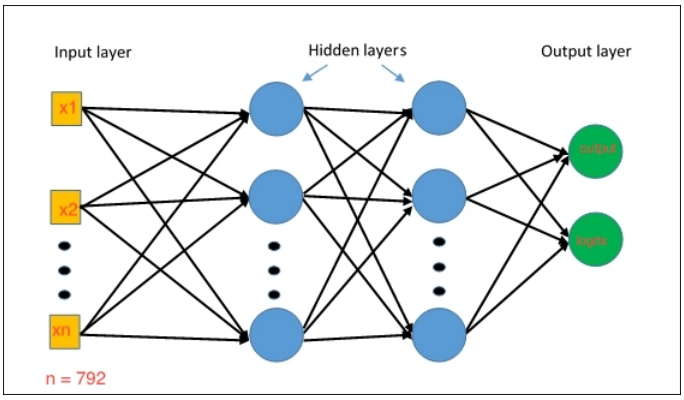

Figure 3.12: Proposed MLP Model

We initiate our research with logistic regression and decision trees to predict activities, establishing a state of art performance. Subsequently, we have developed a neural network model known as the Multi-Layer Perceptron (MLP) designed for classification tasks. The model encompasses a series of transformative stages applied to the input data. This includes linear transformations, dropout for regularization, ReLU activation to introduce non-linearity, and Instance Normalization to stabilize the training process. These transformations work in tandem to reduce the data's dimensionality while capturing crucial features. Subsequently, the processed data is being forwarded through a second pass, further compressing it and generating probabilities for various classes. The model's use of the sigmoid activation function is mapping these outputs to a range between 0 and 1, indicating the likelihood of each class. This architecture empowers the model to discern intricate relationships within the input data, facilitating precise predictions. For the loss function, we are utilizing the Binary Cross-Entropy with Logits (BCEWithLogits) loss, and for optimization, we are using the Stochastic Gradient Descent (SGD) optimizer.
The algorithm 1 outlines the training procedure. It starts by defining a list of activities and folds, with each fold representing a distinct partition of the dataset. For each fold and activity, the train and test data are loaded, and the target variable is converted to binary. The data is split into train, validation, and test sets, and PyTorch datasets and data loaders are created. During training, the model undergoes multiple epochs, where it is trained and validated on the corresponding data subsets. Model performance is evaluated on both the validation and test sets, and learning rate is adjusted using a scheduler. The algorithm employs early stopping to prevent

overfitting and ultimately combines performance metrics for different activities to create a summary.

Algorithm 1 Fairness-Unaware MLP Training

```
activities = ['Standing', 'Walking', 'Sitting', 'Bicycling', 'Running']
define list of folds: folds = ["1", "2", "3"]
for each fold in folds do
    initialize empty lists for training and validation metrics
    for each activity in activities do
        load train and test data for the current fold and activity
        convert target variable to binary
        split into train, validate, and test sets
        create PyTorch datasets and data loaders
        for each epoch in num epochs do
            train model:
            initialize losses and performance metrics
            for each batch in train loader do
                forward pass through the MLP model
                calculate total loss
                backpropagate and update model parameters
            end for
            validate model:
            evaluate on validation set
            track best validation performance and save model if improved
            test model:
            evaluate on test set
            update learning rate using scheduler
            check for early stopping condition based on validation performance
        end for
        save model and performance metrics
    end for
    combine performance metrics for different activities and create summary
end for
```

3.5 Fairness Gap

To calculate the fairness gap, we employ both the Demographic Parity (DP) gap and Equalized Odds (EO) gap techniques. The DP gap is a measure of fairness that evaluates whether a model's predictions are independent of sensitive attributes, such as gender. The DP gap is calculated by comparing the proportion of positive predictions for each gender group [53] [54].

Demographic Parity Gap Calculation

The calculation of the Demographic Parity (DP) gap involves the following steps:

1. **Loading Pre-trained Model:** The function takes inputs such as a data loader, path to a pre-trained model, and the MLP model itself. The pre-trained model is loaded from the specified path, and its mode is set to evaluation.

2. **Predictions and Counting:** The function predicts outcomes for the input data using the model. It tracks the count of positive predictions for each gender group. This is done by thresholding the model's predictions at 0.5 to determine whether a prediction is positive or negative. The count of positive predictions is maintained separately for **each gender group** [55].

3. **Calculating DP Gap:** After processing the data, the DP gap is calculated using the formula:

$$DP\,Gap = \frac{|E(Predictions_{Group\,0}) - E(Predictions_{Group\,1})|}{Total\,Dataset\,Size} \tag{3.12}$$

The DP gap provides insights into potential biases in the model's predictions based on gender, aiming to ensure fairness and equal treatment across different gender groups.
The Equalized Odds (EO) gap is another fairness metric that measures the disparity in model predictions between different gender groups while considering both false positive rates and false negative rates [56].

Equalized Odds Gap Calculation

- **Pre-trained Model:** Load the pre-trained model and set it to evaluation mode.

- **Data Processing:** For each batch in the data loader:

  - Generate predictions and logits from input data using the model.
  - Capture true labels and predicted labels.
  - Record gender information.

- **Prediction Analysis:** Count positive samples and predicted positives for **each class and gender.**

- **Calculating EO Gap:** Calculate the EO gap using:

$$EO\,Gap = \frac{|E(Predictions_{Group\,0}) - E(Predictions_{Group\,1})|}{\text{Total Positive Samples for each class}} \tag{3.13}$$

Here, "Group 0" and "Group 1" refer to the two gender groups. In both methods, we utilize the entire dataset size as the batch size. This choice is made to ensure that positive samples are computed across the entire dataset rather than within subsections.

3.6 Introducing Fair Loss

To assess the fairness of machine learning models in the context of human activity recognition, two key fairness metrics—Equalized Odds and Demographic Parity—are commonly employed.These metrics are fundamental for evaluating the model's behavior across different demographic groups and ensuring equitable treatment. They are quantified using specific loss functions that incorporate mathematical terms to express the extent of fairness achieved by a model. These losses help reduce gaps that indicate unfairness. Smaller gaps mean more fairness. We trained our model twice, once with DP loss and once with EO loss, using the "fairtorch" library in PyTorch.

3.6.1 Equalized Odds Loss

Equalized Odds focuses on the notion of equal treatment in terms of prediction accuracy across different groups. It ensures that the true positive rates and false positive rates are consistent among different subgroups within the dataset. Mathematically, for each demographic group d and each class y, the Equalized Odds Loss (L_{EO}) can be formulated as [57]:

$$L_{EO}(d, y) = \max\left(\left|\text{FPR}_{d,y} - \text{FPR}_{\text{reference},y}\right|, \left|\text{TPR}_{d,y} - \text{TPR}_{\text{reference},y}\right|\right)$$

Where: - $\text{FPR}_{d,y}$ and $\text{TPR}_{d,y}$ are the false positive rate and true positive rate for group d and class y respectively. - $\text{FPR}_{\text{reference},y}$ and $\text{TPR}_{\text{reference},y}$ are the reference false positive rate and true positive rate for class y that the model should strive to achieve.
The Equalized Odds Loss calculates the maximum deviation of false positive rates and true positive rates from the reference rates, thus indicating how well the model maintains fairness in its predictions across demographic groups.

3.6.2 Demographic Parity Loss

Demographic Parity ensures that the proportion of positive predictions remains consistent across different demographic groups. It aims to prevent the model from disproportionately favoring or penalizing any particular group. The Demographic Parity Loss (L_{DP}) can be expressed mathematically as [58]:

$$L_{DP}(d) = |\Pr(y = 1|d) - \Pr(y = 1|\text{reference})|$$

Where: - $\Pr(y = 1|d)$ is the probability of a positive prediction for group d. - $\Pr(y = 1|\text{reference})$ is the reference probability of a positive prediction, indicating the desired fairness threshold.
The Demographic Parity Loss quantifies the difference between the predicted positive probabilities for a particular group and the reference group, thereby capturing the model's adherence to fairness principles.
Both these loss functions provide interpretable metrics to evaluate the fairness of machine learning models in human activity recognition tasks. By minimizing these losses, models can be optimized to make predictions that are not disproportionately biased towards any specific demographic group.

3.6.3 Lagrangian Method

The fundamental idea behind this method is to transform a constrained optimization problem into an unconstrained one by introducing Lagrange multipliers, which act as coefficients that balance the trade-off between optimizing the objective and satisfying the constraints [32].
Consider an optimization problem where we want to minimize a function $f(x)$ subject to some constraints given by $h_i(x) = 0$ for $i = 1, 2, \ldots, m$. The Lagrangian for this problem is defined as follows:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) \tag{3.14}$$

Here, x is the vector of optimization variables, λ is the vector of Lagrange multipliers, and $h_i(x)$ are the constraint functions. The Lagrangian combines the original objective function $f(x)$ with the constraints, weighted by the Lagrange multipliers. The Lagrange multipliers act as scalar coefficients that represent the sensitivity of the objective to changes in the constraints.
In the initial stages, although both the DP and EO gaps exhibited a degree of reduction, we encountered difficulties related to the convergence of the fairness losses. This impeded substantial gap reduction. To circumvent this, we introduced a Lagrangian multiplier. The fair training process, illustrated in Algorithm 2, encapsulates this adaptive approach. The algorithm builds upon a similar class loss function used previously, integrating either the EO or DP loss as an independent fair loss metric. To compute the overall loss, we initially used a combination of the class loss and the fairness loss, weighted by a value called α:

$$(1 - \alpha) \times \text{fairness loss} + \alpha \times \text{class loss}. \tag{3.15}$$

Acknowledging that a static α value could hinder the convergence of the fair loss, we introduced a novel variable, λ. The new combined loss became:

$$\text{class loss} + \lambda \times \text{fairness loss}. \tag{3.16}$$

We have started with λ as 1 and hypertuned the α value, which balanced fairness and accuracy better. As we went through each batch of training data, we updated λ using this equation:

$$\lambda = \lambda + \alpha \times \text{fairness loss}. \tag{3.17}$$

Our experiments show that combining the Lagrangian multiplier with the EO loss produced better results compared to all other methods, as we'll explain in the results section.

Algorithm 2 Fairness-Aware MLP Training

```
activities = ['Standing', 'Walking', 'Sitting', 'Bicycling', 'Running']
define list of folds: folds = ["1", "2", "3"]
for each fold in folds do
    initialize empty lists for training and validation metrics
    for each activity in activities do
        load train and test data for the current fold and activity
        convert target variable to binary
        split into train, validate, and test sets
        create PyTorch datasets and data loaders
        for each epoch in num_epochs do
            train model:
                initialize losses and performance metrics
            for each batch in train_loader do
                forward pass through the MLP model
                calculate class loss (BCEwithLogit) and fair loss (EO or DP)
                calculate: total_loss = class_loss + λ × fair_loss
                backpropagate and update model parameters
                update: λ = λ + a × fair_loss
            end for
            validate model:
                evaluate on validation set
                track best validation performance and save model if improved
            test model:
                evaluate on test set
            update learning rate using scheduler
            check for early stopping condition based on validation performance
        end for
        save model and performance metrics
    end for
    combine performance metrics for different activities and create summary
end for
```

Chapter 4

Result Analysis

4.1 Prediction Model Performance

In this study, we introduced three distinct machine learning algorithms: Logistic Regression (LR) and Decision Tree (DT) served as state of art models, while the Multilayer Perceptron (MLP) was positioned as the primary model. To address fairness considerations, a deep learning approach was necessary, prompting the utilization of the MLP. The Human Activity Recognition (HAR) dataset was collected and processed to perform classification tasks. Our experimental setup involves partitioning the data into k-folds for training and testing. Each fold maintains a balanced distribution of classes, with one male and one female participant in the test set and the remainder forming the 80% training and 20% validation split. The classification algorithm is trained on each subset in sequence, and the error is calculated on both the validation and test sets. This process is repeated k times, and the k error estimates are averaged to yield an overall error estimate.
In Table 4.1, we showcase the state of art accuracy outcomes for both Logistic Regression (LR) and Decision Tree (DT) models. The "Valid" accuracy represents the performance on the validation set, which comprises a portion of the training data. On the other hand, the "Test" accuracy showcases the model's effectiveness on an entirely separate test set, comprising data from disjoint individuals who were not part of the training process. Comparing with the MLP accuracy results (Table 4.2), we observe that the MLP model generally outperforms both LR and DT in terms of accuracy. For instance, in the "Running" activity, the MLP achieves a "Valid Accuracy" of 0.9925 and a "Test Accuracy" of 0.9944, which are notably higher than the LR and DT "Test" accuracies of 0.9914 and 0.9926, respectively.The variation between "Valid" and "Test" accuracy can be due to the generalization capabilities of the model on unseen data, which may differ in distribution from the training data. The "Sitting" and "Standing" activities exhibit lower performance due to the limited amount of available data, resulting in LR, DT and MLP models achieving lower accuracy scores for both "Valid" and "Test" data.
From the confusion matrix of Figure 4.1, it is evident that the model exhibits exceptional performance in accurately identifying instances of the "Running" and "Walking" activities in both the valid and test datasets. These activities are characterized by minimal false positives and negatives, reflecting the model's strong discriminatory abilities. In the "Running" activity, for instance, the model achieves a high true positive count of 142 against only 14 false negatives in the test dataset. Similarly,

Table 4.1: State of Art Accuracy of Logistic Regression (LR) and Decision Tree (DT)

Activities	LR Valid	LR Test	DT Valid	DT Test
Running	0.9548	0.9914	0.9469	0.9926
Sitting	0.8006	0.8226	0.7973	0.8112
Standing	0.7818	0.8073	0.7901	0.8013
Walking	0.9754	0.9751	0.9776	0.9775

Table 4.2: Multi-Layer Perceptron (MLP) Accuracy of Activities

Activities	Valid Accuracy	Test Accuracy
Running	0.9925	0.9944
Sitting	0.8243	0.7850
Standing	0.8350	0.8270
Walking	0.9808	0.9750

in the "Walking" activity, the model demonstrates proficiency with a true positive count of 2223 against 44 false negatives in the test dataset.
The model also demonstrates proficiency in the "Standing" activity, though there are occasional instances of misclassification of non-"Standing" instances. In the valid dataset, the "Standing" activity is represented by a confusion matrix with 3667 true positives, 1037 false positives, and 43 false negatives. However, despite these occasional misclassifications, the model still maintains an overall strong performance.
However, the "Sitting" activity presents a challenge for the model, as it tends to misclassify non-"Sitting" instances and occasionally misses instances of the activity. In the test dataset, the confusion matrix for "Sitting" shows 2854 true positives, 960 false positives, and 27 false negatives. Similarly, in the valid dataset, the confusion matrix records 4018 true positives, 1107 false positives, and 53 false negatives for "Sitting."
Analyzing the confusion matrix values further:

- **Running Activity:** The model's performance is particularly notable, achieving high true positive counts with minimal false positives and negatives in both the valid and test datasets.
- **Standing Activity:** While the model generally performs well in both datasets, there are instances of non-"Standing" instances being misclassified as "Standing," leading to a notable number of false positives.
- **Walking Activity:** The model's proficiency is evident in the test dataset, but there are false negatives indicating instances where "Walking" activities were not recognized. Similar trends are observed in the valid dataset.
- **Sitting Activity:** The model encounters challenges with the "Sitting" activity, as it tends to misclassify non-"Sitting" instances, resulting in a relatively higher number of false positives. Moreover, instances of the "Sitting" activity are occasionally missed.

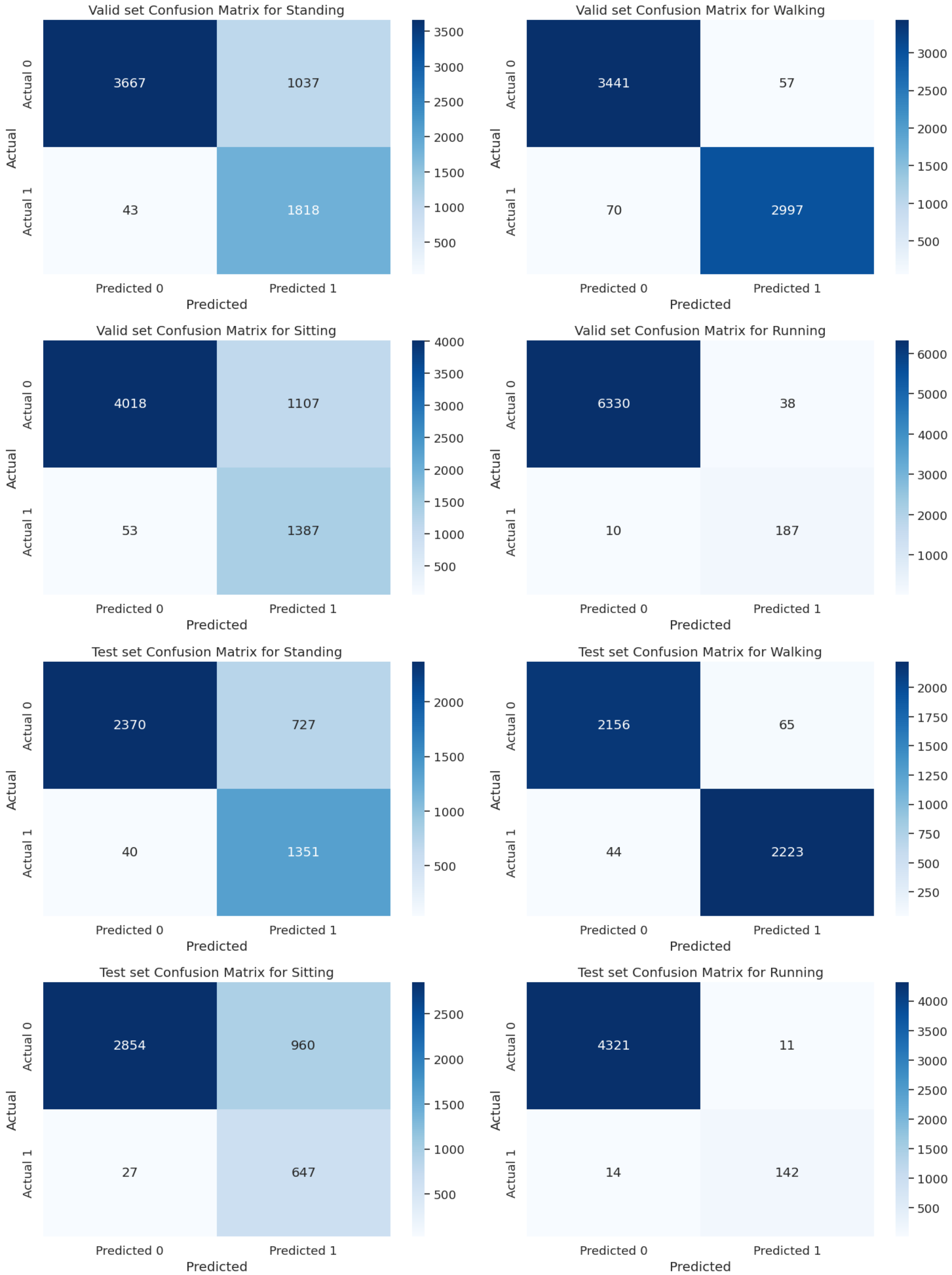

Figure 4.1: Confusion Matrix of MLP Model for Valid and Test Dataset

In Table 4.3, the DP gap values for MLP model are indicative of the disparity in positive predictions between different gender groups for each activity. In the Running activity, the DP gap is quite low at 0.0029 at test and 0.0192 at validate set, suggesting that the model's predictions are fairly balanced between gender groups for that activity. For the Standing activity, the DP gap is 0.0776 and 0.2802, further highlighting the model's efforts to maintain fairness in its predictions. However, in the Walking activity, the DP gap is relatively higher at 0.1094 and 0.2930. This could be due to the specific characteristics of the dataset, potentially including a higher gender imbalance.
Conversely, the EO gap values represent the discrepancies in prediction outcomes between gender groups while considering false positive and false negative rates. Interestingly, the EO gap values are higher than the corresponding DP gap values in most activities. This is likely because the EO gap takes into account both false positives and false negatives, whereas the DP gap focuses solely on the proportion of positive predictions. A higher EO gap could suggest variations in the model's performance across gender groups in terms of false positive and false negative rates. The larger EO gap, compared to the DP gap, indicates that the model might exhibit slight differences in how it handles false positives and false negatives based on gender, while still maintaining overall fairness.
The differences in fairness gaps between "Valid" and "Test" data can be attributed to the dataset composition. The "Test" data is balanced in terms of gender, contributing to the lower fairness gaps. The higher fairness gaps in the "Valid" set might be linked to the fact that the training data is skewed toward a particular gender.

Table 4.3: Fairness Gaps for Different Activities

Activities	Valid EO Gap	Test EO Gap	Valid DP Gap	Test DP Gap
Running	0.4968	0.1399	0.0192	0.0029
Sitting	0.7503	0.2317	0.2809	0.0645
Standing	0.5937	0.1787	0.2802	0.0776
Walking	0.6129	0.2160	0.2930	0.1094

4.2 Fairness Evaluation

In an attempt to enhance fairness, we introduced EO and DP Loss techniques. Applying these techniques aimed to minimize fairness gaps and ensure equitable performance across activities. However, the results were mixed. While EO and DP Loss techniques successfully reduced fairness gaps in most activities, there were instances where the gaps increased. These techniques effectively reduced fairness gaps, aligning with our aim of achieving fairness in predictions. Notably, for the valid set in Table 4.4, both EO and DP Loss techniques managed to decrease the fairness gaps across most activities. Nevertheless, there were certain cases where the gaps increased, posing a challenge to achieving fairness. Particularly, the "Running" activity demonstrated significant difficulties in achieving fairness. In the valid set, applying EO Loss led to an increase in the fairness gap from 0.4968 (in the unfair model) to 0.5986, and a similar pattern was observed with DP Loss, where the fairness gap increased to 0.5782. This underscores the complexity of addressing fairness for certain activities and highlights the limitations of these techniques in specific contexts. For the test set in Table 4.5, fairness gaps were initially lower, indicating a relatively better state of art fairness. Nevertheless, the application of EO and DP Loss techniques yielded slight improvements in fairness gaps for most activities. While the improvements were not as dramatic as in the valid set, the consistency of these enhancements signifies the potential benefits of fairness-aware training.

Table 4.4: Fair MLP Performance Comparison using EO Loss and DP Loss for Validation Data

Activities	DP Loss				EO Loss			
	F1	Acc	EO Gap	DP Gap	F1	Acc	EO Gap	DP Gap
Running	0.9851	0.9835	0.5986	0.0086	0.9852	0.9838	0.5782	0.0077
Sitting	0.7723	0.7491	0.7525	0.1043	0.7809	0.7590	0.7560	0.0950
Standing	0.8171	0.8077	0.5889	0.1033	0.8210	0.8119	0.6519	0.0922
Walking	0.9558	0.9558	0.6094	0.0989	0.9679	0.9679	0.5814	0.0932

Table 4.5: Fair MLP Performance Comparison using EO Loss and DP Loss for Test Data

Activities	DP Loss				EO Loss			
	F1	Acc	EO Gap	DP Gap	F1	Acc	EO Gap	DP Gap
Running	0.9901	0.9897	0.1016	0.0036	0.9915	0.9913	0.1099	0.0032
Sitting	0.7735	0.7364	0.2319	0.0790	0.7794	0.7432	0.2314	0.0492
Standing	0.8296	0.8229	0.1741	0.0772	0.8306	0.8239	0.1717	0.0763
Walking	0.9487	0.9488	0.2170	0.1070	0.9645	0.9645	0.2146	0.1101

The inconsistency in the results might have been due to the constant alpha value used in the fair loss. This fixed value might not have adapted well to the fairness requirements, leading to varying outcomes. To address this, we introduced a lagrangian multiplier technique. Instead of using a constant coefficient for the fair loss, we incorporated two variables, alpha and lambda. Lambda started with a value of 1 and was updated for each class based on alpha multiplied by the fair loss. This adaptive approach ensured that lambda adjusted itself based on the fair loss of the previous class, thus enhancing the fairness of predictions. This improvement is evident in Tables 4.6 and 4.7. Both the valid and test data show enhanced fairness compared to the unfair model in Table 4.3. These results are mean value of three folds we have trained on. The detailed results can be found in Appendix 5.

To substantiate our proposed solution, we integrated an additional dataset featuring accelerometer and gyroscope values from a group of 10 individuals, consisting of 4 females and 6 males having the same type of activities.

After doing all the data inconsistency filtering and feature engineering, Figure 4.2 illustrates the statistical attributes of a randomly generated dataset, employed for

Table 4.6: Fair MLP with Lagrangian Multiplier Performance Comparison for Valid Data

Activities	DP Loss + Lagrangian				EO Loss + Lagrangian			
	F1	Acc	EO Gap	DP Gap	F1	Acc	EO Gap	DP Gap
Running	0.9853	0.9848	0.3710	0.0157	0.9916	0.9915	0.3914	0.0146
Sitting	0.7785	0.7570	0.5638	0.1735	0.7998	0.7817	0.5748	0.0624
Standing	0.8221	0.8134	0.4404	0.2182	0.8265	0.8178	0.4581	0.2210
Walking	0.9619	0.9620	0.4417	0.2118	0.9675	0.9675	0.4681	0.2241

Table 4.7: Fair MLP with Lagrangian Multiplier Performance Comparison for Test Data

Activities	DP Loss + Lagrangian				EO Loss + Lagrangian			
	F1	Acc	EO Gap	DP Gap	F1	Acc	EO Gap	DP Gap
Running	0.9908	0.9911	0.1239	0.0033	0.9878	0.9878	0.0785	0.0024
Sitting	0.7931	0.7603	0.1687	0.0564	0.7912	0.7586	0.1626	0.0297
Standing	0.8320	0.8253	0.1362	0.0128	0.8298	0.8232	0.1300	0.0559
Walking	0.9637	0.9637	0.1665	0.0851	0.9591	0.9591	0.1582	0.0815

the purpose of validating our findings. It's important to note that we have meticulously maintained consistency in the number of features compared to the original dataset. This approach allows us to utilize the same model and assess its performance across diverse scenarios, thus enhancing the robustness of our analysis. The results presented in Table 4.8 and 4.9 illustrate the performance of the same model on this random dataset. It is noteworthy that the fairness metric already exhibits commendable results even before implementing fairness-enhancing measures. This is attributed to the well-balanced distribution of data between genders within this dataset. Furthermore, our utilization of the Lagrangian method has further refined and improved the overall fairness of the model which justifies the model performance.

Table 4.8: Unfair MLP Model Performance Metrics for Random Data

Activities	MLP F1 Score	MLP Test F1 Score	EO Gap	DP Gap
Standing	0.9745	0.9131	0.2563	0.0183
Walking	0.9863	0.9405	0.3198	0.1079
Sitting	0.9773	0.9924	0.2923	0.1274
Running	0.8892	0.9364	0.1984	0.0095

	acc_rf_x_amin	acc_rf_x_amax	acc_rf_x_mean	acc_rf_x_median	acc_rf_x_std	acc_rf_x_skew	acc_rf_x_kurtosis	acc_rf_x_iqr	acc_rf_x_median_abs_deviatio
0	-8296.0	-7676.0	-8053.322835	-8056.0	61.159287	1.616655	11.990163	56.0	28.
1	-8108.0	-7940.0	-8017.750000	-8016.0	34.296605	-0.118592	-0.341147	44.0	24.
2	-8124.0	-7940.0	-8017.840000	-8020.0	33.583729	0.042304	0.145430	44.0	20.
3	-8084.0	-7928.0	-8009.340909	-8012.0	34.400294	0.065610	-0.646873	49.0	28.
4	-8108.0	-7928.0	-8013.340909	-8012.0	34.474632	-0.054829	-0.243237	45.0	24.
...	...	...	...	...	...	...	...	...	.
1669	-32768.0	10040.0	-14631.545455	-10758.0	10108.926069	-0.505906	-0.338865	10388.0	3088.
1670	-32768.0	32767.0	-13615.823864	-10594.0	10708.265961	0.180210	1.952196	8061.0	2010.
1671	-32768.0	15348.0	-13875.446541	-10664.0	10096.585677	-0.358160	0.330676	8020.0	2128.
1672	-32768.0	16904.0	-13679.363636	-10472.0	9699.671476	-0.591397	0.329140	7259.0	2242.
1673	-32768.0	32767.0	-14455.073864	-10964.0	10463.363505	0.080585	1.919679	10598.0	2964.

1674 rows × 797 columns

f_gyro_median_abs_deviation_lag2_logabs	long_rf_gyro_mad_lag1_logabs	long_rf_gyro_mad_lag2_logabs	long_rf_gyro_ran_lag1_logabs	long_rf_gyro_ran_lag2_logabs	activity	sex_b
0.953328	0.947707	0.941931	1.946823	1.912436	Sitting	1
0.743550	0.196823	0.947707	1.163510	1.946823	Sitting	1
0.116916	0.082761	0.196823	0.436059	1.163510	Sitting	1
0.071772	0.085889	0.082761	0.489152	0.436059	Sitting	1
0.074737	0.097568	0.085889	0.545458	0.489152	Sitting	1
...	...	...	...	...	...	...
0.671356	0.818865	0.789996	1.904762	1.887446	Walking	0
0.801788	0.858748	0.818865	1.981771	1.904762	Walking	0
0.886432	0.818066	0.858748	1.934603	1.981771	Walking	0
0.812973	0.848431	0.818066	1.892169	1.934603	Walking	0
0.758145	0.863907	0.848431	1.961984	1.892169	Walking	0

Figure 4.2: Synthetic Data

Table 4.9: Fair MLP with Lagrangian Multiplier Performance Comparison for Randomly Generated Data

Activities	DP Loss + Lagrangian				EO Loss + Lagrangian			
	F1	Acc	EO Gap	DP Gap	F1	Acc	EO Gap	DP Gap
Running	0.8583	0.9264	0.0579	0.0152	0.8832	0.9064	0.0436	0.0139
Sitting	0.9719	0.9802	0.1051	0.0314	0.9256	0.8931	0.0785	0.0285
Standing	0.9320	0.9839	0.1002	0.0081	0.9298	0.9732	0.1013	0.0059
Walking	0.9671	0.9701	0.1665	0.0051	0.9803	0.9912	0.1291	0.0037

From the graph presented in Figure 4.3, it becomes evident that the performance of the fair models solely trained with DP loss or EO loss exhibits a certain degree of fluctuation when transitioning from the validation set to the test set. However, it's worth noting that the model enhanced with a Lagrangian multiplier, particularly when combined with DP loss, demonstrates notable improvements in both the validation and test sets. This enhancement is particularly pronounced in reducing the fairness gap. In this context, a larger fairness gap corresponds to lower levels of fairness. This observed trend underscores the effectiveness of incorporating the Lagrangian multiplier, especially in conjunction with DP loss, in achieving more robust fairness outcomes across different datasets.

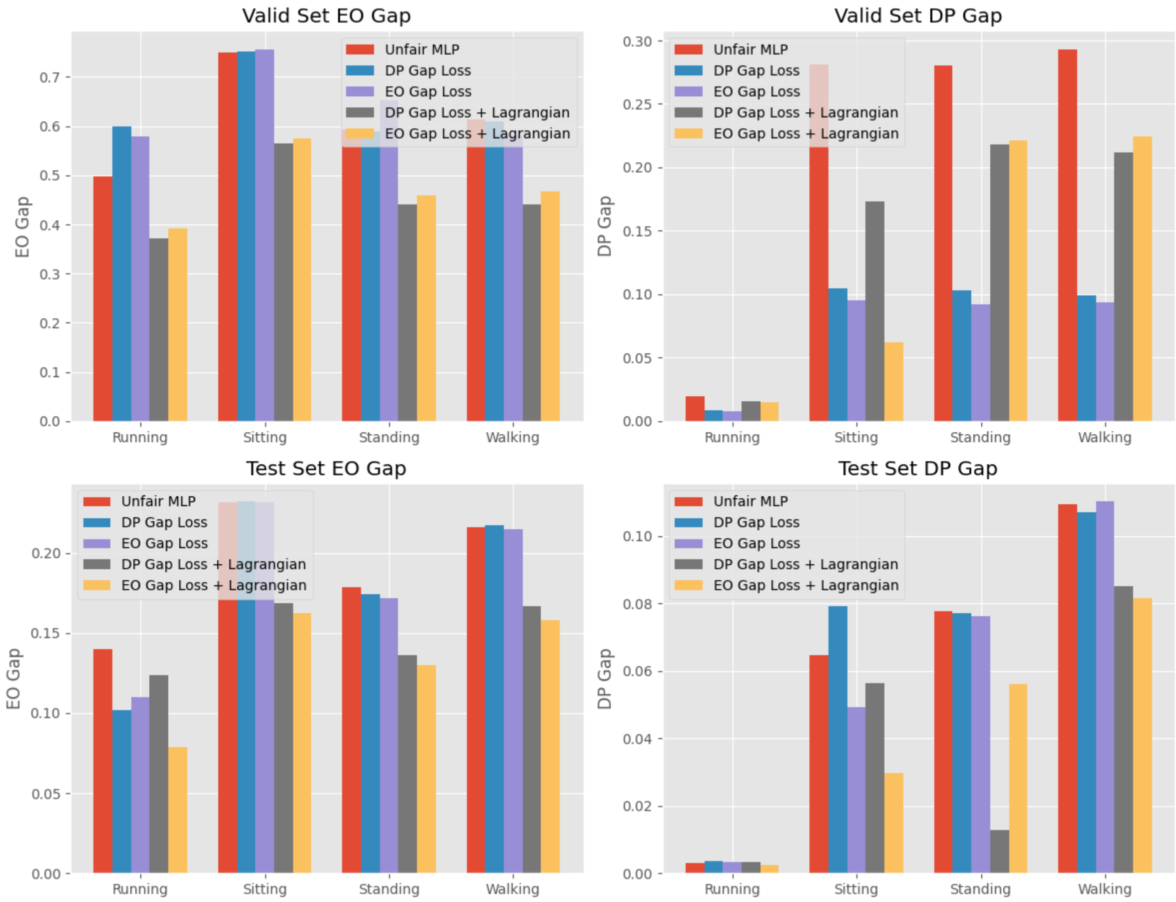

Figure 4.3: Fairness Gap Comparison Graph for Fair and Unfair MLP Models

4.3 Discussion

Deep learning plays a vital role in comprehending and applying fairness metrics within the context of automated decision-making. It serves as a robust tool for analyzing detailed patterns in data, similar to a magnifying glass enhancing observation. In this study, the Multilayer Perceptron (MLP) technique was employed to identify human activities, providing a foundational framework for subsequent fairness considerations. Fairness metrics, such as the Equalized Odds Gap and Demographic Parity Gap, are like special rules we use to check if our decisions are fair. By interpreting the fairness metrics employed, we gain a deeper understanding of the biases and inequities ingrained in the human activity recognition model. It becomes evident that achieving fairness often involves trade-offs with model performance, and a nuanced analysis of these trade-offs sheds light on the intricate balance between accuracy and equity. In considering the real-world applicability of the proposed fairness solutions, we recognize the need for adaptable strategies to navigate dynamic scenarios. Ethical considerations emerge as a significant facet, prompting us to ponder the trade-offs between fairness and potential consequences in domains like healthcare and criminal justice. Moreover, the discussion extends beyond gender-based fairness, delving into the intersectionality of demographic attributes in the context of human activity recognition. However, it is crucial to acknowledge the limitations inherent in this study, which influence the breadth and applicability of its findings. The representative nature of the dataset employed plays a pivotal role in determining the success of the proposed fairness-aware techniques. If the dataset lacks diversity or fails to encompass a wide range of gender-related activities and contexts, the effectiveness of the fairness solutions might not extend to real-world scenarios. While we're mainly looking at fairness related to gender, it's worth noting that recognizing human activities involves many different things about people, such as age, origin, and individual characteristics. These factors collectively impact how computers interpret and describe people's actions. Due to limited dataset availability, we couldn't explore these aspects completely, but with a more detailed dataset, there's potential to delve into this further.

Chapter 5

Conclusion

Within the realm of machine learning, particularly in the sphere of gender-based human activity recognition, the pursuit of fairness emerges as an imperative concern. This study delves comprehensively into the intricate challenges surrounding equity in technology, acknowledging that mere technical accuracy falls short in ensuring impartial outcomes. The biases entrenched within algorithms have the potential to perpetuate gender-biased assumptions, thereby giving rise to disparate and potentially prejudiced results. As technology continues its rapid evolution, our responsibility transcends achieving technical excellence; it extends to utilizing that excellence as a foundation for constructing a future that is both equitable and just. This study's significance lies in its emphasis on integrating fairness into technology, especially within the domain of gender-focused human activity recognition. The research confronts ethical apprehensions associated with biased machine learning models and accentuates the need to transcend the boundaries of technical accuracy, addressing gender-related assumptions and biases head-on. Driven by ethical considerations and grounded in practical implications, the study bridges an existing research gap by presenting tailored strategies to augment fairness within automated systems. By introducing pioneering fairness metrics and innovative loss functions, while harnessing the power of Lagrangian multipliers, this research contributes substantively to the advancement of more equitable machine learning techniques for human activity recognition. The insights garnered through this study serve to illuminate a pathway towards AI systems that mirror the rich tapestry of human experiences, thereby fostering a society that is more inclusive and egalitarian.

However, as individuals increasingly engage in concurrent tasks—like conversing while watching television—a distinct analytical approach is required to detect and comprehend such multifaceted actions, deviating from the methodologies applied to sequential activities. Further exploration in this direction might entail gathering a more diverse dataset in future investigations. Furthermore, modern researchers are progressively integrating federated learning techniques to safeguard privacy and equity in machine learning. These advancements hold promise for enhancing human activity recognition as well.

Bibliography

[1] D. S. Shrestha S, "Exploring gender biases in ml and ai academic research through systematic literature review," *Artif Intell*, 2022, PMID: 36304961. DOI: 10.3389/frai.2022.976838. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9593046/.

[2] F. Zuiderveen Borgesius *et al.*, "Discrimination, artificial intelligence, and algorithmic decision-making," *línea], Council of Europe*, 2018.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, Jul. 2021, ISSN: 0360-0300. DOI: 10.1145/3457607. [Online]. Available: https://doi.org/10.1145/3457607.

[4] Y. Chen, E. W. Clayton, L. L. Novak, S. Anders, and B. Malin, "Human-centered design to address biases in artificial intelligence," *Journal of Medical Internet Research*, vol. 25, e43251, 2023.

[5] L. Boratto, G. Fenu, M. Marras, and G. Medda, "Consumer fairness in recommender systems: Contextualizing definitions and mitigations," in *European Conference on Information Retrieval*, Springer, 2022, pp. 552–566.

[6] A. Cappozzo, U. Della Croce, A. Leardini, and L. Chiari, "Human movement analysis using stereophotogrammetry: Part 1: Theoretical background," *Gait & posture*, vol. 21, no. 2, pp. 186–196, 2005.

[7] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[8] K. Aminian, P. Robert, E. Buchser, B. Rutschmann, D. Hayoz, and M. Depairon, "Physical activity monitoring based on accelerometry: Validation and comparison with video observation," *Medical & biological engineering & computing*, vol. 37, pp. 304–308, 1999.

[9] O. C. Ann and L. B. Theng, "Human activity recognition: A review," in *2014 IEEE international conference on control system, computing and engineering (ICCSCE 2014)*, IEEE, 2014, pp. 389–393.

[10] X. Long, B. Yin, and R. M. Aarts, "Single-accelerometer-based daily physical activity classification," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 6107–6110.

[11] C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," *Sensors*, vol. 10, no. 8, pp. 7772–7788, 2010.

[12] M. J. Roshtkhari and M. D. Levine, “Human activity recognition in videos using a single example,” *Image and Vision Computing*, vol. 31, no. 11, pp. 864–876, 2013.

[13] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, “Evolutionary joint selection to improve human action recognition with rgb-d devices,” *Expert systems with applications*, vol. 41, no. 3, pp. 786–794, 2014.

[14] M. Zhang and A. A. Sawchuk, “Motion primitive-based human activity recognition using a bag-of-features approach,” in *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, 2012, pp. 631–640.

[15] E. Kańtoch and P. Augustyniak, “Human activity surveillance based on wearable body sensor network,” in *2012 computing in cardiology*, IEEE, 2012, pp. 325–328.

[16] E. S. Sazonov, G. Fulk, J. Hill, Y. Schutz, and R. Browning, “Monitoring of posture allocations and activities by a shoe-based wearable sensor,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 4, pp. 983–990, 2010.

[17] V. Q. Viet, G. Lee, and D. Choi, “Fall detection based on movement and smart phone technology,” in *Proceedings of the IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), Ho Chi Minh City, Vietnam*, vol. 27, 2012.

[18] B. P. Clarkson, “Life patterns: Structure from wearable sensors,” Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[19] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” *Pattern recognition letters*, vol. 119, pp. 3–11, 2019.

[20] A. Subasi, K. Khateeb, T. Brahimi, and A. Sarirete, “Human activity recognition using machine learning methods in a smart healthcare environment,” in *Innovation in health informatics*, Elsevier, 2020, pp. 123–144.

[21] C. Jobanputra, J. Bavishi, and N. Doshi, “Human activity recognition: A survey,” *Procedia Computer Science*, vol. 155, pp. 698–703, 2019.

[22] S. Mekruksavanich, A. Jitpattanakul, P. Youplao, and P. Yupapin, “Enhanced hand-oriented activity recognition based on smartwatch sensor data using lstms,” *Symmetry*, vol. 12, no. 9, p. 1570, 2020.

[23] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” *arXiv preprint arXiv:1707.09457*, 2017.

[24] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, “Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction,” in *Proceedings of the 2018 world wide web conference*, 2018, pp. 903–912.

[25] S. Hajian, F. Bonchi, and C. Castillo, “Algorithmic bias: From discrimination discovery to fairness-aware data mining,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 2125–2126.

[26] S. Vasudevan and K. Kenthapadi, "Lift: A scalable framework for measuring fairness in ml applications," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 2773–2780.

[27] N. A. Saxena, "Perceptions of fairness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 537–538.

[28] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–16.

[29] R. Berk, H. Heidari, S. Jabbari, *et al.*, "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*, 2017.

[30] L. Weinberg, "Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ml fairness approaches," *Journal of Artificial Intelligence Research*, vol. 74, pp. 75–109, 2022.

[31] H. F. Menezes, A. S. Ferreira, E. T. Pereira, and H. M. Gomes, "Bias and fairness in face detection," in *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, 2021, pp. 247–254.

[32] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods.* Academic press, 2014.

[33] L. Wang, "Recognition of human activities using continuous autoencoders with wearable sensors," *Sensors*, vol. 16, no. 2, p. 189, 2016.

[34] L. Wang, T. Gu, X. Tao, and J. Lu, "Toward a wearable rfid system for real-time activity recognition using radio patterns," *IEEE Transactions on Mobile Computing*, vol. 16, no. 1, pp. 228–242, 2016.

[35] T. Zebin, P. J. Scully, and K. B. Ozanyan, "Human activity recognition with inertial sensors using a deep learning approach," in *2016 IEEE sensors*, IEEE, 2016, pp. 1–3.

[36] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recognition*, vol. 43, no. 10, pp. 3605–3620, 2010.

[37] F. Chamroukhi, S. Mohammed, D. Trabelsi, L. Oukhellou, and Y. Amirat, "Joint segmentation of multivariate time series with hidden process regression for human activity recognition," *Neurocomputing*, vol. 120, pp. 633–644, 2013.

[38] M. Padala and S. Gujar, "Fnnc: Achieving fairness through neural networks," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization*, 2020.

[39] M. A. U. Alam, "Ai-fairness towards activity recognition of older adults," in *MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2020, pp. 108–117.

[40] M. Yuan, V. Kumar, M. A. Ahmad, and A. Teredesai, "Assessing fairness in classification parity of machine learning models in healthcare," *arXiv preprint arXiv:2102.03717*, 2021.

[41] A. F. Cruz, C. Belém, J. Bravo, P. Saleiro, and P. Bizarro, “Fairgbm: Gradient boosting with fairness constraints,” *arXiv preprint arXiv:2209.07850*, 2022.

[42] R. Chereshnev and A. Kertész-Farkas, “Hugadb: Human gait database for activity recognition from wearable inertial sensor networks,” in *Analysis of Images, Social Networks and Texts: 6th International Conference, AIST 2017, Moscow, Russia, July 27–29, 2017, Revised Selected Papers 6*, Springer, 2018, pp. 131–141.

[43] J.-Y. Yang, J.-S. Wang, and Y.-P. Chen, “Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers,” *Pattern recognition letters*, vol. 29, no. 16, pp. 2213–2220, 2008.

[44] O. D. Lara and M. A. Labrador, “A survey on human activity recognition using wearable sensors,” *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[45] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[46] T. Danka. “How t-sne works.” Accessed on: Aug 21, 2023. (2023), [Online]. Available: https://tivadardanka.com/blog/how-tsne-works.

[47] A. DeMaris, “A tutorial in logistic regression,” *Journal of Marriage and the Family*, pp. 956–968, 1995.

[48] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors),” *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[49] P. H. Swain and H. Hauska, “The decision tree classifier: Design and potential,” *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[50] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artificial Intelligence Review*, vol. 39, pp. 261–283, 2013.

[51] H. Ramchoun, Y. Ghanou, M. Ettaouil, and M. A. Janati Idrissi, “Multilayer perceptron: Architecture optimization and training,” 2016.

[52] L. Noriega, “Multilayer perceptron tutorial,” *School of Computing. Staffordshire University*, vol. 4, no. 5, p. 444, 2005.

[53] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, “Chexclusion: Fairness gaps in deep chest x-ray classifiers,” in *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, World Scientific, 2020, pp. 232–243.

[54] Y. Jin and L. Lai, “Adversarially robust fairness-aware regression,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[55] Y. Zhao, Y. Wang, and T. Derr, “Fairness and explainability: Bridging the gap towards fair model explanations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 11 363–11 371.

[56] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, "The road to explainability is paved with bias: Measuring the fairness of explanations," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1194–1206.

[57] A. Mishler, E. H. Kennedy, and A. Chouldechova, "Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 386–400.

[58] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness gan: Generating datasets with fairness properties using a generative adversarial network," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 3–1, 2019.

Appendix A

The box plots were utilized to visualize the distribution and variability of each feature across different activities. By creating box plots for each feature with respect to all activities, we were able to gain insights into how the data was spread out and the presence of potential outliers or trends. This graphical representation from Figure 1 to Figure 12 allowed us to observe the central tendency (median), spread (interquartile range), and potential skewness or symmetry of the data distribution for each feature within the context of various activities.

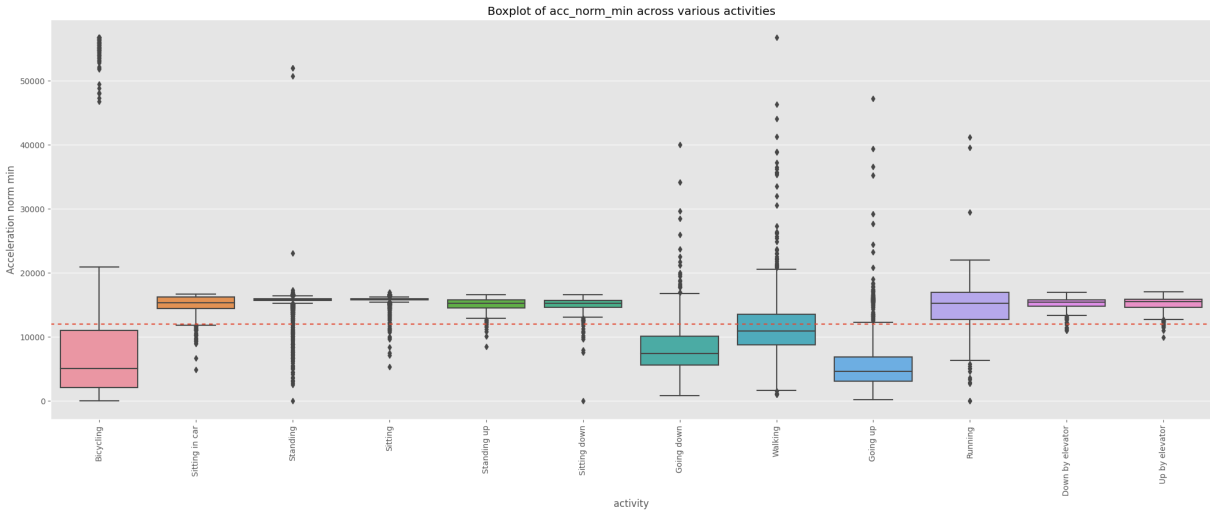

Figure 1: Boxplot of the normalized minimum acceleration (`acc_norm_min`) for all activities

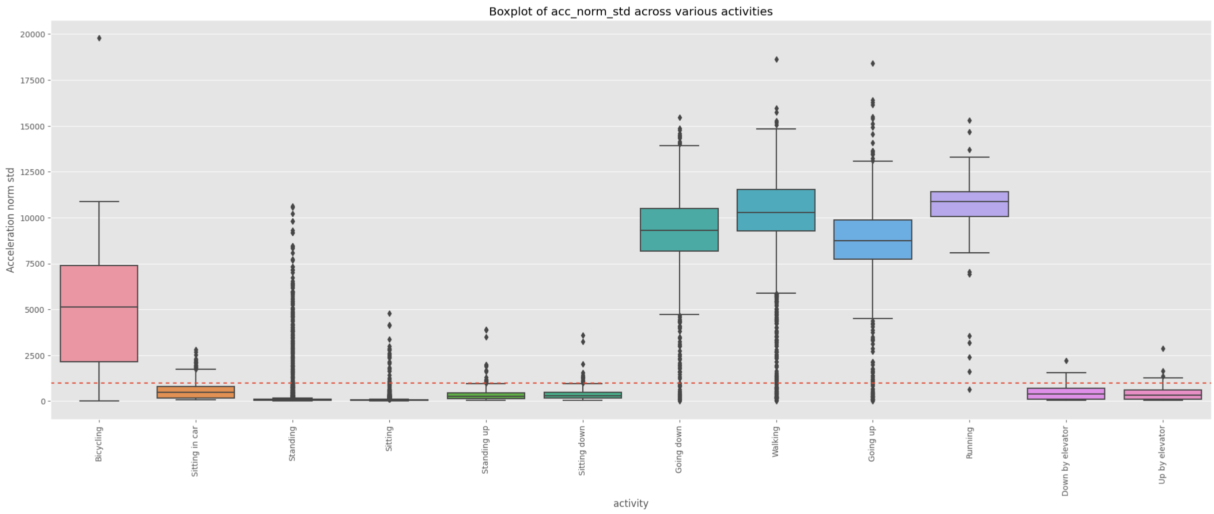

Figure 2: Boxplot of the standard deviation of normalized acceleration (`acc_norm_std`) for all activities

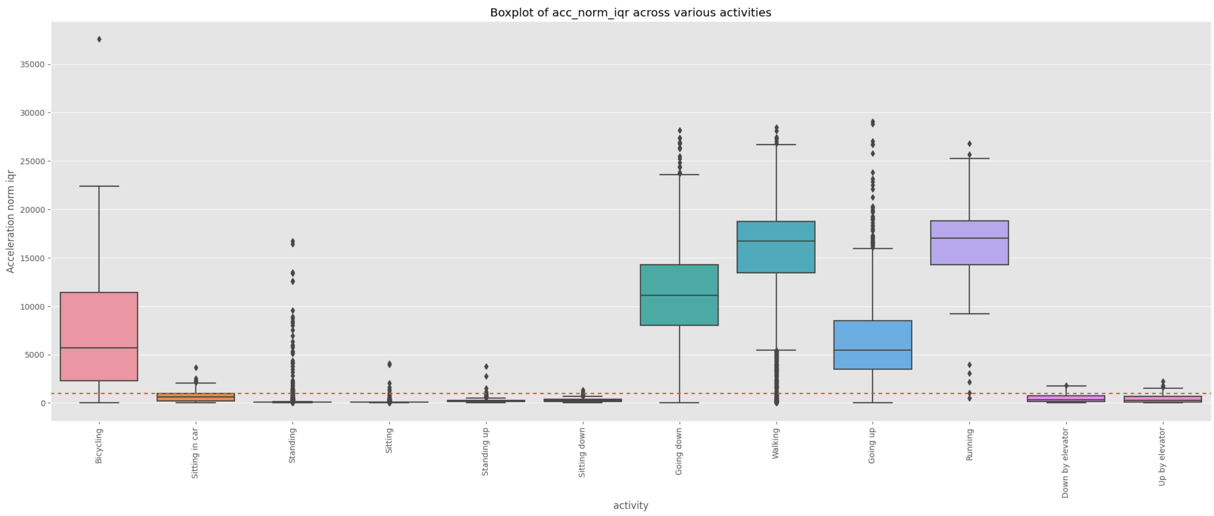

Figure 3: Boxplot of the interquartile range of normalized acceleration (`acc_norm_iqr`) for all activities

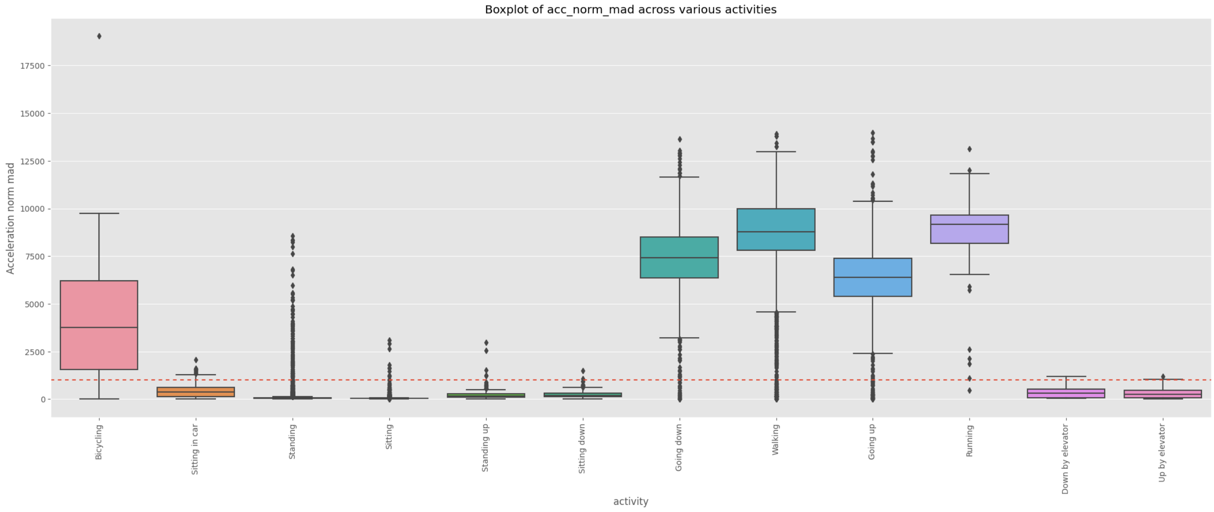

Figure 4: Boxplot of the median absolute deviation of normalized acceleration (`acc_norm_mad`) for all activities

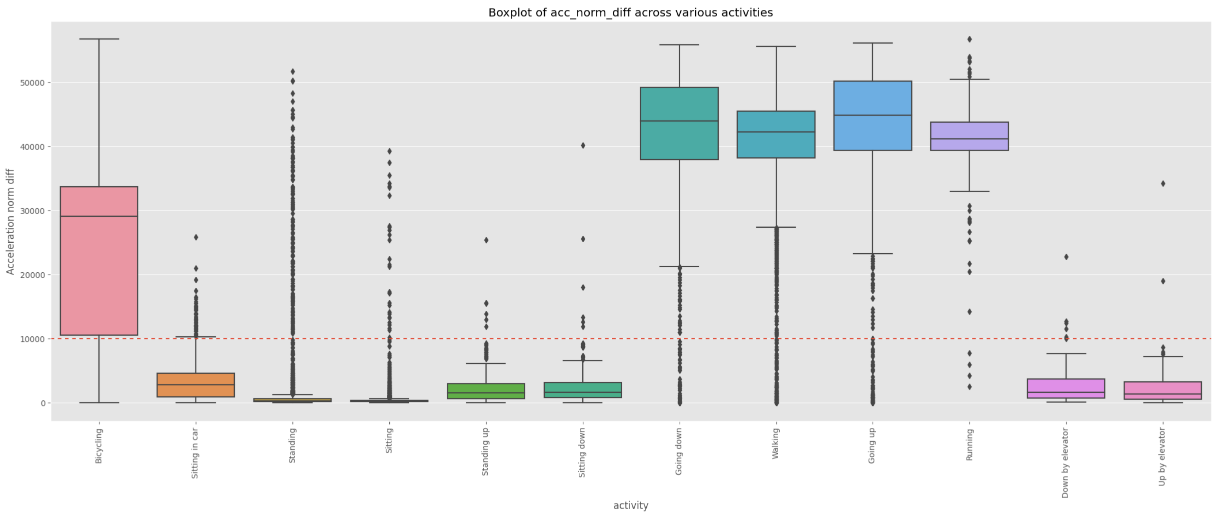

Figure 5: Boxplot of the difference in normalized acceleration (`acc_norm_diff`) for all activities

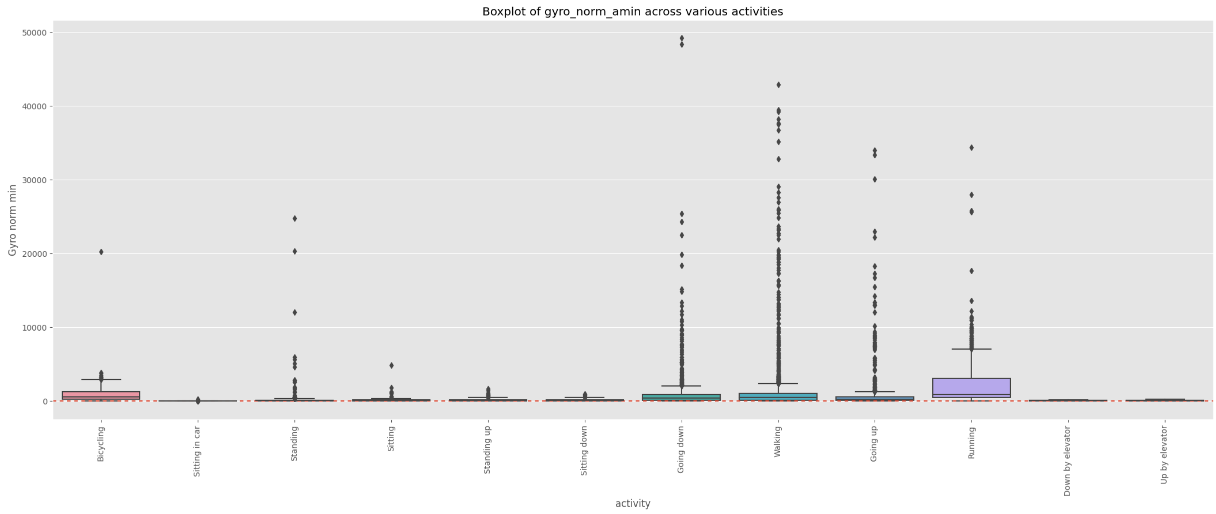

Figure 6: Boxplot of the minimum normalized gyroscope reading (`gyro_norm_amin`) for all activities

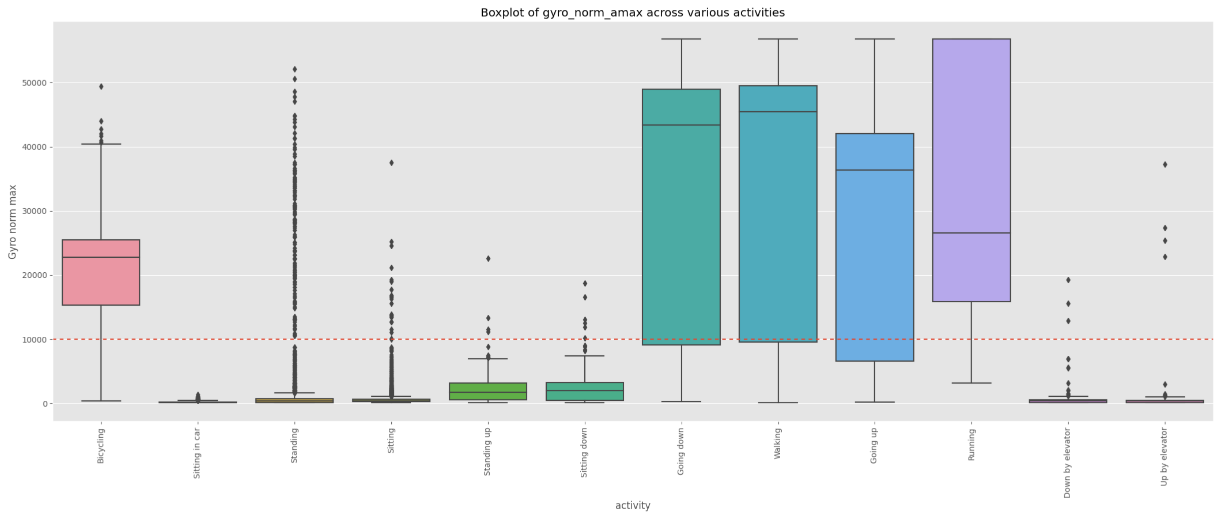

Figure 7: Boxplot of the maximum normalized gyroscope reading (`gyro_norm_amax`) for all activities

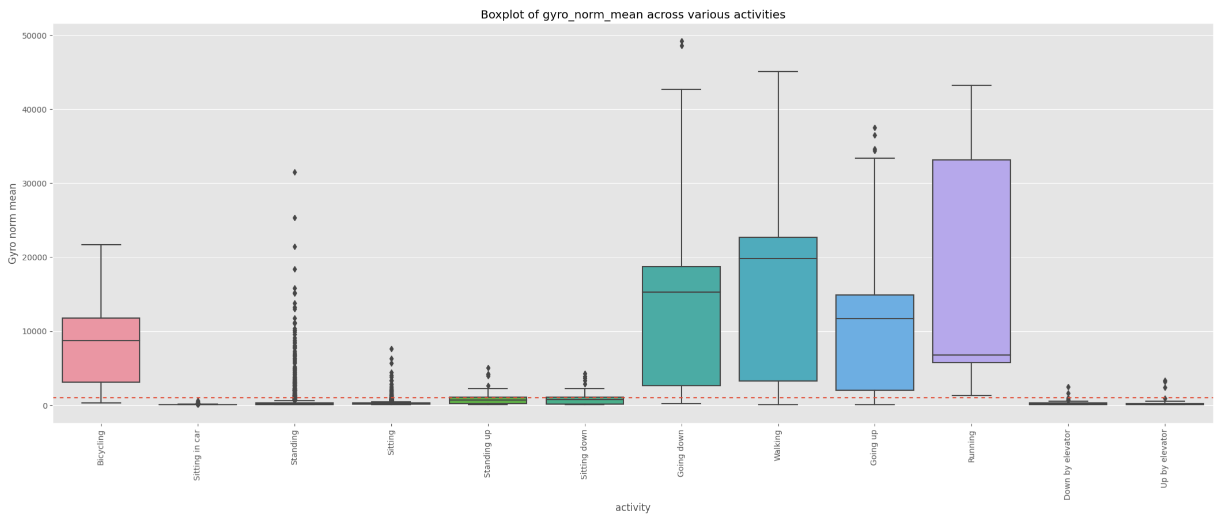

Figure 8: Boxplot of the mean normalized gyroscope reading (`gyro_norm_mean`) for all activities

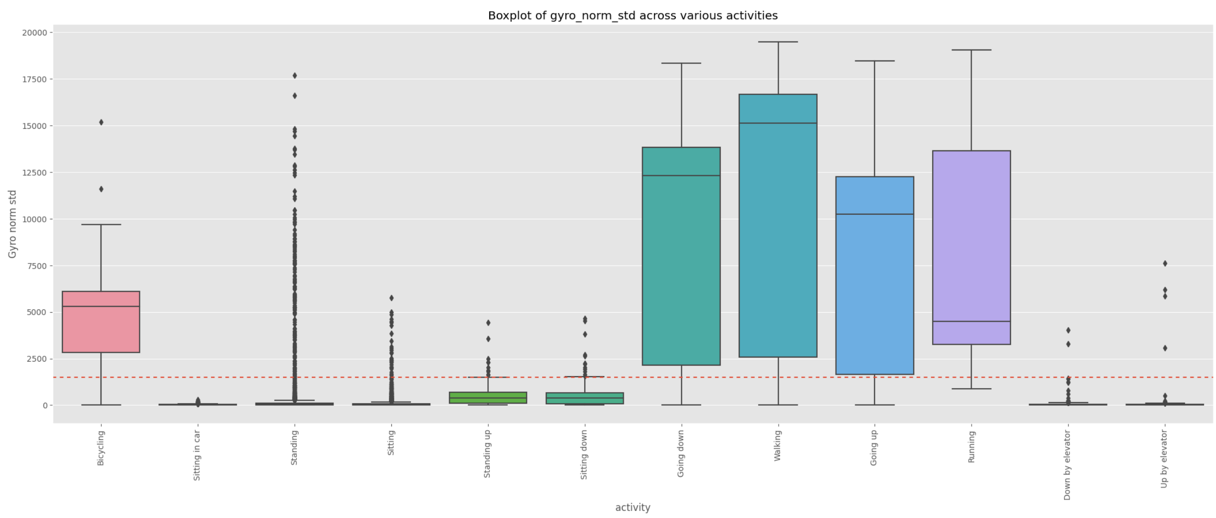

Figure 9: Boxplot of the standard deviation of normalized gyroscope reading (`gyro_norm_std`) for all activities

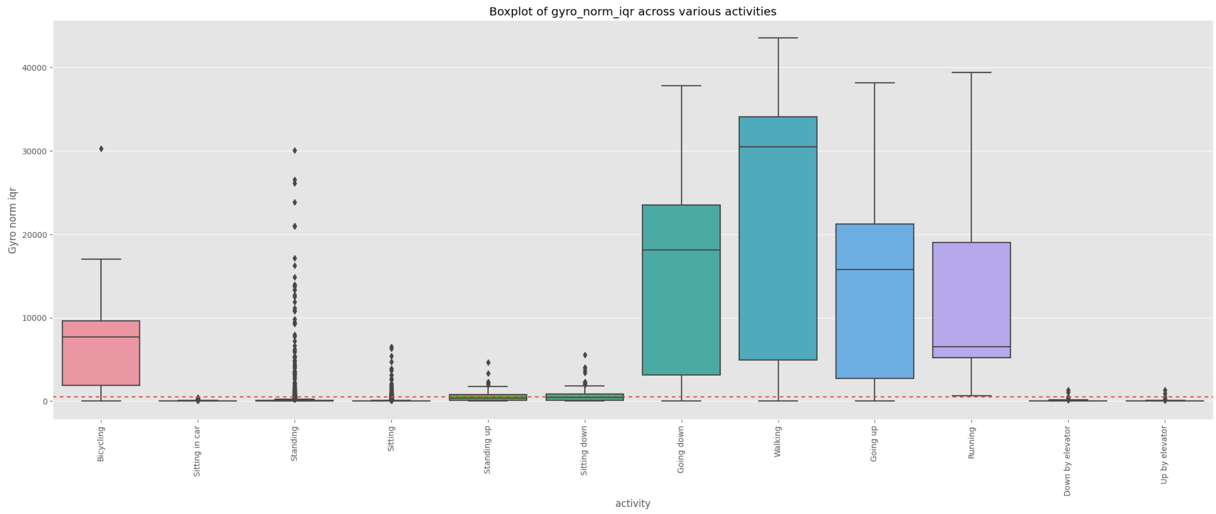

Figure 10: Boxplot of the interquartile range of normalized gyroscope reading (`gyro_norm_iqr`) for all activities

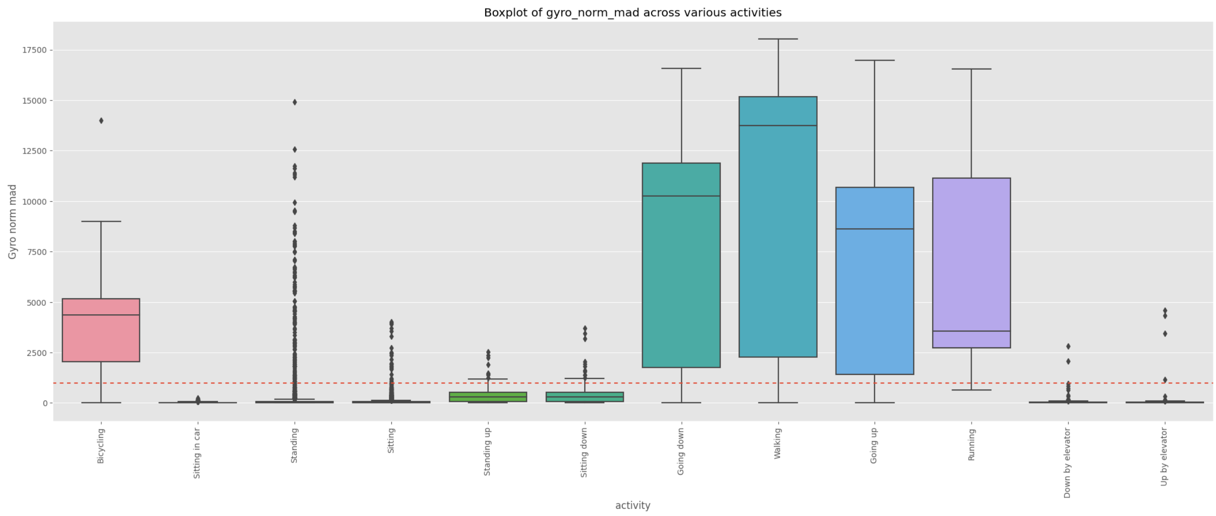

Figure 11: Boxplot of the median absolute deviation of normalized gyroscope reading (`gyro_norm_mad`) for all activities

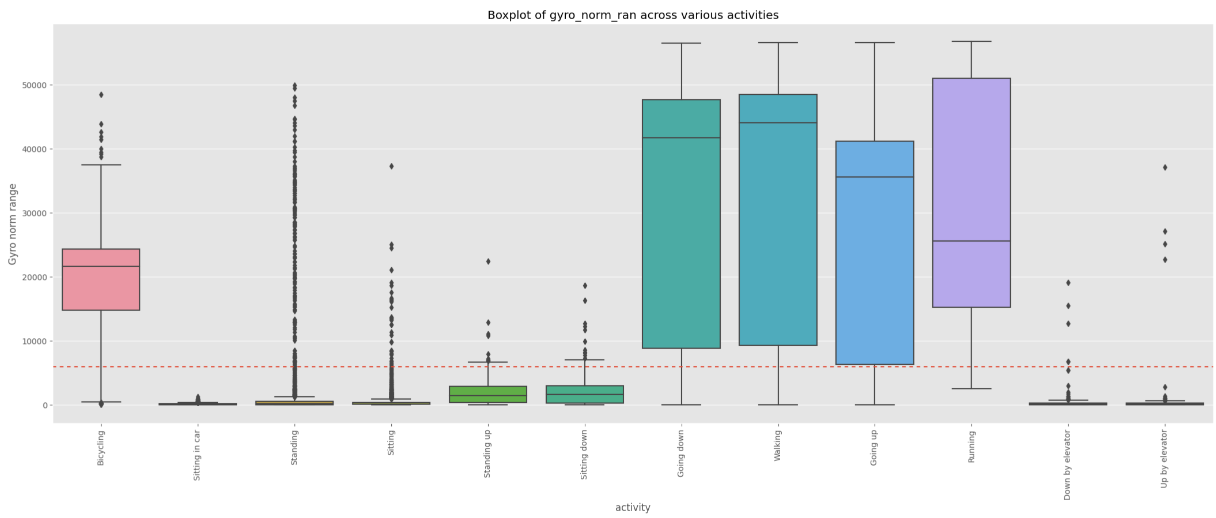

Figure 12: Boxplot of the range of normalized gyroscope reading (`gyro_norm_ran`) for all activities

Appendix B

Figure 13 showcases a condensed representation extracted from a larger dataset comprising 16,038 rows and 792 columns. Each row in this table corresponds to a distinct data entry, while the columns encapsulate an assortment of sensor measurements and derived statistical attributes.
Below, we provide a succinct explanation of several columns found within the table:

- `acc_rf_x_amin`: Represents the minimum recorded value along the x-axis of the accelerometer.
- `acc_rf_x_amax`: Signifies the maximum recorded value along the x-axis of the accelerometer.
- `acc_rf_x_mean`: Denotes the average of readings along the x-axis of the accelerometer.
- `acc_rf_x_median`: Conveys the central value within the readings along the x-axis of the accelerometer.
- `acc_rf_x_std`: Depicts the degree of deviation within the readings along the x-axis of the accelerometer.
- `acc_rf_x_skew`: Illustrates the asymmetry in the distribution of readings along the x-axis of the accelerometer.
- `acc_rf_x_kurtosis`: Portrays the peakedness of the distribution of readings along the x-axis of the accelerometer.
- `acc_rf_x_iqr`: Conveys the range between the first and third quartiles of readings along the x-axis of the accelerometer.
- `acc_rf_x_median_abs_deviation`: Reflects the median-based measure of dispersion along the x-axis of the accelerometer.
- `acc_rf_x_mad`: Represents the mean absolute deviation of readings along the x-axis of the accelerometer.
- ... (as well as other columns)

This table provides a glimpse into the data, focusing on a subset of columns that encompass a diverse range of statistical indicators and sensor-specific insights. These data hold potential significance in various analytical endeavors, including feature engineering, machine learning, and statistical assessments. It is important to emphasize that the table offers only a snapshot of the comprehensive dataset, serving as an overview of its contents and inherent characteristics.

	acc_rf_x_amin	acc_rf_x_amax	acc_rf_x_mean	acc_rf_x_median	acc_rf_x_std	acc_rf_x_skew	acc_rf_x_kurtosis	acc_rf_x_iqr	acc_rf_x_median_abs_deviatio
0	-32768.0000	13436.0000	-15812.8864	-16014.0000	9246.7642	0.3236	0.1420	10443.0000	5310.00
1	-32768.0000	5916.0000	-16043.0227	-18574.0000	10097.3568	0.3187	-1.0803	16908.0000	7404.00
2	-32768.0000	3848.0000	-16569.9773	-18050.0000	10135.6543	0.3297	-1.1142	16818.0000	8156.00
3	-32768.0000	10504.0000	-15919.2045	-17944.0000	12044.0974	0.3049	-1.1567	21773.0000	10296.00
4	-21824.0000	-2896.0000	-13594.0682	-13358.0000	3998.4549	-0.0047	-0.6304	6029.0000	3194.00
...	...	...	...	...	...	...	...	...	
16033	-32768.0000	32767.0000	-10889.5170	-6222.0000	12947.7938	0.0811	1.5188	10885.0000	2170.00
16034	-32768.0000	17900.0000	-11438.4318	-5808.0000	11808.3831	-0.6562	-0.4401	16035.0000	3166.00
16035	-32768.0000	32767.0000	-12348.6648	-6264.0000	12253.2885	-0.3736	0.1002	18140.0000	2800.00
16036	-32768.0000	28580.0000	-10914.5909	-5932.0000	12068.2509	-0.3746	0.4922	10939.0000	2074.00
16037	-32768.0000	18368.0000	-11001.9091	-5528.0000	11302.4587	-0.8506	-0.1145	10146.0000	2028.00

16038 rows × 792 columns

long_rf_gyro_median_abs_deviation_lag1_logabs	long_rf_gyro_median_abs_deviation_lag2_logabs	long_rf_gyro_mad_lag1_logabs	long_rf_gyro_mad_lag2_logabs	long_rf
0.0000	0.0000	0.0000	0.0000	
0.5042	0.0000	0.9454	0.0000	
0.8436	0.5042	0.9853	0.9454	
1.0245	0.8436	0.9911	0.9853	
1.0812	1.0245	0.9828	0.9911	
...	...	...	...	
0.6930	1.0172	1.0062	1.0150	
0.9172	0.6930	0.9631	1.0062	
1.0017	0.9172	1.0128	0.9631	
0.8870	1.0017	0.9978	1.0128	
0.6030	0.8870	0.9430	0.9978	

Figure 13: Selected Features from the Dataset

```
CPU times: user 2 µs, sys: 1e+03 ns, total: 3 µs
Wall time: 5.48 µs
[t-SNE] Computing 151 nearest neighbors...
[t-SNE] Indexed 16038 samples in 0.198s...
[t-SNE] Computed neighbors for 16038 samples in 4.399s...
[t-SNE] Computed conditional probabilities for sample 1000 / 16038
[t-SNE] Computed conditional probabilities for sample 2000 / 16038
[t-SNE] Computed conditional probabilities for sample 3000 / 16038
[t-SNE] Computed conditional probabilities for sample 4000 / 16038
[t-SNE] Computed conditional probabilities for sample 5000 / 16038
[t-SNE] Computed conditional probabilities for sample 6000 / 16038
[t-SNE] Computed conditional probabilities for sample 7000 / 16038
[t-SNE] Computed conditional probabilities for sample 8000 / 16038
[t-SNE] Computed conditional probabilities for sample 9000 / 16038
[t-SNE] Computed conditional probabilities for sample 10000 / 16038
[t-SNE] Computed conditional probabilities for sample 11000 / 16038
[t-SNE] Computed conditional probabilities for sample 12000 / 16038
[t-SNE] Computed conditional probabilities for sample 13000 / 16038
[t-SNE] Computed conditional probabilities for sample 14000 / 16038
[t-SNE] Computed conditional probabilities for sample 15000 / 16038
[t-SNE] Computed conditional probabilities for sample 16000 / 16038
[t-SNE] Computed conditional probabilities for sample 16038 / 16038
[t-SNE] Mean sigma: 1640.986398
[t-SNE] KL divergence after 250 iterations with early exaggeration: 73.369057
[t-SNE] KL divergence after 1000 iterations: 1.153913
```

Figure 14: Dimension Reduction using TSNE (t-Distributed Stochastic Neighbor Embedding) LOGS

Appendix C

Table 1 presents outcomes from the Unfair MLP model along with the associated fairness disparities observed across various human activity recognition activities in three folds. The F1 score and accuracy are reported for each activity, accompanied by the Equalized Odds (EO) gap and Demographic Parity (DP) gap. These performance metrics offer insights into both model effectiveness and the extent of fairness violation in terms of EO and DP gaps.

Table 1: Unfair MLP Results and Fairness Gap for Each Fold

Activities	Valid F1	Test F1	Valid acc	Test acc	Valid EO	Test EO	Valid DP	Test DP	Fold
Standing	0.8275	0.8014	0.8217	0.7928	0.6202	0.4227	0.3141	0.1857	1
Walking	0.9842	0.9655	0.9842	0.9655	0.6717	0.3554	0.3328	0.1757	1
Sitting	0.8518	0.8501	0.8341	0.8335	0.6235	0.3511	0.2153	0.1197	1
Running	0.9899	0.9946	0.9891	0.9946	0.6000	0.1500	0.0237	0.0038	1
Standing	0.8479	0.8520	0.8401	0.8458	0.5684	0.0953	0.2671	0.0185	2
Walking	0.9779	0.9762	0.9779	0.9762	0.5805	0.2487	0.2698	0.1244	2
Sitting	0.8674	0.8366	0.8595	0.8101	0.8004	0.0918	0.2946	0.0410	2
Running	0.9952	0.9928	0.9950	0.9927	0.4688	0.0390	0.0167	0.0000	2
Standing	0.8520	0.8470	0.8431	0.8423	0.5926	0.0181	0.2595	0.0287	3
Walking	0.9802	0.9833	0.9802	0.9833	0.5864	0.0440	0.2763	0.0281	3
Sitting	0.7944	0.7542	0.7793	0.7115	0.8269	0.2521	0.3328	0.0329	3
Running	0.9936	0.9957	0.9935	0.9958	0.4217	0.2308	0.0172	0.0048	3

Table 2 showcases outcomes of the Fair MLP model, with a specific emphasis on Equalized Odds (EO) loss. The presented data include F1 score, accuracy, EO gap, and DP gap for diverse activities across three folds. The Fair MLP model aims to enhance fairness by mitigating the EO gap while preserving strong performance, as evidenced by F1 scores and accuracy.

Table 2: MLP Results and Fairness Gap for Each Fold using EO Loss

Activities	Valid F1	Test F1	Valid acc	Test acc	Valid EO	Test EO	Valid DP	Test DP	Fold
Standing	0.8199	0.8023	0.8138	0.7936	0.6884	0.4124	0.1072	0.1827	1
Walking	0.9728	0.9555	0.9728	0.9555	0.6513	0.3523	0.1210	0.1719	1
Sitting	0.7752	0.8070	0.7422	0.7820	0.6505	0.3600	0.0864	0.1028	1
Running	0.9828	0.9931	0.9802	0.9931	0.8667	0.1500	0.0123	0.0046	1
Standing	0.8276	0.8476	0.8185	0.8412	0.6500	0.0865	0.0905	0.0159	2
Walking	0.9671	0.9589	0.9671	0.9590	0.6080	0.2500	0.0824	0.1343	2
Sitting	0.7734	0.7790	0.7559	0.7386	0.8140	0.0821	0.0986	0.0165	2
Running	0.9887	0.9861	0.9878	0.9854	0.5217	0.0260	0.0072	0.0020	2
Standing	0.8155	0.8418	0.8034	0.8369	0.6171	0.0163	0.0789	0.0305	3
Walking	0.9638	0.9791	0.9638	0.9791	0.4849	0.0416	0.0763	0.0239	3
Sitting	0.7940	0.7522	0.7789	0.7091	0.8036	0.2521	0.1000	0.0281	3
Running	0.9842	0.9952	0.9832	0.9952	0.3462	0.1538	0.0034	0.0030	3

Table 3 and Table 4 extend the Fair MLP results, incorporating the employment of the Lagrangian multiplier coupled with the DP Loss and EO loss. These tables detail F1 score, accuracy, EO gap, and DP gap for various activities across three folds. The Lagrangian multiplier acts as a parameter for balancing the trade-off between classification performance and fairness.

Table 3: Fair MLP Results using DP Loss and Lagrangian Multiplier

Activities	Valid F1	Test F1	Valid acc	Test acc	Valid EO	Test EO	Valid DP	Test DP	Fold
Standing	0.8232	0.8097	0.8173	0.8012	0.4924	0.3232	0.2545	0.0030	1
Walking	0.9783	0.9616	0.9783	0.9616	0.5097	0.2746	0.2469	0.1322	1
Sitting	0.8292	0.8542	0.8074	0.8396	0.4657	0.2325	0.0072	0.1275	1
Running	0.9921	0.9891	0.9921	0.9900	0.4615	0.1346	0.0133	0.0030	1
Standing	0.8267	0.8502	0.8176	0.8438	0.4272	0.0716	0.1999	0.0137	2
Walking	0.9617	0.9516	0.9617	0.9517	0.4466	0.1903	0.2096	0.1028	2
Sitting	0.7784	0.7802	0.7613	0.7406	0.6200	0.0892	0.0925	0.0020	2
Running	0.9931	0.9890	0.9928	0.9887	0.3606	0.0400	0.0152	0.0020	2
Standing	0.8295	0.8360	0.8185	0.8309	0.4545	0.0139	0.2086	0.0216	3
Walking	0.9625	0.9779	0.9625	0.9779	0.4482	0.0348	0.2158	0.0202	3
Sitting	0.7916	0.7451	0.7763	0.7007	0.6388	0.1844	0.0875	0.0395	3
Running	0.9895	0.9943	0.9897	0.9946	0.3522	0.1972	0.0153	0.0051	3

Table 4: Fair MLP Results using EO Loss and Lagrangian Multiplier

Activities	Valid F1	Test F1	Valid acc	Test acc	Valid EO	Test EO	Valid DP	Test DP	fold
Standing	0.8211	0.8039	0.8152	0.7955	0.4696	0.3093	0.2444	0.0807	1
Walking	0.9669	0.9536	0.9669	0.9536	0.4886	0.2621	0.2436	0.1256	1
Sitting	0.7748	0.8529	0.7427	0.8384	0.4752	0.2403	0.1928	0.0733	1
Running	0.9845	0.9881	0.9836	0.9881	0.4444	0.0926	0.0176	0.0028	1
Standing	0.8235	0.8445	0.8144	0.8382	0.4138	0.0673	0.2035	0.0804	2
Walking	0.9589	0.9533	0.9590	0.9533	0.4160	0.1852	0.1942	0.1020	2
Sitting	0.7718	0.7756	0.7545	0.7356	0.5984	0.0608	0.2553	0.0083	2
Running	0.9881	0.9859	0.9878	0.9857	0.3472	0.0289	0.0130	0.0025	2
Standing	0.8217	0.8408	0.8105	0.8361	0.4377	0.0134	0.2066	0.0066	3
Walking	0.9601	0.9705	0.9601	0.9705	0.4205	0.0273	0.1976	0.0168	3
Sitting	0.7890	0.7452	0.7739	0.7017	0.6177	0.1867	0.0725	0.0075	3
Running	0.9834	0.9896	0.9829	0.9896	0.3213	0.1140	0.0166	0.0018	3